Molecular Evolution and Phylogenetics

Masatoshi Nei
Pennsylvania State University

Sudhir Kumar
Arizona State University

OXFORD
UNIVERSITY PRESS
2000

OXFORD
UNIVERSITY PRESS

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Singapore Taipei Tokyo Toronto Warsaw

and associated companies in
Berlin Ibadan

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data

Nei, Masatoshi.
Molecular evolution and phylogenetics / by Masatoshi Nei, Sudhir Kumar.
p. cm.
Includes bibliographical references.
ISBN 0-19-513584-9; ISBN 0-19-513585-7 (pbk.)
1. Evolutionary genetics—Statistical methods. 2. Molecular evolution—Statistical methods.
I. Kumar, S. (Sudhir), 1958– . II. Title.
QH390.N45 2000
572.8'38—dc21 99-39160

9 8 7 6 5 4 3

Printed in the United States of America
on acid-free paper

Preface

> Statistics is a subject of amazingly many uses and surprisingly few effective practitioners. The traditional road to statistical knowledge is blocked, for most, by a formidable wall of mathematics. Our approach here avoids that wall.
>
> Efron and Tibshirani
> (1993)

In 1987, the first author of this book published a work entitled *Molecular Evolutionary Genetics*. The objective of this work was to unify two then separate disciplines of study of evolution at the molecular level: reconstruction of evolutionary histories of organisms and investigation of the mechanism of evolution. During the last ten years, remarkable progress has occurred in both of these research areas, and these two disciplines have now become inseparable. This progress has occurred partly because of the development of biochemical techniques such as PCR that allows rapid DNA sequencing. This technological innovation has generated an enormous amount of DNA sequence data from various groups of organisms and has enabled evolutionists to study molecular evolutionary genetics at an accelerated rate.

Another important factor that has contributed to this progress is the advance in statistical methods of data analysis and progress of computational technology. In recent years, various new methods for phylogenetic analysis and for studying the mechanism of evolution have been developed, and these methods have made the study of molecular evolution more rigorous and often simpler. At the same time, the improvement of high-speed personal computers has made it possible for many investigators to conduct refined statistical analyses of large-scale data.

The purpose of this book is different from that of *Molecular Evolutionary Genetics*. While the latter book was intended to present the framework of the emerging interdisciplinary science in terms of both mathematical theories and experimental observations, the objective of this book is to present statistical methods that are useful in the study of

molecular evolution and to illustrate how to use them in actual data analysis. We present various statistical methods dealing with recent evolutionary problems, but it is not our intention to discuss recent biological findings in molecular evolutionary genetics. These findings have recently been reviewed by Avise (1994), Hartl and Clark (1997), Li (1997), Powell (1997), and Graur and Li (1999), so that the reader may refer to them. In this book, however, we present many numerical examples in which interesting biological problems are discussed. We also present basic information on molecular evolution, particularly in chapters 1 and 10, so that the reader can comprehend the biological issues dealt with in this book. Information on geological time scales that are useful in the study of molecular evolution is presented in the appendices.

This book is written for graduate students and researchers in the field of molecular evolution. We assume that the reader has basic knowledge of molecular biology and evolution as well as of elementary statistics. Although this book is about statistical methods used in molecular evolutionary genetics, it is not intended to present mathematical foundations of statistical methods but rather to show how to use them with a minimum amount of mathematics. Nowadays almost all data analyses are conducted by using computers. Therefore, our explanation of statistical methods is presented with the understanding of this current practice.

Previously we published a computer program package called MEGA (Kumar et al. 1993). This program is no longer up to date, so we have revised it (Kumar et al. 2000). This revised version (MEGA Version 2.0; MEGA2) was used in generating many numerical examples given in this book. We therefore recommend that the readers try to reproduce the results of each data analysis using this program and learn details of the statistical methods presented. This program is available from the website of this book, *http://www.oup-usa.org/sc/0195135857.* Some parts of the data analyses in this book were conducted by PAUP* (Swofford 1998), PHYLIP (Felsenstein 1995), MOLPHY (Adachi and Hasegawa 1996b), PAML (Yang 1995b, 1999), and various computer programs available from the website, *http://mep.bio.psu.edu.* We have found PAUP* particularly useful for phylogenetic analysis of DNA sequences by parsimony and likelihood methods. All the original data used in the numerical examples and other information required for the computation are presented in the website for this book, which also includes some additional numerical examples.

In this book, we first discuss the statistical methods for analyzing protein and DNA sequence data and then present methods for analyzing allele frequency data. We have included various new methods developed in our laboratory as well as in others. Some of the methods and their statistical properties were investigated in the process of writing this book and have been published very recently or for the first time here. Our criterion of inclusion was practical utility rather than mathematical innovation. Preference was also given to statistical methods that are based on realistic biological assumptions. Because the laboratory of the first author has been engaged in this area of study for the last 30 years and some topics have been studied primarily in his laboratory, this book includes a substantial amount of work conducted by his group. In this book, we

included more methods for sequence data than for allele frequency data, simply because sequence data are now used more often than the latter, and many of the latter methods are discussed in the previous book, *Molecular Evolutionary Genetics*. Any reader who wishes to know detailed aspects of statistical methods for analyzing classical allele frequency data may refer to it.

We have used many numerical examples to illustrate data analysis. These examples are used to show not only the computational process but also how to extract biological information from data analysis. Note that the results of the analysis presented are not always the same as those given in the papers from which the original data were cited, because all the data were reanalyzed by ourselves to suit the purpose of this book. This is true even with the examples that were used in our previous papers. To avoid any serious errors, we have chosen numerical examples from among those with which we are relatively well acquainted. For this reason, we are afraid that we have not done justice to many authors who have published interesting biological findings. If this book is to be used in classrooms, we recommend that the instructors choose their own examples, possibly from those obtained in their laboratory. They may also use supplementary materials concerning some theoretical aspects that are not covered here.

This book is a product of collaboration of the two authors for the last four years. The first author (M. N.) is primarily responsible for choosing subject matters and writing the text, and the second author (S. K.) took charge of finding proper numerical examples and data analysis. The second author is also largely responsible for developing the computer software MEGA2, which was used in producing many numerical examples.

However, we could not have written this book without the help of our former and current collaborators in the laboratory of the first author. They have made a concerted effort to solve many challenging problems in statistical study of molecular evolution during the last ten years. Among them, we are especially indebted to Andrey Rzhetsky, Tatsuya Ota, Naoko Takezaki, Koichiro Tamura, Ziheng Yang, Claudia Russo, Tanya Sitnikova, Jianzhi (George) Zhang, and Xun Gu. In addition, such visiting scholars as Naoyuki Takahata, Willem Ferguson, Fumio Tajima, Yoshio Tateno, and Joaquin Dopazo have made significant contributions. We also thank members of the Institute of Molecular Evolutionary Genetics at the Pennsylvania State University, who raised various questions and stimulated our research. We are grateful to many friends and colleagues who read initial drafts of various chapters of this book and provided valuable comments. They include Tom Dowling, Alan Filipski, Rodney Honeycutt, Phil Hedrick, Junhyong Kim, Andrey Rzhetsky, Naruya Saitou, James Lyons-Weiler, Mike Miyamoto, Alex Rooney, Ziheng Yang, George Zhang, and Marcy Uyenoyama. Special thanks go to Sudhindra Gadagkar, Ingrid Jakobsen, and Thomas Whittam, who read almost the entire manuscript and made valuable suggestions to improve the presentation of the book. We are extremely grateful to Joyce White, who patiently typed many versions of the manuscript and helped us in organizing the reference list. Our thanks also go to Barb Backes, who drew the final versions of illustrations.

This work was partly supported by research grants from the National Institutes of Health and the National Science Foundation to M. N. Part of this book was written while M. N. was on a sabbatical leave at the National Institute of Genetics, Mishima, Japan, under the sponsorship of the Japan Society of Promotion of Science (host: Takashi Gojobori). We are grateful to these funding agencies for their generous support.

M. N.
S. K.

Contents

Numerical Examples xiii

1 Molecular Basis of Evolution 3
- 1.1. Evolutionary Tree of Life 3
- 1.2. Mechanism of Evolution 4
- 1.3. Structure and Function of Genes 5
- 1.4. Mutational Changes of DNA Sequences 9
- 1.5. Codon Usage 11

2 Evolutionary Change of Amino Acid Sequences 17
- 2.1. Amino Acid Differences and Proportion of Different Amino Acids 17
- 2.2. Poisson Correction (PC) and Gamma Distances 19
- 2.3. Bootstrap Variances and Covariances 25
- 2.4. Amino Acid Substitution Matrix 27
- 2.5. Mutation Rate and Substitution Rate 29

3 Evolutionary Change of DNA Sequences 33
- 3.1. Nucleotide Differences Between Sequences 33
- 3.2. Estimation of the Number of Nucleotide Substitutions 35
- 3.3. Gamma Distances 43
- 3.4. Numerical Estimation of Evolutionary Distances 45
- 3.5. Alignment of Nucleotide Sequences 46
- 3.6. Handling of Sequence Gaps in the Estimation of Evolutionary Distances 49

4 Synonymous and Nonsynonymous Nucleotide Substitutions 51
- 4.1. Evolutionary Pathway Methods 52
- 4.2. Methods Based on Kimura's 2-Parameter Model 62
- 4.3. Nucleotide Substitutions at Different Codon Positions 67
- 4.4. Likelihood Methods with Codon Substitution Models 69

5 Phylogenetic Trees 73
5.1. Types of Phylogenetic Trees 73
5.2. Topological Differences 81
5.3. Tree-Building Methods 83

6 Phylogenetic Inference: Distance Methods 87
6.1. UPGMA 87
6.2. Least Squares (LS) Methods 92
6.3. Minimum Evolution (ME) Method 99
6.4. Neighbor Joining (NJ) Method 103
6.5. Distance Measures to Be Used for Phylogenetic Reconstruction 111

7 Phylogenetic Inference: Maximum Parsimony Methods 115
7.1. Finding Maximum Parsimony (MP) Trees 116
7.2. Strategies of Searching for MP Trees 122
7.3. Consensus Trees 130
7.4. Estimation of Branch Lengths 131
7.5. Weighted Parsimony 133
7.6. MP Methods for Protein Data 138
7.7. Shared Derived Characters 140

8 Phylogenetic Inference: Maximum Likelihood Methods 147
8.1. Computational Procedure of ML Methods 147
8.2. Models of Nucleotide Substitution 152
8.3. Protein Likelihood Methods 159
8.4. Theoretical Foundation of ML Methods 162
8.5. Parameter Estimation for a Given Topology 163

9 Accuracies and Statistical Tests of Phylogenetic Trees 165
9.1. Optimization Principle and Topological Errors 165
9.2. Interior Branch Tests 168
9.3. Bootstrap Tests 171
9.4. Tests of Topological Differences 175
9.5. Advantages and Disadvantages of Different Tree-Building Methods 178

10 Molecular Clocks and Linearized Trees 187
10.1. Molecular Clock Hypothesis 187
10.2. Relative Rate Tests 191
10.3. Phylogenetic Tests 196
10.4. Linearized Trees 203

11 Ancestral Nucleotide and Amino Acid Sequences 207
11.1. Inference of Ancestral Sequences: Parsimony Approach 207
11.2. Inference of Ancestral Sequences: Bayesian Approach 208

11.3. Synonymous and Nonsynonymous Substitutions in Ancestral Branches 216
11.4. Convergent and Parallel Evolution 221

12 Genetic Polymorphism and Evolution 231
12.1. Evolutionary Significance of Genetic Polymorphism 231
12.2. Analysis of Allele Frequency Data 233
12.3. Genetic Variation in Subdivided Populations 236
12.4. Genetic Variation for Many Loci 244
12.5. DNA Polymorphism 250
12.6. Statistical Tests for Detecting Selection 258

13 Population Trees from Genetic Markers 265
13.1. Genetic Distance for Allele Frequency Data 265
13.2. Analysis of DNA Sequences by Restriction Enzymes 275
13.3. Analysis of RAPD Data 285

14 Perspectives 291
14.1. Statistical Methods 291
14.2. Genome Projects 292
14.3. Molecular Biology and Evolution 294

Appendices
A. Mathematical Symbols and Notations 297
B. Geological Timescale 299
C. Geological Events in the Cenozoic and Mesozoic Eras 301
D. Organismal Evolution Based on the Fossil Record 303

References 305

Index 329

Numerical Examples

Example 2.1. Estimation of Evolutionary Distances and the Rate of Amino Acid Substitution in Hemoglobin α Chains 24
Example 2.2. Standard Errors of Poisson Correction (PC) Distances Obtained by the Analytical and Bootstrap Methods 27
Example 3.1. Number of Nucleotide Substitutions Between the Human and Rhesus Monkey Cytochrome *b* Genes 41
Example 4.1. Positive Darwinian Selection at MHC Loci 59
Example 4.2. Further Analysis of MHC Gene Sequences 66
Example 4.3. $\hat{d}_S$ *and* $\hat{d}_N$ Values for the Mitochondrial Nd5 Gene 67
Example 6.1. UPGMA Tree of Hominoid Species 90
Example 6.2. ME Trees for Hominoid Species 102
Example 6.3. NJ, ME, and BIONJ Trees for Simulated Sequence Data 109
Example 7.1. MP Trees for Five Hominoid Species 121
Example 7.2. Unweighted and Weighted MP Trees for Simulated Sequence Data 134
Example 7.3. Origin of Whales 136
Example 7.4. MP and NJ Trees for Cytochrome *b* Genes 138
Example 8.1. ML Tree for Hominoid Species 151
Example 8.2. ML Trees for Whales and Their Related Species 156
Example 8.3. ML Trees Obtained for Simulated Sequence Data 158
Example 8.4 Protein and DNA ML Trees for Vertebrate Mitochondrial Co1 Genes 160
Example 9.1. P_B and P_C Values for a Few Example Trees 174
Example 10.1 Relative Rate Tests for the Albumin Sequences from Humans, Rats, and Chickens 195
Example 10.2 Tests of the Molecular Clock Hypothesis for the *Drosophila Adh* Genes 200
Example 10.3. Linearized Trees for *Drosophila* Species 204
Example 11.1. Evolution of Color Vision in Mammals 212
Example 11.2. Adaptive Evolution of the ECP Gene after Gene Duplication 219
Example 11.3. Stomach Lysozymes of Foregut-Fermenting Animals 226
Example 12.1. Polymorphism of Mitochondrial DNA in Human Populations 253

Example 12.2. Nucleotide Diversity Within and Between Major Groups of Human Populations 258
Example 13.1. Evolutionary Relationships of Human Populations 273
Example 13.1. Restriction-Site Variation Within and Between Two Species of Chimpanzees 282

MOLECULAR EVOLUTION AND PHYLOGENETICS

1

Molecular Basis of Evolution

1.1. Evolutionary Tree of Life

From the time of Charles Darwin, it has been a dream for many biologists to reconstruct the evolutionary history of all organisms on Earth and express it in the form of a phylogenetic tree (Haeckel 1866). The ideal approach to this problem is to use the fossil record, but since the fossil record is fragmentary and incomplete, most investigators have used the methods of comparative morphology and comparative physiology. Using this approach, classical evolutionists have been able to infer the major aspects of the evolutionary history of organisms. However, the evolutionary change of morphological and physiological characters is so complex that this approach does not produce a clear-cut picture of evolutionary history, and the details of the phylogenetic trees reconstructed have almost always been controversial.

Recent advances in molecular biology have changed this situation drastically. Since the blueprint of all organisms is written in deoxyribonucleic acid (DNA) (ribonucleic acid [RNA] in some viruses), one can study the evolutionary relationships of organisms by comparing their DNA. This approach has several advantages over the classical approach in which morphological and physiological characters are used. First, DNA consists of the four types of nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G), and it can be used for comparing any groups of organisms, including bacteria, plants, and animals. In the classical approach, this is virtually impossible. Second, since the evolutionary change of DNA follows a more or less regular pattern, it is possible to use a mathematical model to formulate the change and compare DNAs from distantly related organisms. The evolutionary change of morphological characters is extremely complicated even for a short evolutionary time. Therefore, it is not clear whether various assumptions required for morphological phylogenetics are really satisfied or not. Third, the genomes of all organisms consist of long sequences of nucleotides and contain a much larger amount of phylogenetic information than morphological characters. For these reasons, molecular phylogenetics is expected to clarify many branching patterns of the tree of life that have been hard to resolve by the classical approach.

Systematics or taxonomy is one of the most controversial areas of biology. The definition of species, genera, families, and others is often subjective, and it is not uncommon that two experts working on the same group of organisms (e.g., *Drosophila*) vehemently disagree about the assignment of organisms to subspecies, species, genera, and so forth. Phylogenetics is less controversial than systematics, because it is primarily concerned with the evolutionary relationships of organisms, and the assignment of a group of organisms to a given taxonomic rank is of secondary importance. Nevertheless, the two areas of biology are closely related to each other, because the classification of organisms is conducted to reflect their evolutionary histories (Darwin 1859; Mayr 1968). In this sense, phylogenetics plays an important role in developing a scientific basis of systematics, though it may not solve all the problems of the latter discipline. Recent advances in molecular phylogenetics have already provided new insights into various aspects of classification of organisms, as will be mentioned later.

1.2. Mechanism of Evolution

The primary cause of evolution is the mutational change of genes. A mutant gene or DNA sequence caused by nucleotide substitution, insertions/deletions, recombination, gene conversion, and so forth may spread through the population by genetic drift and/or natural selection (see, e.g., Nei 1987; Hartl and Clark 1997) and eventually be fixed in a species. If this mutant gene produces a new morphological or physiological character, this character will be inherited by all the descendant species unless the gene mutates again. Therefore, if we establish a valid phylogenetic tree for a group of species, we are able to identify the lineage of species in which any specific character appeared by mutation.

This information is useful in understanding the mechanism of evolution of any specific character of interest. Comparison of the environmental conditions of this lineage of species with those of species lacking the character may suggest whether the character evolved by a particular process of natural selection or by genetic drift. If we can identify the genes involved and study their evolutionary change, we will know what kind of mutational change has generated the particular morphological or physiological character.

This type of study is already being conducted with respect to the enzymes (lysozymes and ribonucleases) that are associated with the evolution of the two-gut digestion system of ruminants and the langur monkey (Stewart et al. 1987; Jermann et al. 1995). The foregut of these animals harbors bacteria that can ferment grasses and tree leaves. These bacteria are then digested by lysozymes in the hind gut, and the RNA released is decomposed by ribonucleases. Fermentation mix, including digested bacteria, provides nutrients to host animals (Barnard 1969). It is now possible to infer the amino acid sequences of the proteins of ancestral organisms by statistical methods (e.g., Fitch 1971; Maddison and Maddison 1992; Yang et al. 1995b) and then reconstruct the ancestral proteins by site-directed mutagenesis. Therefore, we can study the catalytic activity

of the ancient proteins (e.g., Jermann et al. 1995). In this way, it is possible to study the evolutionary change of gene function.

Determination of the relative importance of mutation, natural selection, genetic drift, recombination, and so forth is an important subject in population genetics. For this purpose, population geneticists are now sequencing different alleles at a locus to clarify the evolutionary histories of the alleles. Here the problem is not the phylogenetic tree of different species but of different alleles within species. One of the interesting results obtained from this type of study is that some allelic lineages at the major histocompatibility complex (MHC) loci in mammals have persisted in the population for millions of years (e.g., Figueroa et al. 1988; Lawlor et al. 1988; McConnell et al. 1988; Hughes and Yeager 1998). This observation is consistent with the view that the antigen-recognition site of MHC molecules is subject to overdominant selection (Hughes and Nei 1988). Phylogenetic analyses of polymorphic alleles also have shown that intragenic recombination can occur with an appreciable frequency within species (e.g., Robertson et al. 1995; Fitch 1997).

Phylogenetic analysis of polymorphic alleles may also give important information about the extent of gene flow between two populations. Three decades ago, Prakash et al. (1969) examined the electrophoretic alleles at some enzyme loci in the North American and South American (Bogota, Colombia) populations of *Drosophila pseudoobscura* and found that many alleles are shared by the two populations. From this observation, they concluded that the Bogota population was formed only recently, possibly around 1950, from migrants from North America. However, this conclusion was questioned by Coyne and Felton (1977) in their detailed analysis of electrophoretic alleles. Later Schaeffer and Miller (1991) studied this problem by constructing a phylogenetic tree for the DNA sequences of polymorphic alleles of the alcohol dehydrogenase gene from the North and South American populations and concluded that the Bogota population was probably formed more than 100,000 years ago.

As is clear from the above examples, molecular phylogenetics has become an important tool for studying the mechanism of evolution.

1.3. Structure and Function of Genes

Although the molecular biology of the gene is beyond the scope of this book, let us explain the basic structure and function of genes that are important for understanding this book. In terms of function, genes can be classified into two groups: **protein-coding genes** and **RNA-coding genes.** Protein-coding genes are transcribed into **messenger RNAs (mRNA)**, which are in turn translated into the amino acid sequences of proteins. RNA-coding genes are those that produce **transfer RNAs (tRNA)**, **ribosomal RNAs (rRNA)**, **small nuclear RNAs (snRNA)**, and so on. These nonmessenger RNAs are the final products of RNA-coding genes. Ribosomal RNAs are components of ribosomes that are the core of the machinery of protein synthesis, whereas tRNAs are essential in transferring the genetic information of mRNAs into amino acid sequences of pro-

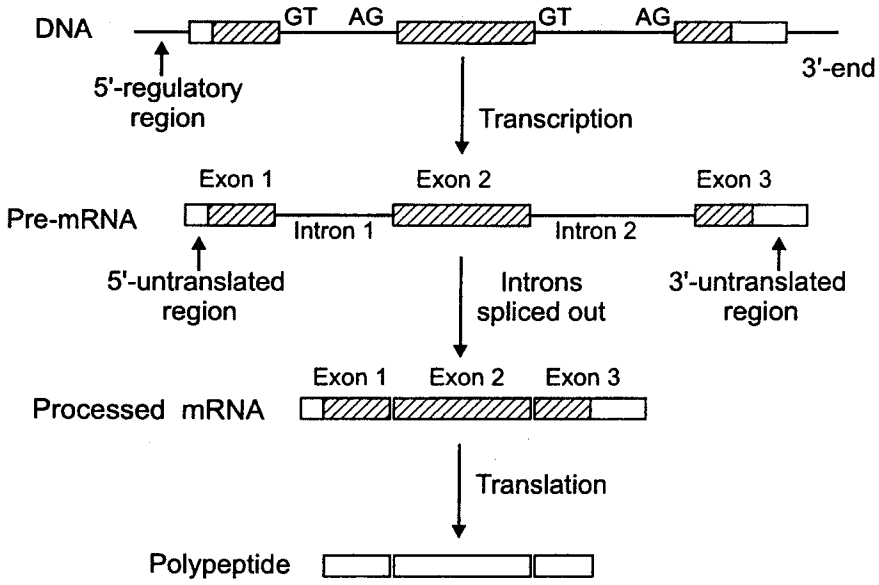


FIGURE 1.1. Basic structure of a eukaryotic protein-coding gene in relation to transcription and translation.

teins. snRNAs are confined to the nucleus, and several of them are involved in intron splicing or other RNA processing reactions.

The basic structure of a protein-coding gene in eukaryotes is presented in Figure 1.1. It is a long linear arrangement of four nucleotides, A, T, C, and G, and consists of a transcribed part of DNA and the 5′ and 3′ nontranscribed flanking regions. The flanking regions are necessary for controlling transcription and processing of **pre-messenger RNAs (pre-mRNA)**. A pre-mRNA consists of coding regions and noncoding regions. Coding regions contain information for encoding amino acids in the polypeptide produced by the gene, whereas noncoding regions contain some information necessary for regulation of polypeptide production. Some segments of noncoding regions are spliced out in the process of production of a **mature mRNA.** These segments are called **introns,** and the remaining regions are called **exons** (Figure 1.1). The number of exons in a gene varies from gene to gene. Prokaryotic genes have no introns, whereas some eukaryotic genes (e.g., muscular dystrophy gene) have as many as 78 introns (Roberts et al. 1992). The functional role of introns is not well understood. Usually, an intron begins with the dinucleotide GT and ends in AG. These dinucleotides provide context for correct intron splicing.

The genetic information carried by the nucleotide sequence of a gene is first transferred to mRNA by a simple process of one-to-one transcription of the nucleotides. The genetic information transferred to mRNA determines the amino acid sequences of the protein produced. Nucleotides of mRNA are read sequentially, three at a time. Each such triplet or **codon** is translated into a particular amino acid in the growing polypeptide chain according to the genetic code.

Table 1.1 Standard or "universal" genetic code.

Codon		Codon		Codon		Codon	
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Ter	UGA	Ter
UUG	Leu	UCG	Ser	UAG	Ter	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

The **genetic code** for nuclear genes seems to be universal for both prokaryotes and eukaryotes with a few exceptions. The same genetic code (**"universal"** or **standard genetic code**) is used for chloroplast genes, but mitochondrial genes use slightly different genetic codes. The standard genetic code is presented in Table 1.1. In this table, amino acids are represented by three-letter codes (see Table 1.2). There are $4^3 = 64$ possible codons for the four different nucleotides, uracil (U), cytosine (C), adenine (A), and guanine (G). (U corresponds to T in nucleotide sequences.) Three of the codons (UAA, UAG, UGA) are, however, **termination** or **stop codons** and do not code for any amino acid. Each of the remaining 61 codons (**sense codons**) codes for a particular amino acid, but since there are only 20 amino acids (Table 1.2) used for making proteins, there are many codons that code for the same amino acid. Codons coding for the same amino acid are called **synonymous codons.** In the genetic code table, codon AUG codes for methionine, but this codon is also used as the **initiation codon.** The methionine encoded by the initiation codon is in a modified form and is later removed from the polypeptide. Recent studies have shown that CUG and UUG are also used as the initiation codons in some nuclear genes (Elzanowski and Ostell 1996). These initiation codons should be excluded in a study of DNA sequence evolution, because they remain unchanged in most cases. Termination codons should also be eliminated.

Table 1.3 shows the genetic code for vertebrate mitochondrial genes. There are a few differences between this genetic code and the standard nuclear genetic code. In the mitochondrial genetic code, codon UGA is not a termination codon but codes for tryptophan. By contrast, codons AGA and AGG are termination codons instead of an arginine codon. AUA, which codes for isoleucine in the nuclear code, is used for encod-

Table 1.2 One- and three-letter amino acid codes.

Name	Code		Property of the Side Chain at pH 7
	1-Letter	3-Letter	
Alanine	A	Ala	Nonpolar (hydrophobic)
Cysteine	C	Cys	Polar
Aspartic acid	D	Asp	Polar (hydrophilic, acidic)
Glutamic acid	E	Glu	Polar (hydrophilic, acidic)
Phenylalanine	F	Phe	Nonpolar (hydrophobic)
Glycine	G	Gly	Nonpolar
Histidine	H	His	Polar (hydrophilic, basic)
Isoleucine	I	Ile	Nonpolar (hydrophobic)
Lysine	K	Lys	Polar (hydrophilic, basic)
Leucine	L	Leu	Nonpolar (hydrophobic)
Methionine	M	Met	Nonpolar (hydrophobic)
Asparagine	N	Asn	Polar (hydrophilic, neutral)
Proline	P	Pro	Nonpolar
Glutamine	Q	Gln	Polar (hydrophilic, neutral)
Arginine	R	Arg	Polar (hydrophilic, basic)
Serine	S	Ser	Polar
Threonine	T	Thr	Polar
Valine	V	Val	Nonpolar (hydrophobic)
Tryptophan	W	Trp	Nonpolar
Tyrosine	Y	Tyr	Polar

ing methionine. The mitochondrial genetic code for vertebrates does not necessarily apply to nonvertebrate organisms. In fact, ascidian, echinoderms, *Drosophila,* yeast, plants, and protozoans are known to have slightly different genetic codes, as shown in Table 1.4. The genetic codes of nuclear genes of ciliated protozoans such as *Tetrahymena* and *Par-*

Table 1.3 Vertebrate mitochondrial genetic code. Differences from the standard genetic code are shown in boldface.

Codon		Codon		Codon		Codon	
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Ter	UGA	**Trp**
UUG	Leu	UCG	Ser	UAG	Ter	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	**Met**	ACA	Thr	AAA	Lys	AGA	**Ter**
AUG	Met	ACG	Thr	AAG	Lys	AGG	**Ter**
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Table 1.4 Some other genetic codes that differ from the standard code.

	Codons						
Organelle/Organisms	UGA	AUA	AAA	AGR	CUN	CGG	UAR
Standard genetic code	Ter	Ile	Lys	Arg	Leu	Arg	Ter
Mitochondrial code							
Vertebrate	Trp	Met	•	Ter	•	•	•
Ascidian	Trp	Met	•	Gly	•	•	•
Echinoderm	Trp	•	Asn	Ser	•	•	•
Drosophila	Trp	Met	•	Ser	•	•	•
Yeast	Trp	Met	•	•	Thr	•	•
Protozoan	Trp	•	•	•	•	•	•
Mold	Trp	•	•	•	•	•	•
Coelenterate	Trp	•	•	•	•	•	•
Nuclear code							
Tetrahymena	•	•	•	•	•	•	Gln
Mycoplasma	Trp	•	•	•	•	•	•
Euplotid	Cys	•	•	•	•	•	•

Note: • Indicates identity with the standard code. R = A or G and N = T, C, A, or G.

amecium are also slightly different from the standard genetic code. Here, UAA and UAG do not appear to be termination codons even in nuclear genes but code for glutamine. Furthermore, in the prokaryotic organism *Mycoplasma capricolum,* the usual termination codon UGA is used for encoding tryptophan (see Osawa 1995).

In plant mitochondrial genes, codon CGG is not directly translated into tryptophan, but the nucleotide C in this codon is converted to U after the mRNA is formed, and this converted codon UGG encodes tryptophan using the standard genetic code. This process is called **RNA editing** (Covello and Gray 1993). In the comparison of amino acid sequences from different plant species, however, one can treat CGG as though it were a tryptophan codon. Actually, RNA editing occurs in some mitochondrial genes of the other eukaryotic kingdoms as well, and one should be cautious in translating DNA sequences into amino acid sequences in these genes.

1.4. Mutational Changes of DNA Sequences

Since all morphological and physiological characters of organisms are ultimately controlled by the genetic information carried by DNA, any mutational changes in these characters are due to some change in DNA molecules. There are four basic types of changes in DNA. They are **substitution** of a nucleotide for another nucleotide (Figure 1.2A), **deletion** of nucleotides (Figure 1.2B), **insertion** of nucleotides (Figure 1.2C), and **inversion** of nucleotides (Figure 1.2D). Insertion, deletion, and inversion may occur with one or more nucleotides as a unit. If insertions or deletions occur in a protein coding gene, they may shift the reading frame of the nucleotide sequence. These insertions and deletions are called **frameshift mutations.**

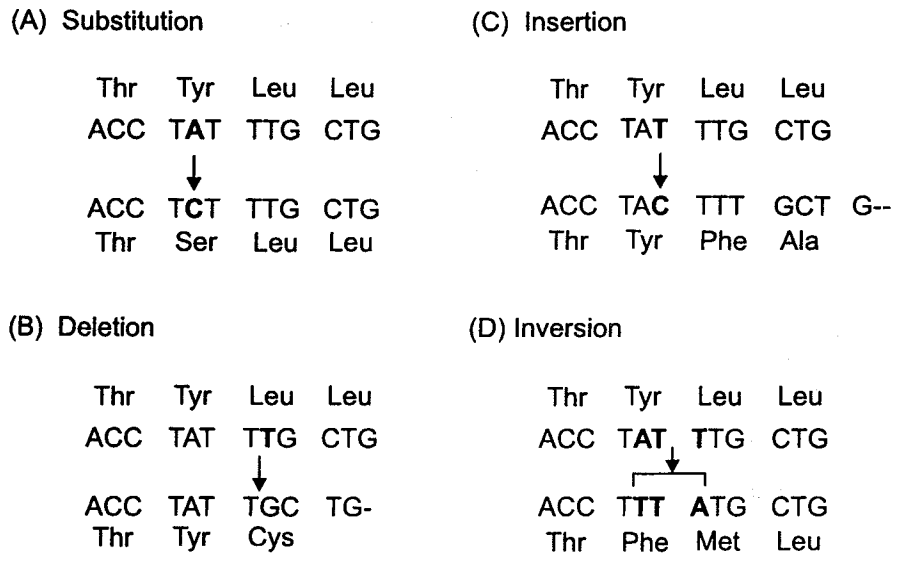


FIGURE 1.2. Four basic types of mutation at the nucleotide level. Nucleotide sequences are presented in units of codons or nucleotide triplets in order to show how the amino acids encoded are affected by the nucleotide changes. The nucleotides affected by the mutational changes are shown in boldface.

Nucleotide substitutions can be divided into two classes: **transitions** and **transversions.** A transition is the substitution of a **purine** (adenine or guanine) for another purine or the substitution of a **pyrimidine** (thymine or cytosine) for another pyrimidine (Figure 1.3). Other types of nucleotide substitutions are called transversions. In most DNA segments, transitional nucleotide substitutions are known to occur more frequently than transversions (e.g., Fitch 1967; Gojobori et al. 1982; Kocher and Wilson 1991). In the case of protein-coding genes, nucleotide substitutions that result in synonymous codons are called **synonymous** or **silent**

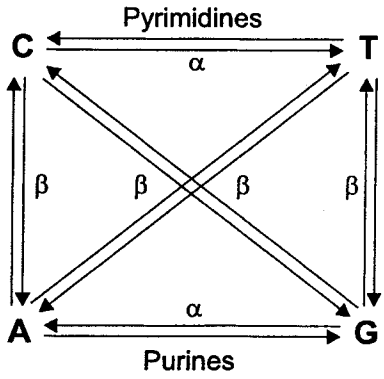


FIGURE 1.3. Transitional (A↔G and T↔C) and transversional (others) nucleotide substitutions. α and β are the rates of transitional and transversional substitutions, respectively.

substitutions, whereas those that result in nonsynonymous codons are called **nonsynonymous** or **amino acid replacement substitutions.** In addition, there are mutations that result in stop codons, and they are called **nonsense mutations.**

Because of the properties of the genetic codes, most synonymous substitutions occur at the third nucleotide position of codons, but some occur at the first position. All nucleotide substitutions at the second position are either nonsynonymous or nonsense mutations. If we assume that all codons are equally frequent in the genome and the probability of substitution is the same for all pairs of nucleotides, the proportions of synonymous, nonsynonymous, and nonsense mutations are about 25, 71, and 4%, respectively (Nei 1975; Li 1997). In practice, of course, the assumption of equal frequencies of codons and random nucleotide substitution is not realistic, but these percentages give a rough idea of the relative frequencies of different mutations at the nucleotide level.

Recent data indicate that insertions or deletions occur quite often, particularly in noncoding regions of DNA. The number of nucleotides involved in an insertion or deletion event varies from a few nucleotides to a large block of DNA. Short insertions or deletions are apparently caused by errors in DNA replication. Long insertions or deletions seem to be mainly due to unequal crossover or DNA transposition. DNA transposition, that is, the movement of a DNA segment from one chromosomal position to another, may occur by the aid of **transposons** or **transposable elements.** Transposable elements are known to cause various sorts of mutational changes (e.g., Kidwell and Lisch 1997). Another possible mechanism of gene insertion is **horizontal gene transfer** between species that is apparently mediated by transposable elements.

The possible role of unequal crossover in increasing the number of genes in the genome has been known for many years (Bridges 1936; Stephens 1951). However, only after the initiation of molecular study of DNA has it been realized that it plays an important role in evolution in increasing or decreasing DNA content (Ohno 1967, 1970; Nei 1969). Particularly in multigene families such as immunoglobulin and ribosomal RNA genes, unequal crossover seems to have played an important role in generating multiple copies of genes. A genetic event related to unequal crossover is **gene conversion.** Gene conversion is the alteration of a segment of DNA that makes the segment identical with another segment of DNA. This event is believed to occur by the repair of mismatched bases in heteroduplex DNA (Radding 1982) and is capable of homogenizing the member genes of a multigene family, but it does not change the number of gene copies.

1.5. Codon Usage

If nucleotide substitution occurs at random at each nucleotide site, every nucleotide site is expected to have one of the four nucleotides, A, T, C, and G, with equal probability. Therefore, if there is no selection and no mutational bias, one would expect that the codons encoding the same amino acid are on average in equal frequencies in protein coding regions

Phe UUU	15 (0.51)	Ser UCU	32 (1.86)	Tyr UAU	18 (0.64)	Cys UGU	5 (1.00)
UUC	44 (1.49)	UCC	38 (2.21)	**UAC**	38 (1.36)	UGC	5 (1.00)
Leu UUA	2 (0.07)	UCA	2 (0.12)	Ter UAA		Ter UGA	
UUG	8 (0.27)	UCG	5 (0.29)	Ter UAG		Trp UGG	8 (1.00)
Leu CUU	11 (0.36)	Pro CCU	9 (0.48)	His CAU	5 (0.36)	Arg **CGU**	89 (3.93)
CUC	18 (0.60)	CCC	0 (0.00)	CAC	23 (1.64)	**CGC**	46 (2.03)
CUA	1 (0.03)	CCA	11 (0.59)	Gln CAA	15 (0.34)	CGA	1 (0.04)
CUG	141 (4.67)	**CCG**	55 (2.93)	**CAG**	73 (1.66)	CGG	0 (0.00)
Ile AUU	29 (0.69)	Thr **ACU**	19 (0.78)	Asn AAU	4 (0.11)	Ser AGU	3 (0.17)
AUC	98 (2.31)	**ACC**	63 (2.57)	**AAC**	66 (1.89)	AGC	23 (1.34)
AUA	0 (0.00)	ACA	3 (0.12)	Lys **AAA**	77 (1.35)	Arg AGA	0 (0.00)
Met AUG	60 (1.00)	ACG	13 (0.53)	AAG	37 (0.65)	AGG	0 (0.00)
Val **GUU**	55 (1.53)	Ala GCU	30 (0.94)	Asp GAU	60 (0.83)	Gly **GGU**	78 (2.40)
GUC	21 (0.58)	GCC	19 (0.59)	GAC	85 (1.17)	GGC	47 (1.45)
GUA	34 (0.94)	**GCA**	30 (0.94)	Glu **GAA**	147 (1.52)	GGA	0 (0.00)
GUG	34 (0.94)	**GCG**	49 (1.53)	GAG	46 (0.48)	GGG	5 (0.15)

FIGURE 1.4. Codon frequencies observed in the RNA polymerase genes (rpo B and D genes) of the bacterium *Escherichia coli*. The codons optimal for the translational system are shown in boldface. Relative synonymous codon usages (RSCU) given in the parentheses were computed by Equation (1.1). Data from Ikemura (1985).

of DNA. For example, amino acid valine (Val) is encoded by four codons, GUU, GUC, GUA, and GUG. So, if we examine a large number of Val codons in a gene, the relative frequencies of GUU, GUC, GUA, and GUG are all expected to be nearly equal to 25%.

In practice, the frequencies of different codons for the same amino acid are usually different, and some codons are used more often than others. Figure 1.4 shows the frequencies of use of each codon (number of times used) in the RNA polymerase of the bacterium *Escherichia coli (E. coli)* (Ikemura 1985). In the case of amino acid valine, the four codons are used nearly equally, though the use of GUU is more than two times higher than that of GUC. In arginine, however, codons CGU and CGC are used almost exclusively, and codons CGA, CGG, AGA, and AGG are almost never used. This type of **codon usage bias** is generally observed in both prokaryotic and eukaryotic genes.

What causes the codon usage bias? There are several factors. First, Ikemura (1981, 1985) showed that in *E. coli* and yeast, the frequency of codon usage in highly expressed genes is correlated with the relative abundance of the isoaccepting tRNAs in the cell. In other words, the tRNAs that correspond to frequently used codons are more abundant than those corresponding to rarely used codons. For example, in the case of arginine codons, CGU and CGC are often used, because the tRNAs corresponding to these two codons are more abundant than the tRNAs corresponding to the other codons. This suggests that the translationary machinery tends to use abundant tRNAs to produce proteins. Ikemura (1985) showed that the abundance of a particular tRNA is correlated with the number of copies of the gene that encodes the tRNA. Therefore, as far as highly expressed genes are concerned, the codon usage bias is essentially the same for all genes in the same organism. The codon usage bias

Table 1.5 Relative synonymous codon usage (*RSCU*) in bacteria (*E. coli*), yeast (*S. cerevisiae*), fruit fly (*D. melanogaster*), and human.

Amino Acid	Codon	Bacteria		Yeast		Fruit fly		Human	
		High[a]	Low[b]	High	Low	High	Low	G+C[c]	A+T[d]
Leu	UUA	0.06	1.24	0.49	1.49	0.03	0.62	0.05	0.99
	UUG	0.07	0.87	**5.34**	**1.48**	0.69	1.05	0.31	1.01
	CUU	0.13	0.72	0.02	0.73	0.25	0.80	0.20	1.26
	CUC	0.17	0.65	0.00	0.51	0.72	0.90	1.42	0.80
	CUA	0.04	0.31	0.15	0.95	0.06	0.60	0.15	0.57
	CUG	**5.54**	**2.20**	0.02	0.84	**4.25**	**2.04**	**3.88**	**1.38**
Val	GUU	**2.41**	**1.09**	**2.07**	**1.13**	0.56	0.74	0.09	1.32
	GUC	0.08	0.99	1.91	0.76	1.59	0.93	1.03	0.69
	GUA	1.12	1.63	0.00	1.18	0.06	0.53	0.11	0.80
	GUG	0.40	1.29	0.02	0.93	**1.79**	**1.80**	**2.78**	**1.19**
Ile	AUU	0.48	1.38	1.26	1.29	0.74	1.27	0.45	1.60
	AUC	**2.51**	**1.12**	**1.74**	**0.66**	**2.26**	**0.95**	**2.43**	**0.76**
	AUA	0.01	0.50	0.00	1.05	0.00	0.78	0.12	0.64
Phe	UUU	0.34	1.33	0.19	1.38	0.12	0.86	0.27	1.20
	UUC	**1.66**	**0.67**	**1.81**	**0.62**	**1.88**	**1.14**	**1.73**	**0.80**

Source: Modified from Sharp et al. (1988).
Note: Codons with the highest usage in the *High* genes and their corresponding usages in the *Low* genes are shown in boldface.

[a]*High* denotes genes with high levels of gene expression.
[b]*Low* denotes genes with low levels of gene expression.
[c]In humans, "G + C" refers to genes in GC-rich regions.
[d]In humans, "A + T" refers to genes in AT-rich regions.

in yeast is known to be quite different from that in *E. coli* (Table 1.5), but this bias can also be explained by the relative contents of tRNAs.

However, the above rule does not necessarily apply to moderately expressed genes (e.g., Thr and Trp synthetase genes in *E. coli*). In these genes, the codon usage tends to be more even for all codons that encode the same amino acid. The reason for this seems to be that the translation does not occur quickly, so that rare isoaccepting tRNAs can be used.

The above observations suggest that nucleotides mutate more or less at random, but the codons that do not correspond to abundant tRNAs are eliminated by **purifying selection** in highly expressed genes, because they are inefficient in protein synthesis. In moderately expressed genes, the selection pressure is apparently so low that many different codons are used. This pattern of natural selection is observed in many single-cell organisms or even in the fruit fly *Drosophila melanogaster* (Table 1.5). However, this rule does not seem to apply to human genes (Sharp et al. 1988).

Although the relative abundance of isoaccepting tRNAs is an important factor, there is another factor that affects the codon usage; it is the **biased mutation pressure.** In bacteria, the relative frequency of nucleotides G and C (GC content) in the genome is known to vary from about 25 to 75% (Osawa 1995), and this variation is believed to be largely due to the difference between the forward and backward mutation rates of the

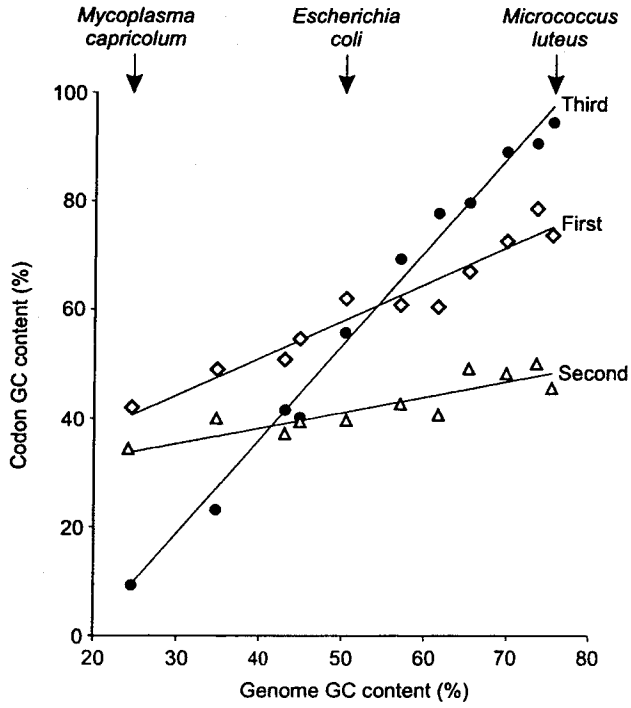


Figure 1.5. Relationships of the total genome GC content and the GC contents of first, second, and third nucleotide positions of genes from 11 different bacterial species, of which three species names are given. Modified from Muto and Osawa (1987).

GC and the AT pairs in the nucleotide sequence (Sueoka 1962). In some bacteria (e.g., *Mycoplasma capricolum*), the mutation pressure from GC to AT is so high that the nucleotides at silent third codon positions are almost always A or T (Muto and Osawa 1987). In some other bacteria (e.g., *Micrococcus luteus*), mutation pressure occurs in the opposite direction (AT → GC), so that the most often used nucleotide at third codon positions is either G or C (Figure 1.5).

Of course, for a protein to maintain its function, even the GC content at third positions is expected to be different from the equilibrium frequency determined by the mutation pressure alone, because some nucleotide substitutions at third positions result in amino acid changes and thus would be subject to purifying selection. Nucleotide substitutions at second positions are all nonsynonymous, so they are primarily controlled by functional constraints rather than by mutation pressure, whereas a small proportion of substitutions at first positions are synonymous, so that the effect of mutation pressure is expected to be intermediate between the effects at third and second codon positions.

Figure 1.5 shows the relationships between the GC content at first, second, and third codon positions of genes and the genome GC content in 11 different bacterial species covering a broad range of genome GC content. At third codon positions, the GC content of genes is nearly equal to

the genome GC content, suggesting that the effect of mutation pressure is very strong. At second positions, however, the slope of the linear relationship with the genome GC content is much lower than that at third positions. This suggests that the effect of mutation pressure is less important at second positions and that the GC content is determined largely by purifying selection due to the functional constraints of the genes, as mentioned above. As expected, the slope of the relationship between the gene and genome GC contents for first positions is intermediate between the slopes for third and second positions. These observations support the view that the codon usage is controlled by both mutation pressure and purifying selection.

The fact that the GC content varies extensively among different groups of bacteria indicates that the pattern of nucleotide substitution is not the same for all groups of bacteria. This introduces complications in the study of phylogenetic relationships of these organisms (Galtier and Gouy 1995, 1998). The different groups of bacteria considered here probably diverged more than one billion years ago, so one might think that this problem is not important for the study of evolution of higher organisms, which evolved more recently. In practice, however, there is evidence that even in a relatively shorter period of evolutionary time the pattern of nucleotide substitution may change (Moriyama and Powell 1998).

In contrast to single-cell organisms, animals and plants are known to have a narrow range of GC content when the total genome is considered (Sueoka 1962). In particular, vertebrate animals all have a GC content of 40–45%. However, codon usage bias is still observed in many genes of higher organisms. In some invertebrates such as *Drosophila* the bias is quite strong, and this bias is apparently caused by the relative abundance of isoaccepting tRNAs, as in the case of microorganisms (Shields et al. 1988; Akashi 1994; Moriyama and Powell 1997).

In vertebrates, this issue is somewhat complicated, because gene expression is tissue-dependent and the genome is heterogeneous in terms of GC content. Bernardi et al. (1985, 1988) have shown that the vertebrate genomes are a mosaic composed of GC-rich regions and GC-poor regions and that some GC-rich regions have a GC content of about 60% and some GC-poor regions about 30%. Each of these GC-rich or GC-poor regions may be as long as 300 kb and contain many functional genes. These GC-rich or GC-poor regions are called **isochores.** Interestingly, the GC content at third codon positions of genes within an isochore is generally close to the GC content of the entire isochore. There are four major groups of isochores (two GC-rich and two GC-poor isochores) in warm-blooded vertebrates such as mammals and birds, but in cold-blooded vertebrates GC-rich isochores are rare or nearly absent. The boundary between the GC-rich and GC-poor isochores is known to be quite narrow (Ikemura and Aota 1988).

The origin of isochores in vertebrates has been a subject of controversy, and no consensus has been reached at the present time. The reader who is interested in this subject may refer to Wolfe et al. (1989), Holmquist and Filipski (1994), and Bernardi (1995). However, it is important to note that the genes located in different isochores are expected to have differ-

ent patterns of codon usage biases, and since codon usage bias affects the rate of nucleotide substitution (Shields et al. 1988; Sharp et al. 1989), they may evolve at different rates.

Statistical Measures of Codon Usage Bias

The numbers of occurrence (frequencies) of different codons for a given amino acid as those shown in Figure 1.4 clearly indicate the extent of codon usage bias if there is any. However, the absolute frequencies of codons are not convenient for comparing the bias for different genes or for different organisms, because the total number of codons examined is not necessarily the same. In this case, a more useful measure of codon usage bias is the **relative synonymous codon usage** (*RSCU*), which is defined as the observed frequency of a codon divided by the expected frequency under the assumption of equal codon usage (Sharp et al. 1986). For a given amino acid, *RSCU* is given by

$$RSCU = X_i/\bar{X} \tag{1.1}$$

where X_i is the observed number of the i-th codon for the amino acid, and $\bar{X}$ is the average of X_i over all codons, that is, $\bar{X} = \Sigma_i X_i/m$, where m is the number of different codons for the amino acid.

As an example, let us consider the codon usage for the RNA polymerase genes, *rpo B* and *D*, of *E. coli* in Figure 1.4. For amino acid proline (Pro) there are four codons ($m = 4$), the observed number of a codon (X_i) varies from 0 to 55, and the average frequency becomes $\Sigma_i X_i/m = 75/4 = 18.75$. Therefore, *RSCU* is 0.48 for codon CCU, 0 for CCC, 0.59 for CCA, and 2.93 for CCG. *RSCU*s for other codons can be obtained in the same way, and they are given in Figure 1.4. These values can now be used for comparing the codon usage patterns for different genes. Table 1.5 shows the *RSCU*s for four amino acids for the genes from bacteria, yeasts, fruit flies, and humans. There is conspicuous codon usage bias for all organisms, but the pattern of the bias varies considerably with organism.

A number of authors have proposed statistical methods for measuring the extent of codon bias for the entire sequence of a gene or a genome. The measures include the **codon adaptation index** (Sharp and Li 1987), **scaled χ^2 measure** (Shields et al. 1988), and **effective number of codons** (Wright 1990). Each measure has advantages and disadvantages, but at this moment it is not clear which method is most useful (Comeron and Aguade 1998).

2

Evolutionary Change of Amino Acid Sequences

Before the invention of rapid methods of DNA sequencing in 1977 (Maxam and Gilbert 1977; Sanger et al. 1977), most studies of molecular evolution were conducted by using amino acid sequence data. Although amino acid sequencing was time consuming and error prone, some important principles of molecular evolution, such as evolution by gene duplication (Ingram 1963; Ohno 1970) and the molecular clock (Zuckerkandl and Pauling 1962; Margoliash 1963), were discovered by the study of amino acid sequence data. At the present time, DNA sequencing is much simpler than amino acid sequencing, and amino acid sequences are usually deduced from nucleotide sequences by using the genetic code. However, amino acid sequences are still useful for evolutionary studies; they are more conserved than DNA sequences and thus provide useful information on long-term evolution of genes or species. They are also almost indispensable for aligning DNA sequences of protein-coding genes. Furthermore, the mathematical model for the evolutionary change of amino acid sequences is much simpler than that of DNA sequences. For these reasons, we first consider the evolutionary change of amino acid sequences.

The primary purpose of this chapter is to present statistical methods for measuring the **evolutionary distance** between two amino acid sequences. Evolutionary distances are fundamental for the study of protein evolution and are useful for constructing phylogenetic trees and estimation of divergence times. In the case of amino acid sequence data, the distance is usually measured by the number of amino acid substitutions, but there are several different measures, depending on the assumptions made.

2.1. Amino Acid Differences and Proportion of Different Amino Acids

Study of the evolutionary change of proteins or polypeptides begins with comparison of two or more amino acid sequences from different organisms. Figure 2.1 shows the amino acid sequences of hemoglobin α chains from the human, horse, cow, kangaroo, newt, and carp. In this figure, all

```
Human     V-LSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHF-DL SHGSAQVKGH  60
Horse     .-..A..... .....S...G .......... .....G.... .......-.. ........A.
Cow       .-..A...G. .........G ..A....... .......... .......-.. ..........
Kangaroo  .-..A...GH ...I.....G .....A..G. ..T.H..... .......-.. ......IQA.
Newt      MK..AE..H. ..TT.DHIKG .EEAL..... F...T.L.A. R....AK-.. .E..SFLHS.
Carp      S-..DK..AA ..I..A.ISP K.DDI..... G..LTVY.Q. ....A.WA.. .P..GP..-.

Human     GKKVA-DALT NAVAHVDDMP NALSALSDLH AHKLRVDPVN FKLLSHCLLV TLAAHLPAEF 120
Horse     .....-.G.. L..G.L..L. G...D..N.. .......... .........S ...V...ND.
Cow       .A...-A... K..E.L..L. G...E..... .......... ......S... ...S...SD.
Kangaroo  ...I.-...G Q..E.I..L. GT..K..... .......... .......... .F....GDA.
Newt      ....M-G..S .....I..ID A..CK...K. .QD.M...A. .PK.A.NI.. VMGI..K.HL
Carp      ....IMG.VG D..SKI..LV GG.AS..E.. .S......A. ..I.ANHIV. GIMFY..GD.

Human     TPAVHASLDK FLASVSTVLT SKYR 144
Horse     .......... ..S..... ....
Cow       .......... ...N..... ....
Kangaroo  ..E..... ...A..... ....
Newt      .YP..C.V.. ..DV.GH... ....
Carp      P.E..M.V.. .FQNLALA.S E...
```

FIGURE 2.1. Amino acid sequences of hemoglobin α-chains from six different vertebrate species. Hyphens (-) indicate the positions of deletions or insertions, and dots (.) show identity with the amino acids of the human sequence.

amino acids are denoted by one-letter codes (Table 1.2). When this type of data is given, there are several ways of measuring the extent of evolutionary divergence of sequences.

One simple measure is the number of amino acid differences (n_d) between two sequences. If the number of amino acids used (n) is the same for all sequences, this number can be used for comparing the extents of sequence divergence for different pairs of sequences. In practice, amino acid sequences often include **insertions** or **deletions (indels)** when many sequences are compared (see Figure 2.1). In this case, all indels or gaps must be eliminated before the computation of n_d. Otherwise, the comparison of n_d's between different pairs of sequences is not meaningful.

Actually, a more convenient measure of the extent of sequence divergence among different proteins is the proportion of different amino acids between two sequences. This proportion (p) can be used for comparing the extents of sequence divergence even when n varies with sequence. It is estimated by

$$\hat{p} = n_d/n \tag{2.1}$$

This proportion is sometimes called the **p distance.** If all amino acid sites are subject to substitution with equal probability, n_d follows the binomial distribution. Therefore, the variance of $\hat{p}$ is given by

$$V(\hat{p}) = p(1 - p)/n \tag{2.2}$$

In actual computation of $V(\hat{p})$, p is replaced by $\hat{p}$.

In the example given in Figure 2.1, the total number of amino acid sites after elimination of all gaps is 140. Therefore, $n = 140$ in this case. Since the n_d values are given above the diagonal of Table 2.1, one can easily compute the $\hat{p}$ values. They are presented below the diagonal of Table 2.1. This table indicates that $\hat{p}$ is greater when the species compared are

Table 2.1 Numbers of amino acid differences (above the diagonal) and the proportions of different amino acids (below the diagonal) between hemoglobin α-chains from different vertebrate species.

	Human	Horse	Cow	Kangaroo	Newt	Carp
Human		17	17	26	61	68
Horse	0.121		17	29	66	67
Cow	0.121	0.121		25	63	65
Kangaroo	0.186	0.207	0.179		66	71
Newt	0.436	0.471	0.450	0.471		74
Carp	0.486	0.479	0.464	0.507	0.529	

Note: Deletions and insertions were excluded from the computation, the total number of amino acids used being 140.

distantly related (e.g., human vs. carp) than when they are closely related (e.g., human vs. horse). This suggests that the number of amino acid substitutions increases as the time after divergence of two species increases. However, as will be shown below, p is not strictly proportional to divergence time (t) (Figure 2.2).

2.2. Poisson Correction (PC) and Gamma Distances

One reason for the nonlinear relationship of p with t is that as multiple amino acid substitutions start to occur at the same sites, the discrepancy between n_d and the actual number of amino acid substitutions gradually increases. One way to estimate the number of substitutions more accu-

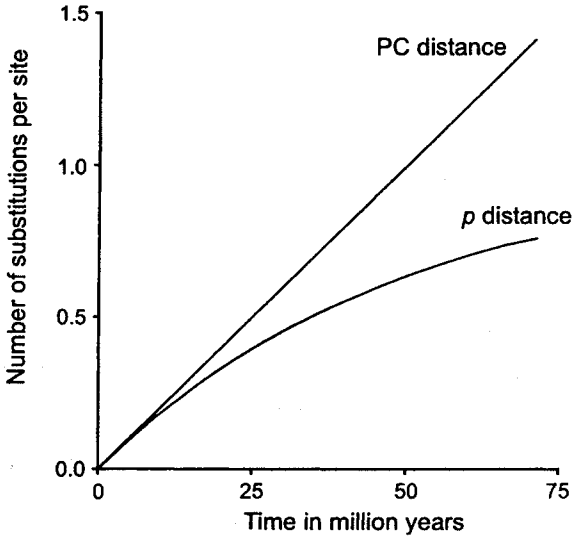


FIGURE 2.2. Relationships of the p distance and the Poisson correction (PC) distance with time. The rate of amino acid substitution (r) is assumed to be 10^{-8} per site per year.

rately is to use the concept of the Poisson distribution. Let r be the rate of amino acid substitution per year at a particular amino acid site and assume for simplicity that it is the same for all sites. This assumption does not necessarily hold in reality, but as will be seen later, the error introduced by this assumption is small unless p is large. The mean number of amino acid substitutions per site during a period of t years is then rt, and the probability of occurrence of k amino acid substitutions at a given site ($k = 0, 1, 2, 3, \ldots$) is given by the following Poisson distribution.

$$P(k;t) = e^{-rt}(rt)^k/k! \tag{2.3}$$

Therefore, the probability that no amino acid change has occurred at a given site is $P(0;t) = e^{-rt}$. If the number of amino acids in a polypeptide is n, the expected number of unchanged amino acids is then given by ne^{-rt}.

In reality, we generally do not know the amino acid sequence for the ancestral species, so Equation (2.3) is not applicable. Therefore, the number of amino acid substitutions is estimated by comparing two homologous sequences that diverged t years ago. Since the probability that no amino acid substitution has occurred during t years at a site of a sequence is e^{-rt}, the probability (q) that neither of the homologous sites of the two sequences has undergone substitution is

$$q = (e^{-rt})^2 = e^{-2rt} \tag{2.4}$$

This probability can be estimated by $1 - \hat{p}$, since $q = 1 - p$. The equation $q = e^{-2rt}$ is approximate because backward mutations and parallel mutations (the same mutations occurring at the homologous amino acid sites in two different evolutionary lines) are not taken into account. However, the effects of these mutations are generally very small unless $\hat{p}$ is large (say, $\hat{p} > 0.3$).

If we use Equation (2.4), the total number of amino acid substitutions per site for the two sequences ($d = 2rt$) is given by

$$d = -\ln(1 - p) \tag{2.5}$$

An estimate ($\hat{d}$) of d can be obtained by replacing p by $\hat{p}$, and the large-sample variance of $\hat{d}$ is given by

$$V(\hat{d}) = p/[(1 - p)n] \tag{2.6}$$

In actual computation of $V(\hat{d})$, p is again replaced by $\hat{p}$. This is true for all the variance formulas for $\hat{d}$ or other estimators in this book. In the following, we call $\hat{d}$ the **Poisson correction (PC) distance.**

In the study of molecular evolution, it is often important to know the rate of amino acid substitution (r). This rate can be estimated by

$$\hat{r} = \hat{d}/(2t) \tag{2.7}$$

if we know the time of divergence between the two sequences from other biological information. Note that $\hat{d}$ is divided by $2t$ rather than by t, be-

cause the rate refers to one evolutionary lineage. The variance of $\hat{r}$ is given by $V(\hat{d})/(2t)^2$. On the other hand, if we know the rate r from previous studies but we do not know the evolutionary time, t can be estimated by

$$\hat{t} = \hat{d}/(2r) \tag{2.8}$$

with variance $V(\hat{d})/(2r)^2$.

In the above formulation, we assumed that the rate of amino acid substitution is the same for all amino acid sites. This assumption usually does not hold, since the rate is often higher at functionally less important sites than at functionally more important sites (Kimura 1983). Indeed, Uzzell and Corbin (1971) have shown that the distribution of the number of amino acid substitutions per site (k) has a variance greater than the Poisson variance and that the number approximately follows the negative binomial distribution. It is known that when the rate of amino acid substitution (r) varies with site according to the gamma distribution, the observed number of substitutions per site (k) will be distributed as a negative binomial distribution (Johnson and Kotz 1969). Therefore, Uzzell and Corbin's (1971) observation suggests that the substitution rate varies from site to site according to the gamma distribution, which is given by

$$f(r) = \frac{b^a}{\Gamma(a)} e^{-br} r^{a-1} \tag{2.9}$$

where $a = \bar{r}^2/V(r)$ and $b = \bar{r}/V(r)$, $\bar{r}$ and $V(r)$ being the mean and variance of r respectively (Johnson and Kotz 1970). Here $\Gamma(a)$ is the gamma function defined by

$$\Gamma(a) = \int_0^\infty e^{-t} t^{a-1} dt \tag{2.10}$$

The shape of the distribution $f(r)$ is determined by a, which is often called the **shape** or **gamma parameter,** whereas b is a scaling factor.

The gamma distribution is known to be very flexible and takes various shapes, depending on the shape parameter a (Figure 2.3). When $a = \infty$, r is the same for all sites. When $a = 1$, r follows the exponential distribution, indicating that r varies extensively from amino acid site to amino acid site. When $a < 1$, the distribution for r is even more skewed, and a substantial proportion of sites show an r value close to 0 and are practically invariable.

Using parsimony analysis (chapter 8), Uzzell and Corbin (1971) estimated that $a = 2$ for a set of cytochrome c sequences from vertebrates. This indicates that r varies considerably with amino acid site (Figure 2.3). Zhang and Gu (1998) estimated the a value for 51 nuclear and 13 mitochondrial proteins from various vertebrate species and showed that it varies from 0.2 to 3.5. This indicates that the variation of r among sites is very high in some genes.

When r varies following the gamma distribution, it is possible to estimate the number of amino acid substitutions per site. To do this, we con-

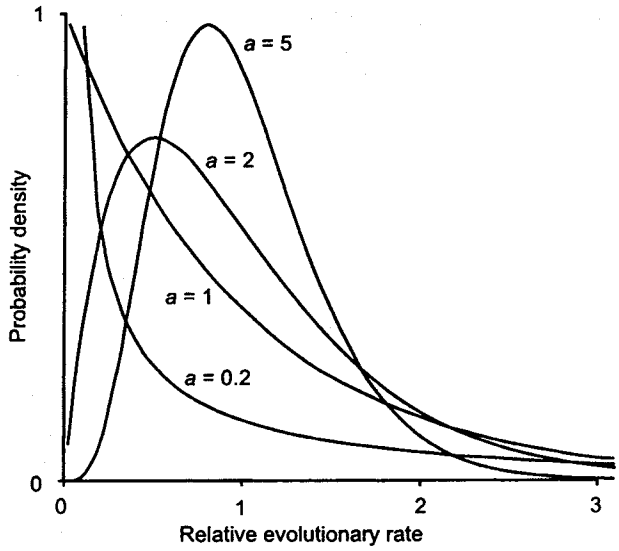


FIGURE 2.3. Gamma distributions of substitution rates among sites for different gamma parameters (a's).

sider the probability of identity of nucleotides at a given site between two sequences at time t, which is given by Equation (2.4). The average of q over all sites is then given by

$$\bar{q} = \int_0^\infty qf(r)dr = \left(\frac{a}{a + 2\bar{r}t}\right)^a \tag{2.11}$$

(Nei et al. 1976a; Ota and Nei 1994b). If we note that the total number of amino acid substitutions per site (d_G) is $2\bar{r}t$ and $\bar{q}$ is given by $1 - p$, we have the following equation for d_G.

$$d_G = a\left[(1 - p)^{-1/a} - 1\right] \tag{2.12}$$

The estimate ($\hat{d}_G$) of d_G is obtained by replacing p by $\hat{p}$. The large-sample variance of $\hat{d}_G$ is given by

$$V(\hat{d}_G) = p\left[(1 - p)^{-(1+2/a)}\right]/n \tag{2.13}$$

In the following, we call $\hat{d}_G$ the **gamma distance** for simplicity. Figure 2.4 shows the relationships between p and d_G for various values of a. The effect of variation in r on the estimate of the number of substitutions is large only when $p > 0.2$ and $a < 0.65$. Therefore, when $p < 0.2$, there is no need to use the gamma distance d_G.

In the above formulation, we considered the variation in substitution rate among different sites. In practice, however, the substitution rate varies not only with amino acid site but also with amino acid pair. For example, amino acids arginine and lysine are both basic and are interchangeable with each other more often than with other amino acids in

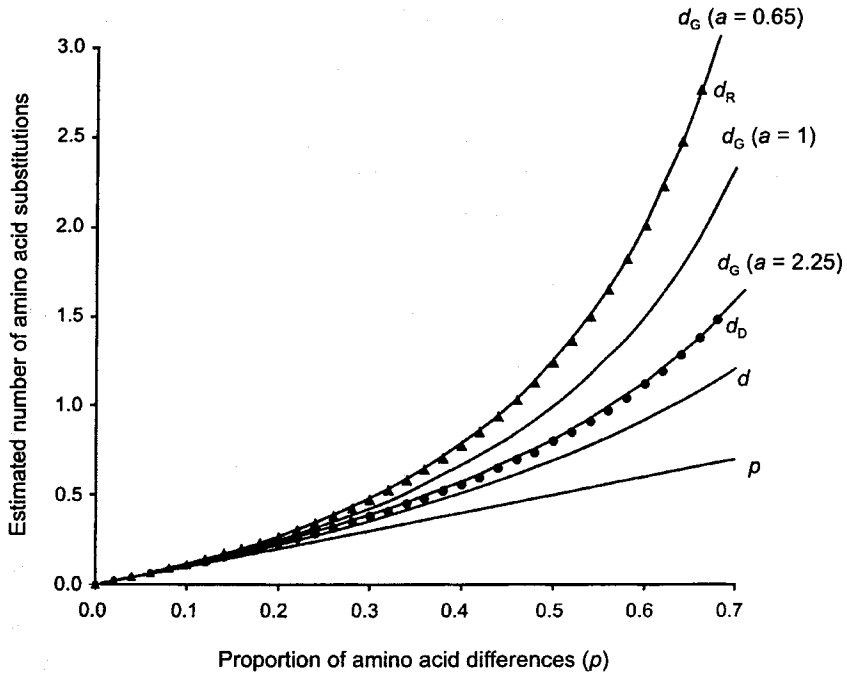


FIGURE 2.4. Relationships of various distance measures with the proportion of amino acid differences (p). Dots indicate the Dayhoff distance (d_D), whereas triangles stand for the Grishin distance (d_R).

the evolutionary process. Grishin (1995) took into account this factor as well as the variation in r and proposed the following formula for estimating the number of amino acid substitutions per site (d_R).

$$q = \frac{\ln(1 + 2d_R)}{2d_R} \tag{2.14}$$

Here, d_R is estimated numerically by solving the above equation for a given value of q. One way of solving the above equation is to use Newton's iteration method. However, the relationship between q and d_R can also be given by Equation (2.12) with $a = 0.65$. In fact, Figure 2.4 shows that the Grishin distance (d_R) can be estimated by

$$\hat{d}_R = 0.65\left[(1 - p)^{-1/0.65} - 1\right] \tag{2.15}$$

as long as $\hat{d}_R \leq 3.0$.

Feng et al. (1997) used the Grishin distance for estimating the time of divergence between eubacteria and eukaryotes, whereas Gogarten et al. (1996) used the gamma distance with $a = 0.7$. Since the Grishin distance is essentially the same as the gamma distance with $a = 0.65$, it is no wonder that the two groups of authors reached very similar conclusions (about 3–4 billion years ago). However, the a value may vary with data set, so it is important to estimate a from the data set under consideration.

For this purpose, one may use Gu and Zhang's (1997) method, which is simple but gives quite accurate estimates.

Example 2.1. Estimation of Evolutionary Distances and the Rate of Amino Acid Substitution in Hemoglobin α Chains

Table 2.1 shows the proportions of different amino acids ($\hat{p}$) for pairwise comparisons of hemoglobin α-chains from six vertebrate species. From these values, we can estimate the PC distance (d) and the gamma distance (d_G). For example, $\hat{p}$ for the human-cow comparison is 0.121. If we put this value into Equation (2.5), we have $\hat{d} = 0.129$. The variance and the standard error of $\hat{d}$ are $V(\hat{d}) = 0.000961$ and $s(\hat{d}) \equiv [V(\hat{d})]^{1/2} = 0.031$, respectively. The values of $\hat{d}$ and $s(\hat{d})$ for other species comparisons are presented in Tables 2.2 and 2.3, respectively. To estimate the rate of amino acid substitution (r), we need the time of divergence (t) in addition to the evolutionary distance (Equation [2.7]). For the human-cow comparison, t is approximately 90 million years (Kumar and Hedges 1998), and $\hat{d} = 0.129$ as computed above. Therefore, $r = 0.129/(2 \times 90 \times 10^6) = 0.717 \times 10^{-9}$ per site per year.

In the human-cow comparison, $\hat{p}$ and $\hat{d}$ are not very different, because $\hat{p}$ is small in this case, and thus the number of multiple hits at a site is small. However, the difference between $\hat{p}$ and $\hat{d}$ increases as $\hat{p}$ increases (see Figure 2.4). This is clear from the human-carp comparison, in which the difference between $\hat{p}(= 0.486)$ and $\hat{d}(= 0.665)$ is much larger than that for the human-cow comparison (Tables 2.1 and 2.2).

The PC distance (d) is derived under the assumption that all amino acid sites evolve at the same rate. If this assumption is not valid, the PC distance would underestimate the number of amino acid substitutions per site, and thus the gamma distance should be used. For computing the gamma distance, we need to know the gamma parameter (a). In the present case, we assume $a = 2$ following Uzzell and Corbin (1971) and compute d_G by using Equation (2.12). For the human-cow comparison, $\hat{p} = 0.121$, and $\hat{d}_G$ becomes 0.134 (Table 2.2). The variance and the standard error of $\hat{d}_G$ are 0.0011225 and 0.034, respectively (Table 2.3). $\hat{d}_G$ is only slightly higher than $\hat{d}$ because $\hat{p}$ is not very large. For the human-carp comparison, however, we have $\hat{d} = 0.665 \pm 0.082$ and $\hat{d}_G = 0.789 \pm 0.115$. (Here, the numbers after the ± sign stand for the standard errors.)

Table 2.2 Poisson-correction (PC) distances (above the diagonal) and gamma distances (below the diagonal) with $a = 2$ for different pairs of hemoglobin α-chains from six different vertebrates.

	Human	Horse	Cow	Kangaroo	Newt	Carp
Human		0.129	0.129	0.205	0.572	0.665
Horse	0.134		0.129	0.232	0.638	0.651
Cow	0.134	0.134		0.197	0.598	0.624
Kangaroo	0.216	0.246	0.207		0.638	0.708
Newt	0.662	0.751	0.697	0.751		0.752
Carp	0.789	0.770	0.733	0.849	0.913	

Table 2.3 Standard error estimates of the PC distances by the analytical (below diagonal) and bootstrap (above diagonal) methods.

	Human	Horse	Cow	Kangaroo	Newt	Carp
Human		0.031	0.031	0.039	0.078	0.083
Horse	0.031		0.030	0.043	0.083	0.081
Cow	0.031	0.031		0.038	0.080	0.079
Kangaroo	0.040	0.043	0.039		0.081	0.084
Newt	0.074	0.080	0.076	0.080		0.090
Carp	0.082	0.081	0.079	0.086	0.089	

Therefore, the difference between $\hat{d}$ and $\hat{d}_G$ increases as $\hat{p}$ increases. The standard error of $\hat{d}$ is smaller than that of $\hat{d}_G$. This is because $\hat{d}$ is based on a one-parameter (r) model, whereas $\hat{d}_G$ is based on a two-parameter (r, a) model. In general, the larger the number of parameters in the model, the greater the variance and the standard error.

2.3. Bootstrap Variances and Covariances

We have shown that there are several ways of estimating the number of amino acid substitutions between two sequences, depending on the mathematical model used. In practice, every model is an approximation to reality and gives only approximate numbers of amino acid substitutions. Therefore, the analytical formulas for computing the variance of distance estimates presented above are also approximate. In the case of $\hat{d}_R$ in Equation (2.14), it is difficult to derive even an approximate formula for the variance. (Of course, if we use Equation [2.15], the variance of $\hat{d}_R$ can be obtained by Equation [2.13] with $a = 0.65$.) Furthermore, when the branch lengths of a phylogenetic tree for many sequences are to be estimated by the least-squares method (see chapter 6), we have to know the variances and covariances of distance estimates for many different pairs of sequences.

In these cases, it is convenient to use the bootstrap method to compute the variances and covariances of various distance measures. The bootstrap method requires no assumptions about the underlying distributions of $\hat{d}$'s except that each amino acid site is assumed to evolve independently (Efron 1982; Efron and Tibshirani 1993). In the present case, this method can be used in the following way.

Suppose we have the following three protein sequences of n amino acids, which are evolutionarily related with one another.

$$\begin{matrix} x_{11}, & x_{12}, & x_{13}, & x_{14}, & x_{15}, & \ldots, & x_{1n} \\ x_{21}, & x_{22}, & x_{23}, & x_{24}, & x_{25}, & \ldots, & x_{2n} \\ x_{31}, & x_{32}, & x_{33}, & x_{34}, & x_{35}, & \ldots, & x_{3n} \end{matrix} \qquad \text{(S2.1)}$$

Here x_{ij} stands for the amino acid at the j-th position of the i-th sequence. We can then compute the $\hat{q}$ values ($\hat{q}_{12}$, $\hat{q}_{13}$, $\hat{q}_{23}$) for the sequences 1 and

2, 1 and 3, and 2 and 3. The PC distance ($\hat{d}_{ij}$) for sequences i and j is given by Equation (2.5) using $\hat{q}_{ij}$. We therefore obtain the estimate of d for each pair of sequences.

In the bootstrap method of computing variances and covariances, a random sample of three amino acid sequences with n amino acid sites is generated from the original data set in (S2.1). This random sample is produced by sampling amino acid sites (columns) with replacement using pseudorandom numbers. This random sample of amino acid sequences may include two or more representations of a given site but no representation of some other sites at all. For example, one random sample produced may have the following sequences with n amino acid sites.

$$\begin{matrix} x_{11}, & x_{12}, & x_{12}, & x_{14}, & x_{15}, & \ldots, & x_{1n-1} \\ x_{21}, & x_{22}, & x_{22}, & x_{24}, & x_{25}, & \ldots, & x_{2n-1} \\ x_{31}, & x_{32}, & x_{32}, & x_{34}, & x_{35}, & \ldots, & x_{3n-1} \end{matrix} \qquad \text{(S2.2)}$$

This is called a *bootstrap resampled data set,* and once this random sample is obtained, an estimate of distance is computed by Equation (2.1), (2.5), (2.12), or (2.15) for each of the three pairs of sequences. If this is repeated B times, we can generate B distance ($\hat{d}$) values for a given pair of sequences. We denote by $\hat{d}_b$ the $\hat{d}$ value for the b-th bootstrap replication. The bootstrap variance is then computed by

$$V_B(\hat{d}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{d}_b - \bar{d})^2 \qquad (2.16)$$

where $\bar{d}$ is the mean of $\hat{d}_b$ over all replications. To compute $V_B(\hat{d})$, it is customary to use about 1000 replications ($B = 1000$).

One assumption often made for the bootstrap is that all sites evolve independently. This assumption of course does not hold in the present case. However, if the number of sites examined is large ($n > 100$) as in the present case, the effect of violation of the assumption is not important, because most sites with different evolutionary rates will be represented in each bootstrap sample.

As mentioned earlier, it is often necessary to compute the covariance of distance estimates for two different pairs of sequences. This covariance can be obtained easily if we use the bootstrap method. It is given by

$$Cov_B(\hat{d}_{ij}, \hat{d}_{kl}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{d}_{ijb} - \bar{d}_{ij})(\hat{d}_{klb} - \bar{d}_{kl}) \qquad (2.17)$$

where subscripts ij and kl refer to sequence pairs i and j and k and l, respectively. This formula is applicable to any distance measure. It is also much simpler than the model-dependent formulas previously proposed by Nei and Jin (1989) and Bulmer (1991).

One advantage of the bootstrap is that it is capable of computing variances and covariances even when no mathematical formulas are available and that it often gives better estimates of the variances and covariances than approximate analytical formulas (Efron and Tibshirani 1993). It is also convenient that the same standard statistical formulas for com-

puting variances and covariances can be used for any distance measure. However, when the original sample is small and is biased, this bias cannot be removed by the bootstrap. In this case, analytical formulas give more accurate variances and covariances than the bootstrap.

Example 2.2. Standard Errors of Poisson Correction (PC) Distances Obtained by the Analytical and Bootstrap Methods

Table 2.3 shows the standard errors of PC distances ($\hat{d}$) obtained by the analytical formula (Equation [2.6]) and the bootstrap method (Equation [2.16]) with $B = 1000$ for the hemoglobin α-chain sequences in Figure 2.1. The $\hat{d}$ values for this data set are given in Table 2.2. It is clear that the standard errors obtained by the two methods are virtually identical. The near-identity of the standard errors obtained by the two methods is also observed for the p and gamma distances. Therefore, the bootstrap method is appropriate for estimating the standard errors of various evolutionary distances.

2.4. Amino Acid Substitution Matrix

Empirical studies of amino acid substitution have shown that substitution occurs more often between amino acids that are similar in terms of biochemical properties such as polarity and volume than between dissimilar amino acids (Dayhoff 1972). In other words, amino acid substitution is generally far from random, and backward and parallel substitutions may occur quite often for similar amino acids. Some amino acids such as cysteine, glycine, and tryptophan rarely change. Unequal rates of substitution at different amino acid sites would also contribute to the inaccuracy of the estimate obtained by the PC distance. To take into account these factors, Dayhoff et al. (1978) proposed another method of estimating evolutionary distance. In this method, the amino acid substitution matrix for a relatively short period of time is considered, and the relationship between the proportion of identical amino acids and the number of amino acid substitutions is derived empirically. The amino acid substitution matrix Dayhoff et al. used was derived from empirical data for many proteins such as hemoglobins, cytochrome c, and fibrinopeptides. They first constructed an evolutionary tree for closely related amino acid sequences and then inferred the relative frequencies of substitutions between different amino acids. From these data, they constructed an empirical amino acid substitution matrix ($\boldsymbol{M}$) for the 20 amino acids (Figure 2.5).

An element (m_{ij}) of this substitution matrix gives the probability that the amino acid in row i changes to the amino acid in column j during one evolutionary time unit. The time unit used in the matrix is the time during which one amino acid substitution per 100 amino acid sites occurs on average. Dayhoff et al. (1978) measured the number of amino acid substitutions in terms of **accepted point mutations (PAM)** and represents one amino acid substitution per 100 amino acid sites. The substitution matrix in Figure 2.5 gives the amino acid substitution probabilities for one PAM.

Replacement

Original	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	1	4	6	1	3	10	21	1	2	3	2	1	1	13	28	22	0	1	13
R Arg	2	9913	1	0	1	9	0	1	8	2	1	37	1	1	5	11	2	2	0	2
N Asn	9	1	9822	42	0	4	7	12	18	3	3	25	0	1	2	34	13	0	3	1
D Asp	10	0	36	9859	0	5	56	11	3	1	0	6	0	0	1	7	4	0	0	1
C Cys	3	1	0	0	9973	0	0	1	1	2	0	0	0	0	1	11	1	0	3	3
Q Gln	8	10	4	6	0	9876	35	3	20	1	6	12	2	0	8	4	3	0	0	2
E Glu	17	0	6	53	0	27	9865	7	1	2	1	7	0	0	3	6	2	0	1	2
G Gly	21	0	6	6	0	1	4	9935	0	0	1	2	0	1	2	16	2	0	0	3
H His	2	10	21	4	1	23	2	1	9912	0	4	2	0	2	5	2	1	0	4	3
I Ile	6	3	3	1	1	1	3	0	0	9872	22	4	5	8	1	2	11	0	1	57
L Leu	4	1	1	0	0	3	1	1	1	9	9947	1	8	6	2	1	2	0	1	11
K Lys	2	19	13	3	0	6	4	2	1	2	2	9926	4	0	2	7	8	0	0	1
M Met	6	4	0	0	0	4	1	1	0	12	45	20	9874	4	1	4	6	0	0	17
F Phe	2	1	1	0	0	0	0	1	2	7	13	0	1	9946	1	3	1	1	21	1
P Pro	22	4	2	1	1	6	3	3	3	0	3	3	0	0	9926	17	5	0	0	3
S Ser	35	6	20	5	5	2	4	21	1	1	1	8	1	2	12	9840	32	1	1	2
T Thr	32	1	9	3	1	2	2	3	1	7	3	11	2	1	4	38	9871	0	1	10
W Trp	0	8	1	0	0	0	0	0	1	0	4	0	0	3	0	5	0	9976	2	0
Y Tyr	2	0	4	0	3	0	1	0	4	1	2	1	0	28	0	2	2	1	9945	2
V Val	18	1	1	1	2	1	2	5	1	33	15	1	4	0	2	2	9	0	1	9901

FIGURE 2.5. Amino acid substitution matrix for the evolutionary distance of 1 PAM (Dayhoff et al. 1978). All elements are multiplied by 10,000.

The amino acid substitution matrix $\boldsymbol{M}$ can be used for predicting amino acid changes for any evolutionary time if we know the initial amino acid frequencies. Let $\boldsymbol{g}_0$ be a row vector of the relative frequencies of 20 amino acids for a polypeptide at time 0. The amino acid frequencies at time t or for t PAMs are then given by

$$\boldsymbol{g}_t = \boldsymbol{g}_0 \boldsymbol{M}_t \tag{2.18}$$

where $\boldsymbol{M}_t = \boldsymbol{M}^t$. Here, we note that element $m_{t(ij)}$ of matrix $\boldsymbol{M}_t$ gives the probability that the amino acid in row i at time 0 will change to the amino acid in column j at time t. In particular, $m_{t(ii)}$ represents the probability that the i-th amino acid at time t is the same as the original one. This probability can be used for relating the proportion of different amino acids between homologous sequences (p) to the number of amino acid substitutions per site (d_D) under the assumption that the amino acid frequencies are at equilibrium and remain the same throughout the evolutionary time. In this case, p is given by

$$p = 1 - \sum_i g_i m_{2t(ii)} \tag{2.19}$$

where g_i is the equilibrium frequency of the i-th amino acid in the sequence under investigation. Here, we use $m_{2t(ii)}$ instead of $m_{t(ii)}$, because we are considering a pair of sequences which diverged t time units ago.

Since $d_D = 0.01 \times t$ (= 0.01 PAMs) and $m_{2t(ij)}$ can be obtained from $\boldsymbol{M}_{2t}$, p can be related to d_D.

In practice, the amino acid frequencies $\boldsymbol{g}$ may vary from protein to protein. To avoid this complication, Dayhoff et al. (1978) used the amino acid frequencies averaged over many different proteins. This does not take into account the specificity of each protein, but it certainly makes the method applicable for many different proteins. Furthermore, if we note that many different proteins have rather similar amino acid frequencies (Dayhoff et al. 1976), this procedure seems to be acceptable for obtaining an estimate of the number of amino acid substitutions.

Using the above method, Dayhoff et al. (1978) derived the relationship between p and d_D. The relationship is given in Figure 2.4. This figure also includes the value of $d = -\ln(1 - p)$ in Equation (2.5) and d_G in Equation (2.12) with $\alpha = 2.25$. As expected, the difference between d_D and d gradually increases as p increases. The d_G value with $\alpha = 2.25$ is very close to d_D for almost all p values, indicating that the d_D can be approximated by d_G with $\alpha = 2.25$.

Although Dayhoff's substitution matrix is still used, Jones et al. (1992) constructed a new matrix based on a large amount of substitution data from many different proteins. Adachi and Hasegawa (1996a) also produced a substitution matrix for 13 mitochondrial proteins in vertebrates. Theoretically, different proteins are expected to have different substitution matrices, so it is desirable to construct a substitution matrix for each group of proteins as more data on amino acid sequences become available. However, for the purpose of measuring the evolutionary distance between a pair of sequences, it is simpler to estimate the gamma parameter α and use the gamma distance given by Equation (2.12). For Jones et al.'s (1992) matrix, the relationship between p and the number of amino acid substitutions is approximately given by Equation (2.12) with $\alpha = 2.4$ (Zhang and Nei 1997).

In recent years, the amino acid substitution matrix has been used in the reconstruction of phylogenetic trees by the maximum likelihood method (Kishino et al. 1990) and in the inference of the amino acid sequences of ancestral proteins (Yang et al. 1995b). These problems will be discussed later.

2.5. Mutation Rate and Substitution Rate

So far we have considered amino acid substitutions as though every mutational change of amino acid were incorporated into the sequence under consideration. In practice, this is not true, because any species consists of a population of many individuals and new mutations occurring in an individual may disappear from the population by chance or by purifying selection. Only when a mutation spreads through the entire population is the mutation incorporated into the genome of the species. This event is called the **fixation** of a mutation in the population. Once a mutation is fixed, every individual of the species has the same mutation. When we compare two amino acid sequences from different species, we

are primarily concerned with amino acid changes that have been incorporated into the genome of the species.

When a mutation occurs, the initial survival of the mutant allele depends largely on chance, whether it is selectively advantageous or not, or whether the population size is small or large. This can be seen in the following way. Let A_1 and A_2 be the wild-type and the mutant alleles in a population, respectively. In a diploid organism, the mutant allele appears first in heterozygous condition (A_1A_2). In a dioecious organism, this individual will mate with a wild-type homozygote (A_1A_1). The mating $A_1A_2 \times A_1A_1$, however, may not produce any offspring for some biological reason other than the effect of the A_2 allele. For example, the mate A_1A_1 may be sterile by chance. The mutant allele will then disappear in the next generation. The survival of the mutant allele is not assured even if $A_1A_2 \times A_1A_1$ produces some offspring. This is because in the offspring the A_1A_2 genotype will appear only with a probability of 1/2. Thus, if two offspring are born from this mating, the chance that no A_1A_2 will appear is 0.25.

When a mutation is neutral and does not affect the fitness of the individual, the relative frequency (x) of the mutant allele may increase or decrease by chance in the population. For simplicity, let us consider an organism with discrete generations such as annual plants or some insects, and let N be the number of adult individuals. The frequency of mutant allele A_2 is initially $x = 1/(2N)$. The fate of this allele is determined purely by chance, and x may increase or decrease. This process will continue until the allele is fixed in or lost from the population (see Nei 1987, Figure 13.6). Since the initial frequency is $1/(2N)$, the probability of fixation is

$$u = 1/(2N) \tag{2.20}$$

whereas the probability of loss is $1 - 1/(2N)$.

Let us now assume that neutral mutations occur at a rate of v per locus per generation, so that the total number of mutations occurring in the entire population is $2Nv$ per locus per generation. Since the proportion of new neutral mutations to be fixed in the population is $1/(2N)$, the **rate of gene substitution** per locus per generation (a) is

$$a = 2Nv \times \frac{1}{2N} = v \tag{2.21}$$

In other words, the rate of gene substitution per locus is equal to the mutation rate. This simple rule was first noticed by Kimura (1968) and King and Jukes (1969).

In the case of amino acid sequence data, we usually consider the substitution rate per amino acid site per year. However, if we redefine the mutation rate as the rate (μ) per amino acid site per year, the above rule obviously applies to neutral mutations. Therefore, the rate of amino acid substitution per site per year (r) is equal to the mutation rate μ. That is,

$$r = 2N\mu \times \frac{1}{2N} = \mu \tag{2.22}$$

As an example, consider a hypothetical polypeptide composed of 100 amino acids and assume that each amino acid mutates to another amino acid with a rate of 10^{-9} per amino acid site per year or 10^{-7} per polypeptide per year. If the effective size of a population is 10^5, the total number of mutations generated for this polypeptide is $2 \times 10^5 \times 10^{-7} = 0.02$ per year. However, since they are fixed with a probability of $1/(2 \times 10^5)$, the rate of amino acid substitution per site is $r = 0.02/(2 \times 10^5 \times 100) = 10^{-9}$, which is equal to the mutation rate per site per year.

What will happen if the mutation is advantageous? For simplicity, let us assume that the relative fitness measured in terms of the number of offspring produced is 1, $1 + s$, and $1 + 2s$ for genotypes A_1A_1, A_1A_2, and A_2A_2, respectively (see Nei 1987; Hartl and Clark 1997). Here s is called the selection coefficient, and it is positive for **advantageous mutations** but negative for **deleterious mutations.** Derivation of the probability of fixation for advantageous mutations is somewhat complicated, but in a large population ($Ns \gg 1$), the probability is given approximately by $2s$ and is independent of N (Crow and Kimura 1970). Therefore, if the rate of advantageous mutations per amino acid site per year is μ, the rate of amino acid substitution by positive selection is given by

$$r = 2N\mu \times 2s = 4Ns\mu \tag{2.23}$$

(King and Jukes 1969).

Suppose $\mu = 10^{-9}$, $N = 10^5$, and $s = 0.01$. Then r becomes 4×10^{-6}, which is 4000 times higher than the rate for neutral mutations. This indicates that the rate of amino acid substitution is enhanced enormously if there is positive selection. Of course, this example is quite artificial. Once a protein function is established in the evolutionary process, most mutations would be deleterious or neutral, and only a few mutations may enhance the activity of the protein. Therefore, the mutation rate for advantageous mutations must be much smaller than the neutral mutation rate, and various statistical studies of protein evolution have suggested that the proportion of amino acid substitutions due to positive selection is quite small (Kimura 1983; Nei 1987). Of course, this issue is somewhat complicated at the DNA level (Kreitman and Akashi 1995), and some of the problems will be discussed later (chapter 12).

It is well known that a large proportion of new mutations is deleterious and that they are quickly eliminated from the population. Therefore, they do not contribute to amino acid substitutions. The problem is the fate of **slightly deleterious mutations.** These mutations may be fixed in the population with a certain (small) probability particularly when the population size N is relatively small (Ohta 1973). However, the continuous accumulation of slightly deleterious mutations in a protein will eventually deteriorate its function. Therefore, if slightly deleterious mutations are fixed, there must be slightly advantageous mutations as well to maintain the function of the protein. In this case, the average rate of amino acid substitution may become more or less constant when long-term evolution is considered.

Under certain conditions, heterozygotes (A_1A_2) for a new mutation may have a higher fitness than either homozygotes (A_1A_1 or A_2A_2). This

type of selection is called **overdominant selection.** In this case, the fitness of A_1A_1, A_1A_2, and A_2A_2 may be written as $1 - s_1$, 1, and $1 - s_2$, respectively. Computation of the fixation probability of overdominant alleles is complicated (Nei and Roychoudhury 1973), but when $s_1 \cong s_2$, the probability is higher than that for neutral mutations. Conducting computer simulations and an analytical study, Maruyama and Nei (1981), Takahata (1990), and Takahata and Nei (1990) have shown that under biologically reasonable conditions overdominant selection can enhance the rate of amino acid substitution (r) several times higher than the neutral rate.

3

Evolutionary Change of DNA Sequences

The evolutionary change of DNA sequences is more complicated than that of protein sequences, because there are various types of DNA regions such as protein-coding regions, noncoding regions, exons, introns, flanking regions, repetitive DNA sequences, and insertion sequences. It is therefore important to know the type and function of the DNA region under investigation. As mentioned in chapter 1, the mutational change of DNA varies extensively with DNA region. Even if we consider protein-coding regions alone, the patterns of nucleotide substitution at the first, second, and third codon positions are not the same. Furthermore, some regions are subject to natural selection more often than other regions, and this also contributes to variation of evolutionary pattern among different regions of DNA.

In this chapter, we are primarily concerned with protein- and RNA-coding regions, because the evolution of these regions is relatively simple and they are important for understanding the general aspects of evolution. Our purpose is to provide statistical methods for studying the evolution of these regions and show how to analyze actual data.

3.1. Nucleotide Differences Between Sequences

When two DNA sequences are derived from a common ancestral sequence, the descendant sequences gradually diverge by nucleotide substitution. A simple measure of the extent of sequence divergence is the proportion (p) of nucleotide sites at which the two sequences are different. This proportion is estimated by

$$\hat{p} = n_d/n \tag{3.1}$$

where n_d and n are the number of different nucleotides between the two sequences and the total number of nucleotides examined, respectively. In the following, we call this the **p distance** for nucleotide sequences. The variance of $\hat{p}$ is given by the same formula as Equation (2.2).

Although the overall nucleotide differences are measured by Equation (3.1), it is often useful to know the frequencies of different nucleotide

Table 3.1 Sixteen different types of nucleotide pairs between sequences X and Y.

Class	Nucleotide Pair				
Identical nucleotides	AA	TT	CC	GG	Total
Frequency	O_1	O_2	O_3	O_4	O
Transition-type pair	AG	GA	TC	CT	Total
Frequency	P_{11}	P_{12}	P_{21}	P_{22}	P
Transversion-type pair	AT	TA	AC	CA	
Frequency	Q_{11}	Q_{12}	Q_{21}	Q_{22}	
	TG	GT	CG	GC	Total
Frequency	Q_{31}	Q_{32}	Q_{41}	Q_{42}	Q

pairs between two sequences, X and Y. Since there are four nucleotides, A, T, C, and G, in each sequence, there are 16 different types of nucleotide pairs. They are given in Table 3.1. In this table, the first letter of each pair of nucleotide symbols stands for the nucleotide for sequence X and the second letter for those of sequence Y. There are four pairs of identical nucleotides (AA, TT, CC, and GG), four transition-type pairs (AG, GA, TC, and CT), and eight transversion-type pairs (all the remaining pairs). We designate the relative frequencies of each pair by the mathematical symbols given in Table 3.1. The total frequencies of identical pairs, transition-type pairs, and transversion-type pairs are denoted by O, P, and Q, respectively. Obviously, we have the relationship $p = P + Q$.

If nucleotide substitution occurs at random among the four nucleotides, Q is expected to be about two times higher than P when p is small. In practice, however, transitions usually occur more frequently than transversions. Therefore, P may be greater than Q. When the extent of sequence divergence is low, the ratio (R) of transitions to transversions, which is often called the **transition/transversion ratio,** can be estimated by

$$\hat{R} = \hat{P}/\hat{Q} \tag{3.2}$$

where $\hat{P}$ and $\hat{Q}$ are the observed values of P and Q, respectively. R is usually 0.5 – 2 in many nuclear genes, but in mitochondrial DNA it can be as high as 15 (Vigilant et al. 1991). Of course, $\hat{R}$ is subject to a large sampling error when the number of nucleotides examined (n) is small, and the variance of $\hat{R}$ is approximately given by

$$V(\hat{R}) = R^2\left(\frac{1}{nP} + \frac{1}{nQ}\right) \tag{3.3}$$

This indicates that unless nP and nQ are large, the variance of $\hat{R}$ can be very large. In actual computation of $V(\hat{R})$, R in the above equation is replaced by $\hat{R}$.

As will be mentioned below, estimation of the number of nucleotide substitutions often depends on the assumption that the nucleotide fre-

quencies in each sequence are in equilibrium and do not change with time. When the nucleotide frequencies in each sequence are in equilibrium, we would expect $P_{11} = P_{12}$, $P_{21} = P_{22}$, $Q_{11} = Q_{12}$, $Q_{21} = Q_{22}$, $Q_{31} = Q_{32}$, and $Q_{41} = Q_{42}$ in Table 3.1. Therefore, the equilibrium condition for nucleotide frequencies can be examined by testing this null hypothesis (Rzhetsky and Nei 1995).

3.2. Estimation of the Number of Nucleotide Substitutions

As in the case of amino acid substitutions, the p distance gives an estimate of the number of nucleotide substitutions per site when the sequences are closely related. However, when p is large, it gives an underestimate of the number, because it does not take into account backward and parallel substitutions. This problem is more serious for nucleotide sequences than for amino acid sequences, because there are only four character states in nucleotide sequences.

To estimate the number of nucleotide substitutions, it is necessary to use a mathematical model of nucleotide substitution. For this reason, many authors have developed different substitution models. Some of the models are presented in the form of substitution rate matrix in Table 3.2. In this book, we consider only simple models that have proved to be use-

Table 3.2 Models of nucleotide substitution.

	A	T	C	G	A	T	C	G
	(A) Jukes-Cantor model				(E) HKY model			
A	-	α	α	α	-	βg_T	βg_C	αg_G
T	α	-	α	α	βg_A	-	αg_C	βg_G
C	α	α	-	α	βg_A	αg_T	-	βg_G
G	α	α	α	-	αg_A	βg_T	βg_C	-
	(B) Kimura model				(F) Tamura-Nei model			
A	-	β	β	α	-	βg_T	βg_C	$\alpha_1 g_G$
T	β	-	α	β	βg_A	-	$\alpha_2 g_C$	βg_C
C	β	α	-	β	βg_A	$\alpha_2 g_T$	-	βg_G
G	α	β	β	-	$\alpha_1 g_A$	βg_T	βg_C	-
	(C) Equal-input model				(G) General reversible model			
A	-	αg_T	αg_C	αg_G	-	ag_T	bg_C	cg_G
T	αg_A	-	αg_C	αg_G	ag_A	-	dg_C	eg_G
C	αg_A	αg_T	-	αg_G	bg_A	dg_T	-	fg_G
G	αg_A	αg_T	αg_C	-	cg_A	eg_T	fg_C	-
	(D) Tamura model				(H) Unrestricted model			
A	-	$\beta\theta_2$	$\beta\theta_1$	$\alpha\theta_1$	-	a_{12}	a_{13}	a_{14}
T	$\beta\theta_2$	-	$\alpha\theta_1$	$\beta\theta_1$	a_{21}	-	a_{23}	a_{24}
C	$\beta\theta_2$	$\alpha\theta_2$	-	$\beta\theta_1$	a_{31}	a_{32}	-	a_{34}
G	$\alpha\theta_2$	$\beta\theta_2$	$\beta\theta_1$	-	a_{41}	a_{42}	a_{43}	-

Note: An element (e_{ij}) of the above substitution matrices stands for the substitution rate from the nucleotide in the i-th row to the nucleotide in the j-th column. g_A, g_T, g_C, and g_G are the nucleotide frequencies. $\theta_1 = g_G + g_C$. $\theta_2 = g_A + g_T$.

ful for actual data analysis. For more sophisticated models, the readers may refer to Zharkikh (1994), Swofford et al. (1996), and Gu and Li (1998) (see also the end of this section).

Jukes and Cantor's Method

One of the simplest models of nucleotide substitution is that of Jukes and Cantor (1969). In this model, it is assumed that nucleotide substitution occurs at any nucleotide site with equal frequency and that at each site a nucleotide changes to one of the three remaining nucleotides with a probability of α per year (or any other time unit) (Table 3.2A). Therefore, the probability of change of a nucleotide to any of the three nucleotides is $r = 3\alpha$. This r is equal to the rate of nucleotide substitution per site per year. Let us now consider two nucleotide sequences, X and Y, which diverged from the common ancestral sequence t years ago. We denote by q_t the proportion of identical nucleotides between X and Y and by p_t $(= 1 - q_t)$ the proportion of different nucleotides. The proportion of identical nucleotides (q_{t+1}) at time $t + 1$ (measured in years) can then be obtained in the following way. First, we note that a site that had the same nucleotide for X and Y at time t will remain the same at time $t + 1$ with probability $(1 - r)^2$ or approximately $1 - 2r$, because r is a small quantity and the term involving r^2 can be neglected. Second, a site that had different nucleotides at time t will have the same nucleotide at time $t + 1$ with probability $2r/3$. This probability is obtained by noting that when sequences X and Y have nucleotides i and j, respectively, at time t, they become identical if i in X changes to j but j in Y remains the same or if j in Y changes to i but i in X remains the same. The probability of occurrence of the first event is $\alpha(1 - r) = r(1 - r)/3$, because i in X has to change to j rather than to one of the other two nucleotides and j in Y has to remain unchanged. The probability of occurrence of the second event is also $r(1 - r)/3$. Therefore, the total probability is $2r(1 - r)/3$, which becomes approximately $2r/3$ if we ignore the term involving r^2.

Therefore, we have the following difference equation.

$$q_{t+1} = (1 - 2r)q_t + \frac{2}{3}r(1 - q_t) \tag{3.4}$$

which can be written as

$$q_{t+1} - q_t = \frac{2r}{3} - \frac{8r}{3}q_t \tag{3.5}$$

Let us now use a continuous time model and represent $q_{t+1} - q_t$ by dq/dt, dropping the subscript t of q_t. We then have the following differential equation.

$$\frac{dq}{dt} = \frac{2r}{3} - \frac{8r}{3}q \tag{3.6}$$

Solution of this equation with the initial condition $q = 1$ at $t = 0$ gives

$$q = 1 - \frac{3}{4}\left(1 - e^{-8rt/3}\right) \tag{3.7}$$

Under the present model, the expected number of nucleotide substitutions per site (d) for the two sequences is $2rt$. Therefore, d is given by

$$d = -(3/4)\ln[1 - (4/3)p] \tag{3.8}$$

where $p = 1 - q$ is the proportion of different nucleotides between X and Y (Jukes and Cantor 1969). An estimate ($\hat{d}$) of d can be obtained by replacing p by the observed value ($\hat{p}$), and the large-sample variance of $\hat{d}$ is given by

$$V(\hat{d}) = \frac{9p(1-p)}{(3-4p)^2 n} \tag{3.9}$$

(Kimura and Ohta 1972). Obviously, the variance of $\hat{d}$ can also be obtained by using the bootstrap method discussed in chapter 2. In MEGA2, the variance of $\hat{d}$ is computed by both the analytical formula and the bootstrap for various substitution models, which are considered in this book.

In the above model, we assumed that the rate of nucleotide substitution is the same for every pair of nucleotides, so the expected frequencies of A, T, C, and G will eventually become equal to 0.25. However, since we have made no assumption about the initial frequencies, Equation (3.8) holds irrespective of the initial frequencies. In other words, there is no need to assume the stationarity of nucleotide frequencies for Equation (3.8) to be applicable (Rzhetsky and Nei 1995).

Kimura's Two-Parameter Method

As mentioned earlier, the rate of transitional nucleotide substitution is often higher than that of transversional nucleotide substitution in real data. Kimura (1980) proposed a method for estimating the number of nucleotide substitutions per site, taking into account this observation. In this model, the rate of transitional substitution per site per year (α) is assumed to be different from that of transversional substitution (2β) (see Figure 1.3 and Table 3.2B). The total substitution rate per site per year (r) is therefore given by $\alpha + 2\beta$. Using this model, Kimura showed that P and Q in Table 3.1 are given by

$$P = (1/4)\left(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}\right) \tag{3.10}$$

$$Q = (1/2)\left(1 - e^{-8\beta t}\right) \tag{3.11}$$

where t is the time after divergence of two sequences X and Y. Therefore, the expected number of nucleotide substitutions per site between X and Y is given by

$$\begin{aligned} d &\equiv 2rt = 2\alpha t + 4\beta t \\ &= -(1/2)\ln(1 - 2P - Q) - (1/4)\ln(1 - 2Q) \end{aligned} \tag{3.12}$$

and the estimate ($\hat{d}$) of d can be obtained by replacing P and Q by the observed values. The variance of $\hat{d}$ is given by

$$V(\hat{d}) = \frac{1}{n}\left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] \tag{3.13}$$

where

$$c_1 = \frac{1}{1 - 2P - Q}, \quad c_2 = \frac{1}{1 - 2Q}, \quad \text{and} \quad c_3 = (c_1 + c_2)/2$$

In the present model it is possible to estimate the numbers of transitional ($s = 2\alpha t$) and transversional ($v = 4\beta t$) nucleotide substitutions per site. The formulas for s and v are given by

$$s = -(1/2)\ln(1 - 2P - Q) + (1/4)\ln(1 - 2Q) \tag{3.14}$$

$$v = -(1/2)\ln(1 - 2Q) \tag{3.15}$$

whereas the variances of the estimates ($\hat{s}$ and $\hat{v}$) of s and v are

$$V(\hat{s}) = \frac{1}{n}\left[c_1^2 P + c_4^2 Q - (c_1 P + c_4 Q)^2\right] \tag{3.16}$$

$$V(\hat{v}) = \frac{1}{n}\left[c_2^2 Q\,(1 - Q)\right] \tag{3.17}$$

where $c_4 = (c_1 - c_2)/2$. Therefore, the transition/transversion ratio is estimated by

$$\hat{R} = \hat{s}/\hat{v} \tag{3.18}$$

With Kimura's model, the equilibrium frequency of each nucleotide is 0.25. However, the above formulas are applicable irrespective of the initial nucleotide frequencies (Rzhetsky and Nei 1995), and in this respect, this model is similar to the Jukes-Cantor model. This property makes the two models applicable to a wider condition than many other models.

Most biologists use the transition/transversion ratio (R) defined either by Equation (3.2) or by Equation (3.18). However, the definition of R varies with mathematical model, and it can be quite complicated when a sophisticated model is used. Theoreticians also tend to prefer to use the transition/transversion rate ratio (k) instead of R. In the case of the Kimura model R is defined as $\alpha/(2\beta)$, whereas k is α/β. Different computer programs often use different definitions of the transition/transversion ratio. Therefore, one should be cautious about the definition used in each computer program.

Tajima and Nei's Method

Tajima and Nei (1984) developed a method of estimating the number of substitutions that seems to be rather insensitive to various disturbing fac-

tors. This method is partly based on the equal-input model (Table 3.2C), which was proposed independently by Felsenstein (1981) and Tajima and Nei (1982). In this method, it is necessary to assume the stationarity of nucleotide frequencies for estimating the number of nucleotide substitution (d), and d is given by

$$d = -b\ln(1 - p/b) \tag{3.19}$$

where

$$b = (1/2)\left[1 - \sum_{i=1}^{4} g_i^2 + p^2/c\right] \tag{3.20}$$

Here, c is given by

$$c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}^2}{2g_i g_j} \tag{3.21}$$

where x_{ij} $(i < j)$ is the relative frequency of nucleotide pair i and j when two DNA sequences are compared. The nucleotide frequencies g_i's are estimated from the two sequences compared. The estimate $(\hat{d})$ of d is then given by replacing p with $\hat{p}$ in Equation (3.19), and the variance of $\hat{d}$ is

$$V(\hat{d}) = \frac{b^2 p(1 - p)}{(b - p)^2 n} \tag{3.22}$$

Note that when the rate of nucleotide substitution is the same for all nucleotide pairs, b is expected to be 3/4 at equilibrium, and Equations (3.19) and (3.22) reduce to Equations (3.8) and (3.9), respectively. In practice, b is usually smaller than 3/4 because of unequal rates of nucleotide substitution, and in this case, Equation (3.19) gives a larger value than the Jukes-Cantor formula.

Tamura's Method

In Kimura's model, the frequencies of the four nucleotides eventually become 0.25, as mentioned earlier. In real data, however, the nucleotide frequencies are rarely equal to one another, and the GC content is often quite different from 0.5. For example, in *Drosophila* mitochondrial DNA, the GC content is known to be about 0.1 (Wolstenholme 1992).

Considering this situation, Tamura (1992) developed a method of estimating d with the substitution model given in Table 3.2D. This model is an extension of Kimura's 2-parameter model to the case of low or high GC content, and d is given by

$$d = -h\ln(1 - P/h - Q) - (1/2)(1 - h)\ln(1 - 2Q) \tag{3.23}$$

where $h = 2\theta(1 - \theta)$, and θ is the GC content.

As in the case of Kimura's method, we can compute $V(\hat{d})$, $\hat{s}$, $V(\hat{s})$, $\hat{v}$, $V(\hat{v})$, $\hat{R}$, and $V(\hat{R})$, but we are not going to present the formulas because they are quite cumbersome (see Kumar et al. 1993). They are incorporated in the computer programs MEGA and MEGA2.

Tamura and Nei's Method

One of the mathematical models often used for phylogenetic inference by the maximum likelihood method is Hasegawa et al.'s (1985a) (Table 3.2E). This (HKY) model is a hybrid of Kimura's 2-parameter model and the equal input model and takes into account both the transition/transversion and GC content biases. However, the formulas for $\hat{d}$ in this model are quite complicated (Rzhetsky and Nei 1995), so we shall not consider this model. Instead, we present Tamura and Nei's (1993) model (Table 3.2F), which includes the HKY model as a special case and permits an analytical solution for d. According to this model, the formula for d is given by

$$\begin{aligned} d = &-\frac{2g_Ag_G}{g_R}\ln\left[1-\frac{g_R}{2g_Ag_G}P_1-\frac{1}{2g_R}Q\right] \\ &-\frac{2g_Tg_C}{g_Y}\ln\left[1-\frac{g_Y}{2g_Tg_C}P_2-\frac{1}{2g_Y}Q\right] \\ &-2\left[g_Rg_Y-\frac{g_Ag_Gg_R}{g_R}-\frac{g_Tg_Cg_R}{g_Y}\right]\ln\left[1-\frac{1}{2g_Rg_Y}Q\right] \end{aligned} \quad (3.24)$$

where P_1 and P_2 are the proportions of transitional differences between A and G and between T and C, respectively, and Q is the proportion of transversional differences (Tamura and Nei 1993). The variance of $\hat{d}$ can be computed by the formula included in MEGA and MEGA2.

Comparison of Different Distance Measures

We considered various distance measures for estimating the number of nucleotide substitutions. Let us now consider the theoretical relationships of these measures assuming $n = \infty$. Figure 3.1 shows the number of nucleotide substitutions estimated by various distance measures when the actual nucleotide substitution follows the Tamura-Nei model. Here $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$, $\alpha_1/\beta = 4$, and $\alpha_2/\beta = 8$ are assumed. Obviously, the estimate obtained by the Tamura-Nei distance is equal to the expected number (d), and all other distance measures give underestimates as the expected number increases. Different distance measures give substantially different results when $d \geq 0.6$. The Tamura distance is virtually identical with the Tamura-Nei distance up to $d = 0.5$, whereas the Tamura, Kimura, and Jukes-Cantor distances are essentially the same as the Tamura-Nei distance when $d \leq 0.25$. Even the p distance becomes very similar to other distance measures when $p \leq 0.1$. Therefore, when one is studying closely related sequences, there is no

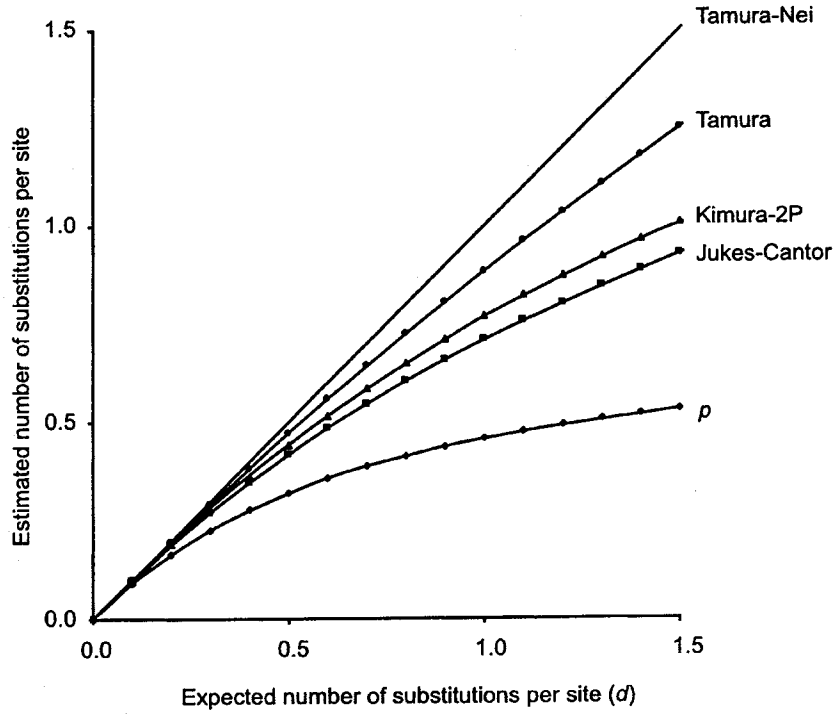


FIGURE 3.1. Estimates of the number of nucleotide substitutions obtained by different distance measures when actual nucleotide substitution follows the Tamura-Nei model. The nucleotide frequencies assumed are $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$; and the two transition/transversion rate ratios assumed are $\alpha_1/\beta = 4$ and $\alpha_2/\beta = 8$.

need to use complex distance measures. In this case, it is better to use a simpler one, because it has a smaller variance.

It should also be noted that for constructing phylogenetic trees from distance measures, sophisticated distances are not necessarily more efficient for obtaining the correct topology than simpler ones even if the mathematical models used fit the data more closely. For estimating branch lengths of a tree, however, a distance measure that fits the data closely usually gives more reliable results. This problem will be discussed in chapter 6.

In recent years, a number of authors proposed more sophisticated distance measures such as the logDet (Steel 1994; Zharkikh 1994) and the paralinear (Lake 1994) distances. The theoretical basis of these distances goes back to Barry and Hartigan (1987) and Rodriguez et al. (1990), but their practical utility still remains unclear.

Example 3.1. Number of Nucleotide Substitutions Between the Human and Rhesus Monkey Cytochrome *b* Genes

The cytochrome *b* gene in animal mitochondrial DNA is highly conserved, and for this reason it is often used for investigating the evolu-

Table 3.3 Observed numbers of the 10 pairs of nucleotides between the DNA sequences for the human and Rhesus monkey mitochondrial cytochrome *b* genes.

	Transition		Transversion				Identical Pair					
Codon Position	TC	AG	TA	TG	CA	CG	TT	CC	AA	GG	n_d	Total (n)
First	21	22	5	1	5	4	68	93	100	56	58	375
Second	20	3	6	1	0	2	140	87	71	45	32	375
Third	60	16	6	5	49	2	11	122	102	2	138	375
All	101	41	17	7	54	8	219	302	273	103	228	1125

Note: The numbers at the first, second, and third codon positions are shown separately.

tionary relationships of distantly related animals. Table 3.3 shows the numbers of ten different types of nucleotide pairs between the DNA sequences of the human and Rhesus monkey cytochrome *b* genes. These numbers are listed separately for the first, second, and third positions of codons. We can easily compute the proportion ($\hat{p}$) of different nucleotides between the two sequences from the data in Table 3.3. For example, the total number of nucleotide differences (n_d) for the first codon positions is 58 (= 21 + 22 + 5 + 1 + 5 + 4), and the total number of nucleotides examined is $n = 375$. Therefore, we have $\hat{p} = 58/375 = 0.155$. Similarly, we can obtain the $\hat{p}$ values for the second and third codon positions (Table 3.4). The $\hat{p}$ value is lowest for the second positions and highest for the third positions. This reflects the fact that no synonymous substitution occurs at the second positions, whereas many synonymous substitutions may occur at the third positions. At the first positions, a minority of potential nucleotide substitutions are synonymous.

Table 3.4 shows the estimates ($\hat{d}$) of the number of nucleotide substitutions obtained by four different methods. The four $\hat{d}$ values for the second positions are very close to one another and are only slightly higher than $\hat{p}$. This indicates that when p is small, the correction for multiple substitutions at the same sites does not really affect the $\hat{d}$ value, irrespective of the method used. The different estimates of d for the first positions are also similar to one another, even though the $\hat{p}$ for the first positions is nearly twice as large as that for the second positions. At the third positions, however, $\hat{p}$ is large, and consequently the correction for multiple substitutions becomes important. Particularly, the Tamura-Nei distance is more than two times larger than $\hat{p}$ and is considerably larger

Table 3.4 Estimates ($\hat{d}$) of the number of nucleotide substitutions per site between the human and Rhesus monkey mitochondrial cytochrome *b* genes for the first, second, and third codon positions ($\hat{d} \times 100$).

Position in Codon	p	Jukes-Cantor	Kimura	Tajima-Nei	Tamura-Nei
First	15.5 ± 1.9	17.3 ± 2.4	17.8 ± 2.5	18.0 ± 2.6	17.9 ± 2.5
Second	8.5 ± 1.4	9.1 ± 1.6	9.2 ± 1.7	9.2 ± 1.7	9.3 ± 1.7
Third	36.8 ± 2.5	50.6 ± 4.9	52.3 ± 5.4	66.5 ± 9.4	87.9 ± 39.0

than the other correction distances. However, these multiple substitutions at the third positions are largely synonymous and do not necessarily change amino acids.

3.3. Gamma Distances

In the evolutionary distances considered above, the rate of nucleotide substitution is assumed to be the same for all nucleotide sites. In reality, this assumption rarely holds, and the rate varies from site to site. In the case of protein-coding genes, this is obvious, because the first, second, and third codon positions have different substitution rates (Table 3.4). The functional constraint of amino acids at the active centers of proteins also contributes to rate variation among nucleotide sites. Rate variation is observed in RNA coding genes as well, because RNAs have functional constraints and usually form a secondary structure consisting of loops and stems that have different substitution rates. Statistical analyses of the rates of substitution at different nucleotide sites have suggested that the rate variation approximately follows the gamma distribution given by Equation (2.9) (Kocher and Wilson 1991; Tamura and Nei 1993; Wakeley 1993, 1994).

For this reason, a number of authors developed gamma distances appropriate for nucleotide substitutions. The gamma distances can be derived by the same mathematical methods as that used for deriving the distance for amino acid sequences.

Gamma Distance for the Jukes-Cantor Model

When the nucleotide substitution at each site follows the Jukes-Cantor model but the substitution rate (r) varies with the gamma distribution, the gamma distance becomes

$$d = \frac{3}{4}a\left[\left(1 - \frac{4}{3}p\right)^{-1/a} - 1\right] \tag{3.25}$$

(Golding 1983; Nei and Gojobori 1986), whereas the variance of the estimate $\hat{d}$ of d is given by

$$V(\hat{d}) = \frac{p(1-p)}{n}\left[\left(1 - \frac{4}{3}p\right)^{-2(1/a+1)}\right] \tag{3.26}$$

(Jin and Nei 1990). Here a is the gamma parameter defined in chapter 2.

Gamma Distance for the Kimura Model

For this model, the gamma distance (d) and the variance of the estimate ($\hat{d}$) of d have been derived by Jin and Nei (1990).

$$d = \frac{a}{2}\left[(1 - 2P - Q)^{-1/a} + \frac{1}{2}(1 - 2Q)^{-1/a} - \frac{3}{2}\right] \tag{3.27}$$

$$V(\hat{d}) = \frac{1}{n}\left[c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2\right] \tag{3.28}$$

where $c_1 = (1 - 2P - Q)^{-(1/a+1)}$, $c_2 = (1 - 2Q)^{-(1/a+1)}$, $c_3 = (c_1 + c_2)/2$, and P and Q are the same as those for Kimura's model.

As in the case of Kimura's distance, we can compute the number of transitional (s) and transversional (v) substitutions per site. They become

$$s = \frac{a}{2}\left[(1 - 2P - Q)^{-1/a} - \frac{1}{2}(1 - 2Q)^{-1/a} - \frac{1}{2}\right] \tag{3.29}$$

$$v = \frac{a}{2}\left[(1 - 2Q)^{-1/a} - 1\right] \tag{3.30}$$

Formulas for obtaining $\hat{R}$ and its variance are included in MEGA and MEGA2.

Gamma Distance for the Tamura-Nei Model

In the control region of mammalian mitochondrial DNA, the rate of nucleotide substitution is known to vary extensively from site to site, and there is a strong transition/transversion bias. The gamma distance for the Tamura-Nei model was developed primarily for the sequence data from this region. There are two hypervariable segments (5′ and 3′ segments in this region), and the middle section is highly conserved. Using human data, Kocher and Wilson (1991) and Tamura and Nei (1993) estimated that a is about 0.11 for the entire control region, whereas Wakeley (1993) obtained $a = 0.47$ for the 5′ hypervariable segment. The gamma distance for the Tamura-Nei model is given by

$$\begin{aligned} d = 2a\Bigg[& \frac{g_A g_G}{g_R}\left(1 - \frac{g_R}{2g_A g_G}P_1 - \frac{1}{2g_R}Q\right)^{-1/a} \\ & + \frac{g_T g_C}{g_Y}\left(1 - \frac{g_Y}{2g_T g_C}P_2 - \frac{1}{2g_Y}Q\right)^{-1/a} \\ & + \left(g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y}\right)\left(1 - \frac{1}{2g_R g_Y}Q\right)^{-1/a} \\ & - g_A g_G - g_T g_C - g_R g_Y\Bigg] \end{aligned} \tag{3.31}$$

A formula for the variance of the estimate ($\hat{d}$) of d is also available (Tamura and Nei 1993). Tamura and Nei derived formulas for the estimates of the average numbers of transitional ($\hat{s}$) and transversional ($\hat{v}$)

substitutions per site, the transition/transversion ratio ($\hat{R}$), and their variances. These formulas are incorporated in MEGA and MEGA2.

In the above formulation, we have assumed that the gamma parameter a is known. In practice, however, we must estimate a from data using many DNA sequences simultaneously. There are several different methods for estimating a (e.g., Kocher and Wilson 1991; Yang 1994b; Sullivan et al. 1995; Yang and Kumar 1996), but the maximum likelihood method (Yang 1994b) is preferable. Yang (1996a) has computed the a value for various nuclear and mitochondrial genes.

Gamma distances are generally more realistic than non-gamma distances, but they have larger variances than the latter. For this reason, they do not necessarily produce better results in phylogenetic inference unless the number of nucleotides used is very large. For estimating branch lengths of a tree, gamma distances generally give better results.

3.4. Numerical Estimation of Evolutionary Distances

We have presented analytical formulas for computing various evolutionary distances. Most of them are based on relatively simple mathematical models of nucleotide substitution, and the estimates are the same as the maximum likelihood estimates under the assumptions made. However, if one wishes to estimate the distances based on complicated models such as the Tamura-Nei gamma or the general reversible model, it is often more convenient to compute the distances numerically by using the maximum likelihood method. This method is particularly useful for estimating the distances for many pairs of sequences. For example, the analytical formula for computing the Tamura-Nei gamma distance (Equation [3.31]) requires information on the nucleotide frequencies as well as of the gamma parameter a. The nucleotide frequencies are usually estimated from the two DNA sequences compared, whereas the a value is obtained from data with additional sequences.

If we use the numerical method, however, it is possible to estimate the nucleotide frequencies as well as the a value from the sequence data used as long as there are a sufficient number of sequences available (Yang 1995b; Swofford 1998). In other words, we can estimate all the substitution parameters as well as the distance value by maximizing the likelihood for the given data set and for a given topology (see chapter 8). This will give the distance estimates for all pairwise comparisons simultaneously. It is also possible to choose the most appropriate substitution model by using the likelihood ratio test (see chapter 9). The computer program for estimating pairwise distances by this method is available in PAUP*.

Theoretically, this method appears to give good distance estimates for DNA sequence data as long as the number of nucleotides examined (n) is very large and the topology used is more or less correct. This is particularly so for estimating branch lengths or for estimating evolutionary times. When n is relatively small, however, the distance estimates obtained in this way do not necessarily give good estimates of the true tree

topology, as will be discussed later. Estimation of tree topologies is a complicated statistical problem, and simple distance measurements often give better results (see chapter 9).

3.5. Alignment of Nucleotide Sequences

So far we have assumed that the two homologous nucleotide sequences compared have no insertions and no deletions and thus can be compared directly. In practice, the numbers of nucleotides of the sequences compared are often different, and we must infer the locations of insertions and deletions to align the two sequences. Both insertions and deletions introduce **gaps** in the DNA sequence alignment. When sequence divergence is low, the positions of gaps can be inferred relatively easily by visual inspection, particularly in protein-coding regions of DNA. However, when sequence divergence is high or when the simultaneous alignment of many sequences is necessary, sequence alignment is not a simple matter and is done by using computer programs. Several methods have been developed for aligning two or more sequences. These methods can be used for both nucleotide and amino acid sequences. Here we present only the principle of sequence alignment.

Alignment of Two Sequences

Let us consider the following two nucleotide sequences.

$$\begin{array}{ll} (1) & \text{ATGCGTCGTT} \\ (2) & \text{ATCCGCGAT} \end{array} \qquad \text{(A1)}$$

Sequence 1 has ten nucleotides and sequence 2 has nine nucleotides. Therefore, at least one gap must be introduced in the alignment of these sequences. A simple way of aligning these sequences is to do a two-dimensional comparison, as given in Figure 3.2. In this comparison (**dot matrix**), dots are given when the nucleotides in sequences 1 and 2 are identical. If the two sequences are identical, there will be a diagonal line of dots. If the sequences are identical except for a gap in one of the two sequences, the diagonal line will shift up or down in the middle of the line. Therefore, we can identify the gap. In practice, there are usually some nucleotide differences between the two sequences in addition to gaps, and this makes it difficult to identify the gaps. For this reason, several mathematical methods have been developed to make a reasonable alignment.

One of the most popular methods of sequence alignment is that developed by Needleman and Wunsch (1970). In this method, the similarity between two sequences is measured by a **similarity index,** and the alignment of the two sequences that maximizes the similarity index is chosen. Later, Sellers (1974) developed another method in which the distance between two sequences (**alignment distance**) is measured by an index and the alignment that minimizes this distance is chosen. However, Smith et al. (1981) have shown that the two methods are essentially the

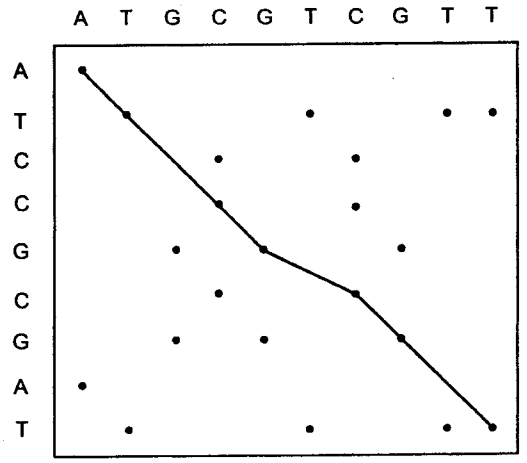


FIGURE 3.2. A dot matrix for aligning two sequences.

same and give the same result in most cases. In the following, we consider the distance method rather than the similarity method.

Consider two DNA sequences A and B of length m and n, respectively. An alignment between sequences A and B is defined as an ordered sequence of nucleotide pairs, each pair containing one nucleotide from each sequence or a nucleotide of either sequence and a null element with the order of the original sequence preserved. Deletions or insertions (**gaps**) are indicated by alignment pairs containing a null element (−). For example, the following alignment

$$\begin{array}{l} \text{ATGC–GTCGTT} \\ \text{AT–CCG–CGAT} \end{array} \tag{A2}$$

contains three gaps of one element (length one), seven pairs of matched elements, and one pair of mismatched elements.

To measure the distance between two sequences, let α be the number of pairs of matched elements, β be the number of pairs of mismatched elements, and γ be the number of gaps irrespective of gap length. The distance between the two sequences may then be measured by

$$E = \text{Min}(w_1\beta + w_2\gamma) \tag{3.32}$$

where Min(•) stands for the minimum value of $w_1\beta + w_2\gamma$ among all possible alignments. Here, w_1 and w_2 are the penalties for a mismatch and for a gap, respectively. However, the assumption that the penalty for a gap is the same irrespective of its length is unrealistic. It is therefore customary to assume that a gap penalty is a function of gap length (Gotoh 1982; Gu and Li 1995). Similarly, nucleotide mismatches can be divided into transitional and transversional mismatches, and different penalties may be given to them. For amino acid sequences, different amino acid mismatches are often given different mismatch penalties based on Day-

hoff's amino acid substitution matrix (Dayhoff et al. 1978). The computation of E is quite complicated, but various algorithms for fast computation have been developed (e.g., Myers and Miller 1988).

Let us now illustrate how to compute the alignment distance using $w_1 = 1$ and $w_2 = 4$ in Equation (3.32). For alignment (A2), we have $E = 1 + 4 \times 3 = 13$, because $\beta = 1$ and $\gamma = 3$. By contrast, the following alignment has two mismatches ($\beta = 2$) and one gap ($\gamma = 1$).

$$\begin{array}{l} \text{ATGCGTCGTT} \\ \text{ATCCG–CGAT} \end{array} \qquad \text{(A3)}$$

Therefore, the alignment distance is $E = 2 + 4 = 6$. This value is smaller than that of (A2), and thus alignment (A3) is considered to be better than (A2). Incidentally, alignment (A3) is the same as that obtained from the dot matrix in Figure 3.2.

The choice of alignment depends considerably on the relative values of w_1 and w_2. If we use a small value of w_2 relative to w_1, we will have an alignment with many gaps and few nucleotide mismatches. However, since deletions and insertions occur less frequently than nucleotide substitutions, such an alignment would be unrealistic. Therefore, it is advisable to make w_2 greater than w_1. For a detailed discussion of this problem, see Vingron and Waterman (1994) and Taylor (1996).

Alignment of Multiple Sequences

When multiple sequences are to be aligned, it is customary to use a **progressive alignment algorithm** (Taylor 1987; Feng and Doolittle 1996; Hein and Støvlbæk 1996; Higgins et al. 1996). In this algorithm, pairs of sequences with small distances are first aligned, and the alignment of more distantly related sequences is done progressively for larger and larger groups. In the initial stage of this procedure, the pairs of sequences are aligned following the algorithm mentioned above. In the next stage, we need to align groups of sequences with each other. This is done by the **profile alignment algorithm,** which is similar to the alignment algorithm for two sequences, except that the average distance is now computed by considering all the nucleotides (or amino acids) at every position in the two groups of sequences. If a gap is introduced, it is inserted in all the sequences of the same group.

In the progressive alignment algorithm, the order in which the sequences are aligned is crucial. In Higgins et al.'s approach (1996), this order is determined by inferring a tree-like relationship among the sequences based on the matrix of pairwise distance scores between sequences. For this purpose, the distance scores (E's) are first estimated, and a phylogenetic tree is constructed by the neighbor joining method described in chapter 6. The progressive alignment algorithm is known to be fast and give reasonable results. However, there is no guarantee that the best alignment has been found.

For protein-coding DNA sequences, alignment at the amino acid sequence level is generally more reliable than that at the nucleotide level,

because amino acid sequences evolve much more slowly. In this case, the nucleotide sequences may be aligned after the alignment of amino acid sequences. Information on the secondary and higher level structure of proteins or RNAs is also useful in multiple sequence alignment. For example, the secondary structure of ribosomal RNAs is routinely used to align sequences from distantly related organisms such as animals, plants, fungi, and bacteria.

3.6. Handling of Sequence Gaps in the Estimation of Evolutionary Distances

The presence of alignment gaps in sequences introduces some complications in the estimation of evolutionary distances. Furthermore, sites with missing information can also occur because of experimental difficulties, and they create the same problem as that for gaps. These sites are generally ignored in the distance estimation, but there are two different ways of treating them. One way is to delete all of these sites from data analysis. This option, called the **complete-deletion** option, is generally desirable because different regions of DNA or amino acid sequences often evolve differently. However, if the number of nucleotides involved in a gap is small and gaps are distributed more or less at random, one may compute a distance for each pair of sequences ignoring only those gaps that are involved in the two sequences compared. This option is called the **pairwise-deletion** option.

To illustrate the two different ways of computing distances, let us consider the following three DNA sequences.

$$\begin{array}{l} \text{A-AC-GGAT-AGGA-ATAAA} \\ \text{AT-CC?GATAA?GAAAAC-A} \\ \text{ATTCC-GA?TACGATA-AGA} \end{array} \qquad \text{(A4)}$$

Here, the alignment gaps are indicated by a hyphen (-), whereas missing-information sites are denoted by a question mark (?). Table 3.5 shows the

Table 3.5 Complete-deletion and pairwise-deletion options.

		Differences/Comparisons		
Option	Sequences	(1,2)	(1,3)	(2,3)
Complete-deletion		1/10	0/10	1/10
(1) A C GA A GA A A A				
(2) A C GA A GA A C A				
(3) A C GA A GA A A A				
Pairwise-deletion		2/12	3/13	3/14
(1) A-AC-GGAT-AGGA-ATAAA				
(2) AT-CC?GATAA?GAAAAC-A				
(3) ATTCC-GA?TACGATA-AGA				

results of computation using the complete-deletion and the pairwise-deletion options. In the former option all gap sites and the ? mark sites are deleted, so that there are 10 nucleotide sites to be compared and the p distances between sequences 1 and 2, 1 and 3, 2 and 3 become 0.1, 0.0, and 0.1, respectively. In the pairwise-deletion option, the number of nucleotides to be compared varies with sequence pair, and the p distance also varies with sequence pair.

4

Synonymous and Nonsynonymous Nucleotide Substitutions

In chapter 3, we have seen that the rate of nucleotide substitution is much higher at the third positions of codons than at the first and second positions. This is caused by the fact that many nucleotide substitutions at the third positions are silent and do not change amino acids. However, not all substitutions at the third positions are silent. Furthermore, some silent substitutions may also occur at the first positions. It is therefore interesting to know the rates of **synonymous** and **nonsynonymous substitution** separately. Since synonymous substitutions are apparently free from natural selection, the rate of synonymous substitution is often equated to the rate of neutral nucleotide substitution (Miyata et al. 1980). Indeed, the rate of synonymous substitution is similar for many genes, unless it is disturbed by codon usage bias and other factors. By contrast, the rate of nonsynonymous substitution is generally much lower than that of synonymous substitution and varies extensively from gene to gene. This is considered to be due to purifying selection, the extent of which varies from gene to gene (Kimura 1983).

However, it is important to note that there are genes in which nonsynonymous substitutions occur at a higher rate than synonymous substitutions (e.g., Hughes and Nei 1988; Lee et al. 1995). These nonsynonymous substitutions are apparently caused by positive Darwinian selection, because under neutral evolution one would expect that the rates of synonymous and nonsynonymous substitution are equal to each other. For these reasons, estimation of the rates of synonymous and nonsynonymous substitution has become an important subject in the study of molecular evolution.

Estimation of the rates of synonymous and nonsynonymous substitution is more complicated than that of the total number of nucleotide substitutions. In most nucleotide sequences there are more nucleotide sites that potentially produce nonsynonymous mutations than sites that potentially produce synonymous mutations, and the numbers of synonymous and nonsynonymous sites vary from gene to gene. Therefore, the rates of synonymous and nonsynonymous substitution should be defined as the number of synonymous substitutions per synonymous site (r_S) and the number of nonsynonymous substitutions per nonsynonymous site (r_N) per year or per generation. In practice, we usually do not

know the time of divergence (t) between two DNA sequences compared. Therefore, it is customary to consider the number of synonymous substitutions per synonymous site ($d_S = 2r_S t$) and the number of nonsynonymous substitutions per nonsynonymous site ($d_N = 2r_N t$) for a pair of sequences.

There are several methods for estimating d_S and d_N. They can be classified into three groups: (1) evolutionary pathway methods, (2) methods based on Kimura's 2-parameter model, and (3) maximum likelihood methods with codon substitution models. These methods are based on different assumptions, and therefore they do not necessarily give the same results. In this chapter, we explain the first two groups of methods in detail, since they are commonly used in the literature. In the following, we consider the standard genetic code, but the same formulation can be made for any genetic code discussed in chapter 1.

4.1. Evolutionary Pathway Methods

This approach was first used by Miyata and Yasunaga (1980). They considered all possible evolutionary pathways between each pair of homologous codons of two DNA sequences and developed a method for estimating d_S and d_N. However, their method is quite complicated, because every nucleotide substitution is weighted by the likelihood of occurrence of the substitution, taking into account the similarity of amino acids encoded. Conducting a computer simulation, Nei and Gojobori (1986) showed that Miyata and Yasunaga's weighting for different pathways is not necessary and that a simple unweighted version gives essentially the same results as those given by Miyata and Yasunaga. We therefore present Nei and Gojobori's (1986) unweighted pathway method and its modifications.

Nei-Gojobori Method

In Nei and Gojobori's method, d_S and d_N are estimated by computing the numbers of synonymous and nonsynonymous substitutions and the numbers of potentially synonymous and potentially nonsynonymous sites. Let us first consider the numbers of potentially synonymous and potentially nonsynonymous sites. In Nei and Gojobori's method, these numbers are computed for each codon under the assumption of equal probabilities of all nucleotide changes. We denote by f_i the proportion of synonymous changes (the ratio of the number of synonymous changes to the sum of synonymous and nonsynonymous changes, excluding nonsense mutations) at the i-th nucleotide position of a codon ($i = 1, 2, 3$). The numbers of potentially synonymous (s) and potentially nonsynonymous (n) sites for this codon are then given by $s = \Sigma_{i=1}^{3} f_i$ and $n = 3 - s$, respectively. For example, in the case of phenylalanine codon TTT, s becomes

$$s = 0 + 0 + \frac{1}{3} \tag{4.1}$$

because no nucleotide changes at the first and second positions result in synonymous codons and at the third position one out of the three possi-

ble changes results in a synonymous codon (TTC). Since all other changes are nonsynonymous, n is given by $3 - 1/3 = 8/3$. When any nucleotide change results in a termination codon, this change is disregarded. For example, a nucleotide change at the third position of the cysteine codon TGT results in a termination codon when T changes to A, but it gives a synonymous codon when T changes to C and a nonsynonymous codon (Trp) when T changes to G. Therefore, $f_3 = 1/2$ in this case. Since $f_1 = f_2 = 0$ for this codon, we have $s = 0.5$ and $n = 2.5$.

To obtain the total numbers of synonymous sites (S) and nonsynonymous sites (N) for the entire sequence, we use the formulas $S = \Sigma_{j=1}^{C} s_j$ and $N = 3C - S$, where s_j is the value of s for the j-th codon and C is the total number of codons. In practice, we compare two sequences, so the average values of S and N for the two sequences are used in actual computation. Note that $S + N = 3C$ is equal to the total number of nucleotides compared.

Let us now compute the numbers of synonymous and nonsynonymous nucleotide differences between a pair of homologous sequences. We compare the two sequences, codon by codon, and count the number of nucleotide differences for each pair of codons compared. When there is only one nucleotide difference, we can immediately decide whether the difference is synonymous or nonsynonymous. For example, if the codon pairs compared are GTT (Val) and GTA (Val), there is one synonymous difference. We denote by s_d and n_d the numbers of synonymous and nonsynonymous differences per codon, respectively. In the present example, $s_d = 1$ and $n_d = 0$. When two nucleotide differences exist between the two codons compared, there are two possible ways to obtain the differences. For example, in the comparison of TTT and GTA there are two possible parsimonious pathways between the two codons. That is,

(1) TTT (Phe) ↔ GTT (Val) ↔ GTA (Val)
(2) TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val)

Pathway (1) involves one synonymous and one nonsynonymous substitutions, whereas pathway (2) involves two nonsynonymous substitutions. We assume that pathways (1) and (2) occur with equal probability. The numbers of synonymous and nonsynonymous differences then become $s_d = 0.5$ and $n_d = 1.5$, respectively. In some comparisons of codons, there are pathways in which termination codons are involved. We eliminate these pathways from the computation.

When there are three nucleotide differences between the codons compared, there are six different possible pathways between the codons, and in each pathway there are three mutational steps. Considering all these pathways and mutational steps, one can again count the numbers of synonymous and nonsynonymous differences in the same way as in the case of two nucleotide differences. For example, if the two codons compared are TTG and AGA, there are following six pathways.

(1) TTG (Leu) ↔ ATG (Met) ↔ AGG (Arg) ↔ AGA (Arg)
(2) TTG (Leu) ↔ ATG (Met) ↔ ATA (Ile) ↔ AGA (Arg)
(3) TTG (Leu) ↔ TGG (Trp) ↔ AGG (Arg) ↔ AGA (Arg)

(4) TTG (Leu) ↔ TGG (Trp) ↔ TGA (Ter) ↔ AGA (Arg)
(5) TTG (Leu) ↔ TTA (Leu) ↔ ATA (Ile) ↔ AGA (Arg)
(6) TTG (Leu) ↔ TTA (Leu) ↔ TGA (Ter) ↔ AGA (Arg)

Pathways (4) and (6) involve a termination codon, so they are disregarded. The numbers of synonymous substitutions in pathways (1), (2), (3), and (5) are 1, 0, 1, and 1, respectively, whereas the numbers of nonsynonymous substitutions are 2, 3, 2, and 2, respectively. Since we assume that the four pathways are equally probable, we have $s_d = 3/4$ and $n_d = 9/4$.

The total numbers of synonymous and nonsynonymous differences for a sequence comparison can be obtained by summing up these values over all codons. That is, $S_d = \Sigma_{j=1}^{C} s_{dj}$ and $N_d = \Sigma_{j=1}^{C} n_{dj}$, where s_{dj} and n_{dj} are the numbers of synonymous and nonsynonymous differences for the j-th codon, and C is the number of codons compared. Note that $S_d + N_d$ is equal to the total number of nucleotide differences between the two DNA sequences compared.

We can, therefore, estimate the proportions of synonymous (p_S) and nonsynonymous (p_N) differences by the equations

$$\hat{p}_S = S_d/S, \qquad \hat{p}_N = N_d/N \tag{4.2}$$

where S and N are the average numbers of synonymous and nonsynonymous sites for the two sequences compared. To estimate the numbers of synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions per site, we use the Jukes-Cantor method (Equation [3.8]) replacing p by $\hat{p}_S$ or $\hat{p}_N$. This method, of course, gives only approximate estimates of d_S and d_N, because the nucleotide substitution at the synonymous and nonsynonymous sites does not really follow the Jukes-Cantor model, as noted by Perler et al. (1980). Despite this theoretical problem, computer simulation has shown that Equation (3.8) gives good estimates of synonymous and nonsynonymous substitutions, as long as the nucleotide frequencies of A, T, C, and G are nearly equal and there is no significant transition/transversion bias (Ota and Nei 1994c).

The approximate large-sample variances of $\hat{d}_S$ and $\hat{d}_N$ can be computed by Equation (3.9) if we replace $\hat{p}$ in the equation by $\hat{p}_S$ or $\hat{p}_N$ and n by S or N (Nei 1987). Theoretically, more accurate large-sample variances [$V(\hat{d}_S)$ and $V(\hat{d}_N)$] of $\hat{d}_S$ and $\hat{d}_N$ are given by

$$V(\hat{d}_S) = V(\hat{p}_S)\Big/\left(1 - \frac{4}{3}p_S\right)^2, \qquad V(\hat{d}_N) = V(\hat{p}_N)\Big/\left(1 - \frac{4}{3}p_N\right)^2 \tag{4.3}$$

where

$$V(\hat{p}_S) = \sum_{i=1}^{C}(s_{di} - p_S s_i)^2/S^2, \qquad V(\hat{p}_N) = \sum_{i=1}^{C}(n_{di} - p_N n_i)^2/N^2 \tag{4.4}$$

(Ota and Nei 1994c). However, computer simulation has shown that the above formulas give nearly the same results as those obtained by Equations (3.9).

Another way of computing the variances of $\hat{d}_S$, $\hat{d}_N$, $\hat{p}_S$, and $\hat{p}_N$ is to use

the bootstrap method explained below. As long as S_d, S_n, S, and N are sufficiently large, the bootstrap is expected to give more accurate variances than the above analytical formulas, because it does not depend on the assumption that the expectations of s_{di} and n_{di} are given by $p_S s_i$ and $p_N n_i$, respectively.

Large-Sample Test of the Difference Between $\hat{d}_S$ and $\hat{d}_N$ or Between $\hat{p}_S$ and $\hat{p}_N$

To detect positive Darwinian selection, it is necessary to show that $\hat{d}_N$ is significantly greater than $\hat{d}_S$. A simple way to test the null hypothesis of $d_N = d_S$ is to compute the difference $\hat{D} = \hat{d}_N - \hat{d}_S$ and the variance $[V(\hat{D})]$ of $\hat{D}$ and conduct the normal deviate or the Z test under the assumption that S_d and N_d are sufficiently large (>10) so that the distribution of $\hat{D}$ approximately follows the normal distribution. In the present case, $V(\hat{D})$ is approximately given by $V(\hat{d}_N) + V(\hat{d}_S)$, because $\hat{d}_N$ and $\hat{d}_S$ are theoretically independent of each other. Therefore, we have

$$Z = \hat{D}/s(\hat{D}) \tag{4.5}$$

where $s(\hat{D}) = [V(\hat{D})]^{1/2}$. Here we are interested in $d_N > d_S$, so that the test will be a one-tail test. The Z values for the significance levels of 5, 1, and 0.1% in this case are 1.64, 1.96, and 2.81, respectively. This test corresponds to the t test with an infinite number of degrees of freedom.

The variance of $\hat{D} = \hat{d}_N - \hat{d}_S$ can also be computed by the bootstrap method. In the bootstrap method, a pair of random codon sequences consisting of the same number of codons as that of the original sequences are generated by the resampling method described in chapter 2. In the present case, however, codons rather than nucleotides are the units of resampling. For the b-th bootstrap sample of codon sequences, we compute estimates ($\hat{p}_{Sb}$, $\hat{p}_{Nb}$, $\hat{d}_{Sb}$, and $\hat{d}_{Nb}$) of p_S, p_N, d_S, and d_N. Therefore, if we repeat this computation about 1000 times, we can compute the variances of these quantities using Equation (2.16). For testing the observed difference $\hat{D} = \hat{d}_N - \hat{d}_S$, we do not really need the variances of $\hat{d}_S$ and $\hat{d}_N$. Instead, we can compute the standard error of $\hat{D}$ directly by using the bootstrap and then use the Z test. An even simpler method would be to compute the number (B_S) of bootstrap replications in which the bootstrap estimate ($\hat{D}_b$) of D is smaller than 0 and compute the proportion of these replications, B_S/B, where B is the total number of bootstrap replications. This proportion is called the **achieved significance level** (*ASL*) to distinguish it from the model-based significance level (Efron and Tibshirani 1993). If *ASL* is less than 5 or 1%, one may conclude that the observed $\hat{D}$ is significantly greater than 0. However, this test seems to be less accurate than the Z test (Efron and Tibshirani 1993).

When the nucleotide sequences are short, $\hat{p}_S$ or $\hat{p}_N$ can be greater than 0.75 by chance or for some other reasons, and $\hat{d}_S$ or $\hat{d}_N$ may not be computable. In this case, the difference between synonymous and nonsynonymous substitutions should be tested by using $\hat{p}_S$ and $\hat{p}_N$ directly. Since the variances of $\hat{p}_S$ and $\hat{p}_N$ can be computed either by Equation (4.4) or by the bootstrap, the null hypothesis of $p_S = p_N$ can be tested by using

Table 4.1 Fisher's exact test for small samples.

	Substitution Sites	Nonsubstitution Sites	Total
Synonymous	S_d (1)	$S - S_d$ (40)	S (41)
Nonsynonymous	N_d (20)	$N - N_d$ (110)	N (130)
Sum	$S_d + N_d$ (21)	$T - S_d - N_d$ (150)	T (171)

Note: The numbers in parentheses refer to those used for testing adaptive evolution at a human MHC (*HLA-A*) locus. $T = S + N$.

the Z test. Actually, $\hat{p}_S$ and $\hat{p}_N$ are better than $\hat{d}_S$ and $\hat{d}_N$ in detecting positive selection, because they require fewer assumptions than the latter.

Small-Sample Test

When the number of nucleotide substitutions per sequence (S_d or N_d) is small, the above large-sample test tends to be too liberal and may be misleading (Zhang et al. 1997). In this case, it is usually possible to count the actual numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions without much error, because most codon differences are caused by one nucleotide substitution. We can then construct a 2×2 contingency table for synonymous and nonsynonymous substitutions and conduct Fisher's exact test as given in Table 4.1. In this table, T stands for the total number of nucleotides examined, i.e., $T = S + N$. An example of Fisher's exact test will be discussed later.

Tests of the Difference Between $\bar{d}_S$ and $\bar{d}_N$ or $\bar{p}_S$ and $\bar{p}_N$

In the study of adaptive evolution, it is often necessary to compare the average values ($\bar{d}_S$ and $\bar{d}_N$) of $\hat{d}_S$'s and $\hat{d}_N$'s for many sequence comparisons, because comparison of a pair of sequences is not always very informative. Hughes and Nei (1988, 1989) used this approach to show that $\bar{d}_N$ is significantly greater than $\bar{d}_S$ in the antigen recognition site of major histocompatibility complex (MHC) genes, and this demonstration led them to conclude that the extremely high degree of polymorphism at MHC loci is primarily due to overdominant selection operating at the antigen recognition site.

To establish $\bar{d}_N > \bar{d}_S$, however, it is necessary to conduct a statistical test of the difference $\hat{D} = \bar{d}_N - \bar{d}_S$. This test can be done by using the standard Z test with the variance $[V(\hat{D})]$ of $\hat{D}$ given by $V(\bar{d}_N) + V(\bar{d}_S) - 2Cov(\bar{d}_N, \bar{d}_S)$, where $V(\bar{d}_N)$, $V(\bar{d}_S)$, and $Cov(\bar{d}_N, \bar{d}_S)$ are the variances and covariance of $\bar{d}_N$ and $\bar{d}_S$. It is not a simple matter to compute $V(\bar{d}_N)$, $V(\bar{d}_S)$, and $Cov(\bar{d}_N, \bar{d}_S)$ analytically, because different sequences are related through evolutionary history. Nei and Jin (1989) developed a method for computing the variances and covariance of $\bar{d}_N$ and $\bar{d}_S$ taking into account the phylogenetic tree of the sequences. This method is useful when the number of sequences used is relatively small but becomes time consuming when the number is large (say, more than 20). Another method for testing the difference $\hat{D}$ is to use the bootstrap method.

Suppose that all sequences consist of C codons. We can then resample C codons with replacement from the original set of sequences and compute $\bar{d}_N$ and $\bar{d}_S$ for this set of samples. If we repeat this computation many times, we can compute the standard error of $\bar{d}_N - \bar{d}_S$ and use it for the Z test. If one is interested in the test of the mean difference $\bar{p}_N - \bar{p}_S$, the same method can be used.

It should be noted that the bootstrap test may lead to an erroneous conclusion when the numbers of synonymous and nonsynonymous substitutions observed are small. To explain this problem, let us consider an extreme case where 6 nonsynonymous ($n = 6$) and 0 synonymous substitutions ($s = 0$) are observed when 60 nonsynonymous sites ($N = 60$) and 30 synonymous sites ($S = 30$) are examined. In this case, Fisher's exact test mentioned above indicates that the null hypothesis $p_N = p_S$ cannot be rejected. However, if we use the bootstrap test, $\hat{p}_N$ would be greater than $\hat{p}_S$ in almost all replications. Therefore, we would conclude that the null hypothesis is rejected. This obviously incorrect conclusion was reached because the original values of s and n were biased by chance and this bias cannot be corrected by bootstrap resampling. It is therefore important to compute the standard error of $\hat{D}$ by analytical formulas when C is small.

Modified Nei-Gojobori Method

Nei and Gojobori's (1986) method assumes random nucleotide substitution among the four nucleotides in computing the number of synonymous and nonsynonymous sites. In practice, this assumption does not necessarily hold, and the rate of transitional change is usually higher than that of transversional change. In this case, the number (S) of potential sites that can produce synonymous substitutions is expected to be greater than the number estimated by Nei and Gojobori's method, because transitional changes at third positions are largely synonymous. Therefore, Nei and Gojobori's method is expected to give overestimates of p_S and d_S and underestimates of p_N and d_N.

To rectify this deficiency, Ina (1995) proposed a method for estimating d_S and d_N using Kimura's (1980) 2-parameter model. His method is quite elaborate, as will be explained later. However, the major problem of Nei and Gojobori's method is the underestimation of S and the overestimation of N. Therefore, if we use appropriate methods of estimating S and N, their approach can still be used (Zhang et al. 1998). In the following, we adapt Ina's method for this purpose.

In Kimura's (1980) model the rates of transitional and transversional changes are given by α and β, respectively (chapter 3), but since any nucleotide can have two different transversional changes, the proportion of transitions among the total changes is given by

$$\frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R} \tag{4.6}$$

where R is the transition/transversion ratio and becomes 0.5 when there is no bias. (Note that R is different from the transition/transversion rate

ratio $k = \alpha/\beta$, which is often used in theoretical papers.) Ina (1995) has shown that the expected number of synonymous changes per codon can be expressed in terms of $R = \alpha/(2\beta)$ for all codons. For example, for codon TTT the number is given by

$$s = 0 + 0 + \frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R} \tag{4.7}$$

because in this case only the third nucleotide position produces synonymous changes and only one (T → C) of the three possible changes is synonymous. For another example, codon CTA (Leu) has the expected number of $s = R/(1 + R) + 1$, because the first, second, and third nucleotide positions of this codon can produce synonymous substitutions with probabilities $R/(1 + R)$, 0, and 1, respectively. In these computations, nonsense mutations are disregarded as before.

It is therefore clear that if we know R, we can compute s for all codons and then estimate S and N $(= 3C - S)$. The problem is how to estimate R from actual data. We suggest that R be estimated by Equation (3.2) or (3.18) in chapter 3 or that the R value obtained from other information be used. Theoretically, when the pattern of nucleotide substitution is complicated, both Equations (3.2) and (3.18) may give underestimates of R (Yang 1995b), and this underestimation of R makes the test of positive selection conservative. However, it is better to use a conservative test for detecting positive selection, because the actual pattern of nucleotide substitution is usually quite complicated and this may inflate $\hat{d}_N$ relative to $\hat{d}_S$ spuriously.

If we use the above method, S is expected to increase, and N is expected to decrease compared with the values obtained by the original Nei-Gojobori method. Let us denote these new S and N by S_R and N_R, respectively. In contrast to S and N, the number of synonymous (S_d) and nonsynonymous (N_d) differences are not seriously affected by the transition/transversion bias, because S_d and N_d are based on the actual number of substitutions observed. Therefore, the proportions of synonymous ($\hat{p}_S$) and nonsynonymous ($\hat{p}_N$) differences are now given by

$$\hat{p}_S = S_d/S_R, \qquad \hat{p}_N = N_d/N_R \tag{4.8}$$

whereas the estimates ($\hat{d}_S$ and $\hat{d}_N$) of d_S and d_N are again approximately given by the Jukes-Cantor formula. Theoretically, there is a better way to estimate d_S and d_N as shown by Ina (1995), but in practice there is not much difference between the estimates obtained by the two methods unless d_S and d_N are very high. (When $d_S > 1.0$ and $d_N > 1.0$, the reliability of $\hat{d}_S$ and $\hat{d}_N$ is very low, because the actual process of synonymous and nonsynonymous substitution is very complicated.) Furthermore, the present methods give smaller variances of $\hat{p}_S$, $\hat{p}_N$, $\hat{d}_S$, and $\hat{d}_N$ than those obtained by Ina's method.

Although the modified Nei-Gojobori method is theoretically better than the original version when Kimura's model with a high R value applies, it should be noted that when the estimate of R is unreliable, it may lead to an erroneous conclusion. Particularly when an overestimate of R

is used, the modified version may conclude that $\hat{d}_N$ is significantly higher than $\hat{d}_S$, even if this is not actually the case. Note that the actual pattern of nucleotide substitution is much more complicated than the Kimura model, and under certain conditions the modified Nei-Gojobori method may give an overestimate of S and an underestimate of N. For this reason, it is always better to use both the original Nei-Gojobori method and the modified version to detect positive selection. If the original version indicates positive Darwinian selection, the conclusion would be safer.

Example 4.1. Positive Darwinian Selection at MHC Loci

Figure 4.1 shows the nucleotide sequences of three alleles from the A locus of the human MHC (HLA) class I α chain genes. The α chain gene encodes three extracellular domains ($\alpha 1$, $\alpha 2$, and $\alpha 3$), a transmembrane portion, and a cytoplasmic tail of the MHC molecule (Klein and Horejsi 1997), and the sequences in Figure 4.1 are for the $\alpha 1$, $\alpha 2$, and $\alpha 3$ extracellular domains. They consist of $C = 274$ codons or $3C = 822$ nucleotides. Comparison of alleles *A*2301* and *A*2501* shows that there are 41 nucleotide differences; 33 of them are from codons showing one nucleotide difference and eight are from codons showing two nucleotide differences. All the codon differences and the s and n values for each codon difference are presented in Table 4.2. From this table, we obtain $S_d = 11.5$ and $N_d = 29.5$.

Nei-Gojobori Method

The total number of synonymous sites (S) can be computed by the methods described above, and it becomes 198 and 195.8 for alleles *A*2301* and *A*2501*, respectively. Therefore, the average of S for the two sequences is 196.9, and the average N is $822 - 196.9 = 625.1$. We can then obtain $\hat{p}_S = 11.5/196.9 = 0.0584$ and $\hat{p}_N = 29.5/625.1 = 0.0472$ from Equation (4.2), and their standard errors become $s(\hat{p}_S) \equiv [V(\hat{p}_S)]^{1/2} = 0.0167 = s(\hat{p}_N) = 0.0085$ from Equation (4.4). Essentially the same standard errors (0.0160 and 0.0087, respectively) are obtained by the bootstrap method. The Z value equivalent to Equation (4.5) is -0.60, which indicates that the difference $\hat{p}_N - \hat{p}_S$ ($= -0.011$) is not statistically significant. (Here the two-tail test should be used.) If we use Equations (3.8) and (4.3), we obtain $\hat{d}_S = 0.0608 \pm 0.0181$ and $\hat{d}_N = 0.0487 \pm 0.0091$ (Table 4.3). The Z value for the difference $\hat{d}_N - \hat{d}_S$ then becomes -0.60, which again indicates that the difference is not significant. Therefore, the Z tests for $\hat{d}_N - \hat{d}_S$ and $\hat{p}_N - \hat{p}_S$ give the same conclusion.

Modified Nei-Gojobori Method

In this method, we first have to estimate the R value. If we use Equation (3.18), we have $R = 0.79$ for alleles *A*2301* and *A*2501*, $R = 0.92$ for alleles *A*2301* and *A*3301*, and $R = 0.82$ for alleles *A*2501* and *A*3301*. Therefore, the average R is approximately 0.85. Using this R value, we have S_R equal to 211.9 and 209.8 for alleles *A*2301* and *A*2501*, respectively, with an average of 210.8. This gives $N_R = 611.2$. Using these

```
          α1
A*2301 GGC TCC CAC TCC ATG AGG TAT TTC TCC ACA TCC GTG TCC CGG CCC GGC CGC GGG GAG CCC  20
A*2501 ... ... ... ... ... ... ... ... .A. ..C ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... ... ... ... A.. ... ... ... ... ... ... ... ... ... ... ...

A*2301 CGC TTC ATC GCC GTG GGC TAC GTG GAC GAC ACG CAG TTC GTG CGG TTC GAC AGC GAC GCC  40
A*2501 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

A*2301 GCG AGC CAG AGG ATG GAG CCG CGG GCG CCG TGG ATA GAG CAG GAG GGG CCG GAG TAT TGG  60
A*2501 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

A*2301 GAC GAG GAG ACA GGG AAA GTG AAG GCC CAC TCA CAG ACT GAC CGA GAG AAC CTG CGG ATC  80
A*2501 ... CG. A.C ... C.. ..T ... ... ... ... ... ... ... ... ... ... .G. ... ... ...
A*3301 ... CG. A.C ... C.. ..T ... ... ... ... ... ... .T. ... ... .T. G.. ... G.. .C.
                                                  α2
A*2301 GCG CTC CGC TAC TAC AAC CAG AGC GAG GCC|GGT TCT CAC ACC CTC CAG ATG ATG TTT GGC 100
A*2501 ... ... ... ... ... ... ... ... ... .A.|... ... ... ... A.. ... .G. ... .A. ...
A*3301 CT. .G. G.. ... ... ... ... ... ... ...|... ... ... ... A.. ... ... ... .A. ...

A*2301 TGC GAC GTG GGG TCG GAC GGG CGC TTC CTC CGC GGG TAC CAC CAG TAC GCC TAC GAC GGC 120
A*2501 ... ... ... ... C.. ... ... ... ... ... ... ... ... ... ..G ... G.. ..T ... ...
A*3301 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..G ... G.. ... ... ...

A*2301 AAG GAT TAC ATC GCC CTG AAA GAG GAC CTG CGC TCT TGG ACC GCG GCG GAC ATG GCG GCT 140
A*2501 ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... T.. ..C ... ... ... ... ... ... ... ... ... ... ... ... ...

A*2301 CAG ATC ACC CAG CGC AAG TGG GAG GCG GCC CGT GTG GCG GAG CAG TTG AGA GCC TAC CTG 160
A*2501 ... ... ... ... ... ... ... ... A.. ... .A. .A. ... ... ... .G. ... ... ... ...
A*3301 ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

A*2301 GAG GGC ACG TGC GTG GAC GGG CTC CGC AGA TAC CTG GAG AAC GGG AAG GAG ACG CTG CAG 180
A*2501 ... ... CG. ... ... ..G T.. ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... ..G T.. ... ... ... C.. ... ... ... ... ... ... ... ... ...
          α3
A*2301 CGC ACG|GAC CCC CCC AAG ACA CAT ATG ACC CAC CAC CCC ATC TCT GAC CAT GAG GCC ACT 200
A*2501 ... ...|... G.. ... ... ..G ... ... ..T ... ... G.T G.. ... ... ... ... ... ..C
A*3301 ... ...|... ... ... .G. ..G ... ... ..T ... ... G.T G.. ... ... ... ... ... ..C

A*2301 CTG AGA TGC TGG GCC CTG GGC TTC TAC CCT GCG GAG ATC ACA CTG ACC TGG CAG CGG GAT 220
A*2501 ... ..G ... ... ... ... A.. ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ..G ... ... ... ... A.. ... ... ... ... ... ... ... ... ... ... ... ... ...

A*2301 GGG GAG GAC CAG ACC CAG GAC ACG GAG CTT GTG GAG ACC AGG CCT GCA GGG GAT GGA ACC 240
A*2501 ... ... ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ..G ...
A*3301 ... ... ... ... ... ... ... ... ..C ... ... ... ... ... ... ... ... ... ... ...

A*2301 TTC CAG AAG TGG GCA GCT GTG GTG GTA CCT TCT GGA GAG GAG CAG AGA TAC ACC TGC CAT 260
A*2501 ... ... ... ... ..G T.. ... ... ..G ... ... ... C.. ... ... ... ... ... ... ...
A*3301 ... ... ... ... ..G T.. ... ... ..G ... ... ... C.. ... ... ... ... ... ... ...

A*2301 GTG CAG CAT GAG GGT CTG CCC AAG CCC CTC ACC CTG AGA TGG 274
A*2501 ... ... ... ... ... ... ... ... ... ... ... ... ...
A*3301 ... ... ... ... ... ..C ... ... ... ... ... ... ...
```

FIGURE 4.1. Nucleotide sequences of three human class I *HLA-A* alleles for the three extracellular domains α_1, α_2, and α_3. A dot (.) shows identity with the first sequence. Exons boundaries are marked with vertical lines. The nucleotides at the antigen recognition site (ARS) are in boldface.

values, we obtain $\hat{d}_S = 0.0566 \pm 0.0169$ and $\hat{d}_N = 0.0499 \pm 0.0093$. Therefore, $\hat{d}_S$ has decreased and $\hat{d}_N$ has increased slightly.

Adaptive Evolution

X-ray diffraction studies have shown that class I MHC molecules form a groove in which a foreign peptide is bound (Bjorkman et al. 1987a,

Table 4.2 Codons that are different between the HLA *A*2301* and *A*2501* alleles.

Codon		s_d	n_d	$s_d + n_d$	Codon		s_d	n_d	$s_d + n_d$
*9	TCC-TAC		1	1	*156	TTG-TGG		1	1
10	ACA-ACC	1		1	*163	ACG-CGG	0.5	1.5	2
*62	GAG-CGG		2	2	*166	GAC-GAG		1	1
*63	GAG-AAC		2	2	*167	GGG-TGG		1	1
*65	GGG-CGG		1	1	184	CCC-GCC		1	1
*66	AAA-AAT		1	1	187	ACA-ACG	1		1
*77	AAC-AGC		1	1	190	ACC-ACT	1		1
90	GCC-GAC		1	1	193	CCC-GCT	1	1	2
*95	CTC-ATC		1	1	194	ATC-GTC		1	1
*97	ATG-AGG		1	1	200	ACT-ACC	1		1
*99	TTT-TAT		1	1	202	AGA-AGG	1		1
105	TCG-CCG		1	1	207	GGC-AGC		1	1
*114	CAC-CAG		1	1	230	CTT-CTC	1		1
*116	TAC-GAC		1	1	239	GGA-GGG	1		1
117	GCC-GCT	1		1	245	GCA-GCG	1		1
127	AAA-AAC		1	1	246	GCT-TCT		1	1
*149	GCG-ACG		1	1	249	GTA-GTG	1		1
*151	CGT-CAT		1	1	253	GAG-CAG		1	1
*152	GTG-GAG		1	1		Total	11.5	29.5	41

Note: Antigen recognition sites are indicated with an asterisk.

1987b). This groove is called the antigen recognition site (ARS) and consists of 57 amino acid sites (boldfaced letters in Figure 4.1). If we apply the Nei-Gojobori method to the 57 amino acid sites of the ARS for alleles *A*2301* and *A*2501*, we obtain $S_d = 0.5$, $N_d = 20.5$, $S = 40.5$, and N

Table 4.3 Numbers of synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions between the HLA *A*2301* and *A*2501* alleles for the extracellular region and the antigen recognition sites (ARS).

	Extracellular Region ($C = 274$)		ARS ($C = 57$)	
Method	$\hat{d}_S$	$\hat{d}_N$	$\hat{d}_S$	$\hat{d}_N$
$R = 0.5$				
NG[a]	6.08 ± 1.81	4.87 ± 0.91	1.24 ± 1.76	17.63 ± 4.03
LWL[b]	6.52 ± 2.02	4.82 ± 0.89	0.03 ± 1.87	17.25 ± 3.99
$R = 0.85$[c]				
Modified-NG	5.66 ± 1.69	4.99 ± 0.93	1.17 ± 1.66	18.06 ± 4.14
PBL[d]	4.59 ± 1.46	4.80 ± 0.90	0.02 ± 1.14	17.04 ± 3.96
Kumar	4.55 ± 1.46	4.74 ± 0.91	0.36 ± 1.19	16.79 ± 4.03
Ina II	4.87 ± 1.47	5.31 ± 0.99	1.50 ± 2.13	16.67 ± 3.81
GY[e]	12.17	4.25	0.02	16.98

Note: $\hat{d}_S$ and $\hat{d}_N$ are multiplied by 100.

[a]NG: Nei-Gojobori.
[b]LWL: Li-Wu-Luo.
[c]$R = 0.85$ was used only for the Modified-NG method. In the other methods, R was computed automatically.
[d]PBL: Pamilo-Bianchi-Li.
[e]GY: Goldman-Yang.

= 130.5. We therefore have $\hat{d}_S = 0.0124 \pm 0.0176$ and $\hat{d}_N = 0.1763 \pm 0.0403$ (Table 4.3). A Z test shows that $Z = 3.7$ and $\hat{d}_N$ is significantly greater than $\hat{d}_S$ at the 0.1% level when a one-tail test is used. In contrast, the modified Nei-Gojobori method gives $S_R = 43.11$ and $N_R = 127.89$ (with $R = 0.85$), so that we have $\hat{d}_S = 0.0117 \pm 0.0166$ and $\hat{d}_N = 0.1806 \pm 0.0414$. The Z test again shows that $\hat{d}_N$ is significantly greater than $\hat{d}_S$ at the 0.1% level. Therefore, both methods show that $\hat{d}_N$ is greater than $\hat{d}_S$, and this strongly suggests that the ARS of class I MHC molecules is the target of positive Darwinian selection.

Small-Sample Tests

In the above tests, we used a large-sample test, which is not really valid because S_d was only 0.5. Let us now use Fisher's exact test. If we use the conservative Nei-Gojobori method, we obtain $S = 41$ and $N = 130$ approximately. We also assume $S_d = 1$ and $N_d = 20$ to make the test even more conservative. We then have the 2×2 contingency table given in parentheses in Table 4.1. Fisher's exact test gives a P value of 0.018. This indicates that $\hat{d}_N$ is significantly greater than $\hat{d}_S$. If we use the modified Nei-Gojobori method, we obtain $S_R = 44$ and $N_R = 127$ (with $R = 0.85$), and Fisher's test gives a P value of 0.012. Therefore, the P values for the small-sample test are higher than those for the large-sample test.

Tests of $\bar{d}_N - \bar{d}_S$ or $\bar{p}_N - \bar{p}_S$

Since the power of detecting positive selection is low in this case because of the small number of codons involved, let us consider the averages of $\hat{d}_N$ and $\hat{d}_S$. In the present case, there are three sequences, so $\hat{d}_N$ and $\hat{d}_S$ can be computed for three pairs of alleles. We can then obtain the averages ($\bar{d}_N$ and $\bar{d}_S$) of these values. If we use the Nei-Gojobori method, they become $\bar{d}_N = (0.1763 + 0.1822 + 0.1479)/3 = 0.1688$ and $\bar{d}_S = (0.0124 + 0.0000 + 0.0124)/3 = 0.0083$. Therefore, the difference $\hat{D} = \bar{d}_N - \bar{d}_S$ is 0.1605. If we use the bootstrap method, the standard error of $\hat{D} = \bar{d}_N - \bar{d}_S$ becomes 0.0322. (This was computed by a program in MEGA2 using 1000 bootstrap replications.) Therefore, we have $Z = 4.98$. If we use Nei and Jin's method, we obtain $Z = 4.80$, which is again highly significant. These results reinforce the conclusion reached by comparison of two sequences.

4.2. Methods Based on Kimura's 2-Parameter Model

Li-Wu-Luo Method

Li et al. (1985) developed another method, based on Kimura's 2-parameter model. They first noted that when the degeneracy of the genetic code is considered, the nucleotide sites of codons can be classified into 4-fold degenerate, 2-fold degenerate, and 0-fold degenerate (nondegenerate) sites with a few exceptions (e.g., isoleucine codons). A site is called 4-

fold degenerate if all possible changes at the site are synonymous, 2-fold degenerate if one of the three possible changes is synonymous, and 0-fold degenerate if all changes are nonsynonymous or nonsense mutations. For example, the third nucleotide positions of the valine codons are 4-fold degenerate sites, and the second positions of all codons are 0-fold degenerate sites. The third positions of the three isoleucine codons are actually 3-fold degenerate sites, but they are regarded as 2-fold degenerate sites to simplify the computation.

Using the above rule, we can compute the numbers of three types of sites for each of the two sequences and denote by L_0, L_2, and L_4 the average numbers of 0-fold, 2-fold, and 4-fold degenerate sites for the two sequences compared, respectively. We then compare the two sequences, codon by codon, and classify each nucleotide difference as either a transition or a transversion. We denote by P_i and Q_i the proportions of transitional and transversional nucleotide differences at the i-th class of nucleotide sites (i = 0, 2, or 4). (Actually, they considered all possible evolutionary pathways between each pair of codons as in the case of the Nei-Gojobori method and computed P_i and Q_i taking into account the likelihood of occurrence of each amino acid substitution. See Li et al. [1985] for the detail.) We can then estimate the numbers of transitional (A_i) and transversional (B_i) substitutions per site for each of the three classes of nucleotide sites. That is,

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i) \tag{4.9a}$$

$$B_i = \frac{1}{2}\ln(b_i) \tag{4.9b}$$

where $a_i = 1/(1 - 2P_i - Q_i)$ and $b_i = 1/(1 - 2Q_i)$.

We note that all substitutions at 4-fold sites are synonymous and all substitutions at 0-fold sites are nonsynonymous. At 2-fold sites, transitional changes (A_2) are mostly synonymous, whereas transversional changes are mostly nonsynonymous. Assuming that nucleotide substitution occurs with equal frequency among the four nucleotides A, T, C, and G, Li et al. (1985) suggested that one third of 2-fold degenerate sites are potentially synonymous sites and two thirds are potentially nonsynonymous sites. With this assumption, they proposed that d_S and d_N be estimated by the following formulas.

$$\hat{d}_S = \frac{3[L_2A_2 + L_4(A_4 + B_4)]}{L_2 + 3L_4} \tag{4.10a}$$

$$\hat{d}_N = \frac{3[L_0(A_0 + B_0) + L_2B_2]}{3L_0 + 2L_2} \tag{4.10b}$$

These formulas depend on a number of assumptions, which are not always satisfied with actual data. First, the type of a given nucleotide site in one sequence may not be the same as that of the homologous site in the other sequence. For example, the type of a given position in one se-

quence may be 2-fold degenerate, but the type of the same position in the other sequence could be 4-fold degenerate. This can happen quite often when sequence divergence is high. In this case, one half of the site is regarded as a 2-fold degenerate site, and the other half as a 4-fold degenerate site. Second, nonsense mutations are counted as nonsynonymous changes. For example, a nucleotide substitution at the third position of tyrosine codon TAT may produce one synonymous codon (TAC) and two nonsense codons (TAA and TAG), but the latter two changes are regarded as nonsynonymous changes. Since nonsense mutations occur with a probability of nearly 4% (chapter 1), this method is expected to give overestimates of d_N. Third, the transitions at the first nucleotide position of four 2-fold degenerate arginine codons (CGA, CGG, AGA, and AGG) are not synonymous but all nonsynonymous with one exception (CGA) that results in a nonsense codon. At the third position of the three isoleucine codons that are 3-fold degenerate, some transversions are synonymous. Despite these problems, Li et al.'s (1985) method seems to give results similar to those obtained by the Nei-Gojobori method when the number of codons is large and sequence divergence is low. When the number of codons used is small (say <100), however, Li et al.'s method may give negative estimates, because a_i and b_i in Equation (4.9) are subject to large sampling errors.

Pamilo-Bianchi-Li Method

Another problem in Li et al.'s (1985) method is the effect of transition/transversion bias, and the error introduced by this bias may be substantial when R is high, as in the case of Nei and Gojobori's method. For this reason, Pamilo and Bianchi (1993) and Li (1993) independently extended Li et al.'s method.

Noting that synonymous transitional changes occur only at 2-fold and 4-fold sites in Li et al.'s model, they proposed that the total number of these changes be estimated by the weighted mean $(L_2A_2 + L_4A_4)/(L_2 + L_4)$. Since the transversions at 4-fold sites are also synonymous, the total number of synonymous substitutions per synonymous site is now estimated by

$$\hat{d}_S = (L_2A_2 + L_4A_4)/(L_2 + L_4) + B_4 \qquad (4.11a)$$

Using the same argument, they also suggested that d_N be estimated by

$$\hat{d}_N = A_0 + (L_0B_0 + L_2B_2)/(L_0 + L_2) \qquad (4.11b)$$

Comeron and Kumar Methods

As mentioned earlier, the treatment of arginine and isoleucine codons in Li et al.'s (1985) method is inaccurate. This is also true with the Pamilo-Bianchi-Li method. This creates a problem when these amino acids are abundant. (In mammalian protamine P1, about 50% of amino acids are arginines; Rooney et al. 2000 n.d.). Comeron (1995) attempted to solve this problem by dividing 2-fold degenerate sites into two groups: 2S-fold

and 2V-fold degenerate sites. The former refer to sites where the two transitional changes are synonymous and the transversional change is nonsynonymous, whereas the latter represent sites where the transitional change is nonsynonymous and the two transversional changes are synonymous. This subdivision of 2-fold degenerate sites certainly help to correct some of Li et al.'s inaccurate classifications of synonymous and nonsynonymous sites (e.g., methionine codons).

However, this does not solve all the problems. For example, a mutation at the first nucleotide position of arginine codon CGG produces TGG (Trp), AGG (Arg), or GGG (Gly). In this case, the transitional change (C → T) results in a nonsynonymous substitution, whereas one transversional change (C → A) results in a synonymous substitution and the other transversional change (C → G) a nonsynonymous substitution. Therefore, this nucleotide site is neither a 2S-fold nor a 2V-fold site. Similarly, the first position of three arginine codons (CGU, CGC, and CGA) and the third position of two isoleucine codons (ATT and ATC) cannot be assigned to any of Comeron's categories.

To take care of these problems, S. Kumar (n.d.) developed another version of the Pamilo-Bianchi-Li method. In this version, nucleotide sites are first classified into 0-fold, 2-fold, and 4-fold degenerate sites, and the 2-fold degenerate sites are further subdivided into simple 2-fold and complex 2-fold degenerate sites. Simple 2-fold degenerated sites are those at which the transitional change results in a synonymous substitution and the two transversional changes generate nonsynonymous substitutions or nonsense mutations. All other 2-fold degenerate sites, including those for the three isoleucine codons, belong to the complex 2-fold sites. Using this classification of sites, Kumar developed a new method for estimating d_S and d_N. This method is included in MEGA2.

Ina's Method

Ina (1995) developed yet another method for estimating d_S and d_N, combining some features of the original Nei-Gojobori and the Pamilo-Bianchi-Li methods. He proposed two methods: method I and method II. In method I, the transition/transversion rate ratio $k = \alpha/\beta$ is estimated by Equation (3.18) in chapter 3 using only data at the third codon positions. This depends on the assumption that the nucleotide substitution at the third positions is largely neutral. S and N are then estimated by using the procedure of the modified Nei-Gojobori method, whereas S_d and N_d are computed by the Nei-Gojobori method. However, Ina divides S_d into synonymous transitional differences (S_{Ts}) and synonymous transversional differences (S_{Tv}) and N_d into nonsynonymous transitional differences (N_{Ts}) and nonsynonymous transversional differences (N_{Tv}). He then estimates $\hat{d}_S$ and $\hat{d}_N$ using formulas analogous to Equation (3.12). In his method II, S and N are estimated from data for all three codon positions, but α and β are estimated by using only synonymous substitutions to reflect the mutation rates before selection. The actual procedure is quite elaborate.

Ina's computer simulation has shown that method II gives slightly more accurate estimates of d_S and d_N than method I when the number of

nucleotides used is large. However, the differences in $\hat{d}_S$ and $\hat{d}_N$ between the two methods or between Ina's methods and the modified Nei-Gojobori method are usually small. Furthermore, when the number of nucleotides used is small and sequence divergence is low, Ina's method I may not be applicable, because no transitions or no transversions may be observed and this makes the estimate of α/β either 0 or ∞. Therefore, some caution is necessary when Ina's methods are to be used.

Example 4.2. Further Analysis of MHC Gene Sequences

In Example 4.1, we computed $\hat{p}_S$, $\hat{p}_N$, $\hat{d}_S$, and $\hat{d}_N$ for human MHC alleles by using the Nei-Gojobori and the modified Nei-Gojobori methods. Let us now compute $\hat{d}_S$ and $\hat{d}_N$ for the alleles *A*2301* and *A*2501* using the methods based on Kimura's model. We will not consider $\hat{p}_S$ and $\hat{p}_N$, since these are not computable in these methods. In the case of the Li-Wu-Luo method, comparison of the two alleles gives $L_0 = 535$, $L_2 = 154.5$, and $L_4 = 132.5$. We also obtain $A_0 = 0.01542$, $B_0 = 0.02985$, $A_2 = 0.00806$, $B_2 = 0.42183$, $A_4 = 0.07359$, and $B_4 = 0.00761$. Therefore, Equations (4.10a) and (4.10b) give $\hat{d}_S = 0.0652 \pm 0.0202$ and $\hat{d}_N = 0.0482 \pm 0.0089$. These values are nearly the same as those obtained by the Nei-Gojobori method (Table 4.3). The Pamilo-Bianchi-Li method gives $\hat{d}_S = 0.0459 \pm 0.0146$ and $\hat{d}_N = 0.0480 \pm 0.0090$. The $\hat{d}_S$ and $\hat{d}_N$ values obtained by the Kumar method and the Ina method II are presented in Table 4.3. The values by the Kumar method are similar to those obtained by the Pamilo-Bianchi-Li method, whereas those by the Ina method II are similar to those obtained by the modified Nei-Gojobori method. Curiously, the $\hat{d}_N$ values obtained by the Pamilo-Bianchi-Li and the Kumar methods are similar to the $\hat{d}_N$ obtained by the Nei-Gojobori method, though they are supposed to be higher than the latter because the transition/transversion bias is taken care of. This is probably caused by the fact that the pattern of nucleotide substitution in MHC genes is much more complicated than Kimura's model (Hughes and Nei 1988).

Let us now compute $\hat{d}_S$ and $\hat{d}_N$ for the ARS using the Li-Wu-Luo, Pamilo-Bianchi-Li, and Kumar methods. The results obtained are presented in Table 4.3 together with the previous results. The $\hat{d}_N$ value obtained by the Li-Wu-Luo method is similar to that obtained by the Nei-Gojobori method, but the $\hat{d}_S$ is much smaller than that obtained by the latter method. This small $\hat{d}_S$ is unreasonable, because the Nei-Gojobori method is based on a parsimonious counting of synonymous substitutions, and therefore it should give a minimum estimate. The unduly low $\hat{d}_S$ value probably occurred because the Kimura model is unlikely to apply to the ARS, where the pattern of nucleotide substitution is complicated and the number of codons involved is small. An unduly small $\hat{d}_S$ value is also obtained by the Pamilo-Bianchi-Li and the Kumar methods, which are also based on the Kimura model. The $\hat{d}_N$ value obtained by the Li-Wu-Luo method is similar to that obtained by the Nei-Gojobori method, but the values obtained by the Pamilo-Bianchi-Li and the Kumar methods are smaller than the $\hat{d}_N$ obtained by the modified Nei-Gojobori method and are virtually the same as the $\hat{d}_N$ obtained by the Li-Wu-Luo method, although they are supposed to be larger. They are even smaller

Table 4.4 Synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions for the mitochondrial *Nd5* gene sequences from humans and chimpanzees.

Method	$\hat{d}_S$	$\hat{d}_N$
$R = 0.5$		
Nei-Gojobori	41.51 ± 3.80	3.79 ± 0.54
Li-Wu-Luo	42.77 ± 4.14	3.78 ± 0.54
$R = 9.21^a$		
Modified Nei-Gojobori	27.30 ± 2.40	4.38 ± 0.62
Ina II	30.31 ± 3.07	4.38 ± 0.63
Pamilo-Bianchi-Li	30.18 ± 2.87	4.38 ± 0.63
Comeron-Kumar	30.18 ± 2.87	4.38 ± 0.63
Goldman-Yang	28.72	4.42

Note: $\hat{d}_S$ and $\hat{d}_N$ are multiplied by 100.

[a] $R = 9.21$ was used only for the Modified-NG method. In the other methods, R is computed automatically. The number of codons used (C) is 603.

than the $\hat{d}_N$ obtained by the parsimonious Nei-Gojobori method. This again suggests that the Kimura model is inappropriate for MHC genes, particularly for the ARS. Table 4.3 also includes the $\hat{d}_S$ and $\hat{d}_N$ obtained by Ina's method II. These values are more similar to those obtained by the Nei-Gojobori method than those obtained by the modified Nei-Gojobori method. This also suggests that the Kimura model is inappropriate for the ARS.

Example 4.3. $\hat{d}_S$ and $\hat{d}_N$ Values for the Mitochondrial *Nd5* Gene

In the above example, the transition/transversion bias ($R = 0.85$) was small, so that the differences in $\hat{d}_S$ and $\hat{d}_N$ between the two assumptions of $R = 0.5$ and $R = 0.85$ were also small. In many nuclear genes, R is 0.5 ~ 2, and the effect of the transition/transversion bias is usually very small. In mitochondrial genes, however, the effect is expected to be large because R is generally high. Let us now consider the mitochondrial NADH dehydrogenase 5 (*Nd5*) gene sequences from humans and chimpanzees (Horai et al. 1995). The total number of codons in this gene is 603, and Equation (3.18) gives an R value of 9.21. The $\hat{d}_S$ and $\hat{d}_N$ values were obtained by the seven different methods discussed above, and they are presented in Table 4.4. In this case, the assumption of $R = 0.5$ certainly gives overestimates of d_S and underestimates of d_N. However, different methods that allow a high R value give very similar estimates of d_S and d_N.

4.3. Nucleotide Substitutions at Different Codon Positions

When relatively closely related species are compared, the number of synonymous substitutions is expected to increase almost linearly with time,

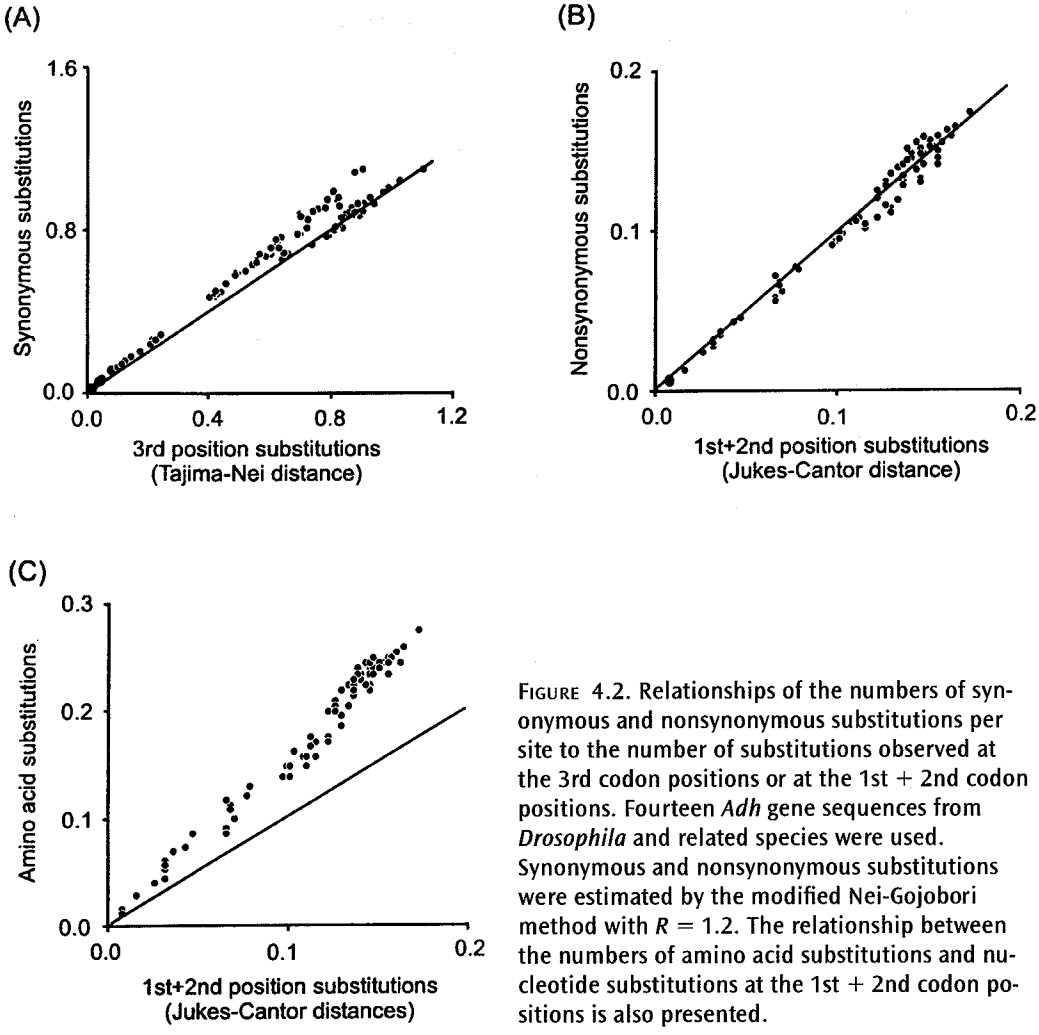


FIGURE 4.2. Relationships of the numbers of synonymous and nonsynonymous substitutions per site to the number of substitutions observed at the 3rd codon positions or at the 1st + 2nd codon positions. Fourteen *Adh* gene sequences from *Drosophila* and related species were used. Synonymous and nonsynonymous substitutions were estimated by the modified Nei-Gojobori method with $R = 1.2$. The relationship between the numbers of amino acid substitutions and nucleotide substitutions at the 1st + 2nd codon positions is also presented.

because they are generally free from selection. However, as the number of substitutions increases, the accuracy of the estimates is expected to decline, because the assumptions used to estimate the number of synonymous substitutions are unlikely to hold for a long time. As mentioned earlier, synonymous and nonsynonymous sites are not fixed but vary with time. For this reason, some authors prefer to use the number of nucleotide substitutions at third codon positions for estimating evolutionary times. At these sites, a certain proportion of nucleotide substitutions are nonsynonymous, but the nucleotide sites are clearly defined and do not change with time. Therefore, the number of substitutions at third positions may be linearly related with evolutionary time.

In practice, the number of synonymous substitutions for a gene is generally greater than the number of third-position substitutions. Figure 4.2A shows the relationship between the number of synonymous substitutions ($\hat{d}_S$) obtained by the modified Nei-Gojobori method and the number of third-position substitutions ($\hat{d}_3$) for the alcohol dehydrogenase

(*Adh*) gene sequences from 14 different *Drosophila* species (see the book's website: *http://www.oup-usa.org/sc/0195135857*). The $\hat{d}_3$ values were obtained by Tajima and Nei's method, because the nucleotide frequencies at third codon positions are substantially different from 0.25. The results show that $\hat{d}_S$ is generally slightly higher than $\hat{d}_3$ as expected, but there is an approximately linear relationship between $\hat{d}_S$ and $\hat{d}_3$ for $\hat{d}_S < 0.8$. This result suggests that either $\hat{d}_S$ or $\hat{d}_3$ can be used for estimating divergence times as long as $d_S < 0.8$ in the present case. In fact, Thomas and Hunt (1993) used $\hat{d}_S$ for estimating the times of divergences of various *Drosophila* species whereas Russo et al. (1995) used $\hat{d}_3$, but their results were virtually the same.

Figure 4.2B shows the relationship between the number of nonsynonymous substitutions ($\hat{d}_N$) and the Jukes-Cantor distance for first and second codon positions ($\hat{d}_{12}$) for the same data set of *Adh* gene sequences. Here the $\hat{d}_N$ and $\hat{d}_{12}$ values are much smaller than the $\hat{d}_S$ and $\hat{d}_3$ values, but $\hat{d}_N$ and $\hat{d}_{12}$ are nearly equal to each other for all sequence comparisons. This indicates that for estimating divergence times either $\hat{d}_N$ or $\hat{d}_{12}$ can be used. Previously, we mentioned that the number of amino acid substitutions often gives a good estimate of divergence time. Figure 4.2C shows the relationship between the Poisson correction distance ($\hat{d}$) for amino acid sequence data and $\hat{d}_{12}$. Here again we can see a good linear relationship, although $\hat{d}$ is greater than $\hat{d}_{12}$ as expected.

4.4. Likelihood Methods with Codon Substitution Models

Goldman and Yang (1994) developed a likelihood method for estimating the rates of synonymous and nonsynonymous nucleotide substitution considering a nucleotide substitution model for 61 sense codons. (Three nonsense codons were eliminated.) Their model is somewhat similar to the Hasegawa-Kishino-Yano model (Table 3.2E) for nucleotide substitution. Let us consider a pair of sequences of C homologous codons and let π_j be the relative frequency of the j-th codon. They assumed that the instantaneous substitution rate (q_{ij}) from codon i to codon j ($i \neq j$) is given by the following equations.

$$q_{ij} = \begin{cases} 0, & \text{if nucleotide change occurs at two or more positions} \\ \pi_j, & \text{for synonymous transversion} \\ k\pi_j, & \text{for synonymous transition} \\ \omega\pi_j, & \text{for nonsynonymous transversion} \\ \omega k\pi_j, & \text{for nonsynonymous transition} \end{cases} \quad (4.12)$$

where k is the transition/transversion rate ratio and ω is the nonsynonymous/synonymous rate ratio. Here k may be written as α/β if the rates of transitional and transversional changes are α and β, respectively. Similarly, ω may be written as r_N/r_S if the rates of synonymous and

nonsynonymous changes are r_S and r_N, respectively. Therefore, if ω is the same for all codon pairs as assumed, it is possible to relate r_N/r_S to d_N/d_S.

There are 61 parameters for π_j, but if we assume that the codon frequencies are in equilibrium, they can be estimated by the observed codon frequencies when the number of codons used (C) is large. Therefore, the only parameters to be estimated are k and ω, and these parameters can be estimated by using the maximum likelihood method (Goldman and Yang 1994). When C is relatively small, however, this approach does not give a reliable estimate of π_j, because π_j is generally very small and thus the sampling error of the estimate of π_j is large. In this case, one may estimate π_j by a product of the observed nucleotide frequencies. In the present approach $\omega < 1$, $\omega = 1$, and $\omega > 1$ represent purifying selection, neutral evolution, and positive selection, respectively. Therefore, if the estimate ($\hat{\omega}$) of ω obtained from the data is significantly greater than 1, positive selection is suggested. Theoretically, this test can be done by using the likelihood ratio test.

Let $\ln L_2$ be the log *maximum likelihood* (ML) value when ω is estimated from the data and $\ln L_1$ be the ML value when $\omega = 1$ (null hypothesis) is assumed. The log likelihood ratio is then given by

$$LR = 2(\ln L_2 - \ln L_1) \tag{4.13}$$

When the numbers of synonymous and nonsynonymous substitutions are sufficiently large and the model used is appropriate, LR is approximately χ^2 distributed with one degree of freedom. Therefore, if $\hat{\omega} > 1$ and $LR \geq 3.84$, one may conclude that the rate of nonsynonymous substitution is significantly higher than that of synonymous substitution at the 5% level and that this is due to positive selection.

One advantage of this approach is that both the transition/transversion rate ratio (k) and the nonsynonymous/synonymous rate ratio (ω) can be estimated simultaneously if the model given in Equation (4.12) is satisfied. Therefore, there is no need to know R ($= 2k$) to estimate d_S and d_N as in the case of the modified Nei-Gojobori method.

However, there seem to be several problems with this approach. First, estimates of π_j's based on the observed frequencies would not be reliable when C is small as mentioned above. Estimation of π_j's by products of nucleotide frequencies would also be unreliable when the codon usage bias exists. Second, the assumption that ω is the same for all codon positions is quite unrealistic as is clear from the pattern of amino acid substitution discussed in chapter 1. This would make $\hat{\omega}$ substantially different from $\hat{d}_N/\hat{d}_S$, because the average of the ratio r_N/r_S is not the same as the ratio of the averages of r_N and r_S. Third, the assumption of independence of k and ω for every codon pair also would not be satisfied in actual data. Therefore, a more careful study is necessary about the effect of violation of the assumption on $\hat{\omega}$.

We have used this method (the default option of the computer program PAML by Yang [1999]) to compute the $\hat{d}_S$ and $\hat{d}_N$ values for the MHC and mitochondrial *Nd5* genes discussed earlier. The results are presented in Tables 4.3 and 4.4. In the extracellular region of the MHC gene, $\hat{d}_N$ is sim-

ilar to the $\hat{d}_N$ values obtained by the other methods, but $\hat{d}_S$ is more than two times higher than the values obtained by the other methods. In the case of the ARS of the human MHC *A* locus, $\hat{d}_N$ was about 1000 times higher than $\hat{d}_S$. In mitochondrial gene *Nd5* the $\hat{d}_N$ and $\hat{d}_S$ are similar to those of the modified Nei-Gojobori method.

Muse (1996) developed a similar likelihood method based on a different codon substitution model. In this method, codon frequencies are estimated by products of nucleotide frequencies, and no transition/transversion bias is assumed. Therefore, the number of parameters to be estimated is less than in the Goldman-Yang model. This method seems to give $\hat{d}_S$'s and $\hat{d}_N$'s similar to those of the Nei-Gojobori method when codon usage bias is small. When this bias and the transition/transversion bias are high, however, Muse's method is expected to give biased estimates.

As the computer technology develops, it is possible to use increasingly complicated mathematical models and conduct statistical analyses based on these models (e.g., Nielsen and Yang 1998). However, as the mathematical model becomes sophisticated, more parameters are required, and the underlying assumptions are likely to be violated quite often. A sophisticated model therefore may give biased estimates of the parameters. In contrast, the evolutionary pathway methods discussed earlier are based on the concept of parsimony analysis and are largely model free. Adaptive amino acid substitutions usually occur at some specific sites for functional reasons, and the pattern of the substitutions are likely to be different from the general pattern of amino acid substitution. Particularly, when d_S and d_N are large (say d_S, $d_N > 0.4$), these methods appear to give less reliable estimates than the simple evolutionary pathway methods, because there are many disturbing factors that affect the estimates of d_S and d_N (Tanaka and Nei 1989; Nei and Hughes 1992).

Another problem with the likelihood approach is the reliability of the likelihood ratio test. This test requires that the assumptions of the mathematical model used are satisfied with real data (Foutz and Srivastava 1977). Zhang (1999) has shown that in the test of evolutionary hypotheses this requirement is often violated and that in this case the test can be either too liberal or too conservative depending on the situation. Note also that the likelihood ratio test is a large-sample test, so that it may give erroneous conclusions when the numbers of synonymous and nonsynonymous substitutions are small. Therefore, caution is necessary in the application of this test.

5

Phylogenetic Trees

Phylogenetic analysis of DNA or protein sequences has become an important tool for studying the evolutionary history of organisms from bacteria to humans. Since the rate of sequence evolution varies extensively with gene or DNA segment (Wilson et al. 1977; Dayhoff et al. 1978), one can study the evolutionary relationships of virtually all levels of classification of organisms (e.g., kingdoms, phyla, families, genera, species, and intraspecific populations) by using different genes or DNA segments. Phylogenetic analysis is also important for clarifying the evolutionary pattern of multigene families (e.g., Atchley et al. 1994; Goodwin et al. 1996; Nei et al. 1997a) as well as for understanding the process of adaptive evolution at the molecular level (e.g., Jermann et al. 1995; Chandrasekharan et al. 1996; Zhang et al. 1998).

Reconstruction of phylogenetic trees by using statistical methods was initiated independently in numerical taxonomy for morphological characters (Sokal and Sneath 1963) and in population genetics for gene frequency data (Cavalli-Sforza and Edwards 1964). Some of the statistical methods developed for these purposes are still used for phylogenetic analysis of molecular data, but in recent years many new methods have been developed. In this book, we will discuss only methods that are useful for analyzing molecular data. For morphological data, the readers may consult Wiley et al. (1991), Maddison and Maddison (1992), and Swofford and Begle (1993). Before discussing tree-building methods, we first consider the types of phylogenetic trees in which molecular evolutionists are interested.

5.1. Types of Phylogenetic Trees

Rooted and Unrooted Trees

Phylogenetic relationships of genes or organisms are usually presented in a tree-like form either with a root (Figure 5.1A) or without any root (Figure 5.1B). The former tree is called a **rooted tree** and the latter an **unrooted tree.** The branching pattern of a tree, whether rooted or unrooted, is called a **topology.** There are many possible rooted and unrooted tree

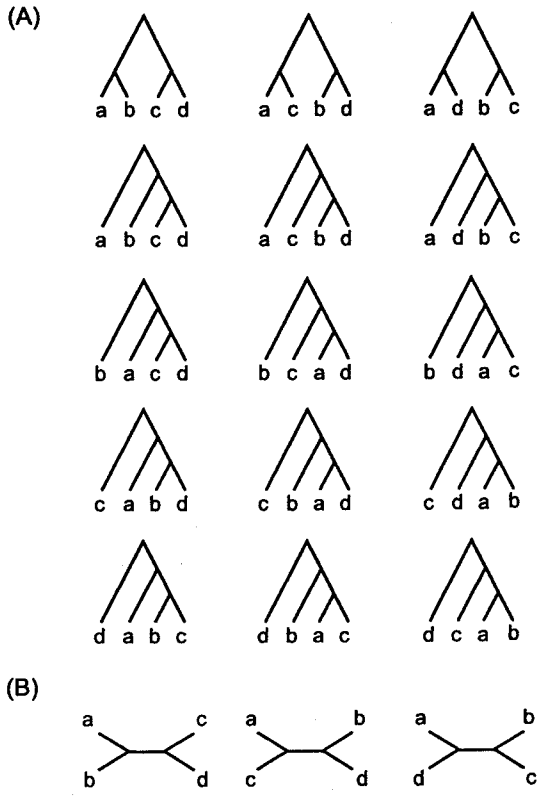


FIGURE 5.1. (A) Fifteen possible rooted trees and (B) three possible unrooted trees for four taxa.

topologies for a sizable number of **taxa** (any kind of taxonomic unit; families, species, populations, DNA sequences, etc.). If the number of taxa (m) is four, there are 15 possible rooted tree topologies and three possible unrooted tree topologies, as shown in Figure 5.1. The number of possible topologies rapidly increases with increasing m. In general, the number of possible topologies for a bifurcating rooted tree of m taxa is given by

$$1 \cdot 3 \cdot 5 \cdot \cdots (2m - 3) = [(2m - 3)!] / [2^{m-2}(m - 2)!] \qquad (5.1)$$

for $m \geq 2$ (Cavalli-Sforza and Edwards 1967). This indicates that the numbers of topologies for $m = 2$, 3, 4, 5, and 6 are 1, 3, 15, 105, and 945, respectively. When $m = 10$, it becomes $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9 \cdot 11 \cdot 13 \cdot 15 \cdot 17 = 34{,}459{,}425$. Only one of these topologies is the true tree. The number of possible topologies for a bifurcating unrooted tree of m taxa is given by replacing m by $m - 1$ in Equation (5.1). This becomes 2,027,025 for $m = 10$. In many cases, a majority of the possible topologies can be excluded from consideration because of obviously unlikely evolutionary relationships or because of other biological information. Nevertheless, it is a very difficult task to find the true tree topology when m is large.

In an unrooted bifurcating tree of m taxa there are $2m - 3$ **branches.** Since there are m **exterior branches** connecting to m extant taxa, the number of **interior branches** is $m - 3$. The number of **interior nodes** is equal to $m - 2$. In a rooted tree, the numbers of interior branches and interior nodes are $m - 2$ and $m - 1$, respectively, and the total number of branches is $2m - 2$.

Theoretically, a DNA sequence splits into two descendant sequences at the time of speciation or gene duplication. Therefore, phylogenetic trees are usually **bifurcating.** However, when a relatively short sequence is considered, some interior branches may show no nucleotide substitution, so that a multifurcating node may appear. This type of tree is often called a **multifurcating tree.** Most tree-building methods are for constructing a bifurcating tree, but the tree obtained can be reduced to a multifurcating tree by eliminating any branch that has zero length. It is also possible that even if the true tree is bifurcating, the reconstructed tree becomes multifurcating because of statistical errors. In reality, it is difficult to distinguish between the two cases.

Phylogenetic relationships of different DNA sequences are sometimes presented in a form of network with some loops. Loops are required when recombination occurs within a sequence or when the resolving power of mutational differences is low (e.g., Bandelt et al. 1995; Fitch 1997; Saitou and Yamamoto 1997; Page and Holmes 1998). In the former case, phylogenetic relationships in network form are a natural representation. In the latter case, the use of an experimental technique with a higher resolving power often resolves the network and reduces it to a bifurcating tree. For example, Avise et al. (1987) obtained a network tree when the variation of mitochondrial DNA in the deer mouse *Peromyscus polionotus* was analyzed by using a single restriction enzyme, but they could produce a bifurcating tree when eight restriction enzymes were used. In practice, network trees are produced only occasionally, so they will not be considered in this book.

Gene Trees and Species Trees

Evolutionists are often interested in a phylogenetic tree that represents the evolutionary history of a group of species or populations. This type of tree is called a **species** or **population tree.** In a species tree, the time of divergence between two species refers to the time when the two species were reproductively isolated. However, when a phylogenetic tree is constructed from one gene from each species, the tree obtained does not necessarily agree with the species tree. In the presence of polymorphic alleles at a locus, the times of divergence of genes sampled from different species are expected to be longer than the time of species divergence (Figure 5.2). The branching pattern of a tree constructed from genes may also be different from that of the species tree. To distinguish this tree from the species tree, we call it a **gene tree** (Nei 1986, 1987). Figure 5.3 shows three different possible relationships between species trees and gene trees for the case of three species. In relationships A and B, the topologies of the species and the gene trees are the same, but in relationship C they are different. If we use the gene genealogy theory in population ge-

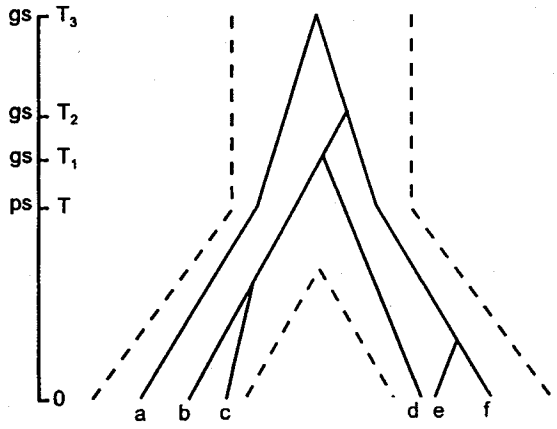


FIGURE 5.2. Diagram showing that the time of gene splitting (gs) is usually earlier than the time of population splitting (ps) when polymorphism exists. From Takahata and Nei (1985).

netics (Tajima 1983), it is possible to compute the probability of occurrence of events A, B, and C (Nei 1987; Pamilo and Nei 1988). The probability of occurrence of relationship C is quite high when the time interval between the first and second species splitting measured in terms of the number of generations (T) is short and the effective population size (N) is large.

Suppose that the long-term effective population size (N) is 10,000 as in the case of some mammals (Nei and Graur 1984) and the interval between two speciation events is one million years. If the generation time of the organism under consideration is 5 years, T becomes 200,000 generations. In this case, the probability [$P(C)$] of occurrence of relationship C is $(2/3)[\exp - (T/2N)] = 0.00003$, which is virtually 0. If N is as large as 100,000 but the generation time is 1 year as in the case of some invertebrate organisms, $P(C)$ becomes 0.004, which is again negligibly small. Therefore, if we consider a group of organisms where speciation event has occurred every one or two million years, the probability that the gene tree is different from the species tree is very small.

By contrast, if $N = 10{,}000$, $T = 100{,}000$, and the generation time is 5 years, we obtain $P(C) = 0.245$, which is substantial. Therefore, for a group of closely related species or intraspecific populations, the chance that the gene tree does not agree with the species or population tree is quite high. This was indeed the case with the DNA sequences from several nuclear genetic loci in a group of recently generated cichlid fish species of Lake Victoria in Africa (Nagl et al. 1998) or with the mitochondrial DNA sequences from several different human populations (Vigilant et al. 1991). To obtain a reliable tree of intraspecific populations or closely related species, interpopulational genetic distances based on a large number of genes from independently evolving (unlinked) loci need to be used (Saitou and Nei 1986; Pamilo and Nei 1988).

It should also be noted that even if the actual pattern of gene splitting

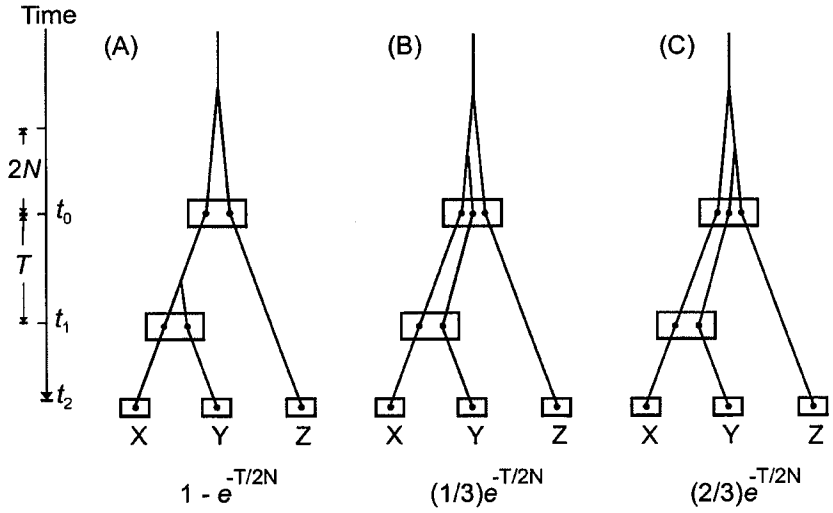


FIGURE 5.3. Three possible relationships between the species and gene trees for the case of three species in the presence of polymorphism. The times of the first and second species splitting are t_0 and t_1, respectively. The probability of occurrence of each tree is given underneath the tree. $T = t_1 - t_0$, and N is the effective population size. From Nei (1987).

agrees with that of species splitting, the branching pattern of a reconstructed gene tree may not agree with that of the species tree if the number of nucleotides or amino acids examined is small. This is because nucleotide or amino acid substitution occurs stochastically, so that the number of substitutions in lineage Z in Figure 5.3A or B may be smaller than that in lineage X or Y. To avoid this type of error, we must examine a large number of nucleotides or amino acids (Saitou and Nei 1986).

When the gene studied belongs to a multigene family, another problem may occur. Suppose that two related species, species 1 and 2, have two duplicate genes a_1 and b_1 and a_2 and b_2, respectively, and that the duplicate genes were generated by gene duplication that occurred before the divergence of the two species (Figure 5.4). In this case, genes a_1 and a_2 or b_1 and b_2 from the different species are called **orthologous genes,** whereas pairs of genes a_1 and b_1, a_2 and b_2, a_1 and b_2, and a_2 and b_2 are called **paralogous genes** (Fitch 1970). To construct a phylogenetic tree of different species, we should use orthologous genes rather than paralogous genes, because only orthologous genes represent speciation events. In practice, however, the distinction between orthologous and paralogous genes is not always easy, particularly when there are many copies of duplicate genes in the genome. We should, therefore, exercise great caution in the inference of species trees from gene trees.

Of course, gene trees are not always produced just to infer species trees. In the study of evolution of multigene families, it is important to know the evolutionary history of member genes and the process of gene duplication. In this case, we must study gene trees.

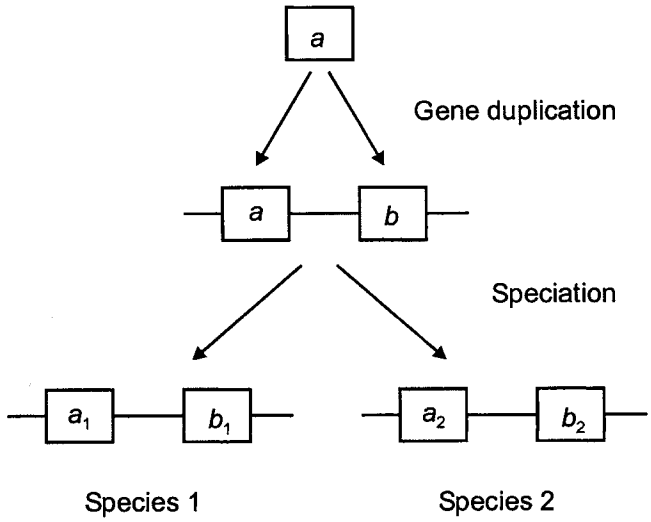


FIGURE 5.4. Duplicate genes from two different species. Genes a_1 and a_2 and b_1 and b_2 are orthologous, whereas pairs of genes a_1 and b_1, a_1 and b_2, a_2 and b_1, etc., are called paralogous.

Expected and Realized Trees

In the theory of phylogenetic inference, it is often assumed that the DNA or protein sequences to be studied are very long (theoretically infinitely long) and that a large number of amino acids or nucleotides that represent a random sample from the long sequences are sampled. While this assumption simplifies the statistical analysis of DNA or protein sequences, investigators are often interested in reconstructing the evolutionary history of a short sequence. For example, if one wants to know the long-term evolution of homeobox genes, he or she must work with a sequence of about 60 codons, because this is the size of the highly conserved homeobox domain (Kappen et al. 1993; Duboule 1994).

If we consider a short gene or a short segment of DNA, the number of nucleotide or amino acid substitutions is subject to large stochastic errors. Therefore, even if the expected number of substitutions increases linearly with time, a phylogenetic tree representing the actual number of substitutions could be very different from what one might expect intuitively. In this case, even the topology of the tree could be different from that of the tree from long DNA sequences. A tree that can be constructed by using infinitely long sequences or the expected number of substitutions for each branch is called an **expected tree,** whereas a tree based on the actual number of substitutions is called a **realized tree** (Nei 1987; Kumar 1996b). Note that both expected and realized trees are often different from the tree reconstructed (**reconstructed** or **inferred tree**) from observed sequence data.

Figure 5.5 shows one example of the differences among the expected, realized, and reconstructed trees when the molecular clock is assumed to work. Tree A in this figure represents an expected tree with each

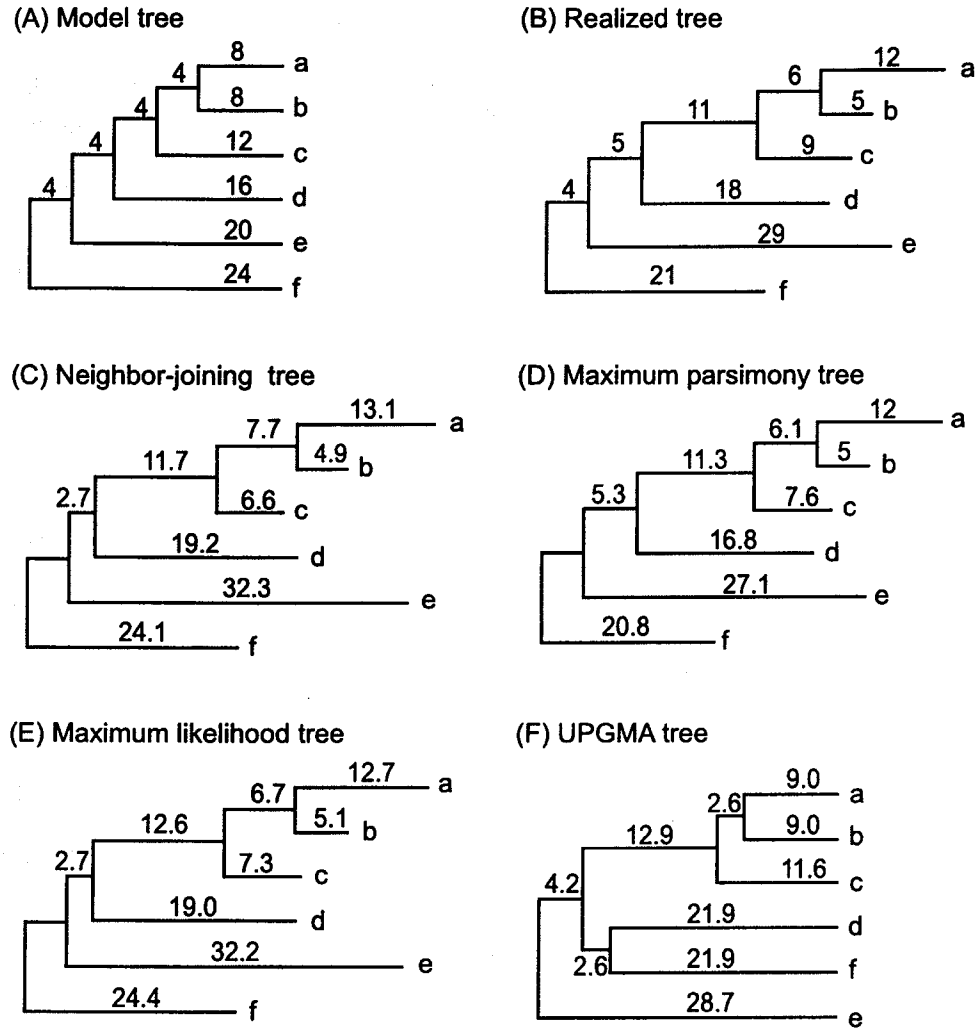


FIGURE 5.5. (A) Model tree, (B) realized tree, and (C–F) reconstructed trees. The neighbor-joining, maximum likelihood, and UPGMA trees were constructed with the Jukes-Cantor model. The branch lengths of the maximum parsimony tree were estimated by the average pathway method.

branch length equal to the expected number of nucleotide substitutions. In this case, the expected number of substitutions from the root to each terminal node is 0.12 per nucleotide site. Therefore, if a sequence of 200 nucleotides is used, the expected number of substitutions per sequence is 24. Tree B is a realized tree obtained in a replication of computer simulation under the assumption that the number of nucleotides used is 200 and nucleotide substitution occurs following the Jukes-Cantor model. The number given for each branch of tree B represents the number of substitutions that actually occurred for that branch. This number is considerably different from the expected value in tree A because of stochastic errors of nucleotide substitution.

Which tree does a tree-building method attempt to reconstruct, the ex-

pected tree or the realized tree? The answer to this question depends on the method of reconstruction, but most methods are intended to reconstruct realized trees. Figure 5.5C shows a tree reconstructed by the neighbor-joining method, which will be explained in chapter 6. The topology of the tree is identical with that of the expected (model) and the realized trees. However, the branch lengths of this tree are very different from those of the model tree and are close to those of the realized tree. This clearly indicates that the neighbor-joining method is intended to infer a realized tree rather than the model tree. Figures 5.5D and 5.5E show the trees constructed by the maximum parsimony and the maximum likelihood methods, respectively. The reconstructed trees are closer to the realized tree than to the model tree, indicating that they are also for inferring a realized tree.

By contrast, the topology of the tree (Figure 5.5F) obtained by the unweighted pair-group method with arithmetic mean (UPGMA) (see chapter 6) is different from that of both the model tree and the realized tree. Because of the incorrect topology, comparison of the branch lengths of the UPGMA tree with those of the model or the realized tree is not very meaningful, but the branch lengths for the correct part (sequences a, b, and c) of the topology are closer to those of the model tree rather than to those of the realized tree. In the present example, the expected number of nucleotide substitutions (4) for each interior branch was small, so that UPGMA could not produce the correct topology because of stochastic errors. However, if the number is two times higher, UPGMA would produce the correct topology with a high probability, and in this case, the branch lengths would have been closer to those of the model tree (Tateno et al. 1982). In other words, UPGMA is intended to infer the model tree or species tree, but unfortunately the topology of the UPGMA tree is disturbed by stochastic errors and other factors more easily than that of the trees obtained by other tree-building methods.

One might argue that what we really want to know is the expected or true tree rather than the realized tree. This is surely the case when a phylogenetic tree for a group of species is to be constructed. In practice, however, it is easier to reconstruct a realized tree than an expected tree, because the sequence data available refer to the realized tree. Note also that the topology of a realized tree is the same as that of the expected tree, unless a realized tree becomes a multifurcating tree because of stochastic errors. A realized tree may become a multifurcating tree when no nucleotide substitution occurs for one or more interior branches of the model tree by chance. As the number of nucleotides examined (n) increases, the realized tree is expected to approach the expected tree.

When one is interested in constructing species or population trees, the expected tree must have branch lengths proportional to evolutionary times, and two evolutionary lineages descendent from an interior node must have the same branch lengths. In practice, it is not easy to reconstruct a species tree defined in this way. Since the evolutionary change of genes is subject to stochastic errors as well as some kinds of selection, even a tree based on many genes could be different from the true species tree. At the present time, many investigators seem to be satisfied if they can reconstruct the correct or nearly correct topology even though the

branch length estimates are not proportional to evolutionary time. In estimating species or population trees, it is important to use as many genes as possible (e.g., Kidd et al. 1974; Nei and Roychoudhury 1974; Doolittle et al. 1996).

5.2. Topological Differences

Topological Distance

Although the true topology is generally unknown in actual data analysis, it is often useful to measure the extent of topological differences between two trees. For example, when one wants to know alternative trees that are closely related to a reconstructed tree, it is necessary to measure the topological distances of the alternative trees from the reconstructed tree. In the measurement of topological distance, it is customary to give no consideration to branch length differences.

The **topological distance** between two different trees is commonly measured by using Penny and Hendy's (1985) method of sequence partitioning. This distance gives the same numerical values as those obtained by Robinson and Foulds' (1981) method but is simpler to compute. For unrooted bifurcating trees, this distance is twice the number of interior branches at which sequence partition is different between the two trees compared. As an example, consider unrooted trees A and B in Figure 5.6. Both trees are for eight sequences and have five interior branches. It is possible to cut the tree at any interior branch and divide the sequences into two groups. Cutting at some interior branch results in the same partition of sequences in both trees A and B but not at other branches. For example, a cut at branch *a* produces two sequence groups (1,2) and (3, 4, 5, 6, 7, 8) in both trees. A cut at branch *c*, however, produces different partitions in trees A and B. That is, the two groups produced by this cut are (1, 2, 3, 4) and (5, 6, 7, 8) in tree A but (1, 2, 3, 5) and (4, 6, 7, 8) in tree B. Similarly, a cut at branch *d* produces different partitions in the two trees. In the present example, only these two cuts result in different partitions. Therefore, the topological distance between the two trees (d_T) is $2 \times 2 = 4$.

In general, if two trees for eight sequences have the same topology, $d_T = 0$, and if all interior branches produce different partitions, $d_T = 10$. However, if the two trees compared have multifurcating nodes, the above rule does not work. In this case, we can use Rzhetsky and Nei's (1992a)

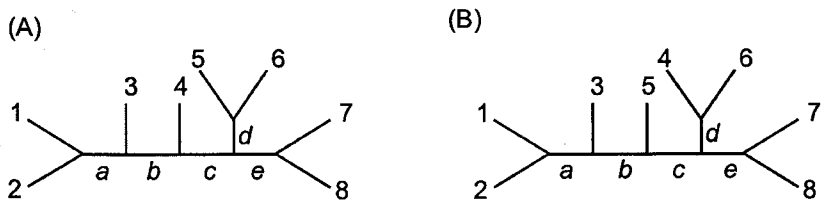


FIGURE 5.6. Two unrooted trees for eight sequences.

general formula for computing d_T for a pair of arbitrary trees for m sequences.

$$d_T = 2[\text{Min}(q_1,q_2) - p] + |q_1 - q_2| \tag{5.2}$$

where q_1 and q_2 are the total numbers of possible partitions (interior branches) for trees 1 and 2, respectively, and p is the number of partitions that are identical for the two trees. q_1 and q_2 may not be the same when multifurcating nodes are involved, and $\text{Min}(q_1,q_2)$ means the smaller value of q_1 and q_2. For bifurcating trees, however, q_1 and q_2 are always the same, and d_T takes only even numbers. In general, an unrooted bifurcating tree for m sequences has $m - 3$ interior branches, so that the maximum possible value of d_T is $2(m - 3)$.

In some methods of phylogenetic inference (see chapter 6) all trees that are different from an initial tree with topological distances $d_T = 2$ and 4 are examined in the search for the most likely tree. (Actually, this process is repeated.) Rzhetsky and Nei (1992a) have shown that the number of trees (topologies) that are different from a given topology by $d_T = 2$ is given by $f(d_T = 2) = 2(m - 3)$, whereas the number for $d_T = 4$ is $f(d_T = 4) = 2(m^2 - 4m + 3m' - 6)$, where m' is the number of tree nodes that are connected to one interior branch and two exterior branches ($m \geq 4$). For example, in tree A of Figure 5.6, $m = 8$ and $m' = 3$. Therefore, $f(d_T = 4) = 70$. This number is fairly large but represents a small portion of the total number of possible topologies (10,395). Therefore, it is much easier to examine the topologies with $d_T = 2$ and 4 rather than all topologies once a plausible tree is found.

Symbolic Expression of Topologies

Although a bifurcating or multifurcating tree can generally be drawn in a two-dimensional space, it is often convenient to use symbolic expressions to represent different tree topologies. Actually, any bifurcating or multifurcating tree can be expressed by a simple symbolic expression. For example, the topology of trees A–E in Figure 5.5 can be expressed as (f(e(d(c(b, a))))) and that of tree F as (e((d, f)(c(b, a)))). A multifurcating tree can also be expressed in the same way. Suppose that taxa a, b, and c are derived from one trifurcating node rather than two bifurcating nodes in trees A–E. The topology of the tree can then be expressed as (f(e(d(c, b, a)))).

In the case of unrooted trees, there are several different ways of describing a topology. One simple method is to subdivide all the taxa into three subgroups of taxa that join at an interior node and then decompose each subgroup consisting of three or more taxa into further subgroups of taxa. For example, in the case of tree A of Figure 5.6, we can first consider three subgroups of taxa (1, 2), 3, and (4, 5, 6, 7, 8). This forms only one topology, but we can further decompose subgroup (4, 5, 6, 7, 8) and write the topology of the entire tree as ((1, 2) 3 (4 ((5, 6), (7, 8)))). When there are multifurcating nodes, we have to use a slightly different expression. Suppose that taxa 5, 6, 7, and 8 are connected through one mul-

tifurcating node in tree A. Then the topology can be written as ((1, 2) 3 (4 (5, 6, 7, 8))) or (((1, 2) 3) 4 (5, 6, 7, 8)).

If we use the above symbolic expressions, all tree topologies can be distinguished from one another. This method of distinction is important in the examination of many different topologies to find the most likely tree, which will be discussed later.

5.3. Tree-Building Methods

There are many statistical methods that can be used for reconstructing phylogenetic trees from molecular data. Commonly used methods are classified into three major groups: (1) distance methods, (2) parsimony methods, and (3) likelihood methods. Details of these methods will be discussed in the next three chapters. Recently, Hendy and his colleagues (Hendy and Charleston 1993; Hendy and Penny 1993; Hendy et al. 1994) proposed the use of Hadamard conjugation for phylogenetic reconstruction (closest tree method). Dopazo and Carazo (1996) also proposed a neural network method of phylogenetic reconstruction. However, the practical utility of these methods has not yet been examined. Therefore, these methods will not be discussed.

It is now customary to consider the reconstruction of a phylogenetic tree as a statistical inference of a true phylogenetic tree, which is unknown. There are two processes involved in this inference: "estimation" of the topology and estimation of branch lengths for a given topology. When the topology is known, estimation of branch lengths is relatively simple, and there are several statistical methods one may use (e.g., least squares and maximum likelihood methods). The problem is the estimation or reconstruction of a topology. When there are a sizable number of DNA or protein sequences (say, 20), the number of possible topologies is enormously large as mentioned above, so that it is generally very difficult to find the true topology among them.

In phylogenetic inference, a certain optimization principle such as the maximum likelihood or the minimum evolution principle is often used for choosing the most likely topology. The theoretical basis of this procedure is not well understood, as will be discussed later, but computer simulations have shown that the optimization principles currently used generally work quite well if the number of nucleotides or amino acids used (n) is large. When this number is small and the number of sequences used is large, the optimization principle tends to give incorrect topologies, as will be discussed in chapter 9.

Some authors (e.g., Felsenstein 1978, 1988) have considered a tree topology as a **parameter** in statistical estimation and regarded a tree-building method as a **statistic** (or **estimator**) for estimating the parameter, as in the case of estimation of the mean of a statistical distribution. Therefore, Felsenstein (1978) used the concept of **inconsistency** in statistics to argue the inferiority of parsimony methods to likelihood methods under certain conditions. In statistical theory, if a statistic approaches the true parameter as the sample size (number of nucleotides

or amino acids in the present case) increases to infinity, the statistic is called a **consistent estimator.** In phylogenetic inference, a tree-building method does not represent any numerical quantity, so it is not a statistic as used in standard statistical theory. Nevertheless, inconsistency is a convenient way of describing a property of a tree-building method, so it is often used in phylogenetics.

If we are allowed to use a similar statistical concept for estimating a topology, we can consider whether a tree-building method gives the correct topology when the same evolutionary process is repeated an infinite number of times with a finite value of n. If a tree-building method gives the correct topology in this case, one may say that the tree-building method is an "unbiased estimator." If we use this definition, it can be shown that all tree-building methods based on the optimization principle are not "unbiased estimators" of the true topology and therefore tend to give incorrect topologies (Nei et al. 1998). In the case of maximum likelihood methods, some authors considered topologies as random variables and attempted to estimate the topology under this statistical framework (Cavalli-Sforza and Edwards 1967; Rannala and Yang 1996). In this case, one has to use a mathematical model for the pattern of species splitting. Rannala and Yang (1996) used the birth-and-death process in probability theory for this purpose, but since the real pattern of species splitting is very complicated, it is unclear how well this approach performs in real data analysis.

At present, the methodology of phylogenetic reconstruction is quite controversial. There seem to be at least three reasons for this controversy, putting aside personal preference. First, some workers were originally trained as systematicists using morphological characters, and they tend to be suspicious about any methods that are based on mathematical models of evolutionary changes, because the evolutionary change of morphological characters is so complex that it does not obey any simple rule. They therefore prefer parsimony methods, which require a minimum number of assumptions. Another group of workers has been trained as geneticists or molecular biologists and tends to prefer using analytical approaches but does not trust highly sophisticated mathematical models. A third group of workers is primarily trained as mathematicians or statisticians and tries to understand the construction of phylogenetic trees as a mathematical problem rather than a practical problem, using abstract mathematical concepts. Since the approaches used by the three groups of workers are quite different, controversies naturally occur.

Second, some scientists are primarily interested in short-term evolution within species or between closely related species, whereas others are interested in long-term evolution dealing with different orders, phyla, or kingdoms. The methodologies used by these two groups of scientists are quite different, and one group tends to feel wary of the approach used by the other group.

Third, in phylogenetic analysis, the true tree is almost always unknown, and it is difficult to test the accuracy of the trees obtained by different tree-building methods. Currently, there are several statistical criteria for evaluating the accuracy, but all of them depend on a number of simplifying assumptions. Therefore, none of them is perfect. Fur-

thermore, the theoretical basis of the statistical methods currently used for phylogenetic reconstruction is not well established, as mentioned above. The mathematical models used for describing sequence evolution are crude approximations to reality, and a sophisticated model does not necessarily give better results. Therefore, there is plenty of room for controversy.

Molecular phylogenetics is still a young scientific discipline, and it is important to realize that every statistical method has some strengths and some weaknesses, and none of the methods is almighty. One cannot reject a method simply because it did not work in a particular study or a particular computer simulation. The overall assessment of superiority of one method over the other should come from broad theoretical and experimental studies. As mentioned above, the evolutionary change of DNA or proteins is so complicated that the mathematical model used is necessarily approximate. Unlike the case in physics, the predictive power of a model in biology is quite low. It seems to us that if the prediction (e.g., a phylogenetic tree reconstructed) of a model is correct in 80% of the cases, it is a good model at least at the present time. In the case of molecular phylogenetics, one can study the phylogeny of a group of organisms using a large number of genes, and this comprehensive study will eventually clarify the evolutionary relationships of organisms.

In the following three chapters, we will discuss various tree-building methods without going into mathematical details and cover only methods that have proved to be useful for practical data analysis. However, the theoretical basis of each method will be discussed with minimum mathematics and verbal arguments as much as possible. In these chapters, we assume that the number of nucleotides or amino acids used (n) is sufficiently large so that phylogenetic inference based on optimization criteria works well. The performance of optimization criteria when n is small will be discussed in chapter 9.

6

Phylogenetic Inference: Distance Methods

In **distance methods** or **distance matrix methods,** evolutionary distances are computed for all pairs of taxa, and a phylogenetic tree is constructed by considering the relationships among these distance values. There are many different methods of constructing trees from distance data. Here we discuss only the methods that have proved to be useful for actual data analysis.

6.1. UPGMA

The simplest method in this category is the **unweighted pair-group method using arithmetic averages (UPGMA).** This method is often attributed to Sokal and Michener (1958), but the method used by these authors is quite different from the currently used version. Its clear-cut algorithm appears in Sneath and Sokal's (1973) book. A tree constructed by this method is sometimes called a **phenogram,** because it was originally used to represent the extent of phenotypic similarity for a group of species in numerical taxonomy. However, it can be used for constructing molecular phylogenies when the rate of gene substitution is more or less constant. Particularly when gene frequency data are used for phylogenetic reconstruction, this model produces reasonably good trees compared with other distance methods (Nei et al. 1983; Takezaki and Nei 1996). In this case, a distance measure that has a smaller coefficient of variation seems to give better trees than other distance measures even if it is not strictly proportional to the number of gene substitutions (Takezaki and Nei 1996). UPGMA is intended to reconstruct a species tree, although topological errors often occur when the rate of gene substitution is not constant or when the number of genes or nucleotides used is small.

Algorithm

In UPGMA, a certain measure of evolutionary distance is computed for all pairs of taxa or sequences, and the distance values are presented in the following matrix form.

Taxon	1	2	3	4
2	d_{12}			
3	d_{13}	d_{23}		
4	d_{14}	d_{24}	d_{34}	
5	d_{15}	d_{25}	d_{35}	d_{45}

Here, d_{ij} stands for the distance between the i-th and j-th taxa. Clustering of taxa starts with a pair of two taxa with the smallest distance. Suppose that d_{12} is smallest among all distance values in the above matrix. Taxa 1 and 2 are then clustered with a branch point located at distance $b = d_{12}/2$. Here, we have assumed that the lengths of the branches leading from this branch point to taxa 1 and 2 are the same (see Example 6.1). Taxa 1 and 2 are then combined into a single composite taxon or cluster [u = (1-2)], and the distance between this u and another taxon $k(k \neq 1, 2)$ is computed by $d_{uk} = (d_{1k} + d_{2k})/2$. Therefore, we have the following new matrix.

Taxon	u = (1-2)	3	4
3	d_{u3}		
4	d_{u4}	d_{34}	
5	d_{u5}	d_{35}	d_{45}

Now suppose that distance d_{u3} is smallest. Then, taxa u and 3 are combined into a new composite taxon or cluster [v = (1-2-3)] with a branch point of $b = d_{u3}/2 = (d_{13} + d_{23})/(2 \times 2)$. The distance between the newly created cluster v and each of the remaining taxa (k's) is now computed by $d_{kv} = (d_{k1} + d_{k2} + d_{k3})/3$. We then have

Taxon	v = (1-2-3)	4
4	d_{v4}	
5	d_{v5}	d_{45}

Let us assume that d_{v4} is smallest in the above distance matrix. We then combine v = (1-2-3) and 4 with a branch point of $b = d_{v4}/2 = (d_{14} + d_{24} + d_{34})/(3 \times 2)$. It is obvious that the last taxon to join the tree is 5, and the branch point is given by $b = (d_{15} + d_{25} + d_{35} + d_{45})/(4 \times 2)$.

It is of course possible that in the second matrix the smallest distance is d_{45} (or any other one) instead of d_{u3}. In this case, taxa 4 and 5 are joined with the branch point of $b = d_{45}/2$, and a new composite taxon, v = (4-5), will be created. The distances between v and other taxa (3 and u are given by $d_{3v} = (d_{34} + d_{35})/2$ and $d_{uv} = (d_{14} + d_{15} + d_{24} + d_{25})/4$. Now suppose that d_{uv} is smallest. Then, taxa u and v are clustered, and taxon 3 will be the last to join the cluster. Of course, if d_{3v} is smallest, taxa 3 and v cluster first.

As is obvious from the above example, the distance between two clusters (A and B) is given by the following formula.

$$d_{AB} = \sum_{ij} d_{ij}/(rs) \tag{6.1}$$

where r and s are the numbers of taxa in clusters A and B, respectively, and d_{ij} is the distance between taxon i in cluster A and taxon j in cluster B. The branch point between the two clusters is given by $d_{AB}/2$. For the purpose of computer programming, however, the above equation is not convenient, and other faster algorithms are used to compute d_{AB} (e.g., Swofford et al. 1996).

Statistical Tests of UPGMA Trees

Rooted and Unrooted UPGMA Trees

A tree obtained by UPGMA is usually presented as a rooted tree, because it is easy to infer the root of the tree under the assumption of a constant rate of evolution. However, UPGMA is a method of inferring both the topology and branch lengths similar to other methods, and we do not have to give the root to a UPGMA tree. In other methods of phylogenetic inference, an uprooted tree is usually constructed, because it is difficult to determine the root when the evolutionary rate varies from branch to branch. We can use the same approach and construct an unrooted UPGMA tree, disregarding the root usually given to a UPGMA tree. When we compare an UPGMA tree with trees constructed by other methods, we should use this unrooted UPGMA tree, because rooting can introduce an additional source of errors in tree building. Unrooted trees are also useful for testing the reliability of the tree obtained by using the bootstrap or other method, as will be discussed below.

Reliability of UPGMA Trees

Since a phylogenetic tree is usually constructed from a limited amount of data, it is important to examine the reliability of the tree obtained. As will be discussed in detail in chapter 9, there are two major methods of testing the reliability of the topology of a tree obtained by distance methods. In the case of UPGMA trees, we can use Nei et al.'s (1985) **interior branch test** or Felsenstein's (1985) **bootstrap test.** Both tests examine the reliability of each interior branch of a tree. If every interior branch length is proved to be positive, the tree is regarded as reliable from the statistical point of view. However, Nei et al.'s test becomes complicated when the number of taxa examined is large. A simpler way of testing the positiveness of an interior branch is to use the bootstrap test considering unrooted UPGMA trees (see chapter 9 for details). In this test, it is customary to compute a quantity equivalent to the probability of confidence (1 − Type I error) rather than the significance level. This value is called the **bootstrap confidence value** (P_B) or **bootstrap value.** If this value is higher than 95% (or 99% depending on the confidence level one wishes to have), the interior branch is considered to be statistically significant (Felsenstein 1985; Efron et al. 1996). In this book, we will use this bootstrap technique extensively.

It is known that when closely related DNA (or protein) sequences are used for constructing UPGMA trees, two or more trees **(tie trees)** may be

produced from the same distance (Kim et al. 1993; Backeljau et al. 1996; Takezaki 1998). These tie trees occur because two or more distance values in a distance matrix occasionally become identical. It is possible to enumerate all these tie trees (Rohlf 1993), but this enumeration is not very meaningful, since these tie trees are primarily caused by sampling errors of distance estimates and they are close to one another. A better way of treating this problem is to construct a **bootstrap consensus tree** (see section 9.3). This consensus tree also has a bootstrap value for each interior branch, and it can be treated in the same way as the above UPGMA tree with bootstrap values. Therefore, we will know the reliability of each branching pattern of the UPGMA tree. When only one UPGMA tree exists for a given data set, the bootstrap consensus tree is usually identical with the original UPGMA tree (Takezaki 1998).

Example 6.1. UPGMA Tree of Hominoid Species

Figure 6.1 shows the nucleotide sequences of a segment (896 nucleotides) of mitochondrial DNA (mtDNA) from humans, chimpanzees, gorillas, orangutans, and gibbons (Brown et al. 1982). In this data set, transitional nucleotide differences are considerably greater than transversional differences. The average transition/transversion ratio (R) obtained by Equation (3.18) is about 6.2. We therefore estimated the number of nucleotide substitutions (d) per site using Equation (3.12) (Kimura distance). The results obtained are presented in Table 6.1. (One site containing an alignment gap was removed.) It is seen that the value between humans and chimpanzees ($\hat{d} = 0.095$) is smallest, so that humans and chimpanzees are the first to be clustered with a branch point at $b_{HC} = 0.095/2 = 0.048$ (Figure 6.2A). Humans and chimpanzees are now combined into a single taxon, (HC). The $\hat{d}$ values between this taxon and gorillas, orangutans, and gibbons become $(0.113 + 0.118)/2 = 0.115$, $(0.183 + 0.201)/2 = 0.192$, and $(0.212 + 0.225)/2 = 0.218$, respectively. The other distance values remain unchanged. The smallest d value in the new d matrix is that (0.115) between (HC) and gorillas. Thus, gorillas join (HC) with a branch point at $b_{G(HC)} = 0.115/2 = 0.058$. If this type of computation is repeated, we finally obtain the phylogenetic tree given in Figure 6.2A.

The tree given in this figure is an unrooted tree, and there are two interior branches. The bootstrap values for the interior branches are written in boldface. In the present case, the bootstrap consensus UPGMA tree is virtually identical with the tree in Figure 6.2A. The branch separating the group of humans, chimpanzees, and gorillas from the two other species shows a bootstrap value of 100%, whereas the branch separating humans and chimpanzees from gorillas has a value of 90%. Therefore, this data set establishes the cluster of humans, chimpanzees, and gorillas but is not sufficient to resolve the branching pattern among these three species at the 95% confidence level. In this data set, application of the statistical test of rate constancy given in chapter 10 does not reject the hypothesis of a molecular clock. Therefore, the use of UPGMA for inferring species trees is justified.

```
Human      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTTACATCCTCATTACTATTCTGCCTAGCAAACTCAAACT  80
Chimpanzee ..............A.T..C.....................................T.....................T.
Gorilla    ..................TG....T.....T.................A......T.......................
Orangutan  ................AC..CC.....G..T.....T.....C.........CC....G.....................
Gibbon     ......T..A..T...AC.G.C.....................A..C..T..CC.G...........T.............

Human      ACGAACGCACTCACAGTCGCATCATAATCCTCTCTCAAGGACTTCAAACTCTACTCCCACTAATAGCTTTTTGATGACTT 160
Chimpanzee .T........C.................T.....C.....................................C...........C
Gorilla    .......A..C.....C.....T.....T.............C.....C.................CC...........
Orangutan  .......A..C.....C.....................C.................C........CC.C.........
Gibbon     .......A........C.............A....G...G..C...G.CT..........G.....C..C........C

Human      CTAGCAAGCCTCGCTAACCTCGCCTTACCCCCCACTATTAACCTACTGGGAGAACTCTCTGTGCTAGTAACCACGTTCTC 240
Chimpanzee .........................C.......T..C.....T..C..A..G........C...........T.A.....
Gorilla    ..G...........C.....................C...........A.....G.....C..A...........A.....
Orangutan  ............A.......T...C....A.....C..C.....T..A...........C..A...A..G...TA.....
Gibbon     GC.....................C.................C..A..T.......TC..A...A.GG..T.C.....

Human      CTGATCAAATATCACTCTCCTACTTACAGGACTCAACATACTAGTCACAGCCCTATACTCCCTCTACATATTTACCACAA 320
Chimpanzee ..........C...........C......T...........A..........G.............G..........
Gorilla    .........C.C...C..TT..........TCT.........A.T........G........T..T.............
Orangutan  T.....T..C.....CA.......A.................A.....A..........T.....T.....C.......
Gibbon     ...GG....C.CT...A.TAC...C..C..G......G.....A....G.............T..........T..T..

Human      CACAATGGGGCTCACTCACCCACCACATTAACAACATAAAACCCTCATTCACACGAGAAAACACCCTCATGTTCATACAC 400
Chimpanzee .......A.....................T.........G.....................T..T.....A..TT.....
Gorilla    .......A...C.......A........C.C.................T.............T......A.....G...
Orangutan  .....C.A..TA...C...A........C.............T..T........C.....T.........C........
Gibbon     .....C.A...A....T..A...........A.................C..............TAT.A..AC.T..G...

Human      CTATCCCCCATTCTCCTCCTATCCCTCAACCCCGACATCATTACCGGGTTTTCCTCTTGTAAATATAGTTTAACCAAAAC 480
Chimpanzee ..........C.....T...........T..T..T.....C..T..A..CA....C.....................
Gorilla    ..........C.........................T..T..C........CA....C.....................
Orangutan  ..........C......T..............AG......CG.T.....CG...AC.....................
Gibbon     ..C.T....C.C.........A..........TA..........T..C...A.TC.C......C.........T......

Human      ATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACCGAGAAAGCTCACAAGAACTGCTAACTCATGCCC 560
Chimpanzee ..........................C.....................................T.T.............T....AT..
Gorilla    ..................T..........C..A.....................GT....G.............A...
Orangutan  ..T.............A.T..T..G.C.CC..A.....................................TCA.T-
Gibbon     ..T.............A....T........CGAA...T...GC.............C.....................CTAT.

Human      CCATGTCTGACAACATGGCTTTCTCAACTTTTAAAGGATAACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGC 640
Chimpanzee .....C.....................................C....G.....................................
Gorilla    ..G..CT.....................................................A.....................
Orangutan  ......G.....................G.....................C...........AT.....................
Gibbon     ......A.....................................................A.....................

Human      AACTCCAAATAAAAGTAATAACCATGCACACTACTATAACCACCCTAACCCTGACTTCCCTAATTCCCCCCATCCTTACC 720
Chimpanzee ..........................T.T.....C.........T.......A...C..T.....T.........C...
Gorilla    ......................T...T..G....C.........T..G....A......T..........T.........
Orangutan  ..................C.G.....TTT..C..C.....TG....C...T.A...........C.........TACCG.T
Gibbon     ....................G.A...T....C..C...G...TT.....G..A..C.................TACAG..

Human      ACCCTCGTTAACCCTAACAAAAAAAACTCATACCCCCATTATGTAAAATCCATTGTCGCATCCACCTTTATTATCAGTCT 800
Chimpanzee .....A.....................T...........G.........A....G...........C..T..C..
Gorilla    ...T..A.C..T.............G..............C........T..C.................C.....C..
Orangutan  .....A.......C............C..........C.........A.GGCCA........G......C......C..
Gibbon     .....TA.......C..T.....G....T......G..C..C......ATG.CCA.T..C..T.........A.....C..

Human      CTTCCCCACAACAATATTCATGTGCCTAGACCAAGAAGTTATTATCTCGAACTGACACTGAGCCACAACCCAAACAACCC 880
Chimpanzee T.....................A.................C.........A.....G........A.................
Gorilla    ...................TC.A.................C.........A.G............A.............TT.
Orangutan  TA....A...........T..C.......GA......ACC..CG..A.A......TG....A.A..C.....G...CTA.
Gibbon     A..T.......T.............AC..........ACC.....T..A.............A.TG..........GCTAG

Human      AGCTCTCCCTAAGCTT 896
Chimpanzee ................
Gorilla    .A..............
Orangutan  .A.....A........
Gibbon     .A..............
```

FIGURE 6.1. Sequences of an 896 bp fragment of primate mitochondrial DNAs. The orangutan sequence has one deletion at position 560. Data from the GenBank.

Table 6.1 Kimura distances for the data shown in Fig. 6.1.

	Human	Chimpanzee	Gorilla	Orangutan
Chimpanzee	0.095 ± 0.011			
Gorilla	0.113 ± 0.012	0.118 ± 0.013		
Orangutan	0.183 ± 0.016	0.201 ± 0.018	0.195 ± 0.017	
Gibbon	0.212 ± 0.018	0.225 ± 0.019	0.225 ± 0.019	0.222 ± 0.018

Note: One site containing an alignment gap was removed from the analysis.

6.2. Least Squares (LS) Methods

When the rate of nucleotide substitution varies from evolutionary lineage to lineage, UPGMA often gives an incorrect topology. In this case, we should use methods that allow different rates of nucleotide substitution for different branches. One group of such methods is least squares (LS) methods. There are several different LS methods, but the most commonly used ones are the ordinary LS and the weighted LS methods.

Topology Construction

In the **ordinary LS method** of phylogenetic inference (Cavalli-Sforza and Edwards 1967), we consider the following residual sum of squares

$$R_S = \sum_{i<j} (d_{ij} - e_{ij})^2 \tag{6.2}$$

where d_{ij} and e_{ij} are the **observed** and **patristic distances** between taxa i and j, respectively. The patristic distance between taxa i and j is the sum of estimates of the lengths of all branches connecting the two taxa in a tree. For example, the patristic distance between humans and gorillas in

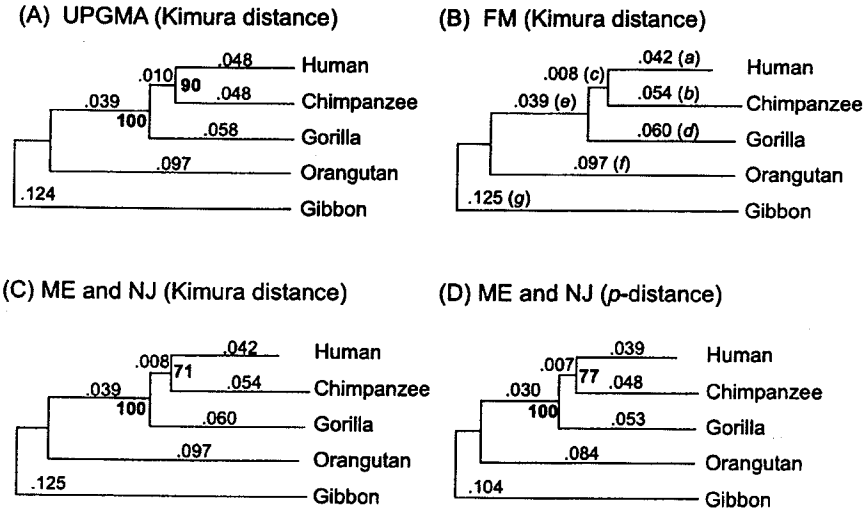


FIGURE 6.2. Evolutionary trees inferred by different distance methods. The UPGMA tree is unrooted. Bootstrap values are in boldface.

the tree of Figure 6.2B is $a + c + d = 0.110$. In the standard LS method, R_S is computed for all plausible topologies, and the topology with the smallest R_S value is chosen as the final tree.

Fitch and Margoliash (1967) used the following R_S value for choosing the final topology.

$$\sum_{i<j}[(d_{ij} - e_{ij})^2 / d_{ij}] \tag{6.3}$$

This procedure is called a **weighted LS method.** In practice, the R_S values defined in Equations (6.2) and (6.3) usually give the same topology or very similar topologies.

Theoretically, a better procedure would be to use the **generalized LS method** of computing R_S, in which both the variance and covariance of d_{ij}'s are taken into account (Cavalli-Sforza and Edwards 1967; Bulmer 1991). However, this method is very time consuming. Furthermore, when the d_{ij} values approach 0, the variance-covariance matrix becomes singular (Rzhetsky and Nei 1992b), and thus this method does not seem to give reliable phylogenetic trees.

Least-Squares Method with the Constraint of Nonnegative Branches

Using computer simulation, Saitou and Nei (1986) and Rzhetsky and Nei (1992a) studied the probability of obtaining the correct tree topology by the ordinary and weighted least-squares methods and showed that the probability is often lower than that of some other distance methods. Part of the reason seems to be that these methods often give negative estimates of branch lengths, which are theoretically unrealistic. Therefore, one way to improve the efficiency of this method would be to use the LS method with the constraint of nonnegative branches (Felsenstein 1995, 1997). Using computer simulation, Kuhner and Felsenstein (1994) indeed showed that this modified method increases the probability of obtaining the correct topology considerably. Estimation of branch lengths with the constraint of nonnegative values requires iterative computation of branch length estimates (Felsenstein 1995, 1997). It is also known that in the case of four taxa this method gives the same topology as that obtained by the neighbor joining method (see section 6.4) (Gascuel, 1994; M. Bulmer, personal communication, 1991).

Estimation of Branch Lengths

Fitch-Margoliash Method

To compute the residual sum of squares, R_S, we must first estimate the branch lengths and the e_{ij}'s for each topology. A simple way to estimate branch lengths is to use Fitch and Margoliash's method (1967). Although the estimates obtained by this method are not always the same as those obtained by the LS method, the differences are usually very small so that the Fitch-Margoliash method is still used. This method takes advantage

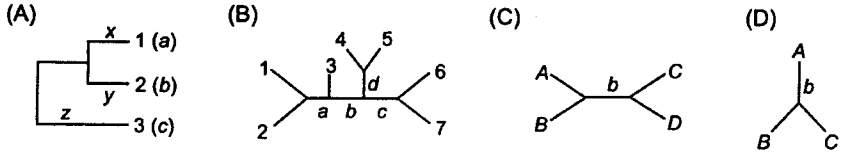


FIGURE 6.3. Estimation of branch lengths.

of the property that when there are only three taxa the estimates of branch lengths for all three taxa can be uniquely determined.

Consider three taxa 1, 2, and 3, of which the evolutionary relationships are given by Figure 6.3A. The evolutionary distances between taxa 1 and 2, 1 and 3, and 2 and 3 are then given by

$$d_{12} = x + y \tag{6.4a}$$

$$d_{13} = x + z \tag{6.4b}$$

$$d_{23} = y + z \tag{6.4c}$$

where x, y, and z are the branch lengths for taxa 1, 2, and 3, respectively. Solving these simultaneous equations gives

$$x = (d_{12} + d_{13} - d_{23})/2 \tag{6.5a}$$

$$y = (d_{12} - d_{13} + d_{23})/2 \tag{6.5b}$$

$$z = (-d_{12} + d_{13} + d_{23})/2 \tag{6.5c}$$

These are LS estimates.

When there are four or more taxa, we first choose the two taxa with the smallest distance and denote them by A and B. All the remaining taxa are combined into a single composite taxon designated by C. The distance between taxa A and B is the same as the original distance (d_{12}), but the distance between taxa A and C is now represented by the simple average of the distances between A and all taxa in C. Similarly, the distance between B and C is the average of the distances between B and all taxa in C. For example, in the distance matrix in Table 6.1, humans and chimpanzees show the smallest distance. Therefore, we denote humans and chimpanzees by A and B, respectively, and the remaining species by C. From the distance estimates given in Table 6.1, we have $d_{AB} = 0.095$, $d_{AC} = (0.113 + 0.183 + 0.212)/3 = 0.169$, and $d_{BC} = (0.118 + 0.201 + 0.225)/3 = 0.181$. The values of x, y, and z therefore become 0.042, 0.054, and 0.124, respectively, from Equations (6.5). Here x and y represent the number of estimated nucleotide substitutions (a and b) for the human and chimpanzee lineages, respectively, and z is the distance between the composite taxon C and the branch point between humans and chimpanzees (Figure 6.2B).

We now combine taxa 1 and 2 and designate the composite taxon as (AB). We then recompute the distances between this composite taxon

(*AB*) and all other taxa and choose the two taxa that show the smallest value among all distances, including those that do not involve (*AB*). These two taxa are again designated by *A* and *B*, whereas *C* represents the composite taxon consisting of all the remaining taxa. The new *x*, *y*, and *z* values are computed by the same procedure. In the case of hominoid data, the distances between (*AB*) and the other taxa (gorillas, orangutans, and gibbons) have already been computed (0.115, 0.192, and 0.218, respectively) when we constructed the UPGMA tree, and the smallest distance in the new matrix is that between (*AB*) and gorillas. Therefore, (*AB*) and gorillas are designated as the new *A* and *B*, respectively, whereas *C* represents orangutans and gibbons. We now have $d_{AB} = 0.115$, $d_{AC} = (0.183 + 0.201 + 0.212 + 0.225)/4 = 0.205$, and $d_{BC} = (0.195 + 0.225)/2 = 0.210$. Therefore, we have $x = 0.055$, $y = 0.060$, and $z = 0.150$ from Equations (6.5). Branch lengths *c* and *d* of the tree in Figure 6.2B are then estimated by using the following relationships.

$$d_{AB} = (a + b)/2 + c + d$$

$$d_{AC} = (a + b)/2 + c + z$$

$$d_{BC} = d + z$$

We know that $(a + b)/2 = 0.048$ and $z = 0.150$. Therefore, we have $c = 0.008$ and $d = 0.060$ (Figure 6.2B). The above procedure is repeated until all branch lengths are estimated. In the case of hominoid data, the estimates of the three remaining branches (*e*, *f*, and *g*) are presented in Figure 6.2B.

We are now in a position to compute e_{ij}'s for all pairs of taxa and then the R_S values in Equations (6.2) and (6.3). The latter values become 0.000047 and 0.002264, respectively. To find the LS tree, however, we must consider all possible or all plausible trees. In practice, the number of topologies is usually very large so that only a small proportion of possible topologies is examined for computing the R_S values. In Fitch-Margoliash's (1967) method, the first topology is constructed by the algorithm described above. Once this topology is obtained, different topologies are examined by various branch-swapping algorithms. These algorithms will be explained in the next chapter, where the algorithms are important in relation to the construction of maximum parsimony trees.

Once the final tree topology is obtained by minimizing R_S, better estimates of branch lengths of the final tree may be obtained by the LS method, which will be described below. Mathematically, the LS estimates are more reliable than those obtained by the Fitch-Margoliash method, but in practice, the differences between them are usually very small when DNA or protein sequences are used.

Least Squares Methods

The standard method of estimating the branch lengths of a tree is to use the LS method. Rzhetsky and Nei (1992a, 1993) developed a fast algorithm for obtaining LS estimates of branch lengths for any given topol-

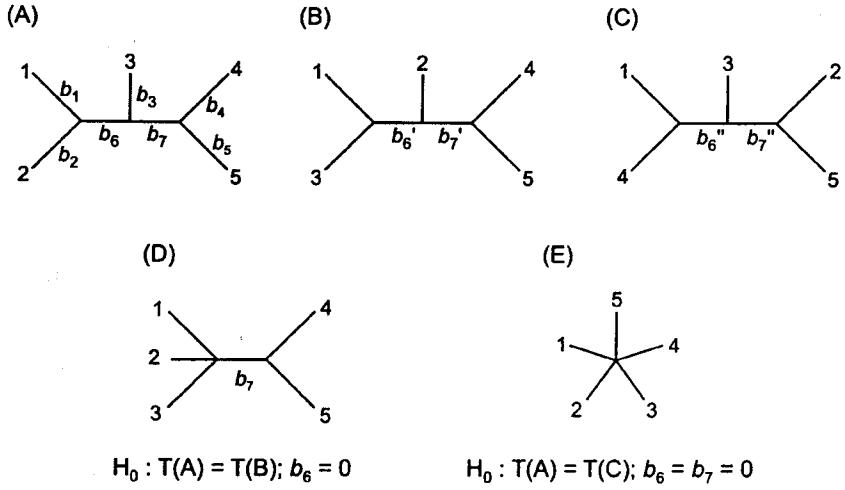


FIGURE 6.4. Three different topologies for five taxa and two "null trees" for testing topological differences.

ogy. Let us consider a hypothetical tree for five sequences given in Figure 6.4A and use the ordinary LS method to estimate the branch lengths denoted by b_1, b_2, . . . , and b_7. We represent an estimate of evolutionary distance between sequences i and j by d_{ij}. We can then write the d_{ij}'s as follows.

$$
\begin{aligned}
d_{12} &= b_1 + b_2 + \epsilon_{12} \\
d_{13} &= b_1 + b_3 + b_6 + \epsilon_{13} \\
d_{14} &= b_1 + b_4 + b_6 + b_7 + \epsilon_{14} \\
d_{15} &= b_1 + b_5 + b_6 + b_7 + \epsilon_{15} \\
d_{23} &= b_2 + b_3 + b_6 + \epsilon_{23} \\
d_{24} &= b_2 + b_4 + b_6 + b_7 + \epsilon_{24} \\
d_{25} &= b_2 + b_5 + b_6 + b_7 + \epsilon_{25} \\
d_{34} &= b_3 + b_4 + b_7 + \epsilon_{34} \\
d_{35} &= b_3 + b_5 + b_7 + \epsilon_{35} \\
d_{45} &= b_4 + b_5 + \epsilon_{45}
\end{aligned}
$$

where ϵ_{ij}'s are sampling errors. We assume that ϵ_{ij} is distributed with mean 0 and variance $V(d_{ij})$. If we use matrix algebra, the above set of equations may be written as

$$\boldsymbol{d} = \boldsymbol{A}\boldsymbol{b} + \boldsymbol{\epsilon} \tag{6.6}$$

where $\boldsymbol{d}$, $\boldsymbol{b}$, and $\boldsymbol{\epsilon}$ are column vectors of d_{ij}'s, b_i's, and ϵ_{ij}'s, respectively; that is, $\boldsymbol{d}^t = (d_{12}, d_{13}, \ldots, d_{45})$, $\boldsymbol{b}^t = (b_1, b_2, \ldots, b_7)$, and $\boldsymbol{\epsilon}^t = (\epsilon_{12}, \epsilon_{13}, \ldots, \epsilon_{45})$. Here t indicates the transpose of a vector or a matarix. Note that vectors $\boldsymbol{d}$ and $\boldsymbol{\epsilon}$ have $r \equiv m(m-1)/2$ elements and $\boldsymbol{b}$ has $T \equiv 2m - 3$ elements, where m is the number of sequences. $\boldsymbol{A}$ is a matrix representing a topology, and in this case (topology [A] in Figure 6.4) it is given by

$$\mathbf{A} = \begin{bmatrix} 1&1&0&0&0&0&0 \\ 1&0&1&0&0&1&0 \\ 1&0&0&1&0&1&1 \\ 1&0&0&0&1&1&1 \\ 0&1&1&0&0&1&0 \\ 0&1&0&1&0&1&1 \\ 0&1&0&0&1&1&1 \\ 0&0&1&1&0&0&1 \\ 0&0&1&0&1&0&1 \\ 0&0&0&1&1&0&0 \end{bmatrix} \tag{6.7}$$

An element of this matrix is 1 when there is a corresponding branch and 0 otherwise (see the equations for d_{ij}'s). The LS estimate of $\boldsymbol{b}$ is then given by

$$\hat{\boldsymbol{b}} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t\boldsymbol{d} = \boldsymbol{L}\boldsymbol{d} \tag{6.8}$$

where $\boldsymbol{L} = (\boldsymbol{A}^t\boldsymbol{A})^{-1}\boldsymbol{A}^t$. Obviously, an estimate of the length of the i-th branch is

$$\hat{b}_i = \boldsymbol{L}_i\boldsymbol{d} \tag{6.9}$$

where L_i is the i-th row of the matrix $\boldsymbol{L}$ (Rzhetsky and Nei 1992a). If we use this formula for the case of topology A in Figure 6.4, we obtain

$$\begin{aligned}
\hat{b}_1 &= \frac{1}{2}d_{12} + \frac{1}{6}(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25}) \\
\hat{b}_2 &= \frac{1}{2}d_{12} - \frac{1}{6}(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25}) \\
\hat{b}_3 &= \frac{1}{4}(d_{13} + d_{23} + d_{34} + d_{35}) - \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25}) \\
\hat{b}_4 &= \frac{1}{2}d_{45} + \frac{1}{6}(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35}) \\
\hat{b}_5 &= \frac{1}{2}d_{45} - \frac{1}{6}(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35}) \\
\hat{b}_6 &= -\frac{1}{2}d_{12} + \frac{1}{4}(d_{13} + d_{23} - d_{34} - d_{35}) + \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25}) \\
\hat{b}_7 &= \frac{1}{4}(d_{34} + d_{35} - d_{13} - d_{23}) + \frac{1}{8}(d_{14} + d_{24} + d_{15} + d_{25}) - \frac{1}{2}d_{45}
\end{aligned} \tag{6.10}$$

Similar expressions can be obtained for any other topology such as topology B or C in Figure 6.4 or for any number of sequences (m).

In practice, however, estimation of branch lengths by Equation (6.10) is not always easy, because a large amount of computational time is required when the number of sequences is large. Rzhetsky and Nei (1993) solved this problem by developing a simple method of estimating branch lengths without using matrix algebra. Consider tree B in Figure 6.3 as an example. If we choose one particular interior branch of this tree, this tree can be drawn in the form of tree C, where A, B, C, and D each represent a cluster of sequences. For example, for the interior branch b of tree B in Figure 6.3, A, B, C, and D represent clusters (3), (1, 2), (4, 5), and (6, 7) respectively. In this case, the branch length b in tree C can be estimated by the following equation

$$\begin{aligned}\hat{b} = \frac{1}{2}\{&\gamma[d_{AC}/(m_A m_C) + d_{BD}/(m_B m_D)] \\ &+ (1-\gamma)[d_{BC}/(m_B m_C) + d_{AD}/(m_A m_D)] \\ &- d_{AB}/(m_A m_B) - d_{CD}/(m_C m_D)\}\end{aligned} \tag{6.11}$$

where

$$\gamma = (m_B m_C + m_A m_D)/[(m_A + m_B)(m_C + m_D)]$$

Here, m_A, m_B, m_C, and m_D are the numbers of sequences in clusters A, B, C, and D, respectively, and d_{AC} is the sum of pairwise distances between cluster A (sequence 3) and cluster C (sequences 4 and 5). The distances d_{BD}, d_{BC}, d_{AD}, d_{AB}, *and* d_{CD} are defined in a similar fashion. By contrast, the LS estimate of the length (b) of an exterior branch of tree D in Figure 6.3 is given by

$$\hat{b} = [d_{AB}/m_B + d_{AC}/m_C - d_{BC}/(m_A m_B)]/2 \tag{6.12}$$

where d_{AB} is the sum of all pairwise distances between sequence A (representing one exterior branch) and all sequences belonging to cluster B, d_{AC} is the sum of distances between A and all sequences belonging to cluster C, d_{BC} is the sum of all pairwise distances between sequences in clusters B and C, and m_B and m_C are the numbers of sequences in clusters B and C, respectively.

The above equations simplify the computation of branch length estimates considerably. For example, $\hat{b}_1$ in Equation (6.10) can be obtained by using Equation (6.12). In this case, the tree is given by Figure 6.4A, and the sequences in clusters A, B, and C are 1, 2, and (3, 4, 5), respectively. Therefore, $d_{AB} = d_{12}$, $d_{AC} = d_{13} + d_{14} + d_{15}$, $d_{BC} = d_{23} + d_{24} + d_{25}$, $m_B = 1$, and $m_C = 3$, and we have $\hat{b}_1 = [d_{12} + (d_{13} + d_{14} + d_{15})/3 - (d_{23} + d_{24} + d_{25})/3]/2$, which is identical to $\hat{b}_1$ in Equation (6.10). Similarly, all the other branch length estimates can be obtained by either Equation (6.11) or (6.12). Once $\hat{b}_i$'s are obtained, e_{ij}'s in Equations (6.2) and (6.3) can easily be obtained by summing the $\hat{b}_i$'s for all the branches that connect sequences i and j, and therefore R_S can be computed.

Bryant (1997), Gascuel (1997b), and Bryant and Waddell (1998) recently developed a fast algorithm for computing $\hat{b}$, using Equations (6.11) and (6.12). The readers who are interested in this algorithm should refer to the original papers. This algorithm is used in PAUP* and MEGA2.

Figure 6.2B shows a tree obtained by the Fitch-Margoliash method with LS estimates of branch lengths. The distance used is the Kimura distance. This tree now has a shorter branch for humans than that of the UPGMA tree and a longer branch for chimpanzees, but the other branches are nearly the same as those of the UPGMA tree. In the present case, Fitch and Margoliash's original algorithm gives essentially the same results.

Theoretical Basis

The LS method is a well-established statistical method of parameter estimation. When the variables are normally distributed, it is as efficient as the maximum likelihood method. In the present case, if the number of nucleotides or amino acids examined is sufficiently large, b_i is expected to follow the normal distribution (Rzhetsky and Nei 1993). Therefore, the LS method is expected to give good estimates of branch lengths (b_i's).

However, our primary interest is to determine the topology of the tree, and if this topology is incorrect, branch length estimates do not have much biological meaning. The mathematical formulation presented in this section is not intended to estimate a topology, because there is no parameter specifying topology in the formula for R_S. What is then the theoretical basis of the LS method for "estimating" the correct topology? At this moment, we do not have a good answer to this question. We can simply argue that a topology of which the estimated branch lengths are closest to the observed ones should be a good topology. Indeed, if unbiased estimates of evolutionary distances are used and the number of nucleotides or amino acids used (n) becomes infinitely large, the R_S value will be 0 only for the correct topology. Therefore, if we regard a tree-building method as a statistic as Felsenstein (1978) did, the LS method is a consistent estimator of the true topology. Computer simulations (e.g., Sourdis and Krimbas 1987; Kuhner and Felsenstein 1994) have shown that the LS method with the constraint of nonnegative branches gives reasonably good results for topology construction when the number of nucleotides used is large.

6.3. Minimum Evolution (ME) Method

Principle

In this method, the sum (S) of all branch length estimates, i.e.,

$$S = \sum_{i}^{T} \hat{b}_i \qquad (6.13)$$

is computed for all or all plausible topologies, and the topology that has the smallest S value is chosen as the best tree. Here $\hat{b}_i$ denotes an esti-

mate of the length of the i-th branch, and T is the total number of branches, that is, $2m - 3$. For example, in the case of tree A of Figure 6.4, S is given by $\hat{b}_1 + \hat{b}_2 + \cdots + \hat{b}_7$, where $\hat{b}_i$ indicates an estimate of b_i. The idea of a minimum evolution method was first put forward by Edwards and Cavalli-Sforza (1963) without giving any justification or algorithm. Later, Kidd and Sgaramella-Zonta (1971) suggested that the total branch lengths [$L(S)$] be computed by summing the absolute values ($|\hat{b}_i|$) of all branch length estimates without any theoretical justification. (In the case of allele frequency data with which Kidd and Sgaramella-Zonta were concerned, LS estimates of b_i's often become negative.) Unfortunately, $L(\hat{S})$ does not have a nice statistical property that permits the fast computation of S values, and the statistical tests as developed by Rzhetsky and Nei (1992a, 1993) are not applicable to $L(S)$. Note also that in the presence of statistical errors estimates of short branch lengths may become negative by chance even for the correct topology (Sitnikova et al. 1995).

The theoretical foundation of the ME method is Rzhetsky and Nei's (1993) mathematical proof that when unbiased estimates of evolutionary distances are used, the expected value of S becomes smallest for the true topology irrespective of the number of sequences (m). This is a good statistical property, but a topology with the smallest S is not necessarily an "unbiased estimator" of the true topology (chapter 9).

Like the LS method, the ME method is supposed to examine all possible topologies and find one that has the smallest S value. For this purpose, one may use the algorithms presented in chapter 7. However, this is very time consuming, and for this reason Rzhetsky and Nei (1992a, 1993) suggested that the neighbor joining (NJ) tree (see section 6.4) be first constructed and then a set of topologies close to the NJ tree be examined to find a tree with a smaller S value (temporary ME tree). A new set of topologies close to this temporary ME tree (excluding previously examined topologies) are now examined to find a tree with an even smaller S value. This process will be continued until no tree with a smaller S is found, and the tree with the smallest S is regarded as the ME tree. The theoretical basis of this strategy is that the ME tree is generally identical or close to the NJ tree when m is relatively small (Saitou and Imanishi 1989; Rzhetsky and Nei 1992a), and thus the NJ tree can be used as a starting tree when m is large.

One way of choosing closely related topologies is to consider all topologies that are different from the temporary ME tree by topological distances $d_T = 2$ and 4. If this is repeated many times, avoiding all topologies previously examined, one can usually obtain the ME tree or a tree close to it. We call this procedure the **close neighbor interchange** (CNI) algorithm.

Computation of S *and* D

In a previous section, we have mentioned that the LS estimates of branch lengths are given by a function of distance estimates (d_{ij}'s), that is, $\hat{\boldsymbol{b}} = \boldsymbol{Ld}$. Therefore, S can be expressed as a linear function of d_{ij} or d_i, where d_{ij}'s are renumbered as d_i for $i = 1, 2, \ldots, m(m-1)/2$. That is,

$$S = \boldsymbol{yd} = \sum_{i=1}^{r} y_i d_i \tag{6.14}$$

where $r = m(m-1)/2$ (Rzhetsky and Nei 1992a). The coefficients y_i's are determined solely by the tree topology, and they can be computed if the topology matrix A in Equation (6.6) is defined. For example, in the case of tree A in Figure 6.4, S becomes

$$\begin{aligned} S_A = {} & d_{12}/2 + d_{13}/4 + d_{14}/8 + d_{15}/8 + d_{23}/4 \\ & + d_{24}/8 + d_{25}/8 + d_{34}/4 + d_{35}/4 + d_{45}/2 \end{aligned} \tag{6.15}$$

Similarly, for tree B in Figure 6.4 we have

$$\begin{aligned} S_B = {} & d_{12}/4 + d_{13}/2 + d_{14}/8 + d_{15}/8 + d_{23}/4 \\ & + d_{24}/4 + d_{25}/4 + d_{34}/8 + d_{35}/8 + d_{45}/2 \end{aligned} \tag{6.16}$$

However, we are primarily interested in the difference in S between two topologies. This difference (D) is given by

$$D = S_B - S_A = \sum_{i=1}^{r} (y_{Bi} - y_{Ai}) d_i \tag{6.17}$$

where y_{Ai} and y_{Bi} are the coefficients of the i-th distance in S for topologies A and B, respectively. Therefore, if y_{Ai}'s and y_{Bi}'s are computed for a pair of topologies, D can easily be obtained. For this purpose, it is not necessary to know individual S values. In the case of trees A and B in Figure 6.4, we know y_{Ai}'s and y_{Bi}'s, so that D is given by

$$D = -d_{12}/4 + d_{13}/4 + d_{24}/8 + d_{25}/8 - d_{34}/8 - d_{35}/8 \tag{6.18}$$

In practice, D may be subject to sampling error, and we are interested in testing the null hypothesis that the expected value [E(D)] of D is 0 for a given type of nucleotide or amino acid substitution. If D is significantly greater than 0, we may conclude that tree A is better than tree B. However, what is the biological meaning of this null hypothesis when two different topologies are compared? Actually, this hypothesis is equivalent to the null hypothesis that the lengths of the interior branches that produce different branching patterns (different partitions of sequences) for the two topologies are 0. Only in this case do trees A and B become identical. The tree corresponding to this null hypothesis is called the **null tree.** For example, in the comparison of trees A and B in Figure 6.4, the null tree is given by tree D, where $b = 0$. Therefore, the test of $E(D) = 0$ for the trees A and B is equivalent to testing the null hypothesis of $b_6 = 0$. Indeed, we can show that

$$D = S_B - S_A = b_6/8 - (2\epsilon_{12} - 2\epsilon_{13} - \epsilon_{24} - \epsilon_{25} + \epsilon_{34} + \epsilon_{35})/8 \tag{6.19}$$

Therefore, when $b_6 = 0$, the expectation of D is 0. In practice, we do not know which of the trees A and B is the correct one. So, D can be positive or negative. Similarly, $D = S_C - S_A$ can be written as

$$D = S_C - S_A = 3(b_6 + b_7)/4 - 3(\epsilon_{12} - \epsilon_{14} - \epsilon_{25} + \epsilon_{45})/8 \quad (6.20)$$

This indicates that we are testing the null hypothesis that both b_6 and b_7 in tree A are 0 and that the null tree for this null hypothesis is tree E in Figure 6.4. This principle applies to any pair of bifurcating trees, irrespective of the number of sequences.

To test the null hypothesis of $E(D) = 0$, we have to know the standard error of D. Rzhetsky and Nei (1992a, 1993) developed a simple algorithm to compute the standard error of D for several substitution models. As long as the number of sequences used is relatively small (say, $m \leq 50$), their method is easily applicable. However, if m is large, it requires a substantial amount of computer time. Another way of testing is to use a bootstrap method (Nei 1991). In this bootstrap method, S is computed for a given pair of topologies (i and j) for each set of resampled sequences (see chapter 9), and $D_{ij} = S_i - S_j$ is computed. If this is repeated many times, we can compute the standard error of D. Therefore, we can test the null hypothesis of $E(D) = 0$ by the Z test given in Equation (4.5). When there are several potentially correct trees, D_{ij} can be computed for all pairs of i and j using the same set of resampled sequences.

Example 6.2. ME Trees for Hominoid Species

In sections 6.1 and 6.2, we constructed the UPGMA and the Fitch-Margoliash trees for five hominoid species using Kimura distances. Let us now construct the ME tree using the same set of pairwise distances (Table 6.1). In the present case, there are only 15 possible topologies, so it is easy to identify the ME tree. The tree obtained is presented in Figure 6.2C. The topology and the branch lengths of the tree are virtually identical with those of the FM tree. We also constructed the ME trees using the p distance, Jukes-Cantor distance, and Kimura gamma distance with $a = 0.53$, but these trees had the same topology, and their branch lengths were similar to those obtained with the Kimura distance.

To see the differences in the S value and branch length estimates between different topologies, let us consider the three possible trees for humans (H), chimpanzees (C), orangutans (O), and gibbons (B) given in Figure 6.5. The S value and the branch length estimates given in these trees were obtained by using Jukes-Cantor distance. Topology I, in which humans and chimpanzees make a cluster, has the smallest S value, and the S values for the other topologies are significantly greater than that for topology I. The bootstrap value (100%) also supports this topology. Therefore, topology I is the most likely tree.

Theoretically, it can be shown that in the absence of sampling error the interior branch of the correct topology for four sequences are always nonnegative, whereas that of an incorrect topology is negative (Rzhetsky and Nei 1992a; Sitnikova et al. 1995). In the present case, the interior branch is positive in topology I but is negative in topologies II and III. These results also support topology I. Incidentally, Figure 6.5 includes the MP and ML trees, which will be discussed later. In these trees, all interior branches are nonnegative, so that the positiveness of interior branches cannot be used for distinguishing between the correct and incorrect

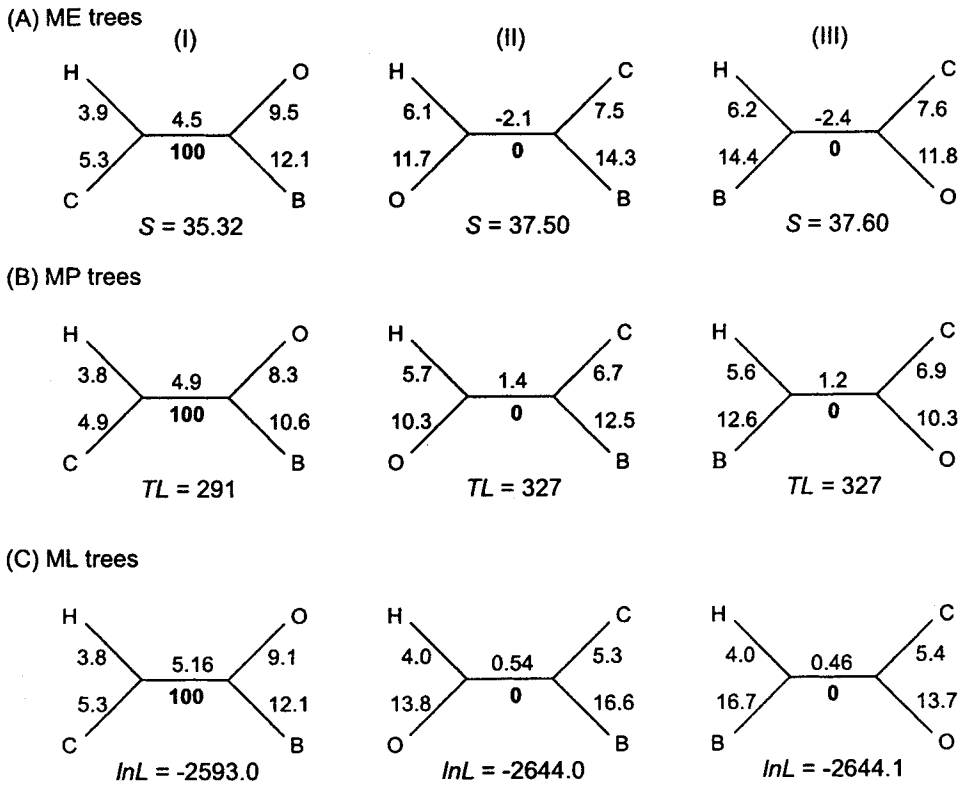


FIGURE 6.5. Estimates of branch lengths obtained by the ME, MP, and ML methods for a tree of humans (H), chimpanzees (C), orangutans (O), and gibbons (B). The Jukes-Cantor model was used in all calculations to make fair comparisons of ME and ML trees with MP trees. The bootstrap values for each case are shown below the interior branch. All branch lengths are in units of the number of substitutions per 100 sites.

topologies, although the interior branch of incorrect topologies tends to be smaller than that of the correct topology. However, the bootstrap test gives essentially the same conclusion as that for the ME tree.

6.4. Neighbor Joining (NJ) Method

Although the ME method has nice statistical properties, it requires a substantial amount of computer time when the number of taxa compared is large. Saitou and Nei (1987) developed an efficient tree-building method that is based on the minimum evolution principle. This method does not examine all possible topologies, but at each stage of taxon clustering a minimum evolution principle is used. This method is called the **neighbor joining (NJ) method** and is regarded as a simplified version of the ME method. When four or five taxa are used, the NJ and ME methods give identical results (Saitou and Nei 1987). There is some similarity between NJ and Sattath and Tversky's (1977) additive tree method (see also Fitch

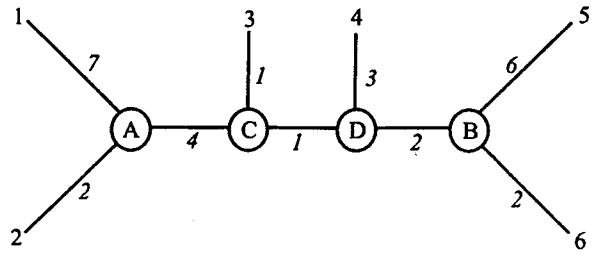


FIGURE 6.6. A phylogeny of six sequences with known branch lengths.

1981), but the former gives both the topology and branch lengths simultaneously.

One of the important concepts in the NJ method is **neighbors,** which are defined as two taxa that are connected by a single node in an unrooted tree. For example, taxa 1 and 2 in the tree of Figure 6.6 are neighbors, because they are connected by the only node A. Similarly, taxa 5 and 6 are neighbors, but all other pairs of taxa are not. However, if we combine taxa 1 and 2 and regard them as a single taxon, the combined taxon (1-2) and taxon 3 are now neighbors. It is possible to define the topology of a tree by successively joining neighbors and producing new pairs of neighbors. For example, the topology of the tree of Figure 6.6 can be described by the following pairs of neighbors: (1, 2), (5, 6), (1-2, 3), and (1-2-3, 4). Therefore, by finding these pairs of neighbors, one can obtain the tree topology.

Algorithm

Construction of a tree by the NJ method begins with a star tree, which is produced under the assumption that there is no clustering of taxa (Figure 6.7A). In practice, this assumption is generally incorrect. Therefore, if we estimate the branch lengths of the star tree and compute the sum of all branches (S_0), this sum should be greater than the sum (S_F) for the true or the final NJ tree. However, if we pick up neighbors 1 and 2 and consider the tree presented in Figure 6.7B, the sum (S_{12}) of all branch lengths should be smaller than S_0, although it may be greater than S_F. In practice, since we do not know which pair of taxa are true neighbors, we consider all pairs of taxa as a potential pair of neighbors and compute the sum of branch lengths (S_{ij}) for the i-th and j-th taxa using a topology similar to that given in Figure 6.7B. We then choose the taxa i and j that show the smallest S_{ij} value. Of course, actual distance values are subject to stochastic errors, so that the neighbors chosen in this way may not always be the true neighbors. Once a pair of neighbors are identified, they are combined into one composite taxon, and this procedure is repeated until the final tree is produced.

Mathematically, S_0 for the star tree is given by

$$S_0 = \sum_{i=1}^{m} L_{iX} = \frac{1}{m-1} \sum_{i<j}^{m} d_{ij}$$
$$= T/(m-1) \qquad (6.21)$$

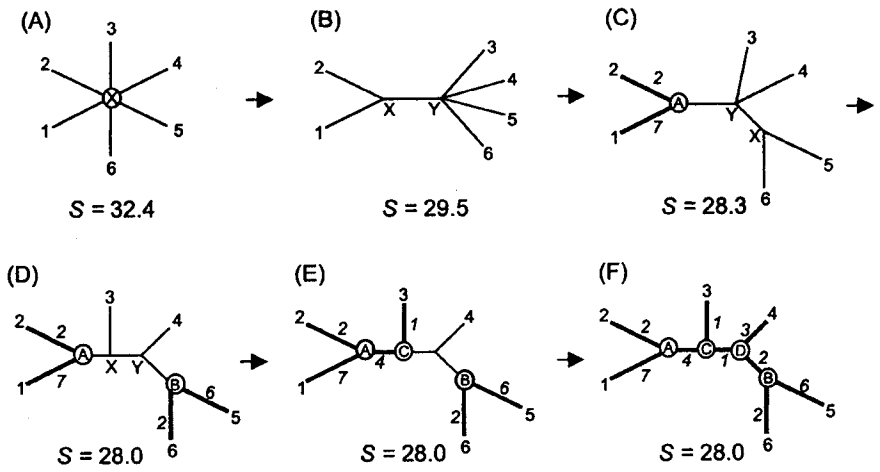


FIGURE 6.7. Illustration of the computational process in the neighbor-joining method.

where L_{iX} is the branch length estimate between nodes i and X, and $T = \Sigma_{i<j} d_{ij}$. In the present case, i stands for the i-th exterior node and X the interior node (Figure 6.7A). By contrast, Figure 6.7B indicates that S_{12} is given by the sum of $L_{1X} + L_{2X}$, L_{XY}, and $\Sigma_{i=3}^{m} L_{iY}$. Here, $L_{1X} + L_{2X} = d_{12}$, and

$$L_{XY} = \frac{1}{2(m-2)}\left[\sum_{i=3}^{m}(d_{1i} + d_{2i}) - (m-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{m} d_{iY}\right]$$

$$\sum_{i=3}^{m} L_{iY} = \frac{1}{m-3} \sum_{3\le i<j} d_{ij}$$

Therefore, we have

$$\begin{aligned} S_{12} &= L_{1X} + L_{2X} + L_{XY} + \sum_{i=3}^{m} L_{iY} \\ &= \frac{1}{2(m-2)} \sum_{i=3}^{m}(d_{1i} + d_{2i}) + \frac{1}{2} d_{12} + \frac{1}{m-2} \sum_{3\le i<j} d_{ij} \end{aligned} \quad (6.22)$$

If we write $R_1 = \Sigma_{i=1}^{m} d_{1i}$ and $R_2 = \Sigma_{i=1}^{m} d_{2i}$, S_{12} can also be expressed as

$$S_{12} = \frac{2T - R_1 - R_2}{2(m-2)} + \frac{d_{12}}{2} \quad (6.23)$$

Obviously, S_{ij} can be computed in the same way if we replace 1 and 2 by i and j, respectively, in the above equations. Equation (6.23) requires less computational time than Equation (6.22). Furthermore, since T is the same for all pairs of i and j, S_{ij} can be replaced by

$$Q_{ij} = (m - 2)d_{ij} - R_i - R_j \tag{6.24}$$

for the purpose of computing the relative value of S_{ij} (Studier and Keppler 1988). Equation (6.24) is used for computer programming to facilitate the computation.

Once the smallest S_{ij} is determined, we can create a new node (A) that connects taxa i and j. The branch lengths (b_{Ai} and b_{Aj}) from this node to taxon i and taxon j are given by

$$b_{Ai} = \frac{1}{2(m-2)}\left[(m-2)d_{ij} + R_i - R_j\right] \tag{6.25a}$$

$$b_{Aj} = \frac{1}{2(m-2)}\left[(m-2)d_{ij} - R_i + R_j\right] \tag{6.25b}$$

(Saitou and Nei 1987; Studier and Keppler 1988). These values are known to be LS estimates for the topology under consideration (Saitou and Nei 1987). The next step is to compute the distance between the new node (A) and the remaining taxa (k; $3 \leq k \leq m$) (Figure 6.7C). This distance is given by

$$d_{Ak} = (d_{ik} + d_{jk} - d_{ij})/2 \tag{6.26}$$

If we compute all the distances using this equation, we have a new $(m - 1) \times (m - 1)$ matrix. From this matrix, we can compute a new S_{ij} matrix using Equation (6.23). However, we denote this new S_{ij} by S_{ij}', because this new S_{ij} does not include the lengths of exterior branches for the first pair of neighbors identified, and thus it is shorter than the real total sum (S_{ij}) of branch lengths at this stage of tree construction. To find the new pair of "neighbors," we choose a pair with the smallest S_{ij}' value. A new node B is then created for this pair of taxa, and a new $(m - 2) \times (m - 2)$ distance matrix is computed by using Equation (6.26). This procedure is repeated until all taxa are clustered in a single unrooted tree. The final tree obtained in this way is the NJ tree.

If one is interested in the reduction in S_{ij} in each cycle of neighbor joining, S_{ij} can be obtained by adding the lengths of all branches eliminated to S_{ij}'. This process of reduction in S_{ij} (represented by S) is shown in Figure 6.7, but in actual practice S_{ij} is rarely computed. In fact, most computer programs use Q_{ij} rather than S_{ij} or S_{ij}'.

To illustrate the computational procedure, let us consider the evolutionary distances given in cycle 1 of Table 6.2. These distances were obtained by adding the branch lengths for each pair of taxa of the tree in Figure 6.6. Therefore, all the distances satisfy the condition of additivity. The total sum of distances is $T = 162$. Therefore, S_0 for the star tree (Figure 6.7A) is 32.4 from Equation (6.21), since $m = 6$ in this case. We now compute S_{ij}'s for all pairs of i and j for topology B in Figure 6.7. For taxa 1 and 2, we have $d_{12} = 9$, $R_1 = 72$, and $R_2 = 52$, so S_{12} is 29.5 from Equation (6.23). Similarly, we compute S_{ij}'s for all other pairs of taxa, and they are shown in cycle 1 of Table 6.2. This table shows that the small-

Table 6.2 Distance and S_{ij}' matrices at sequential steps of the NJ algorithm.

	Distance Matrix					S_{ij} or S_{ij}' Matrix				
Cycle 1										
	1	2	3	4	5	1	2	3	4	5
1										
2	9					**29.5**				
3	12	7				32.5	32.5			
4	15	10	5			33.0	33.0	32.0		
5	20	15	10	11		33.5	33.5	32.5	32.0	
6	16	11	6	7	8	33.5	33.5	32.5	32.0	30.5
Selected Pair (1, 2) with branch lengths (7, 2); A = (1, 2)										
Cycle 2										
	A	3	4	5		A	3	4	5	
A										
3	5					19.7				
4	8	5				20.3	20.3			
5	13	10	11			21.0	21.0	20.7		
6	9	6	7	8		21.0	21.0	20.7	**19.3**	
Selected Pair (5, 6) with branch lengths (6, 2); B = (5, 6)										
Cycle 3										
	A	3	4			A	3	4		
A										
3	5					**11.0**				
4	8	5				11.5	11.5			
B	7	4	5			11.5	11.5	11.0		
Selected Pair (a, 3) with branch lengths (4, 1); C = (A, 3)										

est S_{ij} is S_{12} = 29.5. Thus, we infer that taxa 1 and 2 are neighbors. The fact that these two taxa are indeed a pair of neighbors is seen in Figure 6.6. The branch lengths of taxa 1 and 2 from the new node A in Figure 6.7 can be obtained by Equations (6.25) and becomes 7 and 2, respectively. These branch lengths are also identical with the true values of the tree in Figure 6.6. We now compute the distance between the new node A and taxon k using Equation (6.26). In the next step of neighbor joining (cycle 2 in Table 6.2), taxa 5 and 6 are found to be a pair of neighbors, because S_{56}' = 19.3 is the smallest S_{ij}' value. Therefore, we create a new node B and compute b_{5B} and b_{6B}. They are 6 and 2, respectively, which are again identical with those of the true tree (Figure 6.7). In cycle 3, taxa A and 3 show the smallest S_{ij}' value (= 11.0). Therefore, we now create a new node C. It is then obvious that node B and taxon 4 form a cluster. This creates another node D and completes the entire process of neighbor joining. We can now estimate branch lengths b_{4D}, b_{CD}, and b_{BD} in Figure 6.7F, and they become 3, 1, and 2, respectively. The final tree obtained is shown in Figure 6.7F. Both the topology and branch lengths of this tree are identical with those of the true tree in Figure 6.6.

However, this complete recovery of the original tree occurred because we used additive distances without any backward and parallel mutations. In real data, there are almost always backward and parallel mutations in some sequences, so it is not always easy to reconstruct the true tree. Therefore, it is important to conduct some statistical tests about the reliability of the tree obtained.

Using the sequence data in Figure 6.1, we produced two NJ trees for hominoid species using Kimura and p distances. Tree C in Figure 6.2 represents the NJ tree with Kimura distance, whereas tree D is the NJ tree with p distance. In these trees, the branch lengths were estimated by the ordinary LS method after the topology was determined. Note that the topology of those trees is identical with that of the ME tree obtained for the same data set.

Justification and Modifications

As mentioned above, the NJ method is based on the principle of minimum evolution but generates only one final topology with branch length estimates. Some authors criticized this method for this reason and suggested that the ME method rather than this method be used. Actually, it is possible to modify the NJ method to generate more topologies. Kumar (1996b) developed an algorithm in which not only the minimum S_{ij} but also several S_{ij}'s close to the minimum are considered as indicators of potential neighbors in each cycle of S_{ij} computation. This method generates as many topologies as desired and allows us to compare S values for different topologies. Therefore, one can choose the topology that shows the smallest S value. This is a hybrid method between the ME and NJ methods. A similar method has been proposed by Pearson et al. (1999).

As shown by Rzhetsky and Nei (1993), the ME method is expected to give the correct topology if the number of nucleotides examined (n) is sufficiently large and an unbiased estimate of nucleotide substitutions is used as a distance measure. When n is small and m is large, however, the S value (S_m) of the ME tree tends to be smaller than that (S_c) of the correct tree because of sampling errors. In fact, Nei et al. (1998) have shown that S_m is always equal to or smaller than S_c and that the probability of occurrence of $S_m < S_c$ is quite high when n is small (chapter 9). This indicates that it is not rewarding to spend excessive time to find the true ME tree when n is relatively small, because the true ME tree tends to be incorrect. Computer simulations by Saitou and Imanishi (1989), Rzhetsky and Nei (1992a), Gascuel (1997a, 1997b), and Nei et al. (1998) have also shown that the probability of obtaining the correct topology is nearly the same for both the ME and NJ methods. In other words, NJ is a fast method of constructing phylogenetic trees and is appropriate for analyzing a large data set. It is also capable of conducting bootstrap tests rapidly.

Backeljau et al. (1996) stated that NJ may produce two or more tie trees for the same data set. According to Takezaki (1998), NJ tie trees occur very rarely when the computation is done with high precision (much less than MP tie trees). Furthermore, even if multiple tie trees occur, they do not

pose any serious problem if a bootstrap consensus tree is produced. Therefore, we do not have to worry about them.

Gascuel (1997a) proposed the so-called **BIONJ method** to improve the efficiency of NJ in obtaining the correct topology. The computational algorithm is the same as that of NJ except that different weights are given to d_{ik}, d_{jk}, and d_{ij} in Equation (6.26) to minimize the variance of d_{Ak}. Using computer simulation, he showed that BIONJ is slightly better than NJ when sequence divergence is high. However, our limited experience with actual data analysis has shown that the two methods almost always give the same or very similar trees.

Example 6.3. NJ, ME, and BIONJ Trees for Simulated Sequence Data

To obtain some idea about the accuracy of NJ and BIONJ trees, let us consider the results of a small computer simulation presented in Figure 6.8. With real data, it is usually very difficult to know the true tree, so that it is virtually impossible to compare the reconstructed tree with the true tree. In a computer simulation, we can use a model tree and let a set of DNA sequences evolve following the model tree with a given pattern of nucleotide substitution. We can then reconstruct a tree using the DNA sequences generated and compare the reconstructed tree with the true tree.

The model and the realized trees for 24 DNA sequences used in the present simulation are shown in trees A and B of Figure 6.8, respectively. The number of substitutions in each branch of the realized tree were obtained by using pseudorandom numbers under the assumption that nucleotide substitution occurs following Kimura's model with a transition/transversion ratio (R) of 5 (see Saitou and Nei [1987] for the detail of the simulation). The number of nucleotides per sequence used in this simulation was 500. In both trees A and B, the branch lengths are measured in terms of the number of nucleotide substitutions per site. Note that the branch lengths of the realized tree are much more variable than those of the model tree because of stochastic errors. The topology of the realized tree is identical with that of tree A except for one trifurcating node that occurred because one interior branch corresponding to branch a of the model tree did not have any nucleotide substitution.

Figure 6.8C shows the NJ tree obtained by using the Jukes-Cantor distance for the 24 sequences that were generated by computer simulation. Comparison of this tree with the realized tree shows that tree C has one topological error. That is, the trifurcating node for sequences 8, 9, and 10 in the realized tree is decomposed into two consecutive bifurcating nodes in tree C, though the branch length between the two nodes is very close to 0. This occurred because NJ is designed to construct a bifurcating tree. Except for this minor difference, the topology of this tree is identical with that of the realized tree. The branch length estimates are also very close to those of tree B. Note that here we used the Jukes-Cantor distance instead of the Kimura distance, which is more appropriate in the present case. Yet, the reconstructed tree is very close to the true realized tree. When we used the Kimura distance, we obtained a tree that was very

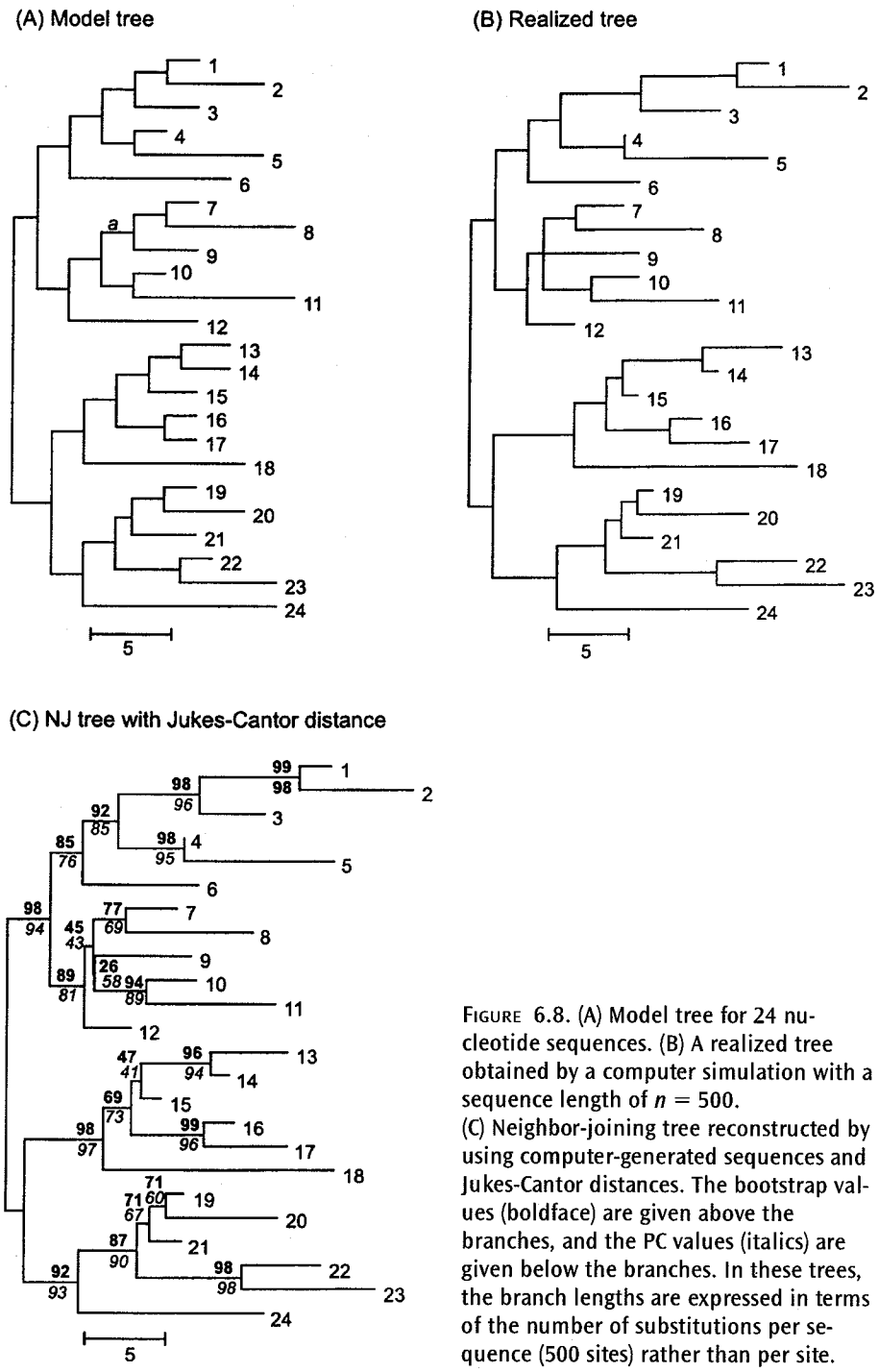


FIGURE 6.8. (A) Model tree for 24 nucleotide sequences. (B) A realized tree obtained by a computer simulation with a sequence length of $n = 500$.
(C) Neighbor-joining tree reconstructed by using computer-generated sequences and Jukes-Cantor distances. The bootstrap values (boldface) are given above the branches, and the PC values (italics) are given below the branches. In these trees, the branch lengths are expressed in terms of the number of substitutions per sequence (500 sites) rather than per site.

similar to tree C, though the Kimura distance tree had slightly longer branch lengths near the root as expected. The similarity of the two reconstructed trees is of course due to the fact that the extent of sequence divergence is low in the present case. We also constructed the ME and BIONJ trees, but they had essentially the same topology and branch lengths as those of tree C.

In the above example, the topological error of a reconstructed tree occurred because one interior branch of the realized tree had no nucleotide substitution, as mentioned above. In fact, zero-length interior branches in realized trees are a source of topological errors in reconstructed trees, particularly when there are many such branches. Unfortunately, we usually do not know such interior branches in real data, and therefore it is difficult to evaluate the effect of this factor, though parsimony methods are capable of identifying such branches under certain conditions (chapter 7). However, such interior branches almost always give low bootstrap values, whether or not the branch pattern obtained is correct. In fact, the interior branch of the NJ trees associated with the zero-length branch in Figure 6.8 has a bootstrap value of 26%. Therefore, if we disregard low-bootstrap interior branches, we can conclude that the NJ, ME, and BIONJ methods reconstruct the true tree quite accurately.

However, this happened partly because the pattern of nucleotide substitution used was relatively simple. In most DNA sequences, the actual substitution pattern is much more complicated than the Kimura model, and this would introduce topological errors even when d's are only moderately large. For this reason, a number of statistical tests of the reliability of an inferred tree have been developed. This problem will be discussed in chapter 9.

6.5. Distance Measures to Be Used for Phylogenetic Reconstruction

In chapters 2–4, we discussed various distance measures for estimating the number of nucleotide or amino acid substitutions (d) considering different mathematical models. In general, a distance measure based on a complex mathematical model requires many parameters to be estimated, and this increases the variance of the estimate of d. Theoretically, it is possible to choose a mathematical model most appropriate for a given set of data using certain statistical criteria. Several such statistical methods are now available (Kishino and Hasegawa 1989; Bulmer 1991; Goldman 1993; Rzhetsky and Nei 1995; Yang 1995a), but in these methods the increment of variance by adding more parameters are not considered. Therefore, the best distance measure identified by these criteria is not necessarily most appropriate for reconstruction of phylogenetic trees, although they are usually useful for branch length estimation.

Generally speaking, the accuracy of an inferred tree depends on at least two factors: (1) the linear relationship of the distance used with the number of substitutions and (2) the standard error or the coefficient of variation of the estimate of the distance measure. For Kimura's (1980) model of nucleotide substitution, several authors have attempted to produce

better distance measures than the original estimator, taking into account these two factors (Schöniger and von Haeseler 1993; Goldstein and Pollock 1994; Tajima and Takezaki 1994), but the practical utility of these distance measures is still unclear.

At the present time, there is no general statistical method for choosing an appropriate distance measure (or mathematical model) for constructing tree topologies. However, computer simulations and empirical studies have led to the following guidelines for the purpose of topology construction (modified from Nei 1996).

1. When the Jukes-Cantor estimate of the number of nucleotide substitutions per site (d) is about 0.05 or less ($d \leq 0.05$), use the p or Jukes-Cantor distance whether there is a transition/transversion bias or not or whether the substitution rate (r) varies with nucleotide site or not. In this case, the Kimura distance and more complicated distance measures give essentially the same value as the p or Jukes-Cantor distance (Figure 3.1), but their variances are greater than those of the latter distances. The p distance tends to give good results, particularly when the number of nucleotides or amino acids used is small.

2. When $0.05 < d < 1.0$ and the number of nucleotides examined is large, use the Jukes-Cantor distance unless the transition/transversion ratio (R) is high, say, $R > 5$. When this ratio is high and the number of nucleotides examined (n) is very large, use the Kimura distance or the gamma distance. However, when the number of sequence is large and n is relatively small, the p distance often gives better results unless the evolutionary rate varies extensively with evolutionary lineage (Takahashi and Nei 2000). In recent years, a number of authors have used maximum likelihood estimates of the HKY gamma distance, apparently because in theory this distance takes care of the GC content and transition/transversion biases as well as the variation in substitution rate among different sites (e.g., Honda et al. 1999). However, computer simulations with 48 nucleotide sequences have shown that with most reasonable model trees this distance generally gives a poorer performance than the p or the Jukes-Cantor distance even if the HKY gamma model is used for generating sequence data (Takahashi and Nei 2000). This is because the HKY gamma distance has a large variance compared with the p or the Jukes-Cantor distance. When the number of nucleotides examined is very large (>10,000) and the rate of nucleotide substitution varies extensively with evolutionary lineage, a complicated distance measure (e.g., HKY gamma distance) may give better results (Takezaki and Gojobori 1999).

3. When $d > 1$ for many pairs of sequences, the phylogenetic tree constructed is generally unreliable for a number of reasons (e.g., large variances of $\hat{d}$'s and sequence alignment errors). We therefore suggest that these data sets should be avoided as much as possible. In this case, one may eliminate the portion(s) of the gene that evolves very fast and use only the remaining region(s) as is often done with immunoglobulin variable region genes (Ota and Nei 1994a; Rast et al. 1994). One may also use a different gene that evolves more slowly.

4. Many distance measures for estimating the number of nucleotide substitutions per site (d) often becomes inapplicable when the distance is very large and n is small. This happens because the mathematical for-

mulas for distance estimation usually involve logarithmic terms, and the arguments of the logarithms often become negative. Theoretically, this problem can be avoided by expanding the logarithmic terms into an infinite series, but the variance of the distance estimated in this way is quite large (Tajima 1993b; Rzhetsky and Nei 1994). Therefore, highly divergent sequences should not be used for topology construction. In this case the p distance is often more efficient for obtaining a reliable topology, because it is always applicable and has a smaller variance.

5. When a phylogenetic tree is constructed from the coding regions of a gene, the distinction between synonymous (d_S) and nonsynonymous (d_N) substitutions may be helpful, because the rate of synonymous substitutions is usually much higher than that of nonsynonymous substitution. When relatively closely related species are studied for a large number of codons and $d_S < 0.5$, one may use d_S for constructing a tree. This procedure is expected to reduce the effect of variation in substitution rate among different sites, because synonymous substitutions are subject to selection less often than nonsynonymous substitutions. However, when relatively distantly related species are studied, d_N or amino acid distances seem to be better. Note also that d_S or p_S sometimes reaches the saturation level rather quickly (chapter 5).

6. As a general rule, if two distance measures give similar distance values for a set of data, use the simpler one because it has a smaller variance. When the rate of nucleotide substitution is nearly the same for all evolutionary lineages and there is no strong transition/transversion bias, the p distance seems to give correct trees more often than other distances, even if sequence divergence is high (Schöniger and von Haeseler 1993; Tajima and Takezaki 1994; Takahashi and Nei 2000). When the substitution rate varies with evolutionary lineage, however, this may not be the case. It is important not to trust computer outputs of tree construction without scrutinizing the pattern of nucleotide or amino acid substitution, differences in nucleotide frequencies among the first, second, and third codon positions, temporal changes of nucleotide frequencies, and so forth. In real data analysis, there are so many unknown factors that the phylogenetic tree produced should be interpreted with caution and common sense.

Note that the above guidelines are for constructing phylogenetic trees. For estimating branch lengths or evolutionary times, unbiased estimators are generally better than biased estimators.

7

Phylogenetic Inference: Maximum Parsimony Methods

Maximum parsimony (MP) methods were originally developed for morphological characters (Henning 1966), and there are many different versions (Wiley 1981; Felsenstein 1982; Wiley et al. 1991; Maddison and Maddison 1992; Swofford and Begle 1993). In this book, we consider only the methods that are useful for analyzing molecular data. Eck and Dayhoff (1966) seem to be the first to use an MP method for constructing trees from amino acid sequence data. Later, Fitch (1971) and Hartigan (1973) developed a more rigorous MP algorithm for nucleotide sequence data. In these MP methods, four or more aligned nucleotide (or amino acid) sequences ($m \geq 4$) are considered, and the nucleotides (amino acids) of ancestral taxa are inferred separately at each site for a given topology under the assumption that mutational changes occur in all directions among the four nucleotides (or 20 amino acids). The smallest number of nucleotide (or amino acid) substitutions that explain the entire evolutionary process for the topology is then computed. This computation is done for all potentially correct topologies, and the topology that requires the smallest number of substitutions is chosen to be the best tree. The theoretical basis of this method is William of Ockham's philosophical idea that the best hypothesis to explain a process is the one that requires the smallest number of assumptions. Sober (1988) states that the less we need to know about the evolutionary process to make a phylogenetic inference, the more confidence we can have in our conclusions. In this chapter, we are primarily concerned with nucleotide sequences, but the same approach can be used for amino acid sequences as well.

If there are no backward and no parallel substitutions (no homoplasy) at each nucleotide site and the number of nucleotides examined (n) is very large, MP methods are expected to produce the correct (realized) tree. In practice, however, nucleotide sequences are often subject to backward and parallel substitutions, and n is rather small. In this case, MP methods tend to give incorrect topologies (chapter 9). Furthermore, Felsenstein (1978) has shown that when the rate of nucleotide substitution varies extensively with evolutionary lineage, MP methods may generate incorrect topologies even if an infinite number of nucleotides are examined. Under certain conditions, this can happen even when the rate of substitution is constant for all lineages (Hendy and Penny 1989;

Zharkikh and Li 1993; Takezaki and Nei 1994; Kim 1996). In this case, long branches (or short branches) of the true tree tend to join together or attract each other in the reconstructed tree (chapter 9). Therefore, this phenomenon is often called **long-branch attraction** (Hendy and Penny 1989) or **short-branch attraction** (Nei 1996). In parsimony analysis, it is also difficult to treat the phylogenetic inference in a statistical framework, because there is no natural way to compute the means and variances of the minimum numbers of substitutions obtained by the parsimony criterion.

Nevertheless, MP methods have some advantages over other tree-building methods. First, they are relatively free from various assumptions that are required for nucleotide or amino acid substitution in distance or likelihood methods. Since any mathematical model currently used is a crude approximation to reality, model-free MP methods may give more reliable trees than other methods when the extent of sequence divergence is low (Miyamoto and Cracraft 1991). In fact, computer simulation has shown that when (1) the extent of sequence divergence is low ($d \leq 0.1$), (2) the rate of nucleotide substitution is more or less constant, and (3) the number of nucleotides examined is large, MP methods are often better than distance methods in obtaining the true topology (Sourdis and Nei 1988; Nei 1991). Furthermore, parsimony analysis is very useful for some types of molecular data such as insertion sequences and insertions/deletions, as will be discussed later.

There are many different versions of MP methods even just for molecular data, but they can be divided into **unweighted MP** and **weighted MP** methods. In unweighted MP methods, nucleotide or amino acid substitutions are assumed to occur in all directions with equal or nearly equal probability. In reality, however, certain substitutions (e.g., transitional changes) occur more often than other substitutions (e.g., transversional changes). It is therefore reasonable to give different weights to different types of substitutions when the minimum number of substitutions for a given topology is to be computed. MP methods incorporating this feature are weighted MP methods. In the following, we first consider unweighted MP methods.

7.1. Finding Maximum Parsimony (MP) Trees

Estimation of the Minimum Number of Substitutions

Let us now explain how to count or estimate the minimum number of substitutions for a given topology. We consider the topology of a rooted tree for six DNA sequences (1, 2, . . . , 6) given in Figure 7.1A and assume that the nucleotides at a given site for the six extant sequences are as given at the exterior nodes of the tree. There are one C, three T's and two A's. From these nucleotides, we can infer the nucleotides for the five ancestral taxa (nodes) *a, b, c, d,* and *e*. The nucleotide at node *a* must be either C or T if we consider the minimum possible number of substitutions. The nucleotide at node *b* is inferred to be T, whereas the nucleotide at

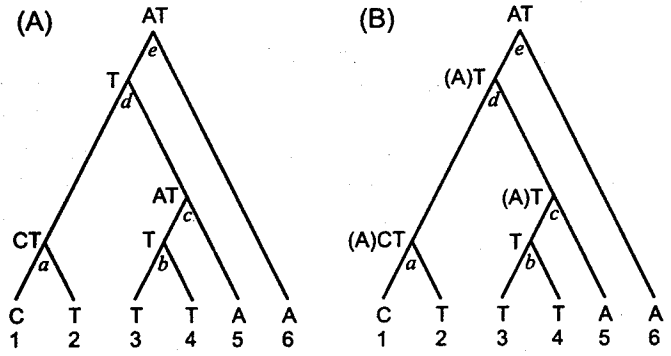


FIGURE 7.1. Nucleotides in six extant sequences and the possible nucleotides in five ancestral sequences.

node *c* must be A or T. Node *d* is expected to have T, because its immediate descendant nodes (*a* and *c*) both have T. Finally, we infer the nucleotide at node *e* to be either A or T. It is now clear that the minimum number of nucleotide substitutions for this set of taxa can be obtained by assuming that all the ancestral nodes had nucleotide T. The number is three. However, this set of nucleotides at the ancestral nodes (pathway) is not the only possible set that explains the evolutionary change of nucleotides.

If we assume that nodes *a, c, d,* and *e* all have A and node *b* has T, the number of substitutions required is again three (see Figure 7.1B). Actually, there are three more pathways that are possible with the same minimum number of substitutions. They are: (a − T, b − T, c − A, d − A, e − A), (a − C, b − T, c − A, d − A, e − A), and (a − T, b − T, c − T, d − T, e − A). These results show that the nucleotides at the ancestral nodes cannot always be determined uniquely, and all the nucleotides listed in Figure 7.1B are parsimonious ones. However, it is possible to count the minimum number of substitutions required. It is three for all of the above cases.

In the above example, we considered a rooted tree. However, the tree can be transformed into an unrooted tree by eliminating the apex node *e*. Elimination of this node does not change the minimum number of substitutions, but the number of possible pathways is reduced. For example, the two possibilities (e − T, d − T, c − T, b − T, a − T) and (e − A, d − T, c − T, b − T, a − T) are no longer distinguishable, because node *e* can be either T or A. In the present case, the total number of pathways for the unrooted tree is four. Because MP methods do not determine the root of the tree, unrooted trees are usually considered.

In the above example, the minimum number of substitutions was three, and there were four equally parsimonious pathways for the unrooted tree. Computation of these numbers was relatively easy in this case, but as the number of taxa increases, it becomes increasingly cumbersome. Therefore, all these computations are done by a computer using the above rule. The basic algorithm for these computations was developed by Fitch (1971) and Hartigan (1973).

Tree Lengths

In the above example, we considered only one topology, but in practice we have to consider all potentially correct topologies and determine the topology that requires the smallest minimum number of substitutions. Let us now consider this problem using the trees given in Figure 7.2 and assuming that taxa 1, 2, and 3 all have nucleotide A but taxa 4, 5, and 6 have G, G, and T, respectively. The trees given in the figure all consist of six taxa, but the topologies are not necessarily the same. We again consider a particular nucleotide site and compute the minimum number of substitutions required. In topology A, this number is obviously two. Topology B, in which taxa 3 and 4 are interchanged, requires at least three substitutions. Of course, there are several equally parsimonious pathways, and another pathway that requires three substitutions is given in tree C. Tree D has a different topology and requires at least three substitutions. However, in the case of six taxa, there are 105 different topologies, so we have to compute the minimum number of substitutions required for all topologies. If this computation is done for all sites and for all topologies, we can compute the sum of the minimum numbers of substitutions over all sites for each topology. This sum (L or TL) is called the **tree length.** The *maximum parsimony* (MP) tree is the topology that has the smallest tree length. In practice, it is possible that two or more different topologies have the same minimum number of substitutions. In this case, we cannot determine the final topology uniquely, and all equally parsimonious MP trees are considered as potentially correct topologies.

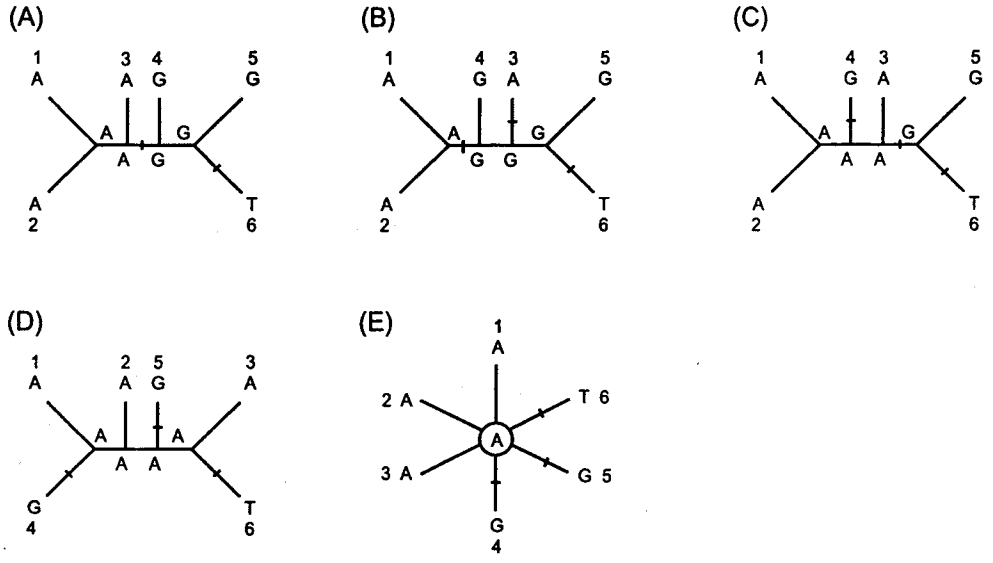


FIGURE 7.2. Assignment of mutations to different branches at a parsimony-informative site.

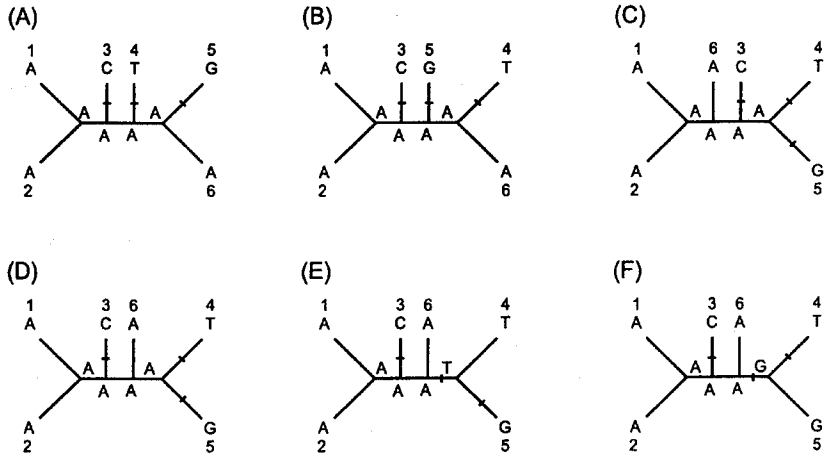


FIGURE 7.3. Assignment of mutations to different branches at a noninformative site.

Informative Sites and Homoplasy

In the search for MP trees, nucleotide or amino acids sites that have the same nucleotide for all taxa (**invariable sites**) are eliminated from the analysis, and only **variable sites** are used. However, not all variable sites are useful for finding an MP tree topology. Any nucleotide site at which only unique nucleotides (singletons) exist is not informative, because the nucleotide variation at the site can always be explained by the same number of substitutions in all topologies. Such a site is called a **singleton site.** For example, tree A in Figure 7.3 has three singleton substitutions C, T, and G and requires three substitutions, but the same number of substitutions is required for any other topology. This can be seen from trees B, C, and D in Figure 7.3. In all these trees (topologies), the three singleton substitutions can be assigned to exterior branches. In some topologies, however, we can assign singleton substitutions to both exterior and interior branches. In Figure 7.3, the topologies of trees D, E, and F are the same but have different assignments of mutational changes to different branches. However, the total number of substitutions is always three. Therefore, this site is not informative for identifying MP trees.

For a nucleotide site to be informative for constructing an MP tree, there must be at least two different kinds of nucleotides, each represented at least two times. These sites are called **informative sites** (Fitch 1977). In trees A, B, C, and D of Figure 7.2, the nucleotide site satisfies this condition, and thus it is useful for finding the topology with the minimum number of substitutions. However, note that singleton sites are informative for topology construction in other tree-building methods. Actually, even invariable sites have some phylogenetic information in distance and maximum likelihood methods. So, Fitch's terminology "phylogenetically informative sites" is not very appropriate. For this reason, we call these sites **parsimony-informative sites.**

In the construction of MP trees, it is sufficient to consider only parsimony-informative sites. However, some authors include singleton sub-

stitutions in the computation of tree lengths. This addition of singleton substitutions to the tree length for parsimony-informative sites does not affect the identification of the MP tree, because the number of singleton substitutions is the same for all topologies. Nevertheless, one should be cautious about the tree length of a published tree and should know whether it is based on only parsimony-informative sites or all variable sites.

Because only informative sites contribute to finding MP trees, it is important to have many informative sites to obtain reliable MP trees. However, when the extent of **homoplasy** (backward and parallel substitutions) is high, MP trees would not be reliable even if there are many informative sites available. For this reason, Kluge and Farris (1969) proposed a quantity called the **consistency index** to measure the extent of homoplasy. This index for a single nucleotide site (i-th site) is given by $c_i = m_i/s_i$, where m_i is the minimum possible number of substitutions at the site for any conceivable topology, and s_i is the minimum number of substitutions required for the topology under consideration. The minimum possible number of substitutions (m_i) is one fewer than the number of different kinds of nucleotides at the site, assuming that one of the observed nucleotides is ancestral. For example, there are three different nucleotides in tree A of Figure 7.2. Therefore, $m_i = 2$. For this topology, s_i is also equal to 2, so $c_i = 1$. This indicates that the nucleotide configuration at this site is supportive of tree A under the MP principle. By contrast, $s_i = 3$ for topologies B, C, and D, so $c_i = 2/3$. Therefore, these topologies are not well supported.

However, the lower bound of the consistency index is not 0, and c_i varies with topology. For this reason, Farris (1989) proposed two more quantities called the **retention index** (see also Archie 1989) and the **rescaled consistency index.** The retention index is given by $r_i = (g_i - s_i)/(g_i - m_i)$, where g_i is the maximum possible number of substitutions at the i-th site for any conceivable tree under the parsimony principle and is equal to the number of substitutions required for a star topology when the most frequent nucleotide is placed at the central node. Diagram E in Figure 7.2 shows such a tree, and in this tree $g_i = 3$. The retention index becomes 0 when the site is least informative for MP tree construction, that is, $s_i = g_i$. In the examples of Figure 7.2, we have $r_i = (3 - 2)/(3 - 2) = 1$ for tree A and $r_i = (3 - 3)/(3 - 2) = 0$ for trees B, C, and D. Therefore, the site under consideration is supportive of tree A but not of the other trees. By contrast, the rescaled consistency index (rc_i) is given by $r_i c_i$. That is,

$$rc_i = \frac{g_i - s_i}{g_i - m_i} \frac{m_i}{s_i} \tag{7.1}$$

This index also is 1 for tree A and 0 for trees B, C, and D. In the present case, therefore, rc_i is identical with r_i, but this is not always the case.

In the above discussion, we considered c_i, r_i, and rc_i for one site. In practice, however, these values are computed for all informative sites, and the **ensemble** or **overall consistency index (*CI*) overall retention index (*RI*),** and **overall rescaled index (*RC*)** for all sites are considered.

These indices are defined as $CI = \Sigma_i m_i/\Sigma_i s_i$, $RI = (\Sigma_i g_i - \Sigma_i s_i)/(\Sigma_i g_i - \Sigma_i m_i)$, and $RC = CI \times RI$, respectively, where i refers to the i-th informative site. These indices should be computed only for informative sites, because for uninformative sites c_i becomes 1 and r_i and rc_i are undefinable. These indices are often used as a measure of accuracy of the topology obtained, particularly for an MP tree obtained from morphological characters. In systematics, $HI \equiv 1 - CI$ is called the **homoplasy index.** When there are no backward and no parallel substitutions, we have $CI = 1$ and $HI = 0$. In this case, the topology is uniquely determined.

Example 7.1. MP Trees for Five Hominoid Species

Let us again consider the DNA sequences given in Figure 6.1 and construct the MP tree. In this data set, if we exclude site 560, in which a deletion exists, there are 281 variable sites of which 90 are parsimony informative. Using these informative sites, we can compute the tree lengths (L) for all the topologies. Only three topologies (B(O(G(C, H))))), (B(O(H(G, C))))), and (B(O(C(G, H)))) have $L = 148$ or less, and all others have much larger L values (Brown et al. 1982). The L values for the three topologies are 147, 145, and 148, respectively, and therefore the topology (B(O(H(G, C)))) is the MP tree. The branch length estimates of this tree are given in Figure 7.4B. The topology of this tree is different from that of the trees obtained by distance methods (Figure 6.2). However, the difference between the two topologies is one branch interchange ($d_T = 2$), and the L value differs only by 2. Therefore, the difference is unlikely to be significant. When the entire sequences of mitochondrial DNA are used, we obtain the same topology as that of the trees in Figure 6.2 (Horai et al. 1995). Another reason why we obtained an erroneous tree seems to be that in this case the transition/transversion ratio is high. In fact, if we use a weighted parsimony method described later, we obtain the same topology as that of the distance trees (Figure 7.4D).

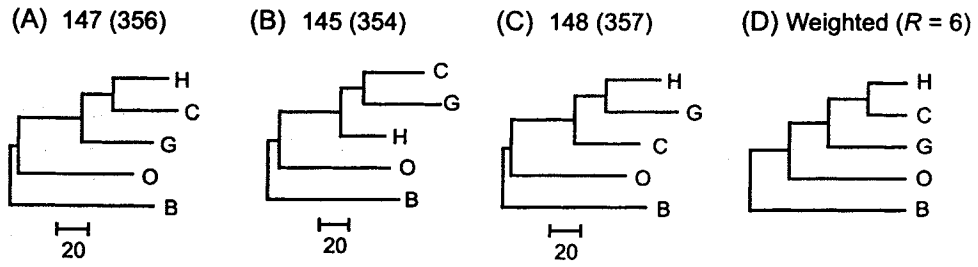


FIGURE 7.4. A–C. Three possible trees (topologies) for the human (H), chimpanzee (C), gorilla (G), orangutan (O), and gibbon (B). These trees were obtained from the DNA sequence data in Figure 6.1. The number given to each topology is the tree length for parsimony-informative sites. The number in parentheses refers to the tree length for all variable sites. D. Maximum parsimony tree obtained when transversions are given six times the weight of the transitional substitutions. Only the branching pattern is shown in D. The overall consistency index (CI) is 0.67, 0.68, and 0.67 for trees A, B, and C, respectively, whereas the overall retention index (*RI*) is 0.47, 0.49, and 0.46 for the three trees.

We have also computed the overall consistency index (*CI*) and the overall retention index (*RI*) for the three topologies. The *CI* values for trees A, B, and C were 0.67, 0.68, and 0.67, respectively. Therefore, the differences in *CI* between the three trees are very small. By contrast, the *RI* values for the three trees were 0.47, 0.49, and 0.46, respectively. These values are negatively correlated with the tree lengths.

7.2. Strategies of Searching for MP Trees

When the number of sequences or taxa (m) is small, say, $m < 10$, it is possible to compute the tree lengths of all possible trees and determine the MP tree. This type of search for MP trees is called the **exhaustive search.** As previously mentioned, the number of topologies rapidly increases as m increases (Equation [5.1]). Therefore, it is virtually impossible to examine all topologies if m is large. However, if we know clearly incorrect topologies, as in the case of the five hominoid species in Figure 6.1, we do not have to compute the L values for them. We can simply compute L's only for potentially correct trees. This type of search is called the **specific-tree search.**

There are two ways of obtaining MP trees when $m > 10$ and the specific-tree search is not applicable. One is to use the **branch-and-bound method** (Hendy and Penny 1982). In this method, the trees that obviously have a tree length longer than that of a previously examined tree are all ignored, and the MP tree is determined by evaluating the tree lengths for a group of trees that potentially have shorter tree lengths. This method guarantees finding of all MP trees, although it is not an exhaustive search. However, even this method becomes very time-consuming if m is about 20 or larger. In this case, one has to use another approach called the **heuristic search.** In this method, only a small portion of all possible trees is examined, and there is no guarantee that the MP tree will be found. However, it is possible to enhance the probability of obtaining the MP tree by using several algorithms.

Branch-and-Bound Search

After the branch-and-bound method was introduced by Hendy and Penny (1982) in parsimony analysis, several different versions were developed (Swofford and Begle 1993). The differences in algorithm and efficiency among them are rather small, and here we present Kumar et al.'s (1993) version. In this version of the branch-and-bound method, the search for an MP tree starts with an initial core tree of three taxa, which has only one unrooted tree (tree A in Figure 7.5). The remaining taxa are added to this core tree one by one according to a certain order, and the tree length of the new tree is computed at each stage of taxon addition. If the addition of a taxon to a particular branch of a core tree results in a tree length greater than a predetermined upper bound of tree length (L_U), this topology and all the subsequent topologies that can be generated by adding more taxa to this core tree are ignored from further consideration.

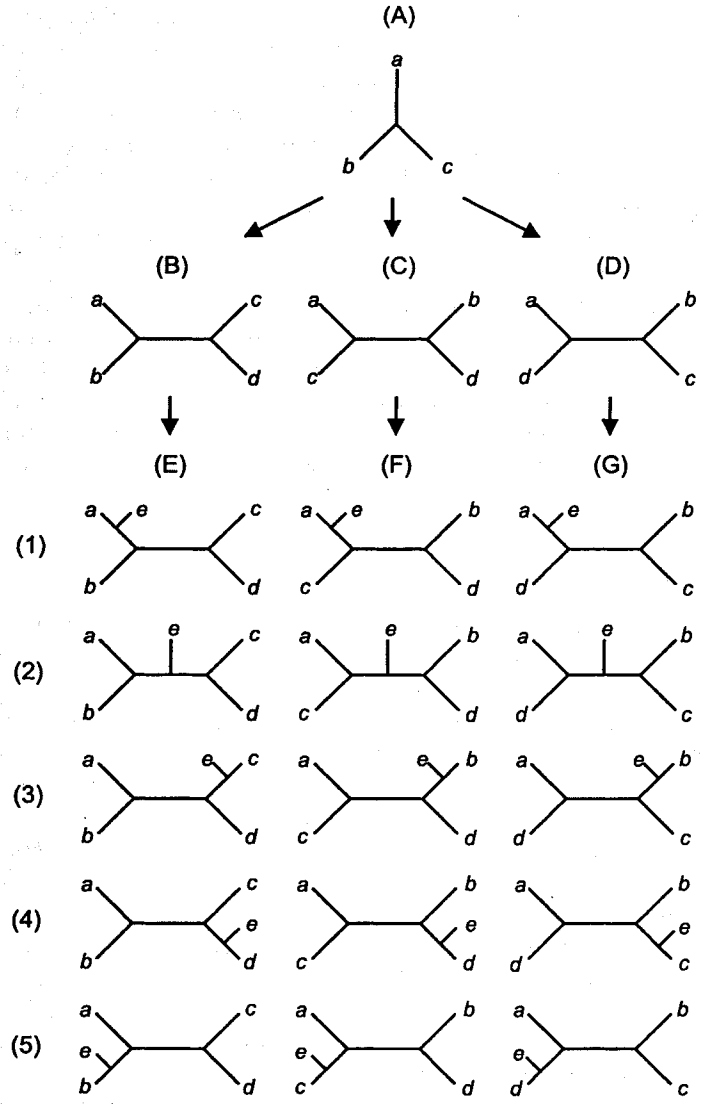


FIGURE 7.5. Diagrams showing the procedures of the branch-and-bound and the heuristic branch-and-bound-like searches.

Core Tree and Order of Taxon Addition

The initial core tree of three taxa is chosen such that the length (L) of the tree is largest or approximately largest among all possible three-taxon trees (Figure 7.5). This is to make L closer to the length (L_M) of the MP tree so that we can reach the MP tree faster. The next step is to determine the order of taxon addition that makes the search for the MP tree faster. To do this, we place one of the remaining taxa on one of the three branches of the initial core tree and compute the tree length by the MP procedure. We repeat this computation for the two remaining branches

and record the minimum value of the three tree lengths. We repeat this procedure for all remaining taxa. We then find the taxon that shows the maximum value of the minimum tree lengths. This taxon is the first to be added to the initial core tree. We call this procedure the maximum-of-the-minimum-values algorithm or simply the **max-mini algorithm.** To find the next taxon, we apply this max-mini algorithm for the remaining taxa using the tree for the first four taxa as the next core tree. In this case, of course, the number of minimum tree lengths to be computed for each taxon is five, because a four-taxon tree has five branches. We can then find a taxon that shows the maximum of the minimum tree lengths. This taxon will be the second one to be added to the initial core tree of three taxa. This process is repeated until the addition order of all taxa is determined. Since the maximum of the minimum values is closer to L_M than many other value (e.g., the minimum of the minimum values), this order of taxon addition is expected to speed up the search for the MP tree.

Search for MP Tree(s)

Once the initial core tree and the order of taxon addition are determined, we are in a position to search for the MP tree. Before starting this search, we must have a predetermined upperbound of tree length, that is, L_U for a **temporary MP tree.** This value is a temporary minimum number of substitutions, which is likely to be slightly larger than the real minimum number, L_M. We determine this value by running the heuristic search called the stepwise addition or the branch-and-bound-like algorithm.

Let us now explain the algorithm for finding the MP tree by using the diagrams in Figure 7.5. We start with the initial core tree in diagram A. In this example of five taxa, taxa *a, b,* and *c* form the initial core tree, and taxa *d* and *e* are added in this order. There are three ways of adding *d* to the core tree (trees B, C, and D). We first compute the tree length (L) for tree B. If this L is greater than L_U, we ignore all the subsequent trees that are generated by adding taxon *e* to this tree (five trees given in column E). If $L \leq L_U$, we add *e* to each of the five branches of tree B to form five different trees with five taxa. We again compute L for each of these five trees and find a tree (or trees) that shows the smallest L value. If this L is greater than L_U, then we move on to tree C. However, if L is equal to L_U for a tree, we save the tree as another potential MP tree and move on to tree C. If a tree (or trees) in column E has an L smaller than L_U, then this tree will become the next temporary MP tree, and L_U is now replaced by this new L value. We then move to tree C. We apply the same procedure to tree C and the trees generated by adding *e* to tree C. If all these trees are examined, we then move to tree D and its descendant trees. Since we adjust L_U whenever we find a tree with an L smaller than the previous L_U, we are assured of finding the MP tree. Of course, there may be two or more equally parsimonious trees, and in this case all these trees are identified by the present method. The same algorithm can be used for the case where the number of taxa (m) is greater than five. This algorithm saves computer time considerably, because many trees need not be examined

if L_U is sufficiently close to the tree length (L_M) of the true MP tree(s). However, even this method becomes time-consuming when $m \geq 20$.

Heuristic Search

Several algorithms of the heuristic search for MP trees are now available (see Maddison and Maddison 1992; Swofford and Begle 1993), but many of them are based on the same principle. In these algorithms, a provisional MP tree is first constructed by using a procedure called the **stepwise addition algorithm,** and this provisional MP tree is then subjected to some kind of **branch swapping** to find a more parsimonious tree. In the following, we first explain the principle of the stepwise addition algorithm and then branch swapping procedures. In addition, we present one more heuristic search algorithm whose principle is different from that of the traditional ones.

Stepwise Addition Algorithms

In this set of algorithms, an initial core tree of three taxa is first formed according to a certain rule, and each of the remaining taxa is then chosen for the next taxon addition. This taxon is connected to one of the three branches of the initial core tree, and the tree lengths of the three resulting trees are evaluated. After this evaluation, the tree of four taxa whose tree length is shortest is saved for the next step of taxon addition. The next taxon is then connected to each of the five branches of the four-taxon tree, and the five-taxon tree whose tree length is shortest is chosen. This process is continued until a tree of all taxa is produced. This final tree is the provisional MP tree. This provisional MP tree usually has a longer tree length than that (L_M) of the MP tree. Therefore, this tree is subjected to branch swapping procedures to find a tree that has a smaller L value. Application of several rounds of branch swapping usually produces a tree whose branch length is considerably shorter than that of the provisional tree, and this tree is regarded as the MP tree.

Swofford and Begle (1993) describe various ways of producing the provisional MP tree considering the order of taxon addition. The simplest one is the "**as is**" option, in which the initial core tree is produced by the first three taxa given in the data set, and the following taxon addition is done according to the taxon order in the data set. Usually this method is not very effective for finding a tree with a small L. The second simplest method is the "**random**" option, where pseudorandom numbers are used to determine the order of taxon addition, and this procedure is applied many times to obtain a provisional MP tree. Another one is called the "**closest**" option, in which the initial core tree is produced by examining all triplets of taxa and choosing the one that shows the smallest L, and in the following steps, a taxon whose addition to the previous core tree shows the smallest increase in L is chosen. The reader should refer to Swofford and Begle (1993) for details of these options. All of these options are included in the software PAUP*, whereas PHYLIP primarily uses the second (Jumble) option. Once a provisional MP tree is produced,

the tree is subjected to one or two of the following algorithms of branch swapping.

Branch Swapping

The most popular algorithms of branch swapping are (1) **nearest neighbor interchanges (NNI)**, (2) **subtree pruning regrafting (SPR)**, and (3) **tree bisection-reconnection (TBR)** (Swofford and Begle 1993). The first algorithm is the same as the examination of all trees that are different from the provisional MP tree by a topological distance of $d_T = 2$ (Figure 7.6A). For example, for the five-taxon tree in Figure 7.6A, there are two alternative trees (interchanges of taxa 2 and 3 and taxa 1 and 3) with a topological distance of $d_T = 2$ from the original tree when the interior branch a is considered. Two more alternative trees can be produced if we consider the interior branch b. This algorithm is obviously related to the *close neighbor interchange* (CNI) algorithm described in relation to the ME method (chapter 6). In the latter algorithm, the trees that are different from the provisional tree by $d_T = 2$ and 4 are examined, and this search is repeated until no tree with a smaller L is found. Therefore, this algorithm examines more trees than the NNI search.

In the SPR algorithm, a branch of a provisional tree is cut into two parts, a pruned subtree and the residual tree. The cutting point of the pruned subtree is then grafted onto each branch of the residual tree to produce a new topology. This is done for all branches of the residual tree to produce more trees to be examined. This is illustrated in Figure 7.6B. In this example, the exterior branch a was cut, and the pruned subtree consists of taxon 1 only, whereas the residual tree is composed of taxa 2, 3, 4, and 5. There are four ways of grafting the subtree to the residual trees in this case. If the interior branch b is cut instead, the subtree is grafted to two exterior branches (4 and 5) of the residual tree to produce two alternative trees. This procedure can be used for a tree of any number of taxa.

In the TBR search, a provisional tree is cut into two subtrees at a branch, and these two subtrees are then reconnected by joining two branches, one from each subtree, to generate a different topology (Figure 7.6C). This is tried for all possible pairs of branches of the two subtrees to generate many different topologies. In the SPR search, only the cutting point of a subtree was regrafted to a branch of the residual tree, whereas in the TBR search all combinations of branches from the two subtrees are considered for reconnection. Therefore, the number of topologies generated is larger than that generated by the SPR search when the number of taxa is greater than five.

Since the TBR search examines a larger number of trees than the NNI and SPR searches, many investigators use this method. However, even this method examines only a limited number of trees when the number of taxa is large (Maddison 1991). One way to increase the number of trees to be examined is to use the "random" option of stepwise addition and the TBR search repeatedly. If this approach is used a large number of times, the chance of finding the MP or a suboptimal MP tree is quite high.

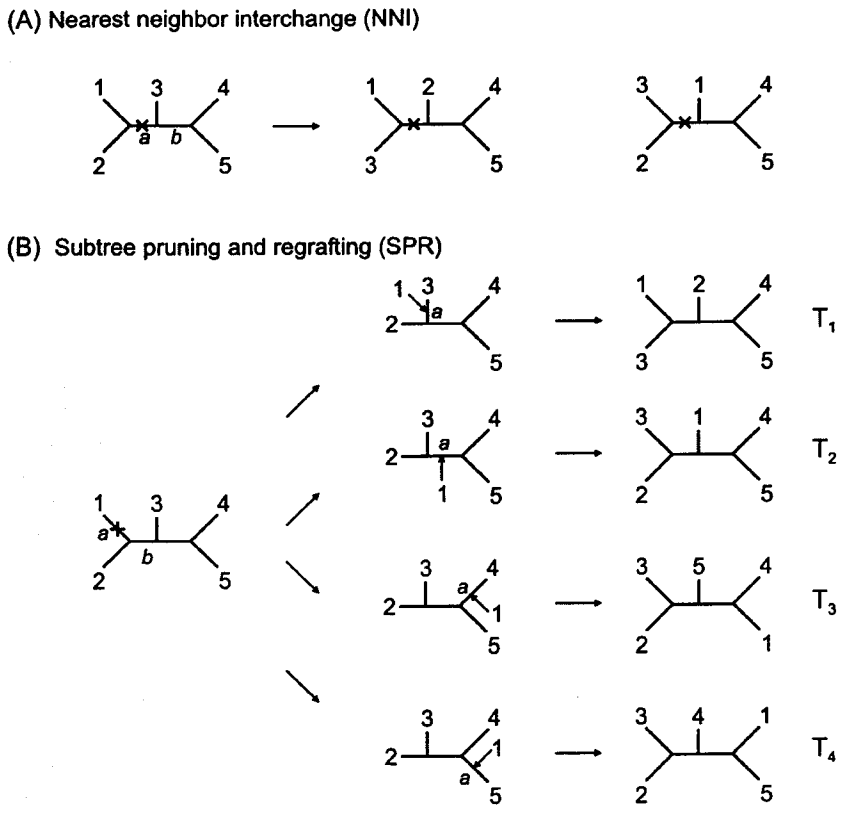


FIGURE 7.6. Three different methods of branch swapping for finding MP trees.

Branch-and-Bound-Like Algorithm

Kumar et al. (1993) proposed a heuristic search algorithm, which is conceptually different from the algorithms mentioned above but is similar to the branch-and-bound method. In this algorithm, we start with an initial core tree of three taxa that is determined as in the case of the branch-and-bound method. The order of taxon addition is also determined in a similar fashion except for the following. In the branch-and-bound

method, we computed the minimum numbers of substitutions for all taxa for each core tree (each step of taxon addition) and then chose the taxon that showed the maximum value among all the minimum values. For this heuristic search, which may be called the **min-mini algorithm,** we choose the minimum of all the minimum values, because we are not going to do a semiexhaustive search as in the case of the branch-and-bound method and want to reach the MP or a suboptimal MP tree relatively quickly.

The algorithm of searching for the MP tree is also similar to that of the branch-and-bound method. Let us again consider Figure 7.5 to explain this algorithm. As before, we start with the core tree A and first connect taxon *d* to branch *c* to produce tree B. We then compute the tree length (*L*) of this tree. We call this *L* value the **local upperbound** (L_1) for the first taxon addition and keep this value for future use. We then connect taxon *e* to branch *a* of tree B to produce tree E (1). We again compute the *L* value of this tree and call it the local upperbound (L_2) for the second taxon addition. If there is another taxon (*f*) to be added, we connect this taxon to branch *a* of tree E (1) and obtain tree E (1, 1) in Figure 7.7. If *f* is the last taxon to be added, we now compute the *L* value not only for tree E (1, 1) but also for all other six trees that can be derived from tree E (1). We then choose the tree that shows the smallest *L* value among the seven trees and call it a temporary MP tree. The *L* for this tree is the temporary upperbound (L_U) in this case.

The next step is to go back to tree E (2) in Figure 7.5 and compute the *L* value. If this *L* is greater than L_2, we neglect all trees that can be generated by adding *f* to this tree. If $L = L_2$, we compute *L* for all the descendant trees. If any of the descendant trees show an *L* equal to L_U, the tree is saved as another potential MP tree. If there is any tree showing an *L* less than L_U, this tree now becomes a new temporary MP tree, and the previous L_U is replaced by this *L*. By contrast, if tree E (2) shows an *L* less than L_2, L_2 is replaced by this *L*. The *L* values for all descendant trees are then computed, and a new potential MP tree or a new temporary MP tree is searched for. This procedure is applied to the remaining three trees E (3), E (4), and E (5) of five taxa, and the temporary MP tree (or trees) that shows the smallest *L* value among the 35 ($= 5 \times 7$) trees derived from tree B is determined.

If the above computation is completed, we now move on to tree C (and tree D) in Figure 7.5 and apply the same procedure to all trees that can be derived from these trees. When this is completed, we have the final tree or trees. When there are more than six taxa, essentially the same algorithm is applied. The only difference is that there are many steps of taxon addition and that at each step of taxon addition the local upperbound ($L_1, L_2, L_3, \ldots, L_{m-3}$, or L_U) is computed, where *m* is the number of taxa. $L_1, L_2, L_3, \ldots, L_{m-4}$, and L_U are then used to determine whether a group of descendant trees should be ignored or not in later computations.

In this algorithm, many trees that are unlikely to have a small *L* value are ignored, and thus the algorithm speeds up the search for the MP tree. However, the final tree or trees obtained by this algorithm may not be the true MP tree(s), because the upperbounds of the *L* values used here are

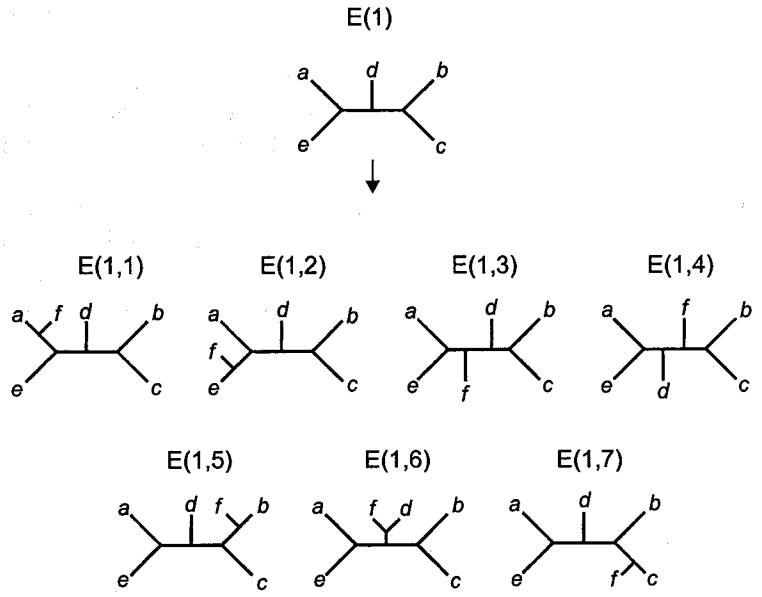


FIGURE 7.7. All possible trees that can be generated by adding taxon *f* to tree E(1) in Figure 7.5.

local upperbounds rather than the global upperbound as used in the branch-and-bound method, and the tree with the global minimum value of L may not have been obtained.

There is a way to improve the efficiency of finding the MP tree. It is to increment the local upperbound at each step of taxon addition. If the local upperbound is large, the number of trees to be examined automatically increases. In the above algorithm, the local upperbound at the i-th step of taxon addition was L_i except for the first step. We now increase L_i by x_i so that the upperbound is given by $L_i' = L_i + x_i$. If x_i is large for all i's, a large number of topologies will be examined. In this case, however, the computational time will be prohibitively large. We call x_i a **search factor.** In the default option of MEGA, $x_i = 2$ is used for all steps, but the user may change it as desired. MEGA2 has another option, in which x_i is defined as $x_i = p(L_{i+1} - L_i)$, where L_i and L_{i+1} are the upperbound of L for the i-th and the $(i + 1)$-th steps, and p is the fraction of $L_{i+1} - L_i$ that one wishes to use. Suppose $L_{i+1} = 200$ and $L_i = 180$, and one wishes to use $p = 0.1$. Then, x_i becomes 2. We call the p value a **proportional search factor.** The optimum p value varies with data set, and the user of MEGA2 may find it by trial and error.

Some Remarks

As mentioned earlier, MP methods tend to give incorrect topologies when the number of sequences used (m) is large and the number of nucleotides used (n) is small. In this case, one may choose an incorrect topology by making an excessive effort to find the real MP tree. When m

is large, some parts of the MP tree (or any other tree) are likely to be incorrect, and a submaximum parsimony tree may be as good as the MP tree in finding the true topology. For this reason, Nei et al. (1998) suggested that a relatively crude method of finding MP or potential MP trees gives essentially the same conclusion about phylogenetic inference as the exhaustive search when the accuracy of the tree obtained is examined by the bootstrap test. In fact, our computer simulation (Takahashi and Nei 2000) has shown that for randomly generated model trees of 48 sequences with n = 1000 the NNI search of MP trees is as efficient as the TBR search in inferring the true tree. This indicates that MP trees are often incorrect and that there is no need to spend an enormous amount of computer time for obtaining MP trees.

7.3. Consensus Trees

Strict and Majority-Rule Consensus Trees

As mentioned above, MP methods often produce several equally parsimonious trees. In this case, it is difficult to present all the trees for publication. One way to solve this problem is to make a composite tree that represents all the trees. Such a composite tree is called a consensus tree.

There are several different types of consensus trees (Swofford and Begle 1993), but the most commonly used ones are the **strict consensus trees** and the **majority-rule consensus trees.** Let us explain these trees using the examples given in Figure 7.8. Suppose that trees A, B, and C are three equally parsimonious trees obtained by an MP method. In a strict consensus tree, any conflicting branching patterns for a set of sequences among the rival trees are resolved by forming a multifurcating branching pattern. Thus, the strict consensus tree for trees A, B, and C is given by tree D. Among the majority-rule consensus trees, the most commonly used is the 50% majority-rule consensus tree. In this tree, a branching

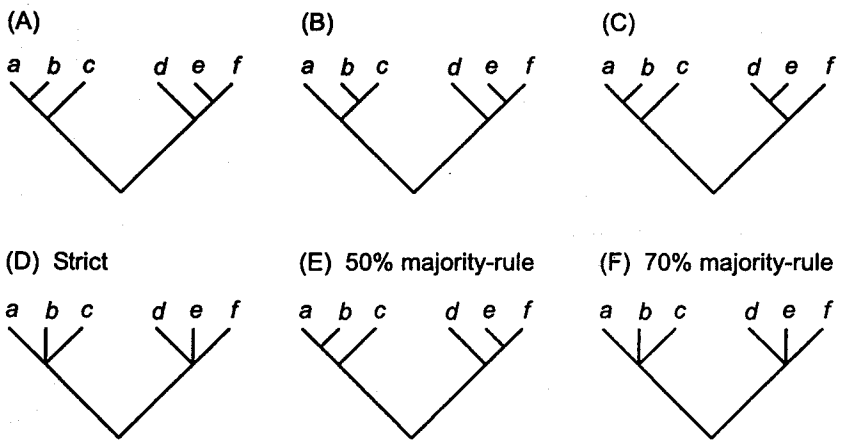


FIGURE 7.8. Examples of consensus trees.

pattern that occurs with a frequency of 50% or more is adopted. In the present example, the branching pattern ((*a*, *b*) *c*) for taxa *a*, *b*, and *c* occurs two times among the three rival trees, so this pattern is adopted. Similarly, branching pattern ((*e*, *f*) *d*) occurs two times among the three trees. Therefore, the 50% majority-rule consensus tree is given by tree E. It is possible to increase the majority-rule percentage. For example, if we use 70%, none of the branching patterns of the two three-taxon clusters reaches 70%. Therefore, the 70% majority-rule consensus tree (tree F) is identical with the strict consensus tree. Note that the 100% majority-rule consensus tree is always identical with the strict consensus tree.

Bootstrap Consensus Trees

One of the effective ways of testing the reliability of an MP tree is to use the bootstrap test, which will be discussed in chapter 9. In this test, the reliability of an inferred tree is examined by using Efron's bootstrap resampling techniques. A set of nucleotide sites is randomly sampled with replacement from the original set, and this random set that has the same number of nucleotide sites as that of the original set is used for constructing a new tree. The topology of this tree may be or may not be the same as that of the inferred tree. This process is repeated many times (over 100 times), and the reliability of the inferred tree is evaluated by the percentage of times in which each branching pattern (sequence partition) is found among all the replicate bootstrap trees (see chapter 9 for details).

Felsenstein (1985) proposed to construct a consensus tree from the replicate bootstrap trees and use it as a new inferred tree. This new inferred tree may be different from the original inferred tree, but since it is an "average" tree of many bootstrap trees, it may be more reliable than the original one, though there is no proof. In the case of MP trees, this procedure also has an advantage to avoid multifurcating trees by producing a low-percentage majority-rule consensus tree. In Figure 7.8, we saw that the 50% majority-rule tree for trees A, B, and C is a bifurcating tree. If we have several hundred bootstrap trees and if we make a 5% majority rule consensus tree, the tree will be almost always a bifurcating tree.

7.4. Estimation of Branch Lengths

MP methods are often used to construct a tree topology without branch lengths. However, it is possible to estimate the branch lengths of a reconstructed tree under certain assumptions, and these estimates should be presented for an MP tree as much as possible.

The branch lengths of an MP tree are estimated by considering all evolutionary pathways at each variable site and computing the average number of substitutions for each exterior or interior branch. When there is only one singleton substitution at a site, this substitution can always be assigned to the exterior branch leading to the taxon that has the substitution. When there are two or more singleton substitutions, there are sev-

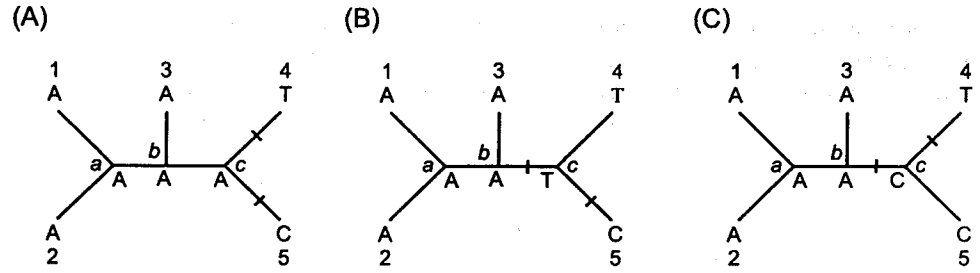


FIGURE 7.9. Assignment of substitutions to different branches when there are two or more singleton substitutions.

eral ways of assigning the substitutions (evolutionary pathways). For example, Figure 7.9 shows a case of two singleton substitutions, in which there are three different evolutionary pathways, and the average number of substitutions for each of the branches $b - c$, $c - 4$, and $c - 5$ is 2/3. In the case of trees D–F in Figure 7.3, there are three singleton substitutions for the same topology, and the number of substitutions for the branch leading to taxon 3 is 1, whereas the number for the branch leading to taxa 4 and 5 is 2/3. In addition, one interior branch has the average number of 2/3 substitutions. Therefore, if we know all the pathways, we can compute the average number of substitutions for each branch.

This is also true for parsimony-informative sites. Let us consider this problem using the tree given in Figure 7.1. We have seen that the nucleotides observed in the six extant taxa generate four equally parsimonious pathways in the unrooted topology of this tree. They are (*a* - T, *b* - T, *c* - T, *d* - T), (*a* - T, *b* - T, *c* - A, *d* - A), (*a* - C, *b* - T, *c* - A, *d* - A), and (*a* - A, *b* - T, *c* - A, *d* - A). The first pathway requires one substitution for each of the branches 1-*a*, 5-*c*, and 6-*d*, and the second requires one substitution for each of the branches 1-*a*, *b*-*c*, and *a*-*d*. Similarly, the third and fourth pathways require one substitution for branches 2-*a*, *b*-*c*, *a*-*d*, and 1-*a*, 2-*a*, *b*-*c*, respectively. We can therefore compute the average numbers of substitutions for all branches. We obtain the numbers 3/4, 2/4, 0, 0, 1/4, 1/4, 2/4, 3/4, and 0 for branches 1-*a*, 2-*a*, 3-*b*, 4-*b*, 5-*c*, 6-*d*, *a*-*d*, *b*-*c*, and *c*-*d*, respectively. The length of a branch can then be obtained by adding all substitutions for that branch at both singleton and informative sites. We call this way of estimating branch lengths the **average pathway method.**

Maddison and Maddison (1992) and Swofford and Begle (1993) also estimate the branch lengths by using two algorithms: **Acctran** and **Deltran.** In Acctran, evolutionary changes of nucleotides are assumed to occur as soon as possible from the root, whereas in Deltran the changes are assumed to occur as late as possible from the root (Swofford and Maddison 1987). As an example, let us consider tree B of Figure 7.1 and assume that the nucleotide (A) for sequence 6 represents the ancestral nucleotide at node *e*. In the Acctran algorithm, A is assumed to change to T at the earliest node *d*, and then the minimum number of nucleotide changes is considered. Therefore, nodes *a*, *b*, and *c* are also assumed to be T. In the Deltran algorithm, however, nucleotide changes are delayed

as much as possible. Therefore, the nucleotides at nodes *a, b, c,* and *d* are assumed to be A, T, A, and A, respectively. (Node *e* is not considered.) This indicates that the nucleotide assignments for the ancestral nodes are considerably different between Acctran and Deltran, and therefore the estimates of branch lengths will also be different. When closely related sequences are examined, however, the difference in branch length estimates between the two methods is not as large as one might suspect.

In general, the estimates of branch lengths obtained by parsimony methods tend to be smaller than the actual values, particularly when sequence divergence is high. One way to avoid this underestimation of branch lengths is to use the least squares or the ML method after the topology of the tree is determined by MP methods. Under certain conditions, MP methods seem to be superior to distance or ML methods for finding the correct topology (see chapter 9). Therefore, this approach may give a more reliable topology and more reliable estimates of branch lengths.

7.5. Weighted Parsimony

As mentioned earlier, MP methods are expected to produce more reliable trees when the number of backward and parallel substitutions (extent of homoplasy) is small than when it is large. Therefore, if a set of sequences used for phylogenetic analysis includes fast-evolving and slow-evolving sites, one would expect that the latter sites are more useful than the former sites for constructing MP trees when distantly related sequences are studied. Therefore, if we give more weight to slow-evolving sites than to fast-evolving sites, a more reliable tree may be obtained than when they are equally weighted (Farris 1969; Swofford et al. 1996).

For example, the nucleotides at the first, second, and third codon positions of protein-coding genes are known to evolve at different rates, those at the third positions evolving fastest and those at the second positions slowest (see Table 3.4). Therefore, one may give such weights as $w_1 = 3$, $w_2 = 5$, and $w_3 = 1$ for the first, second, and third positions, respectively. These weights would of course vary from gene to gene. Generally speaking, functionally less important parts of a gene are known to evolve faster than more important parts (Dickerson 1971; Kimura 1983). Therefore, one may give the former a lower weight than the latter when distantly related sequences are studied.

Weighted parsimony also allows different weights to be given to different types of substitutions at a given site. For example, transitional nucleotide substitutions generally occur more frequently than transversional substitutions, as mentioned earlier. In this case, it is convenient to use a **substitution weight matrix** as given in Figure 7.10A. A weight matrix is sometimes called a **step matrix** (Swofford and Begle 1993). If transitions occur twice as frequently as transversions, we may give $w = 2$. The matrix in Figure 7.10B gives a weight of 0 to all transitional changes and 1 to all transversional changes. Therefore, transitional changes are completely ignored, and only transversional changes are considered. This type of MP method is called **transversion parsimony.**

(A) Weighted parsimony

	A	T	C	G
A		*w*	*w*	1
T	*w*		1	*w*
C	*w*	1		*w*
G	1	*w*	*w*	

(B) Transversion parsimony

	A	T	C	G
A		1	1	0
T	1		0	1
C	1	0		1
G	0	1	1	

FIGURE 7.10. (A) Weight matrix for transitional and transversional substitutions. (B) Weight matrix for transversion parsimony.

Figure 7.4D shows a weighted parsimony tree obtained from the DNA sequence data for the five hominoid species considered earlier (Figure 6.1). Kimura's formula for estimating the transition/transversion ratio (Equation [3.18]) gave $\hat{R} = 6$ for this set of data. This suggests that the transition rate is about six times higher than the transversion rate. We therefore used $w = 6$ in constructing the weighted parsimony tree. Interestingly, the topology of this tree is the same as that of the trees obtained by distance methods (Figure 6.2).

However, there are some problems with weighted parsimony. First, we usually do not know the appropriate weights to be used in actual data analysis. In some cases, information from previous studies can be used, but there is no guarantee that they are appropriate for the data set under consideration. For this reason, a number of authors (Farris 1969; Sankoff and Cedergren 1983; Williams and Fitch 1990) proposed a method called **dynamically weighted parsimony.** In this method, a set of weight parameters that appear to be appropriate a priori are first used to construct an MP tree, and this tree is now used to obtain an improved set of weight parameters. These new parameters are then used to construct a new MP tree. This process is repeated until a stable tree (or trees) is obtained. This is a time-consuming method and does not guarantee convergence to a stable tree. Nevertheless, computer simulation has shown that this method improves the probability of obtaining the correct tree under certain conditions (Fitch and Ye 1991). Second, slowly evolving sites or slowly changing substitution types are informative only when distantly related sequences are studied. When closely related sequences are used, the fast-evolving sites or types of substitutions are obviously more informative. However, actual data often include both distantly related and closely related sequences, and in this case, it is not clear how useful weighted parsimony is. This problem needs more investigation.

Example 7.2. Unweighted and Weighted Trees for Simulated Sequence Data

One of the virtues of MP methods is that when there are no backward or parallel substitutions and there are a sufficiently large number of informative sites, they are able to reconstruct the true tree irrespective of the pattern of nucleotide substitution. This suggests that they will produce a highly reliable tree when the extent of sequence divergence is low. To see whether this is the case or not, we constructed MP trees using the

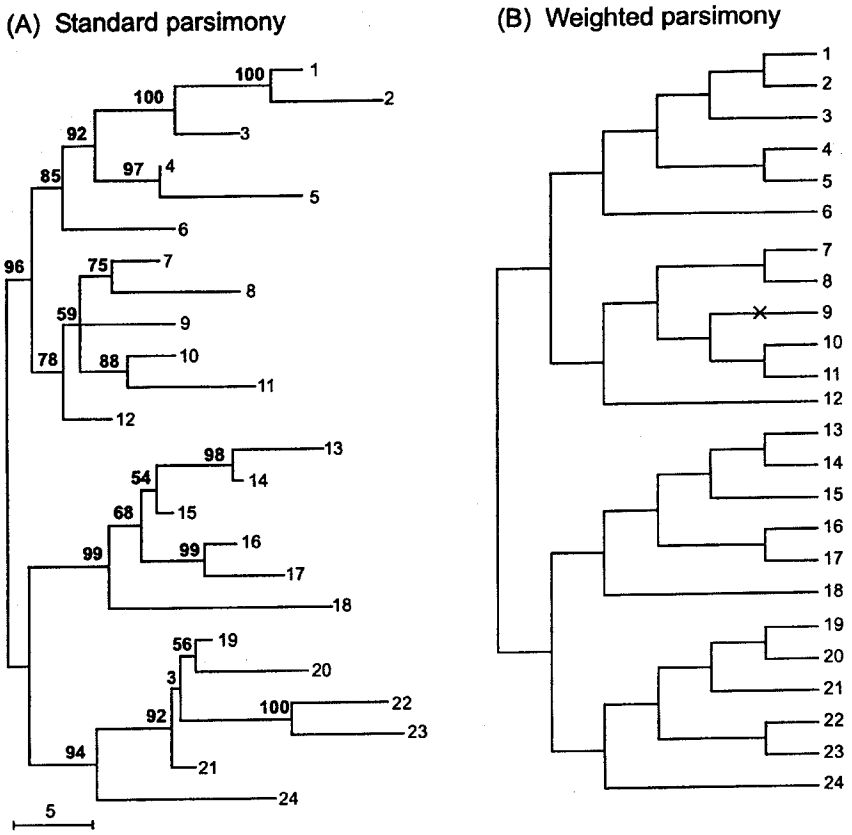


FIGURE 7.11. (A) Standard parsimony tree for simulated sequence data as inferred by the stepwise-addition with "closest" option in PAUP* (no branch swapping). Bootstrap values are shown in boldface (100 replications). (B) Weighted parsimony for the same data set ($w = 5$ was used). The branching pattern of sequence 9 is incorrect (see Figure 6.8A).

24 DNA sequences discussed in chapter 6 (Figure 6.8). We first used the stepwise addition algorithm with the "closest" option in PAUP* to construct the MP tree and the average pathway method to estimate the branch lengths. The tree obtained is presented in Figure 7.11A. Comparison of this tree with that in Figure 6.8B indicates that the MP tree is virtually the same as the true (realized) tree. The only topological difference observed is the interchange of sequence 21 with the cluster of sequences 22 and 23. The branch lengths are also very close to those of the true tree, though there is some tendency for the interior branches of the MP tree to be underestimated because of homoplasy. When we tried a heuristic search with 50 replications of TBR branch swapping, we found two more MP trees, which were different with respect to the splitting pattern of sequences 21, 22, and 23.

These results indicate that when sequence divergence is low, MP trees are very close to the true tree. They are also similar to the NJ tree in Figure 6.8C. It is interesting to note that NJ resolved the branching pattern of sequences 21, 22, and 23 correctly but did not produce a trifurcating node

for sequences 8, 9, and 10. By contrast, MP had no trouble identifying the trifurcating node but did not produce the correct branching pattern for sequences 21, 22, and 23.

We also constructed a weighted MP tree with a transition/transversion ratio (w) of 5. (Note that the DNA sequences used were generated with a transition/transversion ratio [R] of 5.) The tree is presented in Figure 7.11B. The topology of this tree is identical with that of the realized tree B in Figure 6.8 except for one topological error that occurred with respect to the branching pattern of sequence 9. The branch lengths of this tree do not have any biological meaning.

Example 7.3. Origin of Whales

Whales are the largest animals that have ever lived on Earth. They belong to the mammalian order Cetacea that includes whales, dolphins, and porpoises, which are all adapted to aquatic life. The evolutionary origin of cetaceans has been a mystery over a century. In recent years, however, their evolutionary relationships with other mammalian orders are being clarified thanks to molecular data. Although whales were once believed to be related to horses, elephants, or some other mammalian order, it is now generally agreed that they are most closely related with artiodactyls. The order Artiodactyla was traditionally divided into three suborders, Ruminantia (e.g., deer, giraffes, cows, sheep, chevrotains), Tylopoda (e.g., camels), and Suiformes (e.g., pigs, peccaries, and hippopotamuses), and each of these suborders had been considered to be monophyletic. Recent molecular data, however, suggest that the order Cetacea is most closely related to Ruminantia, and therefore Cetacea is included inside the order Artiodactyla (Graur and Higgins 1994; Gatesy 1997; Shimamura et al. 1997).

Here we construct a phylogenetic tree using DNA sequences of the blood-clotting protein γ-fibrinogen gene. This gene consists of 10 exons and spans an 8 kb region of nuclear DNA. Gatesy (1997) sequenced a 523–581 bp fragment of the gene for six species of artiodactyls, three species of cetaceans, two species (horse and Asiatic tapir) of Perissodactyla (odd-toes ungulates), and two species (spotted hyena and coyote) of Carnivora. Adding the human sequence available, Gatesy constructed an MP tree using PAUP* with 50 random taxon replicates and TBR branch swapping. Here we used the branch-and-bound method to construct the MP tree. The number of nucleotides used was 433 after elimination of all alignment gaps. The branch-and-bound method produced three equally parsimonious trees, one of which is presented in Figure 7.12A. The other two trees had different branching patterns for sheep, giraffe, and moose, that is, ([sheep, giraffe] moose) and ([sheep, moose] giraffe) instead of ([moose, giraffe] sheep). All three trees had a tree length of 485 substitutions. When a 30% bootstrap majority-rule consensus tree was constructed from 500 replications, we obtained the same tree as that of Figure 7.12A. However, the bootstrap value of the giraffe-moose cluster is so low that the branching pattern for sheep, giraffe, and moose remains unresolved. We also constructed the NJ and ME trees using Kimura distance for this data set. The NJ tree was identical with

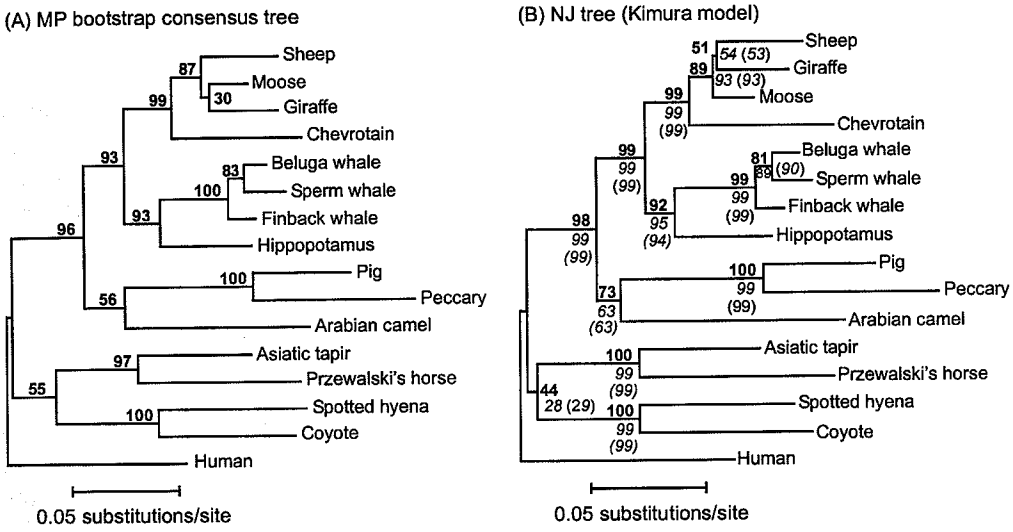


FIGURE 7.12. (A) MP bootstrap consensus tree obtained by the fast-heuristic search of PAUP*. (B) NJ tree. Bootstrap values are shown in boldface (1000 replications), and the PC values (for NJ tree) are shown in italics. The PC values obtained by Dopazo's method are shown in parentheses.

the ME tree and is given in Figure 7.12B. This tree has the same topology as that of tree A except for the branching pattern of sheep, giraffes, and moose, which again has low bootstrap values. Therefore, both MP and NJ trees give essentially the same conclusion as to the phylogenetic relationships of the organisms studied.

Both trees A and B show a branching pattern that is unexpected from the classical taxonomy. That is, cetaceans are close relatives of ruminants, and the other two suborders of Artiodactyla, i.e., Tylopoda and Suiformes, are outgroups of ruminants and cetaceans. This indicates that the order Artiodactyla is not monophyletic but **paraphyletic** (Wiley et al. 1991), because it does not include Cetacea, which is a close relative of the suborder Ruminantia. Since this classification is unnatural, Montgelard et al. (1997) proposed a new mammalian order named Cetartiodactyla that includes both Artiodactyla and Cetacea. This conclusion has been supported by Gatesy et al.'s (1999) further study using a larger set of DNA sequences. It is also supported by the work of Nikaido et al. (1999), who used an entirely different approach (see section 7.7).

7.6. MP Methods for Protein Data

Eck and Dayhoff (1966) used an MP method for protein sequence data. They considered 20 different amino acids as character states and constructed an MP tree, assuming that the evolutionary change can occur in all directions among the 20 amino acids. Dayhoff and her collaborators (Dayhoff 1972) used this method extensively and obtained quite reason-

able trees for various protein sequences. A computer program incorporating this method is available in PAUP* and MEGA2.

Theoretically, this approach is approximate, because some amino acid changes require two or three nucleotide substitutions, whereas other changes can be explained by one substitution. Furthermore, some amino acids are biochemically similar to one another, and substitution occurs more often within each group of similar amino acids than between groups. For this reason, a number of authors (Moore et al. 1973; Fitch and Farris 1974; Sankoff and Rousseau 1975; Felsenstein 1988) have developed various protein parsimony algorithms, taking into account the minimum number of nucleotide substitutions between any pair of amino acids and using the sum of these numbers to compute the tree length. Felsenstein's program PROTPARS in PHYLIP uses one of these algorithms (Felsenstein 1995). However, these algorithms are quite elaborate and depend on a number of simplifying assumptions. Therefore, it is not clear whether they are superior to Eck and Dayhoff's original version. Russo et al.'s (1996) empirical study suggests that Eck and Dayhoff's method is quite efficient in obtaining the true tree.

When DNA sequence data became available in the 1980s, many investigators started to use them for phylogenetic inference. However, it gradually became clear that the evolutionary pattern of DNA sequences is so complex that they are not necessarily better than protein sequences. One problem with DNA sequences is that the substitution pattern is not the same for all nucleotide positions within codons and the GC content in third nucleotide positions often varies with species. For example, the rate of nucleotide substitution at the third codon positions in hominoid mitochondrial genes is so high that the proportion of different nucleotides reaches the saturation level rather quickly (Ruvolo et al. 1994). For these reasons, protein sequences are now again used for phylogenetic reconstruction, and Eck and Dayhoff's simple MP method often gives better results than DNA MP methods.

Example 7.4. MP and NJ Trees from the Cytochrome *b* Gene

The mitochondrial DNA (mtDNA) in vertebrates contains 13 protein coding genes, and the entire sequence of mtDNA is available for a substantial number of organisms. Russo et al. (1996) chose 11 organisms of which the evolutionary relationships are known from paleontological and morphological data and for which complete mtDNA sequences are available and then examined the ability of each gene to reconstruct the correct phylogeny. Here we consider only the cytochrome *b* gene, which is often used for phylogenetic inference. We first constructed the MP tree using amino acid sequence data for cytochrome *b*, which is composed of 377 amino acids. When the 11 species given in Figure 7.13A were used, there were 121 informative sites and 44 uninformative variable sites. The branch-and-bound search produced a single MP tree, which is presented in Figure 7.13A. The topology of this tree is identical with the biological tree we already know. An interesting observation about this tree is that the opossum, chicken, Xenopus, and fish sequences show considerably shorter branch lengths compared with what one would expect under the

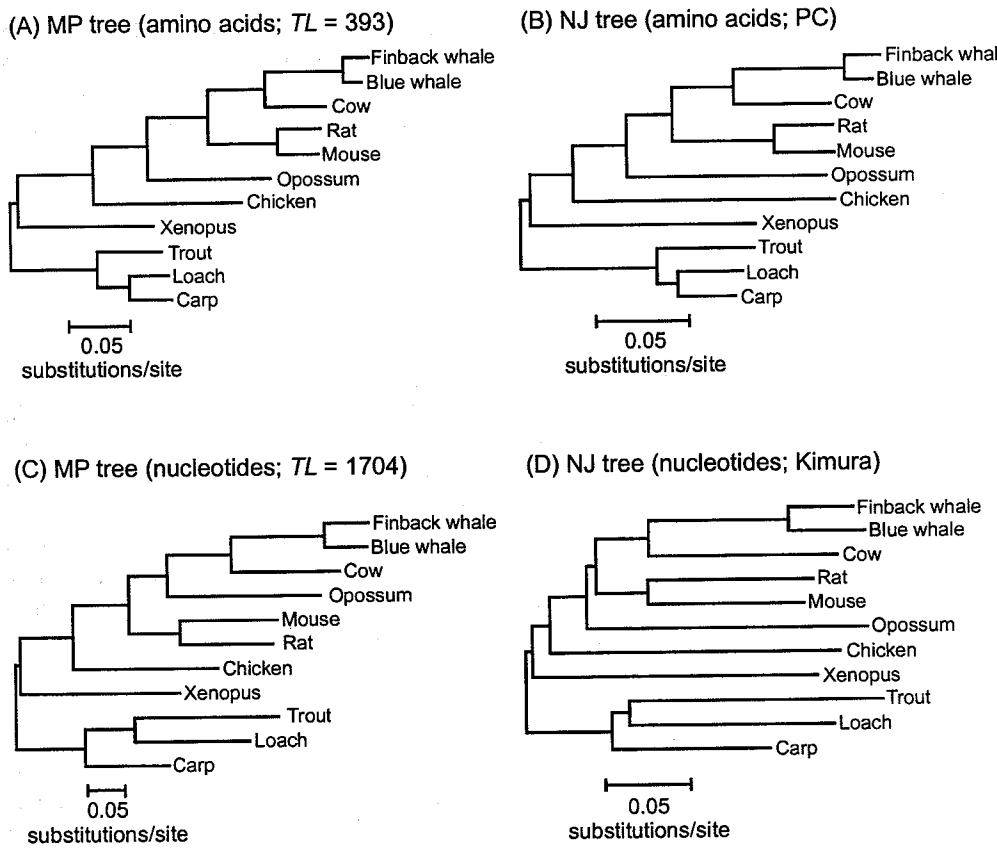


FIGURE 7.13. Inference of a "known" phylogeny of 11 vertebrates using the nucleotide and amino acid sequences of the mitochondrial cytochrome *b* gene. *TL*: Tree length. PC: Poisson correction distance. Kimura: Kimura distance.

molecular clock. Figure 7.13B shows the NJ (and ME) tree obtained by using the Poisson-correction distance. This tree also shows the correct topology, but the branch lengths for the opossum, chicken, and Xenopus are much longer than those of the MP tree.

Figure 7.13C shows the branch-and-bound MP tree for the nucleotide sequence data (1,131 bp long). In this case, there were 494 informative sites and 111 uninformative variable sites. Yet, the topology of the tree is wrong with two branch switches. That is, the opossum should be between chicken and the rodents, and the loach should be closer to the carp rather than to the trout. This indicates that a large number of informative sites alone does not necessarily produce a better tree. The wrong topology of the DNA tree appears to be caused by the fact that nucleotide differences at third codon positions have reached the saturation level, and they introduced noise in phylogenetic construction. However, even when we used only first and second codon position data, the topology was still incorrect with respect to the branching pattern of the three fish species. We also constructed the NJ and ME trees using the Kimura distance for all three codon position data. These trees were identical with

each other but showed one topological error with respect to the three fish species (Figure 7.13D). These results suggest that protein sequences are better than DNA sequences at least in this case. Russo et al. (1996) examined all the 13 protein-coding genes and found that the above conclusion is generally true.

7.7. Shared Derived Characters

Irreversible Shared Derived Characters

If a group of species share a unique and irreversible mutation (mutant character), they must be derived from the same common ancestral species in which this mutation occurred. We call this type of mutations **irreversible shared derived characters.** These characters are very useful for phylogenetic construction (Hennig 1950, 1966). For example, Figure 7.14 shows a phylogenetic tree for four species (1, 2, 3, and 4), in which one mutation ($a \rightarrow a'$, $b \rightarrow b'$, or $c \rightarrow c'$) has occurred in each of the three interior branches. Because these mutations are assumed to be unique and irreversible and each mutation defines a **clade** (a monophyletic group of species), they define the tree unambiguously. Furthermore, since the mutations are directional and the ancestral characters (*a*, *b*, and *c*) are known, we can infer a rooted tree without outgroup species. When the number of species is small, the phylogenetic tree can be easily determined from the distribution of character states among the species. However, when the number of species and the number of characters used are large, the topology and the assignment of mutations for each branch can be cumbersome. In this case, we can use the computer program incorporated in PAUP*.

In **cladistic parsimony,** where the clarification of the evolutionary changes of characters in the phylogeny is emphasized, only shared derived characters or **synapomorphies** are considered to be useful for con-

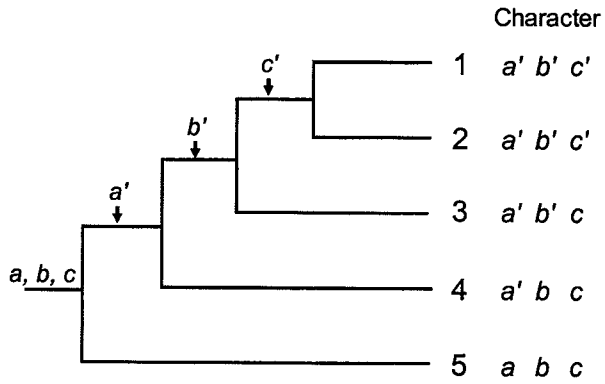


FIGURE 7.14. Phylogenetic tree for five species (1, 2, 3, 4, and 5), which is determined by irreversible mutations ($a \rightarrow a'$, $b \rightarrow b'$, and $c \rightarrow c'$). *a*, *b*, and *c* are the ancestral characters, and *a'*, *b'*, and *c'* are the derived characters. In this case, the root of the tree can also be determined.

structing phylogenetic trees (Henning 1966; Eldredge and Cracraft 1980; Sober 1988; Wiley et al. 1991). However, since the evolutionary changes of morphological characters and the nucleotide substitutions in DNA sequences are usually reversible, most parsimony analyses allow the reversibility of character states. Exceptions are Henning's (1950) strict cladistic analysis and Camin and Sokal's (1965) parsimony. The latter method is different from the Hennigian parsimony in that the same mutation may occur independently in different evolutionary lineages. In the past, however, these methods have rarely been used in actual data analysis because of the apparently unrealistic assumption of irreversibility (Wiley et al. 1991; Swofford et al. 1996).

SINEs and LINEs

However, recent molecular studies have shown that the genomes of higher organisms contain many unique shared derived characters that are apparently irreversible. Among the most well studied are **short interspersed repetitive elements (SINEs)** and **long interspersed repetitive elements (LINEs)** (Singer 1982; Jurka et al. 1988; Britten et al. 1988). SINEs are short sequences of 80–400 nucleotides, whereas LINEs are usually repeats of a few hundred to a few thousand nucleotides. Both SINEs and LINEs are retropseudogenes but are capable of self-replication. Replicated repeat elements are inserted at different locations of the genome, and once they are inserted, they are almost never excised unless they are eliminated by a rare event of large-scale DNA deletion (Hamdi et al.1999; Nikaido et al. 1999). These repeat elements are subject to mutation and minor insertions/deletions and lose their identity in the long run. However, if one is interested in constructing a phylogenetic tree for relatively closely related species (divergence times of up to about 50 million years), these repeat elements can be used as shared derived characters (e.g., Ryan and Dugaiczyk 1989; Okada 1991; Murata et al. 1993; Furano et al. 1994; Verneau et al. 1997).

SINEs and LINEs are identified by appropriate primer DNA sequences, but if they accumulate a substantial number of mutations, the primers may not be able to detect the repeat elements. If this happens for some of the species examined, they are treated as missing characters (Nikaido et al. 1999). When the elements are eliminated by rare deletion events, they are also treated as missing characters. However, since SINEs and LINEs are irreversible mutations, they are still very useful for phylogenetic analysis.

The most well-known family of SINEs is the *Alu* family, which has about 300,000 members in the human and the ape genomes. The members of this family are pseudogenes originally derived from 7SL RNA, one component of the signal recognition particle (Ullu and Tschudi 1984). The SINE families in other organisms are usually pseudogenes derived from various types of t-RNAs rather than 7SL RNA (Kido et al. 1991; Okada et al. 1997). Therefore, there are many different SINE families even in a single species. For example, the salmonid fish have at least three SINE families, and they are confined in this group of fish (Takasaki et al. 1994).

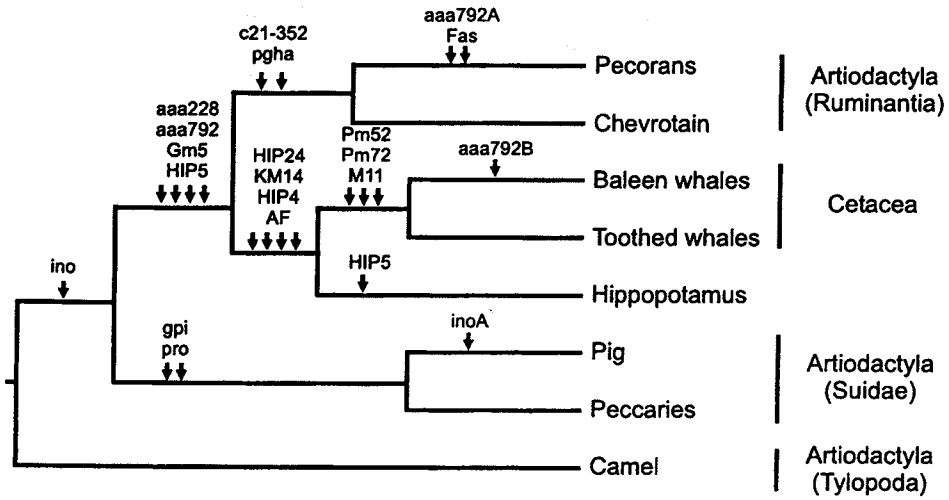


FIGURE 7.15. Evolutionary relationships among cetaceans and artiodactyls as inferred by the presence or absence of 21 different SINE elements. Arrows mark the insertions of SINEs (see Nikaido et al. 1999 for details).

Shimamura et al. (1997) and Nikaido et al. (1999) used mammalian SINE families to study the evolutionary relationships of whales, ruminants, pigs, and camels. The results obtained are presented in Figure 7.15. In this figure, the locations of branches, in which 21 different SINEs were inserted, are presented. As in the case of Figure 7.12, the pattern of insertions of SINEs indicates that whales are a sister group of ruminants and hippopotamuses and are a clade (monophyletic group) within the order Artiodactyla. These results strengthen the conclusion obtained by using the γ-fibrinogen gene (Figure 7.12). They are particularly significant because the SINEs used here are unique and largely irreversible genetic markers, and the tree based on them is unaffected by the error caused by short-branch (or long-branch) attraction (chapter 9). The Hennigian parsimony analysis of the SINE data has shown that this is the most parsimonious tree and that the consistency index (CI) is 1 and the homoplasy index (HI) is 0 (Nidaido et al. 1999). Therefore, the topology given in Figure 7.15 is likely to be correct. SINEs have also been used successfully in clarifying the evolutionary relationships of salmonid species (Murata et al. 1993; Takasaki et al. 1994) and some groups of primate species (Hamdi et al. 1999).

Some might wonder whether the phylogenetic trees based on SINEs are affected by the polymorphism of the presence and absence of SINE insertions in ancestral species (lineage sorting). Theoretically, this polymorphism may generate incongruent phylogenies among different SINE insertions, as in the case of polymorphic DNA sequences in Figure 5.3. This can be seen from the diagrams given in Figure 7.16. In this figure "+" stands for the allele or the genome having a SINE element at a given locus and "−" for the allele lacking it, and the arrow sign indicates the time at which the element was inserted. Here we consider only the cases

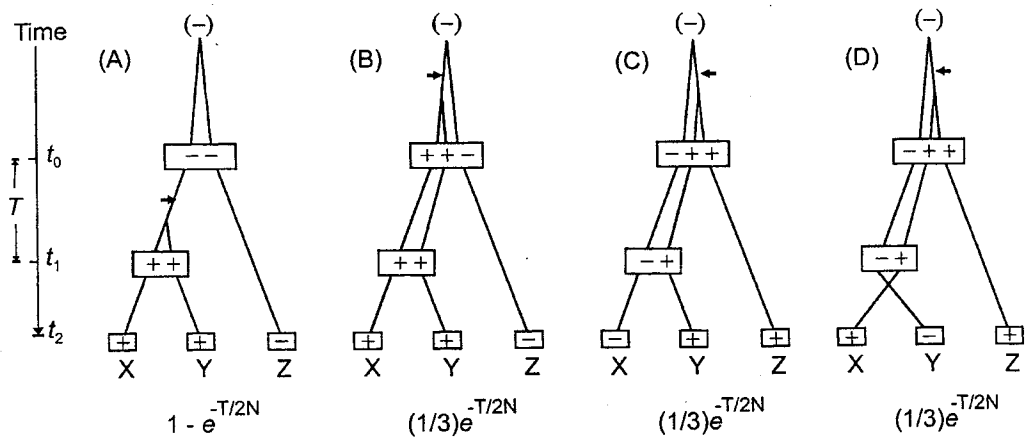


FIGURE 7.16. Four different evolutionary relationships of three species (X, Y, and Z) that may be inferred from SINE element polymorphism (+ and −) exists in ancestral species. The probability of occurrence of each relationship is given underneath the diagram of the relationship. The arrow sign indicates the time of occurrence of SINE insertion. The "−" and "+" are the ancestral and derived character states, respectively. The species tree is the same as that of Figure 5.3.

where two of the three species considered have the SINE element. In diagram (A), the insertion occurred between the times of occurrence of two speciation events (t_0 and t_1), and species X and Y form a clade, which is congruent with the species tree. If we assume that SINE insertion occurs with the same probability for all generations, it can be shown, by using the same mathematical method (coalescence theory) as used by Nei (1987, pp. 401–402), that the probability of occurrence of this event is $1 - \exp(-T/2N)$, where $T = t_1 - t_0$ and N is the effective population size. This probability is the same as that of having evolutionary relationship (A) for DNA sequences in Figure 5.3.

Diagrams (B)–(D) in Figure 7.16 show the cases where the SINE insertion occurred before the first speciation event (time t_0). In this case, the ancestral species can be polymorphic with respect to alleles "+" and "−", and this polymorphism may generate evolutionary relationships that are incongruent with the species tree. Relationship (B) is congruent with the species tree, but relationships (C) and (D) are incongruent. The probability of occurrence of each of these events is given underneath the diagram. Note that if SINE insertion occurs after the second speciation event (time t_1), the SINE element is not informative because it generates a singleton mutation.

Therefore, the effect of ancestral polymorphism of SINE elements is the same as that for DNA sequences discussed in chapter 5, and we have already shown that the probability of occurrence of incongruent relationships is generally very small if T is one million years or greater. (Tachida and Iizuka [1993] studied this problem under different assumptions, but their formulation appears to give an underestimation of the probability of occurrence of incongruent relationships.) Since T is likely to be greater than one million years when different families and genera in mammals

are studied, we can probably dismiss the effect of polymorphism in ancestral species in the case of the phylogenetic tree in Figure 7.15. In fact, all SINE insertions examined in Figure 7.15 are consistent with the topology presented, and there is no indication of incongruent phylogenies for different SINE insertions.

If SINE insertions are irreversible and the effect of ancestral polymorphism is negligible, the tree in Figure 7.15 can be regarded as established without further statistical tests. Application of a bootstrap test (Hillis 1999) to this tree is inappropriate, because every SINE defines a particular clade and exclusion of some SINE insertions (e.g., **ino**) in bootstrap pseudoresamples will make the tree look superficially unreliable. Note that theoretically a single SINE insertion for each interior branch is sufficient to support the topology obtained (Sober 1988).

Of course, this does not mean that no statistical test is needed for actual SINE data analysis. Although homoplasy appears virtually absent in the data of Figure 7.15, it is still too early to exclude homoplasy altogether, because the same SINE insertion may occur independently in different lineages though the probability appears to be very small. In some data sets, the effect of ancestral polymorphism may also generate incongruent phylogenies for different SINEs. In this case, we need some type of statistical test based on the special property of evolution of SINEs. At the present time, we are not sure how to test the reliability of SINE-based trees efficiently, but it is unlikely that the reliability of the tree in Figure 7.15 is questionably low. As a general strategy, it is important to have two or more SINE insertions for each interior branch.

LINEs are longer than SINEs and vary in size rather extensively from copy to copy. Therefore, it is harder to work with LINEs than with SINEs. However, LINEs can be used not only as shared derived characters but also for estimating the time of divergence between species, because LINEs diverge as mutations accumulate and are long enough to give reliable estimates of sequence divergence. Verneau et al. (1997, 1998) used the rodent LINE L1 family (45 rat L1 subfamilies) to clarify the evolutionary relationships and the times of speciation events among 26 rat species of the rodent subfamily Murinae. They showed that these species arose 5–6 million years ago and subsequently underwent different episodes of speciation, the first one occurring about 2.7 million years ago and the second one about 1.2 million years ago.

Other Shared Derived Characters

In addition to SINEs and LINEs, there are a variety of genetic markers that can be used for distinguishing between different groups of organisms. For example, the CMTIA-REP repeat in humans consists of two highly homologous 24 kb sequences (the proximal and the distal CMTIA-REP elements), and some forms of mutation of this repeat result in genetic diseases. Interestingly, this repeat exists only in humans and chimpanzees. However, the distal element of the repeat appears to be present in all primate species but is absent in other mammalian orders (Keller et al. 1999). Apparently, the distal element evolved in the common ancestor of primates, and the proximal element evolved as a result of duplication of the

distal element in the common ancestor of humans and chimpanzees. It is clear that if we find a large number of these shared derived characters they will be very useful for phylogenetic analysis.

Another important class of genetic markers is large-scale insertions or deletions of DNA sequences. For example, the human genome contains a duplication of a large DNA region encompassing about 25 immunoglobulin kappa variable region genes compared with other ape genomes (Zachau 1995). If this DNA duplication had occurred in the ancestor of African apes, it would have been a very useful phylogenetic marker. The class I MHC (HLA) C locus, which is known to be highly polymorphic in humans and chimpanzees, is present only in humans and African apes, so that this locus was apparently generated by gene duplication in the ancestor of these species (Chen et al. 1992). The DY/DI gene clusters in the cattle and pig class II MHC are also apparently confined only to a group of artiodactyl species (Trowsdale 1995). As the genomic structures of many different organisms are studied, we will find many such cladistic characters, and they will be important sources of phylogenetic analysis in the future.

In the study of evolutionary relationships of distantly related organisms, the presence and absence of introns in protein-coding genes will be useful. Although the debate over the intron-early and the intron-late hypotheses is still going on, it is now clear that at least in higher organisms introns are occasionally inserted or deleted, and therefore the presence or absence of introns can be used as cladistic characters. For example, the intron in the protamine P1 gene exists apparently in all mammals but not in other vertebrate species (Rooney et al. 2000). Venkatesh et al. (1999) used information on the presence or absence of introns in seven protein-coding genes for constructing a phylogenetic tree of fish species under the assumption that the probability of independent intron insertion in different lineages is negligibly small. They clarified the difficult-to-ascertain phylogenetic relationships of some ray-finned fishes.

Insertion of an intron is a rare event, and the same intron is almost never inserted twice in the same genomic location. However, the loss of an intron may occur independently in different evolutionary lineages. This property satisfies the conditions required for Farris's (1977) **Dollo parsimony** analysis. In this method, a new mutation (shared derived character) is assumed to be unique, but the loss of the mutation may occur independently in the descendant lineages. Therefore, intron insertion/deletion data can be analyzed by the Dollo parsimony, which is incorporated in PAUP*. However, note that the bootstrap test should not be applied to this type of data, because each intron insertion is a unique and unambiguous event.

8

Phylogenetic Inference: Maximum Likelihood Methods

The idea of using a maximum likelihood (ML) method for phylogenetic inference was first presented by Cavalli-Sforza and Edwards (1967) for gene frequency data, but they encountered a number of problems in implementing the method. Later, considering nucleotide sequence data, Felsenstein (1981) developed an algorithm for constructing a phylogenetic tree by the ML method. Kishino et al. (1990) extended this method to protein sequence data using Dayhoff et al.'s (1978) transition matrix. In ML methods, the likelihood of observing a given set of sequence data for a specific substitution model is maximized for each topology, and the topology that gives the highest maximum likelihood is chosen as the final tree. The parameters to be considered are not the topologies but the branch lengths for each topology, and the likelihood is maximized to estimate branch lengths. In this chapter, we present an outline of these ML methods and discuss some aspects of the theoretical foundation of the methods.

8.1. Computational Procedure of ML Methods

Computation of Likelihood Values

Let us first explain how to compute the likelihood value for a given tree using DNA sequence data. We consider a simple tree of four taxa given in Figure 8.1A and assume that DNA sequences are n nucleotides long and are aligned with no insertions/deletions. We now consider the observed nucleotides for sequences 1, 2, 3, and 4 at a given site (k-th site) and denote them by x_1 x_2, x_3, and x_4, respectively. We do not know the nucleotides at nodes 0, 5, and 6 but assume that they are x_0, x_5, and x_6, respectively. Here x_i takes any of the four nucleotides A, T, C, and G.

Consider a nucleotide site and let $P_{ij}(t)$ be the probability that nucleotide i at time 0 becomes nucleotide j at time t at a given site. Here i and j refer to any of A, T, C, and G. In the ML method, the rate of substitution (r) is allowed to vary from branch to branch, so that it is convenient to measure evolutionary time in terms of the expected number of substitutions ($v = rt$). In the following, we denote the expected number of substitutions for the i-th branch by $v_i \equiv r_i t_i$. In the ML method, the

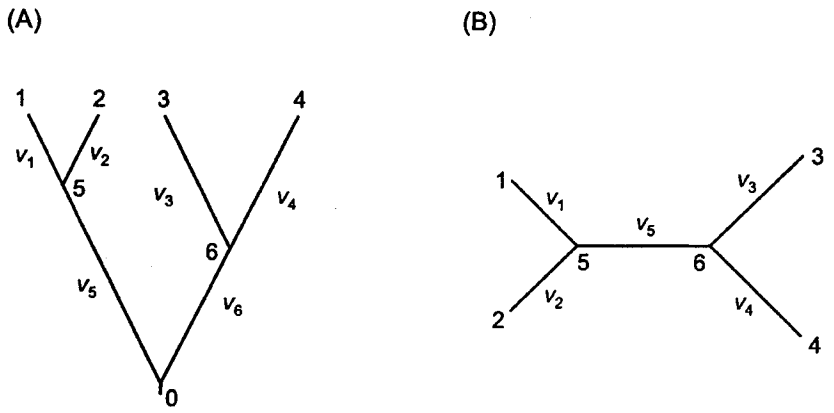


FIGURE 8.1. Rooted and unrooted phylogenetic trees for four taxa to explain the maximum likelihood method of phylogenetic reconstruction. $v_i = r_i t_i$, where r_i is the rate of nucleotide substitution and t_i is the evolutionary time for branch i. In tree B, v_5 represents the sum of v_5 and v_6 in tree A.

branch lengths v_i's are regarded as parameters, and they are estimated by maximizing the likelihood function for a given set of observed nucleotides, as mentioned above. The likelihood function for a nucleotide site (k-th site) is then given by

$$l_k = g_{x_0} P_{x_0 x_5}(v_5) P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_0 x_6}(v_6) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \quad (8.1)$$

where g_{x_0} is the prior probability that node 0 has nucleotide x_0. g_{x_0} is often equated to the relative frequency of nucleotide x_0 in the entire set of sequences, but it may be estimated by the ML method.

To know $P_{ij}(v)$ explicitly, we have to use a specific substitution model. Felsenstein (1981) used the equal-input model mentioned earlier. In this model, $P_{ii}(v)$ and $P_{ij}(v)(i \neq j)$ are given by

$$P_{ii}(v) = g_i + (1 - g_i)e^{-v} \quad (8.2a)$$

$$P_{ij}(v) = g_j(1 - e^{-v}) \quad (8.2b)$$

where g_i is the relative frequency of the i-th nucleotide. When $g_i = 1/4$ and $v = 4rt$, the above equations become identical with those for the Jukes-Cantor model.

In the above formulation, we considered a rooted tree. However, if we use a reversible model of nucleotide substitution for defining $P_{ij}(v)$, there is no need to consider the root (Figure 8.1B). A reversible model means that the process of nucleotide substitutions between time 0 and time t remains the same whether we consider the evolutionary process forward or backward in time. Mathematically, the reversibility condition is given by

$$g_i P_{ij}(v) = g_j P_{ji}(v) \quad (8.3)$$

for all i and j. Equations 8.2 satisfy this condition.

As long as a reversible model is used, the number of nucleotide substitutions ($v_5 + v_6$) between nodes 5 and 6 of tree A remains the same irrespective of the location of the root 0. Therefore, we designate $v_5 + v_6$ in tree A by v_5 in tree B and compute l_k, assuming that evolutionary change starts from some point of the tree. Here we assume that it starts from node 5 for convenience. This simplifies the computation considerably, and we can rewrite Equation (8.1) in the following way.

$$l_k = g_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \tag{8.4}$$

In practice, of course, we do not know x_5 and x_6, so the likelihood will be the sum of the above quantity over all possible nucleotides at nodes 5 and 6. That is,

$$L_k = \sum_{x_5}\sum_{x_6} g_{x_5} P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2) P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4) \tag{8.5a}$$

$$= \sum_{x_5} g_{x_5} [P_{x_5 x_1}(v_1) P_{x_5 x_2}(v_2)] [\sum_{x_6} P_{x_5 x_6}(v_5) P_{x_6 x_3}(v_3) P_{x_6 x_4}(v_4)] \tag{8.5b}$$

So far, we considered only one nucleotide site. In practice, we must consider all nucleotide sites including invariable sites. Since the likelihood (L) for the entire sequence is the product of L_k's for all sites, the log likelihood of the entire tree becomes

$$\ln L = \sum_{k=1}^{n} \ln L_k \tag{8.6}$$

We can now maximize $\ln L$ by changing parameters v_i's. This computation is done numerically by using Newton's method or some other numerical computation. This maximization gives ML estimates of branch lengths (v_i's) for this topology, but we are also interested in the maximum likelihood value for this topology and record it. We now consider the two remaining topologies that are possible for four sequences and compute the ML values for them. The ML tree is the topology that has the highest ML value. The branch lengths for this topology are of course given by the ML estimates of v_i's obtained for this topology.

From this example, it is clear that the construction of an ML tree is very time-consuming because we have to consider all possible nucleotides at each interior node. The number of nucleotide combinations to be examined for a tree of m taxa is given by $4^{(m-2)}$, because there are $m - 2$ interior nodes. For example, if $m = 10$, we must examine 65,536 different combinations of nucleotides. It is possible to reduce the computational time considerably if we rewrite Equation (8.5a) in the form of Equation (8.5b). This treatment is called "pruning" (Felsenstein 1981). In addition, there are several algorithms to speed up the computation of ML values (Adachi and Hasegawa 1996b). However, even if these algorithms are used, the amount of computation is substantial, particularly when the extent of sequence divergence is high. In addition, the number of topologies to be examined increases rapidly with m, as mentioned earlier. This num-

ber is 2,027,025 for $m = 10$. Therefore, it is not very easy to construct a true (global) ML tree when m is large. For this reason, various heuristic search algorithms have been developed, as in the case of MP methods.

In the above formulation, we considered a simple model of nucleotide substitution. However, essentially the same formulation can be used for any kind of substitution model as long as the model is time-reversible (all substitution models except model H in Table 3.2). In general, the likelihood function L for a topology may be written as

$$L = f(\mathbf{x}; \theta) \tag{8.7}$$

where $\mathbf{x}$ is a set of observed nucleotide sequences and θ is a set of parameters such as branch lengths, nucleotide frequencies, and substitution parameters in the mathematical model used (see section 8.2). In the equal-input model, the substitution rate parameter (r_i) is combined with time t_i ($v_i = r_i t_i$), and v_i's are treated as branch length parameters. Therefore, the number of parameters to be estimated are $2m - 2$ if the nucleotide frequencies are estimated from the observed frequencies. If the nucleotide frequencies are estimated by the ML method, we need three additional parameters with the condition of $g_A + g_T + g_C + g_G = 1$. If we use the Hasegawa-Kishino-Yano model (Table 3.2), we need one additional free parameter (α/β) to be estimated. All these parameters can be estimated by maximizing L for a given set of observed data.

As mentioned above, the maximization of $\ln L$ is done numerically, and thus the actual ML value obtained depends on the numerical method used and the extent of accuracy one wishes to have (Tateno et al. 1994). Therefore, different computer programs may give different ML values even for the same set of data. Nevertheless, the relative ML values for different topologies usually remain the same as long as the number of sequences used is small. When a large number of sequences are used, the differences in ML value between different topologies can be small (see Example 8.2), and therefore the accuracy of the method for computing ML values becomes important. Steel (1994) has shown that even for the same topology for four sequences two peaks may arise on the likelihood surface and warned that this may become a real problem in finding ML trees. However, Rogers and Swofford's (1999) computer simulation suggests that this is not a serious problem in actual data analysis as long as the number of sequences used is relatively small. By contrast, Olsen et al. (1994) state that when a large number of sequences are analyzed by ML methods, the existence of multiple peaks becomes a problem. More theoretical study on this problem is necessary.

Search Strategies for ML Trees

Since the search for an ML tree is very time-consuming, various heuristic methods for finding the ML tree have been proposed. Many of them are similar to those (e.g., NNI and TBR) used for obtaining ME or MP trees, and there is no need to repeat them here. However, the efficiencies of these algorithms in obtaining the correct topology are not necessarily the same for the ME, MP, and ML methods.

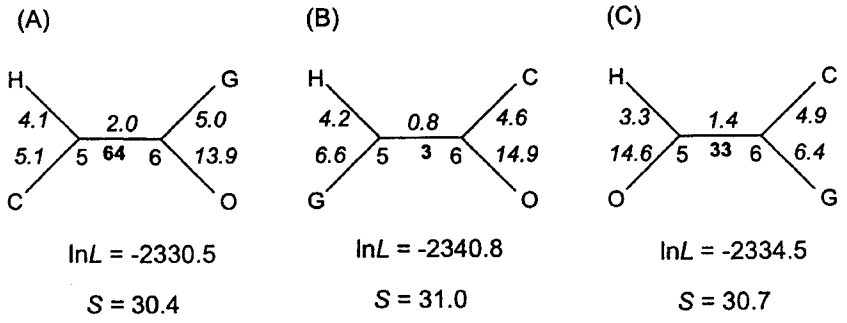


FIGURE 8.2. Unrooted likelihood trees for humans (H), chimpanzees (C), gorillas (G), and orangutans (O). Numbers 5 and 6 stand for the two interior nodes. Branch length estimates given in italics are the number of nucleotide substitutions per 100 sites obtained by using the equal-input model. Bootstrap values are in boldface. *S* is the sum of branch lengths obtained by the ordinary LS method.

For example, the **star-decomposition (SD) method** proposed by Saitou (1988) and Adachi and Hasegawa (1996b) is conceptually the same as the neighbor-joining method and starts with a star tree as shown in Figure 6.7A. The star tree is decomposed into a bifurcating tree step-by-step as in the case of the NJ algorithm by computing the ML value at each step of taxon pairing and by choosing a pair of neighbors that give the highest ML value. In this way, one can obtain an SD ML tree. In likelihood analysis, however, the SD method does not seem to be as efficient as some other heuristic methods (e.g., TBR search) in obtaining the correct topology. For this reason, Nei et al. (1998) suggested that the application of the CNI search (see section 6.3) to the SD ML tree with a few cycles of iterations might find the true topology as often as the extensive ML tree search algorithm. Similarly, a recent computer simulation (Takahashi and Nei 2000) suggested that the stepwise addition + NNI search (see section 7.2) is usually as efficient as the more extensive TBR search in finding the true tree. Like ME and MP methods, ML methods tend to give incorrect topologies when m is large and n is small (chapter 9). Therefore, it is unwise to spend an excessive computer time to search for the ML tree. What is important is to find the true tree or a tree close to it rather than the ML tree.

Example 8.1. ML Tree for Hominoid Species

Let us again consider the DNA sequences in Figure 6.1. For simplicity, we use only the human, chimpanzee, gorilla, and orangutan sequences and denote them by 1, 2, 3, and 4, respectively. In this case, there are three different topologies, A, B, and C, as shown in Figure 8.2. The average nucleotide frequencies of A, T, C, and G for the four sequences are 0.310, 0.254, 0.331, and 0.105, respectively. The nucleotides at the first position of these sequences are all A (Figure 6.1), but we have to consider all the four possible nucleotides at each of nodes 5 and 6. Therefore, if we use node 5 as the starting point, the likelihood of having the observed nucleotides is obtained by Equation (8.5). It becomes

$$
\begin{aligned}
L_1 = {} & g_A P_{AA}(v_1) P_{AA}(v_2) P_{AA}(v_5) P_{AA}(v_3) P_{AA}(v_4) \\
& + g_T P_{TA}(v_1) P_{TA}(v_2) P_{TA}(v_5) P_{AA}(v_3) P_{AA}(v_4) \\
& + g_C P_{CA}(v_1) P_{CA}(v_2) P_{CA}(v_5) P_{AA}(v_3) P_{AA}(v_4) \\
& + g_G P_{GA}(v_1) P_{GA}(v_2) P_{GA}(v_5) P_{AA}(v_3) P_{AA}(v_4) \\
& + \cdots \\
& + g_G P_{GA}(v_1) P_{GA}(v_2) P_{GG}(v_5) P_{GA}(v_3) P_{GA}(v_4)
\end{aligned}
$$

The total number of terms in this equation is 16, because both nodes 5 and 6 can take four different nucleotides. The second nucleotide site also has A for all the sequences, so L_2 is the same as the L_1. The third nucleotide site has G for all the sequences. Therefore, L_3 is written as

$$
\begin{aligned}
L_3 = {} & g_A P_{AG}(v_1) P_{AG}(v_2) P_{AA}(v_5) P_{AG}(v_3) P_{AG}(v_4) \\
& + \cdots \\
& + g_G P_{GG}(v_1) P_{GG}(v_2) P_{GG}(v_5) P_{GG}(v_3) P_{GG}(v_4)
\end{aligned}
$$

This equation also has 16 terms.

If we do this type of computation for all remaining 892 sites, we can now obtain $\ln L$ by $\ln L_1 + \ln L_2 + \ldots + \ln L_{895}$. (Nucleotide site 560 has been eliminated because the orangutan sequence contains a deletion.) The next step of the ML method is to assume $g_A = 0.310$, $g_T = 0.254$, $g_C = 0.331$, and $g_G = 0.105$ and maximize $\ln L$ by varying $v_1, v_2, \ldots,$ and v_5. The v_i's that maximize $\ln L$ are the ML estimates of branch lengths of tree A in Figure 8.2. They become $\hat{v}_1 = 0.041$, $\hat{v}_2 = 0.050$, $\hat{v}_3 = 0.050$, $\hat{v}_4 = 0.139$, and $\hat{v}_5 = 0.020$, and the maximum log likelihood (ML) value for this tree is -2330.5. To determine the ML tree, we have to do the same computation for trees B and C in Figure 8.2 and obtain $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_5$ and the ML value. The results of this computation are given in Figure 8.2, and they show that tree A has the highest ML value and therefore is the ML tree for the data set used here.

In the above example, we did not include the gibbon sequence, because the computation becomes more laborious. In the case of five taxa, there are three interior nodes, and each of these nodes has four possible nucleotides. Therefore, there are $4^3 = 64$ different evolutionary pathways when the nucleotides at the exterior nodes are given. The equation for L_k will then have 64 different terms. Note also that the number of topologies to be examined is now 15 instead of 3. Therefore, the computational burden rapidly increases as the number of taxa increases.

8.2. Models of Nucleotide Substitution

Commonly Used Models

The substitution model given in Equation (8.2) is a simple one and does not take into account various complicating factors such as the transition/transversion bias. Felsenstein (1984) introduced a new substitu-

tion model without clearly specifying the mathematical model, but it takes into account the transition/transversion bias. The mathematical formulation of this model and its statistical properties are discussed by Kishino and Hasegawa (1989) and Tateno et al. (1994). Hasegawa, Kishino, and Yano (1985b) also developed a closely related model. Their model (HKY model) is given by matrix E in Table 3.2. An element, e_{ij}, of this matrix represents the instantaneous rate of substitution from nucleotide i(i-th row) to j (j-th column) (i, j = A, T, C, G). All the elements in each row sum up to 0, so that the diagonal element $e_{ii} = -\Sigma_j e_{ij}$ ($i \neq j$), though this is not presented. The transition and transversion rates from nucleotide i to j are αg_j and βg_j, respectively. Therefore the transition/transversion ratio is given by $R = \alpha/(2\beta)$. This model becomes identical with the equal-input model when $\alpha = \beta$ and with the Kimura model when $g_A = g_T = g_C = g_G = 1/4$.

Felsenstein's (1984) model as described by Kishino and Hasegawa can be written in the following way:

	A	T	C	G
A		βg_T	βg_C	$(\delta/g_R + \beta)g_A$
T	βg_A		$(\delta/g_Y + \beta)g_C$	βg_G
C	βg_A	$(\delta/g_Y + \beta)g_T$		βg_G
G	$(\delta/g_R + \beta)g_A$	βg_T	βg_C	

Here, g_Y and g_R are the relative frequencies of pyrimidines (T and C) and purines (A and G), that is, $g_Y = g_T + g_C$ and $g_R = g_A + g_G$. The parameter β stands for the transversional rate, and δ/g_Y and δ/g_R are parameters that measure the amount of transitional change that exceeds β. In this model, the transition/transversion ratio is given by

$$R = (a_1\delta/\beta + a_2)/a_3 \tag{8.8}$$

where $a_1 = g_T g_C/g_Y + g_A g_G/g_R$, $a_2 = g_T g_C + g_A g_G$, and $a_3 = g_Y g_R$ (Tateno et al. 1994). Note that his model reduces to Kimura's (1980) when g_i = 0.25 and $R = (\delta/\beta + 0.5) = \alpha/(2\beta)$.

In Felsenstein's DNAML program of PHYLIP (Felsenstein 1995), g_i is estimated by the observed frequency in the entire sequences, and R is arbitrarily assigned. Therefore, the only parameters to be estimated by maximizing the likelihood are branch lengths. By contrast, in Adachi and Hasegawa's nucML program of MOLPHY (Adachi and Hasegawa 1996b), the ratio α/β in the HKY model is estimated in addition to branch lengths, as mentioned earlier. Analytical formulas for estimating the evolutionary distance using the Felsenstein and the HKY models are presented by Tateno et al. (1994) and Rzhetsky and Nei (1995), respectively.

Recent computer programs such as PAML (Yang 1999) and PAUP* (Swofford 1998) include various substitution models such as the Jukes-

Cantor, Kimura, and Tamura-Nei models in addition to the above two models. They also include the *general reversible* (REV) model given in Table 3.2G. This is the most general model that satisfies the reversibility condition (Equation [8.3]) and includes eight independent parameters. Three of them refer to the nucleotide frequencies (g_i's with the condition $\Sigma g_i = 1$), but these parameters are often estimated by the observed frequencies. The five parameters, *a, b, c, d,* and *e* have to be estimated by maximizing the likelihood. (f can be assumed to be 1.)

As was mentioned in chapter 6, the rate of nucleotide substitution varies extensively from site to site and approximately follows the gamma distribution. Yang (1993, 1994a) incorporated this feature into the ML method of phylogenetic inference, and the computational algorithm taking into account this feature is included in PAML. Felsenstein and Churchill (1996) used a hidden Markov model to deal with rate variation among different sites, which is expected to give results similar to those by Yang's (1994b) discrete gamma model. However, according to a numerical example presented by Yang et al. (1994), the power of discrimination between different topologies declined when a gamma distribution model was used.

Comparison of Different Models

The actual pattern of nucleotide substitution is obviously very complicated, so one might think that a mathematical model with many parameters is better than a model with fewer parameters for constructing phylogenetic trees. In practice, this is not always the case. A model with many parameters fits the data better than a simpler model, but statistical prediction (or topology estimation) based on a model with many parameters is subject to more errors. It is therefore advisable to use a simple model as long as the model represents the substitution pattern reasonably well.

In the case of ML methods, the goodness of fit of a model to observed data can be examined by using the likelihood ratio test or Akaike's (1974) information criterion (*AIC*). When there are two models, models 1 and 2, and model 1 is a special case of model 2, model 1 is said to be nested in model 2. When the correct topology is known and model 1 is nested in model 2, one can compute the log likelihood ratio by using Equation (4.13), where $\ln L_1$ and $\ln L_2$ are the ML values for models 1 and 2, respectively. Therefore, we can test whether model 2 is significantly better than model 1 or not. For example, the Kimura model is a special case of the HKY model, and the former has one free parameter and the latter four free parameters. (Note that $g_A + g_T + g_C + g_G = 1$.) Therefore, the difference in goodness of fit between the two models can be evaluated by the *LR* or χ^2 test with three degrees of freedom.

In general, the likelihood ratio test cannot be used unless the two models compared are nested. However, it is possible to compare two nonnested models by using *AIC* as long as the topology considered remains the same. *AIC* is defined as

$$AIC = -2\ln L + 2p \tag{8.9}$$

where $\ln L$ is the log likelihood value for a given model and p is the number of free parameters to be estimated. For example, the equal-input model with three free parameters and the Tamura model with two free parameters (Table 3.2) are not nested, but these two models can be compared by *AIC* if we consider the same topology. It has been proposed that the statistical predictability of a model is higher when *AIC* is low than when it is high. Equation (8.9) indicates that even if $-\ln L$ is low, *AIC* can be high when the number of free parameters is large and that a model with a high ML value with a small number of parameters is better.

Although the above statement is generally true, some caution is necessary in the application of *AIC* to phylogenetic analysis. In actual data analysis, the *AIC* value is almost always lower when a sophisticated model (such as the HKY model) is used than when a simple model (such as the Kimura model) is used (see Example 8.2). However, a complex model with a low *AIC* value does not necessarily produce better topologies than a simpler model. In fact, empirical studies with known phylogenies have shown that the *AIC* value has virtually no correlation with the probability of obtaining the correct topology (Russo et al. 1996). Theoretical studies have also indicated that a sophisticated model does not necessarily give the correct topology with a higher probability than a simple model (Gaut and Lewis 1995; Yang 1997).

To study this problem in more detail, Takahashi and Nei (2000) conducted a computer simulation generating artificial extant sequences using the HKY gamma model ($a = 0.5$). The number of sequences used was 48, each sequence consisting of 1000 nucleotides, and the topologies of the model trees were determined by using the branching process. Maximum likelihood trees were then constructed by using the simple Jukes-Cantor model as well as the original (true) HKY gamma model. These ML trees were obtained by using the SA + NNI algorithm incorporated in PAUP*, and the difference between the ML and the true trees (topologies) was measured by the average topological distance ($\bar{d}_T$) for 50 replicate simulations. The results of this simulation showed that the topology of the ML tree obtained is generally closer to that of the true tree when the incorrect Jukes-Cantor model is used than when the true HKY gamma model is used, even though the ML value for the Jukes-Cantor model was always much lower than that for the HKY gamma model. This indicates that when the number of sequences is large a simple model usually gives better results than a complex model as long as the sequence length is relatively short. This situation is similar to the case of NJ and ME methods, where the p-distance often gives better trees than more sophisticated distances (see chapter 6).

This study raises a serious question about the current practice of choosing a substitution model by the likelihood ratio test or by *AIC*. A more careful study is necessary about realistic methods of choosing appropriate substitution model in ML analysis. Recently, a number of authors (e.g., McArthur and Koop 1999; Silberman et al. 1999; Zardoya et al. 1999) have used the HKY gamma or more general models in ML analysis apparently because these models include various special cases such as the Jukes-Cantor model. However, very general models often give poor phylogenetic inference, and this practice should be discouraged.

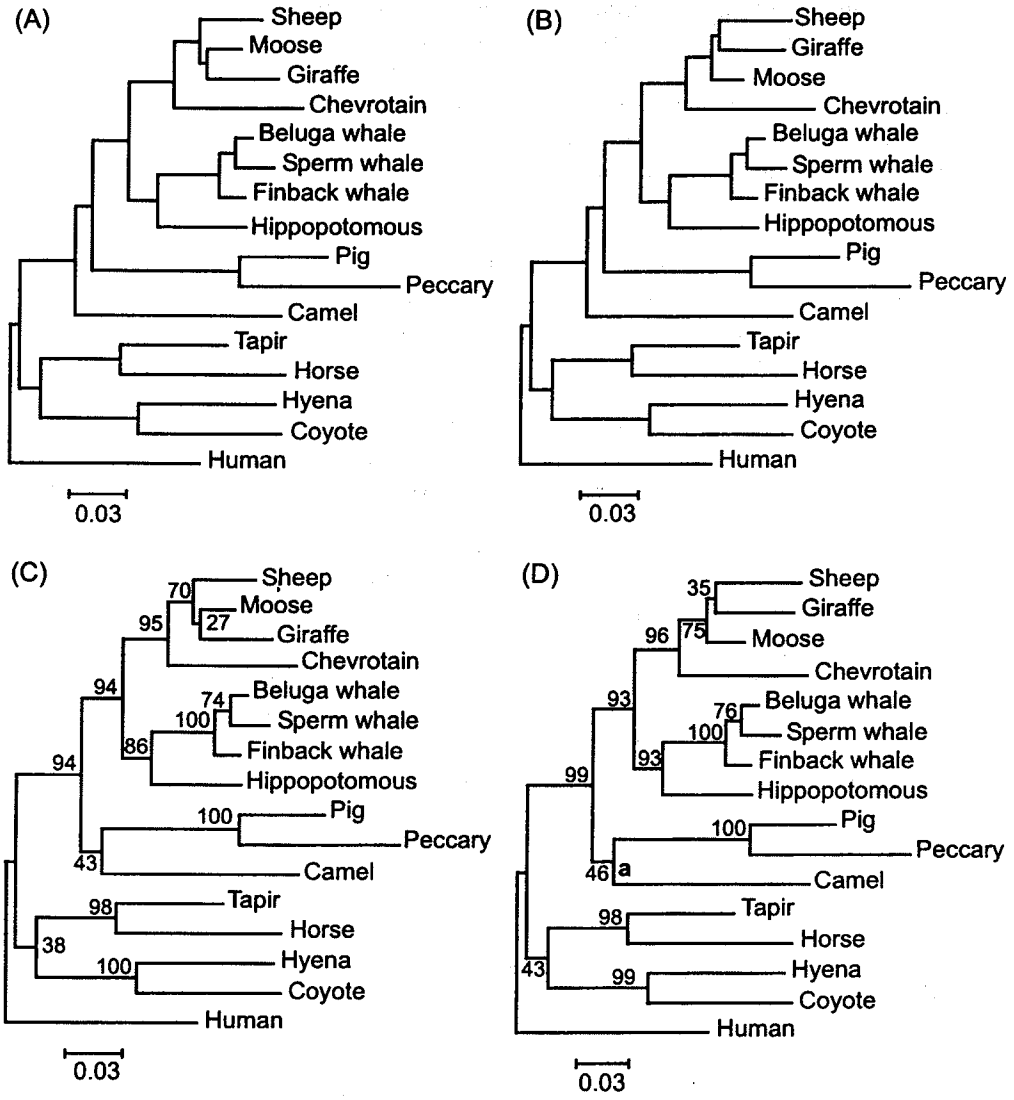


FIGURE 8.3. Maximum likelihood trees for gamma-fibrinogen gene data. (A) Star-decomposition ML tree with the Kimura model. (B) SD + TBR search ML tree with the Kimura model. (C) ML bootstrap consensus tree with the Kimura model (fast stepwise-addition heuristic search; 500 replications). (D) The ML bootstrap consensus tree with the HKY model (fast stepwise-addition heuristic search; 100 bootstrap replications).

Example 8.2. ML Trees for Whales and Their Related Species

We have discussed various substitution models, which may affect the ML tree obtained. Let us now examine this problem using the actual sequence data of the γ-fibrinogen gene from cetartiodactyls utilized in Example 7.3. Here we use the computer programs in PAUP* unless otherwise mentioned. This data set gave a transition/transversion ratio ($\hat{R}$) of about 1.7, so we first used the Kimura model to obtain the ML tree. The

tree obtained by the SD algorithm is presented in Figure 8.3A. The topology of the tree is different from that of the MP and NJ trees in Figure 7.12, and the ML value (−2951.61) for the SD ML tree is slightly lower than that (−2951.52) for the NJ tree (Tree D). When we subjected the SD tree (tree A) to the TBR branch swapping (1473 rearrangements), we obtained a tree that appeared to be the true ML tree (Figure 8.3B). This indicates that the SD algorithm could not find the ML tree, but since the ML value (−2951.41) for this tree is only slightly higher than that for the SD tree, it is not clear whether this tree is better than tree A or not (see Table 8.1). Using PAUP*, we also constructed a bootstrap consensus tree, which is given in Figure 8.3C. The topology of this tree is slightly different from the NJ tree, and the ML value (−2951.67) of this tree is also lower than that for the ML tree.

We then used the HKY and the REV models to construct ML trees using the SD + TBR search. For the HKY model, we first used the observed nucleotide frequencies leaving one free parameter for the substitution model (HKY-f). This time the topology of the tree (identical with Tree D) was the same as that of the NJ tree, but the ML value (−2955.87) was lower than that for the Kimura model (Table 8.1). This looks strange, because there was one free parameter used for both substitution models and in the HKY model additional information about nucleotide frequencies was given. However, it is possible that the observed nucleotide frequencies ($g_A = 0.2888$, $g_T = 0.3063$, $g_C = 0.1886$, and $g_G = 0.2182$) are inappropriate in this case. We therefore estimated these frequencies by the ML method and then searched for the ML tree (HKY-MLf). (The estimated nucleotide frequencies were $g_A = 0.2740$, $g_T = 0.2582$, $g_C = 0.2144$, and $g_G = 0.2534$.) This search produced the same topology as the previous one, but the ML value (−2947.84) was now higher than that for the Kimura model. The log likelihood ratio for the comparison of the two models becomes $LR = 2(-2947.8 + 2951.5) = 7.4$ when the tree (Tree D) for the HKY-MLf model is used. Since the Kimura model is a special case of the HKY model and the latter model has three additional free parameters, LR is approximately distributed as a χ^2 with three degrees of freedoms. This χ^2 test shows that the HKY-MLf model is not significantly

Table 8.1 Log likelihoods and ΔAICs of the three trees for different substitution models.

				$\Delta \ln L^a$			
Model	p^b	$\ln L_{max}$	ΔAIC^c	Tree A	Tree B	Tree C	Tree D
Jukes-Cantor	0	−3032.21	174.2	−0.61	−0.40	−0.15	ML
Kimura	1	−2951.41	14.6	−0.20	ML	−0.26	−0.10
HKY-f	1	−2955.87	23.5	−0.63	−0.30	−0.28	ML
HKY-MLf	4	−2947.84	13.5	−0.31	−0.10	−0.17	ML
REV-f	5	−2940.38	0.6	−0.51	−0.06	−0.42	ML
REV-MLf	8	−2937.10	0.0	−0.40	ML	−0.47	−0.11

[a] $\Delta \ln L = \ln L - \ln L_{max}$.
[b] p = number of free parameters in the substitution model.
[c] $\Delta AIC = AIC_{\ln L_{max}} - AIC_{REV,MLf}$.

better than the Kimura model. Interestingly, when we constructed a bootstrap consensus tree using the HKY-MLf model with four free parameters (100 replications with the SD + NNI search), we obtained tree D. However, the bootstrap value for branch *a* was only 46%, so it is not clear whether this tree is better than trees, A, B, and C.

When we used the REV-MLf model with the TBR search (1419 rearrangements) starting with an initial tree obtained by the NJ procedure, we obtained an ML tree whose topology was the same as that of tree B. The ML value for this tree was −2937.10, which is higher than the value for the tree obtained by the HKY model. The *AIC* value is also smallest. However, the difference in the ML value among the four trees are quite small.

As mentioned earlier, we cannot use the *LR* test for comparing different topologies, but it is obvious that if the differences in the ML value are small, the power of the ML method in discriminating between different topologies is certainly small. It should also be noted that trees A–D in Table 8.1 have only minor topological differences and that all the interior branches associated with the topological differences have low bootstrap values. Therefore, it is not worth spending too much time to find the ML tree with sophisticated models. Note that the Jukes-Cantor model gives the same ML tree as that of the HKY-f and the REV-f models, though the ML value for this model is significantly lower than that for the latter models.

The above example shows that the relationship between the ML value for a substitution model and the inferred tree is quite complicated and illustrates the difficulty in choosing an appropriate model in the ML analysis.

Example 8.3. ML Trees Obtained for Simulated Sequence Data

The above results indicate that estimation of ML trees becomes increasingly difficult as the number of parameters increases. However, the bootstrap consensus ML tree is usually very similar to the bootstrap consensus NJ and MP trees unless the sequence divergence is high and the evolutionary rate varies extensively from branch to branch. To see whether this is the case with the DNA sequence data used in Figure 6.8, we constructed an ML tree for the 24 sequences using PAUP*. As mentioned earlier, this data set was obtained by computer simulation using the Kimura model with $R = 5$. We therefore used the Kimura model with an estimated R value of 5.5 to generate the ML bootstrap consensus tree, which is presented in Figure 8.4. (The number of bootstrap replications was 100.) This ML tree is essentially the same as that of the NJ tree (Figure 6.8C) or the MP tree (Figure 7.11A) if we ignore the minor differences in branching pattern that are not supported by the bootstrap test. The branch length estimates of the correctly identified branches are also nearly the same as those of the NJ tree, though the latter tree was constructed by using the Jukes-Cantor model. When we constructed the ML tree with the Jukes-Cantor model, the topology of the tree was identical with that of the tree with the Kimura model, but the *LR* test showed that the Kimura model with the estimated R value is significantly better than

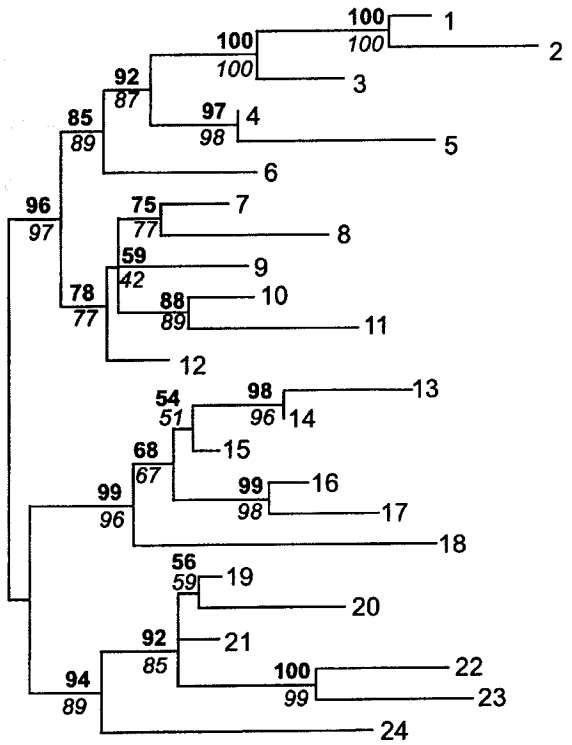


FIGURE 8.4. Maximum likelihood tree for simulated sequence data as inferred by the NJ + NNI search with the Kimura model in PAUP*. Bootstrap values (100 replications) are shown in boldface for the Kimura model and in italics for the Jukes-Cantor model.

Jukes-Cantor model ($\chi^2 = 2\,[-1811.9 + 1918.3] = 210.8$ with one degree of freedom), suggesting that the Kimura model gives better estimates of branch lengths. However, as far as the topology and the bootstrap values are concerned, the two trees were virtually identical (see book website http://www.oup-usa.org/sc/0195135857).

8.3. Protein Likelihood Methods

When the DNA sequences are relatively closely related to one another, DNA likelihood methods seem to work well, particularly when the data for the first and second codon positions are used. However, if they are distantly related protein-coding genes, many complications arise, because the rate of synonymous substitution is generally much higher than that of nonsynonymous substitution. The relative frequencies of the four nucleotides at third codon positions can also vary considerably with species (e.g., Kumar 1996b; Moriyama and Powell 1998), suggesting that the stationary model of nucleotide substitution is not necessarily appropriate. By contrast, the evolutionary change of protein sequences does not suffer very much from these problems and seems to be much simpler

than that of DNA sequences when sequence divergence is relatively high. Noting this property, Kishino et al. (1990) proposed a protein-likelihood method, in which Dayhoff et al.'s (1978) empirical transition matrix for 20 different amino acids is used. Later, Adachi and Hasegawa (1996b) used various transition matrices including the Poisson model, Jones et al.'s (1992) empirical transition matrix for nuclear proteins, and their own matrix for mitochondrial proteins. They applied these methods to various sequence data and obtained reasonably good trees for several groups of vertebrate organisms (Cao et al. 1994a, 1994b).

Algorithms

The basic algorithm for protein ML methods is the same as that for DNA ML methods, but we need a 20×20 matrix of transition probabilities, $P_{ij}(v)$, because there are 20 different amino acids. Kishino et al. (1990) used Dayhoff et al.'s (1978) empirical transition matrix, which was discussed in chapter 2. An element of this matrix is the probability of change of amino acid i to amino acid j for a given evolutionary time or branch length v. Therefore, the likelihood function can be written in the same form as that in Equation (8.7), and we can estimate v_i's by maximizing the likelihood and evaluate the ML value for each topology. The same computation can be done by using Jones et al.'s substitution matrix.

Both Dayhoff et al.'s (1978) (Dayhoff) and Jones et al.'s (1992) (JTT) models are based on the assumption that the amino acid frequencies are in equilibrium and remain the same throughout the evolutionary process. In reality, a given set of protein sequences may have amino acid frequencies different from those for the Dayhoff or the JTT model. Therefore, Adachi and Hasegawa (1996b) suggested that a new Dayhoff or a new JTT model be constructed by incorporating the amino acid frequencies from the data set to be analyzed. These modified Dayhoff and JTT models are called the Dayhoff-*f* and the JTT-*f* models, respectively. Because the amino acid frequencies come from the data set to be analyzed, the Dayhoff-*f* and the JTT-*f* models obviously fit the data better than do the original models. A similar modification (Poisson-*f*) can be made for the Poisson model as well. Among all these models, the JTT-*f* model seems to have the best fit to the data in nuclear proteins (Cao et al. 1994b; Yang 1995b). For the proteins encoded by mitochondrial DNA genes, Adachi and Hasegawa (1996a) and Yang (1999) produced additional substitution models.

Example 8.4. Protein and DNA ML Trees for Vertebrate Mitochondrial Co1 Genes

Mitochondrial genes have been used extensively for studying the evolutionary relationship of vertebrate organisms. Using 11 vertebrate species whose evolutionary relationships are known from paleontological and morphological data, Russo et al. (1996) studied the accuracy of phylogenetic trees reconstructed from various mitochondrial genes. Figure 8.5 shows one example, in which phylogenetic trees were constructed by using the cytochrome oxidase 1 (Co1) gene. The number of amino acids

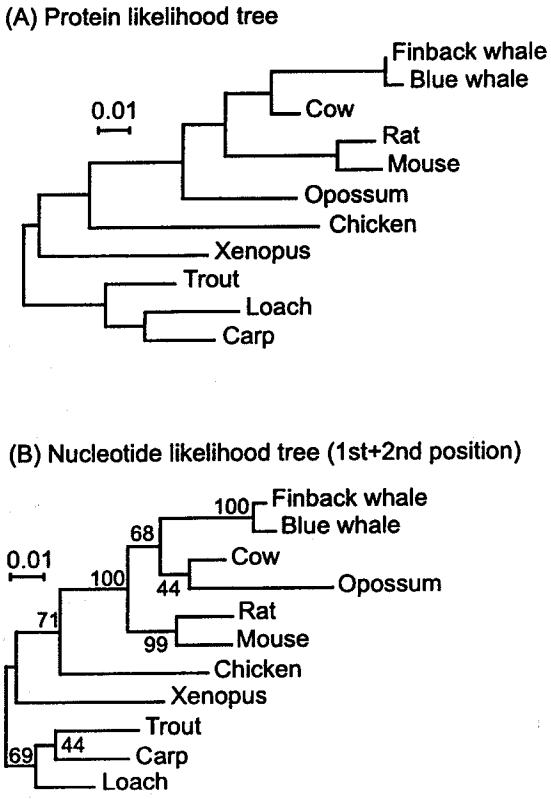


FIGURE 8.5. Maximum likelihood trees for 11 sequences of the mitochondrial cytochrome oxidase subunit 1 gene. (A) Protein likelihood tree with the Poisson model. (B) Nucleotide ML bootstrap consensus tree (1st + 2nd codon position data) with the Kimura model ($R = 2$; 500 bootstrap replications of fast-stepwise heuristic search; $\ln L = -3452.06$). For the true tree (tree A) $\ln L = -3465.46$.

(or codons) used in this case was 511. Tree A was obtained by using the specific-tree search algorithm of Adachi and Hasegawa's (1996b) ProtML program in MOLPHY, whereas tree B was found by PAUP* using the Kimura model with $R = 2$ for the first and second nucleotide positions of 511 codons (1022 nucleotides). Tree A is consistent with the tree biologically accepted, but tree B is clearly incorrect, because opossum is clustered with cow and this cluster joins whales. This suggests that nucleotide sequences are less suitable for tree building than are amino acid sequences when distantly related organisms are used. In fact, Russo et al. reached the same conclusion after examining many other mitochondrial genes.

However, it is premature to extrapolate this conclusion to other genes, particularly various types of nuclear genes. The nuclear genome contains many different types of genes, some being highly conserved and others being less. Therefore, it is unclear whether or not currently available amino acids substitution matrices are appropriate for them. Furthermore, for closely related genes, DNA sequences would be better than protein

sequences, because the former contain many silent substitutions that are phylogenetically informative. It is therefore important to examine the pattern of nucleotide or amino acid substitution before we use protein likelihood methods.

8.4. Theoretical Foundation of ML Methods

The maximum likelihood method is a well-established statistical method of parameter estimation. It is known to give the smallest variance of a parameter estimate when sample size is large. Thus, this method is routinely used when one can formulate a likelihood function involving unknown parameters in a given probability space. Familiar examples are estimation of recombination values in classical genetics and estimation of allele frequencies in population genetics. Estimation of branch lengths by ML methods for a given (true) topology can also be justified as long as the correct substitution model is used.

The problem with ML is the reconstruction of topology. In current ML methods of phylogenetic inference, the likelihood function includes no parameters for topologies. Therefore, we are not estimating a topology by maximizing the likelihood (Nei 1987, 1996; Yang et al. 1995a). We simply choose a topology with the highest ML value under the assumption that the topology with good estimates of branch lengths is likely to be the true tree. This assumption does not necessarily hold. In fact, when the rate of nucleotide substitution varies extensively from branch to branch, incorrect topologies may be chosen more often than the true topology is even when the number of nucleotides examined (n) is large (Huelsenbeck 1995). It is also possible to find examples in which ML methods show a lower probability of obtaining the correct topology than MP and some distance methods (see chapter 9). Note also that the regularity conditions (continuity and differentiability of the likelihood function) required for the asymptotic properties of ML estimators are not satisfied in phylogenetic reconstruction (Yang 1994c, 1995a; Yang et al. 1995a). Nei et al. (1998) also showed that when n is relatively small the ML method tends to choose incorrect topologies as do other methods.

In the original formulation of the maximum likelihood method, Cavalli-Sforza and Edwards (1967) treated a topology as a random variable, assuming that population splitting (speciation) occurs according to the Yule process in probability theory. In reality, speciation events almost never follow the Yule process, so this approach has not been used for real data analysis. Recently, Rannala and Yang (1996) and Yang and Rannala (1997) reconsidered this approach, assuming that speciation events follow the birth–death process in probability theory and inferred topologies by evaluating the posterior probability of each topology by the Bayesian approach. Although the birth–death process is more realistic than the Yule process, the real speciation process is much more complicated. Therefore, it is still unclear whether or not this approach will be useful in the future, although the Rannala and Yang approach alleviates several mathematical problems in topology estimation. Larget and Simon

(1999) also proposed a similar Bayesian method, but its practical utility still remains unclear.

8.5. Parameter Estimation for a Given Topology

Although a sophisticated substitution model does not necessarily improve the reliability of inferred tree topologies, it is expected to improve the accuracy of branch length estimates as measured by the number of nucleotide or amino acid substitutions. For this reason, a number of authors have developed statistical methods for estimating the parameters involved in various substitution models. When all sites evolve independently with the same substitution model, a simple way of estimating substitution parameters is to infer the ancestral sequences by parsimony methods and count the numbers of different types of substitutions. This approach has been used by Dayhoff et al. (1978), Gojobori et al. (1982) and others. It can also be used for estimating the gamma parameter α when the rate of substitution varies following the gamma distribution (Kocher and Wilson 1991; Wakeley 1993; Sullivan et al. 1995; Yang and Kumar 1996).

However, the estimates obtained by parsimony methods may be biased, because multiple substitutions at a site are not taken into account. Theoretically, a better method is to use an ML method for a topology that is likely to be correct. As mentioned earlier, the general likelihood function is given by $L = f(\mathbf{x}; \theta)$, where θ is a set of parameters to be estimated. For example, the Kimura model includes the substitution parameters α and β (see Equation [3.10]) as well as $2m - 3$ parameters for branch lengths. Therefore, one can estimate the parameter set θ by maximizing L. Several computer programs are available for this purpose, but Yang's (1995b, 1999) PAML and Swofford's (1998) PAUP* appear to be particularly convenient.

One important factor that affects the estimates of the number of nucleotide substitutions is variation in substitution rate among different sites. This variation is usually described by the gamma distribution given in equation (2.9) with parameter α. Previously, it was customary to use parsimony methods to estimate the parameter α, but parsimony methods are known to give overestimates of α. Yang (1994b) therefore developed an ML method for estimating α using a discrete version of the gamma distribution. This method seems to give good estimates of α, although the estimate varies considerably with substitution model (Yang 1995b, 1996a; Gu and Zhang 1997). However, Yang's method requires a substantial amount of computer time. To reduce this computer time, Gu and Zhang (1997) proposed another version of the ML method, in which the ancestral sequences are first inferred and then the parameter α is estimated by using information on the ancestral sequences. Computer simulation has shown that their method gives essentially the same α value as that obtained by Yang's method, but the computational time is much shorter. At the present time, Gu and Zhang's method can be used only for amino acid sequence data.

The goodness-of-fit of observed data with a given mathematical model in comparison with another is usually evaluated by the *LR* test, under the assumption that *LR* approximately follows a χ^2 distribution with the number of degrees of freedom equal to the difference in the number of free parameters between the two models. In the case of phylogenetic analysis, this assumption is not always guaranteed. Whelan and Goldman (1999) recently examined this assumption for a variety of mathematical models using computer simulation. They found that the χ^2 approximation is adequate for testing models concerning the transition/transversion ratio but not for testing models concerning the gamma shape parameter. This gives a warning against uncritical use of the LR test with χ^2 approximation in phylogenetic analysis.

9

Accuracies and Statistical Tests of Phylogenetic Trees

When a phylogenetic tree is constructed, it is important to know the reliability of the tree obtained. There are two types of errors in a phylogenetic tree: topological errors and branch length errors. The former errors are differences in branching pattern between an inferred tree and the true tree, and the latter are deviations of estimated branch lengths from the true (realized or expected) branch lengths. If the true topology is known, as in the case of computer simulation, it is possible to evaluate the extent of topological errors by the method described in chapter 5. In practice, however, the true topology is almost never known, and therefore the reliability of the topology obtained is usually tested by examining the statistical confidence of various parts (branching patterns) of the topology.

If the topology of a tree is incorrect, one might think that estimation of branch lengths is meaningless. In practice, however, estimation of branch lengths is closely related to the inference of topology as discussed in the last three chapters, and statistical tests of branch length estimates are important for examining the accuracy of the topology as well as the branch length estimates themselves. The reliability of branch length estimates can be tested by either analytical methods or the bootstrap. In this chapter, we are primarily concerned with topological errors rather than with branch length errors, because the former are more important in current molecular phylogenetics.

In the previous chapters, we indicated that the minimization or maximization principle used in the MP, ME, and ML methods tends to give incorrect topologies when the number of nucleotides or amino acids examined is small. This is caused by stochastic errors in nucleotide or amino acid substitution, and this effect becomes serious when the number of sequences is large. We therefore consider this problem before the discussion of statistical tests of phylogenetic trees.

9.1. Optimization Principle and Topological Errors

A simple way to show that the optimization principle tends to give incorrect topologies is to do a computer simulation. If the number of sequences used is small, we can examine all possible topologies and iden-

tify the MP, ME, and ML trees for each set of sequences generated by computer simulation. We then record the tree length (TL), the sum of branch lengths (S), and the maximum log likelihood value ($A = \ln L$) for the MP, ME, and ML trees, respectively. We also compute the TL, S, and A values for the true tree and denote them by TL_c, S_c and A_c. We can then compare TL, S, and A with TL_c, S_c, A_c, respectively, using the **relative optimality score** defined below. With respect to MP trees, this score is defined as

$$R = (TL - TL_c)/TL_c \tag{9.1}$$

This R is positive when $TL > TL_c$, 0 when $TL = TL_c$, and negative when $TL < TL_c$. For the ME and ML trees, R is defined in the same way using S, S_c, A, and A_c. If the optimization principle works well, we should have $R = 0$ for all replications of the simulation.

Using the model tree given in Figure 9.1A, Nei et al. (1998) conducted the above simulation 500 times. In this simulation, they used Jukes and Cantor's model of nucleotide substitution for generating DNA sequences and for constructing phylogenetic trees. In the construction of MP, ME, and ML trees, they used both the **exhaustive search** and the **quick search algorithms** such as the NJ method for ME trees. The R values obtained in this study are presented in Figure 9.1B. In this figure, the solid bars represent the distribution of the R values for the exhaustive search, whereas the open bars give the distribution of the R values for the quick search. In the case of MP, there were incorrect trees with the same TL as that for the true tree (tie trees). They are represented by the gray columns for the exhaustive search and by the hatched bars for the quick search.

Figure 9.1B shows that when the exhaustive search is used, R is always 0 or negative for all MP, ME, and ML trees. In fact, when the number of nucleotides used (n) is 100, a majority of the MP, ME, and ML trees have an $R < 0$, and thus they are incorrect. This may sound strange, because the optimization principle is usually used to find the true tree. However, the property $R \leq 0$ is obvious if we note that TL, S, and A cannot be greater than TL_c, S_c, and A_c, respectively, when all topologies are examined. For example, if any incorrect topology has a TL greater than TL_c, this topology cannot be the MP tree when all topologies are examined. However, if it has a TL smaller than TL_c, then it may be identified as the MP tree. Therefore, we have the relationship $R \leq 0$. Obviously, this property applies to any model tree irrespective of the number of sequences or whether or not the molecular clock holds.

Note that in Figure 9.1B the probability of obtaining the true tree ($R = 0$) increases as n increases. This indicates that the optimization principle works well when n is large. However, when the number of sequences (m) is large and the extent of sequence divergence is small, R can be negative with a high probability even when n is very large (Takahashi and Nei 2000). In other words, the optimization principle generally gives incorrect topologies when n is small relative to m.

Figure 9.1B also shows the R values obtained by quick search algorithms for finding MP, ME, and ML trees. The quick search MP algorithm used here was the min-mini algorithm with search factor 0 (chapter 7),

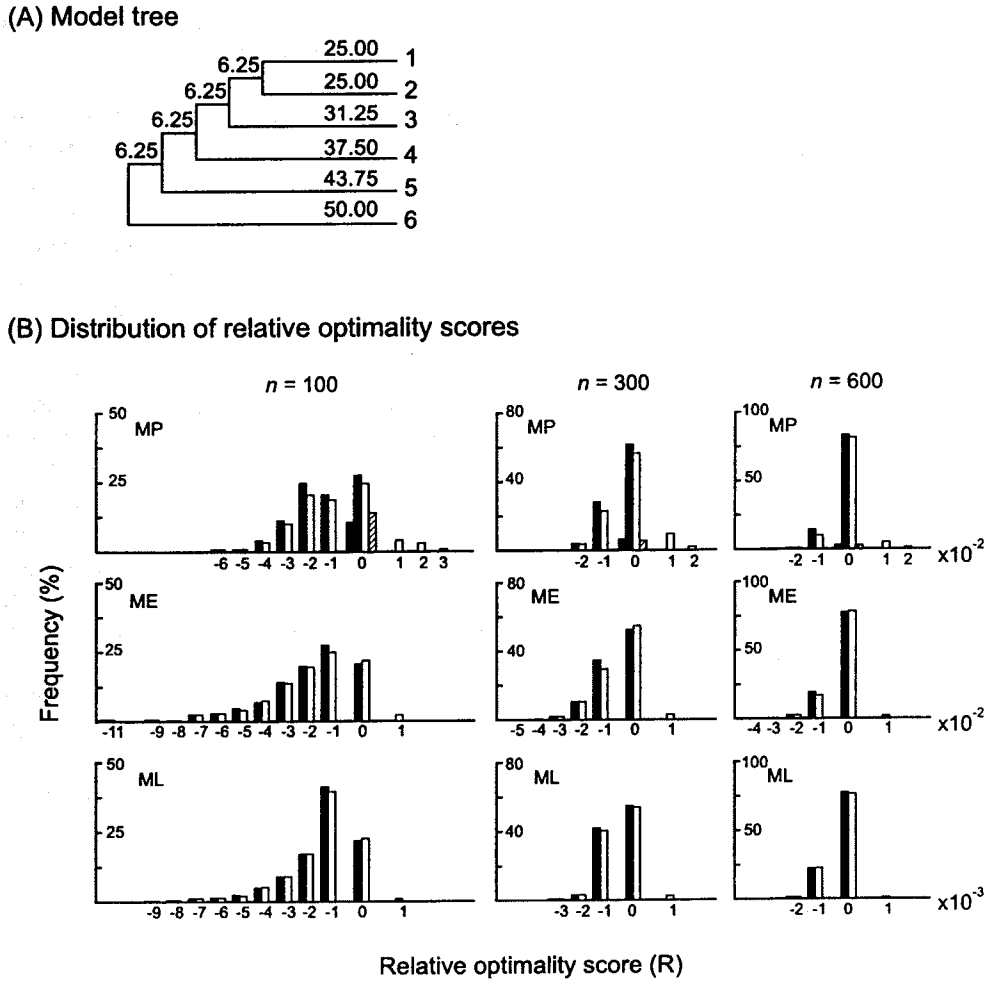


FIGURE 9.1. (A) Model tree used for computer simulation. The branch lengths represent the expected number of nucleotide substitutions per 100 sites. The Jukes-Cantor model of nucleotide substitution was used for generating sequence data. (B) Distributions of relative optimality scores (R) of the MP, ME, and ML trees obtained by the exhaustive search (*solid bars*) and the quick search algorithms (*open bars*). These results were obtained from 500 replications of computer simulation following the model tree. The number of nucleotides used is represented by *n*. See Nei et al. (1998) for details.

and the quick search ME algorithm was the neighbor joining (NJ) method. For ML, the star-decomposition (SD) tree and the stepwise-addition (SA) tree were first constructed, and the topology showing a higher A value was used. This strategy was used because neither the SD nor the SA algorithm alone was very efficient in obtaining the correct topology. These quick search algorithms occasionally gave topologies with $R > 0$, because they did not examine all topologies. Generally speaking, however, the results obtained by these quick search algorithms are very similar to those obtained by the exhaustive search. In particular, the probability of obtaining the correct topology ($R = 0$) is nearly the same for the exhaustive

and the quick searches. This indicates that with the present model tree there is no need to use the exhaustive search; the quick search gives essentially the same results without spending much computational time. Using various model trees of 48 nucleotide sequences and different search algorithms, Takahashi and Nei (2000) also showed that for MP and ML method the NNI and SPR searches are generally as efficient as the TBR search (see chapter 7), and for ME methods the NJ algorithm is as efficient as or more efficient than the TBR search. Similar conclusions were obtained by M. Chase and his colleagues (personal communications) in their statistical analysis of actual DNA sequence data from many different species of plants.

These results indicate that what is important is to develop methods that produce the correct topology with a high probability even for small n, whether the optimization principle is used or not. However, we note that the methods based on the optimization principle work well for large n unless the inconsistency problem arises. Note also that topological errors usually occur when some of the interior branches are short and are subject to stochastic errors, and these branches can be identified by the bootstrap or some other statistical tests. If an interior branch is not well supported, the branching patterns associated with it should be regarded as unresolved. If we adopt this approach, we can still use the methods based on optimization principle. In this case, however, there is no need to use exhaustive or extensive heuristic search algorithms to find the optimal tree. Quick search algorithms appear to be sufficient for statistical inference of phylogenetic trees as long as inferred trees are subjected to statistical tests.

9.2. Interior Branch Tests

Normal Deviate (Z) Test

One way of knowing the reliability of an inferred tree is to test the reliability of each interior branch, as was mentioned previously. This test (**interior branch test**) has been developed for trees constructed by distance methods. Consider the tree for five sequences given in Figure 9.2A. In the case of five sequences, there are 15 possible unrooted bifurcating trees, and each of these trees is composed of five exterior branches and two interior branches. Suppose that topology A is correct and all others are incorrect. One can then show that the expectations of all branch length estimates for the correct topology are 0 or positive and that for incorrect topologies at least one of the interior branches is expected to be negative and this branch generates an incorrect partition of the sequences (Sitnikova et al. 1995). This seems to be true for any number of sequences as long as unbiased distance estimates are used and the branch lengths are estimated by the LS method. Therefore, if a tree has an interior branch whose length estimate is significantly negative, the topology of the tree is likely to be incorrect. The interior branch test utilizes this property.

Suppose that tree A in Figure 9.2 was constructed by a distance method and we wish to test the reliability of this topology. This tree (topology) is

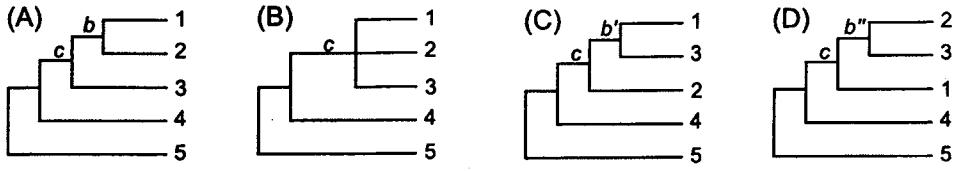


FIGURE 9.2. Three alternative trees that may be generated when the branch length estimate b in tree A is not statistically significant.

regarded as reliable if the two interior branch lengths b and c are shown to be greater than 0. Therefore, testing the null hypothesis that $b \leq 0$ and $c \leq 0$, we can determine whether or not the tree is reliable. In general, the null hypothesis that interior branch length b is equal to or smaller than 0 can be tested by computing the standard error $[s(\hat{b})]$ of an estimate $(\hat{b})$ of b. Since $\hat{b}$ approximately follows the normal distribution when the number of substitutions is sufficiently large (Rzhetsky and Nei 1992a), the null hypothesis $b \leq 0$ can be tested by using the one-tailed test of the normal deviate $Z = \hat{b}/s(\hat{b})$ (Equation [4.5]). Since we are interested in the topology only, there is no need to test exterior branch length estimates.

In phylogenetic analysis, it is often important to test whether or not a group of taxa is monophyletic. For example, if one is interested in testing the monophyly of taxa 1 and 2 (or 3, 4, and 5) in Figure 9.2A, the null hypothesis to be tested is $b \leq 0$ regardless of the value of c. If this null hypothesis is rejected and b is shown to be positive, the monophyly of taxa 1 and 2 (and 3, 4, and 5) is established. Note that under the null hypothesis of $b = 0$ the tree has a trifurcating node (tree B in Figure 9.2). This null hypothesis is also the same as the null hypothesis $b' = 0$ in tree C or $b'' = 0$ in tree D. Therefore, if the null hypothesis $b = 0$ is not rejected, the true topology can be any of trees A, B, C, and D.

When we compare tree A with tree C, the null hypothesis is that trees A and C are identical. This happens only when b and b' are equal to 0, and in this case the null hypothesis is represented by tree B. For this reason, we call tree B a **null tree** (see section 6.3). This null tree can be used for testing the topological difference between trees A and D or trees C and D as well. The validity of tree A, C, or D is established when interior branch b, b', or b'' is shown to be positive. The above rule applies to any comparison of topologies for any number of sequences. In Figure 6.4 we saw that for an unrooted tree of five sequences the null tree for testing two topologies that are different by $d_T = 4$ is the star tree. It is important to realize that whenever two topologies are compared there is always a null tree that corresponds to the null hypothesis to be tested. This is true irrespective of the tree-building method used.

The test of the above null hypothesis can be done relatively easily for a tree obtained by distance methods, particularly by the NJ or ME method, because all the interior branches are expected to become positive only for the correct topology (Figure 6.5). In the case of MP and ML trees, however, all interior branches are positive irrespective of the topology (Figure 6.5), so that it is difficult to develop an analytical method for testing the null hypothesis. However, it is possible to test the null hypothesis using the bootstrap test.

Analytical Method

An analytical method for testing an interior branch was first used by Nei et al. (1985) for a UPGMA tree and then by Li (1989) for an unrooted tree for four or five taxa. Later, Rzhetsky and Nei (1992a, 1993) developed a fast algorithm for computing $s(\hat{b})$ for an ME tree using the ordinary least-squares approach and made it possible to use this test for a large number of sequences. This test requires a specific model of nucleotide substitution to compute $s(\hat{b})$, but a crude model seems to be sufficient unless the extent of sequence divergence is very high (Tateno et al. 1994). In this method, the **confidence probability** (P_C) that $\hat{b} > 0$ rather than the Type I error ($P = 1 - P_C$) is often computed by using the Z test, and if this confidence probability is higher than 95% or 99%, then $\hat{b}$ is considered to be significantly positive. This practice is done to make the test comparable to the bootstrap, although theoretical statisticians may object to it.

The above test is to examine the accuracy of an inferred tree, which was obtained by a particular tree-building method. In practice, however, the inferred tree may be incorrect, and if we consider this possibility, the above test may not give a proper estimate of statistical confidence. In fact, Sitnikova et al. (1995) and Sitnikova (1996) have shown that if this possibility is taken into account, the actual confidence probability becomes lower than the original P_C value. However, since most investigators are interested in the reliability of an inferred tree, we shall not consider this a problem.

It should also be noted that the above interior branch test is applicable only for relatively closely related sequences (see Example 9.1). Distantly related sequences usually have experienced complicated processes of nucleotide or amino acid substitution, so that the assumptions underlying the test may not be satisfied. In this case, the following bootstrap interior branch test is more reliable.

Bootstrap Interior Branch Test

Another interior branch test that is applicable to distance trees is Dopazo's (1994) **bootstrap interior branch test.** This test is different from Felsenstein's bootstrap test and is intended to examine the reliability of each interior branch of a given topology. As in the case of the bootstrap test, the same number of nucleotides (or amino acids) as that of the original sequence is randomly sampled with replacement from the original set of sequences, and the lengths of all branches are estimated by a given tree-building method for the topology that was obtained from the original sequence data. This is repeated several hundred times for the *same topology.* The length estimate ($\hat{b}$) of an interior branch will then vary from replication to replication and may even become negative. We can then compute the mean and standard error of $\hat{b}$ and conduct a Z test.

The results of this test are usually very similar to those obtained by the analytical method mentioned above (Sitnikova 1996). However, this method has one advantage over the analytical method. That is, there is no need to compute the standard error of $\hat{b}$ for each substitution model separately; the standard error can be computed in the same way for all

substitution models. The computational time also does not increase rapidly with the number of sequences. For these reasons, it is easier to use this method than the analytical method. However, when the number of nucleotides or amino acids used is small, this method may give biased estimates of P_C's, because if the original sample is biased, this bias can never be removed by the resampling process. In this case, the analytical method is preferable.

Likelihood Ratio Test

Felsenstein (1988) suggested that the null hypothesis of $b = 0$ could be tested by the likelihood ratio test (see chapter 4), because the trees with $b = 0$ and with $b \geq 0$ are nested. For example, tree B in Figure 9.2 is a special case of tree A. In the case of phylogenetic analysis, however, this test does not seem to be justified. The reason is that the regularity conditions (continuity and differentiability) of the likelihood function required for the asymptotic properties of ML estimators are not satisfied when different topologies are considered (Goldman 1993; Yang 1994c; Yang et al. 1995a). Computer simulations have shown that while the test seems to apply when an appropriate substitution model is used, it may give strong statistical support for a wrong tree when an inappropriate model is used (Tateno et al. 1994; Gaut and Lewis 1995; Zhang 1999). The DNAML program in PHYLIP (Felsenstein 1995) computes the confidence interval of a branch length estimate using the curvature method, but this interval is also unreliable (Tateno et al. 1994). Note that branch length estimates never become negative in ML and MP trees even if the topology is incorrect (Figure 6.5). Therefore, the interior branch test is not meaningful for MP and ML methods.

For MP and ML trees, the most reliable test of interior branch lengths or a monophyly of a group of sequences is probably the bootstrap test mentioned below. The bootstrap is known to be a conservative test, but a conservative test is desirable in phylogenetic analysis, because the actual pattern of nucleotide substitution is very complicated. The bootstrap test is applicable to NJ trees as well.

9.3. Bootstrap Tests

Procedures

One of the most commonly used tests of the reliability of an inferred tree is Felsenstein's (1985) bootstrap test, and we have already used this test in previous chapters. In this test, the reliability of an inferred tree is evaluated by using Efron's (1982) bootstrap resampling technique. Consider the following set of nucleotide sequences.

$$
\begin{array}{l}
x_{11}, x_{12}, x_{13}, \cdots, x_{1n} \\
x_{21}, x_{22}, x_{23}, \cdots, x_{2n} \\
\cdot \quad \cdot \quad \cdot \quad \cdots, \cdot \\
x_{m1}, x_{m2}, x_{m3}, \cdots, x_{mn}
\end{array}
$$

Here x_{ij} represents the nucleotide at the i-th sequence (row) of the j-th site (column), and m and n stand for the number of sequences and the number of nucleotides per sequence, respectively. We first construct a tree from the entire data set by using some tree-building method (e.g., NJ or ML method). This is the inferred tree of which the reliability is to be tested. We now want to test this reliability using the bootstrap resampling technique. In bootstrap resampling, n nucleotide sites are randomly chosen with replacement from the original set of sequences. Therefore, some sites may be chosen two or more times, whereas some other sites may not be chosen at all. These randomly sampled n sites (columns) now constitute a new set of DNA sequences, and this set is used for constructing another tree by the same tree-building method. The topology of this tree is then compared with that of the original tree. Any interior branch of the original tree that gives the same partition of sequences as that of the bootstrap tree (see chapter 5) is given value 1 (identity value), whereas other interior branches are given 0. This process is repeated several hundred times, and the percentage of times each interior branch of the original tree receives identity value 1 is computed. We call this the **bootstrap confidence value** (P_B) or simply the **bootstrap value.** In general, if this P_B is *95% or 99% or higher,* depending on the accuracy one wishes to have, the interior branch is considered to be significantly positive.

When a tree is constructed by distance methods, this bootstrap test is conceptually similar to the interior branch test mentioned above, and the null hypothesis of the test is that the length of each interior branch is 0. Therefore, when closely related sequences are studied, P_B and P_C are usually similar to each other (see Figure 6.8C). Theoretically, however, P_B gives a conservative estimate of statistical confidence of an interior branch and tends to be smaller than P_C (Sitnikova et al. 1995; Sitnikova 1996).

It should be noted that Felsenstein's original version of bootstrap tests is slightly different from the procedure just described. In his method, the test is not for examining the reliability of a tree reconstructed from the original data set but for examining the reliability of a bootstrap consensus tree, which is produced by considering all the trees generated by bootstrap resampling. If a large number of nucleotide or amino acid sites are used and all sites evolve independently with the same substitution model, this consensus tree may be closer to the expected phylogenetic tree mentioned in chapter 5 than is the inferred tree from the original data set (Berry and Gascuel 1996). In this case, the construction of a consensus tree and the computation of P_B values are done simultaneously. Therefore, the null hypothesis for this test is not very clear, but it should be similar to the one previously mentioned. Felsenstein's original version is incorporated in the computer software packages PHYLIP, PAUP*, and MEGA2, whereas the test of an inferred tree is available in MEGA and MEGA2. In practice, these two test procedures give similar results.

In the bootstrap test, a phylogenetic tree needs to be constructed for each set of resampled data. Therefore, the computational time required for constructing a tree becomes an important factor in this test. For this reason, this test is routinely used for NJ trees, but it is very time-consuming for MP and ML trees unless the number of sequences used is

small. This is so even if quick ML and MP search algorithms (e.g., SA + NNI search) are used. Strimmer and von Haeseler's (1996) quartet ML method (PUZZLE algorithm) is a relatively simple algorithm for obtaining approximate ML trees, but it is still time-consuming when a bootstrap test is to be conducted. For this reason, Strimmer and von Haeseler suggested that their relative reliability index may be used as a crude indication of the reliability of an interior branch. However, Cao et al. (1998) and Honda et al. (1999) showed that the index is often too large or too small compared with the bootstrap values depending on the data set.

When the bootstrap test is applied to maximum parsimony trees, it is customary to produce a bootstrap consensus tree and test each interior branch of this tree. This is partly because MP methods often generate two or more parsimonious trees and this complicates the test of an MP tree (or trees) obtained from the original data set (Felsenstein 1985). As mentioned earlier, MP methods can be "inconsistent estimators" of phylogenetic trees more often than distance or ML methods are. If this inconsistency occurs with a data set to be analyzed, a bootstrap test may superficially support a wrong tree as though it were the true tree. Therefore, some caution is necessary about the interpretation of bootstrap values of an MP tree. Since inconsistency is likely to occur more often when the evolutionary rate varies from branch to branch than when it is more or less constant, it is a good idea to examine rate variation among different branches of the MP tree obtained from the entire data set.

Statistical Properties

The statistical properties of the bootstrap test are complicated and are not well understood, though they have been studied by a number of authors (Zharkikh and Li 1992a, 1992b; Felsenstein and Kishino 1993; Hillis and Bull 1993; Sitnikova et al. 1995; Efron et al. 1996). When the test is applied to an NJ tree, however, the interpretation of the test results is relatively simple. If (1) each site of the DNA sequence evolved independently, (2) the distance measure used is an unbiased estimator of the number of nucleotide substitutions, and (3) the numbers of sequences (m) and nucleotides (n) used are sufficiently large, the null hypothesis of the bootstrap test is that the length of each interior branch is 0, and P_B for a branch is supposed to measure the probability (P_C) of the branch length being greater than 0. Efron et al.'s (1996) study suggests that this is indeed the case when m is large. This is also true even when m is small if the true length of an interior branch is substantially large (Sitnikova et al. 1995). However, when m is small and the true length of an interior branch is equal to or close to 0, P_B tends to be an underestimate of P_C when it is close to 1 but an overestimate when it is low (Sitnikova et al. 1995). The actual relationship between P_B and P_C seems to be very complex, and more study should be done.

In the case of MP and ML trees, a somewhat different treatment is necessary for studying the statistical implication of P_B, because $\hat{b}$ never becomes negative in these trees. Nevertheless, the computer simulations by Zharkikh and Li (1992a, 1992b) and Hillis and Bull (1993) suggests that

the statistical property of P_B for MP trees is essentially the same as that for NJ trees.

Although Felsenstein's bootstrap test can be very conservative, we believe that it is a useful method for evaluating the statistical reliability of an inferred tree. The actual pattern of nucleotide substitution is very complicated (Yang 1994a; Kumar 1996a) and often changes with site and evolutionary time (Lake 1994; Galtier and Gouy 1998). Therefore, it is better to use a conservative test for examining trees for distantly related sequences. Of course, when a tree produced is for closely related sequences, one may use mathematically more rigorous methods such as the interior-branch test (Sitnikova et al. 1995) or Zharkikh and Li's (1995) complete and partial bootstrap technique.

Example 9.1. P_B and P_C Values for a Few Example Trees

In the example trees given in chapter 6–8, we have already presented the bootstrap (P_B) or confidence probability (P_C) values or both without going into details of the implication. We have seen that the topology of the NJ tree in Figure 6.8 is virtually identical with that of the true (realized) tree, whether the JC or Kimura distance is used. The P_B value with 1000 replications is also very high (>90%) for many interior branches, but for some interior branches it is quite low. The P_C value is similar to P_B in most interior branches.

The P_B and P_C values are also given for the NJ tree in Figure 7.12B. The P_B and P_C values are again nearly the same for most interior branches. The Dopazo test also gives very similar P_C values. These results indicate that as long as sequence divergence is relatively low, one can use any of the three tests. In practice, the Dopazo test is simplest and is applicable to any model of nucleotide substitution. Therefore, this test should be used more often. However, when the extent of sequence divergence is high, (say $d > 0.5$), P_C may be considerably higher than P_B. In this case, P_C should not be trusted.

The MP tree in Figure 7.12A is a bootstrap consensus tree for cetaceans, artiodactyls, and related species. As discussed earlier, this tree has the same topology as that of the NJ tree (Figure 7.12B) except for the branching pattern for sheep, giraffe, and moose, which is weakly supported by the bootstrap test in both trees. For the other interior branches, the P_B values for the MP tree are virtually the same as those for the NJ tree. In both trees, 8 out of 12 interior branches have a P_B value of 90% or higher.

The ML bootstrap consensus trees for the same set of sequence data are given in Figure 8.3 C, D. These trees were obtained by using the SD + NNI search in PAUP*. Tree C, which was obtained by using the Kimura model, has the same topology as that of the MP tree (Figure 7.12A), and the bootstrap values for the two trees are again very similar. Tree D was generated by using the HKY model with the SD + NNI search, and its topology is different from that of tree C but is the same as that of the NJ tree (Figure 7.12B). However, the interior branch that caused this topological difference has a low P_B value in both trees C and D so that it is

unclear which tree is better than the other. Nevertheless, the interior branches that have a P_B value of 90% or higher in tree D also have a similar value in tree C. Actually, comparison of trees C and D with the MP and NJ bootstrap trees indicates that all the MP, NJ, and ML analyses give essentially the same conclusion about the phylogenetic relationships.

It is also interesting to know that both bootstrap and confidence probability tests statistically establish the monophyly of ruminants, hippopotamus, and whales, and this conclusion is in agreement with the results obtained by retroposon analysis given in Figure 7.15.

Condensed Trees

When a phylogenetic tree has low P_C or P_B values for several interior branches, it is often useful to produce a multifurcating tree assuming that all such interior branches have branch length 0. We call this multifurcating tree a **condensed tree.** In MEGA and MEGA2, this condensed tree can be produced for any level of P_C or P_B value. For example, if there are several branches with P_C or P_B values of 50% or less, a condensed tree with the 50% P_C or P_B level will be a multifurcating tree with all these branch lengths reduced to 0. Since interior branches of low significance are eliminated to form a condensed tree, this tree gives emphasis on the reliable portions of branching patterns. It will affect the branch lengths, but there is no general method for estimating the branch lengths of a condensed tree. Therefore, MEGA and MEGA2 present only the topology without branch lengths.

Note that the condensed tree is different from the consensus tree mentioned earlier, although they may look similar in practice. A consensus tree is produced from many equally parsimonious trees, whereas a condensed tree is merely a simplified presentation of a tree. A condensed tree can be produced for any type of tree (NJ, ME, UPGMA, MP, or ML tree).

9.4. Tests of Topological Differences

Minimum-Evolution Tree

The second class of statistical tests is to compare two topologies in terms of a quantity that is used for an optimization process of phylogenetic inference. Previously we mentioned that the minimum evolution tree is a tree that has the smallest sum (S_M) of branch length estimates. In practice, however, there may be several other trees whose S is greater than S_M but is not significantly different from the latter. As mentioned in chapter 6, these trees are potentially correct trees, and thus one may want to keep them until other data are obtained to identify the true tree. Rzhetsky and Nei (1992a) developed a statistical method for testing the difference ($D = S_B - S_A$) in S between two topologies. This test is equivalent to the test of the lengths of the interior branches at which the two topologies are different. Suppose that topology A in Figure 6.4 is the correct tree and

topologies B and C are incorrect ones. When unbiased distance estimators are used, the expectation [$E(D)$] of D for topologies A and B is given by $b_6/8$, where b_6 is the true length of the left interior branch of topology A (Equation [6.19]). Therefore, if D is significantly greater than 0, we can conclude that topology A is more likely to be correct than topology B. However, if D is significantly smaller than 0, topology B may be the better one. Similarly, comparison of topologies A and C gives the expectation of D ($= S_C - S_A$) equal to $3(b_6 + b_7)/4$ (Equation [6.20]). Therefore, the null hypothesis for the test of D has a clear-cut biological meaning. Using this test, one can identify topologies that are not significantly different from the ME tree. This test is for comparing two topologies of which the validity is unknown. Therefore, it is a two-tailed test.

This method of testing D depends on the mathematical model on which a particular distance measure is based and requires an intensive computation when m is large. However, the hypothesis $E(D) = 0$ can be tested by a bootstrap method. In this test, S is computed for a given pair of topologies (i and j) for each sequence resampling, and $D_{ij} = S_i - S_j$ is computed. If this is repeated many times, we can compute the mean and the standard error of D_{ij}'s. Therefore, we can test the null hypothesis of $E(D_{ij}) = 0$ using the Z test. When there are several potentially correct trees, D_{ij} may be computed for all pairs of i and j by using the same set of resampled sequences.

This D test is clearly related to the interior branch test mentioned earlier, but the exact relationship remains unclear. One might speculate that if every interior branch of an ME tree is significant, the D test will also establish that S_M is significantly smaller than S for any other tree. If this is the case, the interior branch test or the Felsenstein bootstrap test would be simpler than the D test in finding a reliable tree.

As mentioned earlier, the NJ method produces only one final tree rather than several potentially correct trees. However, this is not a serious problem in practice. If all interior branches of a NJ tree are statistically supported, there will be no need to consider other alternative trees. If some of the interior branches are not statistically supported, one may consider the alternative trees that can be generated by changing the branching pattern for each nonsignificant interior branch (see Figure 9.2). This approach would be simpler than the ME method, in which many different topologies are examined by using the D test. Of course, if there are many interior branches whose P_B or P_C values are low, this procedure would require examination of a large number of topologies. In this case, however, there is no need to examine them, because the tree would not be reliable any way whichever tree-building method is used.

ML and MP Trees

One might think that the difference in topology between an ML and a suboptimal tree could be tested simply by the standard likelihood ratio (LR) test with an χ^2 approximation. Unfortunately, this cannot be done, because this test has zero degrees of freedom (Felsenstein 1988), and the regularity conditions of the likelihood function required for asymptotic properties of ML estimators are not satisfied (Yang et al. 1995a).

Huelsenbeck et al. (1996) and Huelsenbeck and Crandall (1997) suggested that the significance level of the *LR* value for testing a preassigned monophyletic group of taxa be computed by a parametric bootstrap. In the parametric bootstrap test, the ML tree inferred from the original data set is assumed to be the true (model) tree, and a new set of sequence data are generated following this model tree with a given substitution model. These sequences are then used for generating a new ML tree, and the *LR* value for testing a preassigned monophyletic group is computed by comparing this tree with the tree with the preassigned monophyly (see the original papers for details). This is repeated many times, and the distribution of *LR* is obtained. Once this distribution is obtained, it is possible to determine the significance level of the original *LR* value.

Although a monophyly or nonmonophyly that is not observed in the original ML tree may be rejected by this test (see Huelsenbeck et al. 1996), this test is not statistically as rigorous as the interior branch test discussed earlier, where the null tree corresponding to each null hypothesis is clearly defined (see Figure 9.2). Actually, the null hypothesis of the parametric bootstrap test is not well defined in terms of phylogenetic trees. Note also that the ML tree obtained from the original data set refers to a realized tree, which can be quite different from the true tree (see Figure 5.5). Therefore, the parametric bootstrap test for a preassigned monophyly or nonmonophyly can be too liberal or too conservative, depending on the taxonomic group included. We also usually do not know the real substitution model, and the use of incorrect substitution model may lead to a too liberal or too conservative test (Zhang 1999). Therefore, a more careful study of the theoretical basis of this test should be conducted. ML tests of topological differences are known to be notoriously complicated, because different topologies represent different probability spaces (Yang et al. 1995a). Actually, a more reliable test of a monophyletic group of sequences observed in the original ML tree is to conduct the standard bootstrap test and evaluate the bootstrap value for the monophyly.

Kishino and Hasegawa (1989) suggested that the difference in log likelihood value between an ML tree and a suboptimal tree can be tested by using the variance of the difference in single-site log likelihood between the two trees. The null hypothesis of this test is obviously that the ML values of the two trees are identical. This can happen only when the nonshared interior branches between the two trees have length 0 as in the case of tree B in Figure 9.2 (or tree D or E in Figure 6.4), and therefore the null hypothesis of this test is the same as that for testing two distance trees. However, Kishino and Hasegawa have not considered their test in relation to the topological difference between the two trees, and therefore it is unclear whether Kishino and Hasegawa's test procedure is appropriate for testing this null hypothesis.

For MP trees, it is difficult to develop any parametric test because of the nonrandom nature of "minimum numbers of substitutions." Templeton (1983) suggested a nonparametric test for comparing two topologies that is similar to Kishino and Hasegawa's (1980) test for ML trees. However, it is again unclear whether this method is appropriate for testing the null hypothesis mentioned above.

9.5. Advantages and Disadvantages of Different Tree-Building Methods

Criteria of Comparison

Because there are many different tree-building methods, one is naturally interested in the advantages and disadvantages of different methods. There are several different criteria for comparing different tree-building methods. Important ones are (1) computational speed, (2) consistency as an "estimator" of a topology, (3) statistical tests of phylogenetic trees, (4) probability of obtaining the true topology, and (5) reliability of branch length estimates.

The computational speed of each tree-building method can be measured relatively easily, although it depends on the algorithm and the computer used. According to this criterion, UPGMA and NJ are superior to most other tree-building methods that are currently used. These methods can handle a large number of sequences ($m > 500$) even with a personal computer, and the bootstrap test can be done very easily. The orthodox MP, LS, ME, and ML methods examine all possible topologies in searching for the MP, LS, ME, and ML trees, respectively. Since the number of possible topologies rapidly increases with m (Equation [5.1]), it is difficult to use these methods when m is large. In the case of ME, the quick-search algorithms such as NJ seem to be as efficient as the exhaustive search in obtaining the correct tree. It is hoped that similar quick-search algorithms will be developed for other optimization methods as well. Note that the vast majority of tree topologies are clearly incorrect when m is large and that there is no need to examine all these trees.

A tree-building method is said to be a "consistent estimator" if the method tends to give the correct topology as the number of nucleotides used (n) approaches infinity (Felsenstein 1978), although this is not a standard usage of the terminology in statistics (see chapter 5). The NJ, ME, and LS methods are all consistent estimators if unbiased estimates of nucleotide substitutions are used as distance measures (Saitou and Nei 1987; DeBry 1992; Rzhetsky and Nei 1992a) and so is the ML method when the correct model of nucleotide substitution is used (Yang 1994c; Rogers 1997). By contrast, MP is sometimes inconsistent, and the phenomenon of short-branch or long-branch attraction occurs, as mentioned in chapter 5. In practice, however, n is usually of the order of hundreds to thousands, and in this case, NJ, ME, LS, and ML may also fail to produce the correct tree when MP fails (Huelsenbeck and Hillis 1993; Huelsenbeck 1995; Nei et al. 1995; Schöniger and von Haeseler 1995). Therefore, consistency is not always a useful criterion for comparing the efficiencies of different tree-building methods.

We have already discussed statistical tests of phylogenetic trees obtained by several different tree-building methods. At present, the statistical methods for testing NJ and ME trees are well established. Solid statistical tests are also available for the trees obtained by the generalized LS method (Bulmer 1991; Uyenoyama 1995). For MP and ML methods, however, there are many complications, as mentioned above. The best

method for testing MP and ML trees is probably Felsenstein's bootstrap test, as long as the problem of inconsistency does not arise.

The probability of obtaining the true topology (P_T) is one of the most important criteria for comparing different tree-building methods, but this is also the most difficult problem to study. This requires knowledge of the true tree, and this knowledge is rarely available except in computer simulation. In computer simulation, it is easy to evaluate P_T for any given topology, but the conclusions obtained by simulation study may not apply to real data. A quantity related to P_T is the topological distance (d_T) of a reconstructed tree from the true tree. Although d_T is generally highly correlated with P_T (Tateno et al. 1982), d_T is more useful when P_T is very small.

Another important criterion for comparing different methods is the reliability of branch length estimates. Once the correct topology is obtained for a given data set, this problem can be studied relatively easily. Theoretically, ML, LS, NJ, and ME are expected to give more reliable estimates of branch lengths than is MP. (For NJ, it is now customary to estimate branch lengths by the LS method once the topology is established.) At the present time MP trees are almost always presented without branch length estimates, probably because MP tends to give underestimates of branch lengths. This is regrettable, because it gives a distorted picture of a phylogenetic tree. Since computer programs (e.g., PAUP*, McClade, and MEGA2) are available for estimating branch lengths of MP trees, and the extent of underestimation is not serious when closely related sequences are studied, MP trees should be presented with branch length estimates as often as possible.

Probability of Obtaining the True Topology

In real data analysis, we almost never know the true phylogeny, so it is difficult to study this problem empirically. However, if we use an appropriate mathematical model, we can simulate the evolutionary changes of DNA sequences following a given model tree. We can then reconstruct a tree by various methods using the artificially generated present-day sequences and compare the topology of the tree obtained with that of the model tree. If this process is repeated many times, we can estimate the probability of obtaining the true topology (P_T), and this probability can be used for comparing the efficiencies of different tree-building methods (Peacock and Boulter 1975; Blanken et al. 1982; Tateno et al. 1982; Nei 1991). Under certain circumstances, this probability can be evaluated analytically as well as empirically.

Theoretical Study

When the number of sequences examined (m) is small (four or five), it is possible to evaluate P_T analytically for the NJ, LS, and MP methods (Saitou and Nei 1986; Zharkikh and Li 1992a; Sitnikova et al. 1995). These studies have shown that when the evolutionary rate is more or less constant for all four or five sequences, NJ has a slightly higher P_T value

than MP, which in turn has a somewhat higher P_T than Fitch and Margoliash's (1967) LS method (Saitou and Nei 1986). It has also been shown that both the ordinary and generalized LS methods are inferior to ME in obtaining the correct topology (Rzhetsky and Nei 1992b). This inferiority seems to be partly due to the fact that the LS methods often generate negative branches as mentioned earlier. However, analytical evaluation of P_T is very difficult when m is large, and the conclusion obtained from the above studies may not apply to a wide variety of situations. No theoretical study has been made for ML even for the case of $m = 4$. For this reason, comparison of P_T among different methods is usually done by computer simulation.

Computer Simulation

If we use computer simulation, it is possible to estimate P_T's for a variety of evolutionary conditions. For this reason, a large number of simulation studies have been done during the last 20 years. The results obtained before 1990 have been reviewed by Nei (1991), but there are many recent studies (e.g., Hasegawa et al. 1991; Rzhetsky and Nei 1992a; Hasegawa and Fujiwara 1993; Kuhner and Felsenstein 1994; Yang 1994c, 1996c; Gaut and Lewis 1995; Huelsenbeck 1995; Nei et al. 1995; Schöniger and von Haeseler 1995; and others). It is not an easy job to summarize these studies, because different authors considered different evolutionary models and used different computer algorithms.

One of the most popular model trees used in computer simulation is the unrooted tree of four sequences in the form given in Figure 9.3A, where a, b, and c represent the expected number of nucleotide substitutions per site. When $a = b = c$ and they are between 0.1 and 0.5, almost any tree-building method produces the correct topology if n is greater than 100. Therefore, this model tree is not useful for discriminating the efficiencies of different methods. For this reason, many authors have assumed $a > b$. If we use the Jukes-Cantor model of nucleotide substitution, the MP method becomes inconsistent when $b = c = 0.05$ and $a \geq 0.394$ (Tateno et al. 1994). Therefore, MP always fails to produce the correct tree when a large number of nucleotides is used. However, NJ and ML usually recovers the correct tree in this case if $a < 0.5$.

Some authors (Hillis et al. 1994) have used cases of an extremely high degree of sequence divergence ($a = 2.83$, which corresponds to $p = 0.65$, and $b = c = 0.05$ corresponding to $p = 0.05$) to show the superiority of ML methods. However, such divergent sequences are almost never used in actual data analysis because of the difficulty of sequence alignment or of a large amount of statistical noise. Therefore, such a study is not biologically meaningful. For the same reason, a large part of the computer simulation conducted by Huelsenbeck and Hillis (1993) and Huelsenbeck (1995) seems to be biologically irrelevant. Although they considered the complete two-dimensional space for a and $b = c$ ($0 \leq p \leq 0.75$; $0 \leq$ corrected distance $d \leq \infty$) for the sake of completeness, actual data used for phylogenetic analysis fall into a relatively small portion of the space near the origin (Nei et al. 1995; Rzhetsky and Sitnikova 1996). When $0.1 < a < 0.5$ and b and c are around 0.05, MP is generally

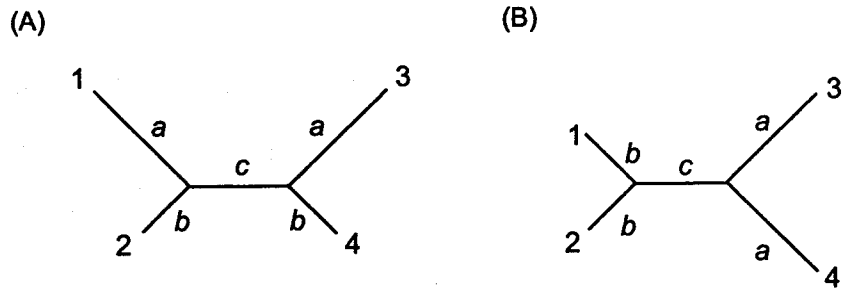


FIGURE 9.3. Four-sequence trees often used for computer simulation.

less efficient than NJ, which is in turn less efficient than ML (Hasegawa et al. 1991; Hasegawa and Fujiwara 1993; Tateno et al. 1994; Huelsenbeck 1995). However, when a, b, and c are all of the order of 0.01 ~ 0.025 and n is about 1,000, all three methods reconstruct the true tree quite easily (Tateno et al. 1994).

Model tree B is different from tree A in that the two long branches with length a are now neighbors and the two short branches with length b are also neighbors. Interestingly, this model tree gives different relative P_T values compared with those for tree A (Nei 1996; Russo et al. 1996). Some results obtained with Kimura's model of nucleotide substitution are given in Table 9.1. In the case of tree A, ML gives the highest P_T value among the three methods ML, MP, and ME irrespective of the number of nucleotides used, and NJ with p distance shows the lowest value. In tree B, however, ML gives the lowest P_T, and NJ with p distance gives the

Table 9.1 Percent probabilities of obtaining the correct tree topology ($P_T \times 100$).

	Tree A					Tree B				
	NJ[a]		MP[b]			NJ		MP		
Nucleotides (n)	p[c]	$K2$[d]	UW[e]	W[f]	ML[g]	p	$K2$	UW	W	ML
100	44	72	47	64	76	98	74	88	96	64
200	41	81	52	80	84	100	82	97	99	76
300	43	88	59	80	92	100	86	98	100	82
500	35	95	62	89	97	100	94	100	100	90
800	29	96	63	94	98	100	96	100	100	94
1000	35	99	66	98	100	100	99	100	100	96

Source: From Nei (1996).
Note: In both trees A and B in Figure 9.3, $a = 0.4$, $b = 0.1$, and $c = 0.05$ were assumed. Sequence data were generated by using Kimura's 2-parameter model with $R = 2$.

[a]NJ: neighbor-joining method.
[b]MP: maximum parsimony method.
[c]p: p distance.
[d]*K2*: modified Kimura distance (Felsenstein 1995).
[e]UW: unweighted.
[f]W: weighted with $w = 4$ in Figure 7.10A.
[g]ML: maximum likelihood method.

higher P_T value. Furthermore, both unweighted and weighted MP show much higher P_T's than ML. These results were obtained apparently because in parsimony and NJ with p distance short branches tend to attract each other. Yang (1996c) has also shown that even when the evolutionary rate is constant, ML can be inferior to unweighted MP. These results indicate the difficulty of obtaining a general conclusion about the relative efficiencies of different tree-building methods even for the simplest case of $m = 4$.

A number of simulation studies have been done for the cases of six or more sequences. It is not easy to consider more than a few dozen sequences in a simulation, because for large m, the interior branch lengths become very small if we make the most divergent sequence pair biologically reasonable (e.g., $d \leq 1.0$). Therefore, P_T becomes very low for any method unless a large number of nucleotides are used and an enormous amount of computer time is spent (Tateno et al. 1982; Kumar 1996b). The model trees considered usually represent the case of constant rate or its modifications (e.g., Saitou and Nei 1987; Sourdis and Krimbas 1987; Sourdis and Nei 1988; Saitou and Imanishi 1989; Kuhner and Felsenstein 1994; Strimmer and von Haeseler 1996). In general, these simulation studies have shown that ML is as good as or better than NJ, which is in turn often better than MP, though the differences are usually quite small when biologically reasonable model trees are considered. However, the number of these studies is quite limited, and it is difficult to extrapolate these results to general cases. The pattern of nucleotide substitutions used in computer simulation is also quite unrealistic. When Håstad and Björklund (1998) used the observed substitution pattern in the mitochondrial cytochrome *b* gene for generating simulated sequence data, MP, NJ, and ML methods were nearly equally efficient.

Despite many computer simulations conducted recently, the interpretation of the results is not as straightforward as was originally expected, and more careful studies seem to be necessary to know the relative efficiencies of different methods. However, it is now clear that no method is perfect and there are situations in which one method performs better than others, and unless the evolutionary rate varies drastically with evolutionary lineages, the three methods considered here generally give the same or similar topologies (Saitou and Imanishi 1989; Hasegawa et al. 1991; Nei et al. 1998). Computer simulations have also indicated that one of the most important factors that affect the accuracy of a reconstructed tree is the number of nucleotides or amino acids used per sequence and that if this number is small, one cannot produce reliable trees.

Tests Based on Known Phylogenies

Although it is generally difficult to know the true topology in real data analysis, there are a few such cases. One is a phylogenetic tree experimentally produced by artificial mutagenesis with T7 phages (Hillis et al. 1992). However, this type of experiment produces only one or a few replications, so it is difficult to compare different methods statistically.

There are a few instances in which the phylogenetic tree for a group of organisms is firmly established on paleontological and morphological

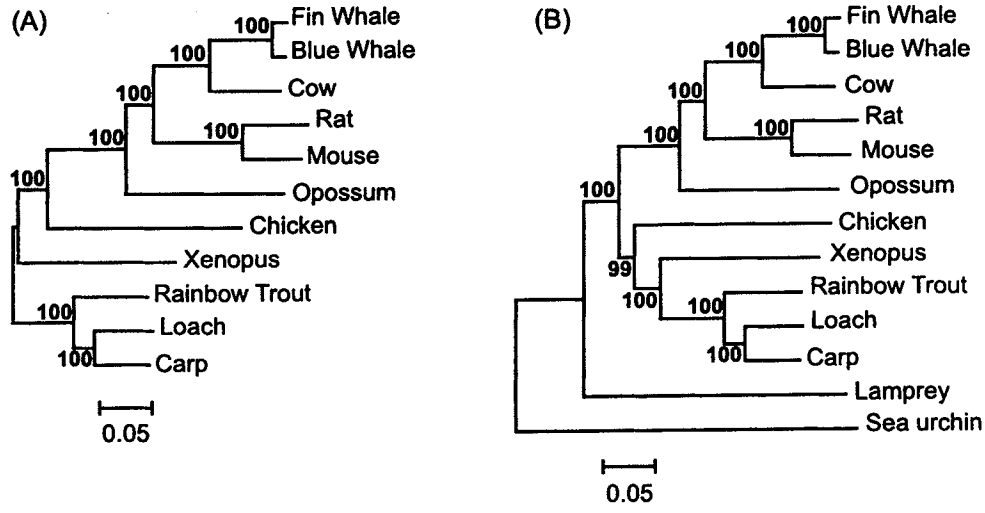


FIGURE 9.4. (A) Known phylogeny of 11 vertebrate species. (B) Inferred phylogeny when lamprey and sea urchin sequences were added. Bootstrap values for the NJ trees constructed using the PC distances are shown for all interior branches.

grounds. One such example is given in Figure 9.4A. The complete nucleotide sequence of mitochondrial DNA (mtDNA) is now available for the 11 vertebrate species given in this figure. The mtDNA in these species contains 13 protein-coding genes, and the number of shared codons among the 11 species varies from 52 to 582, depending on the gene. Russo et al. (1996) constructed a phylogenetic tree for each of these genes and for the entire set of genes (3682 codons) and compared the trees obtained with the true tree. Here we consider their results when amino acid sequences rather than nucleotide sequences were used, because the former produced more reliable trees than the latter.

When all 13 genes were used, all tree-building methods (NJ, ML, and MP) produced the correct tree irrespective of the algorithm used (Table 9.2). A few genes (usually large genes) such as *Nd5, Cytb,* and *Co3* also produced the correct or nearly correct topology. However, some genes (e.g., *Co2, Nd1, Nd3, Nd4ℓ*) almost always produced incorrect trees regardless of the method and algorithm used. This clearly indicates that some genes are more suitable than others in phylogenetic inference and that all tree-building methods tend to produce the same topology whether or not the topology is correct. Similar results were obtained by Kumazawa and Nishida (1995). Since there were only 13 genes studied, it was difficult to evaluate the relative efficiencies of the different tree-building methods. In general, however, sophisticated methods such as the ML method with Jones et al.'s (1992) substitution model were no better than simple methods such as NJ with p distance or ML star-decomposition algorithm. Similar results were obtained by Cao et al. (1994a, 1994b). These results suggest that the pattern and the rate of amino acid substitution vary with a group of organisms and also with evolutionary time and thus sophisticated mathematical models do not necessarily give better results. Somewhat similar studies were conducted by Cummings

Table 9.2 Topological distances (d_T) of reconstructed trees from the true tree for 13 mitochondrial proteins from the vertebrate species given in Figure 9.4A.

	Genes (No. codons)														
	Atp6 (219)	Atp8 (52)	Co1 (511)	Co2 (224)	Co3 (259)	Cytb (377)	Nd1 (312)	Nd2 (342)	Nd3 (112)	Nd4 (457)	Nd4l (97)	Nd5 (582)	Nd6 (138)	All (3682)	n_c[a]
NJ															
p	2	2	2	4	0	0	2	0	2	0	2	0	0	0	6
Poisson	2	0	0	4	2	0	2	2	0	0	2	0	0	0	7
Gamma	2	0	0	4	2	0	2	2	0	0	4	0	0	0	7
ME															
p	2	2	2	4	0	0	2	2	2	0	2	0	0	0	5
Poisson	2	0	2	4	2	0	2	2	2	0	2	0	0	0	5
MP															
Unweighted	1	2	2	2	0	0	2	2	4	0	4	0	2	0	4
Unweighted-b[b]	1	1	1	50	0	1	2	4	0	4	0	2	n.a.[c]	4	
Weighted	2	0	2	2	0	0	2	2	4	0	4	0	2	0	5
Weighted-b[d]	1	0	2	4	0	0	2	2	4	0	4	0	2	n.a.	5
ML star-decomposition															
Poisson	0	0	0	8	0	0	2	2	4	0	2	0	0	0	8
Dayhoff	2	2	0	8	0	0	2	2	2	0	4	0	4	0	5
JTT	0	0	0	8	0	0	2	2	2	2	2	0	4	0	6
JTT-f	0	2	2	8	0	0	2	2	2	2	2	0	2	0	4
ML specific-trees															
Poisson	0	2	0	10	0	0	2	2	4	0	6	0	0	0	7
Dayhoff	2	4	0	10	0	0	2	0	2	0	8	0	0	0	7
JTT	0	2	0	4	0	2	2	0	2	2	4	0	2	0	5
JTT-f	0	6	0	4	0	2	2	0	2	2	4	0	2	0	5

[a] n_c: Numbers of genes that produced the true tree.
[b] Unweighted-b: unweighted bootstrap consensus tree.
[c] n.a.: Because a large anount of computer time was required, the tree was not constructed.
[d] Weighted-b: weighted bootstrap consensus tree.

et al. (1995) and Zardoya and Meyer (1996) though in these studies the true topology was not firmly established.

A surprising result was obtained when the lamprey and sea urchin sequences were added to the 11 sequences in Figure 9.4A. That is, a clearly wrong tree (Figure 9.4B) was obtained by all tree-building methods even when all genes were used, and a bootstrap test showed strong statistical support for this wrong tree! This occurred apparently because wrong substitution models were used. Takezaki and Gojobori (1999) have shown that NJ and ML methods reconstruct the correct topology when a substitution model with a gamma parameter is used. A similar high bootstrap support for an apparently wrong topology has also been observed for a chloroplast gene phylogeny (Lockhart et al. 1992).

Phylogenetic Trees with Bootstrap Values

In recent years, there has been much debate about the relative efficiencies of different tree-building methods, especially the NJ, MP, and ML methods. It is now clear that there is no method that is superior to other methods in all conditions. We have seen that the theoretical basis of reconstruction of phylogenetic trees is not well established and that some methods perform better than others under certain conditions but do worse under other conditions (e.g., Table 9.1). In real data analysis, where the extent of sequence divergence is not very high and a substantial number of sequences are used, NJ, ML, and MP generally give the same or similar topologies. When there are topological differences, the differences are generally caused by interchanges of branches at those interior branches that have weak statistical support, and these interior branches can easily be identified by the bootstrap test (see Figures 7.12 and 8.3).

This suggests that if we construct a tree with bootstrap values any of the NJ, ME, and ML methods give essentially the same conclusion about the phylogenetic relationships of organisms or genes. Similar conclusions have been obtained for phylogenetic trees for various genes from different groups of organisms (Andersson et al. 1999; Honda et al. 1999; Meireles et al. 1999; Tan et al. 1999; Yagi et al. 1999). Unless the bootstrap values for all interior branches are high, we cannot really trust the branching pattern of a tree. In this case, a larger number of nucleotides or more different genes should be used to establish the tree statistically.

When one studies phylogenetic relationships of distantly related organisms or genes (e.g., animal, plants, and fungi), there are many complicating factors, and at the present time it is difficult to make any general conclusion about the relative efficiencies of different tree-building methods. In this case, the substitution pattern is unlikely to remain the same for a long evolutionary time. Investigators are often anxious to prove the validity of the tree obtained. We recommend that any tree should be subjected to various statistical tests, and only when none of the tests rejects it should it be accepted. A phylogenetic tree is a scientific hypothesis that should be subjected to attempts of falsification.

However, this does not mean that phylogenetics tree are worthless unless every interior branch has a high bootstrap value. Actually, every phylogenetic tree constructed is the best tree obtainable under the principle

of reconstruction used. Therefore, even if many interior branches of a tree are not well supported by the bootstrap, the tree should not be discarded. It is a hypothetical tree, but it could be a correct one. Actually, computer simulations have shown that many branching patterns of an inferred tree are correct even if they are not supported by high bootstrap values (e.g., tree C in Figure 6.8). When a tree is constructed for polymorphic gene sequences from a species, the bootstrap values are generally quite low (see Figure 12.3). Yet, the major aspects of the tree appear to be quite reliable (compare the trees in Figure 12.3 and Figure 13.3A). High bootstrap values are necessary when one wants to claim a specific branching pattern excluding other possibilities, as in the case of systematics.

Large Phylogenies for Closely Related Sequences

When the number of sequences (m) is large and the extent of sequence divergence is low, the realized tree may have many interior branches with zero length unless a large number of nucleotides are examined. In this case, it is generally difficult to reconstruct the true tree by any method. However, the bootstrap consensus tree often gives a reasonably good tree. Some authors have used the ML method with a sophisticated mathematical model in hopes of constructing a reliable tree. However, construction of ML trees is time-consuming, particularly when a bootstrap consensus tree is to be constructed. Construction of MP trees with the bootstrap test is also quite time-consuming when m is large. By contrast, if we use the NJ method, it is easy to construct a bootstrap consensus tree even if m is over 200.

The bootstrap consensus trees obtained by the three methods are usually very similar, although weakly supported interior branches may differ. Therefore, there is no need to spend an enormous amount of time for constructing ML or MP bootstrap consensus trees. In this case the same results are obtained by the NJ method very quickly. Note also that in this case the p distance is often better than other distance measures because it has a smaller variance.

10

Molecular Clocks and Linearized Trees

10.1. Molecular Clock Hypothesis

The **molecular clock** hypothesis asserts that the rate of amino acid or nucleotide substitution is *approximately* constant over evolutionary time, although the actual number of substitutions is subject to stochastic errors. Strictly speaking, no gene or no protein would evolve at a constant rate for a long evolutionary time, because the function of a gene is likely to change over time, particularly when the number of genes in the genome increases from simple organisms to complex ones or when environmental conditions change. The mechanisms of DNA damage and its repair may also vary among different groups of the organisms (Britten 1986).

For the above reasons, it would be futile to try to find genes that show a universal molecular clock. However, a molecular clock need not be universal. If it works for a certain group of organisms, it is still very useful for studying the evolutionary relationships of organisms or for estimating the time of divergence between different organisms. Many evolutionists are interested in molecular clocks for this reason.

However, the concept of the molecular clock has a long history of controversy, and it is still hotly debated by evolutionists. This controversy is centered around the accuracy of the clock and the mechanism of evolution. Although this book is not intended to cover this controversy, we briefly touch on it to give some background information for producing linearized trees that will be discussed later. A large part of the molecular clock controversy has been concerned with Kimura's (1983) neutral theory, but we do not discuss this issue because it is beyond the scope of this book.

Early Studies

The approximate constancy of the rate of amino acid substitution was first noted by Zuckerkandl and Pauling (1962, 1965), Margoliash (1963), and Doolittle and Blombäck (1964), who were working with hemoglobin, cytochrome *c*, and fibrinopeptides, respectively. Later similar observations were reported for many other proteins and some RNA molecules,

although the molecular clock did not always work very well (Dayhoff 1972).

The utility of the molecular clock was obvious from the beginning, and it was soon used for estimating the times of divergence of major groups of organisms. For example, Dickerson (1971) studied the evolutionary change of hemoglobin and estimated that plants, animals, and fungi diverged 1,000–1,200 MY ago. Using tRNA and 5S RNA sequence, McLaughlin and Dayhoff (1970) and Kimura and Ohta (1973) estimated that prokaryotes and eukaryotes separated 2,000–2,600 MY ago. Sarich and Wilson's (1966) immunological study of albumin suggested that humans and chimpanzees diverged about 5 MY ago. These estimates are all subject to large stochastic errors, but they are similar to the estimates recently obtained from many other genes (e.g., Horai et al. 1995; Gogarten et al. 1996; Feng et al. 1997; Kumar and Hedges 1998).

Despite these interesting findings, the idea of the molecular clock has been controversial. This controversy has occurred for a number of reasons. First, the constant rate of evolution was unthinkable for classical evolutionists, who had studied the evolution of morphological characters (Simpson 1964; Mayr 1965). At the time when the synthetic theory of evolution or neo-Darwinism was at its height, it was thought that the rate of evolution is controlled by environmental changes and natural selection and therefore should not be constant. Second, the mechanisms underlying the constant rate of amino acid substitution was unclear. Kimura (1968, 1969) and King and Jukes (1969) showed that if most amino acid or nucleotide substitution occurs by neutral mutations and genetic drift and the rate of neutral mutation is constant per year, the molecular clock can be explained. However, this explanation was hard to swallow for many geneticists and evolutionists, who were accustomed to the view that the mutation rate is constant per generation rather than per year in *Drosophila,* mice, humans, and corn (Lewontin 1974). Nei (1975) indicated that almost all mutations studied in classical genetics were deleterious and could be largely due to deletions or frameshift mutations that occur at the time of meiosis and that nondeleterious mutations (e.g., phage-resistant mutations) in bacteria appeared to occur at a constant rate per unit of chronological time rather than per cell division. However, the amount of data on the rate of nondeleterious mutations was too small to convince many geneticists or evolutionists of this view.

Third, as data on amino acid substitutions accumulated, the number of cases in which the molecular clock apparently did not apply increased (e.g., Laird et al. 1969; Jukes and Holmquist 1972; Goodman et al. 1974; Langley and Fitch 1974). Figure 10.1. shows one such example, in which the relationship between the estimated number of amino acid substitutions in hemoglobin α-chain and evolutionary time is presented. The relationship is approximately linear, but the protein divergence between birds and mammals clearly deviates from linearity.

Fourth, the divergence time inferred from paleontological data often unreliable, and this uncertainty introduced errors in the study of molecular clocks (Wilson et al. 1977). For this reason alone, many early studies seem to be unreliable (Easteal et al. 1995). Later, Sarich and Wilson (1967) and Fitch (1976) introduced a relative rate test using three se-

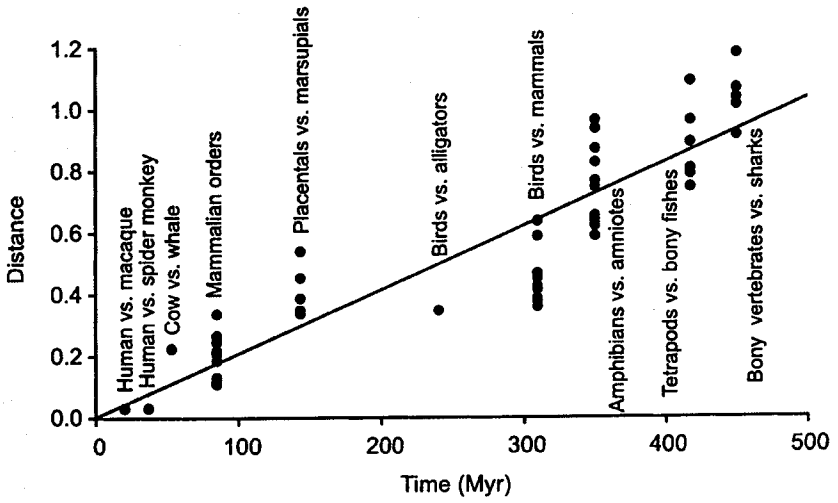


FIGURE 10.1. Relationships between Gamma distance (d_G, $a = 2$) and evolutionary time (t) for the amino acid sequences of vertebrate hemoglobin α chains. The regression line indicates $d = 2.1 \times 10^{-9}t$.

quences, one of which is known to be an outgroup. In this test, no knowledge of geological time estimates is necessary, and application of this method revealed deficiencies in several of the previous studies (Wilson et al. 1977).

Recent Controversies

In the 1980s, the controversy over the molecular clock intensified as DNA sequence data accumulated. Before then, Laird et al. (1969) and Kohne (1970) used DNA-DNA hybridization experiments to study the relationship between the number of nucleotide substitutions and evolutionary time. These studies suggested that mouse and rat DNA evolved much faster than other mammalian DNA, whereas human DNA evolved considerably slower. From these results, the authors proposed that the rate of nucleotide substitution is proportional to generation time rather than to chronological time, and this proposal had a strong influence on the general view about the molecular clock at that time. However, as indicated by Wilson et al. (1977) and Easteal et al. (1995), the divergence time between mice and rats used in these studies (15 MY) was questionable and that between humans and chimpanzees (15 MY) was apparently incorrect. Furthermore, there were some inaccuracies in the estimates of nucleotide substitution obtained from DNA-DNA hybridization data (Nei 1975; Easteal et al. 1995). Britten (1986) later compiled extensive data on DNA-DNA hybridization but still observed variation in evolutionary rate among different taxonomic groups. He proposed that this rate variation is caused by differences in the DNA repair system.

The generation time hypothesis proposed by Laird et al. (1969) was later examined by Wu and Li (1985) using DNA sequence data. They es-

timated the number of synonymous and nonsynonymous substitutions for humans, rodents (mice or rats), and an outgroup species (cows, rabbits, or horses) and applied a refined version of Sarich and Wilson's relative rate test. This study suggested that the rate of synonymous substitution was about two times faster in the rodent lineage than it was in the human lineage, whereas there was no significant difference in the nonsynonymous rate between the two lineages. Since synonymous substitutions would reflect the neutral mutation rate more closely than would nonsynonymous substitutions, they concluded that their findings support the generation time hypothesis.

This conclusion was soon challenged by Easteal (1985), who argued that the gene phylogeny used by Wu and Li was not well-established and rodents could be more distantly related to humans than to the outgroup species used (cows, rabbits, and horses), and that for this reason Wu and Li's conclusion may be incorrect. Analyzing paralogous (duplicate) globin genes, Easteal (1988) later showed that rodents are indeed more distantly related to human than cows, rabbits, and horses are, and that if this phylogeny is used, the evolutionary rate is nearly constant. This new gene phylogeny was then confirmed by Easteal (1990), Li et al. (1990) and Janke et al. (1994) using marsupials or chicken as outgroups. Furthermore, when this phylogeny was used, both the Easteal and the Li groups failed to find a significant difference in synonymous rate between the human and the rodent lineages. However, the nonsynonymous rate was about 1.5 times higher in rodent lineage than in the human lineage. Li et al. (1996) argued that the nonsignificant difference in synonymous rate is probably due to saturation of synonymous substitutions in the comparison between marsupials and humans and between marsupials and rodents.

Another issue, which has been hotly debated, is the slowdown of evolutionary rate in the human lineage, a hypothesis originally proposed by Goodman (1962). This problem has also been studied intensively by the Li and the Easteal groups using the relative rate test. Li et al. (1987) analyzed a contiguous sequence of about 2 kilobases (kb) in the η globin pseudogene region and concluded that the substitution rate was faster in the Old World monkey than in the human lineage. Easteal (1991) confirmed this difference in a sequence of about 10 kb of the same gene region but found no apparent difference for 18 other genes (a total of about 6 kb). Seino et al. (1992) then reported that the substitution rate in intron regions of the insulin gene is higher in Old World monkeys than in humans. Li et al. (1996) extended this study to the intron regions of seven more genes and confirmed their previous conclusion (about 1.3 times higher rate in the Old World monkey than in the human lineage). Herbert and Easteal (1996) reexamined this problem with more data. They confirmed Li et al.'s conclusion for the intron regions (21.2 kb), but their studies of 1,592 synonymous sites and 5,275 nonsynonymous sites showed no significant differences between the Old World and human lineages.

Here we have discussed the debate between the Li and the Easteal groups in some detail to show that the molecular clock is a complex issue and even an intensive study does not necessarily resolve the contro-

versy. The two groups have been persistent on detailed aspects of this issue, primarily because resolution of the controversy is important for understanding the mechanism of molecular evolution. At the present time, it is still unclear whether or not the rate of nondeleterious mutations depends solely on errors of DNA replication. Mutational change of a nucleotide is a complex process consisting of DNA damage and its repair, and this process itself is likely to have undergone evolutionary change. The readers who are interested in more detailed aspects of this controversy should refer to Wilson et al. (1977), Britten (1986), Easteal et al. (1995), and Li et al. (1996). For the purpose of this chapter, it is sufficient to know that when many genes are studied, the extent of rate heterogeneity among different mammalian orders is quite small at least for the genes studied so far (e.g., Kumar and Hedges 1998).

Of course, this does not mean that the heterogeneity of evolutionary rate is small in most groups of organisms. Actually, if one is concerned with a specific gene, it is easy to find cases of varying evolutionary rates. For example, the alcohol dehydrogenase (*Adh*) gene seems to have evolved much faster in Hawaiian *Drosophila* than in *D. pseudoobscura,* but in Hawaiian *Drosophila,* the evolutionary rate appears to have been nearly constant (Russo et al. 1995). The mitochondrial coding genes in fish are also known to have evolved at a much slower rate than those in mammals (Martin et al. 1992; Rand 1994). Figure 9.4A shows the mitochondrial protein tree for 11 animal species. This tree is based on protein sequences of 3,619 amino acids from 13 coding genes. The topology of the tree is the same as that of the biological tree established by paleontological and morphological data, but the branch lengths estimated by the least squares method show that the mitochondrial genes evolved about three times slower in the fish lineage than in the mammalian lineage.

10.2. Relative Rate Tests

The relative rate test is often credited to Sarich and Wilson (1967) or even to Margoliash (1963), but these authors did not conduct a statistical test of the rate difference between two lineages. Fitch (1976) proposed a simple form of statistical test using this approach, but the tests that are more efficient were developed later. Here we first discuss model-based relative rate tests and then nonparametric tests proposed by Gu and Li (1992) and Tajima (1993a).

Model-Based Relative Rate Tests

Let us consider the phylogenetic tree in Figure 10.2, in which sequence 3 is known to be an outgroup for sequences 1 and 2. We denote by d_{12}, d_{13}, and d_{23} the estimates of amino acid or nucleotide substitutions between sequences 1 and 2, 1 and 3, and 2 and 3, respectively, and write $d_{12} = x_1 + x_2$, $d_{13} = x_1 + x_3$, and $d_{23} = x_2 + x_3$, where x_1, x_2, and x_3 are estimates of the number of substitutions for the branches as indicated in Figure 10.2. Therefore, x_1, x_2, and x_3 are given by Equations (6.5). Figure

10.2 indicates that if the rate of substitution is constant, the expectations of x_1 and x_2 are equal to each other. In practice, however, both x_1 and x_2 are subject to statistical errors, and it is necessary to test the null hypothesis $E(x_1) = E(x_2)$. This hypothesis can be tested by examining the statistical significance of

$$D = x_1 - x_2 = d_{13} - d_{23} \tag{10.1}$$

The variance of this D is given by

$$V(D) = V(d_{13}) - 2Cov(d_{13}, d_{23}) + V(d_{23}) \tag{10.2}$$

When the Poisson correction (PC) distance is used for amino acid sequence data, the variance of d_{ij} is given by Equation (2.6), whereas the covariance between the distances for sequences i and j (d_{ij}) and sequences k and l (d_{kl}) is given by

$$Cov(d_{ij}, d_{kl}) = \frac{p_{ij\cdot kl} - p_{ij}p_{kl}}{n(1 - p_{ij})(1 - p_{kl})} \tag{10.3}$$

where p_{ij} is the proportion of different amino acids between sequences i and j, and $P_{ij\cdot kl}$ is the proportion of the sites at which i differs from sequence j and sequence k differs from sequence l (Ota and Nei 1994b). When the gamma distance (d_{Gij}) is used, the variance of the distance is given by Equation (2.13) and the covariance of d_{Gij} and d_{Gkl} is given by Ota and Nei's formula. Similarly, when nucleotide sequence data are used, the variances for various distances can be computed by the formulas given in chapter 3 and the MEGA manual (Kumar et al. 1993), whereas the covariances are obtained by Bulmer's (1991) formula and its extensions (see Rzhetsky and Nei 1992a). For example, the covariance for the Jukes-Cantor distance d_{ij} and d_{kl} is given by

$$Cov(d_{ij}, d_{kl}) = \frac{p_{ij\cdot kl} - p_{ij}p_{kl}}{n(1 - p_{ij}/b)(1 - p_{kl}/b)} \tag{10.4}$$

where $b = 1/4$, and p_{ij} and $p_{ij\cdot kl}$ are equivalent to those defined for amino acid sequences.

Therefore, one can easily compute $V(D)$ in Equation (10.2), and once D and $V(D)$ are obtained, the statistical significance of the difference of D from 0 can be tested by the Z test. However, a simpler way of computing $V(D)$ is to use the bootstrap method as described in chapter 2. In this method, we can compute D for each bootstrap sample and obtain $V(D)$ directly. This approach is convenient, because no analytical formulas are needed.

Wu and Li (1985) used Kimura's (1980) model to compute synonymous and nonsynonymous nucleotide substitutions in the human and mouse lineages and then tested the molecular clock by the Z test. As mentioned earlier, their results are no longer valid, because the outgroup species used were apparently incorrect. However, it is interesting to compare the rates of synonymous and nonsynonymous substitutions between any

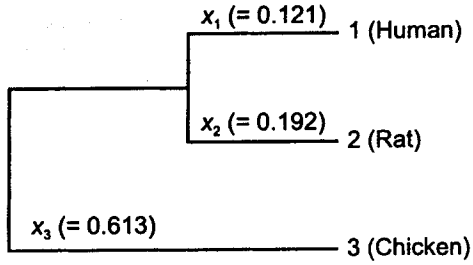


FIGURE 10.2. Relative rate test for three sequences. The example given here is the albumin amino acid sequences for humans, rats, and chickens. Albumin sequences were obtained from the GenBank and the final aligned set of sequences consisted of 608 amino acid sites. The PC distance was used.

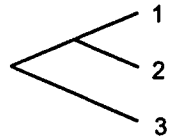
pair of evolutionary lineages, as long as an outgroup species is available. This can be done by using any formulas for d_N and d_S in chapter 4 and computing the variances of these quantities by the bootstrap. As was done by Wu and Li, it is also possible to consider only the 4-fold degenerate sites for synonymous substitutions and the nondegenerate sites for nonsynonymous substitutions.

Muse and Weir (1992) introduced a likelihood method of relative rate tests using Hasegawa et al.'s (1985b) model of nucleotide substitution. In this method, the branch lengths x_1 and x_2 of the tree in Figure 10.2 are estimated by the likelihood method discussed in chapter 8 with and without the assumption of $E(x_1) = E(x_2)$. Let $\ln L_1$ and $\ln L_2$ be the mL values for the case with the assumption and without the assumption, respectively. The null hypothesis $E(x_1) = E(x_2)$ can then be tested by examining the likelihood ratio $LR = 2(\ln L_2 - \ln L_1)$, which approximately follows a χ^2 distribution with one degree of freedom.

This test seems to be appropriate when relatively closely related RNA sequences or noncoding regions of DNA are used. In general, however, the actual pattern of nucleotide substitution is much more complicated than Hasegawa et al.'s model. It is known that when the mathematical model used is inappropriate for the data set to be examined, the likelihood ratio test can be too liberal or too conservative depending on the situation (Zhang 1999). Therefore, one should be cautious about the interpretation of the results of this test.

Nonparametric Tests

The above problem can be avoided if we use a nonparametric test, which requires no mathematical model. Gu and Li (1992) used a simple nonparametric test of a molecular clock for amino acid sequence data, whereas Tajima (1993a) developed a general method for testing the molecular clock hypothesis for both nucleotide and amino acid sequences. The principles of the two methods are the same, but here we present Tajima's general method. Consider three DNA sequences 1, 2, and 3 and assume that sequence 3 is the outgroup. We consider the nucleotide configuration at each nucleotide site for the three sequences. There are five different types of configurations, which are presented in Figure 10.3.



	Nucleotide configuration				
	(a)	(b)	(c)	(d)	(e)
1	*i*	*i*	*j*	*j*	*i*
2	*i*	*j*	*i*	*j*	*j*
3	*i*	*j*	*j*	*i*	*k*
Sum	m_0	m_1	m_2	m_3	m_4
	(238)	(34)	(54)	(207)	(75)

FIGURE 10.3. All possible nucleotide configurations for three sequences. The numbers in parentheses are those obtained for the human (1), rat (2), and chicken (3) albumin sequences.

Configuration type (a) in this figure represents the case where all three sequences have the same nucleotide (i = T, C, A, or G). Configuration type (b) stands for the case where sequence 1 has one kind of nucleotide (say i = T), and sequences 2 and 3 have another kind of nucleotide (j = C, A, or G). Configuration types (c), (d), and (e) are similarly defined. Since sequence 3 is the outgroup, configuration (b) suggests that substitution has occurred in sequence 1, whereas configuration (c) suggests that substitution has occurred in sequence 2. Configurations (a), (d), and (e) are not informative for the comparison of the substitutions in branches 1 and 2 in which we are interested.

Let n_{ijk} be the observed number of sites where sequences 1, 2, and 3 have nucleotides i, j, and k, respectively. Under the molecular clock hypothesis, the expectation of n_{ijk} must be equal to that of n_{jik}, that is

$$E(n_{ijk}) = E(n_{jik}) \tag{10.5}$$

irrespective of the substitution model and whether or not the substitution rate varies with site. Therefore, if we can reject the null hypothesis $E(n_{ijk}) = E(n_{jik})$, we can reject the molecular clock hypothesis. This is the basis of Tajima's test.

To develop a simple test of the molecular clock hypothesis, let us consider the following sums of nucleotide sites with configuration types (b) and (c) (Figure 10.3).

$$m_1 = \sum_i \sum_{j \neq i} n_{ijj} \tag{10.6a}$$

$$m_2 = \sum_i \sum_{j \neq i} n_{jij} \tag{10.6b}$$

Under the molecular clock hypothesis, $E(m_1)$ is equal to $E(m_2)$. Therefore, the molecular clock hypothesis can be tested by using the following X^2 statistic with one degree of freedom.

$$X^2 = \frac{(m_1 - m_2)^2}{m_1 + m_2} \tag{10.7}$$

Strictly speaking, the above quantity is not a χ^2 statistic but approximately follows a χ^2 distribution when m_1 and m_2 are greater than 5 (chapter 4). When m_1 and m_2 are small, we should use Fisher's exact test.

Obviously, the same test can be used for amino acid sequence data, and in this case the test becomes identical with the one used by Gu and Li (1992). This X^2 test looks similar to Fitch's (1976) method, but it can be shown that the latter is less powerful.

The above model-free test can be applied to a more general case, as long as quantities similar to those of Equations (10.6) are definable. Tajima (1993b) considered the case where transitional and transversional nucleotide substitutions are distinguished. In this case, m_i is divided into the number of sites (s_i) with transitional differences and the number of sites (v_i) with transversional differences. Therefore, we consider the following quantities.

$$s_1 = n_{AGG} + n_{GAA} + n_{TCC} + n_{CTT}$$

$$v_1 = m_1 - s_1$$

$$s_2 = n_{AGA} + n_{GAG} + n_{CTC} + n_{TCT}$$

$$v_2 = m_2 - s_2$$

We have assumed that sequence 3 is the outgroup, so the null hypothesis to be tested is $E(s_1) = E(s_2)$ and $E(v_1) = E(v_2)$. We can therefore test this hypothesis using the following X^2 with two degrees of freedom.

$$X^2 = \frac{(s_1 - s_2)^2}{s_1 + s_2} + \frac{(v_1 - v_2)^2}{v_1 + v_2} \tag{10.8}$$

This test is much simpler than the Wu-Li and the Muse-Weir tests but is known to be as powerful as the latter tests (Tajima 1993a). Unlike the latter two tests, Tajima's test is applicable irrespective of the substitution model used; there is no need to specify the substitution model in this test. The only assumption necessary is the stationarity of nucleotide substitution. However, this test cannot be used for synonymous and nonsynonymous substitutions unless the numbers of these substitutions are countable (see chapter 4).

Example 10.1. Relative Rate Tests for the Albumin Sequences from Humans, Rats, and Chickens

Figure 10.2 shows the phylogenetic tree for the albumin protein sequences from humans (1), rats (2), and chicken (3). The distance used here is the PC distance. The distances between the sequences are $d_{12} = 0.312$, $d_{13} = 0.733$, and $d_{23} = 0.804$. We therefore have $D = d_{23} - d_{13} = 0.071$. The variance of d_{13} is 0.0017799 from Equation (2.6). Similarly,

the variance of d_{23} is 0.0020317. $Cov\,(d_B, d_{23})$ becomes 0.0013518 from Equation (10.3). Therefore, we have the standard error $s(D) = 0.033$ and $Z = 2.13$, which is significant at the 5% level. This indicates that albumin evolved faster in the rat lineage than in the human lineage.

The same conclusion can be reached with the nonparametric test. In this case, we have $m_1 = 34$ and $m_2 = 54$, so $X^2 = 4.5$ from Equation (10.7). This is again significant at the 5% level. Since the nucleotide sequences are available in this case, we can test the molecular clock using the first and second codon position data. We exclude the third codon position data, because nucleotide substitutions at the third positions are nearly saturated when the chicken and mammalian sequences are compared. We then obtain $s_1 = 25$, $s_2 = 30$, $v_1 = 26$, and $v_2 = 47$, and X^2 becomes $0.5 + 6.0 = 6.5$. This X^2 is also significant at the 5% level, but this high X^2 value is caused by the fact that transversions are more frequent in rats than in humans.

Some Remarks

As mentioned earlier, the relative rate test is for examining the molecular clock hypothesis for two evolutionary lineages, that is, lineages 1 and 2 of Figure 10.2. It does not give any information about the evolutionary rate of lineage 3. Therefore, even if the null hypothesis is not rejected, we cannot assume that the same rate applies for lineage 3. To test whether the evolutionary rate is the same for the three lineages, we need outgroup sequences for the three sequences.

Note also that the relative rate test may not detect rate variation within lineages. In other words, if both lineages 1 and 2 in Figure 10.2 experienced a period of high evolutionary rate or of low evolutionary rate, this rate variation may not be detected. This could happen when one compares two distantly related lineages such as the primate and rodent lineages with chicken as an outgroup. It is therefore important to use many species to test a molecular clock hypothesis.

10.3. Phylogenetic Tests

The **phylogenetic tests** to be discussed below are designed to avoid the above problems, and a clock hypothesis is tested by using many sequences simultaneously. If we use many sequences, the molecular clock hypothesis can be rejected more easily. A number of authors have developed such tests (see Nei 1996), but here we present simple ones that are useful for the construction of linearized trees.

Two-Cluster Test

Takezaki et al. (1995) developed two simple tests of the molecular clock hypothesis: the **two-cluster test** and the **branch-length test**. These two tests are intended to be applied to a tree of which the topology has been determined by some tree-building method without the assumption of rate constancy, and the root has been located by using outgroup se-

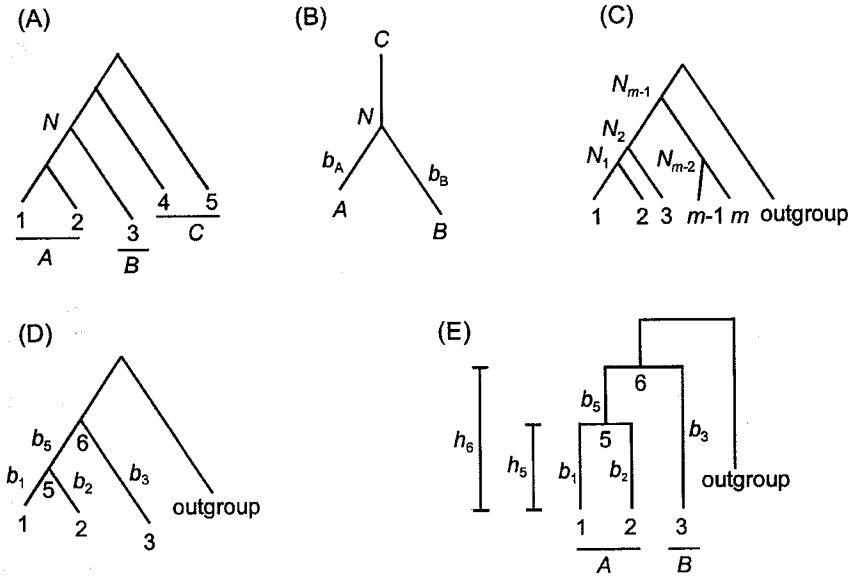


FIGURE 10.4. Tests of the molecular clock hypothesis and construction of a linearized tree. (A) A bifurcating tree connecting three clusters A, B, and C. Cluster C is the outgroup of A and B. (B) Testing whether the average distances (b_A and b_B) from node N to the tips of clusters A and B are significantly different. (C) There are $m - 1$ interior nodes in a tree consisting of m sequences when outgroup sequences are excluded. (D) Test of the difference between the average distances from node N to the tips of clusters A and B. (E) Construction of a linearized tree by estimating the heights of interior nodes.

quences. The two-cluster test can be considered an extension of the relative rate test to the case of multiple sequences and is somewhat similar to Li and Bousquet's (1992) test for two groups of sequences (see also Robinson et al. 1998; Tourasse and Li 1999). The principle of this test is to examine whether or not the average substitution rates for the two clusters of sequences that are created by a node of a given tree are the same. Let us consider the cluster A and B created by node N in the tree in Figure 10.4A. Here, sequences 1 and 2 belong to cluster A and sequence 3 to cluster B. We denote the remaining sequences (4 and 5) by C and regard them as outgroup sequences (Figure 10.4B). In practice, any number of sequences may be included in clusters A, B, and C.

Let b_A and b_B be the estimates of the average numbers of substitutions per site (distance) from node N to the tips of the clusters A and B, respectively. Under the assumption of rate constancy, the expectation of the difference (y) between b_A and b_B is zero. Let L_{AB}, L_{AC}, and L_{BC} be the average distances between clusters A and B, A and C, and B and C, respectively. That is,

$$L_{AB} = \sum_{i \in A; j \in B} \frac{d_{ij}}{n_A n_B} \qquad (10.9a)$$

$$L_{AC} = \sum_{i \in A; j \in C} \frac{d_{ij}}{n_A n_C} \tag{10.9b}$$

$$L_{BC} = \sum_{i \in B; j \in C} \frac{d_{ij}}{n_B n_C} \tag{10.9c}$$

where d_{ij} is the distance between sequences i and j, and n_A, n_B, and n_C are the numbers of sequences that belong to clusters A, B, and C, respectively. Here, the notation $i \in A$, and so forth, mean that sequence i belongs to cluster A, and so on. for the clusters in Figure 10.4A, they are calculated as follows: $L_{AB} = (d_{13} + d_{23})/(2 \times 1)$, $L_{AC} = (d_{14} + d_{15} + d_{24} + d_{25})/(2 \times 2)$, and, $L_{BC} = (d_{34} + d_{35})/(1 \times 2)$. b_A and b_B can then be estimated by

$$b_A = (L_{AB} + L_{AC} - L_{BC})/2 \tag{10.10a}$$

$$b_B = (L_{AB} + L_{BC} - L_{AC})/2 \tag{10.10b}$$

Therefore, y is given by

$$y = b_A - b_B = L_{AC} - L_{BC} \tag{10.11}$$

and the variance $[V(y)]$ of y is

$$\begin{aligned} V(y) = & \left[\sum_{i \in A; j \in C} V(d_{ij}) + 2 \sum_{i,k \in A; j,l \in C} Cov(d_{ij}, d_{kl}) \right] / (n_A n_C)^2 \\ & + \left[\sum_{i \in B; j \in C} V(d_{ij}) + 2 \sum_{i,k \in B; j,l \in C} Cov(d_{ij}, d_{kl}) \right] / (n_B n_C)^2 \\ & - 2 \sum_{i \in A; k \in B; i,l \in C} Cov(d_{ij}, d_{kl}) / (n_A n_B n_C^2) \end{aligned} \tag{10.12}$$

We can use the Z test to evaluate the significance of the difference of y from 0.

If we use the Jukes-Cantor model of nucleotide substitution, the distance (d_{ij}) between a pair of sequences i and j is estimated by Equation (3.8), whereas the variance $V(d_{ij})$ and the covariance $Cov(d_{ij}, d_{kl})$ are given by Equations (3.9) and (10.4), respectively. Therefore, we can compute $V(y)$ for any value of n_A, n_B, and n_C. Note that in the above computation the test for the rate difference between two lineages created by a node does not depend on the branching order of sequences within each of the clusters A, B, and C. As long as the three clusters are definable, we can use this test even if the branching order within each cluster is not very reliable. $V(y)$ can be computed for many other substitution models as long as $V(d_{ij})$ and $Cov(d_{ij}, d_{kl})$ are computable. However, it is much easier to compute $V(y)$ by the bootstrap method.

In the above formulation, we considered the test of y for one pair of clusters. In practice, it is possible to test all y's simultaneously. When we have m sequences excluding the outgroup(s), there are $m - 1$ interior nodes for which we can compute y (see Figure 10.4C). The null hypoth-

esis for this simultaneous test is $E(y_1) = E(y_2) = \cdots = E(y_{m-1}) = 0$, where $E(y_i)$ is the expectation of y for the i-th interior node. Let us denote by $\boldsymbol{y}$ a column vector whose elements are $y_1, y_2, \cdots, y_{m-1}$ and by $\boldsymbol{V} = [v_{ij}]$ its variance-covariance matrix, where $v_{ij} = Cov(y_i, y_j)$. We can then test the null hypothesis using the following statistic.

$$\boldsymbol{U} = \boldsymbol{y}^t \boldsymbol{V}^{-1} \boldsymbol{y} \tag{10.13}$$

where t and -1 stand for the transpose and the inverse of a matrix, respectively (Takezaki et al. 1995). Since the joint distribution of y_i's is close to a multivariate normal distribution, $\boldsymbol{U}$ approximately follows the χ^2 distribution with $m - 1$ degrees of freedom under the null hypothesis. Therefore, we can determine the significance level of $\boldsymbol{U}$.

Branch Length Test

In this test, we examine the deviation of the root-to-tip distance (sum of branch lengths [b_i's] from the root to each sequence) from the average for all sequences except for the outgroup sequence(s). Let us denote by x_i the root-to-tip distance for the i-th sequence. In the tree shown in Figure 10.4D, $x_1 = b_1 + b_5$, $x_2 = b_2 + b_5$, and $x_3 = b_3$. The average ($\bar{x}$) of the x_i's is $\bar{x} = (b_1 + b_2 + b_3 + 2b_5)/3$. If rate constancy holds, the expectation of the difference ($z_i = x_i - \bar{x}$) between x and $\bar{x}$ is zero. Therefore, the deviation of z_i from zero can be tested by the normal deviate statistic (Z).

We estimate the branch lengths of a given tree topology by the ordinary LS method discussed in chapter 6. Since the estimates of branch lengths are a linear combination of pairwise distances (d_{ij}'s), the value of z_i can also be expressed as a linear combination of d_{ij}'s.

$$z_i = \sum_{i<j} a_{ij} d_{ij} \tag{10.14}$$

where a_{ij} is a constant associated with d_{ij}. An analytical formula for computing the variance of z_i is available, but a simpler way of computing the variance is to use the bootstrap. In this method, $V(z)$ is given by

$$V(z_i) = \frac{1}{B-1} \sum_{k=1}^{B} (z_k^* - \bar{z}^*)^2 \tag{10.15}$$

where B is the number of the bootstrap replications, z_k^* is the value of z_i estimated at the k-th bootstrap replication, and $\bar{z}^*$ is the average of z_k^*'s. In each bootstrap replication, we resample the same number of sites as that of the original data with replacement, as in the case of Dopazo's test of interior branch lengths discussed in chapter 9.

To construct a linearized tree, we eliminate sequences that have evolved significantly faster or slower than the average. After elimination of these sequences, the average root-to-tip distance may change. Therefore, we must reestimate the branch length for the remaining sequences by the LS method and conduct the rate constancy test again. This process is repeated until all sequences show no significant rate heterogeneity. Of course, we may retain certain important sequences even if

they evolve significantly faster or slower than the average, as long as they do not distort the linearized tree substantially (see Example 10.2).

This method can be extended to the test of rate constancy (1) among the clusters of sequences by redefining x as the average root-to-tip distance for a cluster and $\bar{x}$ as the average of all x's or (2) between two clusters by letting $z = x_A - x_B$, where x_A and x_B are the average root-to-tip distances within clusters.

As in the case of the two-cluster test, we can test the hypothesis of rate constancy for a set of sequences. That is, the null hypothesis $E(z_1) = E(z_2) = \cdots = E(z_n) = 0$ can be tested by the $\boldsymbol{U}$ statistic in Equation (10.13) with $m - 1$ degrees of freedom.

In the construction of linearized trees, it is more convenient to use the branch-length test rather than the two-cluster test, because the former can easily identify the deviant sequences to be eliminated. In this case, we recommend that the $\boldsymbol{U}$ statistic for testing the null hypothesis $E(z_1) = E(z_2) = \cdots = E(z_n) = 0$ first be computed. If this $\boldsymbol{U}$ is not significant, one may proceed to construction of a linearized tree. If it is significant, the deviant sequences should be identified by testing the deviation of z_i from $\bar{z}$. (See *http://mep.bio.psu.edu*)

Example 10.2. Tests of the Molecular Clock Hypothesis for *Drosophila Adh* Genes

The genus *Drosophila* includes more than 1,300 species of flies and has been used extensively for the study of speciation (Powell 1997). Russo et al. (1995) and Takezaki et al. (1995) studied the phylogeny and the times of divergence of 39 drosophilid species using the alcohol dehydrogenase gene (*Adh*) sequences. Here we consider 23 of the species (see Figure 10.5A) to illustrate how to conduct the two-cluster and the branch length tests.

The results of rate-constancy tests are presented in Figure 10.5A. We first constructed the MP, NJ, and ML trees for the species using 23 aligned *Adh* gene sequences with 759 nucleotides and *Scaptodrosophila lebanonensis* as the outgroup. For constructing the NJ and ML trees, we used the Jukes-Cantor model, because the extent of sequence divergence was relatively small. To topologies of the NJ and ML trees were identical with each other, whereas there were two MP trees, one of which was identical with the NJ/ML tree. In the following analysis, we use the NJ tree because the branch lengths of this tree have been estimated by the LS method.

We first applied the two-cluster test for every pair of sequences or sequence clusters, excluding *S. lebanonensis*. Each interior node corresponding to two clusters that showed a significant rate difference at the 5% or lower level is indicated by the P (Type I error) value. For example, node *a* generates cluster A (*D. mettleri* and *D. hydei*) and cluster B (*D. arizonae*, *D. navojoa*, *D. mojavensis*, and *D. mulleri*), and we obtained $b_A = 0.0539$, $b_B = 0.0692$, $y = b_B - b_A = 0.0153$, and $s(y) = 0.0075$. Therefore, Z is 2.03, which corresponds to a P value of 5%. Thus, this node is given a value of 0.05 in the tree. The tree in Figure 10.5A shows that there are seven nodes showing a P value of 0.05 or less and two nodes

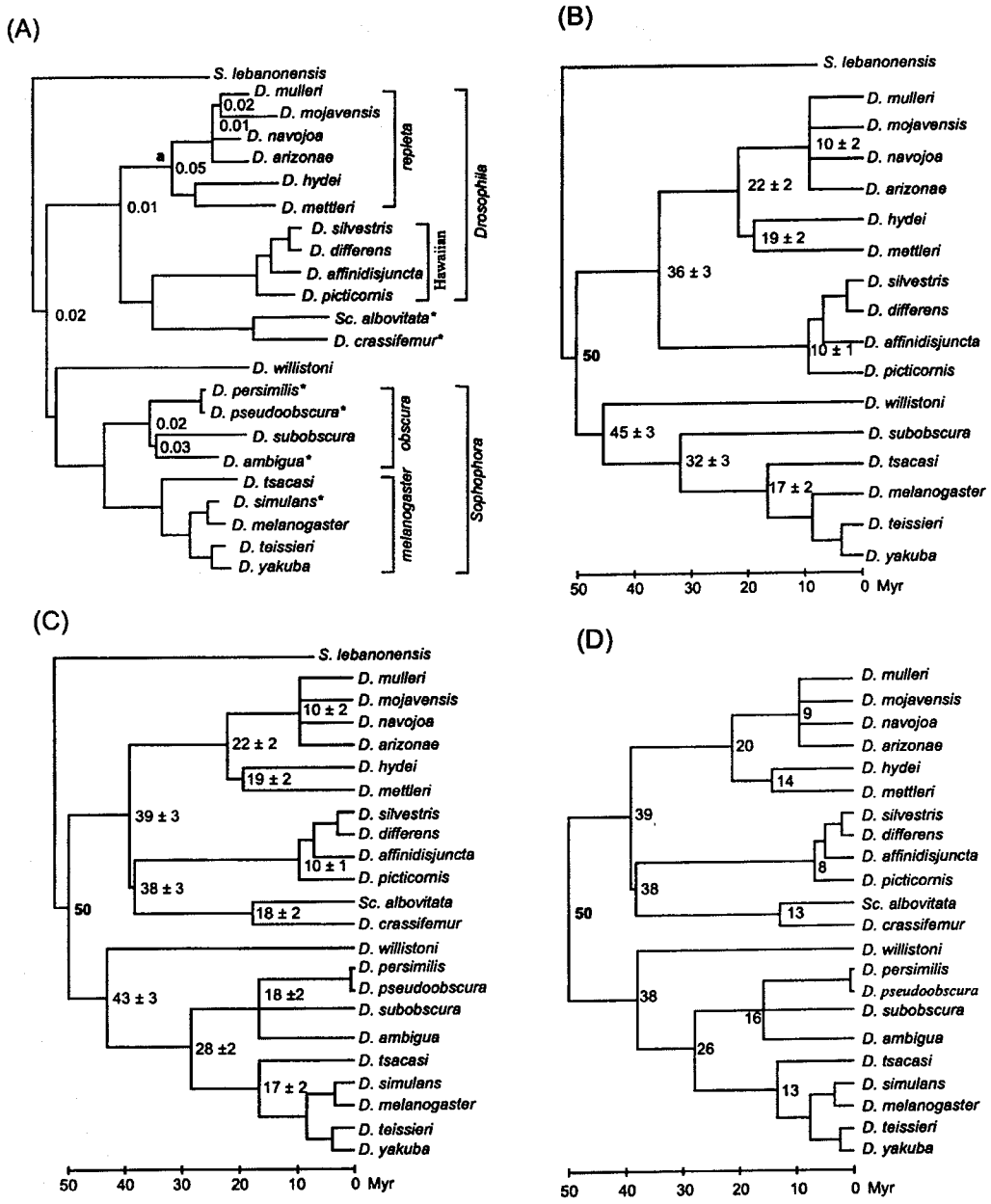


FIGURE 10.5. (A) An NJ tree based on Jukes-Cantor distances for 23 *Adh* gene sequences. Results from two-cluster tests shown next to the corresponding nodes indicate the significance level for the rejection of the molecular clock hypothesis. Species with asterisks evolved significantly slower or faster than the average rate at the 1% level in the root-to-tip tests. (B) Linearized tree for 17 species after elimination of six species. (C) Linearized tree for all 22 species. (The outgroup is excluded.) (D) Maximum likelihood tree for 22 species constructed by assuming a molecular clock. Estimated times of divergence and their standard errors are shown for linearized trees.

showing a P equal to or less than 0.01. Therefore, the molecular clock hypothesis is clearly rejected.

We also applied the branch length test for every sequence, and this test showed that the six sequences indicated by the asterisk (*) evolved significantly faster or slower than the average at the 1% level. Therefore, this test also rejects the molecular clock hypothesis. However, the deviant sequences indicated by this test are not always the same as those indicated by the two-cluster test. For the purpose of constructing a linearized tree, the branch length test is more convenient, but since this test often rejects too many sequences, we may retain all sequences that are not significant at the 1% or even the 0.1% level, depending on the sequence length used.

Previously, we indicated that the null hypothesis of $E(y_1) = E(y_2) = \cdots = E(y_{m-1}) = 0$ can be tested by U in Equation (10.13). The U statistic was 51.7 with 21 degrees of freedom, which is significant at the 1% level. The U statistic for testing $E(z_1) = E(z_2) = \cdots = E(z_{m-1}) = 0$ is known to give the same value as that for the two-cluster test (A. Rzhetsky and P. Xu, unpublished).

Likelihood Ratio and Some Other Tests

The hypothesis of rate constancy can also be tested by computing the ML values with and without the assumption of rate constancy and the likelihood ratio (LR) for the two cases (Felsenstein 1988). Goldman (1993) questioned the χ^2 approximation of the test statistic (LR), but Yang et al.'s (1994) simulation study suggests that the χ^2 approximation is acceptable in most cases, as long as the model of nucleotide substitution used is appropriate (see also Yang and Kumar 1996). This test, however, is not designed to identify deviant sequences that evolve very fast or very slow, so it is cumbersome to use this test for constructing linearized trees when there are many sequences. This test is also more time-consuming than the two-cluster or the branch length test.

Uyenoyama (1995) proposed that the generalized LS method be used for testing the deviation of the root-to-tip distances from the average. Since the generalized LS estimates of branch lengths have a smaller variance than that of the ordinary LS estimates, her test is probably more powerful than Takezaki et al.'s (1995). However, when the number of sequences is very large, application of the method becomes difficult, because the generalized LS estimation of branch lengths requires a large amount of computational time. Furthermore, in the construction of linearized trees, we are not overly concerned with the power of the test of a molecular clock.

Another method (Felsenstein 1984, 1988) for testing the molecular clock is to compare the LS residual sum under the assumption of rate constancy (R_C) with that for the case of no such assumption (R_N) using the following statistic.

$$F = \frac{(R_C - R_N)/(m-2)}{R_N / [m(m-1)/2 - (2m-3)]} \qquad (10.16)$$

This statistic is assumed to follow the F distribution with the degrees of freedom of $m - 2$ and $m(m - 1)/2 - (2m - 3)$, where $m \geq 4$. When the ordinary or weighted LS method is used to compute R_C and R_N (FITCH and KITCH programs in the PHYLIP package), it is implicitly assumed that pairwise distance estimates are independently and normally distributed. Normality may not be seriously violated when the number of substitutions is large. However, pairwise distances are positively correlated because of the historical relationships of the sequences. Therefore, it is unclear how reliable this test is (Felsenstein 1988, 1995).

10.4. Linearized Trees

Although the rate of nucleotide or amino acid substitution would never be strictly constant over long evolutionary time, an approximate molecular clock is still useful for estimating the time of divergence between sequences. At the present time, very little is known about the time of species divergence or gene duplication events, so even rough estimates of divergence times are important for understanding evolutionary processes. In the foregoing sections, we discussed several statistical methods for detecting evolutionary lineages that evolve significantly faster or slower than the average rate. Therefore, if we eliminate these lineages and construct a tree for the remaining ones under the assumption of a molecular clock, we may be able to obtain rough estimates of times of divergence between different pairs of lineages or sequences. The tree constructed in this way is called a **linearized tree.** It is different from a UPGMA tree, which is constructed under the assumption of rate constancy from the beginning.

Construction of a linearized tree consists of the following four steps. (1) A reliable topology for a set of sequences is constructed by using some tree-building method without the assumption of rate constancy, and the root of the tree is established by using outgroup sequences. (2) The hypothesis of rate constancy is tested for the sequences used, and the sequences whose evolutionary rate significantly deviates from the average rate (deviant sequences) are eliminated. (3) A phylogenetic tree for the remaining sequences is then reconstructed under the assumption of rate constancy. (4) If the time of divergence and the extent of sequence divergence are known for a pair of sequences, we can calibrate the evolutionary time.

In the foregoing section, we have already explained the steps (1) and (2), and what remains to be explained are steps (3) and (4). Step (3) is for constructing a linearized tree for the remaining sequences after elimination of deviant sequences. In this case, we must estimate the branch lengths of the remaining sequences for the original tree topology. Let us explain the procedure of branch length estimation using diagrams in Figure 10.4E. Under the assumption of rate constancy, we can compute the height (h) of the branch point of clusters A and B from the tip of the tree and the variance [$V(h)$] of h by

$$h = \frac{L_{AB}}{2} \tag{10.17}$$

$$V(h) = \left[\sum_{i \in A; j \in B} V(d_{ij}) + 2 \sum_{i,k \in A; j,l \in B} Cov(d_{ij}, d_{kl}) \right] / (2n_A n_B)^2 \qquad (10.18)$$

In the case of Figure 10.4E, the heights (h_5 and h_6) of nodes 5 and 6 are given by $d_{12}/2$ and $(d_{13} + d_{23})/4$, respectively. We estimate the heights of all interior nodes below the root. For an exterior branch connected to a node, the branch length is given by the height of the node. Thus, we have $b_1 = b_2 = h_5$ and $b_3 = h_6$ in tree E. For an interior branch, the branch length is estimated by the difference between the heights of the higher and the lower nodes for the branch. In tree E, the length (b_5) of the branch between nodes 5 and 6 is given by $b_5 = h_6 - h_5$. This method of estimating branch lengths is essentially the same as that of UPGMA. Note that UPGMA gives LS estimates of branch lengths when the topology is correct (Chakraborty 1977).

In practice, however, the difference between the heights of the higher and lower nodes may become negative because of the sampling errors or some other disturbing factors. In this case, we assume that the interior branch length is zero and treat the branching of the clusters as a multifurcation. For example, if the estimated height of node 5 is greater than that of node 6 ($h_5 > h_6$) in tree E, we set $b_5 = 0$ and $b_1 = b_2 = b_3 = h_6$. This will generate a multifurcating node. In general, this can be done easily if we start the estimation of node heights from the root of the tree and go down to the tips. Whenever we encounter a node with a height greater than that of the previous node, we replace the height of this node by that of the previous one. This process is continued until the lengths of all interior and exterior branches are estimated. Another way of estimating h_6 is to replace it by the unweighted average height of the sequences concerned. In tree E, this approach gives $h_5 = h_6 = (d_{12} + d_{13} + d_{23})/3$. In practice, both procedures give similar results.

A linearized tree after elimination of deviant sequences can also be constructed by using the LS method with the constraint of no negative branches. This computation can be done with Felsenstein's (1988, 1997) KITCH program in PHYLIP, eliminating the outgroup sequences before the computation. In our experience, this method and the method described above gives essentially the same results.

Another method that can be used for constructing a linearized tree is the ML method. This method is quite flexible for finding an appropriate substitution model and testing the molecular clock hypothesis. Once deviant sequences are eliminated, it is also relatively easy to construct a linearized tree under the assumption of constant rate of evolution (Felsenstein 1988; Yang 1996b). In practice, however, this method gives results similar to those obtained by Takezaki et al.'s (1995) method.

Example 10.3. Linearized Trees for *Drosophila* Species

Let us illustrate the construction of a linearized tree using the *Adh* genes of the *Drosophila* species considered in Example 10.2. In this example, we showed that the six sequences marked with a * sign evolved significantly faster or slower than the average rate. We therefore eliminated these sequences and constructed a new NJ tree for the remaining species.

When we applied the branch length test to this tree, there were three sequences (*D. affinidisjuncta*, *D. differens*, *D. silvestris*) that evolved significantly faster than the average at the 1% level. However, the U statistic for this tree was 25.3 with 15 degrees of freedom and was not statistically significant. We therefore constructed a linearized tree for the 16 *Drosophila* species, which is shown in Figure 10.5B.

This linearized tree has one quadrifurcating node, which occurred because the height of the node for *D. mulleri* and *D. mojavensis* was slightly higher than that for *D. mulleri* and *D. navojoa* or for *D. arizonae* and *D. navojoa*. Calibration of the time scale for *Drosophila* species is difficult, because there is no good fossil record except for a few samples trapped in amber (Grimaldi 1987). However, considering biogeographical information, Powell (1997) suggested that the two major subgroups of *Drosophila*, subgenera *Drosophila* and *Sophophora*, diverged about 50 MY ago. If we accept this suggestion, we can construct a time scale for species divergence in *Drosophila*, as shown below the linearized tree.

This time scale indicates that *D. willistoni* and the *D. melanogaster* group diverged about 46 MY ago and the *D. obscura* and the *D. melanogaster* groups diverged about 30 MY ago. It also indicates that the Hawaiian and the *D. repleta* group species diverged about 36 MY ago. These time estimates are considerably larger than those obtained by Thomas and Hunt (1993) and Russo et al. (1995). These differences occurred because these authors calibrated the molecular clock by using information about the island formation in Hawaii and assuming that *D. picticornis* and the other Hawaiian *Drosophila* diverged about 5 MY ago. However, as argued by Beverley and Wilson (1985), the Hawaiian drosophilids possibly migrated from northeast Asia through the Koko Seamount-Midway Archipelago in the Pacific, and when they arrived in the Hawaiian islands, there might have been several different species. Beverley and Wilson's immunological study of Hawaiian and continental drosophilids support this view. Therefore, the new estimates presented in Figure 10.5B may be more reasonable than the previous ones.

In Figure 10.5B, we eliminated several species that are interesting from the evolutionary point of view. For example, *Scaptomyza albovittata* belongs to a different genus but is very closely related to *D. crassifemur*. It is therefore interesting to estimate the time of divergence between this pair of species and the Hawaiian *Drosophila*. For this reason, we constructed another linearized tree using all 22 drosophilid species (Figure 10.3C). This tree indicates that the estimates of divergence times for the major branch points of the tree are quite similar to those obtained from Figure 10.5B. For example, the divergence time between *D. willistoni* and the other *Sophophora* species is now 43 MY instead of 45 MY, and the divergence time between the Hawaiian and the *D. repleta* groups species is 39 MY instead of 36 MY. This indicates that some variation in evolutionary rate does not alter the estimates of divergence time seriously even if the rate heterogeneity is statistically significant.

Scaptomyza albovittata is morphologically quite different from *Drosophila* species, and for this reason, a different genus name was assigned. However, our tree (Figure 10.5A) clearly shows that this species is closely related to Hawaiian *Drosophila*. There are several species belonging to

this genus in Hawaiian islands, but curiously, *Scaptomyza* species are also found outside Hawaii, especially in Central America, and it is believed that *Scaptomyza* originated in Hawaii and then dispersed to other parts of the world. Tamura et al. (1995) have shown that all the five *Scaptomyza* species studied form a monophyletic group closely related to Hawaiian *Drosophilia*. The linearized tree in Figure 10.5C suggests that they diverged from Hawaiian *Drosophila* about 38 MY ago. (This could be an overestimate because the *S. albovittata* sequence apparently evolved faster than other drosophilid sequences.) This indicates that some species undergo rapid morphological evolution while other species remain without significant changes and stay in the same genus.

We also constructed linearized trees using the KITCH program mentioned earlier and the likelihood method. In these methods, we have to eliminate the outgroup sequence as well as deviant species and then construct a tree under the assumption of a constant rate of evolution. The KITCH program produced a tree that is virtually identical with that obtained by Takezaki et al.'s (1995) method. For the likelihood method, we used the Jukes-Cantor model and tested the molecular clock for the 22 drosophilid species. The *LR* value for this test was 49.1, which is highly significant. However, because it was not simple to identify deviant sequences by this method, we constructed a tree for the 22 sequences under the assumption of a constant rate of evolution. (We used PAUP* for this computation.) The tree obtained is given in Figure 10.5D. Estimates of the times of species divergence obtained by this method tend to be smaller than those obtained by Takezaki et al.'s method, but the differences are small. In the ML method, it is possible to compute the standard error of divergence time by the curvature method (Yang 1996b), but since this method tends to underestimate the standard error (Tateno et al. 1994; Gaut et al. 1996), we did not compute it. (The computer program for Takezati et al.'s method is available from *http://mep.bio.psu.edu*)

11

Ancestral Nucleotide and Amino Acid Sequences

If we can infer the amino acid sequences of ancestral proteins, it is possible to study how the function of a gene has changed in the evolutionary process. Previously, the sequences of ancestral genes and proteins were inferred by parsimony methods in the process of reconstruction of phylogenetic trees, but these sequences were rarely used for studying the evolution of genes. In recent years, it has become possible to reconstruct an ancestral protein by using site-directed mutagenesis and to study biochemical properties of the ancestral protein. This has opened a new field of evolutionary study; we can now study the functional change of proteins in the evolutionary process in the laboratory (e.g., Jermann et al. 1995; Chandrasekharan et al. 1996; Dean and Golding 1997; Yokoyama et al. 1999).

This advance in biochemical technique has renewed interest in statistical methods of inferring ancestral sequences. The parsimony method is quite effective in inferring ancestral sequences when sequence divergence is low. However, when the sequences are distantly related, it gives several possible sequences for the ancestral protein, and it is difficult to determine the most probable ancestral sequence. For this and other reasons, a number of new methods based on likelihood or distance methods have recently been developed.

In this chapter, we discuss a few statistical methods that are useful for this purpose. We will also discuss some statistical methods for studying adaptive evolution using information on ancestral sequences.

11.1. Inference of Ancestral Sequences: Parsimony Approach

We have seen that when the correct topology is known and there is only one amino acid substitution at a site, parsimony methods are able to identify the branch in which the substitution occurred. We therefore can determine the ancestral amino acids uniquely, as shown in Figure 11.1A. When there are two or more amino substitutions at a site, it is not always possible to determine all ancestral amino acids uniquely. We have seen this in Figure 7.9 for five nucleotide sequences, where there are three pos-

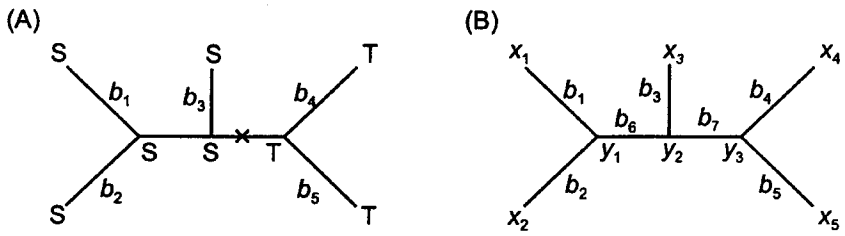


FIGURE 11.1. Trees illustrating the inference of ancestral amino acids.

sible sets of inferred ancestral nucleotides. In general, our ability of determining ancestral amino acids by parsimony depends on the number of sequences, the topology of the tree, and the configuration of amino acids among extant taxa. Parsimony gives good results when the sequences are closely related. Maddison and Maddison (1992) developed an excellent computer program for inferring ancestral amino acids using the principle of maximum parsimony. (Their program seems to work only for rooted trees.) This program includes two algorithms, Acctran and Deltran (see chapter 7). If these two algorithms give the same results, the inferred ancestral amino acids are quite accurate. When they give different results, there are several different possible ancestral sequences that we have to consider. Of course, we do not have to have a unique ancestral sequence to study the evolutionary change of protein function. If there are several possible sequences, we can produce proteins for all of them by site-directed mutagenesis and examine their function.

Nevertheless, it would be better if we could obtain one unique and most likely ancestral sequence. Maddison (1995) presented approximate methods for evaluating the relative importance of different sets of ancestral amino acids under the parsimony principle, and these methods allow us to choose the most likely ancestral sequence. In parsimony methods, however, no consideration is given to branch lengths, so that the evolutionary change of amino acids between two nodes is assumed to occur instantly. In practice, as the branch length increases, the number of substitution increases. Therefore, parsimony methods may give erroneous results when short and long branches are mixed in the phylogenetic tree.

11.2. Inference of Ancestral Sequences: Bayesian Approach

Yang et al. (1995b) developed a Bayesian approach of inferring ancestral sequences. In Yang et al.'s approach, the topology of the tree is assumed to be known, but the branch lengths are estimated by using the likelihood method. In the estimation of branch lengths, various models of substitution can be used. Once the topology and branch lengths are given, the probability of having a given set of ancestral amino acids is evaluated by the Bayesian approach. This is done for all potential sets of ancestral

amino acids, and the set with the highest probability is chosen as the most likely ancestral sequence. However, this method is quite time-consuming when the number of sequences used is large. For this reason, Zhang and Nei (1997) proposed a simplified version of the Bayesian approach, estimating branch lengths by distance methods. Koshi and Goldstein (1996) proposed another Bayesian approach, in which the topology and branch lengths are estimated by the neighbor joining method. In practice, however, the reliability of inferred ancestral sequences is low when the topology is unclear. In this chapter, we therefore consider the case where the topology is established from other information. In the following, we first discuss the distance-based method and then the likelihood-based method. We follow Yang et al.'s (1995b) mathematical approach.

Distance-Based Method

Let us consider a simple example of an unrooted tree for five sequences given in Figure 11.1B. In this tree, x_1, x_2, x_3, x_4, and x_5 represent the amino acids at a given site for sequences 1, 2, 3, 4, and 5, respectively, and y_1, y_2, and y_3 are the amino acids at the three ancestral nodes. The quantities b_1, b_2, $\cdots$, and b_7 are branch length estimates of the seven branches of the tree. In the inference of ancestral sequences, we first estimate the branch lengths by the ordinary LS method described in chapter 6. In the standard LS method, however, some branch length estimates may become negative, though the absolute values are usually very small. We therefore use the LS method with the constraint of nonnegative branches (e.g., Lawson and Hanson 1974; Felsenstein 1997).

Once all branch lengths are estimated, we compute the probability of obtaining the observed set of amino acids for each site under a given mathematical model. We denote the amino acids at the five exterior nodes of the tree in Figure 11.1A by the vector $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5)$ and the amino acids at the three interior nodes by the vector $\boldsymbol{y} = (y_1, y_2, y_3)$ (Figure 11.1B). In the example in Figure 11.1A, $x_1 = x_2 = x_3 = y_1 = y_2 = \mathrm{S}$ (= Ser) and $x_4 = x_5 = y_3 = \mathrm{T}$ (= Thr). If we use a time-reversible model of amino acid substitution (chapter 8), the evolutionary change of amino acids can be assumed to start from any node of the tree. Here we assume that y_1 is the starting amino acid.

The probability of observing a set of amino acids $\boldsymbol{x}$ is given by

$$f(\boldsymbol{x};\boldsymbol{b}) = \sum_{\boldsymbol{y}} f(\boldsymbol{y})\, f(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{b})$$

$$= \sum_{y_1}\sum_{y_2}\sum_{y_3} [g_{y_1} P_{y_1x_1}(b_1)\, P_{y_1x_2}(b_2)\, P_{y_1y_2}(b_6) \times P_{y_2x_3}(b_3)\, P_{y_2y_3}(b_7)\, P_{y_3x_4}(b_4)\, P_{y_3x_5}(b_5)] \qquad (11.1)$$

where $\boldsymbol{b} = (b_1, b_2, b_3, b_4, b_5, b_6, b_7)$ is the vector of estimated branch lengths, $P_{ij}(b_k)$ is the probability of change from amino acid i to j when the branch length b_k is given, and g_{y_1} is the relative frequency of amino acids y_1. Note that $P_{ij}(b_k)$ is equivalent to $m_{t(ij)}$ in the Dayhoff model

(Equation [2.18]) when t is replaced by b_k. In Equation (11.1), $f(\boldsymbol{y})$ is the prior probability of occurrence of $\boldsymbol{y}$ and is given by

$$f(\boldsymbol{y}) = g_{y_1} P_{y_1 y_2}(b_6) P_{y_2 y_3}(b_7) \tag{11.2}$$

The other component of Equation (11.1), $f(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{b})$, is the conditional probability of observing amino acid set $\boldsymbol{x}$ for a given set of ancestral amino acids $\boldsymbol{y}$ and is given by

$$f(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{b}) = P_{y_1 x_1}(b_1) P_{y_1 x_2}(b_2) P_{y_2 x_3}(b_3) P_{y_3 x_4}(b_4) P_{y_3 x_5}(b_5) \tag{11.3}$$

Equation (11.1) is of the same form as that of the likelihood function (Equation [8.5]) but it is not used for estimating parameters. It is just the probability of obtaining data set $\boldsymbol{x}$, when the substitution model and branch lengths are given. Here we are interested in inferring $\boldsymbol{y}$ from the observed data set $\boldsymbol{x}$. For this purpose, we use the Bayesian approach and compute the following posterior probability for each possible set of ancestral amino acids $\boldsymbol{y} = (y_1, y_2, y_3)$.

$$f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{b}) = \frac{f(\boldsymbol{y}) f(\boldsymbol{x}|\boldsymbol{y};\boldsymbol{b})}{f(\boldsymbol{x};\boldsymbol{b})} \tag{11.4}$$

Theoretically, y_1, y_2, and y_3 can be any of the 20 different amino acids. In practice, however, the amino acids that are not observed at the amino acid site under consideration are unlikely to have ever appeared at the site. In fact, if we consider such amino acids, $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{b})$ is usually vanishingly small. Therefore, we exclude them from consideration to speed up the computation.

If we compute $f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{b})$ for all possible or probable combinations of y_1, y_2, and y_3, we will know the set of amino acids that has the highest posterior probability. We can then infer that this set ($\hat{\boldsymbol{y}}$) is the ancestral amino acids. If this is done for all sites of amino acid sequences, we can determine the ancestral sequence at each interior node.

Accuracy of Inferred Amino Acids

It is often useful to know the overall accuracy of inferred ancestral amino acids ($\hat{\boldsymbol{y}}$) for all sites. This accuracy [P(A)] may be measured by the average of Equation (11.4) over all sites.

$$P(A) = \sum_{\boldsymbol{x}} f(\boldsymbol{x}) f(\hat{\boldsymbol{y}}|\boldsymbol{x};\boldsymbol{b}) \tag{11.5}$$

where $f(\hat{\boldsymbol{y}}|\boldsymbol{x};\boldsymbol{b})$ is the posterior probability of the set of inferred ancestral amino acids at a given site and $f(\boldsymbol{x})$ is the relative frequency of set $\boldsymbol{x}$ over all sites (Yang et al. 1995b). Note that Equation (11.4) gives the same set of ancestral amino acids when $\boldsymbol{x}$ is the same.

In the above formulation, we considered the posterior probability of a set of ancestral amino acids. However, it is also possible to compute the posterior probability of a given amino acid (say, amino acid a) assigned

at a given node (say, the i-th node) at a site. This probability is given by the sum of probabilities of all the ancestral amino acid sets ($\mathbf{y}$'s) that have amino acid a at the i-th node. That is,

$$f(y_i = a|\mathbf{x}) = \frac{\sum_{\mathbf{y};y_i=a} f(\mathbf{y})f(\mathbf{x}|\mathbf{y};\mathbf{b})}{f(\mathbf{x};\mathbf{b})} \qquad (11.6)$$

The best assignment at a node will be the amino acid that has the highest probability. It is possible that the best amino acid assignment at a node obtained by the above equation is different from the amino acid assigned for the node in the best set of ancestral amino acids inferred by Equation (11.4). However, such cases are rare, and in this book, Equation (11.4) is used to infer the set of ancestral amino acids, and Equation (11.6) is used to evaluate the accuracy of inferred amino acids.

Models of Amino Acid (Nucleotide) Substitution and Evolutionary Distance

In the above computation, we have to use a certain mathematical model of amino acid substitution so that we can compute $P_{ij}(b_k)$. A model that is often used is Dayhoff et al.'s (1978) or Jones et al.'s (1992) empirical substitution matrix (Dayhoff or JTT model; see section 8.3). In these models, pairwise distances can be computed by using gamma distances (chapter 2). Another model one can use is the simple Poisson model. Our computer simulation and empirical data analysis have shown that the accuracy of the inferred amino acids is not very sensitive to the differences among these models (Yang et al. 1995b; Zhang and Nei 1997). It seems that the JTT model gives sufficiently accurate results for most purposes. However, when the protein under investigation has been subjected to positive selection, these models may not be appropriate, because they are based on amino acid substitution data for conserved proteins. In this case, a more appropriate model would be the Poisson model, which is independent of empirical data. In actual data analysis, the Dayhoff-*f*, the JTT-*f*, and the Poisson-*f* models, which have been discussed in chapter 8, are often used, because the amino acid frequencies vary from protein to protein. (See *http://mep.bio.psu.edu*)

Likelihood-Based Method

In the distance-based Bayesian approach mentioned above, the branch lengths are first estimated from pairwise distances. For computing pairwise distances, one can ideally choose the most appropriate measure for the data set under investigation. This can be done by finding an appropriate model fitting a data set and then computing distances based on the model. In the case of DNA sequence, the fit of a model to a data set can be tested by Rzhetsky and Nei's (1995) methods.

In the likelihood-based method, however, these two processes can be accomplished at the same time by using the likelihood ratio test. In this method, we can set up the following likelihood function

$$L = \sum_{\boldsymbol{y}} f(\boldsymbol{y}) f(\boldsymbol{x}|\boldsymbol{y};\theta) \tag{11.7}$$

which is the same form as Equation (11.1) except that a set of constant $\boldsymbol{b}$ (branch lengths) in the equation is now replaced by the parameter set θ, which is to be estimated from the data. This parameter set may include the parameters for a substitution model as well as for branch lengths. Theoretically, all the parameters in Equation (11.7) can be estimated by maximizing L for the data set used (Schluter 1995; Yang et al. 1995b), but the substitution parameters estimated from a data set are not very reliable because the data set is usually small.

For this reason, it is customary to use substitution parameters that have been estimated from other data sets. Particularly, in the case of amino acid sequence data, the JTT model or its modifications are generally used (Yang et al. 1995b; Koshi and Goldstein 1996). Therefore, we do not have to estimate substitution parameters, and the only parameters that are estimated by the ML method are branch lengths. Once branch lengths (or all necessary parameters) are estimated, the set of ancestral amino acids $\boldsymbol{y}$ is inferred by the same procedure as that mentioned in the section of Distance-Based Method and the posterior probabilities are computed by Equations (11.4)–(11.6).

In this chapter, we have assumed that the topology of the tree is known, so that we do not have to search for the ML topology. However, even if the topology is given, it is still time-consuming to estimate branch lengths by the ML method. Fortunately, branch length estimates obtained by the LS methods are as accurate as those obtained by the ML method (Zhang and Nei 1997). Since the estimation of branch lengths by the LS method is much faster than that by the ML method, it is easier to use the former method when the number of sequences used is large.

Example 11.1. Evolution of Color Vision in Mammals

The color vision of mammals is controlled by photopigments, each of which consists of a protein called opsin and a chromophore 11-*cis* retinal. The diversity of color vision in mammals is due to variation in the number and sequence of opsin genes. Most mammals have a gene for short wavelength (blue)-sensitive opsins and a gene for middle wavelength (green) to long wavelength (red)-sensitive opsins, so that they have dichromatic color vision. In higher primates, however, trichromatic color vision has evolved in two different ways. Hominoids and Old World (OW) monkeys have three different color vision genes; one autosomal gene encoding the blue-sensitive opsin and two X-linked genes encoding the green- and red-sensitive opsins (Nathans et al. 1986). Therefore, they can see a full range of color. Most New World (NW) monkeys have only one X-linked locus in addition to the autosomal blue gene locus, but the X-linked locus is polymorphic and has three different alleles encoding green- and red-sensitive opsins (Mollon et al. 1984; Neitz et al. 1991). Therefore, all males are dichromatic, whereas females are either dichromatic or trichromatic depending on the genotype. The blue and red/green genes seem to have diverged more than 500 million years

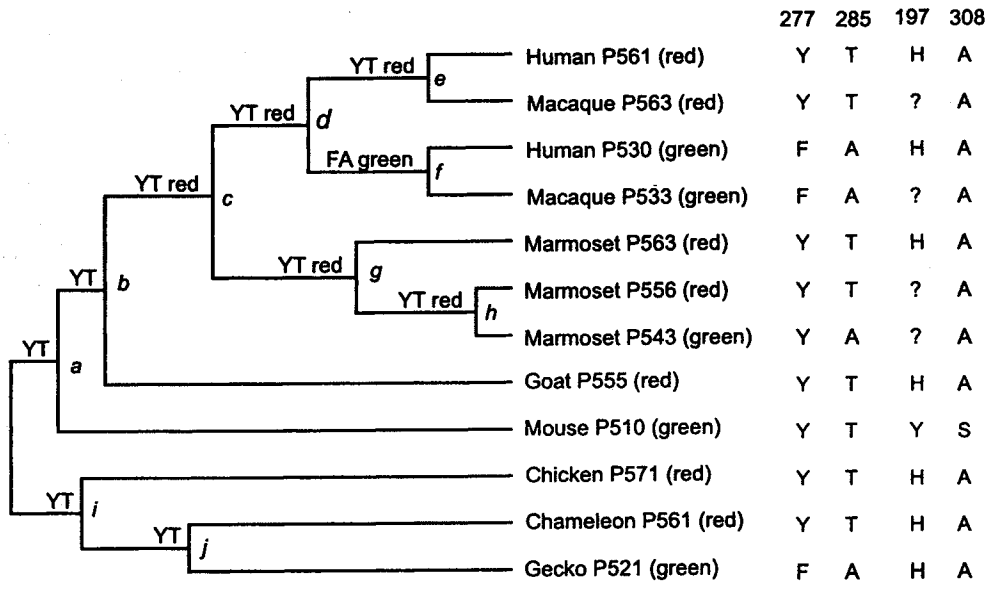


FIGURE 11.2. Phylogenetic trees used and inferred amino acids of ancestral organisms (tree nodes) of higher primates at the four critical amino acid sites for red/green opsins. The number after P in the gene symbol refers to the maximum absorption wavelength (λ_{max}). A = alanine, S = serine, F = phenylalanine, Y = tyrosine, T = threonine, and H = histidine.

(MY) ago, but the green and red genes apparently diverged only after OW monkeys separated from NW monkeys (Nathans et al. 1986; Yokoyama and Yokoyama 1989). Trichromacy in higher primates is believed to have evolved to facilitate the detection of yellow and red fruits against dappled foliage (Mollon 1991). However, the process of evolutionary change from dichromacy to trichromacy is not well understood (Bowmaker 1991; Mollon 1991). To study this problem, Nei et al. (1997b) inferred the amino acid sequences to ancestral organisms and clarified the evolutionary pathways of color vision in higher primates and some other mammals (Figure 11.2).

The spectral sensitivity of opsins is measured by the maximum absorption wavelength (λ_{max}), which is 561 nanometers (nm) and 530 nm for the human red and green opsins, respectively. The genes encoding these opsins are designated as P561 and P530 in Figure 11.2. It is now well established that this difference in λ_{max} between the red and green opsins is primarily controlled by two critical amino acids at sites 277 and 285 in higher primates (Neitz et al. 1991). In red opsins, the positions 277 and 285 have amino acids tyrosine (Y) and threonine (T), respectively, whereas in green opsins they have phenylalanine (F) and alanine (A). This appears to be the case for most primates and a majority of mammalian species. Therefore, if we determine the amino acids of ancestral proteins at these positions, we can infer the type of color vision the ancestral organisms had.

Figure 11.2 shows the phylogenetic tree for the 11 sequences used in

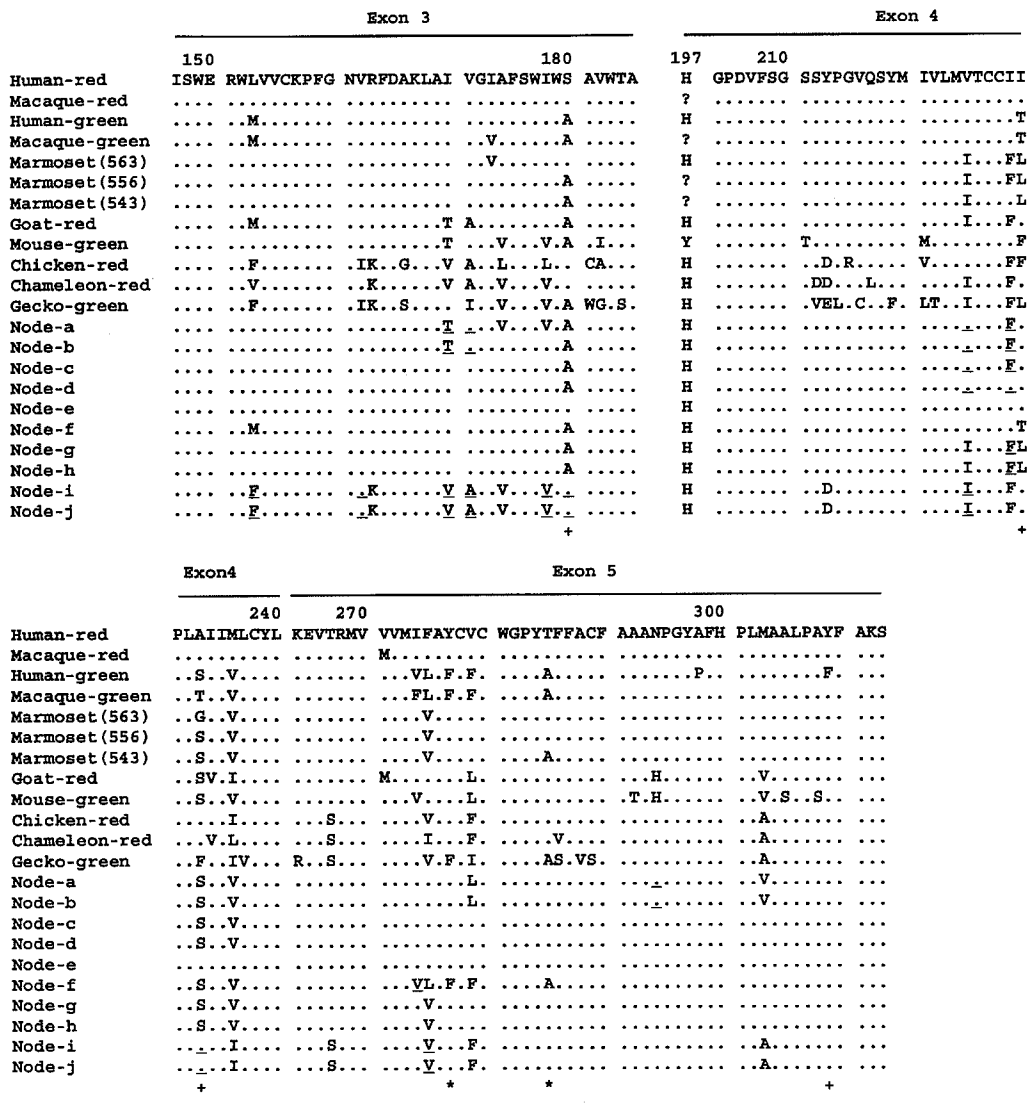


FIGURE 11.3. Partial amino acid sequences of red and green opsins of the present-day and ancestral organisms. The ancestral amino acids that have a posterior probability of 90% or less are underlined. The symbol "?" indicates that the amino acid is unknown. The "*" signs indicate the two critical sites for color vision and the "+" sign shows the four less-important sites in higher primates. Amino acids at sites 197 and 308 are known to affect mouse and rat color vision.

Nei et al.'s (1997b) study plus the mouse sequence. The complete amino acid sequence is not available for all the genes, so we used the partial sequences for exons 3, 4, and 5, as shown in Figure 11.3. This figure also includes the ancestral sequences inferred for all ancestral nodes. These ancestral amino acids were inferred by the distance-based Bayesian method, but the same amino acids were obtained by the likelihood-based method. We used the JTT model as the substitution matrix. The posterior

probabilities of these amino acids were very close to 1 in most cases. Only the amino acids underlined had a probability lower than 0.9. Parsimony analysis generated two or more possible amino acids at ten sites. For example, at position 171, the amino acids inferred for nodes *a* and *b* were either V (valine) or A. According to the Bayesian approach, however, the posterior probabilities of V and A were 0.731 and 0.269, respectively. Therefore, V instead of A is used as an inferred amino acid in Figure 11.3. In general, when parsimony generates two or more alternative amino acids, the Bayesian probability of any inferred amino acid is considerably lower than one. However, even if parsimony generates a single parsimonious amino acid, the Bayesian probability can be relatively low when there are several amino acid changes observed at a site.

The critical amino acids at positions 277 and 285 for the ancestral as well as the present-day organisms are shown in Figure 11.2. As expected, the amino acids at nodes *e* and *f* are YT (red) and FA (green), respectively. However, all other ancestral nodes have YT. This suggests that the common ancestor (node c) of higher primates had one X chromosome gene and it was a red gene. In other words, the green gene in higher primates seems to have been derived from the red gene after gene duplication. Interestingly, even the current human populations include atavistic mutants who lack the green gene. The famous British chemist John Dalton, who discovered color blindness, was one such individual. Hunt et al. (1995) found this when they examined the DNA sequences extracted from Dalton's eyeballs, which had been kept in a museum for nearly 150 years.

Among the present-day sequences presented in Figure 11.2, the mouse gene P510 is unusual. This gene has a λ_{max} of only 510 nm, well below the value for a green opsin, though it clearly belongs to the red/green opsin gene family; the amino acids at position 277 and 285 are YT (Sun et al. 1997). This puzzling observation was explained by Sun et al., who showed that the low λ_{max} value of the mouse green gene is due to replacement of the amino acids H (histidine) and A at positions 197 and 308 in humans by Y and S (serine), respectively. These two amino acids alone are known to reduce λ_{max} by about 44 nm. The amino acids at these positions in other sequences are apparently H and A, though in some organisms the amino acid at position 197 has not been studied (Figure 11.2).

The low λ_{max} value of the mouse green gene makes mice almost colorblind, but this does not matter for them, because they are nocturnal. In fact, another nocturnal species, *Rattus norvegicus,* also has amino acids Y and S at these positions and a λ_{max} of about 510. In these organisms, it seems sufficient for the pigment genes to distinguish day and night.

In most diurnal mammals, however, the red/green opsin has a λ_{max} of about 555 nm (Jacobs 1993). Therefore, it is likely that these organisms do not have Y and S at positions 197 and 308. These observations enhance the reliability of the inferred amino acids YT at positions 277 and 285 for the ancestor of mammals. However, since the ancestors of current mammalian species were apparently nocturnal, it is unclear whether or not they had dichromatic color vision.

Some Remarks

We have indicated that the Bayesian probability approach has an advantage over parsimony in that the relative probabilities of different amino acids at an ancestral node can be computed. It should be noted, however, that the posterior probability assigned for each amino acid in the Bayesian approach is dependent on various assumptions that may not hold in reality. This is particularly so when one is interested in studying adaptive evolution caused by a few amino acid changes, as in the case of the color vision genes. The reason is that the pattern of amino acid substitution due to positive selection is often unique and could be different from that of nonadaptive amino acid substitution for which the Dayhoff or the JTT model was generated. Therefore, the posterior probabilities computed by the distance or likelihood method may not be realistic. In this case, more reliable ancestral amino acids might be found by the model-free parsimony method.

A deficiency of the parsimony method is that two or more alternative amino acid sequences may be inferred at a given node. In the study of adaptive evolution, however, this is not a serious problem, because one can consider all alternative sequences as possible ancestral sequences. In fact, Jermann et al. (1995) constructed all possible ancestral ribonucleases by site-directed mutagenesis and investigated a likely course of evolutionary change of a ribonuclease concerned with the two-gut digestion in ruminants.

For the above reason, it is recommended that both the parsimony and the Bayesian approaches be used in a study of adaptive amino acid substitution.

11.3. Synonymous and Nonsynonymous Substitutions in Ancestral Branches

In chapter 4, we discussed estimation of the numbers of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions per site between two DNA sequences. In the study of adaptive evolution, it is often necessary to estimate the numbers of synonymous (b_S) and nonsynonymous (b_N) substitutions *per site per branch* (e.g., Yu and Irwin 1996). For a pair of sequences 1 and 2, we have the relationships $d_S = b_{S1} + b_{S2}$ and $d_N = b_{N1} + b_{N2}$, where b_{S1} (b_{N1}) and b_{S2} (b_{N2}) are the numbers of synonymous (nonsynonymous) substitutions for sequences 1 and 2 after their divergence, respectively.

Figure 11.4 shows the phylogenetic tree of the ribonuclease duplicate genes (ECP and EDN) for the higher primate species (see Example 11.2 for detailed explanation). The numbers of synonymous (a_S) and nonsynonymous (a_N) substitutions *per sequence per branch* are presented for each branch of the tree. In the present case, the number of potential synonymous sites (S) and the number of potential nonsynonymous sites (N) are virtually the same for all sequence comparisons and are 124 and 347, respectively, when the transition/transversion bias is taken into account (Zhang et al. 1998). Therefore, the b_S and b_N values for each branch

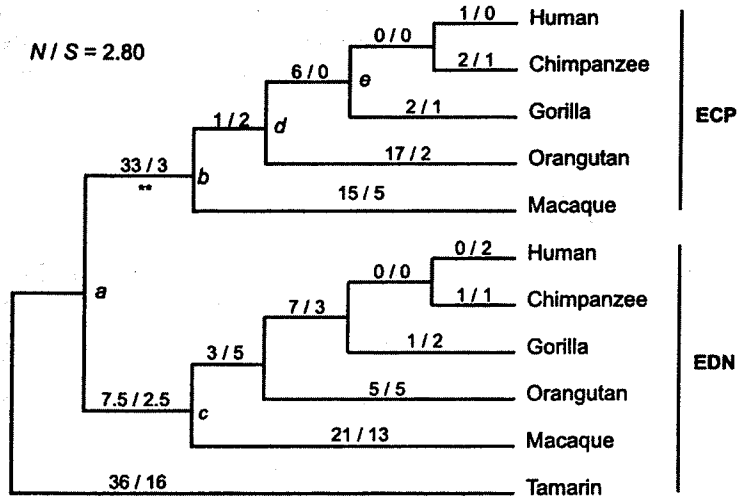


FIGURE 11.4. Numbers of nonsynonymous (a_N) and synonymous (a_S) substitutions per sequence for each branch of the tree for five hominoid and Old World monkey species. The a_N and a_S values are given before and after the "/" sign. N and S are the numbers of potential synonymous and potential nonsynonymous sites, respectively. ** indicates a $P < 0.01$ in the Z test. ECP: Eosinophil cationic protein. EDN: Eosinophil-derived neurotoxin. Computed by the programs available at http://mep.bio.psu.edu.

are estimated by a_S/S and a_N/N, respectively. For example, the estimates ($\hat{b}_S$ and $\hat{b}_N$) of b_S and b_N for the branch between nodes a and b are $\hat{b}_S = 3/124 = 0.024$ and $\hat{b}_N = 33/347 = 0.095$, respectively. So, if we take the face values of $\hat{b}_S$ and $\hat{b}_N$, it appears that nonsynonymous substitutions were accelerated by positive Darwinian selection in this branch. How can we estimate a_S and a_N or b_S and b_N from sequence data and how can we test the statistical significance of the differences between $\hat{b}_S$ and $\hat{b}_N$? These are the problems to be considered in this section.

Estimation of b_S and b_N from Pairwise Distance Data

A simple method of estimating b_S and b_N for each branch is first to estimate d_S and d_N (or p_S and p_N) for all pairs of sequences using Equations (3.8) and (4.8) and then to compute $\hat{b}_S$'s from $\hat{d}_S$'s and $\hat{b}_N$'s from $\hat{d}_N$'s by the LS method (Equation [6.9]) for each branch of a given topology. (Note that positive selection is usually detected only when d_S and d_N are relatively small, and in this case d_S and d_N are approximately equal to p_S and p_N, respectively.) This method gives quite reliable estimates of b_S and b_N when S, N, a_S's and a_N's are large. In most cases, S and N are virtually the same for all sequence comparisons, so that a_S and a_N are given by $a_S = \hat{b}_S S$ and $a_N = \hat{b}_N N$, respectively. To detect positive Darwinian selection for each branch, one may use the Z test for the difference $D = \hat{b}_N - \hat{b}_S$. The variances of $\hat{b}_N$ and $\hat{b}_S$ can be computed by the method described in chapter 6 or by the bootstrap.

However, the accuracies of $\hat{b}_S$ and $\hat{b}_N$ decline as a_S's and a_N's decrease, and the Z test of $\hat{b}_N - \hat{b}_S$ becomes unreliable. In this case, it is better to estimate the ancestral nucleotide sequences for all tree nodes and then compute a_S, a_N, $\hat{b}_S$, and $\hat{b}_S$.

Estimating b_S and b_N from Information on Ancestral Nucleotide Sequences

There are two ways of inferring ancestral nucleotide sequences. The first is to consider each nucleotide site separately and infer the set of ancestral nucleotides for all tree nodes at the site. In this method, one can use either the parsimony or the Bayesian method. The second method is to infer the ancestral amino acids for each tree nodes and then infer the ancestral nucleotides under the constraint of the amino acids inferred (Zhang et al. 1998). (See http://mep.bio.psu.edu.)

The above two methods give similar results when the DNA sequences used are closely related. When sequence divergence is high, however, the first method may generate stop codons in some of the ancestral sequences. Also, since the first, second, and third nucleotide positions of codons often have different substitution patterns, one may have to use different models for them. If we use the second method, there is no danger of having stop codons in the ancestral sequences, and there will be no problem in computing $\hat{b}_S$ and $\hat{b}_N$. These values can be computed between any pair of consecutive branch nodes (ancestors). However, we may still have to use different substitution models for different codon sites once the ancestral amino acids are determined.

With our limited experience, the second method generally gives more reasonable results, although it requires a greater computational time than the first. However, computational time is not a serious problem if we use the distance-based Bayesian approach. Computer simulations have shown that parsimony methods do not give satisfactory results when sequence divergence is high (J. Zhang, unpublished).

Estimation by Counting Methods

When the extent of sequence divergence is low and there are few multiple substitutions (homoplasy) at each codon site, it is possible to determine ancestral DNA sequences (at each node) with a high probability and then identify each synonymous or nonsynonymous substitution for each branch (between sequences at two consecutive nodes) (see chapter 4). We can then count the numbers of synonymous (a_S) and nonsynonymous (a_N) substitutions per sequence. In this case, all sequences have nearly the same numbers of synonymous sites (S) and nonsynonymous sites (N). Therefore, we have $\hat{b}_S = a_S/S$ and $\hat{b}_N = a_N/N$ for every branch. Conceptually, $\hat{b}_S$ and $\hat{b}_N$ correspond to $\hat{p}_S$ and $\hat{p}_N$ in Equation (4.2), but when they are small (say, < 0.10), they are also very close to the actual numbers of substitutions per site. The sums of $\hat{b}_S$ and $\hat{b}_N$ for all the branches connecting two present-day sequences generally exceed $\hat{p}_S$ and $\hat{p}_N$ for the same pair of sequences, respectively.

Example 11.2. Adaptive Evolution of the ECP Gene after Gene Duplication

Eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) in hominoids and OW monkeys belong to the ribonuclease (RNase) superfamily, and they are present in the large specific granules of eosinophilic leukocytes (Rosenberg et al. 1995; Snyder and Gleich 1997). EDN has high catalytic activity as an RNase, but its real physiological function is not well understood. By contrast, ECP has very low RNase activity but is a potent toxin to various pathogenic bacteria and parasites, and its toxicity is unrelated to RNase activity (Rosenberg and Dyer 1995). NW monkeys have only one gene, which encodes a protein whose physiochemical properties are similar to those of hominoid EDN.

Figure 11.4 shows the phylogenetic tree for the ECP and EDN genes from higher primates (Zhang et al. 1998). The ECP gene apparently evolved from the EDN gene through gene duplication that occurred after hominoids and OW monkeys diverged from NW monkeys (e.g., tamarin). We first computed $\hat{d}_N$ and $\hat{d}_S$ for all pairs of sequences using the modified Nei-Gojobori method with $R = 1$ and estimated b_N and b_S for all branches by the LS method mentioned earlier. For a few branches, $\hat{b}_N$ and $\hat{b}_S$ become slightly negative apparently because of sampling errors. In this case, we set them equal to 0. The $\hat{b}_N$ and $\hat{b}_S$ values obtained in this way are given above each branch line of the tree in Figure 11.5. For example, the $\hat{b}_N$ and $\hat{b}_S$ for the branch $a - b$ are 0.086 and 0.021, respectively. Application of the Z test for the difference $D = \hat{b}_N - \hat{b}_S$ for this branch shows that $Z = 3.1$ and is significantly different from 0 at the 0.1% level. This result suggests that positive Darwinian selection operated when the ECP gene was formed from a duplicate copy of the EDN gene. The branch $d - e$ also shows that D is positive and significant ($Z = 2.8$). In this branch, however, $\hat{b}_S = (0.0035)$ was so small that the applicability of the large-sample Z test is questionable.

In fact, the extent of sequence divergence of both the ECP and EDN genes is so small that it is desirable to reexamine the above problem by the ancestral sequence method. We used the Poisson model of amino acid substitution to infer the ancestral amino acid sequences and then inferred the nucleotide sequences under the restriction of inferred ancestral amino acids. Noting that only a single nucleotide substitution occurred at most variable sites of this data set, we used the counting method to estimate a_S's and a_N's. There were four sites in which two nucleotide substitutions apparently occurred. In this case, we used the modified Nei-Gojobori method to count the number of synonymous (a_S's) and nonsynonymous (a_N) substitutions. The results are presented in Figure 11.4. As mentioned earlier, S and N for this data set are approximately 124 and 347, respectively, for all sequence comparisons. Therefore, b_S and b_N are estimated approximately by a_S/S and a_N/N, respectively. The results are presented in Figure 11.5. The $\hat{b}_S$ and $\hat{b}_N$ values are generally close to those obtained from d_S's and d_N's. However, the $\hat{b}_N$ (= 0.0177) for the branch $d - e$ is about half the $\hat{b}_N$ obtained from the $\hat{d}_N$ values, whereas $\hat{b}_S$ is 0 instead of 0.004.

In the present case, the actual numbers of substitutions (a_S and a_N) are so small that the Z test is not appropriate. We therefore used Fisher's ex-

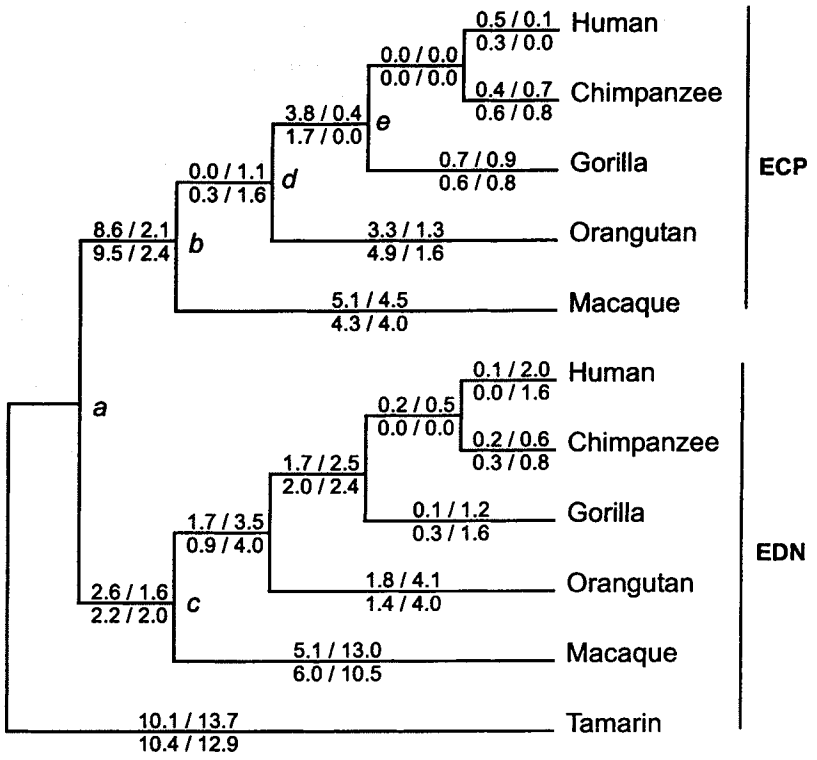


FIGURE 11.5. The b_N and b_S values (×100) for each branch of the phylogenetic trees for the ECP and EDN genes. The numbers before and after the "/" sign refer to b_N and b_S, respectively. The b_N and b_S values obtained by the LS method are given above the branch line, whereas those obtained by the counting method are given below the line.

act test for the a_S and a_N values in Figure 11.4. The results of this test are presented in Table 11.1. In branch $d - e$, six of the 347 nonsynonymous sites experienced substitutions, but none of the 124 synonymous sites had substitutions. Fisher's exact test gives a P value of 0.158, indicating that the difference $\hat{b}_N - \hat{b}_S$ in this branch can be explained by chance alone. This result is different from that of the Z test mentioned earlier, but it is more reliable. However, the application of the same test for a_N and a_S in the branch a-b shows that the difference $\hat{b}_N - \hat{b}_S$ is significant at the 1% level. Therefore, we may conclude that the rate of nonsynonymous substitution was enhanced by positive selection after the for-

Table 11.1 Fisher's exact test of positive selection.

	Branch *d*-*e*				Branch *a*-*b*			
	Nonsyn.	Syn.	Total	*P*	Nonsyn.	Syn.	Total	*P*
Changed sites	6	0	6		33	3	36	
Unchanged sites	341	124	432		314	121	435	
Total	347	124	471	0.158	347	124	471	0.006

```
                                                         40                                          80
Human-ECP     .......... ........M. ......AR.. ...R....A. ...SLN.PR. .I...A.... RW........ .R........
Chimp-ECP     .......... ........M. ......AR.. ...R....A. ...SLN.PR. .I........ RW........ .R........
Gorilla-ECP   .......... ........M. ......AR.. ...R....A. ...SLN.PR. .I........ RW........ .R........
Oran-ECP      .......... ........S. .G....A..R ...R....A. ..VSLN.P.. .T........ .....D.... .R........
Macaque-ECP   .......... ........M. ......AR.. ...K....A. ....VN.PR. .I........ .......... .R....YTA.
Node b        .......... ........M. ......AR.. ...R....A. ....LN.PR. .I........ .......... .R........
Node a        MVPKLFTSQI CLLLLLGLLG VEGSLHVKPP QFTWAQWFEI QHINMTPQQC TNAMRVINNY QRRCKNQNTF LLTTFANVVN
Node c        .......... .......... .......... .......... ......S... ....Q..... .......... ..........
Human-EDN     .......... .........A .......... .........T ......S... ....Q..... .......... ..........
Chimp-EDN     .......... .........A .......... .........T ......S... ....Q..... .......... ..........
Gorilla-EDN   .......... .........A .......... .........T ......S... .......... .......... ..........
Oran-EDN      .......... S........A .D........ .........T ......S... N...Q....F .......... .R........
Macaque-EDN   .......... ........M. ......A..G .......... ......SG.. ....Q..... .......... ......D..H
Tamarin-EDN   .......... .V...F...S ..V..Q...Q ..S.....S. ...QT..LH. .S...A..R. .P........ .H........

                                                         120                                      157
Human-ECP     ....QS.R.. H..T.....R .RFR...L.. D.I.A..... ...DR.GRR. .......... ..S.R..... .......
Chimp-ECP     ....QS.R.. H..T.....Q .RFR...L.. D.I.A..... G..DR.GRR. .......... ..S.R..... .......
Gorilla-ECP   ....QS.R.L H..T.....R .RFR...L.. D.I.A..... ...DR.GRR. .......... Q.S.R..... .......
Oran-ECP      .......... ...T.H...R .RF....L.. ....A..... K..DRTERR. .......... ..S.R..... .......
Macaque-ECP   ..R.ER.R.. ...T.H...R .RYR...L.. D.I.A...T. ...DR.GRR. .....ES... ..S.R..... .......
Node b        .......R.. ...T.H...R .RFR...L.. D.I.A..... ...DR.GRR. .......... ..S.R..... .......
Node a        VCGNPNITCP RNRSLNNCHH SGVQVPLIHC NLTGPQISNC RYAQTPANMF YVVACDNRDP RDPPQYPVVP VHLDTTI
Node c        .......... S......... .......... ...S...... .......... .I........ .......... ....RI.
Human-EDN     ......M... S.KTRK.... ..S....... ...S...... .......... .I.......Q .......... ....RI.
Chimp-EDN     ......M... S.KTRK...Q ..S....... ...S...... .......... .I.......Q .......... ....RI.
Gorilla-EDN   ......M... S.KTRK.... ..S....... ...S...... .......... .I.......Q .......... ....RI.
Oran-EDN      .......... S...R..... .......... ...S...... .......... .I........ .......... ....RI.
Macaque-EDN   .....SMP.. S.T....... .......... ...SRR.... ..T..T..KY .I...N.S.. .......... ....RI.
Tamarin-EDN   ....T..... ..A....... .......TY. .......... V.SS.Q.... .......... .......... .......
```

FIGURE 11.6. Amino acid sequences of present-day and ancestral ECP and EDN proteins. Dots represent the same amino acids as those of the ancestral sequence at node *a*. The arginine (R) residues of ECP and EDN are shown in boldface when they are not identical with those of the ancestral protein at node *a*.

mation of the ECP gene and before the separation of hominoids and OW monkeys.

It is of interest to identify the amino acid substitutions that produced the novel antipathogen function of ECP. Figure 11.6 shows that amino acid sequences of ECP and EDN from the present-day and ancestral organisms. It is clear that in the ECP gene lineage many substitutions to arginine (R) have occurred. Since arginine is the most basic amino acid, ECP is now highly basic in hominoids and OW monkeys. The isolectric point (pI) of ECP ranges from 9.9 to 10.3, whereas the pI of EDN is much lower and ranges from 8.4 to 8.8. Interestingly, the pI value estimated from the amino acid sequences for node *b* is 10.6, whereas the value for node *a* is 8.8. These results suggest that the arginine changes generated antipathogen toxicity in ECP. In fact, there is experimental evidence that ECP kills bacteria and parasites by making holes in their cell membranes through the contact of positively charged arginines to the membranes (Young et al. 1986; Rosenberg et al. 1989).

11.4. Convergent and Parallel Evolution

Convergent Evolution and Adaptation

In morphological evolution, distantly related organisms sometimes develop similar morphological characters due to adaptation to similar en-

vironmental conditions. This is called **convergent evolution**. Good examples are the body shapes of sharks (fish) and dolphins (mammals) and the wings of birds and bats. The similarity of morphological characters caused by convergent evolution is usually superficial, and an anatomical study easily reveals different origins of the characters.

Convergent evolution also occurs quite often at the level of protein function and structure. For example, hemoglobin in vertebrates and hemocyanin in lower sea animals have a similar function, that is, oxygen transportation, but their amino acid sequences are very different (Stewart et al. 1987). Similarly, there are at least three different types of serine proteases: subtilisin, trypsin, and α-β type enzymes. These proteases have similar active sites, but there is no sequence similarity among them (Doolittle 1994). Convergent evolution at the structural level also occurs very often. Thus, similar structural motifs consisting of α helices and β sheets are observed in many different proteins, but the amino acid sequences in different proteins are again very different. Of course, it is not always clear whether this type of structural similarity occurred by convergent evolution or represents a similarity by descent from a common ancestral protein. If the latter is the case, it must have occurred a long time ago.

By contrast, convergent evolution at the amino acid level is rare, and amino acid sequences usually diverge as time goes on (Doolittle and Blombäck 1964). However, a number of authors (e.g., Stewart et al. 1987; Jollès et al. 1989; Yokoyama and Yokoyama 1990; Swanson et al. 1991; Kornegay et al. 1994) reported cases where convergent evolution seems to have occurred. For example, constructing a phylogenetic tree for the amino acid sequences of cow, horse, rat, human, baboon, and langur lysozymes, Stewart et al. (1987) noticed that the tree obtained is different from the evolutionary relationships of the six organisms and that the lysozymes from cow (a ruminant with two guts) and langur (a colobine monkey with two guts) are closely related (Figure 11.7).

In higher vertebrates lysozyme is usually expressed in macrophages, tears, saliva, mammalian milk, and avian egg white as a host defense protein to fight invading bacteria. In two-gut fermenting animals such as cow and langur, however, lysozyme is recruited in stomachs for digesting the bacteria passing through the stomachs to extract the nutrients assimi-

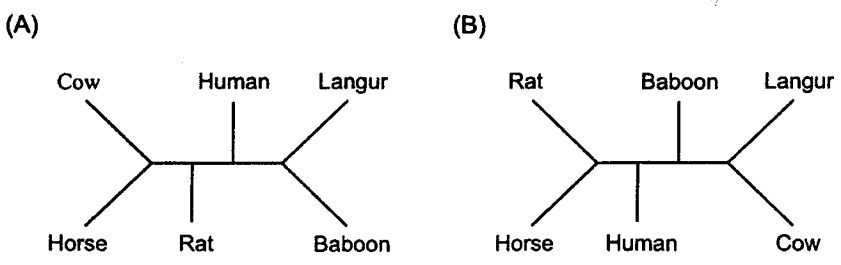


FIGURE 11.7. (A) Assumed biological tree for six species and (B) the inferred maximum parsimony tree from their lysozyme sequences. Adapted from Stewart et al. (1987).

lated by the bacteria. Therefore, Stewart et al. (1987) conjectured that the close relationship of cow and langur lysozymes in Figure 11.7B is caused by convergent evolution of amino acid sequences in these two organisms. They also noticed that the cow and langur sequences share the same amino acid at a number of sites, which are not present in other sequences (uniquely shared sites).

However, later studies of lysozyme sequences from various organisms have shown that the evolution of lysozyme genes is quite complicated. First, the cow and rat genomes were shown to have multiple copies of lysozyme genes (Irwin et al. 1993). Another distantly related two-gut fermenting organism, the hoatzin (an avian species), was also shown to have nine copies of lysozyme genes (Kornegay 1996). Second, when a phylogenetic tree for lysozyme sequences was constructed with more sequences, the tree became more or less the same as the biological tree, although it was not always easy to find orthologous sequences from different groups of organisms (Doolittle 1994; Adachi and Hasegawa 1996b). Third, as the number of sequences increased, the number of uniquely shared amino acid sites declined as expected. However, these observations do not necessarily reject the possibility of convergent evolution at the amino acid sequence level. What is necessary is to evaluate the probability of occurrence of convergent changes by chance and test whether the observed convergent changes can be explained by chance alone. This is not an easy problem, but an approximate method of testing convergent and parallel evolution was developed by Zhang and Kumar (1997).

Convergent and Parallel Amino Acid Changes

In the literature, convergent and parallel amino acid changes are not always distinguished, and whenever these changes are observed, positive Darwinian selection is usually invoked. However, it is important to distinguish between the two types of changes, because the probabilities of their occurrence by chance are quite different. **Convergent changes** of amino acids refer to the event in which two different amino acids change to the same amino acid in two different lineages. Thus, the amino acid changes A → S in lineage 1 (node 6 to node 1) and T → S in lineage 4 (node 8 to node 4) in Figure 11.8A represent an event of convergent changes. By contrast, **parallel changes** of amino acids refer to the case where the same amino acid change occurs in two different lineages at a site. Suppose that node 6 in Figure 11.8A has amino acid T instead of A. Then, the same amino acid change T → S will occur in lineages 1 and 4, and thus these changes are parallel changes. Both convergent and parallel changes may occur by chance, but the probability of occurrence of convergent changes by chance is generally much lower than that of parallel changes.

If the substitution matrix of amino acid changes is known and the phylogenetic tree of a set of amino acid sequences is given, it is possible to infer the amino acids at each ancestral node for each site and compute the probability of occurrence of a given set of convergent or parallel amino acid changes by chance. For example, the probability of occur-

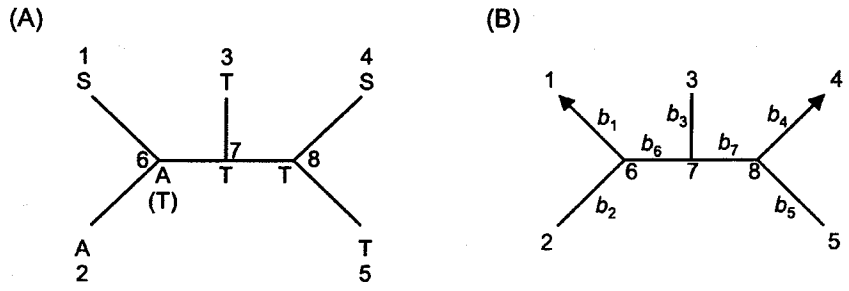


FIGURE 11.8. (A) Convergent changes: A→S in lineage 1 (node 6 to node 1) and T→S in lineage 4 (node 8 to node 4). Parallel changes: T→S in lineage 1 and T→S in lineage 4. (B) Tree for explaining convergent and parallel evolution.

rence of convergent changes A → S in lineage 1 and T → S in lineage 4 in Figure 11.8A can be computed if we know the branch lengths of the tree and use the Dayhoff or the JTT model of amino acid substitution. Therefore, if the same set of convergent changes is observed at s sites for a set of sequence data, we can compute the probability that this evolutionary event occurs s times or more. If this probability is lower than 0.05, we may conclude that the observed number of convergent-change events is too high to be attributable to chance events and that this excess of convergent-change events is probably due to positive selection.

Unfortunately, the probability of occurrence of convergent or parallel amino acid changes is so small that we rarely observe the same convergent or parallel amino acid changes more than once in a given set of sequence data. For this reason, Zhang and Kumar (1997) considered all convergent or all parallel changes and computed the probabilities of occurrence of these changes by chance, assuming that these changes occur with the same probability for all amino acid sites.

Tests of Convergent and Parallel Evolution

Let us first consider the test of convergent evolution. As mentioned above, the test of convergent evolution is to examine whether the number of convergent-change sites observed in a given set of amino acid sequences significantly exceeds the expected number due to chance events alone. To compute the expected number, we first consider the probability of obtaining a given set of amino acids at both exterior and interior nodes using the tree in Figure 11.8B. Let x_i be the amino acid at the i-th node of the tree. It takes any of the 20 different amino acids. If we use a reversible model of amino acid substitution such as the Dayhoff or the JTT model, we can start the evolutionary process from any point of the unrooted tree as mentioned in chapter 8. We therefore assume that the starting point of evolutionary change is node 1 for convenience. The probability of observing a given set of amino acids or a configuration of amino acids $\mathbf{x} = (x_1, x_2, \cdots, x_8)$ at a site is then given by

$$P_x = g_{x_1} P_{x_1x_6}(b_1) P_{x_6x_2}(b_2) P_{x_6x_7}(b_6) \times P_{x_7x_3}(b_3) P_{x_7x_8}(b_7) P_{x_8x_4}(b_4) P_{x_8x_5}(b_5) \quad (11.8)$$

where g_{x_1} is the relative frequency of amino acid x_1 among all sequences and $P_{x_ix_j}(b_k)$ is the probability of change of amino acid x_i to x_j when the branch length is b_k. This equation is similar to Equation (8.1), but here b_k is not a parameter to be estimated by maximum likelihood but is a constant of which the value has already been estimated. We also assume that the ancestral amino acid sequences have already been inferred by the methods discussed earlier in this chapter, and thus we can identify all pairs of convergent or parallel amino acid changes in the phylogenetic tree.

We now focus our attention to the evolutionary lineages in which convergent evolution is likely to have happened for some biological reasons, as in the case of cow and langur stomach lysozymes. We then compute the probability of occurrence of convergent changes by chance. Suppose that our focused lineages are the branches 6 → 1 and 8 → 4 in Figure 11.8B. The probability of occurrence of convergent changes at a site (f) is the sum of the probabilities of occurrence of all amino acid configurations satisfying the conditions $x_1 = x_4$, $x_1 \neq x_6$, $x_4 \neq x_8$, and $x_6 \neq x_8$. Therefore, we have

$$f = \sum_{x_1}\sum_{x_2}\sum_{x_4=x_1}\sum_{x_3}\sum_{x_5}\sum_{x_6\neq x_1}\sum_{x_7}\sum_{x_8\neq x_4,x_6} p_x \quad (11.9)$$

Since the only amino acids at nodes 1, 4, 6, and 8 and the branch lengths b_1, b_4, b_6, and b_7 affect f, Equation (11.9) can be simplified to

$$f = \sum_{x_1}\sum_{x_4=x_1}\sum_{x_6\neq x_1}\sum_{x_8\neq x_4,x_6} g_{x_1} P_{x_1x_6}(b_1) \cdot P_{x_6x_8}(b_6 + b_7) P_{x_8}(b_4) \quad (11.10)$$

If the sequences used are n amino acids long and all sites evolve following the same substitution model, the observed number of convergent-change sites (s) follows a binomial distribution with the mean and variance equal to nf and $nf(1 - f)$, respectively. So, the probability of observing s or more convergent-change sites by chance (P_A) is given by

$$P_A = \sum_{i=s}^{n} \frac{n!}{i!(n-i)!} f^i (1-f)^{n-i} = 1 - \sum_{i=0}^{s-1} \frac{n!}{i!(n-i)!} f^i (1-f)^{n-i} \quad (11.11)$$

This equation is applicable when s is equal to or greater than 1. When s is 0, P_A is obviously 1. We can therefore use the above equation for testing convergent evolution. That is, if P_A is equal to or smaller than 0.01, one may conclude that the number of convergent-change sites is too large

to be explained by chance, and thus the occurrence of convergent evolution is significant at the 1% level.

Essentially the same test can be developed for parallel evolution. The only thing necessary is to change the requirements of amino acid configurations and impose the conditions $x_1 = x_4$, $x_1 \neq x_8$, and $x_4 \neq x_8$ in Equation (11.9). The rest of the computation is the same as before. In the above formulation, we considered only two evolutionary lineages, but it is possible to include more than two lineages, if there is a good biological reason. Of course, if we consider three or more lineages, the P_A value is usually very small. (See *http://mep.bio.psu.edu*)

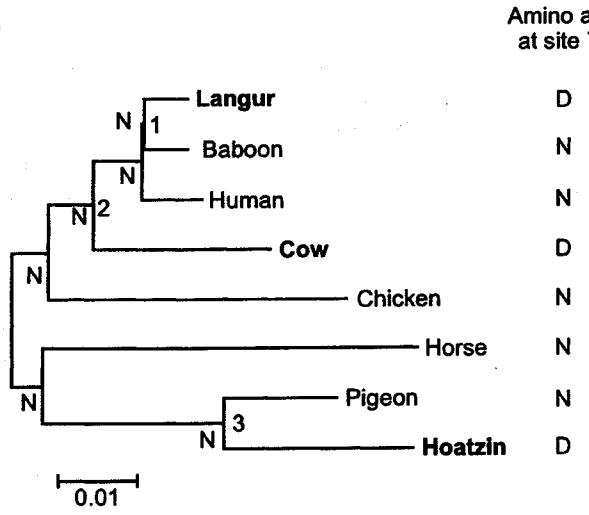
Example 11.3. Stomach Lysozymes of Foregut-Fermenting Animals

As mentioned earlier, ruminants (e.g., cow and deer) and colobine monkeys (e.g., hanuman langur) belong to different orders of mammals, but they share the fermentative foregut, in which grass and leaves are fermented by bacteria. To understand the mechanism of adaptation of the stomach lysozyme, many authors have sequenced stomach and nonstomach lysozymes from various organisms (e.g., Stewart et al. 1987; Irwin et al. 1993; Kornegay 1996; Messier and Stewart 1997). The phylogenetic relationships of lysozyme genes are complicated because of the presence of multiple genes in some species. Zhang and Kumar (1997) conducted a phylogenetic analysis of all the protein sequences available and chose the tree and organisms presented in Figure 11.9A to study convergent and parallel evolution of the lysozyme sequences. (Here we have eliminated the rat sequence from their figure to simplify the explanation.) The topology of this tree is identical with that of Kornegay et al.'s (1994), except that the baboon sequence is included. The langur, cow, and hoatzin sequences are the same as those used by the previous authors. In the following, we assume that this tree represents the true gene tree. Of the eight organisms used in Figure 11.9A, the langur, cow, and hoatzin are foregut fermenters. Therefore, we focus our attention to these three organisms, first pairwise and then all the three together.

We first inferred all ancestral amino acid sequences by the Bayesian methods described earlier and then identified eight amino acid sites at which convergent or parallel changes were observed. Figure 11.9B shows amino acids in the present-day organisms and the ancestral amino acids at nodes 1, 2, and 3 for the sites at which convergent or parallel evolution was observed.

If we focus our attention to the cow and hoatzin lineages, there are three parallel-change sites (sites 75, 76, and 87) and one convergent-change site (site 83). The probability (f) of parallel changes occurring in the two lineages can be computed by the method described above. In this computation, we used the JTT-f model for computing $P_{ij}(b_k)$ and obtained $f = 0.0046$. Since the total number of amino acid sites used (n) was 124, the expected number of parallel-change sites becomes $nf = 0.574$, which is considerably smaller than the observed number. The probability that parallel change occurs at three sites or more can be computed by Equation (11.11) and becomes $P_A = 0.021$. Since this is smaller than 0.05, we may conclude that the parallel amino acid changes are caused by pos-

(A)



(B)

	14	21	41	75	76	83	87	126
Langur	**K**	**K**	**E**	**D**	**A**	**A**	**N**	**K**
Node 1	R	R	Q	N	A	A	D	Q
Baboon	R	R	Q	N	A	A	D	Q
Human	R	R	R	N	A	A	D	Q
Cow	**K**	**K**	**K**	**D**	**G**	**E**	**N**	**E**
Node 2	R	R	Q	N	A	A	D	R
Chicken	R	R	Q	N	L	A	S	R
Horse	A	G	R	N	A	K	E	A
Pigeon	R	V	T	N	A	K	D	R
Hoatzin	**E**	**E**	**E**	**D**	**G**	**E**	**N**	**K**
Node 3	R	E	Q	N	A	K	D	R

FIGURE 11.9. Parallel and convergent amino acid substitutions in the stomach lysozymes. (A) The phylogenetic relationships of the eight vertebrate lysozymes. (B) Parallel-change and convergent-change sites identified in the two-lineage and three-lineage comparisons. Site numbers are based on the human lysozyme sequence. (Some of the amino acid sites were deleted because of insertions/deletions.) The stomach lysozyme sequences of the foregut fermenters are shown in boldface.

itive selection. We can do similar computations for the convergent changes for the same cow and hoatzin lineages and show that the expected number of convergent-change sites is 0.106 and P_A is 0.100. Therefore, it is possible that the occurrence of one event of convergent changes is due to chance effect.

Similar computations were made for other focused lineages, and the results are presented in Table 11.2. In each set of focused lineages, the event of parallel amino acid changes has occurred more often than expected by chance. Therefore, we may conclude that parallel amino acid changes in this data set are caused by positive selection. By contrast, convergent evolution was statistically significant only for the langur and hoatzin lineages.

Some Remarks

In the present analysis, however, some caution is necessary about the interpretation of the results. As mentioned earlier, the Zhang-Kumar formulation is based on the assumption that all amino acid sites follow the same evolutionary pattern, and no special consideration is given to any pair of amino acids, whether or not they are easily interchanged. For example, the amino acids at sites 14 and 21 in Figure 11.9B are either arginine (R) or lysine (K) except a few sequences. Dayhoff et al.'s (1978) matrix of amino acid substitution indicates that arginine and lysine are the amino acids that are most often interchanged in the evolutionary process despite the fact that the interchange requires two nucleotide substitutions. Similarly, asparagine (N) and aspartic acid (D) observed at sites 75 and 87 are also often interchanged. Therefore, it is possible that the f value obtained by considering all amino acids is too small for these particular amino acids, and therefore the test used is too liberal. Nevertheless, the observation that at site 75 only the foregut-fermenting organisms have amino acid D and all other sequences including the ancestral sequences have N is a strong indication of positive selection.

Another problem with the Zhang-Kumar method is that the accuracy of inference of parallel and convergent amino acid changes declines as sequence divergence increases (see Zhang and Kumar 1997). Therefore, one should be cautious about the interpretation of P_A values when the number of amino acid substitutions per site (d) is greater than 0.5. A more rigorous mathematical treatment of convergent and parallel evolution is desirable.

Doolittle (1994) criticized a number of published papers that claimed convergent evolution at the amino acid sequence level. The convergent evolution defined by Zhang and Kumar seems to be very rare. However, parallel evolution at the sequence level, which is also an indication of positive selection, apparently occurs more frequently. We have already shown examples of parallel evolution in the lysozymes of foregut fermenting animals. Another example of parallel evolution is the evolutionary change of color vision genes discussed earlier. Figure 11.2 shows the amino acids of present-day vertebrates and the ancestral amino acids

Table 11.2 Tests of parallel and convergent evolution of stomach lysozyme sequences of the cow, langur, and hoatzin.

Species	s	nf	P_A	Site positions
A. Parallel changes				
Cow-Hoatzin	3	0.574	0.021	75, 76, 87
Cow-Langur	4	0.225	<0.001	14, 21, 75, 87
Langur-Hoatzin	3	0.161	<0.001	41, 75, 87
Cow-Langur-Hoatzin	2	0.010	<0.001	75, 87
B. Convergent changes				
Cow-Hoatzin	1	0.106	0.100	83
Cow Langur	0	0.006	1.000	—
Langur-Hoatzin	1	0.034	0.033	126
Cow-Langur-Hoatzin	0	0.000	1.000	—

inferred at positions 277 and 285. The "red" gene and the ancestral genes all have tyrosine (Y) and threonine (T), whereas the "green" genes mostly have phenylalanine (F) and alanine (A). Application of the Zhang-Kumar test shows that the parallel evolution observed here is highly significant. Yokoyama and Yokoyama (1990) reported that the red and green opsins of the characid fish *Astyanax fasciatus* have the same amino acids as those of human opsins at positions 277 and 285, and the Zhang-Kumar test indicates that this parallel evolution is also highly significant.

In the past, a number of authors have reported cases of convergent evolution at the amino acid level. In our terminology, many of these reports actually refer to parallel evolution rather than to convergent evolution. Some authors have suggested the possibility that two divergent amino acid sequences may become similar because of convergent evolution and that this may introduce errors when phylogenetic trees are constructed. In practice, however, convergent or parallel evolution at the amino acid level occurs very rarely, and the effect of these events on the phylogenetic trees reconstructed is negligibly small (Doolittle 1994).

12

Genetic Polymorphism and Evolution

12.1. Evolutionary Significance of Genetic Polymorphism

Most natural populations of organisms contain a large amount of genetic variation. In sexually reproducing or outbreeding organisms, any pair of individuals is genetically different except identical twins. When a genetic locus is identifiable at the protein level, the locus often contains two or more alleles within a population. Existence of two or more alleles with substantial relative frequencies in a population (usually more than 1%) is called **genetic polymorphism**. Genetic polymorphism at a locus is generated by mutations such as nucleotide substitution, insertions/deletions, gene conversion, and interallelic recombination. However, most of these new mutations are eliminated from the population by genetic drift or purifying selection, and only a minority of them are incorporated into the population by chance, positive selection, or overdominant selection, as mentioned in chapter 2. The main subject of population genetics is to study the generation and maintenance of genetic polymorphism and to understand the mechanisms of evolution at the population level.

In the study of evolution, it is important to know the extent of genetic variation within and between populations (or species) and to understand what kind of variation is important for forming new species. At the present time, we know very little about the mechanism of speciation or evolution of reproductive isolation, but it is relatively easy to evaluate the extent of genetic variation within and between populations at the molecular level (Thorpe 1982; Nei 1987; Hamrick and Godt 1990; Avise 1994). As was pointed out by Kimura and Ohta (1971), long-term evolution of genes and genetic polymorphism are merely two facets of the same evolutionary process.

In the study of molecular population genetics, the neutral theory proposed by Kimura (1968) plays an important role. In this theory, genetic variation at the molecular level is considered to be largely neutral, and the extent of variation is determined primarily by the mutation rate and the effective population size (Kimura and Crow 1964; Nei 1987). Therefore, it is possible to test the hypothesis of neutral evolution by

comparing the observed and predicted amounts of genetic variation. If the discrepancy between the observed and predicted amounts is large, one may invoke some kind of selection.

Although the high degree of genetic polymorphism within species was obvious from the variation of morphological characters, quantitative evaluation of genetic variation became possible only after the technique of protein electrophoresis was introduced in the study of population genetics in the mid 1960s (Harris 1966; Lewontin and Hubby 1966). This technique is capable of detecting only charge changes of proteins, yet application of this technique revealed that most natural populations contain a large amount of genetic variation, and it facilitated the study of population genetics enormously (Lewontin 1974; Nei 1975). The period from 1966 to 1980 was the time when the neutral theory of molecular evolution was hotly debated (Kimura 1983; Nei 1987; Gillespie 1991). Another rejuvenation of population genetics occurred around 1980, when the study of DNA polymorphism was initiated by using restriction enzyme techniques and direct DNA sequencing (Avise et al. 1979; Brown et al. 1979; Kreitman 1983; Cann et al. 1987). In particular, DNA sequencing made it possible to study both synonymous and nonsynonymous polymorphism separately, so one could study adaptive and nonadaptive evolution at the DNA level (Hughes and Nei 1988; McDonald and Kreitman 1991). Progress in this area has been summarized by Avise (1994), Kreitman and Akashi (1995), and Li (1997).

As genetic polymorphism is studied at the DNA level, it has become obvious that a detailed history of polymorphic alleles can be studied by phylogenetic analysis. Thus one can trace the history of disease genes such as the phenylketonuria gene (Wang et al. 1991) or particular alleles such as the 9-base deletion mutant of mitochondrial DNA (Redd et al. 1995). Such gene genealogies are useful even for inferring the evolutionary histories of populations (e.g., Vigilant et al. 1991; Bonatto and Salzano 1997). Phylogenetic analysis of polymorphic genes is also useful for studying the evolutionary mechanism of rapidly evolving genes such as human immunodeficiency virus (HIV) and influenza virus genes (Fitch et al. 1997; Yamaguchi and Gojobori 1997). Furthermore, the recent discovery of a high degree of polymorphism at microsatellite DNA has given the promise that detailed evolutionary relationships of populations or closely related species may soon be clarified (Bowcock et al. 1994; Jorde et al. 1995).

The purpose of this chapter is to discuss statistical methods that are used for studying the problems mentioned above. However, many of the important statistical methods for analyzing genetic polymorphism and their mathematical derivation have already been described in detail in Nei's (1987) book; numerical examples for computing various statistical quantities are also presented. Therefore, we present only the final results and some newly developed methods. In population genetics, there are mathematical theories that are useful for conceptual understanding of the population dynamics of genes. Any reader who is interested in these theories should refer to Kimura (1983), Nei (1987), Gillespie (1991), Hartl and Clark (1997), and Li (1997).

12.2. Analysis of Allele Frequency Data

Alleles at a locus represent different DNA sequences at the molecular level, but when allelic differences are studied by immunological techniques or protein electrophoresis, we do not necessarily detect all the allelic differences at the DNA level. Yet, these techniques are capable of detecting a large amount of genetic variation, and thus they are still used for studying population genetics and evolution. The genetic variation detected by these techniques or the variation observed at microsatellite loci is described by variation in allele frequencies.

Allele Frequencies and Hardy-Weinberg Equilibrium

The relative frequency of a particular allele in a population is called the **allele frequency**. It is a fundamental parameter in the study of evolution, because the genetic change of a population is usually described by the changes in allele frequencies. Consider a locus with two alleles, A_1 and A_2. In diploid organisms, there are three possible genotypes at this locus, that is, A_1A_1, A_1A_2, and A_2A_2. Let N_{11}, N_{12}, and N_{22} be the numbers of genotypes A_1A_1, A_1A_2, and A_2A_2 in a population, respectively, and $N_{11} + N_{12} + N_{22} = N$. The relative frequencies of A_1A_1, A_1A_2, and A_2A_2 are then given by $X_{11} = N_{11}/N$, $X_{12} = N_{12}/N$, and $X_{22} = N_{22}/N$, respectively. On the other hand, the allele frequency of A_1 is given by

$$\begin{aligned} x_1 &= (2N_{11} + N_{12})/(2N) \\ &= X_{11} + X_{12}/2 \end{aligned} \tag{12.1}$$

Obviously, the allele frequency of A_2 is $x_2 = 1 - x_1$.

Theoretically, the **genotype frequencies** X_{11}, X_{12}, and X_{22} can take any value between 0 and 1 with the restriction of $X_{11} + X_{12} + X_{22} = 1$. In many organisms, however, mating occurs roughly at random, and genotypes are produced approximately by random union of male and female gametes. In this case, the genotype frequencies are approximately given by the expansion of $(x_1 + x_2)^2$. That is,

$$X_{11} = x_1^2, \qquad X_{12} = 2x_1x_2, \qquad X_{22} = x_2^2 \tag{12.2}$$

This property was first noted by Hardy (1908) and Weinberg (1908) independently, so it is called the Hardy-Weinberg principle.

The relationship between allele and genotype frequencies when more than two alleles exist at a locus is essentially the same as that for the case of two alleles. Let q be the number of alleles existing at a locus and denote the i-th allele by A_i. There are q possible homozygotes and $q(q-1)/2$ possible heterozygotes, the total number of genotypes being $q(q+1)/2$. We represent the frequency of genotype A_iA_j by X_{ij}. Then, the frequency of the i-th allele is given by

$$x_i = X_{ii} + \frac{1}{2}\sum_{j \neq i} X_{ij} \tag{12.3}$$

where $\Sigma_{j \neq i}$ indicates the summation of X_{ij} over all j's except for $j = i$. For example, when there are three alleles, the frequency of A_1 is given by $x_1 = X_{11} + (X_{12} + X_{13})/2$.

The genotype frequencies under random mating are given by the expansion of $(x_1 + x_2 + \cdots + x_q)^2$. Therefore, the frequencies of homozygote A_iA_i and heterozygote A_iA_j are

$$X_{ii} = x_i^2 \quad \text{and} \quad X_{ij} = 2x_ix_j \tag{12.4}$$

respectively. These are the Hardy-Weinberg proportions for the case of multiple alleles.

Estimation of Allele Frequencies in a Random Mating Population

When the population size is large, it is not easy to examine the genotypes of all individuals in order to determine the allele frequencies. Therefore, we must sample a certain number of individuals from the population and estimate the population allele frequencies from this sample. Surely the accuracy of the estimates of allele frequencies depends on sample size, but it is also affected by dominance, mating system, and the allele frequencies themselves. This problem has been discussed in many textbooks (e.g., Li 1976; Weir 1996; Hartl and Clark 1997). In the following, we consider only the simple case of codominant alleles.

When all alleles are codominant and consequently all genotypes are identifiable, the allele frequencies can be estimated by counting the number of genes in the sample. Suppose that there are q alleles at a locus and m diploid individuals are sampled from the population. Let m_{ij} be the number of individuals with genotype A_iA_j, the total number of individuals being m, that is, $\Sigma m_{ij} = m$. The allele frequency of A_i is then estimated by

$$\hat{x}_i = \left(2m_{ii} + \sum_{j \neq i} m_{ij}\right) / (2m) \tag{12.5}$$

Most protein polymorphisms detectable by electrophoresis or those at microsatellite loci are controlled by codominant alleles. At the red-cell acid phosphatase locus in humans, there are three major alleles, A, B, and C, all being codominant. Hopkinson and Harris (1969) examined the genotypes of 880 individuals in England and obtained the results given in Table 12.1. From these results we can obtain the estimates ($\hat{x}_1$, $\hat{x}_2$, and

Table 12.1 Observed and expected numbers of genotypes at the red cell acid phosphatase locus in humans in a survey of an English population.

Genotype	AA	BB	CC	AB	AC	BC	Total
Observed number	119	282	0	379	39	61	880
Expected number	122.3	286.4	2.8	374.2	37.3	57.0	880

Source: Cited from Mourant et al. (1976).

$\hat{x}_3$) of allele frequencies of A, B, and C. They become $\hat{x}_1 = 0.373$, $\hat{x}_2 = 0.570$, and $\hat{x}_3 = 0.057$.

The expected numbers of genotypes under random mating are obtained by replacing x_i's in Equation (12.4) by their respective estimate $\hat{x}_i$'s and multiplying X_{ii} and X_{ij} by m. The results are presented in Table 12.1. The agreement between the observed and expected numbers of genotypes can be tested by the X^2 statistic. The general formula for X^2 is

$$X^2 = \sum (O - E)^2 / E \tag{12.6}$$

where O and E stand for the observed and expected numbers, and Σ indicates summation over all genotypes. In the present case, there are six genotypes, but the expected number of genotype CC is less than three, and no individuals of this genotype were observed. We therefore combine this genotype and genotype BB and conduct the X^2 test. We then have $X^2 = 3.41$. The number of degrees of freedom for this X^2 is two, because there are four independent observations (genotype frequencies) and we have estimated two independent parameters (allele frequencies). This X^2 is not significant at the 5% level, implying that the agreement between the observed and expected frequencies is satisfactory.

When the number of alleles at a locus is very large and the number of individuals examined is relatively small, Equation (12.6) is not very powerful. In this case, one may use Guo and Thompson's (1992) exact test with a computer resampling procedure. This test will be useful for such genetic loci as microsatellite DNA and major histocompatibility complex loci, where the number of alleles within a population can be as large as 20.

Deviations from Hardy-Weinberg Proportions

Although Hardy-Weinberg equilibrium approximately holds in outbreeding organisms, it can be disturbed by a number of factors such as inbreeding, assortative mating, natural selection, and population subdivision. In this section, we discuss this problem without going into details.

A Pair of Alleles

When there are two alleles at a locus, any deviation from Hardy-Weinberg proportions may be measured by a single parameter (F) called the **fixation index** (Wright 1951, 1965). If we use this index, the genotype frequencies are given by

$$X_{11} = (1 - F)x_1^2 + Fx_1 \tag{12.7a}$$

$$X_{12} = 2(1 - F)x_1x_2 \tag{12.7b}$$

$$X_{22} = (1 - F)x_2^2 + Fx_2 \tag{12.7c}$$

The fixation index F can be positive or negative, depending on the case.

From Equation (12.7b), we obtain

$$F = (2x_1x_2 - X_{12})/(2x_1x_2) \tag{12.8}$$

If we note that $2x_1x_2$ is the expected frequency of heterozygotes under random mating (h) and x_{12} is the observed frequency of heterozygotes in the population (h_0), Equation (12.8) can be written as

$$F = (h - h_0)/h \tag{12.9}$$

This indicates that F is positive when h_0 is smaller than h and negative when h_0 is greater than h. In the presence of inbreeding such as consanguineous mating or selfing, the observed frequency of heterozygotes declines, so that F becomes positive.

Multiple Alleles

When there are q alleles at a locus, we generally need $q(q-1)/2$ fixation indices to specify all genotype frequencies in terms of allele frequencies and fixation indices. However, if the deviations from Hardy-Weinberg equilibrium occur solely by inbreeding, the deviations can be described by a single fixation index. In this case, the frequency (X_{ij}) of homozygote $A_iA_i(i = 1, 2, \cdots, q)$ is given by

$$X_{ii} = (1 - F)x_i^2 + Fx_i \tag{12.10}$$

whereas the frequency (X_{ij}) of heterozygote A_iA_j is

$$X_{ij} = 2(1 - F)x_ix_j \tag{12.11}$$

The expected frequency of heterozygotes under the assumption of Hardy-Weinberg equilibrium is given by $h = 2\sum_{i<j} x_ix_j$, whereas the observed frequency is $h_0 = \sum_{i<j} X_{ij}$. Therefore, Equation (12.9) holds, even for a locus with multiple alleles. In population genetics, h and h_0 are often called the **expected** and the **observed heterozygosities,** respectively.

12.3. Genetic Variation in Subdivided Populations

Wahlund's Principle

So far, we have considered a single population, whether or not the population is inbred. In practice, however, most natural populations are subdivided into many different breeding units or subpopulations, though these subpopulations are not completely isolated. In this case, it is important to study the genetic variation within and between populations.

Let us consider a population divided into s subpopulations, in each of which Hardy-Weinberg equilibrium holds. Let x_k be the frequency of allele A_1 in the k-th subpopulation, so that the frequencies of genotypes

A_1A_1, A_1A_2, and A_2A_2 in this subpopulation are given by x_k^2, $2x_k(1 - x_k)$, and $(1 - x_k)^2$, respectively. We denote by w_k the relative size of the k-th subpopulation with $\Sigma\, w_k = 1$. The mean frequencies of A_1A_1, A_1A_2, and A_2A_2 in the entire population are then given by

$$X_{11} = \sum_{k=1}^{s} w_k x_k^2 = \bar{x}^2 + \sigma^2 \tag{12.12a}$$

$$X_{12} = 2\sum_{k=1}^{s} w_k x_k(1 - x_k) = 2\bar{x}(1 - \bar{x}) - 2\sigma^2 \tag{12.12b}$$

$$X_{22} = \sum_{k=1}^{s} w_k(1 - x_k)^2 = (1 - \bar{x})^2 + \sigma^2 \tag{12.12c}$$

where $\bar{x} \equiv \Sigma\, w_k x_k$ and $\sigma^2 \equiv \Sigma\, w_k\,(x_k - \bar{x})^2$ are the mean and variance of allele frequency among subpopulations. Comparison of these equations with those in Equation (12.7) shows that σ^2 corresponds to $F\bar{x}\,(1 - \bar{x})$. Therefore,

$$F = \sigma^2/[\bar{x}(1 - \bar{x})] \tag{12.13}$$

This indicates that if a population is subdivided into many breeding units, the frequency of homozygotes tends to be higher than the Hardy-Weinberg proportion. This property was first noted by Wahlund (1928) and is called **Wahlund's principle**. Note that F is 0 when the allele frequency x_k is the same for all subpopulations ($\sigma^2 = 0$) and is 1 when every subpopulation is fixed with allele A_1 or A_2 [$\sigma^2 = \bar{x}(1 - \bar{x})$].

In the above formulation, we assumed that the relative population size, w_k, is known. In practice, we rarely know w_k, so that we usually assume $x_k = 1/s$. This assumption is reasonable, because population size is transitory and we are usually interested in the genetic differentiation of populations without regard to population size. (For example, the British population has remained nearly the same for the last 50 years, whereas the Chinese and Indian populations have doubled or tripled.)

In the presence of multiple alleles at a locus, we denote the frequency of the i-th allele in the k-th subpopulation by x_{ki} and the relative size of the k-th population by w_k with $\Sigma\, w_k = 1$. The mean frequencies (X_{ii} and X_{ij}) of genotypes A_iA_i and A_iA_j over subpopulations are then given by

$$X_{ii} = \sum\nolimits_k w_k x_{ki}^2 = \bar{x}_i^2 + \sigma_i^2 \tag{12.14a}$$

$$X_{ij} = 2\sum\nolimits_k w_k x_i x_j = 2\bar{x}_i\bar{x}_j + 2\sigma_{ij} \tag{12.14b}$$

where $\bar{x} = \Sigma_k w_k x_{ki}$, $\sigma_i^2 = \Sigma_k w_k(x_{ki} - \bar{x}_i)^2$, and $\sigma_{ij} = \Sigma_k w_k(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$ (Nei 1965). However, the mean frequencies of genotypes A_iA_i and A_iA_j can also be written as

$$X_{ii} = (1 - F_{ii})\bar{x}_i^2 + F_{ii}\bar{x}_i \tag{12.15a}$$

$$X_{ij} = 2(1 - F_{ij})\bar{x}_i\bar{x}_j \tag{12.15b}$$

by analogy to Equations (12.10) and (12.11). Here F_{ij} may vary with genotype.

Comparison of Equations (12.14a) and (12.15a) and Equations (12.14b) and (12.15b) gives

$$F_{ii} = \sigma_i^2/[\bar{x}_i(1 - \bar{x}_i)] \tag{12.16a}$$

$$F_{ij} = -\sigma_{ij}/\bar{x}_i\bar{x}_j \tag{12.16b}$$

If the differentiation of subpopulations has occurred at random, we expect $F_{ii} = F_{ij} = F$. Therefore, the hypothesis of random differentiation of allele frequencies among subpopulations can be tested by examining $F_{ii} = F_{ij} = F$. However, when there are a large number of alleles at a locus, as in the case of microsatellite loci, and sample size is relatively small, the estimates of F_{ij}'s may be subject to large sampling errors. In fact, the estimate of F_{ij} may become negative and even lower than -1 (National Research Council 1996). Large negative estimates of F_{ij}'s are almost always associated with low frequency alleles and are mainly due to sampling errors.

Fixation Indices

In actual populations, the genotype frequencies in each subpopulation do not necessarily follow Hardy-Weinberg equilibrium, and therefore F in each subpopulation may not be 0. Wright (1943, 1951, 1965, 1978) proposed that the deviations of genotype frequencies in a subdivided population be measured in terms of three parameters, F_{IS}, F_{IT}, and F_{ST}, which are called **fixation indices**. In Wright's definition, F_{IS} and F_{IT} are the correlations between the two uniting gametes relative to the subpopulation and relative to the total population, respectively, whereas F_{ST} is the correlation between two gametes drawn at random from each subpopulation and measures the degree of genetic differentiation of subpopulations. They are related by the following formula

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST}) \tag{12.17}$$

The fixation indices are useful for understanding the breeding structure of populations or the pattern of selection associated with polymorphic alleles.

The basic idea of Wright's formulation is to assume that the populations or subpopulations under investigation are derived from a common ancestral population at the same time and that all populations are equally related to one another whether or not there is migration among the populations (Figure 12.1A). In other words, the populations are considered as a random sample from a set of infinitely many equally related populations. This idealized population structure approximately applies to a set of populations that are artificially generated in laboratories (Buri 1956) or in agricultural experiments. In natural populations, however,

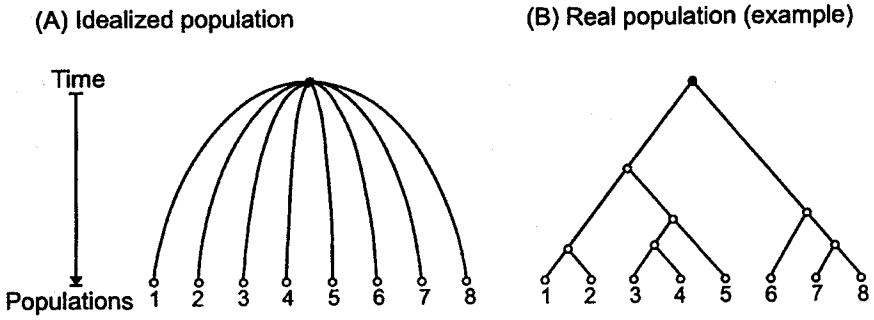


FIGURE 12.1. (A) Idealized model of population structure. (B) Realistic model of population structure (an example). Mathematical formulations by Wright (1951) and Cocharham (1969) apply only to model A, whereas Nei's (1977) formulation applies to both models A and B.

this model almost never applies. Most natural populations have some phylogenetic (historical) relationships as illustrated in Figure 12.1B, and the phylogenetic relationship depends on the populations under investigation (see Figure 12.3 for an actual example). In other words, the populations are not a random sample from a set of many equally related populations. Note also that the population size usually varies extensively among populations, and, if there is migration, the migration rate is not the same for all pairs of populations. Therefore, the concept of correlations of uniting gametes hardly applies to real populations.

For this reason, Nei (1977) redefined the fixation indices without using the concept of correlation of uniting gametes and showed that Equation (12.17) holds for any situation regardless of the phylogenetic relationships, migration pattern, and the number of alleles and whether or not there is selection. In particular, he showed that all fixation indices can be defined by using the observed and expected heterozygosities for the population under investigation.

In this theory, fixation indices F_{IS}, F_{IT}, and F_{ST} are defined in terms of three heterozygosities, that is, **observed within-population heterozygosity** (h_0), **expected within-population heterozygosity** (h_S), and **expected total heterozygosity** (h_T). In this context, heterozygosity is often called **gene diversity**, as will be discussed later. The observed within-population heterozygosity is defined by

$$h_0 = \sum_k^s w_k \sum_{i \neq j}^q X_{kij} \tag{12.18}$$

where X_{kij} $(i \neq j)$ is the frequency of A_iA_j in the k-th population. The expected within-population heterozygosity is

$$h_S = 1 - \sum_k^s w_k \sum_i^q x_{ki}^2 \tag{12.19}$$

whereas the expected total heterozygosity is given by

$$h_T = 1 - \sum_i^q \bar{x}_i^2 \tag{12.20}$$

where $\bar{x}_i = \Sigma_k w_k x_{ki}$. As mentioned earlier, it is customary to assume $w_k = 1/s$ because information on w_k is not usually available.

The fixation indices can now be defined by using h_0, h_S, and h_T in the following way.

$$F_{IS} = (h_S - h_0)/h_S \tag{12.21a}$$

$$F_{IT} = (h_T - h_0)/h_T \tag{12.21b}$$

$$F_{ST} = (h_T - h_S)/h_T \tag{12.21c}$$

For the derivation of these equations, see Nei (1987).

Note that F_{IS} and F_{IT} can be negative when h_0 is unusually high for some reason. However, F_{ST} is nonnegative, since h_T is always greater than or equal to h_S. Note also that the above equations apply to the case of two alleles at a locus, and in this case F_{ST} in Equation (12.21c) becomes equal to F in Equation (12.13). Actually, Wright's formulation of fixation indices was developed for the case of two alleles, and for this reason F_{ST} defined for the case of multiple alleles is often denoted by G_{ST}, which was originally called the **coefficient of gene differentiation** (Nei 1973). It should be noted that the fixation indices defined above apply to any situation whether or not there is selection, because they are defined in terms of the present allele and genotype frequencies. In the case of two alleles, G_{ST} becomes identical with F in Equation (12.13) and takes a value between 0 and 1. However, when there are many alleles at a locus and h_S is high, G_{ST} may be considerably smaller than 1 even if there are alleles that are not shared by different subpopulations (Jin and Chakraborty 1995; Hedrick 1999). As mentioned earlier, microsatellite DNA loci are highly polymorphic with high h_S and h_T. Therefore, for these loci G_{ST} is generally quite small (National Research Council 1996).

The quantity d_{ST} $(= h_T - h_S)$ in Equation (12.21c) can be written as

$$d_{ST} = (\Sigma_k \Sigma_l d_{kl})/s^2$$

where $d_{kl} = \Sigma_i (x_{ki} - x_{li})^2/2$ (see Nei 1987, p. 189). Obviously, $d_{kk} = 0$, so that we have $d_{ST} = (s-1)d'_{ST}/s$, where

$$d'_{ST} = \sum_{k \neq l} d_{kl}/[s(s-1)] \tag{12.22}$$

Here, the summation is taken over all d_{kl} except d_{kk}'s. (d'_{ST} is equal to $\bar{D}_m$ in Equation [12.46].) Therefore, the average allele frequency differentiation between populations may be measured by d'_{ST} rather than by d_{ST}. We can then use

$$F'_{ST} = d'_{ST}/h'_T = (h'_T - h_S)/h'_T \tag{12.23}$$

as another measure of the extent of genetic differentiation of subpopulations, where $h'_T = h_S + d'_{ST}$. This has an advantage over F_{ST} in that it is

independent of s. However, h'_T is no longer the heterozygosity for the total population. In this case, one must redefine F_{IT} as

$$F'_{IT} = (h'_T - h_0)/h'_T \tag{12.24}$$

in order to maintain the relationship in Equation (12.17), while F_{IS} remains the same. In practice, the difference between F_{ST} and F'_{ST} or F_{IT} and F'_{IT} is very small if $s \geq 5$.

Estimation of Fixation Indices

In the above formulation, we defined fixation indices in terms of population allele and genotype frequencies. In practice, we must estimate these indices from sample allele and genotype frequencies.

We assume that m_k diploid individuals are randomly chosen from the k-th subpopulation. Let $\hat{x}_{ki}$ and $\hat{X}_{kij}$ be the frequencies of allele A_i and genotype A_iA_j in the sample from the k-th subpopulation, respectively. Since the fixation indices are defined in terms of h_0, h_S, and h_T (or h'_T), we must have estimates of these quantities. For the reason mentioned earlier, we assume $w_k = 1/s$. Estimation of h_0 is simple, because $\hat{X}_{ii}$ is an unbiased estimate of X_{ii} under our assumption. An unbiased estimate of h_0 in the k-th subpopulation is $1 - \Sigma_i \hat{X}_{kii}$, and thus h_0 may be estimated by

$$\hat{h}_0 = 1 - \sum_{k=1}^{s} \sum_{i=1}^{q} \hat{X}_{kii}/s \tag{12.25}$$

Derivation of an unbiased estimator of h_S is more complicated, since $\hat{x}_{ki}^2$ is not an unbiased estimate of x_{ki}^2. However, it can be shown that an unbiased estimator is

$$\hat{h}_S = \frac{\tilde{m}}{\tilde{m} - 1}\left[1 - \sum_{i=1}^{q} \overline{\hat{x}_i^2} - \frac{\hat{h}_0}{2\tilde{m}}\right] \tag{12.26}$$

where $\overline{\hat{x}_i^2} = \Sigma_k \hat{x}_{ki}^2/s$, and $\tilde{m}$ is the harmonic mean of m_k. Similarly, it can be shown that an unbiased estimate of h_T is

$$\hat{h}_T = 1 - \sum_{i=1}^{q} \overline{x}_i^2 + \frac{\hat{h}_s}{\tilde{m}s} - \frac{\hat{h}_0}{2\tilde{m}s} \tag{12.27}$$

where $\overline{x}_i = \Sigma_k \hat{x}_{ki}/s$. Therefore, the estimates of F_{IS}, F_{IT}, and F_{ST} are

$$\hat{F}_{IS} = 1 - \hat{h}_0/\hat{h}_S \tag{12.28a}$$

$$\hat{F}_{IT} = 1 - \hat{h}_0/\hat{h}_T \tag{12.28b}$$

$$\hat{F}_{ST} = 1 - \hat{h}_S/\hat{h}_T \tag{12.28c}$$

Similarly, one can estimate F'_{IT} and F'_{ST} by estimating h'_T. This estimate $(\hat{h}'_T)$ of h'_T is given by $\hat{h}_S + \hat{d}'_{ST}$, where $\hat{d}'_{ST}$ is $s(\hat{h}_T - \hat{h}_S)/(s - 1)$.

Note that $\hat{h}_0$ is not affected by sample size $\tilde{m}$, whereas $\hat{h}_S$ and $\hat{h}_T$ are. Furthermore, the effects of $\tilde{m}$ on $\hat{h}_S$ and $\hat{h}_T$ are negligibly small if $\tilde{m}$ is sufficiently large, say, $\tilde{m} > 30$. Therefore, the above sample size corrections are necessary only when $\tilde{m}$ is small.

$\hat{F}_{ST}$ is of special importance, because this quantity measures the extent of genetic differentiation of subpopulations. The statistical significance of the deviation of $\hat{F}_{ST}$ from zero can be tested by using the usual X^2 test of heterogeneity of gene frequencies (Nei 1987).

Some Remarks

Cockerham (1969, 1973) and Weir and Cockerham (1984) reformulated F_{ST} in terms of the coancestry coefficient (θ) (Falconer 1981) in the inbreeding theory and developed a statistical method for estimating θ (see Weir 1996). The coancestry coefficient is essentially the same as Wright's correlation of uniting gametes, and in this method, the populations used are regarded as a random sample from infinitely many populations that are equally related to one another, as shown in Figure 12.1A. As mentioned earlier, this model is not appropriate for natural populations, which are usually characterized by a complex phylogenetic relationship (Figure 12.1B). By contrast, Nei (1973, 1977) makes no assumption about the history of the populations, so that his theory is applicable to any set of populations. Another difference between Cockerham's and Nei's theories is that in the former the average allele frequencies for the total populations are computed by weighting the allele frequencies in subpopulations with relative sample sizes (u_k), whereas in the latter they are computed without weighting ($u_k = 1$). In real data analysis, sample size often varies extensively with subpopulation, particularly when allele frequency data for different populations are obtained by different authors. This problem is similar to that of weighting for different populations, and for most biological purposes, the unweighted version seems to be more meaningful.

Pons and Petit (1995) recently extended Nei's approach to the case of haploid organisms under Wright's model of population structure (Figure 12.1A). Their mathematical formation is elegant and simpler than Cockerham's and gives results intermediate between Cockerham's and Nei's; they do not use sample sizes as weights for computing the average allele frequencies. In our view, Nei's theory is more appropriate for studying the extent of population differentiation, because Wright's model is not realistic.

Nagylaki (1998) pointed out that the fixation indices defined by Nei are random variables rather than parameters when long-term evolution is considered. This is true, but the population structure often changes drastically from generation to generation. Therefore, it is not always meaningful to consider the fixation indices averaged over generations. Nagylaki redefined Nei's fixation indices by considering conceptual replications of populations. While these redefined fixation indices can be regarded as parameters, he did not show how to estimate them. In Nei's approach the fixation indices are actually parameters, because they are defined in terms of population gene and genotype frequencies in each

generation and they are estimated by sampling finite numbers of genes and genotypes.

Despite these conceptual differences between the different formulations of F_{ST}, all the formulas described here give very similar results for real data unless the sample size varies extensively with subpopulations. Therefore, one may use the simple mathematical treatment presented above. For numerical examples of $\hat{F}_{IS}$, $\hat{F}_{IT}$, and $\hat{F}_{ST}$ obtained by various methods, the readers may consult the papers by Chakraborty and Danker-Hopfe (1991) and Pons and Petit (1995).

Expected Value of F_{ST} or G_{ST}

In population genetics theory, it is customary to present the expected value of F_{ST} and G_{ST} under certain assumptions. For example, if a large number of isolated populations are generated with the same initial allele frequencies and the same effective size (N) for all populations (see Figure 12.1A) and the effect of mutation is negligible, F_{ST} or G_{ST} for the entire population in generation t is given by

$$F_{ST} = G_{ST} = 1 - e^{-t/(2N)} \tag{12.29}$$

(see Nei 1987, p. 359). By contrast, if gene flow occurs among the populations following the island model (Wright 1931) and mutation occurs with a rate of v per generation, G_{ST} eventually reaches an equilibrium value, which is approximately given by

$$G_{ST} = 1/\left[1 + 4N\left(\frac{s}{s-1}\right)(m+v)\right] \tag{12.30}$$

where m is the migration rate per generation (see Takahata and Nei [1984] for the precise definition of m), and s is the number of subpopulations.

Equation (12.29) is useful for a conceptual understanding of the effect of genetic drift on the differentiation of allele frequencies and for predicting the F_{ST} value in experiments to study the effect of genetic drift in small populations as was done by Buri (1956) in *Drosophila melanogaster*. It is also useful for estimating the average effective population size of natural populations when the number of generations since population splitting is known. However, it is very difficult to know the demographic history of natural populations in any species. Furthermore, as was discussed earlier, most natural populations have a unique phylogenetic structure, and the effective sizes of populations may vary drastically with time. Therefore, the applicability of Equation (12.29) is quite limited. We also usually do not know when a set of populations diverged or whether no gene migration has occurred among subpopulations.

Justification of the applicability of Equation (12.30) is a little easier than that of Equation (12.29). It is known that when the migration rate is a few percent or higher ($m \geq 0.01$), G_{ST} reaches the equilibrium value relatively rapidly (in about $1/m$ generations) (Nei et al. 1977; Crow and Aoki 1984). Therefore, one would expect that Equation (12.30) is applicable for estimating migration parameter Nm when $v \ll m$. Indeed, a

large number of authors have estimated Nm in many different organisms. However, the reliability of the estimate of Nm again depends on the population history. In order to obtain a reliable estimate of Nm, it is necessary that the model of population differentiation given in Figure 12.1A approximately applies, and the same population structure with the same N and m must have been maintained at least for about $1/m$ generations. This means that if $m = 0.02$, the population structure should have been the same for about 50 generations. In most species, this assumption is rarely satisfied. For example, the geographical distribution of *Drosophila pseudoobscura* in western United States is known to show a drastic temporal change (Jones et al. 1981). For this reason, we must be cautious about the interpretation of the estimate of Nm obtained in this way.

In general, fixation indices are determined by a complex history of the populations studied, and it is difficult to estimate any population parameters such as N, m, and v from them. However, the fixation indices were developed primarily for quantifying the extent of genetic differentiation of populations, and for this purpose, they are useful. By computing G_{ST} for different sets of populations, we will know which set of populations has the largest or smallest degree of genetic differentiation. In general, one cannot extrapolate an estimate of G_{ST} obtained from a set of populations to another set, because G_{ST} is population specific.

Theoretically, a more informative approach to the study of genetic differentiation of populations would be to conduct a phylogenetic analysis of populations using data from many loci. This analysis gives more information about the genetic relationships of populations than G_{ST} analysis, whether they are caused by genealogical history or partial geographical isolation. Most phylogenetic analyses show that some populations are more closely related than others, as shown in Figure 12.1B. However, F_{ST} or G_{ST} is useful as an overall measure of population differentiation. Particularly when the genealogical history of populations is largely erased by recent migration, F_{ST} and G_{ST} would be a useful measure of comparing the extent of population differentiation among different sets of populations. The phylogenetic analysis of allele frequency data from many loci will be discussed in chapter 13.

12.4. Genetic Variation for Many Loci

So far, we have considered only one locus, but to evaluate the amount of genetic variation in a population, it is necessary to examine many loci. Ideally, one should examine all genetic loci (say, protein-coding loci), but this is virtually impossible, since the number of loci in the genome is very large. For this reason, genetic variation in a population is usually studied by sampling different loci at random from the genome. In electrophoretic studies of protein polymorphism, the choice of a locus generally depends on the availability of a suitable staining technique for the protein. Because this staining technique has nothing to do with the extent of polymorphism, the genetic loci chosen by this criterion are generally considered to be a random sample (Hubby and Lewontin 1966).

Similarly, a set of microsatellite DNA loci may be considered as a random sample from a large number of such loci in the genome.

Average Gene Diversity (Heterozygosity)

When allele frequencies are studied at many loci, the extent of genetic variation in a population is usually measured by **average gene diversity**, which is often called **average heterozygosity**. Here, the heterozygosity is not the observed frequency of heterozygotes in the population but the expected heterozygosity under the assumption of Hardy-Weinberg equilibrium. The reason why we use the expected rather than the observed heterozygosity is that it depends only on allele frequencies and can be used irrespective of the mating pattern of the population, which is transitory or species specific. It also does not depend on the ploidy of the organism studied and thus can be used even for haploid organisms or for X-linked loci in mammals. For this reason, it is better to call it average gene diversity rather than average heterozygosity. However, since the latter terminology is widespread, we use both terminologies in this book depending on the situation.

The gene diversity at a locus is defined as

$$h = 1 - \sum_{i=1}^{q} x_i^2 \tag{12.31}$$

where x_i is the population frequency of the i-th allele and q is the number of alleles. Average gene diversity is the average of this quantity over all loci.

In reality, we do not know the allele frequencies in the population and need to estimate them by sampling m individuals from the population. In the case of diploid organisms the frequency of allele A_i at a codominant locus is estimated by

$$\hat{x}_i = \hat{X}_{ii} + \sum_{j<i} \hat{X}_{ij}/2 \tag{12.32}$$

where $\hat{X}_{ij}$ is the sample frequency of genotype A_iA_j. The gene diversity at this locus can then be estimated by

$$\hat{h} = 2m\left(1 - \sum \hat{x}_i^2\right)/(2m - 1) \tag{12.33}$$

(Nei and Roychoudhury 1974). Therefore, when L loci are studied, the average gene diversity is estimated by

$$\hat{H} = \sum_{j=1}^{L} \hat{h}_j/L \tag{12.34}$$

where $\hat{h}_j$ is the value of $\hat{h}$ at the j-th locus. The variance of $\hat{H}$ is given by

$$V(\hat{H}) = \sum_{j=1}^{L} (\hat{h}_j - \hat{H})^2/[L(L-1)] \tag{12.35}$$

If $\hat{H}$ is computed for random sets of loci from two unrelated species (e.g., humans and mice) and one wishes to test the difference in $\hat{H}$ between the two species, a Z test should be used with the above variance. However, if the same set of loci is used in both species, we should use the following t test because gene diversity values are usually locus dependent. That is, we should first compute the difference in gene diversity for the i-th locus between populations X and Y by $d_i = \hat{h}_{Xi} - \hat{h}_{Yi}$ for all polymorphic loci, where subscripts X and Y refer to populations X and Y, respectively. We can then compute the mean ($\bar{d}$) and the variance [$V(\bar{d})$] of d_i by the following formulas

$$\bar{d} = \sum_{i=1}^{L'} d_i / L' \tag{12.36}$$

$$V(\bar{d}) = \sum_{i=1}^{L'} (d_i - \bar{d})^2 / [L'(L' - 1)] \tag{12.37}$$

where L' is the number of polymorphic loci. The difference in $\hat{H}$ between the two populations may then be tested by

$$t_{L'-1} = \bar{d}/s(\bar{d}) \tag{12.38}$$

with $L' - 1$ degrees of freedom, where $s(\bar{d})$ is the square root of $V(\bar{d})$.

It should be noted that the h values for closely related populations or species are correlated not only because different loci have different mutation rates or different selection intensities (purifying selection) but also because the same loci have similar h values for historical reasons (see Nei 1987, p. 184). Therefore, we recommend that Equation (12.38) be used as a general test of the differences in average gene diversities. Examples of testing $\bar{d}$ values are given in Nei (1987). (The first printing of Nei's book contained some numerical errors, and these errors are corrected in the paperback edition published in 1989.) The test of $\bar{d}$ can also be done by using the bootstrap standard error of $\bar{d}$.

Expected Value of $\hat{H}$

Infinite-Allele Model

It is useful to know the expected value of H for neutral alleles when the effects of mutation and genetic drift are balanced. In population genetics, it is customary to use two different models of generation of mutant alleles: **infinite-allele model** and **stepwise mutation model**. In the infinite-allele model, every new mutation is assumed to create a new allele (Wright 1939; Kimura and Crow 1964). Under this model, the expected gene diversity is given by

$$E(\hat{H}) = \frac{4Nv}{1 + 4Nv} \tag{12.39}$$

where N and v are the effective population size and the mutation rate per generation, respectively (Kimura and Crow 1964). The infinite-allele model approximately applies to alleles at the nucleotide or amino acid sequence level. It is also believed to apply to protein polymorphism detectable by electrophoresis. Electrophoretic alleles were originally thought to mutate following the stepwise mutation model (Ohta and Kimura 1973), which will be mentioned below. However, empirical studies have suggested that the actual pattern is closer to the infinite-allele model than to the stepwise mutation model (Li 1976; Ramshaw et al. 1979; Fuerst and Ferrell 1980).

Nei and Graur (1984) compared the observed average gene diversity ($\hat{H}$) for protein loci and the expected value [$E(H)$] in many different organisms and showed that $\hat{H}$ is almost always smaller than $E(H)$. They explained this difference by the fact that if N fluctuates over the evolutionary time, the actual effective size is closer to the minimum size (Wright 1978). Another explanation is that the $\hat{H}$ values in current organisms have not reached the equilibrium values because the population size rapidly expanded during the last 10,000 years after the ice age in the Pleistocene and there has not been enough time for new mutations to accumulate.

Stepwise Mutation Model

In recent years, many authors have estimated the average gene diversity for microsatellite DNA loci (e.g., Bowcock et al. 1994; Deka et al. 1995; Jorde et al. 1995). **Microsatellite DNA** loci are segments of repeated DNA with a short repeat length, usually 1–6 nucleotides. For example, an allele for a CA repeat locus may be represented by CACACACACACACA, where the dinucleotide CA is repeated seven times. For this reason, microsatellite loci are also called **short tandem repeat (STR)** loci. Microsatellite loci are believed to be subject to a mutational change following the slippage model of duplication or deletion of repeat units and are usually highly polymorphic. Therefore, there may be alleles with 6, 7, 8, and 9 repeats of CA in the above example. Empirical studies have suggested that alleles are generated roughly according to Ohta and Kimura's (1973) stepwise mutation model. In this model, alleles are represented by the numbers of nucleotide repeats (allele size), and a mutation is assumed to increase or decrease the allele size or allelic state by one, as shown in Figure 12.2. With this model, the expected value of $\overline{H}$ in an equilibrium population is given by

$$E(\overline{H}) = 1 - \frac{1}{\sqrt{1 + 8Nv}} \tag{12.40}$$

(Ohta and Kimura). When Nv is the same, this formula gives a somewhat smaller value than Equation (12.39). Of course, some authors (Shriver et al. 1993; Di Rienzo et al. 1994) have questioned the applicability of this model to real data. The real situation seems to be somewhere between this model and the infinite-allele model.

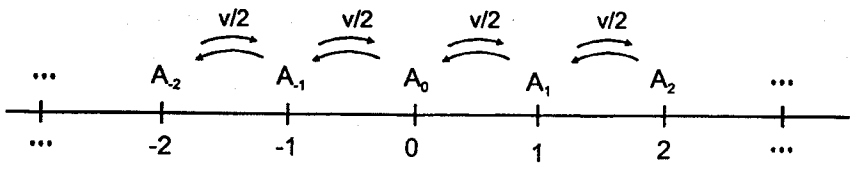


FIGURE 12.2. Stepwise mutation model.

Gene Diversity Analysis in Subdivided Populations

In the preceding section, we discussed the extent of gene diversity within populations. However, natural populations are often subdivided into a number of subpopulations, and in this case, it is useful to study the gene diversities within and between subpopulations. The analysis of gene diversity in the total population into its components can be made by using Nei's (1973) method. In this method, gene diversity is not related to genotype frequencies unless random mating occurs in each subpopulation. In other words, we disregard the distribution of genotype frequencies within subpopulations and consider the decomposition of genomic variation into interpopulational and intrapopulational variation.

The following theory is intended to be applied to the average gene diversity for many loci, but for simplicity we consider a single locus. The results obtained are directly applicable to the average gene diversity. For this reason, we use the notation for average gene diversities rather than the notation for single locus gene diversities.

Consider a population that is divided into s subpopulations. Let x_{ki} be the frequency of the i-th allele in the k-th subpopulation. The gene diversity in this subpopulation is given by

$$H_k = 1 - \sum_i^q x_{ki}^2 \tag{12.41}$$

whereas the average gene diversity within subpopulations is

$$H_S = 1 - \sum_k^s w_k \sum_i^q x_{ki}^2 \tag{12.42}$$

As mentioned earlier, w_k is usually unknown, and it is customary to assume $w_k = 1/s$. By contrast, the gene diversity for the entire population can be written as

$$H_T = 1 - \sum_i^q \bar{x}_{ki}^2 = H_S + D_{ST} \tag{12.43}$$

where $\bar{x}_{ki}$ is the average of x_{ki} over all subpopulations and

$$D_{ST} = \sum_k \sum_l D_{kl}/s^2 \tag{12.44}$$

with $D_{kl} = \Sigma_i (x_{ki} - x_{li})^2/2$ (Nei 1987).

The relative magnitude of gene differentiation among subpopulations may be measured by

$$G_{ST} = D_{ST} / H_T = (H_T - H_S) / H_T \tag{12.45}$$

This G_{ST} corresponds to F_{ST} previously discussed.

Although G_{ST} is a good measure of the degree of relative gene differentiation among subpopulations, it is highly dependent on the value of H_T. When this is small, G_{ST} may be large even if the extent of absolute gene differentiation is small. Note also that D_{ST} includes comparisons of subpopulations with themselves, that is, the cases of $D_{kk} = 0$. To measure the extent of absolute gene differentiation, we should exclude these cases. Nei (1973), therefore, proposed that the following quantity be used as a measure of absolute gene differentiation.

$$\begin{aligned} \bar{D}_m &= \sum_{k \neq l} D_{kl} / [s(s-1)] \\ &= sD_{ST} / (s-1) \end{aligned} \tag{12.46}$$

It should be noted that G_{ST} defined in Equation (12.45) depends on the number of subpopulations used (s) even if H_S and the extent of absolute gene differentiation ($\bar{D}_m$) remain the same. One can remove this dependence by replacing D_{ST} by $\bar{D}_m$ and redefining the total gene diversity by $H'_T = H_S + \bar{D}_m$, as in the case of fixation indices. We then have $G'_{ST} = \bar{D}_m / H'_T$, which is independent of s. However, the simple concept of decomposition of the total gene diversity no longer holds. In practice, if s is greater than 5 there is not much difference between G_{ST} and G'_{ST}. So, as long as $s \geq 5$, we suggest that G_{ST} be used instead of G'_{ST}.

In the above formulation, we have considered the analysis of gene diversity in terms of population gene frequencies. As long as 50 or more individuals are studied for each subpopulation, the theory developed applies to sample allele frequencies as well. However, if the sample size is smaller than this, certain biases may arise in the estimation of gene diversities, as in the case of single random mating populations discussed earlier. Unbiased estimates of H_S, H_T, and D_{ST} may be obtained by applying Equations (12.26) and (12.27) for each locus and taking the averages ($\hat{H}_S$ and $\hat{H}_T$) of $\hat{h}_S$ and $\hat{h}_T$ over all loci as with Equation (12.34). When Hardy-Weinberg equilibrium applies in each subpopulation, $H_S = \hat{H}_0$ may be assumed. In this case,

$$\hat{H}_S = 2m(1 - \overline{\sum \hat{x}_i^2}) / (2m - 1) \tag{12.47}$$

$$\hat{H}_T = 1 - \sum \bar{x}_i^2 + \hat{H}_s / (2ms) \tag{12.48}$$

These estimators generally have smaller sampling variances than those in Equations (12.26) and (12.27).

Once $\hat{H}_S$ and $\hat{H}_T$ are obtained, $\hat{G}_{ST}$ can be estimated by

$$\hat{G}_{ST} = (\hat{H}_T - \hat{H}_S) / \hat{H}_T \tag{12.49}$$

The variances of $\hat{H}_T$, $\hat{H}_S$, and $\hat{G}_{ST}$ can be obtained by the bootstrap or the jackknife method (see the next section). In this case, bootstrap resampling will be conducted among different loci used.

12.5. DNA Polymorphism

For the study of genetic variation in natural populations, DNA sequences are much more informative than protein sequences or electrophoretic variation of proteins, because a large part of DNA sequences is not encoded into proteins, and there is degeneracy of the genetic code. Genetic variation in noncoding regions of DNA (introns, flanking regions, etc.) or silent nucleotide substitutions in the coding regions can be studied only by examining DNA sequences. DNA sequences also reveal detailed information about the polymorphism due to nucleotide substitution, insertion/deletion, gene conversion, unequal crossing over, horizontal gene transfer, and so forth. In the following, we concentrate primarily on nucleotide polymorphism, for which many statistical methods have been developed.

Variation in DNA Sequence

The extent of DNA polymorphism may be measured in several different ways (Nei 1987), but the most commonly used measures are the **number of segregating sites per nucleotide site** and **nucleotide diversity** (or **heterozygosity at the nucleotide level**).

Number of Segregating Sites

We consider a given region (locus) of DNA and assume that m copies (genes) are randomly sampled from a population. If the DNA region consists of n nucleotides, we then have a matrix of n nucleotides by m sequences, as in the case of MHC allelic sequences in Figure 4.1. Any nucleotide site that shows two or more nucleotides among the m sequences is called a **segregating site**. We denote by S the total number of segregating sites in the data set. The number of segregating sites per nucleotide site (p_s) is then given by $p_s = S/n$, where n is the total number of nucleotides examined. S and p_s are clearly sample-size dependent, and as m increases, they increase. Let us now consider the expected value of p_s under the assumption that there is no recombination between any pair of nucleotide sites and new mutations always occur at nonsegregating nucleotide sites. A genetic model satisfying this assumption is called the **infinite-site model**. If we make a further assumption that there is no selection and the population is in mutation-drift balance, the expected value of p_s is given by

$$E(p_s) = a_1\theta \tag{12.50}$$

where $a_1 = 1 + 2^{-1} + 3^{-1} + \cdots + (m-1)^{-1}$ and $\theta = 4N\mu$ (Watterson 1975). Here N and μ are the effective population size and the mutation

rate per site, respectively. Note that the mutation rate per sequence is given by $v = n\mu$. In this case, $4Nv$ is often denoted by M. It is clear that $E(p_s)$ increases as m increases. The theoretical variance of p_s is known to be

$$V(p_s) = E(p_s)/n + a_2\theta^2 \tag{12.51}$$

where $a_2 = 1 + 2^{-2} + 3^{-2} + \cdots + (m-1)^{-2}$. Therefore, the variance of p_s also increases with m.

Here we note that θ is a more basic parameter of genetic variation than p_s, because it is the product of mutation rate and population size and is independent of sample size. This quantity can be estimated by

$$\hat{\theta} = p_s/a_1 \tag{12.52}$$

and the variance of $\hat{\theta}$ is given by

$$V(\hat{\theta}) = V(p_s)/a_1^2 \tag{12.53}$$

However, this equation is valid only when neutral mutations are considered and the population size remains constant throughout the evolutionary time.

Nucleotide Diversity

A measure of DNA polymorphism that does not depend on sample size m is the average number of nucleotide differences per site between two sequences or **nucleotide diversity** (Nei and Li 1979). This is defined by

$$\pi = \sum_{ij}^{q} x_i x_j d_{ij} \tag{12.54}$$

where q is the total number of alleles (different allelic sequences), x_i is the population frequency of the i-th allele, and d_{ij} is the number of nucleotide differences or substitutions per site between the i-th and j-th alleles. In a random mating population, π is simply heterozygosity at the nucleotide level. It can be estimated by

$$\hat{\pi} = \frac{q}{q-1}\sum_{ij} \hat{x}_i \hat{x}_j d_{ij} \tag{12.55}$$

or by

$$\hat{\pi} = \sum_{i<j}^{m} d_{ij}/c \tag{12.56}$$

where m, $\hat{x}_i$, and c are the total number of DNA sequences examined, the frequency of the i-th allele in the sample, and the total number of sequence comparisons $[m(m-1)/2]$, respectively. In Equation (12.56), i and j refer to the i-th and j-th sequences rather than to the i-th and j-th al-

leles. In other words, all DNA sequences are labeled differently in this equation, whether they are identical alleles or not. In this definition, d_{ij} is 0 whenever sequences i and j are identical, but this equation is more convenient than Equation (12.55) for numerical computation. In practice, d_{ij} may be estimated by $\hat{d}_{ij} = -b\ln(1 - \hat{p}_{ij}/b)$, where $b = 3/4$ and $\hat{p}_{ij}$ is the proportion of different nucleotides. When $\hat{p}_{ij}$ is small, however, $\hat{d}_{ij}$ is approximately equal to $\hat{p}_{ij}$, so that one may use $\hat{p}_{ij}$ instead of $\hat{d}_{ij}$.

The variance of $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \left[\sum_{i\neq j}^{m} V(d_{ij}) + \sum_{i\neq k}^{m}\sum_{k\neq l}^{m} Cov(d_{ij}, d_{kl})\right] / c^2 \qquad (12.57)$$

where $V(d_{ij})$ is the variance of d_{ij} and is given by

$$V(d_{ij}) = \frac{p_{ij}(1 - p_{ij})}{n(1 - p_{ij}/b)^2} \qquad (12.58)$$

for the Jukes-Cantor model, and $Cov(d_{ij}, d_{kl})$ is the covariance of d_{ij} and d_{kl} and is approximately given by

$$Cov(d_{ij}, d_{kl}) = \frac{p_{ij\cdot kl} - p_{ij}p_{kl}}{n(1 - p_{ij}/b)(1 - p_{kl}/b)} \qquad (12.59)$$

where $p_{ij\cdot kl}$ is the proportion of sites at which sequence i differs from sequence j and sequence k differs from sequence l, and p_{ij} is the proportion of sites at which sequence i differs from sequence j (Bulmer 1991; Rzhetsky and Nei 1992a). When the p distance is used, $V(p_{ij})$ and $Cov(p_{ij}, d_{kl})$ are given by

$$V(p_{ij}) = \frac{p_{ij}(1 - p_{ij})}{n}, \qquad Cov(p_{ij}, p_{kl}) = \frac{p_{ij\cdot kl} - p_{ij}p_{kl}}{n} \qquad (12.60)$$

When the number of sequences (m) is 50 or more, the above computation can be slow. In this case $V(\hat{\pi})$ may be computed by the jackknife or the bootstrap method. In the jackknife method, we first compute $\hat{\pi}$ from the original data set using Equation (12.55) or (12.56). We then eliminate data for the first nucleotide site and again compute $\hat{\pi}$ from this new set of data. We denote this $\hat{\pi}$ by $\hat{\pi}_1$. We do the same computation for the second, third, . . . , and n-th nucleotide sites and denote them by $\hat{\pi}_2$, $\hat{\pi}_3$, . . . , and $\hat{\pi}_m$, respectively, where n is the total number of nucleotides. Once all $\hat{\pi}_i$'s are computed, the variance of $\hat{\pi}$ is given by

$$V(\hat{\pi}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - \bar{\pi})^2 \qquad (12.61)$$

The bootstrap variance of $\hat{\pi}$ is computed by the method described in chapter 2. Theoretically, the bootstrap variance seems to be more accurate than the jackknife variance (Efron and Tibshirani 1993).

The mean and variance of $\hat{\pi}$ discussed above are empirical ones, and therefore they do not assume any particular model of population dy-

namics. However, if the infinite-site model with neutral mutations applies and the effects of mutation and drift are in equilibrium, the theoretical mean $[E(\hat{\pi})]$ and variance $[V'(\hat{\pi})]$ of $\hat{\pi}$ under the stochastic process become

$$E(\hat{\pi}) = 4N\mu = \theta \tag{12.62}$$

(Haldane 1939; Kimura 1968; Watterson 1975) and

$$V'(\hat{\pi}) = \frac{m+1}{3(m-1)n}\theta + \frac{2(m^2+m+3)}{9m(m-1)}\theta^2 \tag{12.63}$$

(Tajima 1983; Nei 1987).

The above equation gives the variance of $\hat{\pi}$ among different populations, which are evolving independently. Interestingly, this variance becomes $(2/9)\,\theta^2$ as n and m increase to infinity. By contrast, the variance of $V(\hat{\pi})$ in Equation (12.57) is the within-population variance that is caused by the fact that only m sequences with n nucleotides are examined. Therefore, $V(\hat{\pi})$ becomes 0 as $n \to \infty$. In general, $V'(\hat{\pi})$ is much greater than $V(\hat{\pi})$, because the former represents the variance due to the stochastic change of π in long-term evolution. $V(\hat{\pi})$ is the proper variance when one wants to know the standard error of the observed value of π in a population. However, if we consider π as an estimator of Θ, then $V'(\hat{\pi})$ is the appropriate variance.

Note that the variance of θ obtained from segregating nucleotide sites (Equation [12.53]) is smaller than $V'(\hat{\pi})$. This indicates that the number of segregating sites is more informative than nucleotide diversity in estimating θ. In reality, however, the population size fluctuates extensively when long-term evolution is considered, and p_s is more sensitive to population size change than $\hat{\pi}$. Therefore, both p_s and $\hat{\pi}$ have advantages and disadvantages. Fortunately, empirical studies have shown that p_s and $\hat{\pi}$ generally gives similar estimates of θ. For estimating θ, Felsenstein (1992) developed a phylogenetic approach, which is more efficient than the above two methods (see also Fu 1994). However, this method also depends on the assumption of constant population size.

Example 12.1. Polymorphism of Mitochondrial DNA in Human Populations

Figure 12.3 shows the phylogenetic tree (NJ tree) for 70 human mtDNA sequences sampled from sub-Saharan Africans, Europeans, Asians, and New Guineans (a subset of the data obtained by Vigilant et al. 1991). This tree was obtained from a 663-bp hypervariable segment of the mtDNA control region. Because this is a gene tree, sequences from different continents are intermingled, though the African sequences tend to form distinct clusters. The numbers of sequences from Africans, Europeans, Asians, and New Guineans are 30, 15, 20, and 5, respectively. Table 12.2 shows the values of a_1, a_2, m, S, $\hat{\theta}$, $s(\hat{\theta}) = \sqrt{V(\hat{\theta})}$, $\hat{\pi}$, $s(\hat{\pi}) = \sqrt{V(\hat{\pi})}$, and $s'(\hat{\pi}) = \sqrt{V'(\hat{\pi})}$ for the four human populations. The standard error $s(\hat{\theta})$ was computed by using Equation (12.53), and the standard error $s(\hat{\pi})$ was computed by the bootstrap.

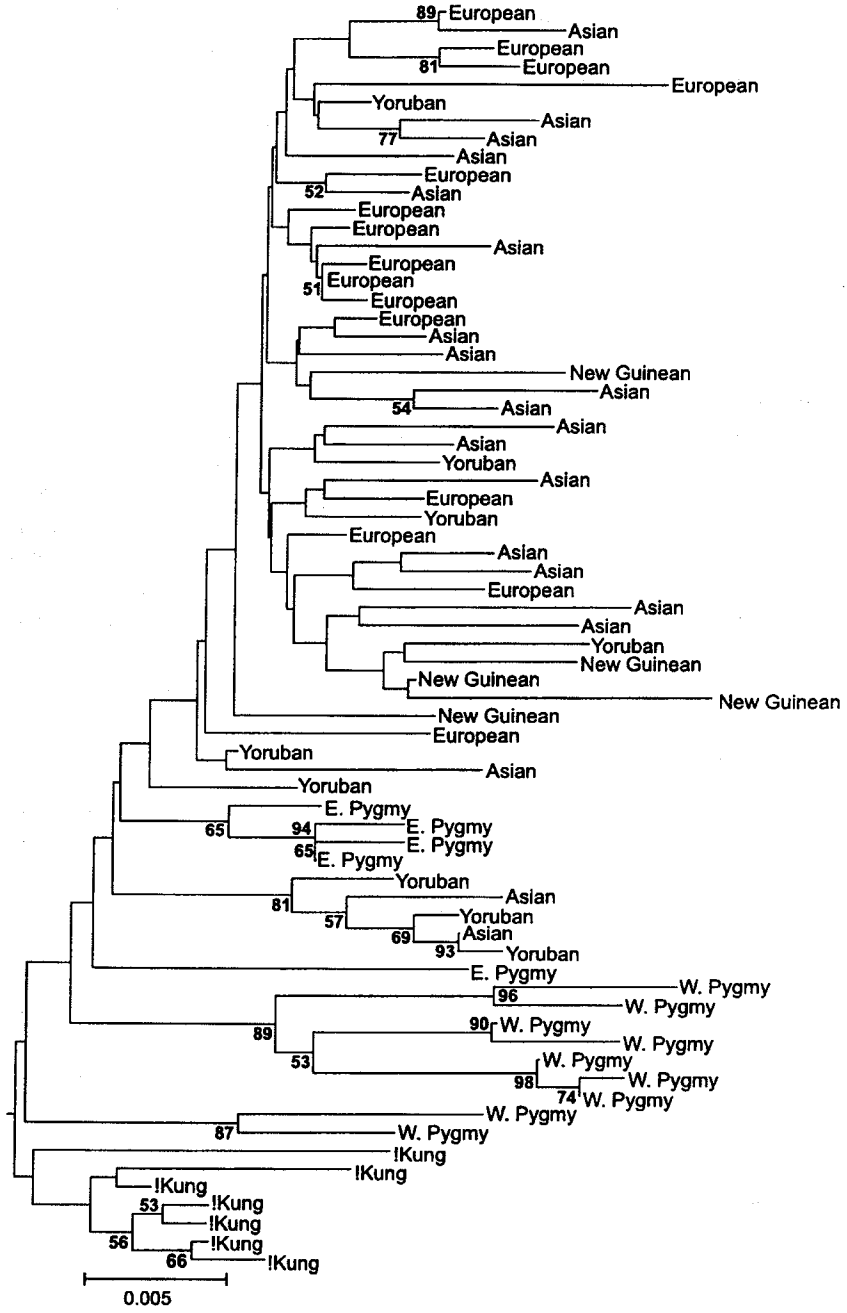


FIGURE 12.3. The neighbor-joining tree obtained by using p-distances for 70 human mtDNA sequences from sub-Saharan Africa, Europe, Asia, and New Guinea. The root of the tree was determined by using chimpanzees as the outgroup. Bootstrap values are in boldface. The bootstrap values lower than 50% are not shown.

Table 12.2 Values of a_1, a_2, m, S, $\hat{\theta} \pm s(\hat{\theta})$, $\hat{\pi} \pm s(\hat{\pi})$, and $s'(\hat{\pi})$ (multiplied by 100) for the African, Asian, European, and New Guinean populations.

Population	a_1	a_2	m	S	$\hat{\theta}$	$\hat{\pi}$	$s'(\hat{\pi})$
Africans	3.96	1.61	30	73	2.9 ± 1.0	2.4 ± 0.4	1.4
Asians	3.55	1.59	20	74	3.3 ± 1.2	1.9 ± 0.3	1.6
Europeans	3.25	1.58	15	38	1.8 ± 0.7	1.1 ± 0.2	0.9
N. Guineans	2.08	1.42	5	25	1.9 ± 1.1	1.8 ± 0.3	1.2

Note: A 633 bp segment of the D-loop region of mitochondrial DNA was used.

As expected from Figure 12.3, Africans have the highest nucleotide diversity ($\hat{\pi} = 0.024$), but it is not significantly different from that of Asians (0.019) or New Guineans (0.018). The Z statistic for testing the difference in $\hat{\pi}$ between Africans (0.024) and Europeans (0.011) is 3.2, so the difference is significant at the 1% level. This lower nucleotide diversity in Europeans suggests that Europeans are relatively recently derived populations. Table 12.2 also includes the theoretical standard error $[s'(\hat{\pi})]$ of $\hat{\pi}$ obtained under the assumption that the population is in mutation-drift equilibrium and the DNA sample is drawn randomly from a single random-mating population (Equation [12.63]). This standard error is three to six times higher than $s(\hat{\pi})$. This large difference between $s(\hat{\pi})$ and $s'(\hat{\pi})$ is caused by the fact that $s'(\hat{\pi})$ includes the stochastic error caused by mutation and drift during the entire evolutionary process and this stochastic error cannot be reduced to 0 even if the sample size (m) and the number of nucleotides examined (n) are infinitely large (Tajima 1983). By contrast, $s(\hat{\pi})$ approaches 0 as m becomes infinite. In practice, however, the assumptions required for $s'(\hat{\pi})$ are rarely satisfied, and since we are usually interested in the sampling error of $\hat{\pi}$ for the sequences actually studied, we recommend that $s(\hat{\pi})$ be used for testing the difference between two $\hat{\pi}$ values.

As mentioned above, the expectations of $\hat{\pi}$ and $\hat{\theta}$ are both equal to $4N\mu$ if the population is a single random mating population and is in mutation-drift equilibrium. The $\hat{\theta}$ values in Table 12.2, however, tend to be greater than $\hat{\pi}$, particularly in Asians and Europeans. This suggests that the assumptions required for the estimation of θ are not satisfied with the present data set. In fact, the DNA sequences used in Vigilant et al.'s (1991) study were obtained from diverse geographical areas for each population and do not satisfy the assumption of a random sample from a random mating population. Nevertheless, $\hat{\theta}$ and $\hat{\pi}$ are highly correlated with each other, as expected.

DNA Divergence Between Populations

In chapter 3, we discussed methods of estimating the number of nucleotide substitutions between two DNA sequences obtained from different species, ignoring the effect of polymorphism. To estimate the extent of DNA divergence between two populations or two closely related species, however, we must consider the effect of polymorphism. This can be done in the following way.

Nucleotide Differences Between Populations

Suppose that there are q different alleles at a particular DNA region (locus) and m_X and m_Y sequences have been sampled from populations X and Y, respectively. Let $\hat{x}_i$ and $\hat{y}_i$ be the sample frequencies of the i-th allele for populations X and Y, respectively. The average number of nucleotide substitutions for a randomly chosen pair of alleles in population X (d_X; nucleotide diversity) can be estimated by

$$\hat{d}_X = \frac{q}{q-1} \sum_{ij}^{q} \hat{x}_i \hat{x}_j \hat{d}_{ij} \tag{12.64}$$

where $\hat{d}_{ij}$ is an estimate of the number of nucleotide substitutions per site between the i-th and j-th alleles. When all DNA sequences are different, $\hat{x}_i = 1/m_X$. The average number of nucleotide substitutions (d_Y) for Y can be estimated in the same way. Note that $\hat{d}_X$ and $\hat{d}_Y$ can also be obtained by Equation (12.56). The average number (d_{XY}) of nucleotide substitutions per site between alleles from X and Y can be estimated by

$$\hat{d}_{XY} = \sum_{ij}^{q} \hat{x}_i \hat{y}_{ij} \hat{d}_{ij} \tag{12.65}$$

where $\hat{d}_{ij}$ is an estimate of the nucleotide substitutions between the i-th allele from X and the j-th allele from Y. This can also be written as

$$\hat{d}_{XY} = \sum_{ij} d_{ij} / (m_X m_Y) \tag{12.66}$$

where $\hat{d}_{ij}$ is an estimate of the number of nucleotide substitutions between the i-th sequence (not allele) sampled from population X and the j-th sequence from population Y. Note that there are m_X sequences sampled from X and m_Y sequences from Y. The number of net nucleotide substitutions between the two populations (d_A) is then estimated by

$$\hat{d}_A = \hat{d}_{XY} - (\hat{d}_X + \hat{d}_Y)/2 \tag{12.67}$$

(Nei and Li 1979).

If the populations are in equilibrium with respect to the effects of mutation and genetic drift throughout the evolutionary process and the time since divergence between the two populations is T, then the expected value [$E(d_{XY})$] of $\hat{d}_{XY}$ is given by

$$E(d_{XY}) = \theta + 2rT \tag{12.68}$$

where r ($= \mu$) is the mutation rate per nucleotide site (= substitution rate) (chapter 3). Under the present condition, $E(\hat{d}_X) = E(\hat{d}_Y) = \theta$, so that we have

$$E(d_A) = 2rT \tag{12.69}$$

The variance of $\hat{d}_{XY}$ is given by

$$V(\hat{d}_{XY}) = \left[\sum_{ij} V(d_{ij}) + \sum_{ij}\sum_{kl} Cov(d_{ij}, d_{kl})\right] / (m_X m_Y)^2 \quad (12.70)$$

where $V(d_{ij})$ and $Cov(d_{ij}, d_{kl})$ can be obtained by Equations (12.58) and (12.59), respectively. On the other hand, the variance of $\hat{d}_A$ is

$$V(\hat{d}_A) = V(\hat{d}_{XY}) + \frac{1}{4}[V(\hat{d}_X) + V(\hat{d}_Y)] - [Cov(\hat{d}_{XY}, \hat{d}_X) + Cov(\hat{d}_{XY}, \hat{d}_Y)] \quad (12.71)$$

All the variances and covariances of $\hat{d}_{XY}$, $\hat{d}_X$, and $\hat{d}_Y$ on the right-hand side of this equation can be computed by Equations (12.58) and (12.59).

When the number of sequences examined is large, the computation of $V(\hat{d}_{XY})$ and $V(\hat{d}_A)$ is time-consuming. In this case, one may use the jackknife or the bootstrap method to compute $V(\hat{d}_{XY})$ and $V(\hat{d}_A)$.

d_A in Equation (12.69) is useful for estimating evolutionary time T when r is known or for computing the substitution rate when T is known. It is also useful for constructing phylogenetic trees when there are many populations. However, the variance of $\hat{d}_A$ is usually greater than that of $\hat{d}_{XY}$ in Equation (12.65). For this reason, $\hat{d}_{XY}$ often gives more reliable trees than $\hat{d}_A$ (Nei 1987).

Coefficient of Nucleotide Differentiation

When there are many populations, it is often useful to have a measure of population differentiation similar to F_{ST} or G_{ST} for gene frequency data. We again assume that there are s subpopulations and the relative size of the k-th subpopulation is w_k. (In practice, we assume $w_k = 1/s$ as before.) The average nucleotide diversity (π_S) within populations can then be estimated by

$$\hat{\pi}_S = \sum_{k=1}^{s} w_k \hat{\pi}_k \quad (12.72)$$

where $\hat{\pi}_k$ is the estimate of π in the k-th subpopulation (Equation [12.55]). On the other hand, one can estimate the nucleotide diversity for the entire population disregarding population structure. In this case, we denote by $\bar{x}_i$ the estimate of average frequency of the i-th allele in the entire population. The nucleotide diversity for the entire population (π_T) is then estimated by

$$\hat{\pi}_T = \frac{q}{q-1}\sum^{q} \bar{x}_i \bar{x}_j \hat{d}_{ij} \quad (12.73)$$

where q is the total number of alleles examined. The estimate of interpopulational nucleotide diversity (δ_{ST}) is given by

$$\hat{\delta}_{ST} = \hat{\pi}_T - \hat{\pi}_S \quad (12.74)$$

whereas the estimate of the proportion of interpopulational diversity (N_{ST}) is

Table 12.3 Values of $\hat{d}_X$, $\hat{d}_{XY}$, $\hat{d}_A$, and $\hat{N}_{ST}$ for the African, Asian, European, and New Guinean populations.

	Africans	Asians	Europeans	New Guineans
Africans	2.39 ± 0.35	2.52 ± 0.32	2.28 ± 0.33	2.53 ± 0.30
Asians	0.35 ± 0.10	1.94 ± 0.23	1.56 ± 0.20	1.97 ± 0.27
Europeans	0.52 ± 0.15	0.04 ± 0.03	1.12 ± 0.19	1.65 ± 0.26
N. Guineans	0.44 ± 0.13	0.11 ± 0.06	0.19 ± 0.08	1.79 ± 0.34
$\hat{\pi}_S = 1.81 \pm 0.02$		$\hat{\pi}_T = 2.17 \pm 0.02$		$\hat{N}_{ST} = 16.7 \pm 3.7$

Note: A 633 bp segment of D-loop region of mitochondrial DNA was used. All values are multiplied by 100. Standard errors were computed by the bootstrap method. $\hat{d}_X$'s are on the diagonal; d_{XY}'s are above the diagonal; $\hat{d}_A$'s are below the diagonal.

$$\hat{N}_{ST} = \hat{\delta}_{ST} / \hat{\pi}_T \tag{12.75}$$

(Nei 1982; Takahata and Palumbi 1985; Pannell and Charlesworth 1999). Obviously, π_S, π_T, δ_{ST}, and N_{ST} correspond to H_S, H_T, D_{ST}, and G_{ST} in gene diversity analysis, and we call N_{ST} the **coefficient of nucleotide differentiation**. In actual computation, we recommend $w_k = 1/s$ for the reason mentioned earlier. Lynch and Crease (1990) and Pons and Petit (1996) developed methods for computing the variances of $\hat{\pi}_S$, $\hat{\pi}_T$, and $\hat{N}_{ST}$, considering the processes of population and allelic sampling. Their formulas are somewhat complicated. We suggest that the variances of these quantities be computed by the jackknife or the bootstrap method.

Example 12.2. Nucleotide Diversity Within and Between Major Groups of Human Populations

Previously, we computed nucleotide diversity $\hat{\pi}$ for mtDNA sequences from four major human populations. Let us now compute $\hat{d}_{XY}$'s and $\hat{d}_A$'s for the four populations. They are presented in Table 12.3. The standard errors of $\hat{d}_X$'s, $\hat{d}_{XY}$'s, and $\hat{d}_A$'s were obtained by the bootstrap method. As expected from fig. 12.3. $\hat{d}_{XY}$'s and $\hat{d}_A$'s are higher for the comparisons between Africans with non-Africans than for the comparisons of non-African populations, supporting the idea that modern humans originated in Africa (Cann et al. 1987; Wilson and Cann 1992). Table 12.3 also includes $\hat{\pi}$, $\hat{\pi}_T$, and $\hat{N}_{ST}$. $\hat{N}_{ST}$ indicates that the extent of nucleotide differentiation among the major groups of humans is slightly higher than that (0.1) of classical genetic markers (Nei 1982). However, note that the standard error of $\hat{N}_{ST}$ is quite high.

In Table 12.3, we have also presented the standard errors of $\hat{d}_X$'s (or $\hat{\pi}_S$'s) obtained by the bootstrap for the four human populations. Comparison of these values with those obtained by Equation (12.57) (Table 12.2) indicates that the two methods give similar results.

12.6. Statistical Tests for Detecting Selection

Natural selection is obviously an important factor in the formation of a new species. Since the genetic variation of all morphological or physio-

logical characters is ultimately controlled by variation at the protein or DNA level, it is important to identify the changes in protein or DNA sequences that are caused by natural selection. Furthermore, because the basic process of evolution by natural selection is the replacement of one allele by another with a higher fitness in a population, many evolutionists are interested in obtaining direct proof that this process actually occurs in nature.

In practice, however, this is not a simple problem for a number of reasons (see Lewontin 1974; Nei 1987; Kreitman and Akashi 1995; Hartl and Clark 1997). First, in many higher organisms, the generation time is too long to determine whether an observed change in allele frequency is due to selection or genetic drift. Second, since all genes are linked with some other genes in a chromosome, it is difficult to isolate the effect of allelic substitution at one locus from that of other loci. Third, when different genes interact with one another, we may have to examine the effects of allelic substitutions at two or more loci simultaneously, and this often poses technical difficulty. Fourth, the environmental factors that affect selection coefficients or population size never stay constant in nature, and these factors make it difficult to interpret the results of studies of allele frequency changes in nature.

Fortunately, the neutral theory of molecular evolution is capable of making a number of predictions about the relationships among the number of polymorphic alleles per locus, heterozygosity, extent of genetic divergence between different populations, and so on (e.g., Kimura 1983; Nei 1987; Gillespie 1991; Li 1997). Therefore, it is possible to examine the internal consistency of various quantities predicted under the neutral theory. For example, in the case of protein polymorphism, it is possible to compute the theoretical distribution of single-locus heterozygosity (h) if we know the average heterozygosity (H) for many loci and then compare this distribution with the observed distribution of single-locus heterozygosity ($\hat{h}$) (Nei et al. 1976b). If the agreement between the two distributions is satisfactory, one cannot reject the neutral theory. Another example is Watterson's (1977) test (see also Ewens 1972), in which the observed heterozygosity for a single-locus heterozygosity is compared with the theoretical heterozygosity predicted from the number of alleles observed.

These tests have been applied to protein polymorphism data obtained from hundreds of different species, as was summarized by Kimura (1983), Nei and Graur (1984), and others. Most of these tests could not reject the null hypothesis of neutral evolution, although there were some notable exceptions (e.g., Hedrick and Thomson 1983). This is partly due to the relatively low power of the tests used, but these studies do suggest that the statistical properties of protein polymorphism are in rough agreement with those expected from the neutral theory when temporal fluctuation of effective population size is taken into account (Nei 1987).

However, this does not mean that the neutral theory has been shown to be correct. Ohta's (1973, 1992) theory of slightly deleterious mutations is also capable of explaining several aspects of protein polymorphism. In fact, her theory has several appealing features (Kreitman and Akashi 1995). However, it should be noted that if slightly deleterious mutations

are abundant and are fixed in the population continuously, the genes will gradually deteriorate and eventually lose their original function. This is a serious problem for the theory of slightly deleterious mutations.

Recent studies of this problem at the DNA level have complicated the issue, because we now have to deal with synonymous and nonsynonymous nucleotide substitutions and codon usage bias as well as coding and noncoding regions. However, it is not within the scope of this book to discuss these new biological findings on the maintenance of DNA polymorphism. The reader who is interested in this issue may consult Brookfield and Sharp (1994), Kreitman and Akashi (1995), Hartl and Clark (1997), and Li (1997).

Here we present two statistical methods that are often used for testing the neutrality of nucleotide variation and discuss problems associated with them. There are several other statistical methods for testing neutrality, but they are recently reviewed in detail by Li (1997), so there seems to be no need to include all of them here. It should be mentioned at this stage that all of these tests are for detecting any deviations from strict neutrality. Since Kimura's neutral theory does not assert that all alleles are neutral, detection of selection by the following tests does not necessarily reject the neutral theory. Only when a majority of amino acid and nucleotide substitutions are shown to be positively selected, is the theory rejected.

Tests Based on DNA Polymorphism

Tajima Test and Its Extensions

Tajima's (1989b) test is conceptually similar to Watterson's (1977) homozygosity test for protein polymorphism. It examines the relationship between the number of segregating sites and nucleotide diversity. Suppose that a random sample of m sequences with n nucleotides is chosen at a locus in a random mating population. We can then compute the number of segregating sites per site (p_s) and nucleotide diversity ($\hat{\pi}$) for this data set. We have shown that if all alleles are selectively neutral and the population is in mutation-drift equilibrium, $\theta = 4N\mu$ can be estimated either by p_s/a_1 or by $\hat{\pi}$. However, if there are many deleterious alleles in the population, we would expect that p_s/a_1 is inflated, whereas $\hat{\pi}$ is not seriously affected, because the latter quantity is determined largely by high-frequency alleles. By contrast, if there is balancing selection that increases allele frequencies, we expect that $\hat{\pi}$ is enhanced, whereas p_s/a_1 is not. Therefore, by testing the difference between p_s/a_1 and $\hat{\pi}$, we may be able to detect the presence of deleterious mutations or balancing selection.

In practice, the number of segregating sites per sequence ($S = np_s$) and nucleotide diversity (heterozygosity) per sequence $\hat{k} = n\hat{\pi}$ instead of p_s/a_1 and $\hat{\pi}$ are used for this purpose because of mathematical convenience. Therefore, the following test statistic is used.

$$D = \frac{\hat{k} - S/a_1}{[V(\hat{k} - S/a_1)]^{1/2}} \tag{12.76}$$

Here the variance of $\hat{k} - S/a$ is given by

$$V(\hat{k} - S/a_1) = e_1 S + e_2 S(S-1) \tag{12.77}$$

where

$$e_1 = (a_1 b_1 - 1)/a_1^2 \tag{12.78}$$

$$e_2 = \frac{b_2 - (m+2)/(a_1 m) + a_2/a_1^2}{a_1^2 + a_2} \tag{12.79}$$

with a_1 and a_2 defined in Equations (12.50) and (12.51), respectively, and

$$b_1 = \frac{m+1}{3(m-1)} \tag{12.80}$$

$$b_2 = \frac{2(m^2 + m + 3)}{9m(m-1)} \tag{12.81}$$

The mean and variance of D are approximately 0 and 1, but the distribution of D is quite different from the normal distribution. Using a computer simulation, Tajima (1989b) showed that the distribution of D is close to a beta distribution and suggested that the beta distribution be used for computing the significance level. However, because the beta distribution is a crude approximation, Simonsen et al. (1995) suggested that the significance level be determined by computer simulation in which a set of allelic sequences is generated according to the coalescent theory (Kingman 1982; Hudson 1983; Tajima 1983). Computer simulation has shown that Simonsen et al.'s approach is more powerful than Tajima's approximate approach.

Fu and Li (1993) proposed a slightly different approach to test neutral evolution. They considered the case where outgroup sequences are available, so that a rooted phylogenetic tree can be constructed for a given set of allelic sequences. If we have such a tree, it is possible to count the number of mutations for each interior or exterior branch. Let y_i and y_e be the total numbers of mutations for all interior branches and for all exterior branches, respectively, and $y = y_i + y_e$. It can be shown that the expected values of y and y_e are $E(y) = a_1 M$ and $E(y_e) = M$, respectively, where $M = 4Nn\mu$. Therefore, we can test the null hypothesis of neutral evolution by examining the statistical significance of $D = (y - a_1 y_e)/[V(y - a_1 y_e)]^{1/2}$. The distribution of D is again complicated, and Fu and Li suggest that the test should be done by computing the significance level with computer simulation.

Fu and Li (1993) developed two more test statistics, but these tests seem to be less powerful than Tajima's test (Simonsen et al. 1995). Simonsen et al. also developed a series of other test statistics, but they concluded that these tests are less powerful than Tajima's.

As mentioned earlier, these tests of neutrality depend on the assumption that the population has been in mutation-drift balance for a long evolutionary time. This assumption is unlikely to hold in most natural pop-

ulations. Tajima (1989a) has shown that if a population goes through a bottleneck, D may become significantly positive or significantly negative depending on the population history. Therefore, a careless use of the test may lead to erroneous conclusions. Another assumption of these tests is that all nucleotides are equally mutable and are subject to the same population dynamics. This assumption usually does not hold when the above tests are applied to coding regions of DNA, because the extent of polymorphism is not the same for the first, second, and third codon positions, and codon usage biases further complicate the mutation pattern. Therefore, although the mathematical theory is correct under the assumptions made, one should always be cautious about the interpretation of results obtained by the test. If this test suggests any positive selection, it is advisable to confirm it with some experimental work. The same comments apply to all other methods related to Tajima's test.

Despite these complications, many authors have applied the Tajima test to DNA polymorphism data (e.g., Schaeffer and Miller 1992; Brookfield and Sharp 1994; Hilton and Hey 1996; Moriyama and Powell 1996). The general conclusion from these studies is that the null hypothesis of neutral evolution cannot be easily rejected, although purifying selection is sometimes detected (Wise et al. 1998). This conclusion was reached partly because DNA polymorphisms are largely synonymous and partly because the Tajima test is not very powerful.

The HKA and the McDonald-Kreitman Tests

In the neutral theory of molecular evolution, the long-term change of genes in the evolutionary process and genetic polymorphism in current natural populations are two different aspects of the same evolutionary process, as mentioned earlier. In this theory, both the evolutionary rate and the extent of polymorphism of a gene are primarily determined by the rate of neutral mutations. Therefore, if there are two genetic loci that evolve with different rates, the locus with a higher rate is expected to show a higher degree of polymorphism than the other. Hudson, Kreitman, and Aguade (1987) used this idea to develop a statistical method of testing the neutral evolution of DNA sequences. In this test, which is known as the **HKA test**, the extent of genetic divergence between species and the degree of polymorphism are measured in terms of the number of variable or polymorphic sites. Since the expected numbers of variable sites per sequence within and between populations can be expressed as a function of mutation rate per sequence (v), sample size, and the time of divergence between the two species (T), it is possible to test the neutral theory by examining the consistency of sequence divergence and polymorphism. However, this test is not used very often, partly because it requires an extensive survey of DNA sequences and partly because the required assumption that the effective population size remains the same for the entire evolutionary process is often violated. For some of the results obtained by this test, see Hudson et al. (1987), Moriyama and Powell (1996), and Wells (1996).

McDonald and Kreitman (1991) developed a simpler method of testing neutral evolution, considering the relationships of synonymous and non-

synonymous nucleotide differences between and within populations. The basic idea of this test is the same as that of the HKA test, but this test examines whether the ratio of nonsynonymous to synonymous nucleotide differences between species is the same as that within populations.

Consider a protein coding region of DNA and assume that m_X and m_Y sequences of this region are obtained from two closely related species, X and Y. We then examine whether each nucleotide site is variable or not and disregard all invariable sites. If a nucleotide site is variable and has a nucleotide (say, A) for all sequences from species X, but another nucleotide (say, C) for all sequences from species Y, this site is called a **fixed site.** All other variable sites are called **polymorphic sites.** These polymorphic sites are polymorphic either in one of the two species or in both species. Both fixed sites and polymorphic sites are now divided into two groups: sites in which the nucleotide differences are synonymous and sites in which the differences are nonsynonymous. Let s_F, n_F, s_P, and n_P be the numbers of sites with fixed synonymous, fixed nonsynonymous, polymorphic synonymous, and polymorphic nonsynonymous differences, respectively. The null hypothesis of the McDonald-Kreitman test is $E(n_F)/E(s_F) = E(n_P)/E(s_P)$, where E stands for an expectation operator. This null hypothesis can be tested by using a 2×2 contingency table (Table 12.4). If we assume that n_P/s_P reflects the ratio of the rates of synonymous and nonsynonymous mutation, an n_F/s_F ratio that is significantly higher than n_P/s_P suggests that some of the nonsynonymous nucleotide substitutions between the two species are caused by positive selection. By contrast, if n_F/s_F is significantly lower than n_P/s_P, it suggests that purifying selection has operated to reduce the number of nonsynonymous substitutions between the species. This reasoning depends on the assumption that the rate of synonymous mutation remains the same during the entire evolutionary process. This test is applicable even for three or four species if they are closely related.

McDonald and Kreitman (1991) applied this test to DNA sequences of the coding region of the alcohol dehydrogenase (*Adh*) gene from *Drosophila melanogaster, D. simulans,* and *D. yakuba.* The s_F, n_F, s_P, and n_P values obtained are presented in Table 12.4. Fisher's exact test indi-

Table 12.4 2×2 contingency table for the numbers of synonymous and nonsynonymous nucleotide differences for fixed and polymorphic sites. The numbers in the parentheses are results from the Adh gene when *D. melanogaster* (12), *D. simulans* (6), and *D. yakuba* (24) sequences were compared (see text).

	Sites		
Differences	Fixed	Polymorphic	Total
Synonymous	s_F (17)	s_P (42)	$s_F + s_P$ (59)
Nonsynonymous	n_F (7)	n_P (2)	$n_F + n_P$ (9)
Sum	$s_F + n_F$ (24)	$s_P + n_P$ (44)	68

cates that the ratio n_F/s_F (= 7/17) is significantly higher than the ratio n_P/s_P (= 2/42) (P = 0.006). Therefore, this result suggests that nonsynonymous nucleotide substitutions between species have been enhanced by positive selection. Similar results were obtained by Eanes et al. (1993) for the G6pd gene in *D. melanogaster* and *D. simulans* (but see Eanes et al. 1996).

However, the interpretation of results of this test is not as straightforward as it looks. Whittam and Nei (1991) indicated that McDonald and Kreitman's test does not take into account multiple substitutions at the same sites and that if these substitutions are taken into account, their original data set does not reject neutrality. Akashi (1995) pointed out that if there is strong codon usuage bias, n_F/s_F may become greater than n_P/s_P. For example, in the case of the *Adh* gene in the *D. melanogaster* group, codon *GCC* is used most often among the codons encoding alanine. Polymorphism data indicate that this codon mutates to other codons but the mutant codons do not stay for a long time because of purifying selection. This suggests that synonymous changes have a relatively low probability of being fixed in the population. Even if they are fixed, they may mutate back to the preferred codon. By contrast, neutral or nearly neutral amino acid mutations are apparently rare compared with synonymous mutations, but once they are fixed in the population, they do not easily mutate back to the original amino acid. For this reason, the n_F/s_F ratio may become higher than n_P/s_P. In fact, *D. melanogaster, D. simulans,* and *D. yakuba* have a highly biased codon usage, the *GC* content at the third codon position being about 82%. Therefore, we cannot ignore the effect of codon usage bias in McDonald and Kreitman's (1991) data analysis.

A higher n_F/s_F ratio than n_P/s_P may also occur when purifying selection is relaxed for some reason in the early stage of gene evolution but is again intensified in recent times. This situation is likely to occur when a new gene is produced by gene duplication (Li and Gojobori 1983). Ohta (1993) also indicated that n_F/s_F may become higher than n_P/s_P when the population size change occurs in the presence of slightly deleterious mutations. As suggested by Eyre-Walker (1997), a temporal change of codon usage due to mutational bias may also disturb the null hypothesis of the McDonald-Kreitman test.

It is therefore important to examine these factors when the McDonald-Kreitman test is used. In this test, the null hypothesis used is $E(n_F)/E(s_F) = E(n_P)/E(s_P)$ as mentioned earlier, whereas in the $\hat{d}_N - \hat{d}_S$ (or $\hat{b}_N - \hat{b}_S$) test mentioned in chapters 4 and 11 the null hypothesis is $E(d_N) = E(d_S)$. Since a ratio of two quantities is disturbed more easily by various factors than a difference, the $d_N - d_S$ test seems to be more robust than the McDonald-Kreitman test. Because purifying selection and mutation bias are allowed in Kimura's neutral theory, the McDonald-Kreitman test does not necessarily reject the neutral theory, even if it gives a significant result. Caution is necessary in the interpretation of the results of the McDonald-Kreitman test (Brookfield and Sharp 1994; Moriyama and Powell 1996).

13

Population Trees from Genetic Markers

To construct a phylogenetic tree for a group of populations or closely related species, it is customary to use allele frequency data from many different genetic loci. This is because the genetic differences between populations are small and they are usually measured in terms of allele frequency differences rather than nucleotide or amino acid differences. Of course, any allelic difference is caused by differences in nucleotide sequence between alleles. Therefore, the ideal approach to this problem is to consider the nucleotide differences as well as the frequency differences among populations.

Cavalli-Sforza and Edwards (1964) seem to be the first to investigate the evolutionary relationships of human populations using allele frequency data. They constructed a phylogenetic tree for various human populations and inferred the history of human evolution on Earth. In the 1960s and 1970s, a large number of authors studied protein polymorphism using electrophoresis and examined the extent of genetic differentiation of natural populations or closely related species in various groups of organisms. Actually, research using this simple technique is still going on. Nei (1975, 1987) summarized the progress in this area from the theoretical point of view, whereas biologically interesting findings have been reviewed by Ayala (1975), Thorpe (1982), Avise and Aquadro (1982), and Avise (1994). In this chapter, we are primarily concerned with statistical methods that are useful for analyzing these data as well as for new molecular data such as RFLP, RAPD, and microsatellite DNA data. A formal discussion of the genetic distance theory is presented in Nei's (1987) book, and therefore the reader may refer to it for further details.

13.1. Genetic Distance for Allele Frequency Data

Although **genetic distance** means the extent of genetic differences (genomic difference) between two populations, it usually refers to the genetic difference as measured by a function of allele frequencies. As in the case of evolutionary distances discussed in chapters 3 and 4, genetic distance can be used for estimating the time of divergence between popula-

tions and constructing phylogenetic trees of populations. The concept of genetic distance was first used by Sanghvi (1953), who proposed to measure the distance between two populations by a mathematical quantity that is similar to the X^2 statistic for testing allele frequency differences. Later, various authors proposed many different measures from various points of view. Here we present only a few measures that are often used for actual data analysis.

Distance Measures Based on Geometric Consideration

Rogers' Distance

Suppose that there are q alleles at a locus, and let x_i and y_i be the frequencies of the i-th allele in populations X and Y, respectively. Each allele frequency may take a value between 0 and 1. Therefore, it is possible to represent populations X and Y in a q-dimensional space. The distance between the two populations in the space is then given by

$$d_R = \left[\sum_{i=1}^{q}(x_i - y_i)^2\right]^{1/2} \tag{13.1}$$

This distance takes a value between 0 and $\sqrt{2}$, the latter value being obtained when the two populations are fixed for different alleles. This property is not very desirable. So, Rogers (1972) proposed the following measure, which takes a value between 0 and 1.

$$D_R = \left[\frac{1}{2}\sum_{i=1}^{q}(x_i - y_i)^2\right]^{1/2} \tag{13.2}$$

When allele frequency data are available for many loci, the average of this value is used, and the standard error of the average D_R is computed from the variance of D_R among loci. Note, however, that this measure has one deficiency. When the two populations are both polymorphic but share no common alleles, D_R is given by $[(\Sigma x_i^2 + \Sigma y_i^2)/2]^{1/2}$. This value can be much smaller than 1 even if the populations have entirely different sets of alleles. For example, when there are five nonshared alleles in each population and all allele frequencies are equal ($x_i = 1/5$; $y_i = 1/5$), we have $D_R = 0.45$. This property is clearly undesirable.

Bhattacharyya's Angular Transformation and Its Modifications

Representing two populations on the surface of a multidimensional hypersphere, Bhattacharyya (1946) suggested that the extent of differentiation of populations be measured in terms of the angle (θ) between the two lines projecting from the origin to the two populations (X and Y) on the hypersphere (Figure 13.1). When there are q alleles, we consider a q-dimensional hypersphere with radius 1 and let each axis represent

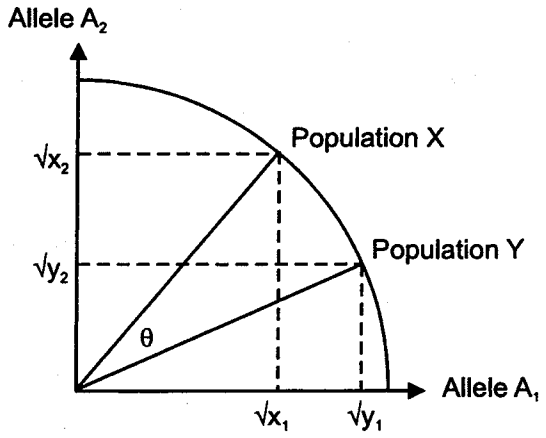


FIGURE 13.1. Bhattacharyya's geometric representation of populations X and Y for the case of two alleles.

the square root of the allele frequency, that is, $\xi_i = \sqrt{(x_i)}$ and $\eta_i = \sqrt{(y_i)}$. Therefore, $\Sigma \xi_i^2 = \Sigma \eta_i^2 = 1$. When there are only two alleles, populations X and Y can be represented on a circle, as shown in Figure 13.1. Elementary geometry shows that in the case of q alleles the angle θ is given by

$$\cos\theta = \sum_{i=1}^{q} \xi_i \eta_i = \sum_{i=1}^{q} \sqrt{x_i y_i} \tag{13.3}$$

Bhattacharyya proposed that the distance between two populations be measured by

$$\begin{aligned} \theta^2 &= \left[\arccos\left(\sum \sqrt{x_i y_i}\right)\right]^2 \\ &\approx \frac{1}{2} \sum_{i=1}^{q} \frac{(x_i - y_i)^2}{(x_i + y_i)} \end{aligned} \tag{13.4}$$

This measure takes a value between 0 and 1.

Cavalli-Sforza and Edwards (1967) also used an angular transformation. They proposed that the genetic distance between two populations be measured by the chord length between points X and Y on the q-dimensional hypersphere. This chord length is given by $[2(1 - \cos\theta)]^{1/2}$. In practice, they used

$$d_C = (2/\pi)\left[2\left(1 - \sum_{i=1}^{q} \sqrt{x_i y_i}\right)\right]^{1/2} \tag{13.5}$$

as a distance measure, because $\theta = \pi/2$ corresponds to the case of complete gene substitution.

In a computer simulation, Nei et al. (1983) noted that the following distance measure is quite efficient in recovering the true topology of an evolutionary tree when it is reconstructed from allele frequency data.

$$D_A = \sum_{k=1}^{L}\left(1 - \sum_{i=1}^{q_k}\sqrt{x_{ik}y_{ik}}\right)/L \tag{13.6}$$

where q_k and L are the number of alleles at the k-th locus and the number of loci examined, respectively. This measure takes a value between 0 and 1, the latter value being obtained when the two populations share no common alleles. Since the maximum value of D_A is 1, D_A is nonlinearly related to the number of gene substitutions. When D_A is small, however, it increases roughly linearly with evolutionary time.

The standard error of D_A or the difference in D_A between two populations can again be computed by the bootstrap method if it is based on many loci. In this case, a bootstrap sample will represent a different set of loci, which have been chosen at random with replacement. Similarly the standard errors of average D_R, θ^2, and d_C can be computed by the bootstrap.

Distance Measures Based on Evolutionary Models

F_{ST}^* Distance

In chapter 12, we have seen that the allele frequencies of different populations may differentiate by genetic drift alone without any selection. When a population splits into many populations of effective size N in a generation, the extent of differentiation of allele frequencies in subsequent generations can be measured by F_{ST} in Equation (12.29). When there are only two populations but allele frequency data are available for many loci, it is possible to develop a statistic whose expectation is equal to F_{ST}. Latter (1972) developed such a statistic, which is given by

$$F_{ST}^* = [(\hat{J}_X + \hat{J}_Y)/2 - \hat{J}_{XY}]/(1 - \hat{J}_{XY}) \tag{13.7}$$

where $\hat{J}_X$, $\hat{J}_Y$, and $\hat{J}_{XY}$ are unbiased estimators of the means (J_X, J_Y, and J_{XY}) of Σx_i^2, Σy_i^2, and $\Sigma x_i y_i$ over all loci, respectively. For a single locus, unbiased estimates of Σx_i^2, Σy_i^2, and $\Sigma x_i y_i$ are given by

$$\hat{j}_X = \left(2m_X\sum \hat{x}_i^2 - 1\right)/(2m_X - 1) \tag{13.8}$$

$$\hat{j}_Y = \left(2m_Y\sum \hat{y}_i^2 - 1\right)/(2m_Y - 1) \tag{13.9}$$

$$\hat{j}_{XY} = \sum \hat{x}_i\hat{y}_i \tag{13.10}$$

where m_X and m_Y are the numbers of diploid individuals sampled from populations X and Y, respectively, and $\hat{x}_i$ and $\hat{y}_i$ are the sample frequencies of allele A_i in populations X and Y (Nei 1987). Therefore, $\hat{J}_X$, $\hat{J}_Y$, and $\hat{J}_{XY}$ are the means of $\hat{j}_X$, $\hat{j}_Y$, and $\hat{j}_{XY}$ over all loci, respectively. As mentioned in the previous chapter, the expectation of F_{ST}^* is given by

$$E(F_{ST}^*) = 1 - e^{-t/(2N)} \tag{13.11}$$

Therefore,

$$D_L = -\ln(1 - F_{ST}^*) \tag{13.12}$$

is expected to be proportional to t when the number of loci used is large [$E(D_L) = t/(2N)$.] This indicates that when evolutionary time is short and new mutations are negligible, one can estimate t by $2ND_L$ if N is known (Latter 1972). In practice, however, new mutations always occur, and this will disturb the linear relationship between D_L and t when a relatively long evolutionary time is considered. N is also usually unknown.

Standard Genetic Distance

Nei (1972) developed a genetic distance measure (called **standard genetic distance**) whose expected value is proportional to evolutionary time when both effects of mutation and genetic drift are taken into account. It is defined by

$$D = -\ln I \tag{13.13}$$

where

$$I = \hat{J}_{XY} / \sqrt{\hat{J}_X \hat{J}_Y} \tag{13.14}$$

The variances of I and D can be computed by the formulas given by Nei (1978, 1987).

When the populations are in mutation-drift balance throughout the evolutionary process and all mutations result in new alleles (infinite-allele model), the expectation of D increases in proportion to the time after divergence between two populations. That is,

$$E(D) = 2\alpha t \tag{13.15}$$

where α is the rate of mutation or gene substitution (Nei 1972). Therefore, if we know α, we can estimate divergence time from D.

The α value varies with genetic locus and the type of data used. For the genetic loci that are commonly used in protein electrophoresis, Nei (1975, 1987) suggested that α is approximately 10^{-7} per locus per year. If this is correct, the time after divergence between two populations is estimated by

$$t = 5 \times 10^6 D \tag{13.16}$$

This formula is based on the assumption that all loci have the same rate of gene substitution. In practice, this assumption is unlikely to hold. In fact, Nei et al. (1976a) showed that α varies among loci approximately following the gamma distribution given by Equation (2.9). They then showed that the average value of I over loci is given by

$$I_A = \left[\frac{a}{a + 2\bar{\alpha}t}\right]^a \tag{13.17}$$

where a is the gamma parameter in Equation (2.9) and $\bar{\alpha}$ is the mean of α over loci. Therefore, the number of gene substitutions per locus is given by

$$D_v = 2\bar{\alpha}t = a[(1 - I_A)^{-1/a} - 1] \tag{13.18}$$

When $a = 1$, this becomes

$$D_v = (1 - I_A)/I_A \tag{13.19}$$

Here I_A is estimated by Equation (13.14). In the case of $a > 0$, t can be estimated by replacing D in Equation (13.16) by D_v. Note that D_v is nearly equal to D when $I_A \geq 0.8$ and $a = 1$.

Nei (1987) examined various data sets where the D_v values and information on approximate divergence times (t's) are available and showed that when a sufficiently large number of protein loci are used, the relationship $t = 5 \times 10^6 D_v$ with $a = 1$ approximately holds. Of course, when D_v is greater than 0.5, the reliability of $\hat{t}$ declines, because the variance of $\hat{D}_v$ becomes very large. The standard error of $\hat{t}$ is given by

$$s(\hat{t}) = 5 \times 10^6 \times s(\hat{D}_v) \tag{13.20}$$

where $s(\hat{D}_v)$ is the standard error of $\hat{D}_v$ and can be computed by the bootstrap method.

$(\delta\mu)^2$ and Its Related Distances

As mentioned in chapter 12, microsatellite DNA loci are highly polymorphic with respect to the number of repeats of short nucleotides, and therefore they are useful for studying phylogenetic relationships of populations. Goldstein et al. (1995a) and Slatkin (1995) proposed a genetic distance measure that increases linearly with evolutionary time when the mutation-drift balance is maintained throughout the evolutionary process and the **stepwise mutation model** (Figure 12.2) applies. This distance (D_G) for a locus is given by

$$D_G = \sum_i \sum_j (i - j)^2 x_i y_j \tag{13.21}$$

where x_i and y_j are the frequency of the i-th allele in population X and the frequency of the j-th allele in population Y, respectively. Here, the alleles (i, j) are numbered according to the number of repeats (allele size). Unfortunately, this distance measure has a large variance even if many loci are used. For this reason, Goldstein et al. (1995b) proposed that the following distance measure be used for microsatellite DNA data.

$$(\delta\mu)^2 = \sum_k^L (\mu_{Xk} - \mu_{Yk})^2/L \tag{13.22}$$

where μ_{Xk} ($= \Sigma_i x_{ik}$) and μ_{Yk} ($= \Sigma_i y_{jk}$) are the mean numbers of repeats at the k-th locus in populations X and Y, respectively. The expectation of $(\delta\mu)^2$ is given by $E(\delta\mu)^2 = 2\alpha t$, where α is the mutation rate per generation. Therefore, t can be estimated by $(\delta\mu)^2/(2\alpha)$.

In practice, however, there are a number of problems with this method. First, the α value apparently varies considerably with locus and organism, and it is not a simple matter to estimate α for each locus. In humans, Weber and Wong's (1993) experimental data for 15 dinucleotide loci have suggested that α is approximately 5×10^{-4} per locus per generation (see Goldstein et al. 1995a), whereas Mahtani and Willard (1993) estimated α to be of the order of 0.01 per locus per generation. In *Drosophila,* Schug et al. (1997) estimated that the average mutation rate is about 10^{-5} per generation. At the present time, our knowledge about the mutation rate for microsatellite DNA is very poor. Second, the variance or the coefficient of variation of $(\delta\mu)^2$ is very high compared with that of other distance measures such as d_C and D_A (Takezaki and Nei 1996). Therefore, a large number of loci must be used to obtain a reliable estimate of t even if α is known. Third, there is evidence that the actual mutational pattern is irregular and deviates considerably from the stepwise mutation model (section 12.3) and that there seems to be an upper limit on the number of repeats (Weber and Wong 1993; Forbes et al. 1995; Garza et al. 1995). Furthermore, some microsatellite loci are highly polymorphic in some populations or species but monomorphic in others (Bowcock et al. 1994; Taylor et al. 1999). This raises a question about long-term stability of microsatellite loci. Therefore, more studies about the mutation pattern of microsatellite loci seem to be necessary before we can use $(\delta\mu)^2$ for estimating evolutionary time.

Genetic Distance and Phylogenetic Trees

A linear relationship of a distance measure with evolutionary time is important for estimating the time of divergence between two populations. It is also a nice property for constructing phylogenetic trees, other things being equal. In practice, however, different distance measures have different variances, and for this reason a distance measure that is linear with time is not necessarily better than a nonlinear distance in obtaining true trees (topologies) as in the case of distances based on nucleotide or amino acid sequences.

Takezaki and Nei (1996) and Rao et al. (1997) studied this problem by computer simulation. Takezaki and Nei considered a model tree for eight populations, as shown in Figure 13.2, and assumed that the populations are in mutation-drift balance throughout the evolutionary process, except when a bottleneck effect was intentionally introduced to see the effect of different rates of gene substitution in different evolutionary lineages. The actual procedure of the simulation was as follows.

(1) The ancestral population was split into two populations, and, at a later time, one of the two populations was again split into two populations. This process was continued until the eight populations in the model tree were generated.

(2) Starting with the ancestral population, the allele frequency change

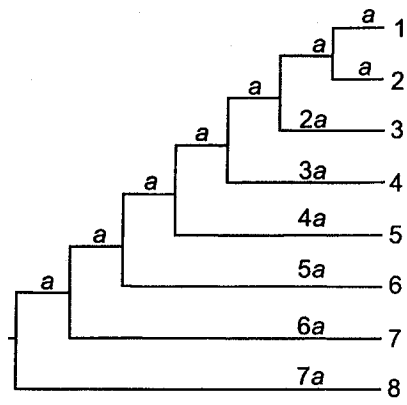


FIGURE 13.2. Model tree used for computer simulation. *a* is a branch length in the unit of expected number of mutations accumulated per locus (vt). The *a* value used was 0.1 and 0.04 for the infinite-allele model and 0.4 and 0.04 for the stepwise-mutation model.

in each population was followed by introducing mutation and sampling alleles at random for the next generation, and at the end of the evolutionary change the allele frequencies for all populations were recorded. The number of loci examined in each replication was 100 for the infinite-allele model and 300 for the stepwise mutation model.

(3) Matrices of pairwise distances for the eight populations were computed for the first 10, 20, 30, 40, . . . , 100 loci for each of the distance measures examined.

(4) For each of these distance matrices, phylogenetic trees were constructed by using the NJ method and UPGMA, and the topology of each reconstructed tree was compared with that of the model tree.

(5) This simulation was repeated 100 times for each set of branch lengths considered (Figure 13.2), and the proportion of replications in which the correct topology was obtained (probability of obtaining the true topology) was computed.

The general conclusions obtained from this simulation and Rao et al.'s (1997) are as follows.

(1) For all distance measures, the probability of obtaining the true topology (P_T) was very low when the number of loci used was less than ten but gradually increased with increasing number of loci. In general, P_T is lower for the stepwise mutation model than for the infinite-allele model. This indicates that a larger number of loci should be used for microsatellite DNA data than for electrophoretic data when the level of average heterozygosity is the same.

(2) When the rate of gene substitution is constant and is the same for all populations, UPGMA often gives a slightly higher P_T than NJ does. However, when the rate is not constant, UPGMA generally shows a lower P_T value than NJ does, as expected.

(3) Distance measures D_A and d_C are generally more efficient in obtaining the true topology than other distance measures under many different conditions examined for both the infinite-allele and the stepwise

mutation model. When the extent of population divergence is high, D_A is slightly better than d_C. The high efficiencies of D_A and d_C are primarily due to the fact that they have smaller variances or smaller coefficients of variation (CV) than others. However, CV's are not the only factor that determines the P_T value. The slightly lower efficiency of d_C than D_A is apparently due to the fact that d_C initially increases rapidly with increasing time but the rate of increase rapidly declines in comparison with D_A.

(4) The mean values of D and $(\delta\mu)^2$ increase linearly with time under the infinite-allele and the stepwise mutation models, respectively, but the efficiencies of these measures in obtaining the true tree are quite low primarily because of the high CV's of these distances. The P_T value for $(\delta\mu)^2$ is generally lower than that for D even for the stepwise mutation model except when the extent of genetic differentiation of populations is extremely high. For $(\delta\mu)^2$, more than 100 loci are required to have a reasonably high P_T value.

(5) When the total number of individuals to be studied is fixed, it is generally better to examine more loci with a smaller number of individuals per locus rather than fewer loci with a large number of individuals in order to have a high P_T value, as long as the number of individuals per locus is greater than about 25. When the average heterozygosity is as high as 0.8, however, a large number of individuals per locus need to be studied (Nei 1978; Archie 1985; Takezaki and Nei 1996).

Example 13.1. Evolutionary Relationships of Human Populations

Bowcock et al. (1994) examined microsatellite DNA (mostly CA repeats) polymorphisms for 30 loci from 14 human populations and for 25 loci from one chimpanzee species. They constructed a phylogenetic tree for the human populations without using chimpanzee data, whereas Nei and Takezaki (1996) constructed a tree for 25 loci using chimpanzees as the outgroup. The D_A and $(\delta\mu)^2$ distances for the 25 loci shared by humans and chimpanzees are presented in Table 13.1, and the phylogenetic trees obtained by the neighbor joining method are given in Figure 13.3. The tree obtained by D_A distances (Figure 13.3A) is identical with that of Nei and Takezaki and is also virtually the same as that obtained by Bowcock et al. with d_C distances.

This tree shows that Africans (Pygmies and Bantu) first separated from the rest of the human groups and that the bootstrap values for the interior branches connecting Africans and chimpanzees and non-Africans and chimpanzees are both very high. This result supports the currently popular view that modern humans originated in Africa (Cann et al. 1987; Vigilant et al. 1991). The same tree shows that Europeans first diverged from the non-African people and then the group of New Guineans and native Australians separated from the remaining group. The first separation of Europeans from the rest of non-Africans is well supported by bootstrap values, but the next separation of New Guineans and Australians is less clear, because the bootstrap value for one of the two interior branches involved is only 53%. In fact, a similar study using classical markers (blood group and allozyme data) has suggested that New Guineans and Australians are genetically close to southeastern Asians (Indonesians,

Table 13.1 D_A distances (× 100; below diagonal) and $(\delta\mu)^2$ distances (above diagonal) for 25 microsatellite loci from one chimpanzee and 14 human populations.

Population Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Pygmy (CAR)[a]		5.1	14.1	22.9	32.3	44.2	30.5	16.2	20.7	48.9	19.8	20.3	29.3	22.2	103.0
2. Pygmy (Lisongo)	25.3		14.6	20.5	28.2	39.5	26.7	14.5	19.5	45.9	18.1	16.8	28.4	17.2	99.7
3. Pygmy (Zaire)	27.4	27.0		24.5	34.4	46.8	42.4	33.2	32.1	52.8	31.7	29.8	42.6	32.1	100.6
4. Cambodian	36.9	36.3	38.2		3.3	6.2	9.8	17.3	8.5	13.4	8.3	8.1	6.4	9.9	85.7
5. Chinese	43.8	39.4	40.5	14.3		5.4	10.5	23.1	11.6	10.4	8.8	10.8	9.6	11.5	83.5
6. Japanese	41.7	39.1	40.1	17.9	15.6		8.1	26.9	13.2	9.1	15.8	17.7	5.2	16.8	80.6
7. Amerindian (K)[b]	51.8	49.3	53.1	32.1	30.6	30.2		12.8	7.9	17.2	12.4	13.2	6.6	9.0	70.7
8. Amerindian (S)[c]	41.7	41.6	43.9	25.3	25.1	26.0	19.4		5.2	37.7	12.4	10.6	17.5	6.9	98.1
9. Amerindians (M)[d]	39.5	40.5	41.5	21.7	22.0	21.9	26.5	17.9		20.5	6.7	9.4	8.4	7.2	89.8
10. North Italian	33.9	33.7	40.1	26.1	25.3	27.8	35.8	29.9	26.8		12.8	20.2	10.3	29.2	89.1
11. North European	36.6	33.7	39.0	21.8	22.4	23.7	35.4	30.7	24.7	14.4		5.3	9.6	10.6	95.3
12. Australian	41.1	39.0	47.5	22.8	24.3	28.3	38.8	30.5	26.3	24.5	23.7		11.4	5.1	101.6
13. Melanesian	43.2	44.9	49.4	24.7	26.6	28.6	42.5	34.3	27.6	28.6	28.8	22.9		13.7	85.3
14. New Guinean	45.5	43.7	50.8	26.9	27.2	28.6	36.8	38.2	29.0	34.2	28.4	21.8	26.1		92.9
15. Chimpanzee	61.0	62.1	64.5	64.0	67.1	59.4	68.3	67.5	66.3	64.4	61.3	71.4	69.2	69.4	

[a]CAR: Central African Republic.
[b]K: Karitiana.
[c]S: Surui.
[d]M: Maya.

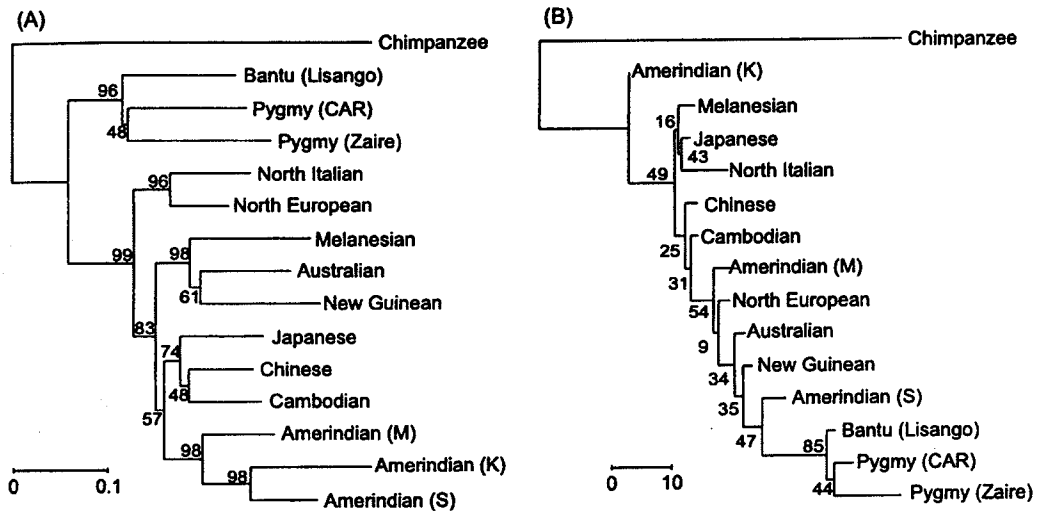


FIGURE 13.3. Neighbor-joining trees of human populations obtained by using Bowcock et al.'s (1994) data of 25 microsatellite loci. (A) Tree obtained by D_A distance. (B) Tree obtained by $(\delta\mu)^2$ distance. The number for each interior branch is the bootstrap value from 1000 replications. M: Maya. K: Karitiana. S: Surui. CAR: Central African Republic. Constructed by the computer program available at *http://mep.bio.psu.edu.*

Filipinos, Thais) (Nei and Roychoudhury 1993). To clarify this aspect of evolutionary relationships, it seems necessary to examine many more loci.

Figure 13.3B shows the tree obtained by $(\delta\mu)^2$ distances. The topology of this tree is very different from that for D_A distances and is poorly supported by the bootstrap test. This unreliable tree was obtained mainly because the sampling error of $(\delta\mu)^2$ is very large, as mentioned earlier.

13.2. Analysis of DNA Sequences by Restriction Enzymes

A restriction enzyme recognizes a specific sequence of nucleotide pairs, generally four or six pairs in length, and cleaves it. Therefore, if a circular DNA such as animal mitochondrial DNA has n recognition (restriction) sites, it is fragmented into n segments after digestion by this enzyme. The number and location of restriction sites vary with nucleotide sequence. The higher the similarity of the two DNA sequences compared, the closer the cleavage patterns. Therefore, it is possible to estimate the number of nucleotide substitutions between two homologous DNA sequences by comparing the locations of restriction sites. The number of nucleotide substitutions can also be estimated from the proportion of DNA fragments that are shared by the two DNA sequences. This problem has been studied by a number of authors (Upholt 1977; Gotoh et al. 1979; Kaplan and Langley 1979; Nei and Li 1979; Nei and Tajima 1983). When many DNA sequences are sampled from each population, it is also pos-

sible to estimate the nucleotide diversity defined by Equation (12.54) and the number of net nucleotide differences between two populations (Equation [12.67]) from restriction-site data.

This method gives rather crude estimates of the number of nucleotide substitutions compared with the method of direct nucleotide sequencing. However, it is a fast and inexpensive method for estimating the number of nucleotide substitutions and gives fairly accurate results when the sequences are closely related. For this reason, it is used primarily for the study of genetic differentiation of intraspecific populations or closely related species (Avise 1994).

In the study of population genetics, restriction enzymes are often used to identify genetic polymorphism at a nuclear DNA locus, and the alleles detected in this case usually represent different lengths of DNA fragments identified by a restriction enzyme (Botstein et al. 1980). This type of polymorphism is called the **restriction fragment length polymorphism** or **RFLP.** RFLPs have been used for constructing genetic linkage maps, but they are also useful for evolutionary studies. The number of alleles detected at a genetic locus (a segment of DNA) is usually small, and population differences are measured in terms of allele frequency differences (Mountain and Cavalli-Sforza 1994). Therefore, RFLP data can be analyzed by the statistical methods described in the previous section. In this case, the underlying mutation model is unclear, and we recommend that distance measure D_A be used for data analysis (Nei and Takezaki 1996). Of course, when a large number of RFLP loci are examined for closely related populations or species, it is possible to estimate the number of nucleotide differences per site using the restriction fragment method as mentioned below.

In this section, we discuss simple statistical methods for estimating the number of nucleotide substitutions, nucleotide diversity, and so forth from restriction-site or restriction-fragment data without going into details of the theoretical basis. Any reader who is interested in the theoretical foundations should refer to Nei (1987).

Estimation of the Number of Nucleotide Substitutions

Restriction-Site Data

To estimate the number of nucleotide substitutions from restriction-site data, we have to make some simplifying assumptions. One of such assumptions is that the four types of nucleotides (A, T, C, G) are randomly arranged in the DNA sequence, and all nucleotide sites have the same probability of substitution. This assumption does not always hold, but since the restriction enzyme method is used only when the number of nucleotide substitutions per site (d) is relatively small (say, $d < 0.2$), violation of the assumption does not introduce serious errors (Tajima and Nei 1982; Kaplan 1983). Under this assumption, the expected number of restriction sites (n) for a restriction enzyme with a recognition sequence of b nucleotides (usually $b = 4$ or 6) is given by

$$E(n) = n_T a \tag{13.23}$$

where n_T is the total number of nucleotides and a is the probability that a sequence of b nucleotides in the DNA is a restriction site.

Under our assumption, a is given by

$$a = (g_A)^{b_A}(g_T)^{b_T}(g_C)^{b_C}(g_G)^{b_G} \tag{13.24}$$

where g_A, g_T, g_C, and g_G are the frequencies of nucleotides A, T, C, and G in the DNA sequence, respectively, and b_A, b_T, b_C, and b_G are the numbers of A, T, C, and G in the recognition sequence, respectively ($b_A + b_T + b_C + b_G = b$). For example, the recognition sequence of enzyme *Eco*RI is GAATTC, so that $b_A = 2$, $b_T = 2$, $b_C = 1$, and $b_G = 1$. When a restriction enzyme identifies more than one type of recognition sequence (e.g., *Hae*I), a is given by a somewhat different formula, as will be discussed later. Usually, a is much smaller than 1. For example, in the human mitochondrial DNA, $n_T = 16,569$, $g_A = 0.25$, $g_T = 0.31$, $g_C = 0.14$, and $g_G = 0.30$, so that $a = 0.00025$ and $E(n) = 4.2$ for *Eco*RI. This expected number of restriction sites per DNA sequence agrees quite well with the observed number (3.8) (Ferris et al. 1981b). Obviously, a is generally greater for four-base restriction enzymes than for six-base enzymes. However, when the nucleotide frequencies deviate considerably from 1/4, a can be small even for a four-base enzyme.

Let us now consider two DNA sequences (X and Y) that diverged t years (or generations) ago and compare all restriction sites of the two sequences. We note that there are n_T possible restriction sites in a circular DNA of n_T nucleotides. For a linear DNA, the possible number of restriction sites is $n_T - b + 1$, but since n_T is usually much larger than b, the possible number is again approximately n_T. In the comparison of restriction sites between two DNA sequences, there are four different cases. A sequence of b nucleotides at a particular position of the DNA is a restriction site (1) for both X and Y, (2) for X but not for Y, (3) for Y but not for X, or (4) for neither X nor Y. Let n_X and n_Y be the numbers of restriction sites for DNA sequences X and Y and n_{XY} be the number of restriction sites shared by the two sequences. The numbers of observations of the above four different cases are then given by n_{XY}, $n_X - n_{XY}$, $n_Y - n_{XY}$, and $n_T - n_X - n_Y + n_{XY}$, respectively.

In order to derive a formula for estimating the d value, it is necessary to know the probability (S) that sequences X and Y share the same recognition sequence at a given nucleotide position. Derivation of this probability is somewhat complicated, but it is approximately given by

$$S = e^{-2brt} \tag{13.25}$$

where r is the rate of nucleotide substitution per site per year and t is the divergence time in years. (A more accurate formula is given in Nei 1987.)

It is now possible to estimate S by the maximum likelihood method. The estimator ($\hat{S}$) and its variance [$V(\hat{S})$] become

$$\hat{S} = \frac{2n_{XY}}{n_X + n_Y} \tag{13.26}$$

$$V(\hat{S}) = \frac{S(1 - S)(2 - S)}{2n_T a} \tag{13.27}$$

Since the expected number of nucleotide substitutions per site is given by $d = 2rt$, this number can be estimated by

$$\hat{d} = [-\ln\hat{S}]/b \tag{13.28}$$

with variance

$$V(\hat{d}) = \frac{(2 - S)(1 - S)}{2b^2\bar{n}S} \tag{13.29}$$

where $\bar{n}$ is equal to $n_T a$ and can be estimated by $\bar{n} = (n_X + n_Y)/2$ (Nei 1987).

In the above formulation, we considered restriction-site data obtained by a single restriction enzyme. However, the above formulas apply to data from many different restriction enzymes as long as they have the same b value. The only thing one has to do in this case is to take the summation of n_X, n_Y, and n_{XY} for all enzymes.

Restriction Enzymes with Multiple Recognition Sequences

Most restriction enzymes used for evolutionary studies recognize a unique sequence of 4 or 6 nucleotides. However, there are enzymes that recognize multiple sequences of a given number of nucleotides. For example, *Hind*II recognizes the sequence GTPyPuAC, where Py is either T or C and Pu is either A or G. In this case, a is given by $(g_A + g_C)(g_T + g_C)g_A g_T g_C g_G$, and $S = 1 - e^{-2brt}$ approximately, if we redefine $b = 16/3$ and $d \leq 0.2$ (Nei 1987). Another six-base enzyme with multiple recognition sequences is *Hae*I, which recognizes $\binom{A}{T}GGCC\binom{A}{T}$. In this case, $a = (g_A + g_T)^2 g_C^2 g_G^2$, and $S = 1 - e^{-2brt}$ approximately with $b = 16/3$. However, since $16/3 = 5.33$ is close to 6 and g_A, g_T, g_C, and g_G are usually different from one another, it seems that we had better use $b = 6$ for these enzymes to simplify mathematical computation. Therefore, we follow this simplification in this book. Note that there are enzymes with $b = 5$, but these enzymes are rarely used. (See program RESTDATA at *http://mep.bio.psu.edu.*)

Restriction Fragment Data

In the above theory, we assumed that all restriction sites can be mapped on the DNA sequence so that shared and unshared restriction sites can be determined. Although mapping of restriction sites is not difficult, it is time-consuming when a large-scale population survey has to be made. In the case of mitochondrial or chloroplast DNA, where the total length is

fixed, there is a simpler method. That is, one can estimate d by comparing the electrophoretic patterns of DNAs digested by a restriction enzyme between the two DNA sequences compared. The rationale for this is that the degree of genetic divergence between two DNA sequences is correlated with the proportion of DNA fragments shared by them. Three different methods (Upholt 1977; Nei and Li 1979; Engels 1981) have been developed for estimating d from this information. According to Kaplan's (1983) computer simulation, however, all of them give essentially the same results. In the following, let us discuss Nei and Li's method, because it is simple and mathematically a little more rigorous than the others.

Nei and Li have shown that the expected proportion of shared DNA fragments (F) can be expressed by the following approximate formula

$$F \cong G^4/(3 - 2G) \tag{13.30}$$

where G is e^{-brt}. Therefore, $d \equiv 2rt$ can be estimated from F by the above equation. To estimate d, we first note that F can be estimated by

$$\hat{F} = 2n_{XY}/(n_X + n_Y) \tag{13.31}$$

where n_X and n_Y are the numbers of restriction fragments in DNA sequences X and Y, respectively, and n_{XY} is the number of fragments shared by the two sequences. In practice, many restriction enzymes with the same b value are used, so that n_X, n_Y, and n_{XY} represent the sums of the numbers of DNA fragments for all enzymes. If $\hat{F}$ is obtained, we can estimate G by the iteration formula,

$$\hat{G} = [\hat{F}(3 - 2\hat{G}_1)]^{1/4} \tag{13.32}$$

where $\hat{G}$ is an estimate of G and $\hat{G}_1$ is a trial value of $\hat{G}$. This iterative computation is done until $\hat{G} = \hat{G}_1$ is obtained. Usually a few cycles of iterations are sufficient. Nei (1987) suggested that $\hat{F}^{1/4}$ be used as the first trial value of $\hat{G}_1$. Once $\hat{G}$ is obtained, d can be estimated by

$$\hat{d} = -(2/b)\ln\hat{G} \tag{13.33}$$

Suppose that one obtains $n_X = 40$, $n_Y = 36$, and $n_{XY} = 20$ for a pair of mtDNA sequences using 10 six-base restriction enzymes. We then have $\hat{F} = 0.5263$ and $\hat{G}_1 = \hat{F}^{1/4} = 0.8517$. Putting these into Equation (13.32), we obtain $\hat{G} = 0.9089$. Now regarding this value as $\hat{G}_1$ in Equation (13.32), we have $\hat{G} = 0.8882$. Repeating this process a few more times, we obtain $\hat{G} = 0.8938$ as the final estimate. We therefore have $\hat{d} = 0.037$ from Equation (13.33).

It should be noted that the above method is useful only for the case of small d, because F rapidly declines as d increases. When F is small, it is affected by random errors so much that the estimate of d is unreliable. In general, the estimate obtained by this method (length-difference method) has a larger variance than that obtained by the previous site-difference method. Furthermore, estimates of d obtained by the length-difference method are more susceptible to errors caused by undetectable restriction

fragments or fragment length differences than those obtained by the site-difference method. Nevertheless, the length-difference method is simple to use and gives a fairly accurate estimate of d when d is smaller than 0.05. Therefore, in the estimation of d for a pair of highly homologous DNA sequences, it is a useful method. Avise (1994) describes many cases in which this technique has been used effectively in the study of genetic differentiation of populations.

Estimation of the Number of Nucleotide Substitutions from Many Different Enzymes

When different kinds of enzymes with different b values are used, estimation of d becomes a little more complicated. Nei and Tajima (1983) developed a maximum likelihood estimator of d, but it requires an iterative solution. However, when $d \leq 0.2$, for which this method is intended to be used, the following simple formula gives sufficiently accurate estimates (Nei and Miller 1990).

$$\hat{d} = \frac{\sum_k n_k b_k \hat{d}_k}{\sum_k n_k b_k} \tag{13.34}$$

where subscript k refers to the k-th class of restriction enzymes, and summation is taken over all different enzyme classes. In practice, only two different classes with $b = 4$ and 6 are used in most cases. Note that this equation is applicable for both restriction-site and restriction-fragment data, as long as the estimates for different enzyme classes are available. It is not simple to derive a formula for the variance of $\hat{d}$ in Equation (13.34), but the variance of $\hat{d}$ can be obtained by the jackknife or bootstrap method.

Nucleotide Diversity and Nucleotide Divergence

Restriction-Site Data

In chapter 12, we discussed nucleotide diversity (π) as a measure of sequence variation within species. This measure is the average value of d_{ij}'s for all pairwise comparisons. Therefore, if we compute $\hat{d}_{ij}$'s from restriction data, we can compute the estimate ($\hat{\pi}$) by using Equation (12.55) or (12.56).

However, the above computation becomes time-consuming when the number of sequences (m) is large and many different enzymes are used. A simpler method for this case is to compute a single S value for each enzyme class and then estimate π by using a formula equivalent to Equation (13.28). In this case, the number of restriction sites for the i-th sequence and the number of restriction sites between the i-th and j-th sequences may be tabulated as shown in Table 13.2. We can then compute the following quantity

$$\tilde{S} = \frac{2\sum_{i<j} n_{ij}}{\sum_{i<j} n_{i(j)} + \sum_{i<j} n_{(i)j}} \tag{13.35}$$

Table 13.2 Number of restriction sites in a sequence (n_i) and the number of shared restriction sites between sequences (n_{ij}).

	1	2	3	4	...	m
1	n_1					
2	n_{21}	n_2				
3	n_{31}	n_{32}	n_3			
4	n_{41}	n_{42}	n_{41}	n_4		
...	...	...	...	...	...	
m	n_{m1}	n_{m2}	n_{m3}	n_{m4}	...	n_m

for each enzyme class, where $n_{i(j)}$ and $n_{(i)j}$ are the n_i and n_j values, respectively, when sequences i and j are compared (see Table 13.2). Note that $\widetilde{S}$ is an average of S_{ij} weighted with $(n_i + n_j)/2$. Therefore, if $(n_i + n_j)/2$ is the same for all i's and j's, $\widetilde{S}$ becomes the arithmetic mean of S_{ij}. Actually, the denominator of Equation (13.35) can be written as $(m - 1)\Sigma_i n_i$. Therefore, we can rewrite Equation (13.35) as

$$\widetilde{S} = \frac{2\sum_{i<j} n_{ij}}{(m-1)\sum_i n_i} \tag{13.36}$$

The π value for the k-th enzyme class can then be estimated by

$$\hat{\pi}_k = [-\ln \widetilde{S}_k]/b_k \tag{13.37}$$

where subscript k refers to the k-th enzyme class.

Once $\hat{\pi}_k$ for this class is obtained for all classes, π can be estimated by

$$\hat{\pi} = \frac{\sum_k \bar{n}_k b_k \pi_k}{\sum_k \bar{n}_k b_k} \tag{13.38}$$

where $\bar{n}_k$ is the average $(\Sigma_i n_i/m)$ of n_i for the k-th class of enzymes. The $\hat{\pi}$ value in the above equation has advantages and disadvantages compared with $\hat{\pi}$ obtained by Equation (12.56). When the number of restriction sites examined is large, Equation (12.56) is expected to give a more reliable estimate. However, when this number is small, the above method seems to be better, because the averaging of n_i and n_{ij} would reduce the effect of sampling errors. The computational process in the latter method is also much simpler than that in the former, as mentioned earlier.

Nucleotide Divergence

In chapter 12, we have shown that the extent of nucleotide divergence (average number of net nucleotide substitutions per site) between two populations, X and Y, can be measured by $d_A = d_{XY} - (d_X + d_Y)/2$, where

d_X and d_Y are the π values in populations X and Y, respectively, whereas d_{XY} is the average number of nucleotide substitutions per site between X and Y. In the case of restriction-site data, we can estimate d_X and d_Y by Equation (13.38). A similar equation for estimating d_{XY} can also be developed (Nei and Miller 1990). In this case, we first compute $\tilde{S}_{XY}$ defined by

$$\tilde{S}_{XY} = \frac{2\sum_{ij} n_{XiYj}}{m_Y \sum_i n_{Xi} + m_X \sum_j n_{Yj}} \tag{13.39}$$

where n_{XiYj} is the number of restriction sites shared by the i-th sequence from population X and the j-th sequence from Y, whereas m_X and m_Y are the numbers of sequences examined in populations X and Y, respectively, and n_{Xi} and n_{Yj} are the number of restriction sites for the i-th sequence from population X and the j-th sequence from population Y, respectively. We can then estimate d_{XY} for the k-th enzyme class by

$$\hat{d}_{XYk} = [-\ln \tilde{S}_{XYk}]/b_k \tag{13.40}$$

and the estimate of d_{XY} for all enzyme classes is given by

$$\hat{d}_{XY} = \frac{\sum_k \bar{n}_{XYk} b_k \hat{d}_{XYk}}{\sum_k \bar{n}_{XYk} b_k} \tag{13.41}$$

where $\bar{n}_{XYk}$ is the average of n_k for populations X and Y. The d_A value is then estimated by

$$\hat{d}_A = \hat{d}_{XY} - (\hat{d}_X + \hat{d}_Y)/2 \tag{13.42}$$

Example 13.2. Restriction-Site Variation Within and Between Two Species of Chimpanzees

Ferris et al. (1981a, 1981b) studied the restriction site polymorphism among 10 mtDNAs from common chimpanzees and 3 mtDNAs from pygmy chimpanzees (bonobos). They used 15 restriction enzymes with $b = 6$ and one enzyme with $b = 4$. The n_i and n_{ij} values for the 13 mtDNAs are given in Table 13.3. Let us first compute $\hat{\pi}$ for pygmy chimpanzees using Equation (13.38). To obtain this $\hat{\pi}$, we must first compute $\tilde{S}$ separately for enzymes with $b = 6$ and $b = 4$. For enzymes with $b = 6$, we have $(m - 1)\Sigma_i n_i = 2 \times (45 + 44 + 45) = 268$, and $2\Sigma_{i<j} n_{ij} = 2 \times (41 + 45 + 41) = 254$. Therefore, $\tilde{S}_1 = 254/268 = 0.948$. Similarly, we obtain $\tilde{S}_2 = 0.962$ for $b = 4$. From these values, we obtain $\hat{\pi}_1 = 0.0089$ and $\hat{\pi}_2 = 0.0098$ for $b = 6$ and 4, respectively, using Equation (13.37). From Table 13.3, we also obtain $\bar{n}_1 = 44.67$ and $\bar{n}_2 = 8.67$. Therefore, $\hat{\pi}$ is 0.0090 from Equation (13.38). The standard error of $\hat{\pi}$ can be obtained by the bootstrap method, and it becomes 0.0027.

A similar computation for common chimpanzees gives a $\hat{\pi}$ value of

Table 13.3 Number of restriction sites in a sequence (n_i on diagonal) and the number of shared restriction sites between sequences (n_{ij} below diagonal) for 10 mtDNAs from common chimpanzees (1–10) and 3 mtDNAs from pygmy chimpanzees (11–13).

	1	2	3	4	5	6	7	8	9	10	11	12	13
1.	46,6												
2.	43,6	46,6											
3.	43,6	46,6	46,6										
4.	42,6	43,6	43,6	50,7									
5.	42,6	43,6	43,6	50,7	50,7								
6.	46,6	43,6	43,6	42,6	42,6	46,6							
7.	43,6	44,6	44,6	42,6	42,6	43,6	46,6						
8.	42,6	43,6	43,6	50,7	50,7	43,6	42,6	50,7					
9.	44,6	42,6	42,6	42,6	42,6	44,6	42,6	42,6	45,6				
10.	43,6	46,6	46,6	43,6	43,6	43,6	44,6	43,6	42,6	46,6			
11.	38,6	37,6	37,6	38,6	38,6	38,6	36,6	38,6	39,6	37,6	45,9		
12.	37,6	38,6	38,6	39,6	39,6	37,6	37,6	39,6	38,6	38,6	41,8	44,8	
13.	38,6	37,6	37,6	38,6	38,6	38,6	36,6	38,6	39,6	37,6	45,9	41,8	45,9

Note: First numbers are for 6-base enzymes and the second numbers are for 4-base enzymes. Computed by the program RESTDATA available at *http://mep.bio.psu.edu.*

0.0133 with standard error 0.0033. This indicates that mtDNAs are more polymorphic in common chimpanzees than in pygmy chimpanzees, but the difference between the two $\hat{\pi}$'s is not statistically significant.

We can also compute $\hat{d}_{XY}$ and $\hat{d}_A$ between the two species. They become $\hat{d}_{XY} = 0.0358 \pm 0.0075$ and $\hat{d}_A = 0.0244 \pm 0.0067$. The $\hat{d}_A$ value is significantly greater than 0 at the 0.1% level, even though the sample sizes are very small.

Restriction-Fragment Data

When the populations are closely related, it is more convenient to use restriction-fragment data to estimate the number of nucleotide differences. In this case, the number of nucleotide substitutions per site for a pair of DNA sequences is estimated by solving for $\hat{G}$ in Equation (13.32), using information on the fraction of shared fragments between the two sequences ($\hat{F}$).

When many DNA sequences are examined in a population, π can be estimated by Nei and Li's (1979) method, but Nei and Miller's (1990) method is again simpler to use. In this method, the following single F value is computed.

$$\widetilde{F} = \frac{2\sum_{i<j} n_{ij}}{(m-1)\sum n_i} \tag{13.43}$$

where n_i and n_{ij} are the number of fragments and the number of shared fragments that are equivalent to those in Table 13.2. Once $\widetilde{F}$ is obtained, the π for the k-th enzyme class is estimated by

$$\hat{\pi}_k = -(2/b)\ln\tilde{G} \tag{13.44}$$

where $\tilde{G}$ is the value of $\hat{G}$ given by Equation (13.32) when $\tilde{F}$ is used. One can then compute $\hat{\pi}$ using Equation (13.38). In this case, however, $\bar{n}_k$ represents the average number of restriction fragments per sequence for the k-th class of enzymes. The $\hat{d}_{XY}$ and $\hat{d}_A$ values can be obtained in the same way.

However, note that $\tilde{F}$ declines to 0 more rapidly than $\tilde{S}$ does as π or d_{XY} increases. Therefore, this method should be applied only to intraspecific populations or to very closely related species.

Restriction-Site Data and Phylogenetic Trees

We have discussed various methods of estimating the number of nucleotide substitutions between a pair of DNA sequences (d) or between a pair of populations (d_{XY} or d_A). If this estimate is obtained for many different sequences or populations, it is possible to construct phylogenetic trees (e.g., Cann et al. 1987; Harris and Hey 1999). In this case, one can use various distance methods of phylogenetic inference such as UPGMA and the NJ method. For restriction-site data, however, it is also possible to use parsimony (MP) methods, because the presence and absence of a restriction site at a particular DNA position can be coded as 1 and 0 (DeBry and Slade 1985).

Using a computer simulation, Jin and Nei (1991) studied the efficiencies of obtaining the correct topology from restriction-site data when UPGMA, NJ, and the standard MP methods are used. In this simulation, model trees of six DNA sequences with and without the molecular clock were used. The results indicated that NJ and MP are considerably more efficient than UPGMA, whether the molecular clock holds or not, and that NJ is slightly better than MP. However, the model trees examined are limited, so that more studies are necessary to derive a definitive conclusion.

AFLP Data

The **amplified fragment length polymorphism** (AFLP) technique developed by Vos et al. (1995) is a powerful method of DNA fingerprinting of nuclear genomes. It is a selective PCR amplification of restriction fragments from a total digest of genomic DNA. In this method, the entire DNA is first digested with two restriction enzymes (*Eco*RI and *Mse*I in the original protocol), and double-stranded oligonucleotide adapters are ligated to both ends of a restriction fragment. PCR primers complementary to the adapters and parts of the restriction site are then used for amplification of the fragment that is flanked by the adapters. Only those restriction fragments that perfectly match the primer sequences are amplified. Usually, 50–100 restriction fragments are amplified and detected on denaturing polyacrylamide electrophoretic gels. However, if a new mutation occurs at a restriction recognition site, the primer will not recognize the site, and a restriction fragment corresponding to the restriction site will not show up on the gel. Therefore, AFLP data are qualitatively sim-

ilar to the restriction-fragment data for mitochondrial DNA mentioned earlier, and the presence and absence of a band among different individuals indicate polymorphism at this locus. This polymorphism can be used to measure the intraspecific as well as interspecific variation.

There are two different ways of analyzing AFLP data to study intra- and interpopulational genetic variation. One way is to treat the presence and absence of a band at a locus as two allelic forms and use the allele frequency analysis as mentioned in sections 12.4 and 13.1 (e.g., Travis et al. 1996). Some authors (e.g., Janssen et al. 1996; Keim et al. 1997; Lopez et al. 1999) have reported that there are AFLP alleles that can be used as diagnostic markers of different strains of subspecies. Another way is to estimate the number of nucleotide substitutions (d) by a method analogous to Equation (13.44) and compute $\hat{d}_X$, $\hat{d}_{XY}$, $\hat{d}_A$, and $\hat{N}_{ST}$. Because two restriction enzymes are used in this case, some modification of the mathematical formulas involved is necessary. Such a modification has been done by Innan et al. (1999).

13.3. Analysis of RAPD Data

Another fast and inexpensive method of studying DNA variation within and between populations is **random amplification of polymorphic DNA (RAPD)** by polymerase chain reaction (PCR) (Welsh and McClelland 1990; Williams et al. 1990; Hadrys et al. 1992). In this method, the total DNA from an organism is subjected to PCR by using short oligonucleotides of random sequence. This method is different from the standard PCR in that only a single random oligonucleotide primer (usually about ten nucleotides long) is used. When the primer is short, there is a high probability that the genome contains several primer sites that are close to one another in the chromosome and are in inverted orientation (Figure 13.4). The PCR technique scans a genome for these inverted repeats and amplifies intervening DNA segments of various lengths. In practice, the amplification products (DNA fragments) of length 400–2,000 base pairs (bp) are identified as an electrophoretic band on agarose gels (Taberner et al. 1997). DNA fragments shorter than 400 bp or longer than 2,000 bp usually do not show up on gels.

Suppose that one nucleotide (say, the first nucleotide) of the primer site AAGACCCCTC in Figure 13.4 mutates to another nucleotide (say, G). Then the primer no longer recognizes the primer site. Therefore, the intervening DNA segment will not be amplified. It may recognize the following primer site with the correct nucleotide sequence, but in this case the intervening DNA segment is likely to be too long to be amplified by PCR. Thus, the mutant form will not show up on electrophoretic gels. Similarly, any mutation that occurs in the primer site will not be recognized by electrophoresis. Therefore, in haploid or homozygous diploid organisms, the polymorphism at a locus (DNA segment) is detected when some individuals show an electrophoretic band and others do not.

Although the principle of RAPD analysis is simple, there are some practical problems. First, the assumption that a primer fails to match the primer site whenever one or more mutations occur may not hold, and the

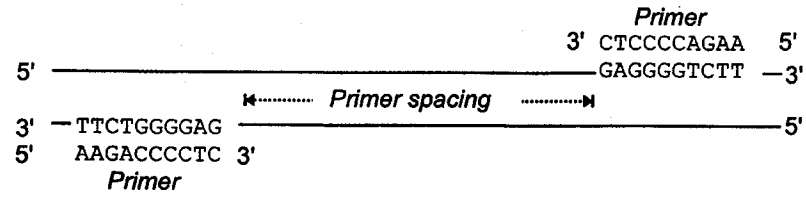


FIGURE 13.4. Random amplication of polymorphic DNA (RAPD) by PCR.

matching failure may occur only when two or more mutations have occurred. Second, when this technique is applied to diploid organisms, the difficulty of distinguishing between homozygotes and heterozygotes introduces some problems in statistical analysis, which will be discussed later. Third, under certain circumstances, experimental results are not reproducible (Hadrys et al. 1992). However, this failure of reproducibility depends on experimental conditions, and careful work seems to enhance the reproducibility. Taberner et al. (1997) reported that in their experiment, 3,396 bands (99.2%) out of the 3,422 bands reexamined were reproducible. Furthermore, Tibayrenc et al. (1993) state that even if only reproducible RAPD data are used, useful information can be obtained for a study of genetic differentiation of populations. Note also that the utility of this technique depends on the organism studied. In bacterial species, which are haploid and reproduce asexually, this technique seems to be quite useful (Wang et al. 1993; Whittam 1995). In this book, we consider only statistical methods for analyzing RAPD data.

Number of Nucleotide Substitutions Between Two Genomes

We first consider haploid organisms with asexual reproduction or selfing diploid organisms, in which the unit of inheritance is the entire genome, since there is virtually no recombination between different chromosomes or genes. In these organisms, a locus is identified by the presence of an electrophoretic band, and if every individual has this band, the locus is regarded as monomorphic. If the band at a locus is present in some individuals but not in others, this locus is polymorphic, and the individuals having the band are recorded as "1" and those lacking the band as "0". Therefore, if n bands are identified when m individuals are studied, we will have a matrix of element 1 or 0 arranged in m rows and n columns. It is then possible to estimate the number of nucleotide substitutions per site (d) between two individuals, X and Y. As indicated by Clark and Lanigan (1993), a simple way of estimating d is to use Equations (13.31)–(13.33), which were originally developed for restriction fragment data. Here n_X and n_Y in Equation (13.31) represent the numbers of RAPD fragments observed in individuals X and Y, respectively, and n_{XY} is the number of fragments shared by X and Y.

In most experimental studies, about ten or less bands are scored for each individual when a primer of about ten nucleotides is used (Taberner et al. 1997). As in the case of restriction-fragment data (section 13.2), two

haploid individuals with the shared genomic sequences will have the same set of DNA fragments, but as the extent of nucleotide differences increases, the proportion (F) of identical DNA fragments will decline.

Using the same mathematical argument as that used by Nei and Li (1979) for restriction fragments, Clark and Lanigan (1993) have shown that the expected value of F is given by Equation (13.30) and therefore G and d can be estimated by Equations (13.32) and (13.33), respectively. In this case, b in Equation (13.33) represents the number of nucleotides in the primer site. In practice, of course, several primers with the same number of nucleotides are used for each individual, and the $\hat{F}$ value between two individuals is computed by Equation (13.31).

Nucleotide Diversity and Nucleotide Divergence

Haploid or Inbred Populations

In haploid or selfing diploid organisms, the nucleotide diversity (π) and nucleotide divergence (d_A) can also be estimated in the same way as that for restriction fragment data. Thus, $\hat{\pi}$ for one population can be computed by Equation (13.38) if we redefine n_{ij}, $n_{i(j)}$, and $n_{(i)j}$ in Equation (13.35) in terms of the numbers of RAPD fragments. Similarly, d_{XY} and d_A can be estimated by the same procedure.

Nei and Takezaki (1994), however, developed another method to estimate d_A. In haploid or selfing diploid populations, the frequencies of the allele (A) with the electrophoretic band and the allele (B) without the band at a locus can easily be determined. Let p_{Xi} and p_{Yi} be the frequencies of allele A at the i-th locus in populations X and Y, respectively. If we choose one haploid genome from each of populations X and Y, respectively, the probability of obtaining the same band (fragment) at locus i is $p_{Xi}p_{Yi}$. Therefore, the expected number of the shared bands for all loci is

$$E(n_{XY}) = \sum_i p_{Xi}p_{Yi} \tag{13.45}$$

The expected numbers of bands at the i-th locus in populations X and Y are p_{Xi} and p_{Yi} per haploid genome. Therefore, the expected numbers for all loci for X and Y are $E(n_X) = \Sigma_i p_{Xi}$ and $E(n_Y) = \Sigma_i p_{Yi}$, respectively. The proportion of shared bands between the two populations is then given by

$$Q_{XY} = \frac{2\sum_i p_{Xi}p_{Yi}}{\sum_i p_{Xi} + \sum_i p_{Yi}} \quad \text{or} \quad Q_{XY} = \frac{\sum_i p_{Xi}p_{Yi}}{\sqrt{\sum_i p_{Xi}\sum_i p_{Yi}}} \tag{13.46}$$

Here we have assumed that $\Sigma_i p_{Xi}$ and $\Sigma_i p_{Yi}$ are more or less the same. In fact, if the populations are in mutation-drift equilibrium, $E(n_X) = E(n_Y)$.

However, Equation (13.46) includes the effect of the within-population variation that existed at the time of the divergence between populations X and Y, and to correct for this effect we must normalize it by the proportions of shared bands within populations. The proportions of shared

bands between two randomly chosen haploid genomes within populations X and Y (Q_X and Q_Y, respectively) are given by

$$Q_X = \sum_i p_{Xi}^2 / \sum_i p_{Xi} \tag{13.47}$$

$$Q_Y = \sum_i p_{Yi}^2 / \sum_i p_{Yi} \tag{13.48}$$

Therefore, the normalized proportion of shared bands becomes

$$Q = \frac{\sum_i p_{Xi} p_{Yi}}{\sqrt{\sum_i p_{Xi}^2 \sum_i p_{Yi}^2}} \tag{13.49}$$

if we use the second formula in Equation (13.46) for Q_{XY} (Nei and Takezaki 1994).

This Q can be estimated by replacing p_{Xi}'s and p_{Yi}'s by their sample estimates $\hat{p}_{Xi}$'s and $\hat{p}_{Yi}$'s, respectively. Once the estimate ($\hat{Q}$) of Q is obtained, we can estimate G by Equation (13.32) and d_A by Equation (13.33).

Equation (13.49) is similar to the gene identity (I) in Equation (13.14) and becomes 1 when the two populations are identical with $p_{Xi} = p_{Yi}$ (at time 0) and 0 when the populations do not have any shared bands. Note also that Q may be expressed as $Q_{XY}/(Q_X Q_Y)^{1/2}$. Therefore, Q_X, Q_Y, and Q_{XY} correspond to J_X, J_Y, and J_{XY} in Equation (13.14), respectively. Clark and Lanigan's (1993) estimate of DNA sequence divergence corresponds to $\hat{d}_{XY}$ rather than $\hat{d}_A$. An approximate value of $\hat{d}_{XY}$ may be obtained from Q_{XY} rather than Q.

Equation (13.49) depends on the assumption that even a single mutation in a primer site prevents the matching of a primer and its primer site. If this assumption is incorrect, we will have to redefine the b value. In practice, it is difficult to know the real b value, but even if the b value used is incorrect, d_A can still be used as a relative distance measure. It should also be noted that although the above estimate of d_A has a good theoretical basis, it has a relatively large variance. Therefore, for the purpose of construction of phylogenetic trees $D_Q = 1 - Q$ may give better results. One may also use Q_{XY} or $D_{QXY} = 1 - Q_{XY}$.

Diploid Populations

In sexually reproducing diploid organisms, it is not straightforward to estimate even the number of nucleotide substitutions between two genomes, because a band on a gel may represent either a homozygote or a heterozygote. To estimate the relative frequencies of homozygotes and heterozygotes for a band, we need to study a substantial number of individuals and assume Hardy-Weinberg proportions. In this case, therefore, only nucleotide diversity (π) and nucleotide divergence between populations (d_A) can be estimated. Of course, in random mating populations it is meaningless to estimate the number of nucleotide substitutions between two genomes, because the genome of an individual is subject to recombination and is not the unit of inheritance.

Realizing this problem, Clark and Lanigan (1993) developed a new

method for estimating n_{XY}'s, n_X's, and n_Y's for random mating populations. Their formulas are quite complicated and seem to give large variances of the estimates of π and d_{XY}. Nei and Takezaki (1994) suggested that Equations (13.46) and (13.49) be used for estimating d_{XY} and d_A. In diploid organisms, p_i cannot be directly determined but can be estimated by $1 - z_i^{1/2}$, where z_i is the frequency of individuals without the band (00) at the i-th locus. For the estimate of p_i to be reliable, the number of individuals examined must be large. However, Clark and Lanigan's computer simulation suggests that even if the number of individuals examined in a population is about 10, this estimate of p_i gives rather good estimates of d_{XY} (and d_A) when the number of loci examined is large. We therefore recommend the use of Equations (13.46) and (13.49) with this estimate of p_i.

Gene Diversity Analysis in Subdivided Populations

Lynch and Milligan (1994) developed a number of ways of analyzing gene diversity using RAPD data. The theoretical basis of their analysis is similar to that presented in chapter 12, and two alleles (A and B) at each locus are considered. Unfortunately, their statistical methods are quite complicated, so we do not present them here. Instead, we would like to indicate that a similar gene diversity analysis can be done rather easily if we use Q_X and Q_Y defined in Equations (13.47) and (13.48).

We have already indicated the similarity between Q_X and J_X. Therefore, the extent of genetic variability of a population may be measured by

$$K_X = 1 - Q_X \tag{13.50}$$

When there are many subpopulations, we can compute K_X for each subpopulation for many primers and take the average for all subpopulations, which we denote by K_S. As in the case of usual gene diversity analysis, we can also compute the total gene diversity (K_T) by using average allele frequencies ($\bar{p}_i$) over all populations. The extent of population differentiation can then be measured by the following **coefficient of population differentiation.**

$$R_{ST} = (K_T - K_S)/K_T \tag{13.51}$$

In practice, of course, we have to estimate K_S and K_T from sample allele frequencies.

14

Perspectives

We have discussed various statistical methods that are useful for the study of molecular evolution. We hope these methods will help researchers to analyze their data objectively and obtain reasonable answers to various problems of evolution. Needless to say, current statistical methods are not necessarily perfect, and many of them need further refinements. New statistical methods will also be required to deal with new biological problems that will surely arise in the future. In this chapter, we present an overview on the use of statistical methods in molecular evolutionary genetics with an emphasis on future studies.

14.1. Statistical Methods

Although most of the statistical methods presented in this book are currently used by a large number of investigators, their theoretical foundations are not necessarily well established, and there is an urgent need to clarify the foundations. This is particularly so with the methods of constructing and testing phylogenetic trees. The statistical problems concerned with molecular evolution are much more complicated than those considered in traditional statistics, and it is important to take into account detailed aspects of the evolutionary change of DNA and protein sequences in developing efficient statistical methods. Application of existing statistical principles without consideration of the biological processes involved may lead to inappropriate statistical methods. Statistical methods that are too sophisticated with many parameters may also give poor results in actual data analysis, because there are always some uncontrollable elements involved in the evolution of DNA or protein sequences.

Recent progress in computer technology has been spectacular, and this has led many authors to write computer programs for analyzing molecular data. These computer programs are now essential for the study of molecular evolution. In the future, more convenient programs will be developed as the technology advances further, and molecular evolutionists will be heavily dependent on computers. While this progress in computational capability will facilitate the study of molecular evolution enor-

mously, this may create a situation in which the computer becomes a black box; researchers may just feed their data into a computer and receive final results of a statistical analysis without knowing how the data are analyzed. If this happens, researchers may reach erroneous conclusions whenever they accidentally use wrong options in the computer program. It is therefore important to know the details of the statistical methods used in each computer program.

As mentioned in the previous chapters, it is common that there are several methods of analyzing molecular data for the same purpose and some investigators prefer a particular method to another. For this reason, many computer program packages include several options so that the users may be able to choose their preferred method. Although this practice is useful for many investigators, some programs include obsolete statistical methods or methods whose validity has been questioned in the literature. It is therefore important for the user to choose a method that is most appropriate for their data analysis. Of course, if there is controversy over the method to be used, it is advisable to try several other methods and derive the most reasonable conclusion.

14.2. Genome Projects

In the past, relatively small data sets have been used in the study of molecular evolution, simply because it has been difficult for a single researcher or a single research group to obtain large data sets. This situation will soon change. It is now possible to use automated sequencing machines to generate a large amount of sequence data. Total genome sequencing has also been initiated in many different organisms. In bacteria, there are already about 20 species for which the complete genome sequence is available, and many other species are being studied. The complete sequence is also available for some eukaryotic species such as the yeast *Saccharomyces cerevisiae,* and the nematode *Caenorhabditis elegans,* and the genome projects in *Drosophila melanogaster,* mice, and humans will soon be completed. The genome projects for economically important animals and plants such as cattle and rice have also been initiated. Although the primary purpose of these genome projects is to use genetic information contained in the genome for enhancing human welfare, this will give a good opportunity for evolutionists to study various problems in evolution.

In fact, genome projects are expected to resolve various controversies concerning molecular evolution. As mentioned in chapter 1, gene duplication is one of the most important mechanisms of evolution and is responsible for the increase of the number of genes from bacteria to higher organisms. The bacterial species *Mycoplasma genitalium* has one of the smallest genome sizes and contains about 470 genes (NCBI, *http://www.ncbi.nlm.gov/COG/*), whereas the mammalian genome has been estimated to have about 70,000 genes (Fields et al. 1994). The large number of genes in higher organisms has been brought about by genome or tandem (or individual) gene duplication (Ohno 1967, 1970; Nei 1969). However, the relative importance of these two processes still remains un-

clear and is hotly debated at the present time (Endo et al. 1997; Kasahara 1997; Hughes 1999). Theoretically, this problem can be solved by phylogenetic analysis of DNA sequences, but currently available computer programs cannot handle this problem properly, because many orthologous and paralogous sequences have to be examined simultaneously.

Another problem is the evolution of multigene families such as the ribosomal RNA and immunoglobulin gene families. A number of authors (e.g., Smith 1974; Hood et al. 1975; Ohta 1980; Zimmer et al. 1980; Arnheim 1983) proposed that these multigene families are generally subject to **concerted evolution,** and the member genes are homogenized by unequal crossing over or gene conversion. However, Nei and colleagues (Nei and Hughes 1992; Nei et al. 1997a) have challenged this view, showing that at least the immunoglobulin and the major histocompatibility complex gene families are subject to evolution by a **birth-and-death process.** In this model of evolution, new genes are created by repeated gene duplication, and some duplicate genes are maintained in the genome for a long time, but others are deleted or become nonfunctional by deleterious mutations (see also Klein et al. 1993; Michelmore and Meyers 1998; Robertson 1998). These two hypotheses are not mutually exclusive, but it is important to determine the general pattern of evolution of multigene families. This problem can be studied by phylogenetic analysis of DNA sequences, but it requires a new way of statistical analysis.

One of the important problems in current genome analysis is the identification of the function of potential genes or open reading frames revealed by DNA sequencing (Chervitz et al. 1998). Although the function should eventually be determined by biochemical experiments, it can be inferred by finding their homologous genes in other organisms. However, this homology search is not always easy when the sequence similarity level is low or when there are many similar genes. One way to facilitate this homology search would be to conduct a phylogenetic analysis of all related genes from many different organisms (Eisen 1998; Pellegrini et al. 1999). At the present time, however, the statistical technique that can be used for this purpose is not well developed.

Genome projects are also expected to contribute to the study of phylogenetic relationships of organisms. At the present time, molecular phylogenies are studied by using a small number of genes, and it is often difficult to have a definitive conclusion. For example, there are 20 different orders of placental mammals (Novacek 1992), but the evolutionary relationships of these orders are largely unknown. The merit of genome projects is that they will provide information on duplicate genes, large DNA deletions/insertions, gene translocations, transposon insertions, and so forth, as well as on DNA sequences. Many of these genomic changes can be used as derived shared characters for clarifying the evolutionary histories of different taxa (chapter 7). If there are a sufficient number of shared derived characters with no or few reversible changes, they will facilitate the study of phylogenies enormously. Of course, this does not mean that DNA sequence data are not needed for phylogenetic inference in the future. Rather, sequence data will always be useful, because approximate phylogenetic relationships can be quickly found by sequence data and they would be sufficient for many closely related taxa. DNA se-

quence data are also essential for estimating the times of divergence for various groups of organisms. If we use a large number of genes, the accuracy of the estimates of divergence times is expected to increase (Doolittle et al. 1996; Hedges et al. 1996; Kumar and Hedges 1998).

As mentioned earlier, the complete genome sequence is already available for many bacterial species, and their phylogenetic relationships are now under investigation. Unfortunately, a substantial amount of interspecific gene transfer seems to have occurred between different bacterial species so that it is not easy to determine their phylogenetic relationships (Koonin et al. 1997; Woese 1998). In higher organisms, however, the extent of interspecific gene transfer is apparently very low. Therefore, we will probably be able to determine the evolutionary relationships of these organisms.

14.3. Molecular Biology and Evolution

In the last four decades, the study of evolution has been strongly influenced by the development and progress of molecular biology. Every time a new concept in molecular biology is developed or a new molecular technique is introduced, evolutionists have adopted it to enhance the analytical power of studying evolution. In contrast, the progress of molecular biology has rarely been affected by new developments in evolutionary biology. This situation is now changing, and molecular biologists have started to use the evolutionary approach to have a better understanding of molecular biology. This is particularly true in the study of developmental biology and immunology. In these areas, interspecific comparison of genes and gene regulation systems has already contributed to the understanding of the mechanisms of morphogenesis (Carroll 1995; Gehring 1996) and vertebrate immune systems (Klein and Horejsi 1997), and it is now a common practice for molecular biologists to construct phylogenetic trees to find the orthologous and paralogous genes from different organisms.

There are usually a large number of genes involved in the development of a particular morphological character or in the function of an immune system, and these genes apparently coevolved as parts of an interactive genetic system. There are also several different biochemical pathways in the development of a specific morphological character (e.g., eyes and wings) or several different mechanisms of defending hosts from invading parasites. How these systems or mechanisms have evolved is largely unknown at the present time. However, it is obvious that to attack this problem, interdisciplinary research between molecular biology and evolutionary biology is necessary.

Molecular biology has played a central role in unifying various branches of biology. We have seen that all life processes such as development, physiology, and reproduction are essentially the same for all organisms at the molecular level. Another important discipline that unifies all areas of biology is evolutionary biology. It is now clear that all life forms on Earth are descended from a single progenote (ancestor for all organisms) that existed some 4 billion years ago and thus all organisms

are historically related, as speculated by Charles Darwin (1859). The extensive biodiversity on Earth is a product of mutation and natural selection that enabled different groups of organisms to adapt to different environmental niches. However, the evolutionary relationships of different organisms are still poorly understood. The detailed mechanisms of evolution that produced sophisticated organisms such as mammals and flowering plants are also unclear. These problems are now being studied by molecular evolutionists using a quantitative approach. It is in this context that the statistical and computational study of molecular evolution plays an important role in biological sciences.

Appendices

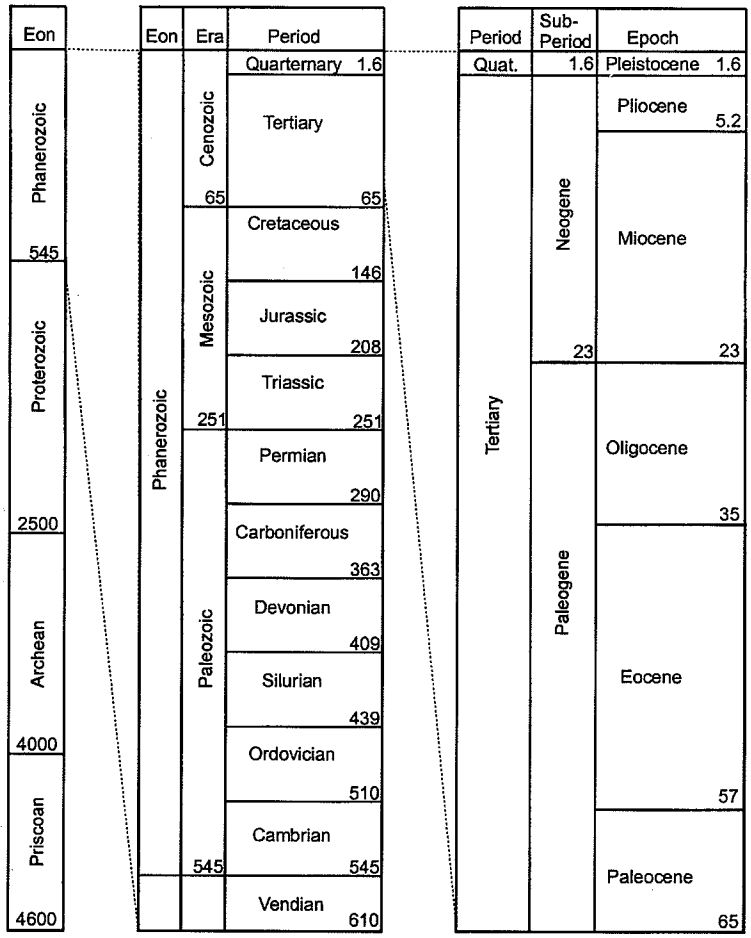
A. Mathematical Symbols and Notations

Symbol	Description
α	Rate of nucleotide (gene) substitution; transition rate
β	Transversion rate
μ	Mutation rate per nucleotide or amino acid site
π	Nucleotide diversity
θ	G + C content, $4N\mu$
$(\delta\mu)^2$	$(\delta\mu)^2$ distance
χ^2	Chi-square statistic
π_j	Frequency of the j-th codon in a sequence
π_S	Nucleotide diversity within populations
π_T	Nucleotide diversity for the entire population
a	Gamma parameter
b	Branch length
C	Number of codons in a gene
CI	Consistency index
d	Number of nucleotide or amino acid substitutions per site
D	Nei's standard distance
d_A	Net nucleotide divergence between two populations
D_A	D_A distance
d_C	Chord distance
d_G	Gamma distance
D_L	Latter's distance
d_N	Number of nonsynonymous substitutions per nonsynonymous site
d_R	Grishin distance for amino acid sequences
D_R	Roger's distance
d_S	Number of synonymous substitutions per synonymous site
d_X	Nucleotide diversity in a population
d_{XY}	Nucleotide divergence between two populations

Symbol	Description
F	Fixation index, proportion of shared DNA fragments
F_{IS}, F_{IT}	Fixation indices
F_{ST}	Fixation index
G_{ST}	Coefficient of gene differentiation; fixation index
g_i	Nucleotide or amino acid frequencies in sequences
h	Single-locus heterozygosity
H	Average heterozygosity, gene diversity
I	Genetic identity between two populations
J_X	Genetic identity of genes within a population
k	Transition/transversion rate ratio
K_X	Extent of genetic variability in a population in RAPD analysis
L	Number of genetic locii; tree length in parsimony analysis; likelihood value
LR	Likelihood ratio test statistic
m	Number of sequences; number of individuals examined; migration rate per generation
n	Number of nucleotides or amino acids in a sequence
N	Number of nonsynonymous sites per sequence; effective population size
N_{ST}	Coefficient of nucleotide differentiation
p	Proportion of different nucleotides or amino acids between two sequences
P	Proportion of sites with transitional difference
PAM	Accepted point mutations
P_B	Bootstrap confidence value (bootstrap value)
P_C	Confidence probability (1 − Type I error) that the length of a branch in a tree is non-negative

Symbol	Description	Symbol	Description
$P_{ij}(t)$	Probability that nucleotide i at time 0 changes to j at time t	RI	Retention index
p_N	Proportion of nonsynonymous differences per nonsynonymous site	$RSCU$	Relative synonymous codon usage
p_S	Proportion of synonymous differences per synonymous site	R_{ST}	Coefficient of population differentiation in RAPD analysis
p_s	Number of segregating sites per site	s	Number of transitional substitutions per site; number of subpopulations
q	Proportion of identical nucleotides or amino acids; number of polymorphic alleles	S	Number of synonymous sites per sequence; sum of branch lengths of a tree; number of segregating sites per sequence; proportion of shared restriction sites
Q	Proportion of sites with transversional difference; normalized proportion of shared bands between two populations	t, T	Divergence time
		TL	Tree length in parsimony analysis
r	Rate of nucleotide or amino acid substitution	v	Number of transversional substitutions per site; mutation rate per locus; expected number of substitutions
R	Transition/transversion ratio; relative optimality score	x_i	Frequency for allele i
		Z	Normal deviate test statistic

B. Geological Timescale



C. Geological Events in the Cenozoic and Mesozoic Eras

Geological Event	Time (mya)
Pangea	180
Gondwanaland (India + Australia + Antarctica + Africa + S. America) and Laurasia (N. America + Greenland + Eurasia) separate	160–170
Gondwana splits into two parts: east (India + Australia + Antarctica) and west (Africa + S. America)	120–148
India separates from Australia + Antarctica	130–140
Eastern and Western N. Americas separate by waterway	105–120
Africa and S. America separate	95–105
India and Madagascar separate	88–95
Australia separates from Antarctica	80–85
Terminal Cretaceous Extinction	65
Disappearance of the waterway separating east and west of N. America	60–70
India and Eurasia collision begins	50–55
Panama collides with North West Colombia	7
Kauai island formation	5.1
Oahu island formation	4
Formation of the Isthmus of Panama	3.5
Hawaii island formation	0.5–0.8

Note: Data from Harland (1989) and Smith (1994).

D. Organismal Evolution Based on the Fossil Record

Era	Period	Time (my)	Fossils
Cenozoic (65)	Quarternary (1.6)	0.0–1.6	Humans evolve
	Tertiary (63.4)	1.6–65	Early Anthropoids appear Rodents, primates appear Flowering plants found Modern birds appear
Mesozoic (180)	Cretaceous (81)	65–146	Extinction of dinosaurs Early grasses appear Flowering plants proliferate Marsupials appear
	Jurassic (62)	146–208	First birds appear Early mammals appear Teleost fish radiate Giant dinosaurs appear Early flowering plants
	Triassic (37)	208–245	Conifers dominate Therapsid reptiles diversify

continued

Era	Period	Time (my)	Fossils
Paleozoic (300)	Permian (45)	245–290	Extinction of Trilobites
	Carboniferous (73)	290–363	Conifers appear Early reptiles emerge Amphibians and insects develop Gymnosperms appear
	Devonian (46)	363–409	Shark fossils First land vertebrates appear Wingless insects
	Silurian (30)	409–439	Land plants appear Oldest fungal fossils Terrestrial invertebrates appear
	Ordovician (71)	439–510	First land plants Invertebrate fauna proliferates
	Cambrian (35)	510–545	Emergence of fish First fossil evidence for most animal phyla

Note: Data from Benton (1993) and Graham (1993).

References

Adachi, J. and M. Hasegawa. (1996a) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42: 459–468.

Adachi, J. and M. Hasegawa. (1996b) *MOLPHY: Programs for molecular phylogenetics.* Institute of Statistical Mathematics, Tokyo.

Akaike, H. (1974) A new look at the statistical model identifications. *IEEE Trans. Automat. Contr.* AC-19: 716–723.

Akashi, H. (1994) Synonymous codon usage in *Drosophila melanogaster:* Natural selection and translational accuracy. *Genetics* 136: 927–935.

Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139: 1067–1076.

Andersson, S. G. E., D. R. Stothard, P. Fuerst and C. G. Kurland. (1999) Molecular phylogeny and rearrangement of rRNA genes in *Rickettsia* species. *Mol. Biol. Evol.* 16: 987–995.

Archie, J. W. (1985) Statistical analysis of heterozygosity data: Independent sample comparison. *Evolution* 39: 623–637.

Archie, J. W. (1989) A randomization test for phylogenetic information in systematic data. *Syst. Zool.* 28: 239–252.

Arnheim, N. (1983) Concerted evolution of multigene families. In *Evolution of genes and proteins* (M. Nei and R. K. Koehn, eds.), pp. 38–61. Sinauer Associates, Sunderland, MA.

Atchley, W. R., W. M. Fitch and M. Bronner-Fraser. (1994). Molecular evolution of the MyoD family of transcription factors. *Proc. Natl. Acad. Sci. USA* 91: 11,522–11,526.

Avise, J. C. (1994) *Molecular markers, natural history and evolution.* Chapman & Hall, New York.

Avise, J. C. and C. F. Aquadro. (1982) A comparative summary of genetic distances in the vertebrates: Patterns and correlations. *Evol. Biol.* 15: 151–185.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel et al. (1987) Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. *Annu. Rev. Ecol. Syst.* 18: 489–522.

Avise, J. C., R. A. Lansman and R. O. Shade. (1979) The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. I. Population structure and evolution in the genus *Peromyscus. Genetics* 92: 279–295.

Ayala, F. J. (1975) Genetic differentiation during the speciation process. *Evol. Biol.* 8: 1–78.

Backeljau, T., L. DeBruyn, H. DeWolf, K. Jordaens, S. Van Dongen and B. Winnepenninckx. (1996) Multiple UPGMA and neighbor-joining trees

and the performance of some computer packages. *Mol. Biol. Evol.* 13: 309–313.

Bandelt, H.-J., P. Forster, B. C. Sykes and M. B. Richards. (1995) Mitochondrial portraits of human populations using median networks. *Genetics* 141: 743–753.

Barnard, E. A. (1969) Biological function of pancreatic ribonuclease. *Nature* 221: 340–344.

Barry, D. and J. A. Hartigan. (1987) Asynchronous distance between homologous DNA sequences. *Biometrics* 43: 261–276.

Benton, M. J. (1993) *The fossil record 2.* Chapman & Hall, New York.

Bernardi, G. (1995) The human genome: Organization and evolutionary history. *Annu. Rev. Genet.* 29: 445–476.

Bernardi, G., D. Mouchiroud, C. Gautier and G. Bernardi. (1988) Compositional patterns in vertebrate genomes: Conservation and change in evolution. *J. Mol. Evol.* 28: 7–18.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny et al. (1985) The mosaic genome of warm-blooded vertebrates. *Science* 228: 953–958.

Berry, V. and O. Gascuel. (1996) On the interpretation of bootstrap trees: Appropriate threshold of clade selection and induced gain. *J. Mol. Evol.* 13: 999–1011.

Beverley, S. M. and A. C. Wilson. (1985) Ancient origin for Hawaiian Drosophilinae inferred from protein comparisons. *Proc. Natl. Acad. Sci. USA* 82: 4753–4757.

Bhattacharyya, A. (1946) On a measure of divergence between two multinomial populations. *Sankhya* 7: 401–406.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger and D. C. Wiley. (1987a) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* 329: 512–518.

Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger and D. C. Wiley. (1987b) Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* 329: 506–512.

Blanken, R. L., L. C. Klotz and A. G. Hinnebusch. (1982) Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. *J. Mol. Evol.* 19: 9–19.

Bonatto, S. L. and F. M. Salzano. (1997) Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *Am. J. Hum. Genet.* 61: 1413–1423.

Botstein, D., R. L. White, M. Skolnick and R. W. Davis. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32: 314–331.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza. (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.

Bowmaker, J. K. (1991) The evolution of vertebrate visual pigments and photoreceptors. In *Evolution of the eye and visual system* (J. R. Cronly-Dillon and R. L. Gregory, eds.), pp. 63–81. CRC Press, Boca Raton, FL.

Bridges, C. B. (1936) Genes and chromosomes. *Teaching Biology* 11: 17–23.

Britten, R. J. (1986) Rates of DNA sequence evolution differ between taxonomic groups. *Science* 231: 1393–1398.

Britten, R. J., W. F. Baron, D. B. Stout and E. H. Davidson. (1988) Sources and evolution of human *Alu* repeated sequences. *Proc. Natl. Acad. Sci. USA* 85: 4770–4774.

Brookfield, J. F. and P. M. Sharp. (1994) Neutralism and selectionism face up to DNA data. *Trends Genet.* 10: 109–111.

Brown, W. M., G. Matthew and A. C. Wilson. (1979) Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 76: 1967–1971.

Brown, W. M., E. M. Prager, A. Wang and A. C. Wilson. (1982) Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *J. Mol. Evol.* 18: 225–239.

Bryant, D. J. (1997) Building trees, hunting for trees and comparing trees—Theory and method in phylogenetic analysis. Ph.D. dissertation. University of Canterbury, Christchurch, New Zealand.

Bryant, D. J. and P. Waddell. (1998) Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Mol. Biol. Evol.* 15: 1346–1359.

Bulmer, M. (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol. Biol. Evol.* 8: 868–883.

Buri, P. (1956) Gene frequency in small populations of mutant *Drosophila*. *Evolution* 10: 357–402.

Camin, J. H. and R. R. Sokal. (1965) A method for deducing branching sequences in phylogeny. *Evolution* 19: 311–326.

Cann, R. L., M. Stoneking and A. C. Wilson. (1987) Mitochondrial DNA and human evolution. *Nature* 325: 31–36.

Cao, Y., J. Adachi and M. Hasegawa. (1994a) Eutherian phylogeny as inferred from mitochondrial DNA sequence data. *Jpn. J. Genet.* 69: 455–472.

Cao, Y., J. Adachi and M. Hasegawa. (1998) Comment on the quartet puzzling method for finding maximum-likelihood tree topologies. *Mol. Biol. Evol.* 15: 87–89.

Cao, Y., J. Adachi, A. Janke, S. Pääbo and M. Hasegawa. (1994b) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *J. Mol. Evol.* 39: 519–527.

Carroll, S. B. (1995) Homeotic genes and the evolution of arthropods and chordates. *Nature* 376: 479–485.

Cavalli-Sforza, L. L. and A. W. F. Edwards. (1964) Analysis of human evolution. In *Proc. 11th Intl. Cong. Genet.*, pp. 923–933. Pergamon, New York.

Cavalli-Sforza, L. L. and A. W. F. Edwards. (1967) Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.* 19: 233–257.

Chakraborty, R. (1977) Estimation of time of divergence from phylogenetic studies. *Can. J. Genet. Cytol.* 19: 217–223.

Chakraborty, R. and H. Danker-Hopfe. (1991) Analysis of population structure: A comparative study of different estimators of Wright's fixation indices. In *Handbook of statistics.* Vol. 8, *statistical methods in biological and medical sciences* (C. R. Rao and R. Chakraborty, eds.), pp. 203–254. Elsevier Science, New York.

Chandrasekharan, U. M., S. Sanker, M. J. Glynias, S. S. Karnik and A. Husain. (1996) Angiotensin II-forming activity in a reconstructed ancestral chymase. *Science* 271: 502–505.

Chen, Z. W., S. N. McAdam, A. L. Hughes, A. L. Dogon, N. L. Letvin and D. I. Watkins. (1992) Molecular cloning of orangutan and gibbon MHC class I cDNA. The HLA-A and -B loci diverged over 30 million years ago. *J. Immunol.* 148: 2547–2554.

Chervitz, S. A., L. Aravind, G. Sherlock, C. A. Ball, E. V. Koonin, S. S. Dwight et al. (1998) Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* 282: 2022–2028.

Clark, A. G. and C. M. S. Lanigan. (1993) Prospects for estimating nucleotide divergence with RAPDs. *Mol. Biol. Evol.* 10: 1096–1111.

Cockerham, C. C. (1969) Variance of gene frequencies. *Evolution* 23: 72–84.

Cockerham, C. C. (1973) Analyses of gene frequencies. *Genetics* 74: 679–700.

Comeron, J. M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* 41: 1152–1159.

Comeron, J. M. and M. Aguade. (1998) An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* 47: 268–274.

Covello, P. S. and M. W. Gray. (1993) On the evolution of RNA editing. *Trends Genet.* 9: 265–268.

Coyne, J. A. and A. A. Felton. (1977) Genetic heterogeneity at two alcohol dehydrogenase loci in *Drosophila pseudoobscura* and *Drosophila persimilis*. *Genetics* 87: 285–304.

Crow, J. F. and K. Aoki. (1984) Group selection for a polygenic behavioral trait: Estimating the degree of population subdivision. *Proc. Natl. Acad. Sci. USA* 81: 6073–6077.

Crow, J. F. and M. Kimura. (1970) *An introduction to population genetics theory.* Harper & Row, New York.

Cummings, M. P., S. P. Otto and J. Wakeley. (1995) Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* 12: 814–822.

Darwin, C. (1859) *On the origin of species.* Murray, London.

Dayhoff, M. O. (1972) *Atlas of protein sequence and structure.* National Biomedical Research Foundation, Silver Springs, MD.

Dayhoff, M. O., R. E. Dayhoff and L. T. Hung. (1976) Composition of proteins. In *Atlas of protein sequence and structure* (M. O. Dayhoff, ed.), pp. 301–310. National Biomedical Research Foundation, Silver Springs, MD.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt. (1978) A model of evolutionary change in proteins. In *Atlas of protein sequence and structure* (M. O. Dayhoff, ed.), pp. 345–352. National Biomedical Research Foundation, Silver Spring, MD.

Dean, A. M. and G. B. Golding. (1997) Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proc. Natl. Acad. Sci. USA* 94: 3104–3109.

DeBry, R. W. (1992) The consistency of several phylogeny-inference methods under varying evolutionary rates. *Mol. Biol. Evol.* 9: 537–551.

DeBry, R. W. and N. A. Slade. (1985) Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. *Sys. Zool.* 34: 21–34.

Deka, R., L. Jin, M. D. Shriver, L. M. Yu, S. DeCroo, J. Hundrieser et al. (1995) Population genetics of dinucleotide $(dC\text{-}dA)_n$. $(dG\text{-}dT)_n$ polymorphisms in world populations. *Am. J. Hum. Genet.* 56: 461–474.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin and N. B. Freimer. (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91: 3166–3170.

Dickerson, R. E. (1971) The structure of cytochrome c and the rates of molecular evolution. *J. Mol. Evol.* 1: 26–45.

Doolittle, R. F. (1994) Convergent evolution: The need to be explicit. *Trends Biochem. Sci.* 19: 15–18.

Doolittle, R. F. and B. Blombäck. (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: Evolutionary implications. *Nature* 202: 147–152.

Doolittle, R. F., D.-F. Feng, S. Tsang, G. Cho and E. Little. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271: 470–477.

Dopazo, J. (1994) Estimating errors and confidence intervals for branch lengths in phylogenetic trees by a bootstrap approach. *J. Mol. Evol.* 38: 300–304.

Dopazo, J. and J. M. Carazo. (1996) Phylogenetic reconstruction using an unsupervised growing neutral network that adopts the topology of a phylogenetic tree. *J. Mol. Evol.* 44: 226–233.

Duboule, D. (1994) *Guidebook to the homeobox genes.* Oxford University Press, New York.

Eanes, W. F., M. Kirchner and J. Yoon. (1993) Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. *Proc. Natl. Acad. Sci. USA* 90: 7475–7479.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, I. N. Wang, M. A. McCartney et al. (1996) Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans. Genetics* 144: 1027–1041.

Easteal, S. (1985) Generation time and the rate of molecular evolution. *Mol. Biol. Evol.* 2: 450–453.

Easteal, S. (1988) Rate constancy of globin gene evolution in placental mammals. *Proc. Natl. Acad. Sci. USA* 85: 7622–7626.
Easteal, S. (1990) The pattern of mammalian evolution and the relative rate of molecular evolution. *Genetics* 124: 165–173.
Easteal, S. (1991) The relative rate of cDNA evolution in primates. *Mol. Biol. Evol.* 8: 115–127.
Easteal, S., C. Collet and D. Betty. (1995) *The mammalian molecular clock.* R. G. Landes, Austin, TX.
Eck, R. V. and M. O. Dayhoff. (1966) *Atlas of protein sequence and structure.* National Biomedical Research Foundation, Silver Springs, MD.
Edwards, A. W. F. and L. L. Cavalli-Sforza. (1963) The reconstruction of evolution. *Heredity* 18: 553.
Efron, B. (1982) *The jackknife, the bootstrap and other resampling plans.* Society for Industrial and Applied Mathematics, Philadelphia, PA.
Efron, B., E. Halloran and S. Holmes. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* 93: 7085–7090.
Efron, B. and R. J. Tibshirani. (1993) *An introduction to the bootstrap.* Chapman & Hall, New York.
Eisen, J. A. (1998) Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8: 163–167.
Eldredge, N. and J. Cracraft. (1980) *Phylogenetic patterns and the evolutionary process.* Columbia University Press, New York.
Elzanowski, A. and J. Ostell. (1996) *The genetic codes.* National Center for Biotechnology Information (NCBI), Bethesda, MD.
Endo, T., T. Imanishi, T. Gojobori and H. Inoko. (1997) Evolutionary significance of intra-genome duplications on human chromosomes. *Gene* 205: 19–27.
Engels, W. R. (1981) Estimating genetic divergence and genetic variability with restriction endonucleases. *Proc. Natl. Acad. Sci. USA* 78: 6329–6333.
Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3: 87–112.
Eyre-Walker, A. (1997) Differentiating between selection and mutation bias. *Genetics* 147: 1983–1987.
Falconer, D. S. (1981) *Introduction to quantitative genetics.* Longman, London.
Farris, J. S. (1969) A successive approximations approach to character weighting. *Syst. Zool.* 18: 374–385.
Farris, J. (1977) Phylogenetic analysis under Dollo's law. *Syst. Zool.* 26: 77–88.
Farris, J. S. (1989) The retention index and the rescaled consistency index. *Cladistics* 5: 417–419.
Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27: 401–410.
Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17: 368–376.
Felsenstein, J. (1982) Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.* 57: 379–404.
Felsenstein, J. (1984) *PHYLIP: Phylogeny inference package, ver. 2.51.* University of Washington, Seattle, WA.
Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791.
Felsenstein, J. (1988) Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* 22: 521–565.
Felsenstein, J. (1992) Estimating effective population size from samples of sequences: Inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet. Res.* 59: 139–147.
Felsenstein, J. (1995) *PHYLIP: Phylogeny inference package, ver. 3.572.* University of Washington, Seattle, WA.

Felsenstein, J. (1997) An alternating least squares approach to inferring phylogenies from pairwise distances. *Syst. Biol.* 46: 101–111.

Felsenstein, J. and G. A. Churchill. (1996) A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13: 93–104.

Felsenstein, J. and H. Kishino. (1993) Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.* 42: 193–200.

Feng, D.-F., G. Cho and R. F. Doolittle. (1997) Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA* 94: 13,028–13,033.

Feng, D.-F. and R. F. Doolittle. (1996) Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. In *Methods in Enzymology* (R. F. Doolittle, ed.), pp. 368–382. Academic, San Diego, CA.

Ferris, S. D., W. M. Brown, W. S. Davidson and A. C. Wilson. (1981a) Extensive polymorphism in the mitochondrial DNA of apes. *Proc. Natl. Acad. Sci. USA* 78: 6319–6323.

Ferris, S. D., A. C. Wilson and W. M. Brown. (1981b) Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 78: 2432–2436.

Fields, C., M. D. Adams, O. White and J. C. Venter. (1994) How many genes in the human genome? *Nat. Hist.* 7: 345–346.

Figueroa, F., E. Günther and J. Klein. (1988) MHC polymorphism pre-dating speciation. *Nature* 335: 265–267.

Fitch, W. M. (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.* 26: 499–507.

Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19: 99–113.

Fitch, W. M. (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* 20: 406–416.

Fitch, W. M. (1976) Molecular evolutionary clocks. In *Molecular evolution* (F. J. Ayala, ed.), pp. 160–178. Sinauer Associates, Sunderland, MA.

Fitch, W. M. (1977) On the problem of discovering the most parsimonious tree. *Am. Nat.* 111: 223–257.

Fitch, W. M. (1981) A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* 18: 30–37.

Fitch, W. M. (1997) Networks and viral evolution. *J. Mol. Evol.* 44: S65–S75.

Fitch, W. M., R. M. Bush, C. A. Bender and N. J. Cox. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* 94: 7712–7718.

Fitch, W. M. and J. S. Farris. (1974) Evolutionary trees with minimum nucleotide replacements from amino acid sequences. *J. Mol. Evol.* 3: 263–278.

Fitch, W. M. and E. Margoliash. (1967) Construction of phylogenetic trees. *Science* 155: 279–284.

Fitch, W. M. and J. Ye. (1991) Weighted parsimony: Does it work? In *Phylogenetic analysis of DNA sequences* (M. M. Miyamoto and J. Cracraft, eds.), pp. 147–154. Oxford University Press, New York.

Forbes, S. H., J. T. Hogg, F. C. Buchanan, A. M. Crawford and F. W. Allendorf. (1995) Microsatellite evolution in congeneric mammals: Domestic and bighorn sheep. *Mol. Biol. Evol.* 12: 1106–1113.

Foutz, R. V. and R. C. Srivastava. (1977) The performance of the likelihood ratio test when the model is incorrect. *Annu. Statis.* 5: 1183–1194.

Fu, Y.-X. (1994) A phylogenetic estimator of effective population size or mutation rate. *Genetics* 136: 685–692.

Fu, Y.-X. and W.-H. Li. (1993) Statistical tests of neutrality of mutations. *Genetics* 133: 693–709.

Fuerst, P. A. and R. D. Ferrell. (1980) The stepwise mutation model: An experimental evaluation utilizing hemoglobin variants. *Genetics* 94: 185–201.

Furano, A. V., B. E. Hayward, P. Chevret, F. Catzeflis and K. Usdin. (1994) Amplification of the ancient murine Lx family of long interspersed repeated DNA occurred during the murine radiation. *J. Mol. Evol.* 38: 18–27.

Galtier, N. and M. Gouy. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA* 92: 11,317–11,321.

Galtier, N. and M. Gouy. (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15: 871–879.

Garza, J. C., M. Slatkin and N. B. Freimer. (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* 12: 594–603.

Gascuel, O. (1994) A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.* 11: 961–963.

Gascuel, O. (1997a) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* 14: 685–695.

Gascuel, O. (1997b) Concerning the NJ algorithm and its unweighted version, UNJ. In *Mathematical hierarchies and biology* (B. Mirkin, F. R. McMorris, F. S. Roberts and A. Rzhetsky, eds.), pp. 149–170. American Mathematical Society, Providence, RI.

Gatesy, J., M. Milinkovitch, V. Waddell and M. Stanhope. (1999) Stability of cladistic relationships between cetacea and higher-level artiodactyl taxa. *Syst. Biol.* 48: 6–20.

Gatesy, J. C. (1997) More DNA support for a Cetacea/Hippopotamidae clade: The blood-clotting protein gene gamma-fibrinogen. *Mol. Biol. Evol.* 14: 537–543.

Gaut, B. S. and P. O. Lewis. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* 12: 152–162.

Gaut, B. S., B. R. Morton, B. C. McCaig and M. T. Clegg. (1996) Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL. Proc. Natl. Acad. Sci. USA* 93: 10,274–10,279.

Gehring, W. J. (1996) The master control gene for morphogenesis and evolution of the eye. *Genes Cells* 1: 11–15.

Gillespie, J. (1991) *The causes of molecular evolution.* Oxford University Press, Oxford, U.K.

Gogarten, J. P., L. Olendzenski, E. Hilario, C. Simon and K. E. Holsinger. (1996) Dating the cenancester of organisms. *Science* 274: 1750–1751.

Gojobori, T., W.-H. Li and D. Graur. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* 18: 360–369.

Golding, G. B. (1983) Estimates of DNA and protein sequence divergence: An examination of some assumptions. *Mol. Biol. Evol.* 1: 125–142.

Goldman, N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36: 182–198.

Goldman, N. and Z. Yang. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 725–736.

Goldstein, D. B. and D. D. Pollock. (1994) Least squares estimation of molecular distance—noise abatement in phylogenetic reconstruction. *Theor. Pop. Biol.* 45: 219–226.

Goldstein, D. B., A. Ruiz-Linares, L. L. Cavalli-Sforza and M. W. Feldman. (1995a) An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139: 463–471.

Goldstein, D. B., A. Ruiz-Linares, L. L. Cavalli-Sforza and M. W. Feldman. (1995b) Microsatellite loci, genetic distances, and human evolution. *Proc. Natl. Acad. Sci. USA.* 92: 6723–6727.

Goodman, M. (1962) Immunochemistry of the primates and primate evolution. *Ann. N.Y. Acad. Sci.* 102: 219–234.

Goodman, M., G. W. Moore and J. Barnabas. (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. *J. Mol. Evol.* 3: 1–48.

Goodwin, R. L., H. Baumann and F. G. Berger. (1996) Patterns of divergence during evolution of α_1-proteinase inhibitors in mammals. *Mol. Biol. Evol.* 13: 346–358.

Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162: 705–708.

Gotoh, O., J.-I. Hayashi, H. Yonekawa and Y. Tagashira. (1979) An improved method for estimating sequence divergence between related DNAs from changes in restriction endonuclease cleavage sites. *J. Mol. Evol.* 14: 301–310.

Graham, L. E. (1993) *Origin of land plants.* John Wiley, New York.

Graur, D. and D. G. Higgins. (1994) Molecular evidence for the inclusion of cetaceans within the order Artiodactyla. *Mol. Biol. Evol.* 11: 357–364.

Graur, D. and W. H. Li. (1999) *Fundamentals of molecular evolution.* Sinauer Associates, Sunderland, MA.

Grimaldi, D. A. (1987) Amber fossil Drosophilidae (Diptera), with particular reference to the Hispaniolan taxa. *Am. Mus. Novitates* 2880: 1–23.

Grishin, N. V. (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* 41: 675–679.

Gu, X. and W.-H. Li. (1992) Higher rates of amino acid substitution in rodents than in humans. *Mol. Phyl. Evol.* 1: 211–214.

Gu, X. and W.-H. Li. (1995) The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty of sequence alignment. *J. Mol. Evol.* 40: 464–473.

Gu, X. and W. H. Li. (1998) Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. *Proc. Natl. Acad. Sci. USA* 95: 5899–5905.

Gu, X. and J. Zhang. (1997) A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.* 15: 1106–1113.

Guo, S. W. and E. A. Thompson. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48: 361–372.

Hadrys, H., M. Balick and B. Schierwater. (1992) Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Mol. Ecol.* 1: 55–63.

Haeckel, E. (1866) *Generelle Morphologie der Organismen.* Georg Riemer, Berlin.

Haldane, J. B. S. (1939) The equilibrium between mutation and random extinction. *Ann. Eugen.* 9: 400–405.

Hamdi, H., H. Nishio, R. Zielinski and A. Dugaiczyk. (1999) Origin and phylogenetic distribution of *Alu* DNA repeats: Irreversible events in the evolution of primates. *J. Mol. Biol.* 289: 861–871.

Hamrick, J. L. and M. J. W. Godt. (1990) Allozyme diversity in plant species. In *Plant population genetics, breeding and genetic resources* (A. H. D. Brown, M. T. Clegg, A. T. Kahler and B. S. Weir, eds.), pp. 43–63. Sinauer Associates, Sunderland, MA.

Hardy, G. H. (1908) Mendelian proportions in a mixed population. *Science* 28: 49–50.

Harland, W. B., R. L. Armstrong, A. V. Cox, L. E. Craig, A. G. Smith and D. G. Smith. (1989) *A geological time scale.* Cambridge University Press, Cambridge, U.K.

Harris, E. E. and J. Hey. (1999) X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. USA* 96: 3320–3324.

Harris, H. (1966) Enzyme polymorphisms in man. *Proc. R. Soc. Lond. B.* 164: 298–310.

Hartigan, J. A. (1973) Minimum evolution fits to a given tree. *Biometrics* 29: 53–65.

Hartl, D. L. and A. G. Clark. (1997) *Principles of population genetics.* Sinauer Associates, Sunderland, MA.

Hasegawa, M. and M. Fujiwara. (1993) Relative efficiencies of the maximum likelihood, maximum-parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phyl. Evol.* 2: 1–5.

Hasegawa, M., Y. Iida, T. Yano, F. Takaiwa and M. Iwabuchi. (1985a) Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. *J. Mol. Evol.* 22: 32–38.

Hasegawa, M., H. Kishino and N. Saitou. (1991) On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* 32: 443–445.

Hasegawa, M., H. Kishino and T. Yano. (1985b) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160–174.

Håstad, O. and M. Björklund. (1998) Nucleotide substitution models and estimation of phylogeny. *Mol. Biol. Evol.* 15: 1381–1389.

Hedges, S. B., P. H. Parker, C. G. Sibley and S. Kumar. (1996) Continental breakup and ordinal diversification of birds and mammals. *Nature* 381: 226–229.

Hedrick, P. W. (1999) Perspective: Highly variable loci and their interpretation in evolution and conservation. *Evolution* 53: 313–318.

Hedrick, P. W. and G. Thomson. (1983) Evidence for balancing selection at HLA. *Genetics* 104: 449–456.

Hein, J. and J. Støvlbæk. (1996) Combined DNA and protein alignment. In *Methods in enzymology* (R. F. Doolittle, ed.), pp. 402–418. Academic, San Diego, CA.

Hendy, M. D. and M. A. Charleston. (1993) Hadamard conjugation: A versatile tool for modeling nucleotide sequence evolution. *New Zealand J. Bot.* 31: 231–237.

Hendy, M. D. and D. Penny. (1982) Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* 59: 277–290.

Hendy, M. D. and D. Penny. (1989) A framework for the quantitative study of evolutionary trees. *Syst. Zool.* 38: 297–309.

Hendy, M. D. and D. Penny. (1993) Spectral analysis of phylogenetic data. *J. Classification* 10: 5–24.

Hendy, M. D., D. Penny and M. A. Steel. (1994) A discrete Fourier analysis of evolutionary trees. *Proc. Natl. Acad. Sci. USA* 91: 3339–3343.

Hennig, W. (1950) *Grundzüge einer Theorie der phylogenetischen Systematik.* Deutscher Zentralverlag, Berlin.

Hennig, W. (1966) *Phylogenetic systematics.* University of Illinois Press, Urbana.

Herbert, G. and S. Easteal. (1996) Relative rates of nuclear DNA evolution in human and Old World monkey lineages. *Mol. Biol. Evol.* 13: 1054–1057.

Higgins, D. G., J. D. Thompson and T. J. Gibson. (1996) Using CLUSTAL for multiple sequence alignments. In *Methods in enzymology* (R. F. Doolittle, ed.), pp. 383–401. Academic, San Diego, CA.

Hillis, D. M. (1999) SINEs of the perfect character. *Proc. Natl. Acad. Sci. USA* 96: 9979–9981.

Hillis, D. M. and J. J. Bull. (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42: 182–192.

Hillis, D. M., J. J. Bull, M. E. White, M. R. Badgett and I. J. Molineus. (1992) Experimental phylogenetics: Generation of a known phylogeny. *Science* 255: 589–592.

Hillis, D. M., J. P. Huelsenbeck and D. L. Swofford. (1994) Hobgoblin of phylogenetics? *Nature* 369: 363–364.

Hilton, H. and J. Hey. (1996) DNA sequence variation at the *Period* locus reveals the history of species and speciation events in the *Drosophila virlis* group. *Genetics* 144: 1015–1025.

Holmquist, G. P. and J. Filipski. (1994) Organization of mutations along the

genome: A prime determinant of genome evolution. *Trends Ecol. Evol.* 9: 65–69.

Honda, D., A. Yokota and J. Sugiyama. (1999) Detection of seven major evolutionary lineages in Cyanobacteria based on 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *J. Mol. Evol.* 48: 723–739.

Hood, L., J. H. Campbell and S. C. R. Elgin. (1975) The organization, expression, and evolution of antibody genes and other multigene families. *Annu. Rev. Genet.* 9: 305–353.

Hopkinson, D. A. and H. Harris. (1969) Red cell acid phosphatase, phosophoglucomutase, and adenylate kinase. In *Biochemical methods in red cell genetics* (G. Yunis, ed.), pp. 337–375. Academic, New York.

Horai, S., K. Hayasaka, R. Kondo, K. Tsugane and N. Takahata. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA* 92: 532–536.

Hubby, J. L. and R. C. Lewontin. (1966) A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics* 54: 577–594.

Hudson, R. R. (1983) Testing the constant rate neutral allele model with protein sequence data. *Evolution* 37: 203–217.

Hudson, R. R., M. Kreitman and M. Aguade. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.

Huelsenbeck, J. P. (1995) The performance of phylogenetic methods in simulation. *Syst. Biol.* 44: 17–48.

Huelsenbeck, J. P. and K. A. Crandall. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28: 437–466.

Huelsenbeck, J. P. and D. M. Hillis. (1993) Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* 42: 247–264.

Huelsenbeck, J. P., D. M. Hillis and R. Nielsen. (1996) A likelihood-ratio test of monophyly. *Syst. Biol.* 45: 546–558.

Hughes, A. L. (1999) Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* 48: 565–576.

Hughes, A. L. and M. Nei. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167–170.

Hughes, A. L. and M. Nei. (1989) Nucleotide substitution at major histocompatibility complex II loci: Evidence for overdominant selection. *Proc. Natl. Acad. Sci. USA* 86: 958–962.

Hughes, A. L. and M. Yeager. (1998) Natural selection and the evolutionary history of major histocompatibility complex loci. *Front. Biosci.* 3: d509–d516.

Hunt, D. M., K. S. Dulai, J. K. Bowmaker and J. D. Mollon. (1995) The chemistry of John Dalton's color blindness. *Science* 267: 984–988.

Ikemura, T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* 146: 1–21.

Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2: 13–34.

Ikemura, T. and S. Aota. (1988) Global variation in G+C content along vertebrate genome DNA. Possible correlation with chromosome band structures. *J. Mol. Biol.* 203: 1–13.

Ina, Y. (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40: 190–226.

Ingram, V. M. (1963) *The hemoglobins in genetics and evolution*. Columbia University Press, New York.

Innan, H., R. Terauchi, G. Kahl and F. Tajima. (1999) A method for estimating nucleotide diversity from AFLP data. *Genetics* 151: 1157–1164.

Irwin, D. M., R. T. White and A. C. Wilson. (1993) Characterization of the cow stomach lysozyme genes: repetitive DNA and concerted evolution. *J. Mol. Evol.* 37: 355–366.

Jacobs, G. H. (1993) The distribution and nature of color vision among the mammals. *Biol. Rev.* 68: 413–471.

Janke, A., G. Feldmaier-Fuchs, W. K. Thomas, A. von-Haeseler and S. Pääbo. (1994) The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics* 137: 243–256.

Janssen, P., R. Coopman, G. Huys, J. Swings, M. Bleeker, P. Vos et al. (1996) Evaluation of the DNA fingerprinting method AFLP as a new tool in bacterial taxonomy. *Microbiology* 142: 1881–1893.

Jermann, T. M., J. G. Opitz, J. Stackhouse and S. A. Benner. (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374: 57–59.

Jin, L. and R. Chakraborty. (1995) Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. *Heredity* 74: 274–285.

Jin, L. and M. Nei. (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7: 82–102.

Jin, L. and M. Nei. (1991) Relative efficiencies of the maximum-parsimony and distance-matrix methods of phylogeny construction for restriction data. *Mol. Biol. Evol.* 8: 356–365.

Johnson, N. L. and S. Kotz. (1969) *Distributions in statistics: Discrete distributions.* Houghton Mifflin, Boston, MA.

Johnson, N. L. and S. Kotz. (1970) *Distributions in statistics: Continuous univariate distributions—1.* Houghton Mifflin, Boston.

Jollès, J., P. Jollès, B. H. Bowman, E. M. Prager, C.-B. Stewart and A. C. Wilson. (1989) Episodic evolution in the stomach lysozymes of Ruminants. *J. Mol. Evol.* 28: 528–535.

Jones, D. T., W. R. Taylor and J. M. Thornton. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8: 275–282.

Jones, J. S., S. H. Bryant, R. C. Lewontin, J. A. Moore and T. Prout. (1981) Gene flow and the geographical distribution of a molecular polymorphism in *Drosophila pseudoobscura. Genetics* 98: 157–178.

Jorde, L. B., M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley, P. A. Krakowiak et al. (1995) Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57: 523–538.

Jukes, T. H. and C. R. Cantor. (1969) Evolution of protein molecules. In *Mammalian protein metabolism* (H. N. Munro, ed.), pp. 21–132. Academic, New York.

Jukes, T. H. and R. Holmquist. (1972) Evolutionary clock: Nonconstancy of rate in different species. *Science* 177: 530–532.

Jurka, J. and T. Smith. (1988) A fundamental division in the *Alu* family of repeated sequences. *Proc. Natl. Acad. Sci. USA* 85: 475–478.

Kaplan, N. (1983) Statistical analysis of restriction enzyme map data and nucleotide sequence data. In *Statistical analysis of DNA sequence data* (B. S. Weir, ed.), pp. 75–106. Marcel Dekker, New York.

Kaplan, N. and C. H. Langley. (1979) A new estimate of sequence divergence of mitochondrial DNA using restriction endonuclease mapping. *J. Mol. Evol.* 13: 295–394.

Kappen, C., K. Schughart and F. H. Ruddle. (1993) Early evolutionary origin of major homeodomain sequence classes. *Genomics* 18: 54–70.

Kasahara, M. (1997) New insights into the genomic organization and origin of the major histocompatibility complex: Role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas* 127: 59–65.

Keim, P., A. Kalif, J. Schupp, K. Hill, S. E. Travis, K. Richmond et al. (1997)

Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. *J. Bact.* 179: 818–824.
Keller, M. P., B. A. Seifried and P. F. Chance. (1999) Molecular evolution of the CMT1A-REP region: A human- and chimpanzee-specific repeat. *Mol. Biol. Evol.* 16: 1019–1026.
Kidd, K. K., P. Astolfi and L. Cavalli-Sforza. (1974) Error in the reconstruction of evolutionary trees. In *Genetic distance* (J. F. Crow and C. Denniston, eds.), pp. 131–136. Plenum, New York.
Kidd, K. K. and L. A. Sgaramella-Zonta. (1971) Phylogenetic analysis: Concepts and methods. *Am. J. Hum. Genet.* 23: 235–252.
Kido, Y., M. Aono, T. Yamaki, K. Matsumoto, S. Murata, M. Saneyoshi et al. (1991) Shaping and reshaping of salmonid genomes by amplification of tRNA-derived retroposons during evolution. *Proc. Natl. Acad. Sci. USA* 88: 2326–2330.
Kidwell, M. G. and D. Lisch. (1997) Transposable elements as sources of variation in animals and plants. *Proc. Natl. Acad. Sci. USA* 94: 7704–7711.
Kim, J. (1996) General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. *Syst. Biol.* 45: 363–374.
Kim, J., F. J. Rohlf and R. R. Sokal. (1993) The accuracy of phylogenetic estimation using the neighbor-joining method. *Evolution* 47: 471–486.
Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217: 624–626.
Kimura, M. (1969) The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. USA* 63: 1181–1188.
Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111–120.
Kimura, M. (1983) *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, U.K.
Kimura, M. and J. F. Crow. (1964) The number of alleles that can be maintained in a finite population. *Genetics* 49: 725–738.
Kimura, M. and T. Ohta. (1971) Protein polymorphism as a phase of molecular evolution. *Nature* 229: 467–469.
Kimura, M. and T. Ohta. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* 2: 87–90.
Kimura, M. and T. Ohta. (1973) Eukaryotes-prokaryotes divergence estimated by 5S ribosomal RNA sequences. *Nature* 243: 199–200.
King, J. L. and T. H. Jukes. (1969) Non-Darwinian evolution. *Science* 164: 788–798.
Kingman, J. F. C. (1982) On the genealogy of large populations. *J. Appl. Prob.* 19A: 27–43.
Kishino, H. and M. Hasegawa. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topology from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29: 170–179.
Kishino, H., T. Miyata and M. Hasegawa. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 31: 151–160.
Klein, J. and V. Horejsi. (1997) *Immunology.* Blackwell, London.
Klein, J., H. Ono, D. Klein and C. O'hUigin. (1993) The accordion model of Mhc evolution. *Prog. Immunol.* 8: 137–143.
Kluge, A. and J. Farris. (1969) Quantitative phyletics and the evolution of anurans. *Syst. Zool.* 18: 1–32.
Kocher, T. D. and A. C. Wilson. (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In *Evolution of life* (S. Osawa and T. Honjo, eds.), pp. 391–413. Springer-Verlag, New York.
Kohne, D. E. (1970) Evolution of higher-organism DNA. *Quart. Rev. Biophys.* 3: 327–375.

Koonin, E. V., A. R. Mushegian, M. Y. Galperin and D. R. Walker. (1997) Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol. Microbiol.* 25: 619–637.

Kornegay, J. R. (1996) Molecular genetics and evolution of stomach and nonstomach lysozymes in the hoatzin. *J. Mol. Evol.* 42: 676–684.

Kornegay, J. R., J. W. Schilling and A. C. Wilson. (1994) Molecular adaptation of a leaf-eating bird: Stomach lysozyme of the hoatzin. *Mol. Biol. Evol.* 11: 921–928.

Koshi, J. M. and R. A. Goldstein. (1996) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42: 313–320.

Kreitman, M. (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster. Nature* 304: 412–417.

Kreitman, M. and H. Akashi. (1995) Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26: 403–422.

Kuhner, M. K. and J. Felsenstein. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11: 459–468.

Kumar, S. (1996a) Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* 143: 537–548.

Kumar, S. (1996b) A stepwise algorithm for finding minimum evolution trees. *Mol. Biol. Evol.* 13: 584–593.

Kumar, S. and S. B. Hedges. (1998) A molecular timescale for vertebrate evolution. *Nature* 392: 917–919.

Kumar, S., K. Tamura, I. Jakobsen and M. Nei. (2000) *MEGA: Molecular evolutionary genetics analysis, ver. 2.* Pennsylvania State University, University Park, and Arizona State University, Tempe.

Kumar, S., K. Tamura and M. Nei. (1993) *MEGA: Molecular evolutionary genetics analysis* (with a 130-page printed manual). Pennsylvania State University, University Park.

Kumar, S., K. Tamura and M. Nei. (1994) MEGA: Molecular evolutionary genetics analysis software for microcomputers. *Comput. Appl. Biosci.* 10: 189–191.

Kumazawa, Y. and M. Nishida. (1995) Phylogenetic utility of mitochondrial transfer RNA genes for deep divergence in animals. In *Current topics on molecular evolution* (M. Nei and N. Takahata, eds.), pp. 23–35. The Pennsylvania State University, University Park.

Laird, C. D., B. L. McConaughy and B. J. McCarthy. (1969) Rate of fixation of nucleotide substitutions in evolution. *Nature* 224: 149–154.

Lake, J. A. (1994) Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* 91: 1455–1459.

Langley, C. H. and W. M. Fitch. (1974) An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* 3: 161–177.

Larget, B. and D. L. Simon. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16: 750–759.

Latter, B. D. H. (1972) Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics* 70: 475–490.

Lawlor, D. A., F. E. Ward, P. D. Ennis, A. P. Jackson and P. Parham. (1988) *HLA-A* and *B* polymorphisms predate the divergence of humans and chimpanzees. *Nature* 335: 268–271.

Lawson, C. L. and R. J. Hanson. (1974) *Solving least squares problems.* Prentice Hall, Englewood Cliffs, NJ.

Lee, Y.-H., T. Ota and V. D. Vacquier. (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol. Biol. Evol.* 12: 231–238.

Lewontin, R. C. (1974) *The genetic basis of evolutionary change.* Columbia University Press, New York.

Lewontin, R. C. and J. L. Hubby. (1966) A molecular approach to the study of

genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54: 595–609.

Li, C. C. (1976) *First course in population genetics*. Boxwood Press, Pacific Grove, CA.

Li, P. and J. Bousquet. (1992) Relative-rate test for nucleotide substitutions between two lineages. *Mol. Biol. Evol.* 9: 1185–1189.

Li, W.-H. (1989) A statistical test of phylogenies estimated from sequence data. *Mol. Biol. Evol.* 6:424–435.

Li, W.-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36: 96–99.

Li, W.-H. (1997) *Molecular evolution*. Sinauer Associates, Sunderland, MA.

Li, W.-H., D. L. Ellsworth, J. Krushkal, B. H.-J. Chang and D. Hewett-Emmett. (1996) Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phyl. Evol.* 5: 182–187.

Li, W.-H. and T. Gojobori. (1983) Rapid evolution of goat and sheep globin genes following gene duplication. *Mol. Biol. Evol.* 1: 94–108.

Li, W.-H., M. Gouy, P. M. Sharp, C. O'hUigin and Y.-W. Yang. (1990) Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA* 87: 6703–6707.

Li, W.-H., M. Tanimura and P. M. Sharp. (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* 25: 330–342.

Li, W.-H., C.-I. Wu and C.-C. Luo. (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2: 150–174.

Lockhart, P. J., D. Penny, M. D. Hendy, C. J. Howe, T. J. Geanland and A. W. D. Larkum. (1992) Controversy on chloroplast origins. *FEBS Lett.* 301: 127–131.

Lopez, J. V., R. Kersanach, S. A. Rehner and N. Knowlton. (1999) Molecular determination of species boundaries in corals: Genetic analysis of the *Montastraea annularis* complex using amplified fragment length polymorphisms and a microsatellite marker. *Biol. Bull.* 196: 80–93.

Lynch, M. and T. J. Crease. (1990) The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* 7: 377–394.

Lynch, M. and B. G. Milligan. (1994) Analysis of population genetic structure with RAPD markers. *Mol. Ecol.* 3: 91–99.

Maddison, D. R. (1991) The discovery and importance of multiple islands of most-parsimonious trees. *Syst. Biol.* 40: 315–328.

Maddison, W. P. (1995) Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Syst. Biol.* 44: 474–481.

Maddison, W. P. and D. R. Maddison. (1992) *MacClade: Analysis of phylogeny and character evolution*. Sinauer Associates, Sunderland, MA.

Mahtani, M. M. and H. F. Willard. (1993) A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: Implications for mechanisms of mutation at short tandem repeat loci. *Hum. Mol. Genet.* 2: 431–437.

Margoliash, E. (1963) Primary structure and evolution of cytochrome c. *Proc. Natl. Acad. Sci. USA* 50: 672–679.

Martin, A. P., G. J. P. Naylor and S. R. Palumbi. (1992) Rates of mitochondrial DNA evolution in sharks are slow compared to mammals. *Nature* 357: 153–155.

Maruyama, T. and M. Nei. (1981) Genetic variability maintained by mutation and overdominant selection in finite populations. *Genetics* 98: 441–459.

Maxam, A. M. and W. Gilbert. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* 74: 560–564.

Mayr, E. (1965) Discussion. In *Evolving genes and proteins* (V. Bryson and H. J. Vogel, eds.), pp. 293–294. Academic, New York.
Mayr, E. (1968) The role of systematics in biology. *Science* 159: 595–599.
McArthur, A. G. and B. F. Koop. (1999) Partial 28S rDNA sequences and the antiquity of hydrothermal vent endemic gastropods. *Mol. Phyl. Evol.* 13: 255–274.
McConnell, T. J., W. S. Talbot, R. A. McIndoe and E. K. Wakeland. (1988) The origin of MHC class II gene polymorphism within the genus *Mus*. *Nature* 322: 651–654.
McDonald, J. H. and M. Kreitman. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654.
McLaughlin, P. J. and M. O. Dayhoff. (1970) Eukaryotes versus Prokaryotes: An estimate of evolutionary distance. *Science* 168: 1469–1471.
Meireles, C. M., J. Czelusniak, M. P. C. Schneider, J. A. P. C. Muniz, M. C. Brigido, H. S. Ferreira et al. (1999) Molecular phylogeny of Ateline New World monkeys (*Platyrryhini, Atelinae*) based on γ-globin gene sequences: Evidence that *Brachyteles* is the sister group of *Lagothrix*. *Mol. Phyl. Evol.* 12: 10–30.
Messier, W. and C. B. Stewart. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385: 151–154.
Michelmore, R. W. and B. C. Meyers. (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* 8: 1113–1130.
Miyamoto, M. M. and J. Cracraft. (1991) Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. In *Phylogenetic analysis of DNA sequences* (M. M. Miyamoto and J. Cracraft, eds.), pp. 3–17. Oxford University Press, New York.
Miyata, T. and T. Yasunaga. (1980) Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* 16: 23–36.
Miyata, T., T. Yasunaga and T. Nishida. (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc. Natl. Acad. Sci. USA* 77: 7328–7332.
Mollon, J. (1991) The uses and evolutionary origins of primate color vision. In *Evolution of the eye and visual system* (J. R. Cronly-Dillon and R. L. Gregory, eds.), pp. 306–319. CRC, Boca Raton, FL.
Mollon, J. D., J. K. Bowmaker and G. H. Jacobs. (1984) Variations of color vision in a New World primate can be explained by a polymorphism of retinal photopigments. *Proc. R. Soc. Lond. B.* 222: 373–399.
Montgelard, C., F. M. Catzeflis and E. Douzery. (1997) Phylogenetic relationships of artiodactyls and cetaceans as deduced from the comparison of cytochrome b and 12S rRNA mitochondrial sequences. *Mol. Biol. Evol.* 14: 550–559.
Moore, G. W., J. Barnabas and M. Goodman. (1973) A method for constructing maximum parsimony ancestral amino acid sequences on a given network. *J. Theor. Biol.* 38: 459–485.
Moriyama, E. N. and J. R. Powell. (1996) Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* 13: 261–277.
Moriyama, E. N. and J. R. Powell. (1997) Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* 45: 514–523.
Moriyama, E. N. and J. R. Powell. (1998) Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* 26: 3188–3193.
Mountain, J. L. and L. L. Cavalli-Sforza. (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc. Natl. Acad. Sci. USA* 91: 6515–6519.
Murata, S., N. Takasaki, M. Saitoh and N. Okada. (1993) Determination of the phylogenetic relationships among Pacific salmonids by using short inter-

spersed elements (SINEs) as temporal landmarks of evolution. *Proc. Natl. Acad. Sci. USA* 90: 6995–6999.
Muse, S. (1996) Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* 13: 105–114.
Muse, S. V. and B. S. Weir. (1992) Testing for equality of evolutionary rates. *Genetics* 132: 269–276.
Muto, A. and S. Osawa. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA* 84: 116–119.
Myers, E. W. and W. Miller. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.* 4: 11–17.
Nagl, S., H. Tichy, W. E. Mayer, N. Takahata and J. Klein. (1998) Persistence of neutral polymorphisms in Lake Victoria cichlid fish. *Proc. Natl. Acad. Sci. USA* 95: 14,238–14,243.
Nagylaki, T. (1998) Fixation indices in subdivided populations. *Genetics* 148: 1325–1332.
Nathans, J., D. Thomas and D. S. Hogness. (1986) Molecular genetics of human color vision: The genes encoding blue, green, and red pigments. *Science* 232: 193–202.
National Research Council. (1996) *The evaluation of forensic DNA evidence.* National Academy Press, Washington, DC.
Needleman, S. G. and C. D. Wunsch. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48: 443–453.
Nei, M. (1965) Variation and covariation of gene frequencies in subdivided populations. *Evolution* 19: 256–258.
Nei, M. (1969) Gene duplication and nucleotide substitution in evolution. *Nature* 221: 40–42.
Nei, M. (1972) Genetic distance between populations. *Am. Nat.* 106: 283–292.
Nei, M. (1973) Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70: 3321–3323.
Nei, M. (1975) *Molecular population genetics and evolution.* North-Holland, Amsterdam, The Netherlands.
Nei, M. (1977) *F*-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* 41: 225–233.
Nei, M. (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89: 583–590.
Nei, M. (1982) Evolution of human races at the gene level. In *Human genetics, part A: The unfolding genome* (B. Bonné-Tamir, ed.), pp. 167–181. Alan R. Liss, New York.
Nei, M. (1986) Stochastic errors in DNA evolution and molecular phylogeny. In *Evolutionary perspectives and the new genetics* (H. Gershowitz, ed.), pp. 133–147. Alan R. Liss, New York.
Nei, M. (1987) *Molecular evolutionary genetics.* Columbia University Press, New York.
Nei, M. (1991) Relative efficiencies of different tree making methods for molecular data. In *Recent advances in phylogenetic studies of DNA sequences* (M. M. Miyamoto and J. L. Cracraft, eds.), pp. 133–147. Oxford University Press, Oxford, U.K.
Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* 30: 371–403.
Nei, M., R. Chakraborty and P. A. Fuerst. (1976a) Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci. USA* 73: 4164–4168.
Nei, M., A. Chakravarti and Y. Tateno. (1977) Mean and variance of F_{ST} in a finite number of incompletely isolated populations. *Theor. Pop. Biol.* 11: 291–306.
Nei, M., P. A. Fuerst and R. Chakraborty. (1976b) Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. *Nature* 262: 491–493.
Nei, M. and T. Gojobori. (1986) Simple methods for estimating the numbers

of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418–426.

Nei, M. and D. Graur. (1984) Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol.* 17: 73–118.

Nei, M., X. Gu and T. Sitnikova. (1997a) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proc. Natl. Acad. Sci. USA* 94: 7799–7806.

Nei, M. and A. L. Hughes. (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In *11th histocompatibility workshop and conference* (K. Tsuji, M. Aizawa and T. Sasazuki, eds.), pp. 27–38. Oxford University Press, Oxford, U.K.

Nei, M. and L. Jin. (1989) Variances of the average numbers of nucleotide substitutions within and between populations. *Mol. Biol. Evol.* 6: 290–300.

Nei, M., S. Kumar and K. Takahashi. (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proc. Natl. Acad. Sci. USA* 95: 12,390–12,397.

Nei, M. and W.-H. Li. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* 76: 5269–5273.

Nei, M. and J. C. Miller. (1990) A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* 125: 873–879.

Nei, M. and A. K. Roychoudhury. (1973) Probability of fixation and mean fixation time of an overdominant mutation. *Genetics* 74: 371–380.

Nei, M. and A. K. Roychoudhury. (1974) Genetic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am. J. Hum. Genet.* 26: 421–443.

Nei, M. and A. K. Roychoudhury. (1993) Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* 10: 927–943.

Nei, M., J. C. Stephens and N. Saitou. (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. *Mol. Biol. Evol.* 2: 66–85.

Nei, M. and F. Tajima. (1983) Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. *Genetics* 105: 207–217.

Nei, M., F. Tajima and Y. Tateno. (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.* 19: 153–170.

Nei, M. and N. Takezaki. (1994) Estimation of genetic distances and phylogenetic trees from DNA analysis, pp. 405–412 in *Proc. 5th World Congress Genet. Appl. Livestock Production,* edited by C. Smith, University of Guelph, Canada, Vol. 21.

Nei, M. and N. Takezaki. (1996) The root of the phylogenetic tree of human populations. *Mol. Biol. Evol.* 13: 170–177.

Nei, M., N. Takezaki and T. Sitnikova. (1995) Assessing molecular phylogenies. *Science* 267: 253–255.

Nei, M., J. Zhang and S. Yokoyama. (1997b) Color vision of ancestral organisms of higher primates. *Mol. Biol. Evol.* 14: 611–618.

Neitz, M., J. Neitz and G. H. Jacobs. (1991) Spectral tuning of visual pigments underlying red-green color vision. *Science* 252: 971–974.

Nielsen, R. and Z. Yang. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 920–936.

Nikaido, M., A. P. Rooney and N. Okada. (1999) Phylogenetic relationships among cetartiodactyls based on evidence from SINEs and LINEs: Hippopotamuses are the closest extant relatives of whales. *Proc. Natl. Acad. Sci. USA* 96: 10,261–10,266.

Novacek, M. J. (1992) Mammalian phylogeny: Shaking the tree. *Nature* 356: 121–125.

Ohno, S. (1967) *Sex chromosomes and sex-linked genes.* Springer-Verlag, Berlin.

Ohno, S. (1970) *Evolution by gene duplication.* Springer-Verlag, Berlin.

Ohta, T. (1973) Slightly deleterious mutant substitution in evolution. *Nature* 246: 96–98.

Ohta, T. (1980) *Evolution and variation of multigene families.* Springer-Verlag, Berlin.

Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* 23: 263–286.

Ohta, T. (1993) An examination of generation-time effect on molecular evolution. *Proc. Natl. Acad. Sci. USA* 90: 10,676–10,680.

Ohta, T. and M. Kimura. (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22: 201–204.

Okada, N. (1991) SINEs. *Curr. Opin. Genet. Dev.* 1: 498–504.

Okada, N., M. Hamada, I. Ogiwara and K. Ohshima. (1997) SINEs and LINEs share common 3′ sequences: A review. *Gene* 205: 229–243.

Olsen, G. J., H. Matsuda, R. Hagstrom and R. Overbeek. (1994) Fast DNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* 10: 41–48.

Osawa, S. (1995) *Evolution of the genetic code.* Oxford University Press, Oxford, U.K.

Ota, T. and M. Nei. (1994a) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin V_H gene family. *Mol. Biol. Evol.* 11: 469–482.

Ota, T. and M. Nei. (1994b) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. *J. Mol. Evol.* 38: 642–643.

Ota, T. and M. Nei. (1994c) Variances and covariances of the number of synonymous and nonsynonymous substitutions per site. *Mol. Biol. Evol.* 11: 613–619.

Page, R. D. M. and E. C. Holmes. (1998) *Molecular evolution: A phylogenetic approach.* Blackwell Science, Oxford, U.K.

Pamilo, P. and O. Bianchi. (1993) Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes. *Mol. Biol. Evol.* 19: 271–281.

Pamilo, P. and M. Nei. (1988) Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5: 568–583.

Pannell, J. R. and B. Charlesworth. (1999) Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution* 53: 664–676.

Peacock, D. and D. Boulter. (1975) Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. *J. Mol. Biol.* 95: 513–527.

Pearson, W. R., G. Robins and T. Zhang. (1999) Generalized neighbor-joining: More reliable phylogenetic tree reconstruction. *Mol. Biol. Evol.* 16: 806–816.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg and T. O. Yeates. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* 96: 4285–4288.

Penny, D. and M. D. Hendy. (1985) The use of tree comparison metrics. *Syst. Zool.* 34: 75–82.

Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner and J. Dodgeson. (1980) The evolution of genes: The chicken preproinsulin gene. *Cell* 20: 555–566.

Pons, O. and R. J. Petit. (1995) Estimation, variance and optimal sampling of genetic diversity. I. Haploid locus. *Theor. Appl. Genet.* 90: 462–470.

Pons, O. and R. J. Petit. (1996) Measuring and testing genetic differentiation with ordered versus unordered alleles. *Genetics* 144: 1237–1245.

Powell, J. R. (1997) *Progress and prospects in evolutionary biology: The Drosophila model.* Oxford University Press, New York.

Prakash, S., R. C. Lewontin and J. L. Hubby. (1969) A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. *Genetics* 61: 841–858.

Radding, C. M. (1982) Strand transfer in homologous genetic recombination. *Annu. Rev. Genet.* 16: 405–437.

Ramshaw, J. A. M., J. A. Coyne and R. C. Lewontin. (1979) The sensitivity of gel electrophoresis as a detector of genetic variation. *Genetics* 93: 1019–1037.

Rand, D. M. (1994) Thermal habit, metabolic rate and the evolution of mitochondrial DNA. *Trends Ecol. Evol.* 9: 125–131.

Rannala, B. and Z. Yang. (1996) Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *J. Mol. Evol.* 43: 304–311.

Rao, B. K., S. B. Sil and P. P. Majumder. (1997) How useful are microsatellite loci in recovering short-term evolutionary history? *J. Genet.* 76: 181–188.

Rast, J. P., M. K. Anderson, T. Ota, R. T. Litman, M. Margittai, M. J. Shamblott et al. (1994) Immunoglobulin light chain class multiplicity and alternative organizational forms in early vertebrate phylogeny. *Immunogenetics* 40: 83–99.

Redd, A. J., N. Takezaki, S. T. Sherry, S. T. McGarvey, A. S. M. Sofro and M. Stoneking. (1995) Evolutionary history of the COII/tRNA^{Lys} intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Mol. Biol. Evol.* 12: 604–615.

Roberts, R. G., A. J. Coffey, M. Bobrow and D. R. Bentley. (1992) Determination of the exon structure of the distal portion of the dystrophin gene by vectorette PCR. *Genomics* 13: 942–950.

Robertson, D. L., B. H. Hahn and P. M. Sharp. (1995) Recombination in AIDS viruses. *J. Mol. Evol.* 40: 249–259.

Robertson, H. M. (1998) Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res.* 8: 449–463.

Robinson, D. F. and L. R. Foulds. (1981) Comparison of phylogenetic trees. *Math. Biosci.* 53: 131–147.

Robinson, M., M. Gouy, C. Gautier and D. Mouchiroud. (1998) Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.* 15: 1091–1098.

Rodriguez, F., J. L. Oliver, A. Marin and J. R. Medina. (1990) The general stochastic model of nucleotide substitution. *J. Theor. Biol.* 142: 485–501.

Rogers, J. S. (1972) Measures of genetic similarity and genetic distance. In *Studies in genetics VII,* pp. 145–153. Publication 7213, University of Texas, Austin, TX.

Rogers, J. S. (1997) On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.* 46: 354–357.

Rogers, J. S. and D. L. Swofford. (1999) Multiple local maxima for likelihoods of phylogenetic trees: A simulation study. *Mol. Biol. Evol.* 16: 1079–1085.

Rohlf, F. J. (1993) *NTSYS-pc: Numerical taxonomy and multivariate analysis system, Edit. 1.80.* Applied Biostatistics, Setauket, NY.

Rooney, A. P., J. Zhang and M. Nei. (2000) An unusual form of purifying selection in a sperm protein. *Mol. Biol. Evol.* 17: 278–283.

Rosenberg, H. F., S. J. Ackerman and D. G. Tenen. (1989) Human eosinophil cationic protein. Molecular cloning of a cytotoxin and helminthotoxin with ribonuclease activity. *J. Exp. Med.* 170: 163–176.

Rosenberg, H. F. and K. D. Dyer. (1995) Eosinophil cationic protein and eosinophil-derived neurotoxin. Evolution of novel function in a primate ribonuclease gene family. *J. Biol. Chem.* 270: 21,539–21,544.

Rosenberg, H. F., K. D. Dyer, H. L. Tiffany and M. Gonzalez. (1995) Rapid evolution of a unique family of primate ribonuclease genes. *Nat. Genet.* 10: 219–223.

Russo, C. A. M., N. Takezaki and M. Nei. (1995) Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* 12: 391–404.

Russo, C. A. M., N. Takezaki and M. Nei. (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* 13: 525–536.

Ruvolo, M., D. Pan, S. Zehr, T. Goldberg, T. R. Disotell and M. von Dornum. (1994) Gene trees and hominoid phylogeny. *Proc. Natl. Acad. Sci. USA* 91: 8900–8904.

Ryan, S. C. and A. Dugaiczyk. (1989) Newly arisen DNA repeats in primate phylogeny. *Proc. Natl. Acad. Sci. USA* 86: 9360–9364.

Rzhetsky, A. and M. Nei. (1992a) A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol.* 9: 945–967.

Rzhetsky, A. and M. Nei. (1992b) Statistical properties of the ordinary least-squares, generalized least-squares and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* 35: 367–375.

Rzhetsky, A. and M. Nei. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* 10: 1073–1095.

Rzhetsky, A. and M. Nei. (1994) Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. *J. Mol. Evol.* 38: 295–299.

Rzhetsky, A. and M. Nei. (1995) Tests of applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12: 131–151.

Rzhetsky, A. and T. Sitnikova. (1996) When is it safe to use an oversimplified substitution model in tree-making? *Mol. Biol. Evol.* 13: 1255–1265.

Saitou, N. (1988) Property and efficiency of the maximum likelihood method for molecular phylogeny. *J. Mol. Evol.* 27: 261–273.

Saitou, N. and M. Imanishi. (1989) Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic reconstructions in obtaining the correct tree. *Mol. Biol. Evol.* 6: 514–525.

Saitou, N. and M. Nei. (1986) The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *J. Mol. Evol.* 24: 189–204.

Saitou, N. and M. Nei. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406–425.

Saitou, N. and F. Yamamoto. (1997) Evolution of primate ABO blood group genes and their homologous genes. *Mol. Biol. Evol.* 14: 399–411.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes et al. (1977) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature* 265: 687–695.

Sanghvi, L. D. (1953) Comparison of genetical and morphological methods for a study of biological differences. *Am. J. Phys. Anthr.* 11: 385–404.

Sankoff, D. and R. J. Cedergren. (1983) Simultaneous comparison of three or more sequences related by a tree. In *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (D. Sankoff and J. B. Kruskal, eds.), pp. 253–263. Addison-Wesley, Reading, MA.

Sankoff, D. D. and P. Rousseau. (1975) Locating the vertices of a Steiner tree in arbitrary space. *Math. Programming* 9: 240–246.

Sarich, V. M. and A. C. Wilson. (1966) Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation. *Science* 154: 1563–1566.

Sarich, V. M. and A. C. Wilson. (1967) Immunological time scale for hominid evolution. *Science* 158: 1200–1203.

Sattath, S. and A. Tversky. (1977) Additive similarity trees. *Psychometrika* 42: 319–345.

Schaeffer, S. W. and E. L. Miller. (1991) Nucleotide sequence analysis of Adh genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. USA* 88: 6097–6101.

Schaeffer, S. W. and E. L. Miller. (1992) Estimates of gene flow in *Drosophila*

pseudoobscura determined from nucleotide sequence analysis of the alcohol dehydrogenase region. *Genetics* 132: 471–480.
Schluter, D. (1995) Uncertainty in ancient phylogenies. *Nature* 377: 108–109.
Schöniger, M. and A. von Haeseler. (1993) A simple method to improve the reliability of tree reconstructions. *Mol. Biol. Evol.* 10: 471–483.
Schöniger, M. and A. von Haeseler. (1995) Performance of the maximum likelihood, neighbor joining, and maximum parsimony methods when sequence sites are not independent. *Syst. Biol.* 44: 533–547.
Schug, M. D., T. F. Mackay and C. F. Aquadro. (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster. Nat. Genet.* 15: 99–102.
Seino, S., G. I. Bell and W.-H. Li. (1992) Sequences of primate insulin genes support the hypothesis of a slower rate of molecular evolution in humans and apes than in monkeys. *Mol. Biol. Evol.* 9: 193–203.
Sellers, P. H. (1974) On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* 26: 787–793.
Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe and F. Wright. (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens:* A review of the considerable within-species diversity. *Nucleic Acids Res.* 16: 8207–8211.
Sharp, P. M. and W.-H. Li. (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* 4: 222–230.
Sharp, P. M., D. C. Shields, K. H. Wolfe and W. H. Li. (1989) Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* 246: 808–810.
Sharp, P. M., T. M. Tuohy and K. R. Mosurski. (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* 14: 5125–5143.
Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright. (1988) "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. *Mol. Biol. Evol.* 5: 704–716.
Shimamura, M., H. Yasue, K. Ohshima, H. Abe, H. Kato, T. Kishiro et al. (1997) Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* 388: 666–670.
Shriver, M. D., L. Jin, R. Chakraborty and E. Boerwinkle. (1993) VNTR allele frequency distributions under the stepwise mutation model: A computer simulation approach. *Genetics* 134: 983–993.
Silberman, J. D., C. G. Clark, L. S. Diamond and M. L. Sogin. (1999) Phylogeny of the Genera *Entamoeba* and *Endolimax* as deduced from small-subunit ribosomal RNA sequences. *Mol. Biol. Evol.* 16: 1740–1751.
Simonsen, K. L., G. A. Churchill and C. F. Aquadro. (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141: 413–429.
Simpson, G. G. (1964) Organisms and molecules in evolution. *Science* 146: 1535–1538.
Singer, M. F. (1982) SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. *Cell* 28: 433–434.
Sitnikova, T. (1996) Bootstrap method of interior-branch test for phylogenetic trees. *Mol. Biol. Evol.* 13: 605–611.
Sitnikova, T., A. Rzhetsky and M. Nei. (1995) Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evol.* 12: 319–333.
Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139: 457–462.
Smith, A. G., D. G. Smith and B. M. Funnell. (1994) *Atlas of Mesozoic and Cenozoic coastlines.* Cambridge University Press, New York.
Smith, G. P. (1974) Unequal crossover and the evolution of multigene families. *Cold Spring Harbor Symp. Quant. Biol.* 38: 507–513.

Smith, T. F., M. S. Waterman and W. M. Fitch. (1981) Comparative biosequence metrics. *J. Evol. Biol.* 18: 38–46.

Sneath, P. H. A. and R. R. Sokal. (1973) *Numerical taxonomy.* Freeman, San Francisco, CA.

Snyder, M. R. and G. J. Gleich. (1997) *Ribonucleases: Structures and functions.* Academic, New York.

Sober, E. (1988) *Reconstructing the past: Parsimony, evolution, and inference.* MIT Press, Cambridge, MA.

Sokal, R. R. and C. D. Michener. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 28: 1409–1438.

Sokal, R. R. and P. H. A. Sneath. (1963) *Principles of numerical taxonomy.* Freeman, San Francisco, CA.

Soodyall, H., L. Vigilant, A. V. Hill, M. Stoneking and T. Jenkins. (1996) mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub Saharan Africans. *Am. J. Hum. Genet.* 58: 596–608.

Sourdis, J. and C. Krimbas. (1987) Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.* 4: 159–166.

Sourdis, J. and M. Nei. (1988) Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Mol. Biol. Evol.* 5: 298–311.

Steel, M. (1994) The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biol.* 43: 560–564.

Stephens, S. G. (1951) Possible significance of duplication in evolution. *Adv. Genet.* 4: 247–265.

Stewart, C. B., J. W. Schilling and A. C. Wilson. (1987) Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* 330: 401–404.

Strimmer, K. and A. von Haeseler. (1996) Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13: 964–969.

Studier, J. A. and K. J. Keppler. (1988) A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* 5: 729–731.

Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* 48: 582–592.

Sullivan, J., K. E. Holsinger and C. Simon. (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmontine rodents. *Mol. Biol. Evol.* 12: 988–1001.

Sun, H., J. P. Macke and J. Nathans. (1997) Mechanisms of spectral tuning in the mouse green cone pigment. *Proc. Natl. Acad. Sci. USA* 94: 8860–8865.

Swanson, K. W., D. M. Irwin and A. C. Wilson. (1991) Stomach lysozyme gene of the Langur monkey: Tests for convergence and positive selection. *J. Mol. Evol.* 33: 418–425.

Swofford, D. L. (1998) *PAUP*: Phylogenetic analysis using parsimony (and other methods).* Sinauer Associates, Sunderland, MA.

Swofford, D. L. and D. P. Begle. (1993) *PAUP: Phylogenetic analysis using parsimony, ver. 3.1. user's manual.* Illinois Natural History Survey, Champaign, IL.

Swofford, D. L. and W. P. Maddison. (1987) Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* 87: 199–299.

Swofford, D. L., G. J. Olsen, P. J. Waddell and D. M. Hillis. (1996) Phylogenetic inference. In *Molecular systematics,* 2nd ed. (D. M. Hillis, C. Moritz and B. K. Mable, eds.), pp. 407–514. Sinauer Associates, Sunderland, MA.

Taberner, A., J. Dopazo and P. Castañera. (1997) Genetic characterization of populations of a de novo arisen sugar beet pest, *Aubeonymus mariaefranciscae* (Coleoptera, Curculionidae) by RAPD analysis. *J. Mol. Evol.* 44: 24–31.

Tachida, H. and M. Iizuka. (1993) A population genetic study of the evolution of SINEs. I. Polymorphism with regard to the presence or absence of an element. *Genetics* 133: 1023–1030.

Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460.
Tajima, F. (1989a) The effect of change in population size on DNA polymorphism. *Genetics* 123: 597–601.
Tajima, F. (1989b) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
Tajima, F. (1993a) Simple methods for testing molecular clock hypothesis. *Genetics* 135: 599–607.
Tajima, F. (1993b) Unbiased estimate of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 10: 677–688.
Tajima, F. and M. Nei. (1982) Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J. Mol. Evol.* 17: 115–120.
Tajima, F. and M. Nei. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.* 1: 269–285.
Tajima, F. and N. Takezaki. (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. *Mol. Biol. Evol.* 11: 278–286.
Takahashi, K. and M. Nei. (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol.* (in press).
Takahata, N. (1990) A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. *Proc. Natl. Acad. Sci. USA* 87: 2419–2423.
Takahata, N. and M. Nei. (1984) F_{ST} and G_{ST} statistics in the finite island model. *Genetics* 107: 501–504.
Takahata, N. and M. Nei. (1985) Gene genealogy and variance of interpopulational nucleotide difference. *Genetics* 110: 325–344.
Takahata, N. and M. Nei. (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124: 967–978.
Takahata, N. and S. R. Palumbi. (1985) Extranuclear differentiation and gene flow in the finite island model. *Genetics* 109: 441–457.
Takasaki, N., S. Murata, M. Saitoh, T. Kobayashi, L. Park and N. Okada. (1994) Species-specific amplification of tRNA-derived short interspersed repetitive elements (SINEs) by retroposition: A process of parasitization of entire genomes during the evolution of salmonids. *Proc. Natl. Acad. Sci. USA* 91: 10,153–10,157.
Takezaki, N. (1998) Tie trees generated by distance methods of phylogenetic reconstruction. *Mol. Biol. Evol.* 15: 727–737.
Takezaki, N. and T. Gojobori. (1999) Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol. Biol. Evol.* 16: 590–601.
Takezaki, N. and M. Nei. (1994) Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *J. Mol. Evol.* 39: 210–218.
Takezaki, N. and M. Nei. (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 144: 389–399.
Takezaki, N., A. Rzhetsky and M. Nei. (1995) Phylogenetic test of the molecular clock and linearized tree. *Mol. Biol. Evol.* 12: 823–833.
Tamura, K. (1992) The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol. Biol. Evol.* 9: 814–825.
Tamura, K. and M. Nei. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10: 512–526.
Tamura, K., G. Toba, J. Park and T. Aotsuka. (1995) Origin of Hawaiian drosophilids inferred from alcohol dehydrogenase gene sequences. In *Current topics on molecular evolution* (M. Nei and N. Takahata, eds.), pp. 9–18. The Pennsylvania State University, University Park.
Tan, I. H., J. Blomster, G. Hansen, E. Leskinen, C. A. Maggs, D. G. Mann et al.

(1999) Molecular phylogenetic evidence for a reversible morphogenetic switch controlling the gross morphology of two common genera of green seaweeds, *Ulva* and *Enteromorpha*. *Mol. Biol. Evol.* 16: 1011–1018.

Tanaka, T. and M. Nei. (1989) Positive Darwinian selection observed at the variable region genes of immunoglobulins. *Mol. Biol. Evol.* 6: 447–459.

Tateno, Y., M. Nei and F. Tajima. (1982) Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* 18: 387–404.

Tateno, Y., N. Takezaki and M. Nei. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* 11: 261–277.

Taylor, J. S., J. M. H. Durkin and F. Breden. (1999) The death of a microsatellite: A phylogenetic perspective on microsatellite interruptions. *Mol. Biol. Evol.* 16: 567–572.

Taylor, W. R. (1987) Multiple sequence alignment by a pairwise algorithm. *Comput. Appl. Biosci.* 3: 81–87.

Taylor, W. R. (1996) Multiple protein sequence alignment: Algorithms and gap insertion. In *Methods in enzymology* (R. F. Doolittle, ed.), pp. 343–367. Academic, San Diego, CA.

Templeton, A. R. (1983) Phylogenetic inference from restriction cleavage site maps with particular reference to the evolution of humans and the apes. *Evolution* 37: 221–244.

Thomas, R. H. and J. A. Hunt. (1993) Phylogenetic relationships in *Drosophila:* A conflict between molecular and morphological data. *Mol. Biol. Evol.* 10: 362–374.

Thorpe, J. P. (1982) The molecular clock hypothesis: Biochemical evolution, genetic differentiation and systematics. *Annu. Rev. Ecol. Syst.* 13: 139–168.

Tibayrenc, M., K. Neubauer, C. Barnabé, F. Guerrini, D. Skarecky and F. J. Ayala. (1993) Genetic characterization of six parasitic protozoa: Parity between random-primer DNA typing and multilocus enzyme electrophoresis. *Proc. Natl. Acad. Sci. USA* 90: 1335–1339.

Tourasse, N. J. and W. H. Li. (1999) Performance of the relative-rate test under nonstationary models of nucleotide substitution. *Mol. Biol. Evol.* 16: 1068–1078.

Travis, S. E., J. Maschinski and P. Keim. (1996) An analysis of genetic variation in *Astragalus cremnophylas var. cremnophylax,* a critically endangered plant, using AFLP markers. *Mol. Ecol.* 5: 735–745.

Trowsdale, J. (1995) "Both man & bird & beast": Comparative organization of MHC genes. *Immunogenetics* 41: 1–17.

Ullu, E. and C. Tschudi. (1984) *Alu* sequences are processed 7SL RNA genes. *Nature* 312: 171–172.

Upholt, W. B. (1977) Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucleic Acids Res.* 4: 1257–1265.

Uyenoyama, M. (1995) A generalized least-squares estimate of the origin of sporophytic self-incompatibility. *Genetics* 139: 975–992.

Uzzell, T. and K. Corbin. (1971) Fitting discrete probability distributions to evolutionary events. *Science* 172: 1089–1096.

Venkatesh, B., Y. Ning and S. Brenner. (1999) Late changes in spliceosomal introns define clades in vertebrate evolution. *Proc. Natl. Acad. Sci. USA* 96: 10,267–10,271.

Verneau, O., F. Catzeflis and A. V. Furano. (1997) Determination of the evolutionary relationships in *Rattus* sensu lato (Rodentia: Muridae) using L1 (LINE-1) amplification events. *J. Mol. Evol.* 45: 424–436.

Verneau, O., F. Catzeflis and A. V. Furano. (1998) Determining the dating recent rodent speciation events by using L1 (LINE-1) retrotransposons. *Proc. Natl. Acad. Sci. USA* 95: 11,284–11,289.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson. (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253: 1503–1507.

Vingron, M. and M. S. Waterman. (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* 235: 1–12.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes et al. (1995) AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407–4414.

Wahlund, S. (1928) Zusammensetzung von Populationen und Korrelationsercheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas* 11: 65–106.

Wakeley, J. (1993) Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. *J. Mol. Evol.* 37: 613–623.

Wakeley, J. (1994) Substitution rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11: 436–442.

Wang, G., T. S. Whittam, C. M. Berg and D. E. Berg. (1993) RAPD (arbitrary primer) PCR is more sensitive than multilocus enzyme electrophoresis for distinguishing related bacterial strains. *Nucleic Acids Res.* 21: 5930–5933.

Wang, T., Y. Okano, R. C. Eisensmith, W. H. Lo, S. Z. Huang, Y. T. Zeng et al. (1991) Identification of a novel phenylketonuria (PKU) mutation in the Chinese: Further evidence of multiple origins of PKU in Asia. *Am. J. Hum. Genet.* 43: 628–630.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* 7: 256–276.

Watterson, G. A. (1977) Heterosis or neutrality? *Genetics* 85: 789–814.

Weber, J. L. and C. Wong. (1993) Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2: 1123–1128.

Weinberg, W. (1908) Über den Nachweis der Vererbung beim Menschen. *Jahresb. Verein f. vaterl. Naturk. Württem* 64: 368–382.

Weir, B. S. (1996) *Genetic data analysis II.* Sinauer Associates, Sunderland, MA.

Weir, B. S. and C. C. Cockerham. (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Wells, R. S. (1996) Nucleotide variation at the *Gpdh* locus in the genus *Drosophila. Genetics* 143: 375–384.

Welsh, J. and M. McClelland. (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res.* 18: 7213–7218.

Whelan, S. and N. Goldman. (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 16: 1292–1299.

Whittam, T. and M. Nei. (1991) Neutral mutation hypothesis test. *Nature* 354: 115–116.

Whittam, T. S. (1995) Genetic population structure and pathogenicity in enteric bacteria. In *Population genetics of bacteria* (S. Baumberg, J. P. W. Young, S. R. Saunders and E. M. H. Wellington, eds.), pp. 217–245. Cambridge University Press, London.

Wiley, E. O. (1981) *Phylogenetics: The theory and practice of phylogenetic systematics.* Wiley, New York.

Wiley, E. O., D. R. Brooks, D. Siegel-Causey and V. A. Funk. (1991) *The Compleat cladist: A primer of phylogenetic procedures.* Museum of Natural History, University of Kansas, Lawrence.

Williams, J. G. K., A. R. Kublelik, K. J. Livak, J. A. Rafalski and S. V. Tingey. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531–6535.

Williams, P. L. and W. M. Fitch. (1990) Phylogeny determination using dynamically weighted parsimony method. In *Methods in enzymology* (R. F. Doolittle, ed.), pp. 615–626. Academic, San Diego, CA.

Wilson, A. C. and R. L. Cann. (1992) The recent African genesis of humans. *Sci. Amer.* 266: 68–73.

Wilson, A. C., S. S. Carlson and T. J. White. (1977) Biochemical evolution. *Annu. Rev. Biochem.* 46: 573–639.

Wise, C. A., M. Sraml and S. Easteal. (1998) Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. *Genetics* 148: 409–421.

Woese, C. (1998) The universal ancestor. *Proc. Natl. Acad. Sci. USA* 95: 6854–6859.

Wolfe, K. H., P. M. Sharp and W.-H. Li. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337: 283–285.

Wolstenholme, D. R. (1992) Animal mitochondrial DNA: Structure and evolution. In *International review of cytology: Mitochondrial genomes* (D. R. Wolstenholme and K. W. Jeon, eds.), pp. 173–216. Academic, San Diego, CA.

Wright, F. (1990) The effective number of codons used in a gene. *Genetics* 87: 23–29.

Wright, S. (1931) Evolution of Mendelian populations. *Genetics* 16: 167–176.

Wright, S. (1939) The distribution of self-sterility alleles in populations. *Genetics* 24: 538–552.

Wright, S. (1943) Isolation by distance. *Genetics* 28: 114–138.

Wright, S. (1951) The genetical structure of populations. *Ann. Eugen.* 15: 323–354.

Wright, S. (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution* 19: 395–420.

Wright, S. (1978) *Evolution and the genetics of populations.* University of Chicago Press, Chicago, IL.

Wu, C.-I. and W.-H. Li. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* 82: 1741–1745.

Yagi, T., G. Sasaki and H. Takebe. (1999) Phylogeny of Japanese papilionid butterflies inferred from nucleotide sequences of the mitochondrial ND5 gene. *J. Mol. Biol.* 48: 42–48.

Yamaguchi, Y. and T. Gojobori. (1997) Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proc. Natl. Acad. Sci. USA* 94: 1264–1269.

Yang, Z. (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10: 1396–1402.

Yang, Z. (1994a) Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* 39: 105–111.

Yang, Z. (1994b) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39: 306–314.

Yang, Z. (1994c) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.* 43: 329–342.

Yang, Z. (1995a) Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *J. Mol. Evol.* 40: 689–697.

Yang, Z. (1995b) *PAML: Phylogenetic analysis by maximum likelihood, ver. 1.0.* Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park.

Yang, Z. (1996a) Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11: 367–372.

Yang, Z. (1996b) Maximum-likelihood method for combined analyses of multiple sequence data. *J. Mol. Evol.* 42: 587–596.

Yang, Z. (1996c) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42: 294–307.

Yang, Z. (1997) How often do wrong models produce better phylogenies? *Mol. Biol. Evol.* 14: 105–108.

Yang, Z. (1999) *PAML: Phylogenetic analysis by maximum likelihood, ver. 2.0.* University College London, London.

Yang, Z., N. Goldman and A. Friday. (1994) Comparison of models for nu-

cleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11: 316–324.

Yang, Z., N. Goldman and A. E. Friday. (1995a) Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst. Biol.* 44: 384–399.

Yang, Z. and S. Kumar. (1996) Approximate methods for estimating the pattern of nucleotide substitution rates among sites. *Mol. Biol. Evol.* 13: 650–659.

Yang, Z., S. Kumar and M. Nei. (1995b) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141: 1641–1650.

Yang, Z. and B. Rannala. (1997) Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. *Mol. Biol. Evol.* 14: 717–724.

Yokoyama, R. and S. Yokoyama. (1990) Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus,* and human. *Proc. Natl. Acad. Sci. USA* 87: 9315–9318.

Yokoyama, S. and R. Yokoyama. (1989) Molecular evolution of human visual pigments genes. *Mol. Biol. Evol.* 6: 186–197.

Yokoyama, S., H. Zhang, F. B. Radlwimmer and N. S. Blow. (1999) Adaptive evolution of color vision of the Comoran coelacanth (*Latimeria chalumnae*). *Proc. Natl. Acad. Sci. USA* 96: 6279–6284.

Young, J. D., C. G. Peterson, P. Venge and Z. A. Cohn. (1986) Mechanism of membrane damage mediated by human eosinophil cationic protein. *Nature* 321: 613–616.

Yu, M. and D. M. Irwin. (1996) Evolution of stomach lysozyme: The pig lysozyme gene. *Mol. Phyl. Evol.* 5: 298–308.

Zachau, H. G. (1995) The human immunoglobulin κ genes. In *Immunoglobulin genes* (T. Honjo and F. W. Alt, eds.), pp. 173–191. Academic, San Diego, CA.

Zardoya, R., P. S. Economidis and I. Doadrio. (1999) Phylogenetic relationships of Greek Cyprinidae: Molecular evidence for at least two origins of the Greek Cyprinid Fauna. *Mol. Phyl. Evol.* 13: 122–131.

Zardoya, R. and A. Meyer. (1996) Phylogenetic performance of mitochondrial protein coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.* 13: 933–942.

Zhang, J. (1999) Performances of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol. Biol. Evol.* 16: 868–875.

Zhang, J. and X. Gu. (1998) Correlation between the substitution rate and rate variation among sites in protein evolution. *Genetics* 149: 1615–1625.

Zhang, J. and S. Kumar. (1997) Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* 14: 527–536.

Zhang, J., S. Kumar and M. Nei. (1997) Small-sample tests of episodic adaptive evolution: A case study of primate lysozymes. *Mol. Biol. Evol.* 14: 1335–1338.

Zhang, J. and M. Nei. (1997) Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods. *J. Mol. Evol.* 44: S139–S146.

Zhang, J., H. F. Rosenberg and M. Nei. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* 95: 3708–3713.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 39: 315–329.

Zharkikh, A. and W.-H. Li. (1992a) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. *Mol. Biol. Evol.* 9: 1119–1147.

Zharkikh, A. and W.-H. Li. (1992b) Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. *J. Mol. Evol.* 35: 356–366.

Zharkikh, A. and W.-H. Li. (1993) Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. *Syst. Biol.* 42: 113–125.

Zharkikh, A. and W.-H. Li. (1995) Estimation of confidence in phylogeny: Complete-and-partial bootstrap technique. *Mol. Phyl. Evol.* 4: 44–63.

Zimmer, E. A., S. L. Martin, S. M. Beverley, Y. W. Kan and A. C. Wilson. (1980) Rapid duplication and loss of genes coding for the chains of hemoglobin. *Proc. Natl. Acad. Sci. USA* 77: 2158–2162.

Zuckerkandl, E. and L. Pauling. (1962) Molecular disease, evolution, and genetic heterogeneity. In *Horizons in biochemistry* (M. Kasha and B. Pullman, eds.), pp. 189–225. Academic, New York.

Zuckerkandl, E. and L. Pauling. (1965) Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (V. Bryson and H. J. Vogel, eds.), pp. 97–166. Academic, New York.

Index

accepted point mutations, 27
acctran algorithm, 132
achieved significance level (ASL), 55
Adh, 5, 68, 69, 191, 200, 204, 263
AFLP, 284
Akaike's information content (AIC), 154
albumin, 195
alcohol dehydrogenase. *See Adh*
algorithm(s)
  branch swapping, 126
  close neighbor interchange, 100
  max-mini, 124
  min-mini, 128
  nearest neighbor interchange (NNI), 126
  star-decomposition (SD), 151, 157
  stepwise addition, 125
  subtree pruning and regrafting (SPR), 126
  tree bisection-reconnection (TBR), 126
alignment distance, 46
alignment gaps, 47
allele(s)
  codominant, 234
  frequencies of, 233, 234
Alu, 141
amplified fragment length polymorphism (AFLP), 284
ancestral sequence inference
  accuracy of, 210
  distance-based method, 209
  likelihood-based method, 211
  parsimony method, 208
average gene diversity, 245
average heterozygosity, 245
average pathway method, 132
beta distribution, 261
BIONJ method, 109
birth-and-death process, 293
bootstrap
  confidence value, 89, 172
  consensus trees, 90, 131
  covariance, 26
  interior branch test, 170
  statistical properties, 173
  test of phylogeny, 89, 171
  variance, 26
  value, 89, 172
branch length estimation
  acctran, 132
  by maximum parsimony, 131
  deltran, 132
  Fitch-Margoliash method, 94
  least-squares method, 98
  neighbor-joining method, 106

clade, 140
cladistic parsimony, 140
coancestry coefficient, 242
codon, 6
codon usage, 12
codon usage bias, 12
  tRNA abundance, 12
  effect on evolutionary distance, 51, 70
  RSCU, 16
codons, types of, 7
coefficient of nucleotide differentiation, 258
color vision, 212
complete-deletion option, 49
concerted evolution, 293
condensed tree, 175
confidence probability, 170

consistency index, 120
consistent estimator, 178
convergent changes, 223
convergent evolution, 222
core tree, 123, 127
cytochrome *b*, 41, 138
cytochrome *c*, 21
cytochrome oxidase 1, 160

Dayhoff
  distance, 28
  matrix, 27
deleterious mutations, 31
deletion, 9, 11
deltran algorithm, 132
distance
  alignment, 46
  Dayhoff, 28
  gamma. *See* gamma distance
  genetic. *See* genetic distance
  Grishin, 23
  Jukes-Cantor, 37
  Kimura two-parameter, 37
  logDet, 41
  maximum likelihood, 45
  nonsynonymous, 52
  paralinear, 41
  patristic, 92
  Poisson correction (PC), 20
  synonymous, 52
  Tajima-Nei, 39
  Tamura, 39
  Tamura-Nei, 40
  topological, 81
  transitional, 38
  transversional, 38
distance matrix methods, 87
D-loop region, 44
DNA, 3
DNA polymorphism, 250
Dollo parsimony, 145
dot matrix, 46
dynamically weighted parsimony, 134

eosinophil cationic protein, 219
eosinophil-derived neurotoxin, 219
evolutionary distance, 17
evolutionary process
  birth-and-death evolution, 293
  concerted evolution, 293
  gene conversion, 11
  horizontal gene transfer, 11
  transposition, 11
  unequal crossover, 11
exhaustive search, 122, 166
exons, 6
expected tree, 78
exterior branches, 75

Fisher's exact test, 56
Fitch-Margoliash method, 93
fixation indices, 238
  estimation of, 241
  multiple alleles, 236
fixed site, 263
frameshift mutations, 9

gamma distance
  for proteins, 22
  Jukes-Cantor, 43
  Kimura 2-parameter, 43
  Tamura-Nei, 44
gamma distribution, 21, 269-270
gamma parameter, 21
gaps, 46
  complete-deletion, 49
  handling, 49
  pairwise-deletion, 49
  penalty, 47
GC content, 13, 15
gene conversion, 11, 293
gene diversity
  expected, 246, 247
  subdivided populations, 248
gene tree, 75
generalized least squares method, 93
genes
  orthologous vs. paralogous, 77
  protein-coding, 5, 6
  RNA-coding, 5
genetic codes, 7
genetic distance, 265
  chord distance, 267
  D_A, 268
  F_{ST}, 268
  Rogers', 266
  standard, 269
genetic drift, 4, 188
genetic markers
  AFLP, 284
  allele frequencies, 265
  RAPD, 285
  restriction sites, 275
  RFLP, 276
genetic polymorphism, 231
genome projects, 292
genotype frequencies, 233
  observed vs. expected, 235

G_{ST}, 240
  expected value of, 243
γ-fibrinogen, 136, 156

Hardy-Weinberg principle, 233
hemoglobin, 17, 24, 27, 188
heterozygosity
  average, 245
  observed and expected, 236
  within-population, 239
heuristic search, 122, 125
HKA test, 262
homoplasy, 115, 120
  index, 121
horizontal gene transfer, 11

inconsistency, 83
indels, 18
inferred tree, 78
infinite-allele model, 246
infinite-site model, 250
informative site, 119
initiation codons, 7
insertion, 9, 11, 18
interior-branch test, 89, 168
  bootstrap, 170
interior branches, 75
interior nodes, 75
introns, 6
invariable site, 119
inversion, 9
isochores, 15

likelihood ratio test, 154, 171
  for trees, 176
linearized tree, 203
LINEs, 141
local upperbound, 128
long-branch attraction, 116
lysozyme, 4, 222, 226

major histocompatibility complex. *See* MHC
majority-rule consensus trees, 130
maximum likelihood
  parameter estimation, 163
  protein sequences, 159
  theoretical foundation, 162
  tree search algorithms, 150, 160
maximum parsimony method, 115
  estimation of tree length, 116
  estimation of branch lengths, 131
McDonald-Kreitman test, 263
MHC, 5, 56, 59, 66, 71
  antigen recognition site, 61
microsatellite DNA, 247
minimum evolution method, 99
mismatch penalty, 47
missing data. *See* gaps
model
  equal-input, 35
  Felsenstein, 38, 153
  general reversible (REV), 35, 154
  Hasegawa et al. (HKY), 40
  Jukes-Cantor, 43
  Kimura two-parameter, 37
  Tamura, 39
  Tamura-Nei, 40
modified Nei-Gojobori method, 57
molecular clock, 21
  generation time hypothesis, 189
  linearized tree, 203
  relative rate test, 191–196
  phylogenetic tests, 196, 200
mRNA, 5–6
multifurcating tree, 75
mutation
  advantageous, 31
  biased mutation pressure, 13
  deleterious, 31
  fixation, 4, 29, 231
  frameshift, 9
  gene, 4
  neutral, 30
  nonsense, 11
  slightly deleterious, 31
  types of, 9

NADH dehydrogenase 5, 67, 70
negative binomial distribution, 21
Nei-Gojobori method, 52
neighbor joining (NJ) method, 103
neighbors, 104
neutral theory, 187, 231, 232
  d_N/d_S tests, 56
  HKA test, 262
  McDonald-Kreitman test, 263
  Tajima test, 260–261
nonsense mutations, 11
nonsynonymous substitutions, 51
nucleotide diversity, 251
null tree, 101, 169

opsin, 212, 229
optimization principle, 165
ordinary least-squares method, 92
orthologous genes, 77
overdominant selection, 56

p distance, 18, 33
pairwise-deletion option, 49

PAM, 27
parallel changes, 223
paralogous genes, 77
paraphyly, 137
parsimony-informative sites, 119
PC distance, 20
phenogram, 87
phenylketonuria, 232
phylogenetic tests, 196–200
  branch length test, 199–200
  two-cluster test, 196–199
Poisson distribution, 20
polymorphic sites, 263
polymorphism, 231
population tree, 75
positive Darwinian selection, 55
pre-mRNA, 6
pruning, 149
purifying selection, 13
purine, 10
pyrimidine, 10

random amplification of polymorphic DNA (RAPD), 285
realized tree, 78
reconstructed tree, 78
relative optimality score, 166
relative rate test, 191–196
  likelihood method, 193
  nonparameteric, 193
  Tajima, 195
replacement substitutions, 11
rescaled consistency index, 120
restriction enzymes, 275
  multiple recognition sequences, 278
restriction fragment data, 278, 283
restriction site data, 280
retention index, 120
reversible model, 148, 158
RFLP, 276
ribonucleases, 4
RNA editing, 9
RNA polymerase, 12
rooted tree, 73
rRNA, 5
RSCU, 16

search factor, 129
segregating sites, 250
selection
  d_N/d_S test, 55, 56
  Fu-Li test, 261
  HKA test, 262
  McDonald-Kreitman test, 263
  positive, 31
  purifying, 13, 231
  Tajima test, 260
sequence alignment, 48
shared derived characters, 140
short tandem repeats (STR), 247
short-branch attraction, 116
silent substitutions, 11
similarity index, 46
SINEs, 141
singleton site, 119
small nuclear RNA, 6
species tree, 75
specific tree search, 122
standard genetic code, 7
standard genetic distance, 269
step matrix, 133
stepwise addition algorithm, 125
stepwise mutation model, 247, 270
stop codons, 7
strict consensus trees, 130
subdivided populations, 236
substitution
  nonsynonymous, 11, 51
  nucleotide, 9
  rate of amino acid, 20
  rate of gene, 30
  replacement, 11
  silent, 11
  synonymous, 10
  transitional, 10
  transversional, 10
substitution matrix
  amino acid, 27
  Dayhoff, 28
  Jones et al., 29
  mitochondrial proteins, 29
subtree pruning and regrafting (SPR), 126
synapomorphies, 140
synonymous
  codons, 7
  substitutions, 10, 51

taxa, 74
taxonomy, 4
temporary MP tree, 124
termination codons, 7
tie trees, 89
topological distance, 81
topology, 73
transition/transversion ratio (R), 34, 38
transition/transversion rate ratio (k), 38

transitional substitutions, 10
transposable elements, 11
transversion parsimony, 133
transversional substitutions, 10
tree bisection and reconnection (TBR), 126
tree length, 118
trees
  bifurcating, 75
  bootstrap consensus, 90, 131
  condensed, 175
  consensus, 130
  expected, 78
  gene, 75
  multifurcating, 75
  number of branches, 75
  number of interior nodes, 75
  number of topologies, 74
  population, 75
  realized, 78
  reconstructed, 78
  rooted and unrooted, 73
  species, 75
  symbolic expression of, 82
tRNA, 5

unequal crossover, 11
uniquely shared sites, 223
universal genetic code, 7
unrooted tree, 73
unweighted parsimony, 116
UPGMA, 87

variable site, 119

Wahlund's principle, 237
weight matrix, 133, 134
weighted least-squares method, 93
weighted parsimony, 116, 133

Z-test, 199, 217

η-globin, 190