University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Agnieszka Mykowiecka
Student no. 292484

Inference of Credible Associations between Genes and Genomes

PhD's dissertation
in COMPUTER SCIENCE

Supervisors:

dr hab. Paweł Górecki
Institute of Informatics, University of Warsaw

June 2022

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of PhD of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Inference of Credible Associations between Genes and Genomes

Hypothesis of the course of gene and species evolution can be represented as a phylogenetic tree, which illustrates the ancestor–descendant relationships. Events such as gene duplications, losses and horizontal gene transfers (HGT) can lead to the incongruence of the gene and its species tree. To locate such events, one can use tree reconciliation. However, this method is prone to topology errors, therefore, assessing the credibility of evolutionary events and reliable inference of reticulate evolution are important issues in phylogenetics.

In this dissertation we propose a novel approach to assess credibility of gene duplications and speciations. We developed a linear time algorithm, based on unrooted reconciliation and non-parametric bootstrap, which calculates support values for evolutionary events. We also show the application of our method to the rooting and supertree problems.

To address the needs of metagenomic and microbial studies, we investigated the problem of the inference of well-supported HGT events. We propose a new measure based on non-parametric bootstrap, called transfer support. Then, we use it to design a new, efficient heuristic algorithm that iteratively infers acyclic and well-supported transfer scenarios. Our method, based on a novel square time HGT-reconciliation algorithm, postulates the most probable locations based on the extended tree reconciliation and credibility of inferred HGTs.

Another challenge in metagenomic studies is the gene-species assignment problem, i.e., the problem of mapping of genes of unknown origin to a particular species after shotgun sequencing. To address the problem, we propose the first HGT-reconciliation based approach to infer such mappings with two tractable HGT-models: time consistent (tcDTL) and general (DTL). The algorithm for the DTL model runs in square time, while for the tcDTL model, we describe a cubic time solution with several improvements and generalizations.

Finally, we propose a novel network-based approach to datasets containing sequences whose high similarity prevents a credible phylogenetic tree inference. We apply the methods to BCR receptor sequences from B-cells of follicular lymphoma patients, which allowed us to model tumor evolution and observe subclonal selection driven by BCR mutations.

Rekonstrukcja Wiarygodnych Relacji Między Genami a Genomami

Hipotezę dotyczącą przebiegu ewolucji genów i gatunków można przedstawić w postaci drzew filogenetycznych, które ilustrują relacje przodek–potomek. Zdarzenia duplikacji, strat oraz horyzontalnego transferu genów (HGT) mogą prowadzić do niezgodności pomiędzy topologiami drzew genów i gatunków. Metoda uzgadniania drzew pozwala na zlokalizowanie takich zdarzeń, jednak jej ograniczenia oraz duża wrażliwość na błędy w topologiach drzew sprawia, że wiarygodność zdarzeń ewolucyjnych oraz opracowanie wiarygodnych metod rekonstruowania zdarzeń retykulacyjnych, wciąż stanowią otwarty problem w dziedzinie filogenetyki.

W niniejszej rozprawie zaproponowaliśmy nowe podejście do oceny wiarygodności duplikacji i specjacji. Zdefiniowaliśmy miarę wsparcia dla tych zdarzeń i opracowaliśmy liniowy algorytm, oparty na nieukorzenionym uzgadnianiu drzew i nieparametrycznym bootstrapie, do jej obliczania. Pokazaliśmy również zastosowanie naszej metody do problemów ukorzeniania drzew i budowy superdrzew.

Aby odpowiedzieć na potrzeby badań metagenomicznych i mikrobiologicznych, podjęliśmy temat lokalizowania wiarygodnych zdarzeń HGT. Zaproponowaliśmy nową miarę opartą na nieparametrycznym bootstrapie, zwaną wsparciem transferu, i wykorzystaliśmy ją do stworzenia nowego i wydajnego algorytmu heurystycznego, który iteracyjnie znajduje acykliczne i dobrze wspierane transfery genów. Nasza metoda, oparta na kwadratowym algorytmie uzgadniania, postuluje najbardziej prawdopodobne miejsca transferów na podstawie ich wiarygodności.

Innym wyzwaniem, które pojawia się w badaniach metagenomicznych, jest problem przyporządkowania genów do gatunków po wykorzystaniu metody sekwencjonowania typu *shotgun*, w której te przyporządkowania mogą zostać utracone. Do rekonstrukcji relacji gen–gatunek zaproponowaliśmy pierwsze tego rodzaju podejście, oparte na uzgadnianiu z transferami, umożliwiające zastosowanie dwóch modeli: spójnego czasowo (tcDTL) i ogólnego (DTL). Algorytm dla modelu DTL działa w czasie kwadratowym, natomiast dla modelu tcDTL opisujemy rozwiązanie w czasie sześciennym z kilkoma ulepszeniami i uogólnieniami.

Na końcu, skupiliśmy się na przypadku, gdy drzewa filogenetyczne są niewystarczające do przedstawienia złożonych relacji ewolucyjnych. Nasze podejście oparte na sieciach zastosowaliśmy do zbiorów danych zawierających sekwencje, których duże podobieństwo uniemożliwia zbudowanie wiarygodnych drzew filogenetycznych. Sieci otrzymane dla sekwencji receptora BCR, pochodzących z limfocytów B pobranych od pacjentów z chłoniakiem pęcherzykowym, pozwalają modelować ewolucję nowotworu i obserwować selekcję subklonalną indukowaną przez mutacje BCR.

Keywords

tree reconciliation, phylogenetic tree, gene tree, species tree, gene duplication, speciation, horizontal gene transfer, phylogenetic network, bootstrap, credibility of evolutionary events

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

Applied computing → Life and medical sciences →
→ Genomics
→ Bioinformatics
→ Computational biology → Computational genomics
Mathematics of computing → Discrete mathematics →
→ Graph theory → Trees
→ Graph theory → Graph algorithms

Tytuł pracy w języku polskim

Rekonstrukcja Wiarygodnych Relacji Między Genami a Genomami

List of publications of major results from the thesis:

Mykowiecka, A., & Górecki, P. (2016). Bootstrapping algorithms for gene duplication and speciation events. *In International Conference on Algorithms for Computational Biology* (pp. 106-118). Springer, Cham.

Mykowiecka, A., Szczęsny, P., & Górecki, P. (2017). Inferring gene-species assignments in the presence of horizontal gene transfer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5), 1571-1578.

Mykowiecka, A., & Górecki, P. (2018). Credibility of evolutionary events in gene trees. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3), 713-726.

Mykowiecka, A., Muszewska, A., & Górecki, P. (2018). Inferring time-consistent and well-supported horizontal gene transfers. In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 79-83). IEEE.

van Bergen, C. A., Koning, M. T., Quinten, E., Mykowiecka, A., Sepulveda, J., Monajemi, R., et al. & Veelken, H. (2019). High-Throughput BCR Sequencing and Single-Cell Transcriptomics Reveal Distinct Transcriptional Profiles Associated with Subclonal Evolution of Follicular Lymphoma. *Blood*, 134, 298.

List of selected publications in phylogenetics:

Szczesny, P., Mykowiecka, A., Pawłowski, K., & Grynberg, M. (2013). Distinct protein classes in human red cell proteome revealed by similarity of phylogenetic profiles. *PloS one*, 8(1), e54471.

Górecki, P., Paszek, J., & Mykowiecka, A. (2016). Mean values of gene duplication and loss cost functions. *In International Symposium on Bioinformatics Research and Applications* (pp. 189-199). Springer, Cham.

Górecki, P., Mykowiecka, A., Paszek, J., & Eulenstein, O. (2019). Mathematical properties of the gene duplication cost. *Discrete Applied Mathematics*, 258, 114-122.

Wawerka, M., Dąbkowski, D., Rutecka, N., Mykowiecka, A., & Górecki, P. (2021). Conflict Resolution Algorithms for Deep Coalescence Phylogenetic Networks. In 21st International Workshop on Algorithms in Bioinformatics (WABI 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Wawerka, M., Dąbkowski, D., Rutecka, N., Mykowiecka, A., & Górecki, P. (2022). Embedding gene trees into phylogenetic networks by conflict resolution algorithms. *Algorithms for Molecular Biology*, 17(1), 1-23.

Acknowledgements:

Many thanks to my supervisor Paweł Górecki for his time and guidance. Thank you for your commitment to our research, tons of great ideas and inspirations, as well as your humor and patience. I hope that we can count the snow-capped peaks of Huangshan as Mont Blanc.

I would also like to thank my collaborators, colleagues, and family for their invaluable support.

Contents

List of Figures

List of Tables

1 Introduction

Phylogenetics by the most general definition is the study of the evolutionary history and relationships between individuals, groups of organisms or genes and genomes. The term phylogeny from the German *Phylogenie* was introduced by German biologist Ernst Heinrich Haeckel in *Generelle Morphologie der Organismen* in the 19th century and it originates from the Greek φυλή/φῦλον (phylé/phylon) *tribe*, *race*, and γενετικός (genetikós) *origin*, *source*, *birth*.

The word was coined in the 19th century, but classifying the natural world into meaningful and useful categories dates back at least to ancient Greece. For centuries the notion of a *Great Chain of Being* or Latin *Scala Naturae* was the prevailing theory throughout the Western world. This concept was derived from Plato and Aristotle's *Historia Animalium* in which the philosopher ranked animals over plants due to their ability to move and their senses. Animals were graded by their reproductive mode, laying eggs being lower in the chain than live birth, and possession of blood, warm-blooded mammals and birds again being higher than "bloodless" invertebrates. This non-religious concept of higher and lower organisms was further developed by natural philosophers and became the basis of the *Scala Naturae*. *Scala* ordered beings from God to angels, humans, animals and plants to minerals. In medieval times, the great chain was seen as decreed by God and unchangeable. More modern classification in which physical components of the world were di-

vided into the three kingdoms of animals, plants and minerals was introduced by Carl Linnaeus in 18th century. This Swedish botanist, zoologist and physician is usually regarded as the founder of modern taxonomy. In his books he laid the foundations for biological nomenclature, developed principles for assigning names to plants and animals using binomial nomenclature, and introduced the standard hierarchy of class, order, genus, and species. Although Linnaeus's classification was a step toward more scientific approach, it was still based on Aristotle's idea of the essential features of living organisms and on general physical similarity. The revolution came with Charles Darwin's publication of the theory of evolution in his work *On the Origin of Species by Means of Natural Selection*. Since then, phylogenetics and the search of evolutionary relationships began to be based on archaeological, and historical studies, and after further discoveries, also on genetic and molecular data.

Molecular era

Great many fields of science and medicine have emerged after the discoveries of the existence of proteins and genes. Among the most important scientific disciplines are genetics, which focuses on genes, genetic variation and heredity in organisms, and molecular biology. The latter lies between genetics and biochemistry covering issues concerning molecular basis of biological processes and interactions inside of cells and between them.

The term *molecular biology* was introduced in 1938 by Warren Weaver, however, the field itself was already established in 1930 [Alberts *et al.* (2003); Morange (2000)]. At that time it already had a solid foundation. Existence of certain discrete inheritance units was suggested by Gregor Mendel, who studied edible pea plants and discovered inheritance patterns by observing visual characteristics of the crossbred plants. His research allowed him to describe dominant and recessive traits, the concept of homozygote and heterozygote, and the phenomenon of discontinuous inheritance. Mendel's work was published in 1866 but it was not until the the end of 19th century that his findings were recognized and confirmed by other researchers [Henig (2000)]. Shortly thereafter, at the beginning of the 20th century terms *gene* and *genetics* were introduced by Wilhelm Johannsen and William Bateson [Johannsen (1909); Bateson *et al.* (1906)].

DNA helix and the genetic code

Another milestone in advancing molecular biology and genetics was the discovery of the deoxyribonucleic acid (DNA) and its helix structure. DNA is a polymer consisting of two polynucleotide chains forming a double helix. Nucleotides are

composed of a deoxyribose sugar molecule, a phosphate group, and one of four nitrogenous bases: a purine (adenine (A) or guanine (G)) or a pyrimidine (cytosine (C) or thymine (T)). A single DNA strand is connected by covalent bonds, while two strands are joined in a helix by hydrogen bonds [Alberts *et al.* (2003)].

DNA was isolated from the cell nucleus for the first time in 1869 and it took over 70 years to discover its structure and role in genetic inheritance. In the early 20th century, the chemical structure was identified [Cohen and Portugal (1974)] and later in 1953 James D. Watson and Francis Crick [Watson and Crick (1953)] published their model of the DNA as a double stranded helix. The first DNA sequence was read in early 1970s and since then teh sequencing methods were developed and improved. Today's methods allow to sequence much longer sequences and in much shorter time, however, the problem of read errors has not yet been completely solved.

In the course of DNA research, a second nucleic acid was also discovered. Ribonucleic acid (RNA) structure is very similar to DNA with a few differences. Instead of the deoxyribose, the RNA molecule contains ribose and the thymine is replaced by uracil. Although RNA can form double-stranded molecules, unlike DNA it is usually single-stranded. There are many types of RNA molecules, such as mRNA, tRNA, rRNA and some of their functions and related processes have only been discovered in modern times [Berg *et al.* (2007)].

DNA is organized into long structures called chromosomes, which in eukaryotic cells are located in the nucleus, whereas in prokaryotic cells chromosomes are circular and stored in cytoplasm. The role of DNA in all organisms and many viruses is to code genetic information responsible for the development, function and reproduction of the organism. The complete set of genetic information in an organism is called *genome*. Genetic code is based on three-letter, non-overlapping fragments called *codons*, which are translated into the amino acids in the process of *gene expression*. Amino acids form long chains, i.e. proteins, which are an essential part of all living organisms, being structural component and playing an important role in many processes as enzymes or elements of the immune system. Gene expression consists of two major steps. The first one, called *transcription*, is the process of copying genetic information contained in the genes from DNA into mRNA. Next, during *translation* the coding fragments from mRNA are used as a template for protein synthesis. The principle stating that information can only be transferred from nucleic acid to nucleic acid and from nucleic acid to protein was established as the central dogma of molecular biology by Francis Crick in 1958 [Watson and Crick (1958)]. The simplified version that says that DNA is transcribed to RNA and RNA

is translated to proteins is incorrect due to the existence of processes such as reverse transcription, which transcribes information from RNA to DNA.

The genetic information in DNA is contained in *genes*, the nucleotide sequences encoding a specific protein or RNA. They are the basic unit of heredity and the information they contain determines the phenotype, i.e., a set of observable characteristics or traits of an organism. Genes therefore influence traits such as hair color or blood type, but also the structural correctness of proteins, which ensures the correctness of biochemical processes occurring in the organism.

In addition to the protein coding part, the gene structure also includes regulatory sequences that are essential for their expression. To initiate the transcription process, the promoter sequence is required. It is the binding site for *RNA polymeraze*, which synthesizes RNA from a DNA template. The *stop codon* is located on the opposite side from the promoter. Stop codons cause the polymeraze to disassociate and terminate transcription. There are three universal stop codons in the standard genetic code. Additional regulatory sequences such as enhancers and silencers that increase and decrease gene expression levels, respectively. These regulatory regions can be locates within the gene sequence or in a very distant DNA region [Maston *et al.* (2006)].

In order for the genetic information to be passed on, it has to be replicated. It is one of the most essential processes during cell division, which in turn is the basis for the development of all organisms. DNA replication is catalyzed by the enzyme called *DNA polymerase*. During this process, DNA helix is unwound and hydrogen bonds are temporarily broken, enabling polymerase to bind to the DNA and synthesizes a new nucleotide chain using a single DNA strand as a template [Alberts *et al.* (2003)]. As a result of replication, two new double-stranded DNA molecules are created.

Mutations and errors

Despite the remarkable precision of DNA replication and the ability of the DNA polymerase to perform proof-reading and error correction, some errors do occur during the process. The error rate varies by species and is much higher in viruses than in eukaryotic cells [Drake *et al.* (1998)]. Errors in the genetic material can occur also during cell divisions, namely mitosis and meiosis, or due to damage caused, for example, by radiation or chemicals. Changes in DNA can be also a result of the mobile genetic elements movements, which can cause insertions or deletions. Alterations in the DNA sequence are called *mutations*.

The effect of a mutation on the organism depends on the size of the change, its

type and location. Errors that occur in a non-coding and non-regulatory parts of DNA are neutral for the organism. Small mutations, involving change of a single or several bases within gene coding sequence, can be silent if a codon changes to other codon that encodes the same amino acid. In such case the substitution is synonymous. A nonsynonymous substitution results in a codon encoding a different amino acid (missense mutation) or a stop codon (nonsense mutation). The first one changes the protein sequence, hence it influences the structure and possibly function of the protein. A nonsense mutation, in turn, causes the termination of the protein development. In both cases the resulting proteins are usually nonfunctional. Less often, we can talk about gain-of-function mutations, which changes the function of the protein. Aside from the nucleotide substitution, a mutation can be also caused by an insertion or deletion. Such an event changes the reading frame, i.e., the grouping of codons. The earlier in the sequence the mutation occur, the more the resulting protein is altered. The occurrence of a mutation within the regulatory part of the gene may affect the level of gene expression [Alberts *et al.* (2003)].

Another type of mutations are large-scale mutations, which include deletions and duplications of one or more genes, duplications of entire sets of chromosomes, and chromosomal rearrangement, i.e., large-scale changes in chromosome structure [Hastings *et al.* (2009)].

The rate of mutation varies for different mutations, genome regions and organisms. Regions characterized with very low mutation rate are called *conserved*.

Molecular evolution

Mutations as a consequence of errors and DNA damages may seem like just an imperfection in the DNA machinery. Indeed, mutations can cause protein alterations, which may lead to a loss of function and consequently stop biologically important processes. Such alterations may cause various types of diseases or even be lethal. Results of other mutations may not be immediately noticeable, however, the accumulation of certain types of mutations can cause malignant transformation from normal cell to cancer cell [Alberts *et al.* (2003)]. However, aside from the the negative impact, mutations are the source of all genetic variation and together with natural selection are the main driving force of evolution. Some mutations can be advantageous by increasing gene expression or positively affecting the function of protein. The main source of new genes is gene duplication, which creates a redundant gene copy. Duplicated genes (*paralogs*) are no longer under the selection pressure and may evolve by changing sequence and acquiring new function [Alberts *et al.* (2003)].

Mutations may or may not be inherited by offspring, depending on the type of organism and the cell in which the mutation occurred. In simple single-cell organisms mutations are present in all progeny cells. In multicellular organisms, on the other hand, we distinguish somatic mutations, which occur in somatic cells and are not passed on, and germline mutations, that are located in reproductive cells and consequently are observable in all cells of the offspring.

Homology and similarity of sequences

The inheritance of genetic sequences along with the mutations allows to trace back the evolution by comparing DNA sequences. The more similar the sequences of two genes are, the more likely they are evolutionarily related. A group of genes that share a common sequence ancestor form a gene family and we say that their sequences are homologous. Genes from the gene family located within one genome or in genomes of different species are called *paralogs* and orthologs, respectively [Demuth *et al.* (2006)].

To determine similarity and potential evolutionary relationships, the sequences must be compared in some way. The comparison can be made by matching two sequences of nucleotides in a way that gaps are inserted between the nucleotides or amino acids so the identical characters are aligned in successive columns. Such a comparison is called *alignment*. Characters that do not match can also be paired if it will result in a better score. The score function for DNA and RNA sequences simply adds a set value for a match, and subtracts for a mismatch or a gap. In case of protein alignments the substitution matrix is often used to assign scores to amino-acid matches or mismatches. We distinguish two types of alignments – global and local. Global methods search for the optimal alignment for entire sequences. In turn, local alignments matches only the most similar fragments. Extension of the classical alignment method, called *multiple sequence alignment* (MSA), allows to align more than two sequences at a time. Such an alignment identifies similar regions in a set of sequences and helps in the analysis of evolutionary relationships. [Gollery (2005); Needleman and Wunsch (1970); Smith *et al.* (1981)].

Phylogenetic trees: concepts and methodologies

Phylogenetic tree is a diagram showing evolutionary relations between species, genes or proteins. The relationships can be determined by physical or genetic similarities. In a phylogenetic tree, all nodes represent species or sequences of genes or proteins, although usually the internal nodes are unknown and the tree contains only data from the present day represented by tree leaves [Felsenstein and Felenstein (2004)].

In a species tree each branching corresponds to a *speciation*, i.e., the formation of new distinct species. Branching in the gene tree usually represents speciation or duplication of a gene. Trees can be either binary or multifurcated. Multifurcations occur mainly when a few sequences are too similar to correctly determine their evolutionary relationship. Tree can be also rooted or unrooted. A more natural way to represent evolutionary relationships is a rooted tree, where the root corresponds to the common ancestor of all of the nodes in the tree and the direction of evolution is clear. An unrooted tree shows relations, however, for certain nodes the ancestor–descendant relationship may be impossible to determine. There are several methods for rooting unrooted trees, such as outgroup rooting or midpoint rooting, but the rooting location determined by these methods is not always correct [Boykin *et al.* (2010)].

There exist several methods for phylogenetic tree inference. The first category is distance-matrix methods, which includes methods such as UPGMA and neighbour-joining (NJ). Distance-matrix methods are based explicitly on the evolutionary distance between sequences, calculated with MSA. UPGMA is a agglomerative, hierarchical clustering method, which returns rooted trees. At each step the algorithm connects two closest groups of nodes and iterates until the tree is completely resolved. Neighbour-joining is similar with two main differences: it does not assume the same evolution rate for all branches and it returns and unrooted tree. The advantage of the NJ is its speed comparing to other methods [Saitou and Nei (1987)].

Maximum parsimony method minimizes the number of character-state changes observed along the branches [Fitch (1971)]. It is very intuitive method, however, existing algorithms are very slow, which makes its application to large data sets impossible. Another disadvantage of this method is the fact that inferred trees often underestimate the actual number of mutations that has occurred [Felsenstein (1978)].

The next method, unlike the previous ones, is a probability-based method. Maximum likelihood methods aim to find the most likely tree given the available data. For this method a substitution model is required to calculate the probability of particular mutations. Maximum likelihood seeks a similar tree as a maximum parsimony method, but it assume different evolution rates on branches. Due to its complexity heuristic solutions are used for the tree inference.

Non-tree evolution and phylogenetic tree extensions

Although many evolutionary relationships can be represented by phylogenetic trees, not all phenomena can be visualized in a tree-like structure. Reticulation events

cause the appearance of new, non-tree branches and introduce the need to define phylogenetic networks [Linder *et al.* (2004)]. The processes driving the reticulate evolution include symbiosis and symbiogenesis, which are forms of coexistence between two distinct species that benefit both or one of them. A special case of symbiosis in which one species resides in the second one is called symbiogenesis. The evolutionary theory of symbiogenesis explains the origin of eukaryotic cells and organelles such as mitochondria and chloroplasts, with cells formerly living in endosymbiosis. Such close interaction of two organisms may lead to evolutionary changes in both species.

Another example of a reticular event is horizontal gene transfer (HGT). In the HGT event, the genetic material is transferred between two separate organisms in a horizontal manner rather than vertically from parent to offspring. HGT mechanisms include bacterial conjugation between two bacterial cells, bacterial gene transfer agents, transformation due to the uptake of foreign DNA from the environment, and transduction whereby DNA is transmitted between bacteria by viruses. Event of HGT are though to be the main mechanism for the spread of antibiotic resistance in bacteria and play important role in their evolution. A gene found in a different species as a result of HGT is called *xenolog*.

While horizontal gene transfers refers to the transmission of the DNA fragments at the cellular level, the hybridization phenomenon concerns the cross-breeding of two parental organisms from subspecies or two distinct species to produce offspring. Hybrid offspring may be infertile, however, they are fertile in most cases and able to breed with both parental line [Gontier (2015)].

Genes, genomes and evolutionary events

As mentioned before, genome is a repository of an organism's genetic information, containing genes and the regulatory regions that control the gene expression. Both the evolution of species and the evolution of the gene families encoded in their genomes can be represented by phylogenetic trees. However, the presence of evolutionary events, such as duplications and losses, may cause the topologies of the gene and species tree to be incongruent. Therefore, the question remains, how to represent the joint evolution of genes and species.

A method linking two topologies and explaining the deferences between them is the tree reconciliation. Informally, reconciliation approach infers the evolutionary scenario for given gene and species trees, by embedding the gene tree into the species tree [Goodman *et al.* (1979); Page (1994)]. The algorithm minimizes the cost of the embedding, so a model defining allowed evolutionary events and their costs

must be specified. The most commonly used model is the duplication-loss model. The embedding shows duplications and losses needed to fit the gene tree into the species tree, indicating in the gene tree the nodes where these events likely occurred.

Credibility of evolutionary relationships

Phylogenetic trees inferred based on the DNA sequence similarity may be incorrect for a variety of reasons. Among the most common are DNA sequencing errors, high similarity of the studied sequences, and limitations of the tree inference methods. Incorrect tree topologies can also be caused by the presence of reticulation events, such as horizontal gene transfer (HGT) or hybridization, which introduce disruptions to the sequences.

Incorrect topologies can cause further problems in data analysis. The correctness of evolutionary events inferred by the topology-based reconciliation method depends tremendously on the correctness of the reconciled trees. The high error sensitivity and limitations of the reconciliation method make the credibility of evolutionary events, and the development of reliable methods for the inference of reticulation events, still a significant problem in the field of phylogenetics.

2
Basic definitions

The Chapter *Basic definitions* introduces several basic concepts from the field of phylogenetics. We start with the formal definitions of the gene and species tree structures. Further, we explain the concept of the tree reconciliation and reconciliation cost, at first only for rooted trees and a set of evolutionary events limited to gene duplications and losses. Then, we show how to extend this model to the unrooted trees. Finally, we raise the issue concerning the credibility of inferred phylogenetic trees and present a method to address this problem.

2.1. Phylogenetic trees

Gene and species trees

A *species tree* we define as a rooted tree whose leaves are identified with *species*. For a tree T by $\mathcal{L}_T$ we denote the set of all leaves/species, and by I_T the set of all internal nodes present in T. A rooted gene tree G over a species tree S is a triple $\langle V_G, E_G, \Lambda_G \rangle$ and $\langle V_G, E_G \rangle$ is a rooted tree in which an internal node has at least two children, and $\Lambda_G \colon \mathcal{L}_G \to \mathcal{L}_S$ is the leaf labelling function expressing the relationship between leaves from G representing genes and species from the species tree S, called simply *labelling*. For vertices $a, b \in V_G$, we use the binary order relation $a \preceq b$ if b is a vertex

on the path between a inclusively, and the root of G. We write that a and b are *comparable* if $a \preceq b$ or $a \succeq b$. If $a \preceq b$ and $a \neq b$, we use the notation $a \prec b$. By $\mathsf{par}(a)$ we denote the parent of a non-root node a. A node is binary if it has exactly two children; it is non-binary otherwise. We call a tree binary if all of its nodes are binary. A child a of a binary node v is a sibling of b, denoted $a = \mathsf{sib}(b)$, if $\mathsf{par}(a) = \mathsf{par}(b) = v$. The set of all children of a node a is denoted by $\widehat{a}$. If a is a leaf then $\widehat{a} = \emptyset$. By $\|a, b\|$ we denote the number of edges on the shortest path connecting a and b. A *cluster* for v is the set of all leaves present in the subtree of T rooted at v.

In this work, gene and species trees are usually denoted by G and S, respectively. We denote trees by using the standard nested parenthesis notation with the extension that allows to encode labelling. For instance, in Figure 2.1, $G = ((a, a), (b, (b, c)))$ is a five-leaf gene tree over a species tree $((a, b), c)$ such that two leaves of the gene tree labelled with a are assigned to species a.

For convenience, we also assume that a species tree is leaf labelled, where the labellig is given by the identity function, i.e., each leaf/species in S is labelled. This convention is used in Chapters 3 and 4.

Unrooted gene tree

An unrooted gene tree G over a species tree S is a triple $\langle V_G, E_G, \Lambda_G \rangle$ such that $\langle V_G, E_G \rangle$ is an undirected acyclic connected graph in which the degree of each internal node is 3, all leaves has degree 1, and $\Lambda_G \colon \mathcal{L}_G \to \mathcal{L}_S$ is a leaf labelling function, where $\mathcal{L}_T$ is the set of all leaves in an unrooted tree T. A *split* $A|B$ is a partition of X, i.e., A and B are two disjoint non-empty sets such that $A \cup B = X$. We say that a split $A|B$ is present in an unrooted tree if there is an edge e in G, such that removing e from G induces two subtrees of G having A and B as the set of its leaves, respectively. Splits are the unrooted equivalent of clusters.

2.2. Tree reconciliation under the duplication-loss model

Events such as gene duplications and losses can lead to the incongruence of the gene and its species tree topologies. To locate such events, one can use tree reconciliation. Below we introduce the tree reconciliations method, and show its extension to unrooted trees.

2.2.1. Tree reconciliation

Tree reconciliation is a method linking two tree topologies. By embedding the gene tree into the species tree, this method infers the evolutionary scenario that explains the incongruences between given trees, by postulating duplication and loss events.

For the definition of the tree reconciliation, we need the notion of the lca-mapping. Let $\mathcal{L}_T$ be the set of all leaf labels from a tree T. Let $G = \langle V_G, E_G \rangle$ be a rooted gene tree, such that $\mathcal{L}_G \subseteq \mathcal{L}_S$, where $S = \langle V_S, E_S \rangle$ is the rooted species tree. *The least common ancestor mapping*, or *lca-mapping* is a function $\mathsf{M} : V_G \to V_S$, such that:

$$\mathsf{M}(g) = \begin{cases} s & \text{if } g \text{ is a leaf,} \\ \mathsf{M}(g') \oplus \mathsf{M}(g'') & \text{otherwise,} \end{cases}$$

where g' and g'' are children of the internal node g and $v \oplus u$ is the lowest common ancestor of v and u. The internal node $g \in V_G$ is called a *duplication* (D) if $\mathsf{M}(g) = \mathsf{M}(g')$ for some child g' of the node g. Remaining nodes are called *speciation* nodes. An example of tree reconciliation is depicted in Figure 2.1. Lca-mapping is shown in Figure 2.1 on the left and the corresponding embedding of the gene tree G into the species tree S representing evolutionary scenario is on the right side.

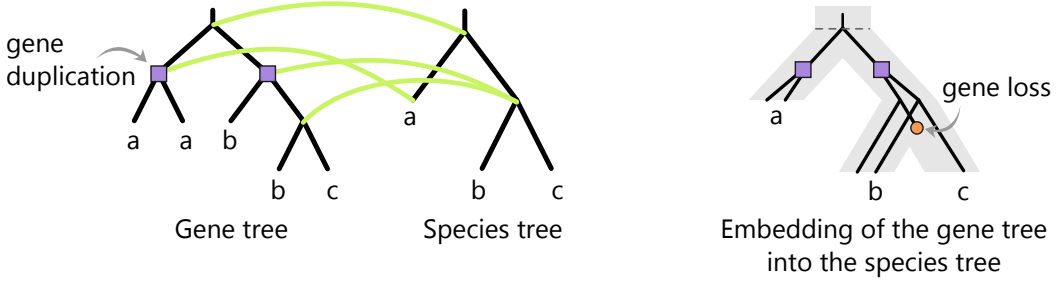


Figure 2.1: *An example of lca-mapping and evolutionary scenario. Left:* The lca-mapping between the gene tree G and the species tree S (leaves mappings are omitted). *Right:* The embedding representing evolutionary scenario that corresponds to the lca-mapping. Here, for reconciling G and S two duplications and one gene loss were needed.

2.2.2. Duplication-loss cost

For the tree reconciliation method, a cost function determining the allowed types of evolutionary events and their costs, must be specified. In the duplication-loss model (DL model), as the name suggests, there are two types of events: duplications and losses. Gene duplications are the main mechanism by which new genetic material is created in the process of molecular evolution [Taylor and Raes (2004)]. Gene losses, on the other hand, results in the loss of some genetic information. The effect of

these two phenomena on DNA makes a model based on them appear to be a good approximation of evolution.

For a given tree T and S, the *duplication cost*, denoted by $\mathrm{D}(T,S)$, is the total number of duplications required to reconcile T and S [Page (1997)]. The total number of *gene losses* in T can be defined by:

$$\mathrm{L}(T,S) = 2\,\mathrm{D}(T,S) + \sum_{\substack{g \text{ is internal} \\ a,b \text{ children of } g}} (\|\mathsf{M}(a),\mathsf{M}(b)\| - 2),$$

where $\|a,b\|$ is the number of edges on the path connecting a and b in S [Ma *et al.* (2000a)]. Finally, we can define the *duplication-loss cost* (DL cost):

$$\mathrm{DL}(T,S) = \mathrm{D}(T,S) + \mathrm{L}(T,S).$$

2.2.3. Unrooted tree reconciliation

In Section 2.2.1 we presented the notion of tree reconciliation. Although the classical model applies only to rooted trees, it can be extended to reconcile an unrooted gene tree with a rooted species tree by seeking a rooting of the unrooted gene tree that invokes the minimal duplication-loss cost [Górecki and Tiuryn (2007a); Yu *et al.* (2011)].

For an unrooted gene tree G and an edge $e \in E_G$, by G_e, we denote the *rooting*, i.e., a rooted gene tree, obtained from G by placing the root on e. By M_e we denote the lca-mapping between G_e and S. For a species tree S, such a rooting induces the duplication-loss cost $\mathrm{DL}(G_e,S)$. The set of all edges with the minimal duplication-loss cost, or *optimal edges*, is called *plateau*. Rootings of optimal edges are also called *optimal*. We now introduce types of edges and stars in unrooted gene trees. Both notions are crucial in comparing gene and species trees in unrooted framework [Górecki *et al.* (2013)].

Without loss of generality we assume that every root of a gene tree is mapped into the root of S, denoted by $\top$, and both trees are non-trivial. An edge $e = \{v,w\}$ of G is *empty* if the root of G_e is a speciation. We call e *double* if $\mathsf{M}_e(v) = \top = \mathsf{M}_e(w)$. Otherwise, e is called *single*. A single edge e is called *v-incoming* or *w-outgoing* if $\mathsf{M}_e(v) \neq \top = \mathsf{M}_e(w)$. An edge is called *symmetric* if it is either empty or double. Let v be an internal node of G. Then a *star* with a *center* v consists of three edges sharing a common node v incident to nodes a, b and c, respectively (see Figure 2.2). There are five types of possible star topologies:

1. the $S1$ star has one v-incoming edge and two v-outgoing edges,

2. the $S2$ star has exactly two v-outgoing edges and one empty edge,

3. the $S3$ star has two v-outgoing edges and one double edge,

4. the $S4$ star all 3 edges are double,

5. and the $S5$ star has one v-outgoing edge and two double edges. An example of the $S5$ is shown in Figure 2.3.

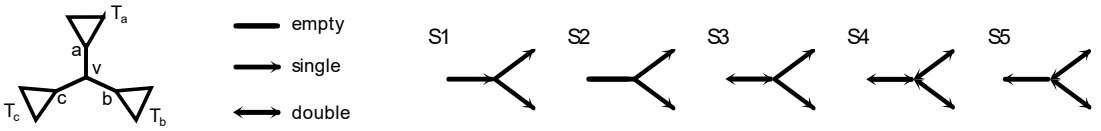


Figure 2.2: Stars in unrooted reconciliation. A star in a gene tree and possible types of edges and stars topologies. Subtrees T_a, T_b and T_c are rooted at nodes a, b and c, respectively.

The main result on unrooted reconciliation is presented below.

Theorem 1 (Adopted from Górecki and Tiuryn (2007a))**.** *For a given unrooted gene tree G, we have: (1) either G has exactly one empty edge or G has at least one double edge, (2) if the plateau of G consists of exactly one edge then this edge is symmetric, and all other edges are single, or (3) if the plateau of G has more than one edge then it is composed only of all edges present in stars $S4$ and $S5$, and all other edges are single.*

It follows from this theorem and the properties of stars that the the graph induced by the plateau is an unrooted binary tree. See also [Górecki *et al.* (2013); Górecki and Tiuryn (2007a)] for more details. An example of the unrooted reconciliation is depicted in Figure 2.3.

2.3. Felsenstein's phylogenetic bootstrap

Phylogenetic trees are often inferred from partial and noisy data due to the character of the sequencing methods. Moreover, results obtained by tree inference methods for the same data set often vary, and the correctness of the resulting trees is not guaranteed. Thus, a necessary step in phylogenetic research is to assess the credibility of the inferred trees. A commonly used method is a non-parametric bootstrap, proposed by Felsenstein [[Felsenstein (1985)]], which allows determining whether the given phylogenetic tree is a good approximation of the evolution of given sequences.The method is divided into three steps. The first one consists of creating new alignments by drawing with replacements columns from the original alignment. Obtained alignments are used to infer *sample trees*, which are compared with

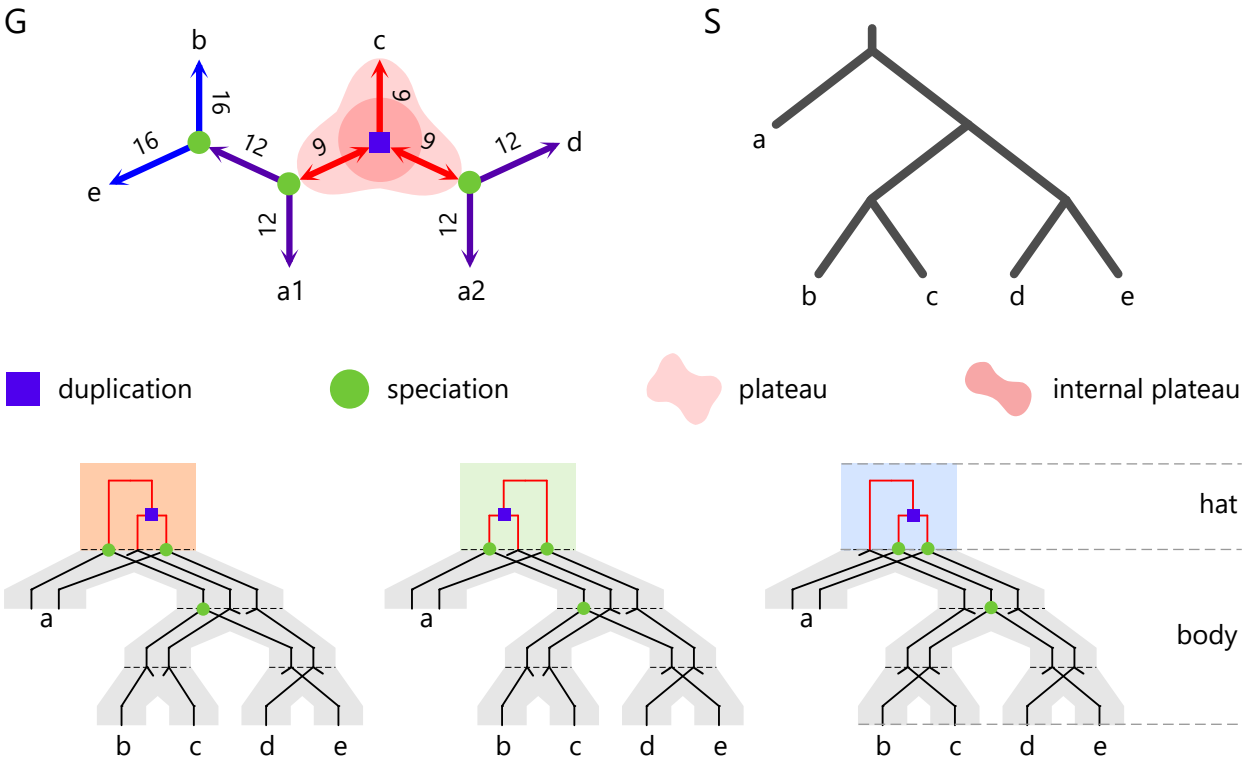


Figure 2.3: *Top*: An example of unrooted gene tree G, reconciled with S, with three optimal rootings. Each edge is decorated with the duplication-loss cost of the corresponding rooting (optimal cost is 9). G has three non-leaf speciation clusters (marked by green circles) and two duplication clusters (one marked by purple square + the cluster of the root). Every plateau edge, colored in red, has a label E1, E2 or E3, which relates to one of the embeddings visible below. *Bottom*: Three embeddings of all optimal rootings of G into S. Corresponding edges in embeddings are color coded. Every embedding has 2 duplications and 7 gene losses.

the original tree in the third and final step. The results can be interpreted as an indication of the influence of arbitrary changes that do not resemble the evolutionary pattern, such as sequencing errors, on the topology of the phylogenetic tree. Support values can be computed for both branches or clusters present in the tree and they are calculated as a percentage of sample trees in which an identical branch or cluster was found. The visualization of the bootstrap method and the calculation of the support values are presented in Figure 2.4.

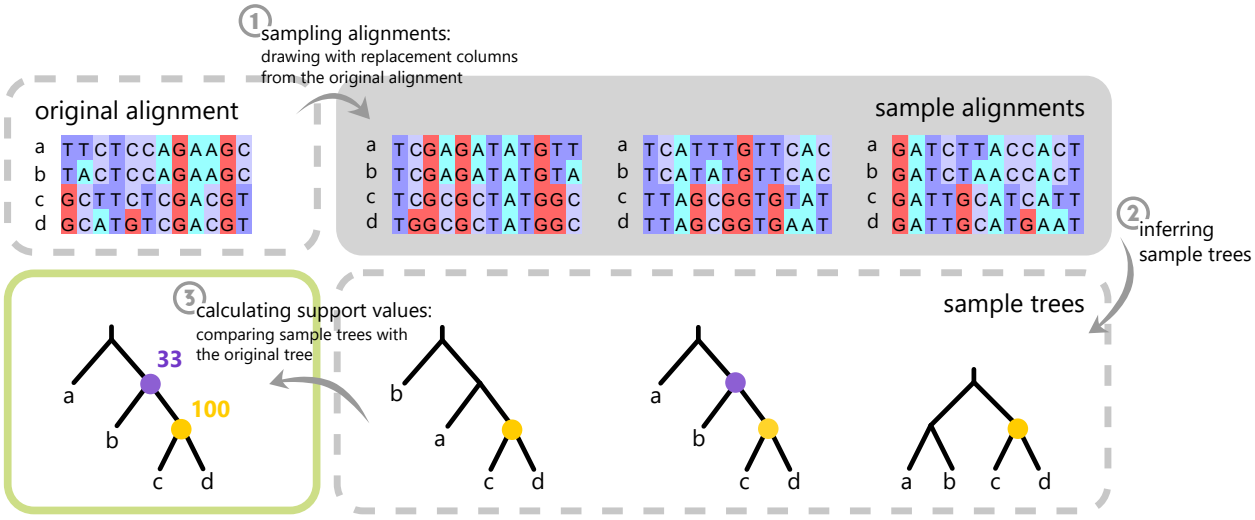


Figure 2.4: Schema showing the three steps of calculating support values for phylogenetic trees using the bootstrap method.

3

Credibility of Duplication and Speciation Events

Duplication and loss events in the evolutionary history of a gene family can be located using the tree reconciliation method. However, while the classical reconciliation model applies only to rooted trees, most standard tree inference methods return trees that are unrooted. One solution to this problem is to root the tree. There are several rooting methods, but the problem of identifying a credible rooting placement is not trivial [Górecki and Eulenstein (2012a)]. The outgroup rooting can result in incorrect rootings if there is heterogeneity in a gene tree. Another approach is to root gene trees under the molecular clock assumption, or similarly by using midpoint method. In both cases the results may be incorrect if there is a molecular rate variation throughout the tree [Holland *et al.* (2003); Huelsenbeck *et al.* (2002)]. In [[Chaudhary *et al.* (2012); Durand *et al.* (2006)]] the proposed solution was to use tree edit operations to correct trees before reconciliation. For example, in [Chaudhary *et al.* (2012); Górecki and Eulenstein (2012a)] a gene tree is considered to have an error if there is a tree with improved reconciliation cost in the local neighborhood of the given gene tree. Another method proposed in [Beretta and Dondi (2014); Swenson *et al.* (2012); Dondi *et al.* (2014)] consisted in preprocessing a set of gene trees by removing nodes that cause inconsistencies. A more statistical approach was presented in

[Wu *et al.* (2012)]. Certain refinement methods can be also applied to trees with polytomies [Lafond *et al.* (2014); Noutahi *et al.* (2016)].

An approach to the problem from a different angle is to extend tree reconciliation to reconcile an unrooted gene tree with a rooted species tree. Such solutions were presented in [[Górecki and Tiuryn (2007a); Yu *et al.* (2011)]], where the reconciliation method seeks a rooting of the unrooted gene tree that invokes the minimal duplication-loss cost. However, this is not a complete solution to the problem. In addition to the incorrect root location, there may be other errors in the gene tree topology due to sequencing errors or limitations of the tree inference methods. For this reason, the credibility of tree reconciliation results remains an vital issue in phylogenetic studies.

A commonly used method to address the issue of the credibility of inferred phylogenetic trees, in non-parametric bootstrap [[Felsenstein (1985)]]. So far, bootstrap methods in tree reconciliation were mainly focused on rooted trees. For example, in [[Park *et al.* (2010)]], the authors proposed to estimate the support of inferred horizontal gene transfers. An early approach to integrate reconciliation and bootstrapping for unrooted trees [Zmasek and Eddy (2002)] relied on rooting unrooted trees by choosing the midpoint rooting with the minimal duplication-loss cost. By aggregating gene duplication locations from all rooted sample trees, the authors were able to compute support values for every duplication from the input gene tree. However, in such a method, information can be lost as the selected midpoint rooting may be incorrect.

In this Chapter, we present our approach to the issue of credibility of tree reconciliation results. We propose a solution combining methods of unrooted reconciliation and non-parametric bootstrap. We show the properties of optimal rootings and we define the main notion of support values for duplication and speciation events. Based on the theoretical properties we propose an algorithm for the computation of the support values. Next, we present a comparative study of tree rooting methods and evaluation of their performance using both simulated data and real yeast genomes. Finally, we examined how the inclusion of support values for bootstrapped trees affects the results of supertree inference.

3.1. Theoretical results

In this Section we present theoretical results related to reconciliation with bootstrapping. First, we introduce support values for branching events, i.e., duplications and speciations in a gene tree. Next, we show our linear time algorithm for the

computation of bootstrap values, and finally, we explain the correspondence of our approach with the classical non-parametric bootstrapping.

3.1.1. Properties of optimal rootings

We start by introducing some important notions and theorems concerning unrooted gene trees and the *plateau* property. Please recall, that the *plateau* is the set of all edges with the minimal duplication-loss cost.

Lemma 1. *If e and e' are elements of the plateau of an unrooted gene tree G, then $\mathsf{M}_e(w) = \mathsf{M}_{e'}(w)$ for any $w \in V_G$.*

Proof. Without loss of generality we assume that e and e' are two elements of a star with center v. It is clear that for any node w from T_a, T_b and T_c (see notation from Figure 2.2), we have $\mathsf{M}_e(w) = \mathsf{M}_{e'}(w)$. It remains to show that lca-mappings of v are the same in both rootings. Both e and e' are in the plateau, thus at least one edge, say e, is double and the second edge is either v-outgoing or double. Now, from the types of edges we have $\mathsf{M}_e(v) = \top$ and the same holds for the other rooting. This completes the proof for incident edges. The rest of the proof follows easily by induction. ■

Note that the root of any rooting of the unrooted gene tree is mapped to the same node in S. Despite the formal complication with the root (which is formally a new node in a rooting of unrooted tree), we conclude that lca-mappings of optimal rootings are identical. The next result follows from Lemma 1.

Theorem 2 (Homology in unrooted trees)**.** *For a node of an unrooted gene tree G, its type is the same in every optimal rooting of G.*

Proof. It follows from Lemma 1. ■

Clusters in an unrooted gene tree G are inherited from its rootings as follows: $A \subseteq L_G$ is a cluster in G if there is a rooting T of G and a node in T such that A is the cluster of this node.

Lemma 2. *Let G be an unrooted gene tree and $A \subsetneq \mathcal{L}_G$ be a cluster of G. Then, if there are rootings T and T' of a node from G and nodes $v \in T$ and $v' \in T'$ such that the clusters of v and v' are equal then $v = v'$.*

Proof. Any v is element of V_G by the definition of rooting. Assume that two rootings T and T' of G have cluster A and $v \neq v'$. Hence, there is a non-empty path in G connecting these two nodes. Next, by rooting G somewhere on this path, we obtain a rooting that has two nodes with the same cluster A. This is a contradiction with the definition of cluster as a set of leaves (not labels). ■

The above result shows a correspondence between every proper cluster A from G and a node in G, denoted by $G \downarrow A$. For completeness, the cluster consisting of all leaves will be called *root-cluster* and it is not associated with any node. Examples of clusters are presented in Figure 3.1.

Now we introduce a notion of *type* of cluster in unrooted trees. A proper cluster A of G is called a *duplication* if $G \downarrow A$ is a duplication in an optimal rooting of G (or, by Theorem 2, equivalently in all optimal rootings). Similarly, we call the root-cluster a duplication if the root of any rooting is a duplication. Analogically, we define a *speciation cluster* in G. Note, that there may exists clusters without type, e.g., $\{a1, a2, c, d\}$ in the example from Figure 2.3.

Lemma 3. *There are four disjoint kinds of speciation and duplication clusters in unrooted gene trees: (1) clusters of internal nodes of the plateau, (2) clusters of leaves of the plateau, (3) clusters of nodes disjoint with the plateau, and (4) the root-cluster composed of all leaves. The last three kinds are present in every optimal rooting.*

Proof. It follows from the properties of optimal rootings, Theorem 2 and Lemma 2. ■

Now we summarize the properties of unrooted reconciliation. Optimal evolutionary scenarios can be represented by embeddings of an optimal rooting into the species tree [Górecki and Tiuryn (2006)]. From Lemma 1 and Theorem 3, we conclude that these scenarios differ only in their rooting edges (*hat*) while the remaining parts of all the trees (*body*) are identical as indicated in Figure 2.3 (see also [Górecki and Tiuryn (2007a)]).

3.1.2. Support values for evolutionary events

Our method is partially based on the classical non-parametric bootstrap proposed by Felsenstein. We are given a set X of n gene sequences and a multiple sequence alignment A (of dimension n rows and k columns) of sequences from X. First, N *bootstrap alignments* are constructed, where each bootstrap alignment is formed by randomly selecting k columns from A with replacement. Next, for each bootstrap alignment, an unrooted gene tree, called *sample tree*, is inferred by using some standard tree-building tool, e.g. PhyML [Guindon *et al.* (2009)]. Finally, a gene tree G is inferred from the alignment A. The frequency of clusters/splits present in sample trees indicates the support for the corresponding clusters/splits in G.

Based on the non-parametric bootstrapping we provide the main notion of duplication and speciation (D/S) support values.

Definition 1 (D/S support values)**.** *Given: a rooted species tree S and a collection of unrooted sample trees $\mathcal{U}$ such that all trees from $\mathcal{U}$ have the same set of leaves X (a set of genes) and the same labelling*[1] *$\Lambda\colon X \to \mathcal{L}_S$. Then, for a cluster $A \subseteq X$, the duplication support for A is defined as $b^{Dup}(A,\mathcal{U}) = \frac{1}{|\mathcal{U}|}|\{T \in \mathcal{U} : A \text{ is a duplication cluster in } T\}|$, and the speciation support as $b^{Spec}(A,\mathcal{U}) = \frac{1}{|\mathcal{U}|}|\{T \in \mathcal{U} : A \text{ is a speciation cluster in } T\}|$.*

Similarly to the standard non-parametric bootstrapping of phylogenetic trees we analyze the support values for the clusters of the gene tree inferred from the input alignment. An example is depicted in Figure 3.1.

Problem 1. *Given: a rooted species tree S, an unrooted gene tree G over S and a collection of unrooted sample trees $\mathcal{U}$ such that all trees from $\mathcal{U} \cup \{G\}$ have the same set of leaves X and the same labelling $\Lambda\colon X \to \mathcal{L}_S$. For each duplication and speciation cluster A in G compute:*

$$\sigma_G(A,\mathcal{U}) = \begin{cases} b^{Dup}(A,\mathcal{U}) & \text{if } A \text{ is a duplication cluster in } G, \\ b^{Spec}(A,\mathcal{U}) & \text{if } A \text{ is a speciation cluster in } G. \end{cases}$$

For a set of edges E_G in G, let $\hat{E}_G = \{\langle v,w\rangle, \langle w,v\rangle\colon \{v,w\} \in E_G\}$ and for a directed edge $\langle v,w\rangle \in \hat{E}_G$, by $c(v,w,G)$ we denote the cluster of v in the rooting $G_{\langle v,w\rangle}$[2]. There is one-to-one correspondence between clusters and directed edges, therefore, due to computational efficiency in our algorithm we assign support values to the directed edges only.

3.1.3. Algorithm

Algorithm 1 calculates the D/S support values. It relies on recognizing the type of a cluster determined by a directed edge on the basis of lca mapping adopted to unrooted trees. Let S be a rooted tree, G an unrooted tree over S and $\langle v,w\rangle \in \hat{E}_G$.

We start with two auxiliary functions defined in Algorithm 1 in the line 1. The function m represents lca-mappings in G while τ will be used to assign clusters types to nodes of rootings of G. In the first four lemmas we show several properties of these functions.

Lemma 4. *For a rooted tree S, an unrooted tree G over S and $\langle v,w\rangle \in \hat{E}_G$ we have $m(v,w,G,S) = \mathsf{M}_{\{v,w\}}(v)$, where $\mathsf{M}_{\{v,w\}}$ is the lca mapping between $G_{\{v,w\}}$ and S.*

Proof. The proof is by induction on the structure of G. If v is a leaf then the equality is straightforward. Otherwise if v is an internal node, then:

$$m(v,w,G,S) = m(x,v,G,S) \oplus m(y,v,G,S) = \mathsf{M}_{\{x,v\}}(x) \oplus \mathsf{M}_{\{y,v\}}(y) = \mathsf{M}_{\{v,w\}}(v).$$

[1] Note that in this definition all trees in $\mathcal{U}$ are over S.

[2] Recall that $G_{\langle v,w\rangle}$ denotes the rootings of G on the edge $\langle v,w\rangle$.

This completes the proof.

■

Recall that an edge is symmetric if it is double or empty.

Lemma 5. *The cluster $c(v,w,G)$ is present in an optimal rooting of G if and only if for $e=\{v,w\}$ one of the following cases holds:*

(a) e is v-incoming,

(b) v is an internal node of the plateau,

(c) or e is symmetric.

Proof. ($\Leftarrow$) Assume that *(a)-(c)* hold. We show that $c(v,w,G)$ is present in an optimal rooting of G:

(a) If e is a v-incoming edge, then we have two cases (see Theorem 1). If w is a center of star $S5$ then e is optimal. Otherwise, e is not an element of the plateau, hence w is located on the path connecting v and the root of every optimal rooting. Thus, cluster is always present in plateau rootings. We conclude that in both cases $c(v,w,G)$ is present in some optimal rooting.

(b)-(c) In both cases e belongs to the plateau (Theorem 1) thus, $c(v,w,G)$ is present in an optimal rooting of G_e.

This completes the first part of the proof.

($\Rightarrow$) For the next part let us assume that $c(v,w,G)$ is present in an optimal rooting. Then e is either within the plateau or outside it. If e is in the plateau then either e is symmetric (condition *(c)*), or e is single. If e is single then either v is internal node of the plateau (condition *(b)*) or it is located on the border; in this case e is v-incoming (condition *(a)*).

If e is not in the plateau then by Theorem 1 e is single. In addition, the subtree of G_e rooted at w contains all nodes from the plateau of G. Therefore, $M_{\langle v,w\rangle}(w)=\top$. We conclude that $M_{\langle v,w\rangle}(v)\prec\top$. Hence, e is a v-incoming edge (condition (a)). ■

Lemma 6. *We have the following properties of the predicates from the line 1 of Alg. 1.*

(a) `incoming`(v,w) *is satisfied iff $\langle v,w\rangle$ is an v-incomming edge.*

(b) `symmetric`(v,w) *is satisfied iff $\langle v,w\rangle$ is a symmetric edge.*

(c) `insideplateau`(v) *is satisfied iff v is an internal node of the plateau (i.e., it is not a plateau leaf).*

(d) `inoptrooting`(v, w, P, Q) *is satisfied iff the cluster* $c(v, w, P)$ *is present in some optimal rooting of* P *when reconciling with the species tree* Q.

Proof. Both (a) and (b) follow easily from the definition of incomming/symmetric edges and Lemma 4. By Theorem 1, a node is internal in the plateau if and only if it a center of star S4 or S5. Only such stars contain at least two symmetric edges. Hence, v has two neighbours x and y such that $\langle v, x\rangle$ and $\langle v, y\rangle$ are symmetric if and only if v is a center of a star $S4$ or $S5$. This completes the proof of (c). The last property (d), follows from Lemma 5. ■

Cluster types are represented in Alg. 1 by τ:

Lemma 7. *For a rooted tree* S*, an unrooted tree* G *over* S *and* $\langle v, w\rangle \in \hat{E}_G$ *value of* $\tau(v, w, G, S)$ *is:*

- *Dup iff* $c(v, w, G)$ *is a duplication cluster,*
- *Spec iff* $c(v, w, G)$ *is a speciation cluster,*
- *or None iff* $c(v, w, G)$ *is not present in an optimal rooting of* G.

Proof. It follows from Lemma 5 and Lemma 6 that $\tau(v, w, G, S) = None$ if and only if $c(v, w, G)$ is not present in an optimal rooting of G. Now, for the rest of the proof we assume that $c(v, w, G)$ is in an optimal rooting of G.

Observe that $v = G \downarrow c(v, w, G)$ (see Section 3.1.1). Thus, v is a duplication in $G_{\{v,w\}}$ if and only if $\mathsf{M}_{\{v,w\}}(v) = \mathrm{lca}_Q(\mathsf{M}_{\{v,w\}}(x), \mathsf{M}_{\{y,v\}}(y)) = \mathsf{M}_{\{v,w\}}(y)$. By Lemma 4 this is exactly the condition from the definition of τ. This completes the proof of first case.

Analogously it can be proven that $\tau(v, w, G, S) = Spec$ if and only if $c(v, w, G)$ is a speciation cluster. ■

Lemma 8. *Given two unrooted trees* G *and* T *with set of leaves* X *and two adjacent nodes* v *and* w *from* T*. Test if* $c(v, w, T)$ *is a cluster present in* G *can be done in constant time after linear time preprocessing.*

Proof. Let us fix one leaf, say ω, in G. Let R be a rooted tree obtained from G by placing the root on the edge incident to ω. To verify whether $c(v, w, T)$ is present in G we use efficient lca-queries [Bender and Farach-Colton (2000)] between a gene tree $T' = \langle V_T, E_T, id_X\rangle$ over a species tree $R = (\omega, R')$ (i.e. T' is obtained from T by introducing identity labelling). For simplicity we denote $c(v, w, T)$ by A.

We have two cases depending on whether ω is present in A. If ω is not present in A then A is a cluster in R' and $\mathsf{M}_{\{v,w\}}(v)$ is also a node from R' , where $\mathsf{M}_{\{v,w\}}$

Algorithm 1 Computing D/S Support Values

1: **Auxiliary definitions**. For a rooted tree Q, an unrooted gene tree P such that $L \subseteq \mathcal{L}_Q$ and $\langle v, w\rangle$ in $\hat{E}_P$:

$$m(v,w,P,Q) := \begin{cases} \Lambda_P(v) & v \text{ is a leaf in } P, \\ m(x,v,P,Q) \oplus m(y,v,P,Q) & v \text{ is internal and } \{x,y\} = ch(v,w), \end{cases}$$

$$\tau(v,w,P,Q) := \begin{cases} None & \text{not } \texttt{inoptrooting}(v,w,P,Q), \\ Dup & \texttt{inoptrooting}(v,w,P,Q),\ \{x,y\} = ch(v,w) \text{ and} \\ & m(v,w,P,Q) = m(x,v,P,Q) \text{ or } m(v,w,P,Q) = m(y,v,P,Q), \\ Spec & \text{otherwise}, \end{cases}$$

where for an internal node $v \in V_P$

- $ch(v,w) = \{x,y\}$ such that $\{x,y,w\}$ is the set of all neighbours of v;
- here $\top$ is the lowest node in Q whose cluster contains $\Lambda_P(L_Q)$;
- $\texttt{incoming}(v,w) := m(v,w,P,Q) \neq \top = m(w,v,P,Q) = \top$;
- $\texttt{symmetric}(v,w) := m(v,w,P,Q) = \top = m(w,v,P,Q) = \top$ OR $m(v,w,P,Q) \neq \top \neq m(w,v,P,Q)$;
- $\texttt{insideplateau}(v) :=$ ∃ siblings x and y of v such that: $x \neq y$ AND $\texttt{symmetric}(v,x)$ AND $\texttt{symmetric}(v,y)$
- and $\texttt{inoptrooting}(v,w,P,Q) := \texttt{incoming}(v,w)$ OR $\texttt{insideplateau}(v)$ OR $\texttt{symmetric}(v,w)$.

2: **Input/output:** See Problem 1. Let $\mathcal{U} = \{T_1, T_2, \ldots, T_N\}$.
3: Fix $\omega \in X$. Let $R := G_e$, where e is the edge incident to ω. For a node $g \in V_G$, let $\pi(g)$ denote the parent of g in R if it is not the root of R, otherwise $\pi(g)$ is the sibling of g. Note that $\pi(g)$ is an element of V_G and $\{g, \pi(g)\} \in E_G$. For each i, let T_i' be the unrooted gene tree over R obtained from T_i by replacing the labelling with the identity function on X.
4: Init lca-structures for S and R. For $\langle v,w\rangle \in \hat{E}_G$, $\#(v,w) := 0$ // reset cluster counters
5: **For** each $i \in 1, 2, \ldots, N$
6: **For** each $\langle v,w\rangle \in \hat{E}_{T_i}$ such that $\tau(v,w,T_i,S) \neq None$
7: **If** $\omega \notin c(v,w,T_i)$ **Then**
8: $g := m(v,w,T_i',R)$
9: **If** $|c(v,w,T_i)| = |c(g,\pi(g),G)|$ AND $\tau(v,w,T_i,S) = \tau(g,\pi(g),G,S)$
10: **Then** $\#(v,w)++$
11: **Else**
12: $g := m(w,v,T_i',R)$
13: **If** $|c(w,v,T_i)| = |c(g,\pi(g),G)|$ AND $\tau(v,w,T_i,S) = \tau(g,\pi(g),G,S)$
14: **Then** $\#(v,w)++$
15: **Return** $\#(v,w)/|\mathcal{U}|$ for each $\langle v,w\rangle \in \hat{E}_G$ such that $\tau(v,w,G,S) \in \{Dup, Spec\}$.

is the lca-mapping between $T'_{\{v,w\}}$ and R. If the cluster of $\mathsf{M}_{\{v,w\}}(v)$ has the same number of elements as A then both clusters are equal. For the second case, we use the observation that if A is present in G and T then $X \setminus A$ is also present in both trees. Therefore if ω is present in A it is equivalent to test the cluster $c(w, v, T) = X \setminus A$ analogously to the first case by using $\mathsf{M}_{\{v,w\}}(w)$.

The size of clusters in both trees can be computed once in linear time. Lca-mappings stored in $m(v, w, T', R)$ require $\mathcal{O}(|R|)$ preprocessing. Having these data structures, the test can be completed in $\mathcal{O}(1)$ time. ■

Theorem 3. *Algorithm 1 computes D/S support values in linear time.*

Proof. Correctness. In the main part of the algorithm (lines 5-8) we increase the counter of events $\#(v, w)$ when the cluster $c(v, w, G)$ is present in one of the sample trees (T_i) and has the same type. The test if $c(v, w, T_i)$ is present in G is composed of two cases according to the cases from Lemma 8. To check types of clusters we use τ. See Lemma 7 for the correctness of τ. Lemma 4 describes calculating of mapping values. For every cluster present in an optimal rooting the algorithm returns the number of sample trees in which it occurs with the same type divided by the number of bootstrap trees. This is exactly the D/S support value for this cluster as defined in Definition 1.

Time complexity. The main loop of the algorithm iterates over all directed edges of all bootstrap trees T_i. For each edge we check the type of the cluster defined by this edge in T_i which can be done in constant time. Afterwards, the existence of that particular cluster in the tree G is tested. According to Lemma 8 this can be done in a constant time with linear preprocessing. Next, the type of the cluster in G is checked again in constant time. Thus the total algorithm cost is linear.

■

3.1.4. Correspondence to classical bootstrap

Now we present the correspondence between D/S support values and the support values from Felsenstein's bootstrapping. We use the notation from Section 3.1.2 and Definition 1. For a split $A|B$ the support for $A|B$ in $\mathcal{U}$ is defined by $s^u(A|B, \mathcal{U}) = \frac{1}{|\mathcal{U}|}|\{T \in \mathcal{U} : \ A|B \text{ is a split in } T\}|$.

Theorem 4. *For a collection of unrooted gene trees $\mathcal{U}$ and a split $A|B$, $s^u(A|B, \mathcal{U}) \geq b^{Dup}(A, \mathcal{U}) + b^{Spec}(A, \mathcal{U})$.*[3]

[3]Observe that B is not present in the right side of the inequality.

Proof. It is sufficient to show that the inequality holds when $\mathcal{U}$ consists of a single gene tree T. If $A|B$ is present in T, then $s^u(A|B,\mathcal{U}) = 1$. If A is present in an optimal rooting of T, then it is either a speciation or a duplication and both sides of the inequality are equal. Otherwise, the support for duplication/speciation is zero. Finally, if $A|B$ is not present in T then $s^u(A|B,\mathcal{U}) = 0$ and A cannot be a cluster in any rooting of T. In such a case both supports are 0. ■

Similarly, we define the support for a cluster A in a collection of rooted trees $\mathcal{R}$: $s^r(A,\mathcal{R}) = \frac{1}{|\mathcal{R}|}|\{T \in \mathcal{R}\colon \text{ A is a cluster in } T\}|$. The D/S support values can be naturally extended to collections of rooted gene trees by replacing the term "cluster" with "node" in Definition 1. We omit the straightforward definitions.

Theorem 5. *For a collection of rooted gene trees $\mathcal{R}$ over the same set of leaves, and a cluster A, we have $s^r(A,\mathcal{R}) = b^{Dup}(A,\mathcal{R}) + b^{Spec}(A,\mathcal{R})$.*

Proof. As in the proof of the previous theorem, it is sufficient to show the equality for a singleton collection of trees. Let $\mathcal{R} = \{T\}$. We have two cases depending on the value of $s^r(A,\mathcal{R})$. If $s^r(A,\mathcal{R}) = 1$ then A is a cluster in T and it is either a duplication or a speciation. In such a case the equality holds trivially. Similarly, we have the equality when $s^r(A,\mathcal{R}) = 0$. ■

3.2. Experiments

We performed two computational experiments with bootstrapping and tree reconciliation on simulated and empirical data. In the first one, we show a comparative study of several rooting methods, conducted using our method to evaluate the correctness of each of the rootings. We also propose to solve the problem of rooting of an unrooted gene tree. The purpose of the second experiment is to test the quality of supertree inference from collections of well-supported gene trees.

3.2.1. A comparative study of rooting methods

Here we present how to use bootstrapping to evaluate the rooting problem, i.e., find the best possible rooting location for a given unrooted tree.

Simulated data preparation. In the first step, *model species trees* were generated using Mesquite [Maddison and Maddison (2015)], with topology generation performed according to the Yule-Harding distribution. The procedure is similar to this proposed in [Chaudhary *et al.* (2014)] with tree height set to 115 Myr, and the number of leaves equal 16.

Simulated gene trees were created from model species trees using a continuous time birth-death process [Arvestad *et al.* (2004)] with the gene duplication and gene loss events. On each lineage, an occurrence of gene duplication (bifurcation) or loss (termination) was drawn with a probability defined by a constant rate. As duplication should not change the height of a tree, a duplication node was added precisely at the point of the model tree edge in which a duplication event was postulated. In [Rasmussen and Kellis (2012)], three different values of rates of duplication and losses were proposed: 0.002, 0.004 and 0.008 events/gene per Myr. For greater diversity of gene trees, in our experiment, we additionally tested the rate of 0.012. For each simulated model tree, 1000 simulated gene trees were generated. For each of them, we simulated a nucleotide sequence alignment of length 100 under the GTR + Gamma + I model using Seq-Gen [Rambaut and Grassly (1997)]. Next, for each parameter rate $\lambda \in \{0.002, 0.004, 0.008, 0.012\}$, we obtained a set Sim_λ consisting of 1000 *unrooted gene family trees* inferred by *PhyML* program [Guindon and Gascuel (2003)] from the corresponding alignments. Finally, for each from each Sim_λ, we inferred a species tree S_λ by using the program *fasturec* [Górecki and Eulenstein (2012b)].

Empirical data preparation. We downloaded the set of 9 yeast genomes consisting of 4617 protein families from [The Génolevures Consortium (2009)]. After removing families with only two genes, we inferred 4141 gene trees by using *PhyML* with the standard parameter setting. Plateau sizes for all datasets are depicted in Figure 3.2.

Bootstrap processing. Further steps were performed for all datasets. For each alignment we created 100 bootstrap alignments by *Seqboot* from *PHYLIP* package [Felsenstein]. Finally, for each bootstrap alignment we inferred a sample tree by *PhyML*.

Experiment. In our study we compared five rooting methods by using the rooting score based on the D/S support values as follows. Given an optimal edge e from a gene tree G and a set of sample trees $\mathcal{U}$, *a rooting score for* e is the average support value of all non trivial (non leaf/root) clusters A from G_e. Formally,

$$r(e,\mathcal{U}) = \frac{1}{n-2} \sum_{\substack{A \text{ is a non-leaf/root} \\ \text{cluster in } G_e}} \sigma_G(A,\mathcal{U}).$$

We claim that the edges of the plateau having the maximal rooting score are the best candidates for rooting. We need two additional definitions. The *edge distance* between two nodes is the number of edges on the shortest path connecting these nodes. In the case when the gene tree has branch lengths, the *branch length distance* between two nodes is the total branch length of all edges on the shortest path connecting these nodes. We have three types of standard rooting methods. Two of

them take into consideration all tree edges [Farris (1972)] while the last one uses only edges included in the plateau [Zmasek and Eddy (2002)].

- **Midpoint edge rooting**: the root is placed in a half-way between two the most edge distant leaves.
- **BL-Midpoint rooting**: the root is placed in a half-way between two the most branch length distant leaves.
- **Midpoint plateau rooting**: the root is placed in a half-way between two the most distant nodes from the plateau.

Note that in our model of binary trees, the midpoint rootings may be non unique. For instance, if an unrooted tree has three leaves a, b and c, then the midpoint edge rooting can be $(a,(b,c))$, $(b,(a,c))$ or $(c,(b,a))$. The same property holds for the BL-midpoint rootings. Additionally, for a control, we tested two random rootings.

- **Random edge rooting**: the root is placed on the edge uniformly chosen from the set of all edges of a gene tree.
- **Random plateau rooting**: the root is placed on the edge uniformly chosen from the set of all edges of the plateau.

Results. The summary of results is depicted in Table 3.1. Our results suggest that the midpoint edge and BL-midpoint rooting methods indicate generally poorly supported rootings for the simulated datasets. Even the random edge rooting method performs better than these two methods. This observation partially holds for the yeast dataset with the difference that BL-midpoint rootings are generally better supported (1282 well supported rootings).

For the plateau based methods, the number of well supported rootings is usually high due to a large number of singleton plateaux present in our datasets. For example, in the dataset $X_{0.002}$, 843 out of 1000 trees have a unique rooting candidate in the plateau. Therefore, to compare these methods we analyzed non-singleton plateaux (see columns C). In the first simulated dataset, the ratio of optimal bootstrap rootings is 58% for the midpoint plateau rootings. This property can be explained by the fact that relatively large portion of trees has the plateau of size 3 (see Figure 3.2). In consequence, in such a case the midpoint plateau rooting method gives all three possible rootings which include the rooting maximal score. Next, the first dataset performed better than the other simulated datasets, which is due to usually more complex plateaux as indicated in Figure 3.2. In the empirical dataset, the midpoint plateau method inferred 46% rootings with the maximal score. However,

a better ratio for the non-singleton plateau trees was obtained for the BL-midpoint method. On the other hand, the latter method performed poorly for the trees with singleton plateaux.

In summary, the midpoint edge method is generally the worst, even the random edge method seems to be a better choice. Based on the simulation data, the same conclusion can be stated for the BL-midpoint method, however, we observed a better performance for the empirical dataset.

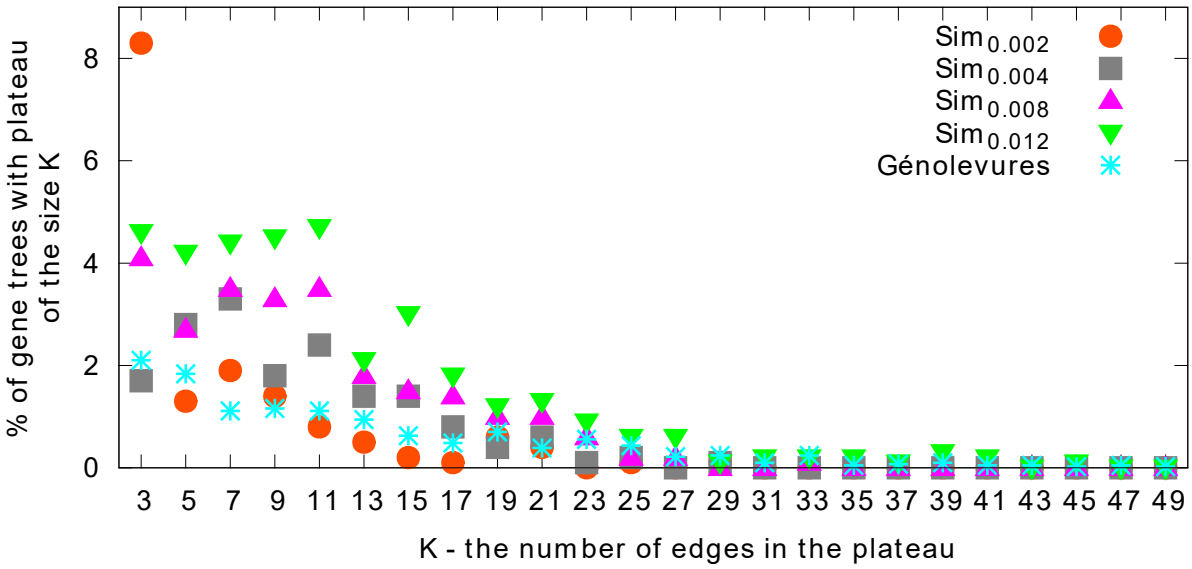
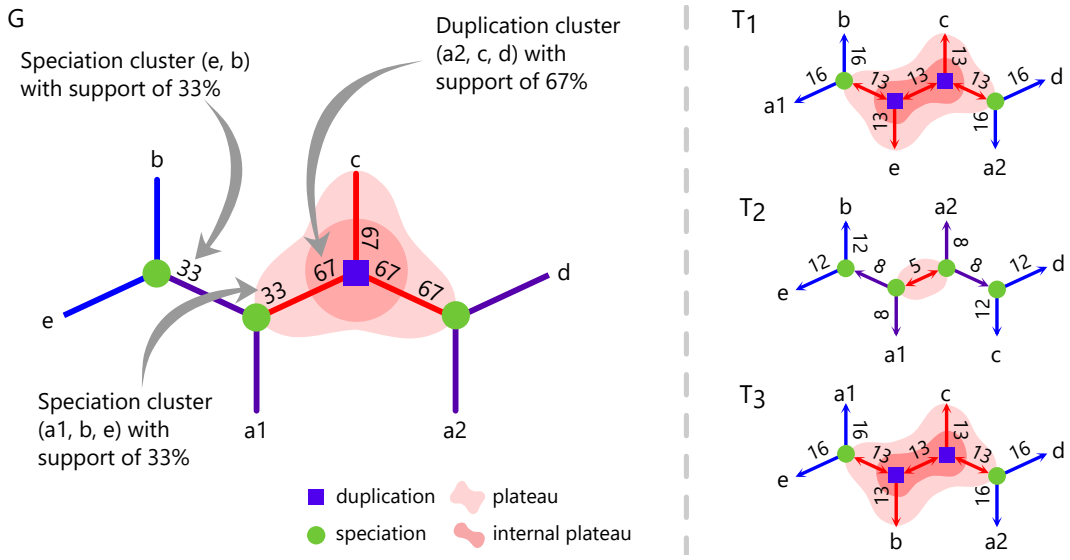


Figure 3.2: Frequency diagram of plateau sizes. The numbers of gene trees having singleton plateau (for $K = 1$), omitted here, are present in the second row of Table 3.1.



Cluster	T_1	T_2	T_3	b^{Dup}	b^{Spec}	G
a1 a2 b c d e*	Dr	Dr	Dr	1.0	0.0	Dr
a1 b e	Di	Sb	Di	0.67	0.33	Sb
a2 c d	Di	Sb	Di	0.67	0.33	Di
a2 d	Sb	-	Sb	0.0	0.67	Sb
a1 b c e	Di	-	Di	0.67	0.0	Di
a1 a2 b d e	Di	-	Di	0.67	0.0	Di
b e	-	So	-	0.0	0.33	So
a1 a2 b c d	Di	-	-	0.33	0.0	-
a1 b	Sb	-	-	0.0	0.33	-
a2 c d e	Di	-	-	0.33	0.0	-
c d	-	So	-	0.0	0.33	-
a1 a2 c d e	-	-	Di	0.33	0.0	-
a1 e	-	-	Sb	0.0	0.33	-
a2 b c d	-	-	Di	0.33	0.0	-

* the cluster of the root

Figure 3.1: *An artificial bootstrapping example (see Figure 2.3). Top left*: a gene tree G with D/S support values shown for non-leaf clusters present in optimal rootings (e.g., $\{a1, a2, c, d\}$ has no support). *Top right*: trees T_1, T_2 and T_3 - sampled from G. Edges of T_i's are decorated with the rooting cost (DL). *Bottom*: D/S support values for all non-leaf clusters from rootings of G, T_1, T_2 and T_3. Cluster type is denoted by S (speciation) and D (duplication). Additionally, r (root), i (plateau internal), b (plateau border) and o (outside plateau) denote the location of the cluster. For example, the duplication cluster $\{a2, c, d\}$ from G is present as a duplication located inside of the plateau of T_1, which is denoted by Di, while the same cluster determines a speciation located on the border of the plateau in T_2 (denoted by Sb).

	$Sim_{0.002}$			$Sim_{0.004}$			$Sim_{0.008}$			$Sim_{0.012}$			Génolevures		
Dataset size	1000			1000			1000			1000			4141		
# singleton plateau trees	843			830			751			645			3601		
Rooting method	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Midpoint edge	101	2	1%	147	14	2%	202	17	3%	302	19	3%	726	88	6%
BL-midpoint	62	13	6%	135	35	8%	180	45	6%	270	52	7%	1407	1282	65%
Midpoint plateau	1000	934	58%	1000	856	15%	1000	800	20%	1000	707	17%	4141	3847	46%
Random edge	131	74	10%	162	81	8%	193	75	10%	263	69	6%	1121	807	9%
Random plateau	1000	884	24%	1000	845	9%	1000	779	11%	1000	693	14%	4141	3660	11%

Table 3.1: Summary of rootings of gene trees from simulated and empirical datasets. A - the number of gene trees from a given dataset having rooting inside DL-plateau. B - the number of rootings having the maximal rooting score. C - the percent of gene trees with non-singleton plateau having the maximal rooting score. In case of ambiguity, we assume a match if there is a non-empty intersection between sets of corresponding rootings.

3.2.2. Supertree inference from well supported gene trees

In biological datasets, many of the gene family trees have topology different than their species tree because of the presence of evolutionary events. However, sometimes their incongruence is a result of errors in the gene tree inference process. Such trees may be not reliable and affect the supertree inference. The supertree problem under the duplication-loss cost is defined as follows [Ma *et al.* (2000b); Górecki and Tiuryn (2007a); Page (1997)]:

Problem 2. *Given a collection of unrooted gene trees $\mathcal{U}$, find a species tree S, called* optimal species tree, *that minimizes the total duplication-loss cost $\sum_{G\in\mathcal{U}} \mathrm{uDL}(G,S)$, where* $\mathrm{uDL}(G,S) = \min_{e\in E_G} \mathrm{DL}(G_e,S)$.

The supertree problem for the duplication cost is defined similarly.

In this experiment, our goal is to check whether bootstrapping can improve supertree inference results. Since trees with incorrect topology should have lower support values than correct ones, we propose to filter the gene tree dataset according to their support values. We expect that a supertree inferred from more reliable trees better represent species evolutionary history.

Data preparation. From species present in TreeFam [Ruan *et al.* (2008)] dataset, we selected 14 out of 109 species from different taxa: *Arabidopsis thaliana* (artha), *Saccharomyces cerevisiae* (sacer), *Amphimedon queenslandica* (amque), *Nematostella vectensis* (nevec), *Daphnia pulex* (dapul), *Drosophila melanogaster* (drmel), *Helobdella robusta* (herob), *Lottia gigantea* (logig), *Pelodiscus sinensis* (pesin), *Bos taurus* (botau), *Macaca mulatta* (mamul), *Homo sapiens* (hosap), *Mus musculus* (mumus) and *Danio rerio* (darer). Then we downloaded a collection of 15321 gene family sequences from TreeFam v9.0. Each family was contracted to the set of selected species. We also removed families containing more than 40 genes and families with less than 4 genes or species. We obtained 9443 gene families. Then we aligned them using the T-coffee program and we inferred gene trees by using *PhyML* program with standard parameters setting. Finally, we applied the bootstrap procedure from Section 5.7.

In order to compute bootstrap values, we need a species tree, called here a *model tree*. In experiments we have 5 model trees:

- S^* – the TreeFam tree have based on NCBI taxonomy,
- S^1, S^2 – trees highly similar to S^*,
- S^f – inferred from all 9443 input gene trees by *fasturec* under DL cost,

- S^{fr} – inferred similarly to S^f with the difference that the root is fixed, i.e., S^{fr} is the optimal species tree among trees having the leaf labelled *A. thaliana* (artha) as one of the children of the root.

The model trees and all trees inferred in our experiments are depicted in Figure 3.6.

In our experiment, we introduce two positive real parameters α and β. Let G be a gene tree and S be a model species tree such that $\mathcal{U}$ is the set of bootstrap trees for G. We call a duplication or a speciation cluster A of a gene tree G, weak if $b^{Dup}(A,\mathcal{U}) < \alpha$ or $b^{Spec}(A,\mathcal{U}) < \alpha$, respectively. We say that G is *well supported* if it contains less than β weak clusters. For a given model tree $S \in \{S^*, S^1, S^2, S^f, S^{fr}\}$, $\alpha \in \{0.1, 0.15, \ldots, 0.9\}$ and $\beta \in \{1, 2, \ldots, 6\}$ by $\mathcal{U}(\alpha, \beta, S)$ we denote the set of well supported gene trees induced by α and β in the context of S. In total we generated $17 \cdot 6 \cdot 5 = 510$ sets of well supported gene trees by using bootstrap filtering. Next, for each $\mathcal{U}(\alpha, \beta, S)$ we inferred supertrees for costs D and DL and for fixed and non-fixed root by using *fasturec* program.

For each model tree we depict 8 diagrams depending on the cost and fixed root. Additionally we show aggregated results for optimal and close to optimal trees whose scores differ by less than 100 for D and 400 for DL from the best score.

Discussion. Observe that in our diagrams the left-upper corner denote the most restrictive parameter setting (i.e., low α and high β), while the right-lower corner contains results for almost whole set of gene trees (around 9000). Note, that the datasets for the restrictive parameters represent the most *credible* gene trees.

Results of experiments are depicted in Figures 3.3-3.10. In Figures 3.3-3.5, representing results for the model tree S^*, the most credible datasets induce S^* as the optimal tree. Only for S^*/DL the optimal tree is different but still close to the S^*. S^f and S^{fr} are optimal for a large range of parameters for DL and DL/root, respectively, under all model trees. In DL and DL/root experiments under all model trees the S^r and S^{fr}, respectively, are inferred for the large range of parameters. The results for S^1 and S^2 are very similar to results for S^* (see Figures 3.7-3.8). In Figure 3.9 with the model tree S^f, there is no S^* present in results, however, we have only one case when S^f is the optimal tree for well supported dataset (see S^f/D in Figure 3.9). The results for S^{fr} in Figure 3.10 are mostly compatible with the results for S^f. Topology S^h is frequently present as optimal under all model trees, however, for the restrictive datasets it is optimal only two times for S^f/DL/root and S^{fr}/DL/root.

Under the assumption that S^* is a biologically correct species tree, we observe that model trees highly similar to S^* support S^* for the most credible datasets.

In unrooted reconciliation, the plateau depends only on the top split of the species tree, i.e., the clusters of the children of the root. Therefore, if our method is biased towards the model tree, then, for model trees with different top splits, the results would be significantly different and support independently the corresponding model tree more often. However, the results for the model trees S^f and S^{fr} having different top splits are similar. This suggests that the correct species tree can be inferred from well supported gene trees even if the model tree is inferred in an approximate way.

Runtime. Experiments were performed on a server with 256GB RAM and 8 AMD Opteron processors. The total runtime for calculating bootstrap alignments and trees was about two weeks. Following data processing including tree reconciliation using *URec* [Górecki and Tiuryn (2007b)], computing support values and inferring supertrees took about 5 hours.

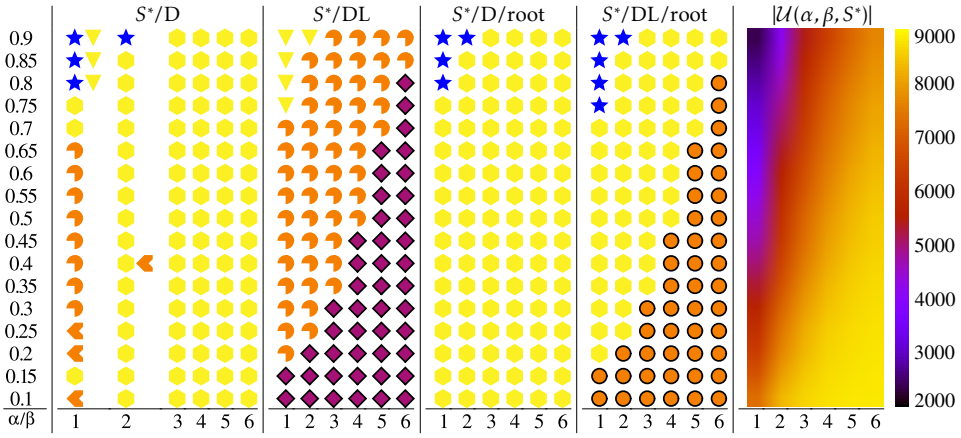


Figure 3.3: Summary of supertree inference experiments for the model S^* with the supertrees having the best score inferred from $\mathcal{U}(\alpha, \beta, S^*)$ by *fasturec* program. Note that in some cases more than one optimal supertree exists. Trees corresponding to used marks are shown in Figure 3.6. The heatmap on the right shows the size of $\mathcal{U}(\alpha, \beta, S^*)$. From the left: S^*/D - supertrees for the D cost, S^*/DL - supertrees for the DL cost, S^*/D/root - supertrees for the D cost with fixed root, S^*/DL/root - supertrees for the DL cost with fixed root.

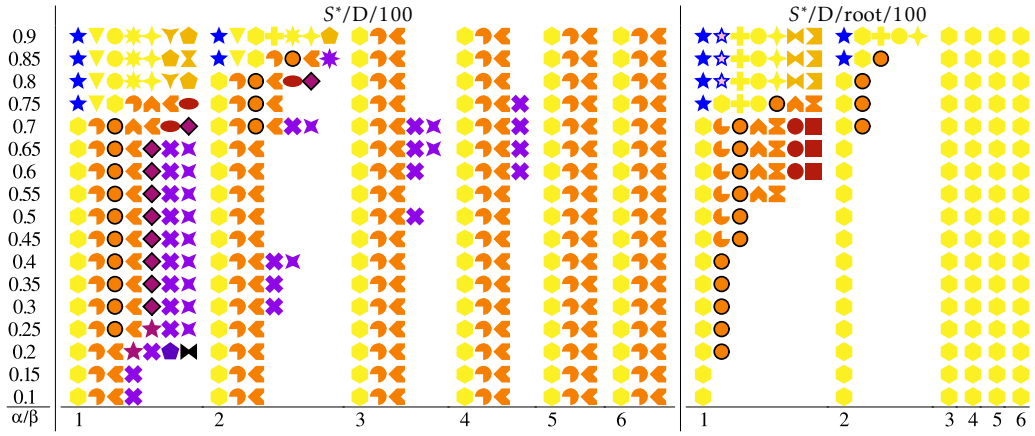


Figure 3.4: Diagrams continued from Figure 3.3. From the left: S^*/DL/400 - supertrees for the DL cost whose scores differ by less than 400 form the best score, S^*/DL/root/400 - supertrees for the DL cost with fixed root whose scores differ by less than 400 form the best score.

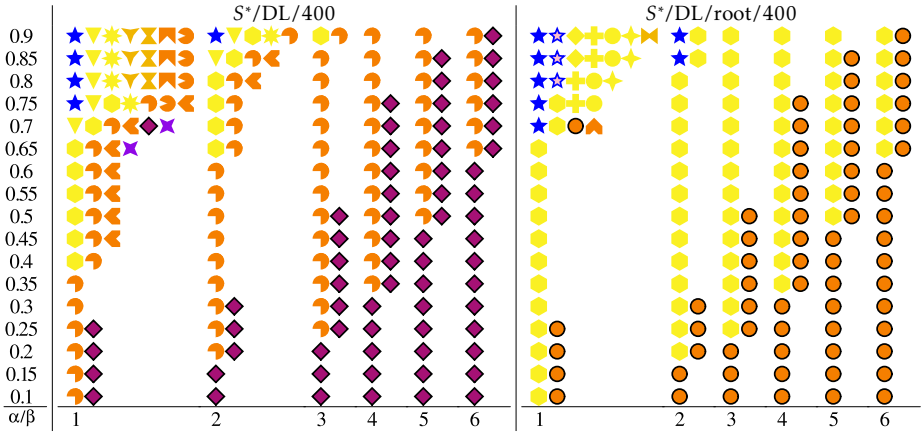


Figure 3.5: Diagrams analogous to those from Figure 3.4 for DL cost and cutoff set to 400.

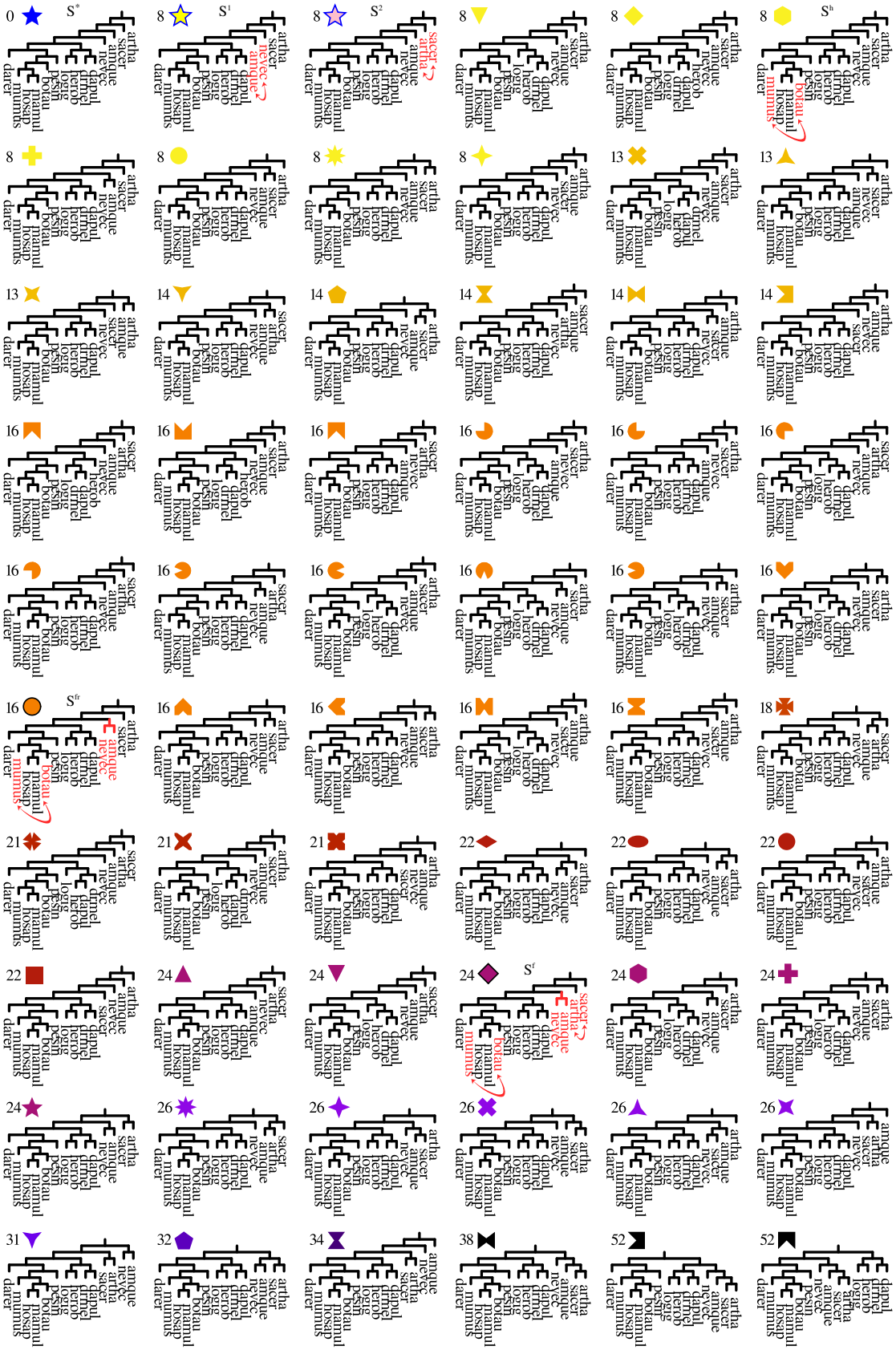


Figure 3.6: Supertrees inferred in four supertree experiments. S^* denotes the species tree from TreeFam database. Trees S^1 and S^2 are among the most similar to S^*. S^{fr} and S^f are supertrees inferred under the DL cost from the whole set of gene trees with fixed and non-fixed root, respectively. S^h is a frequently observed topology in presented experiments. For each species tree S shown here, the number denotes similarity of S to S^* measured as symmetrical DL cost: $\mathrm{DL}(S, S^*) + \mathrm{DL}(S^*, S)$.

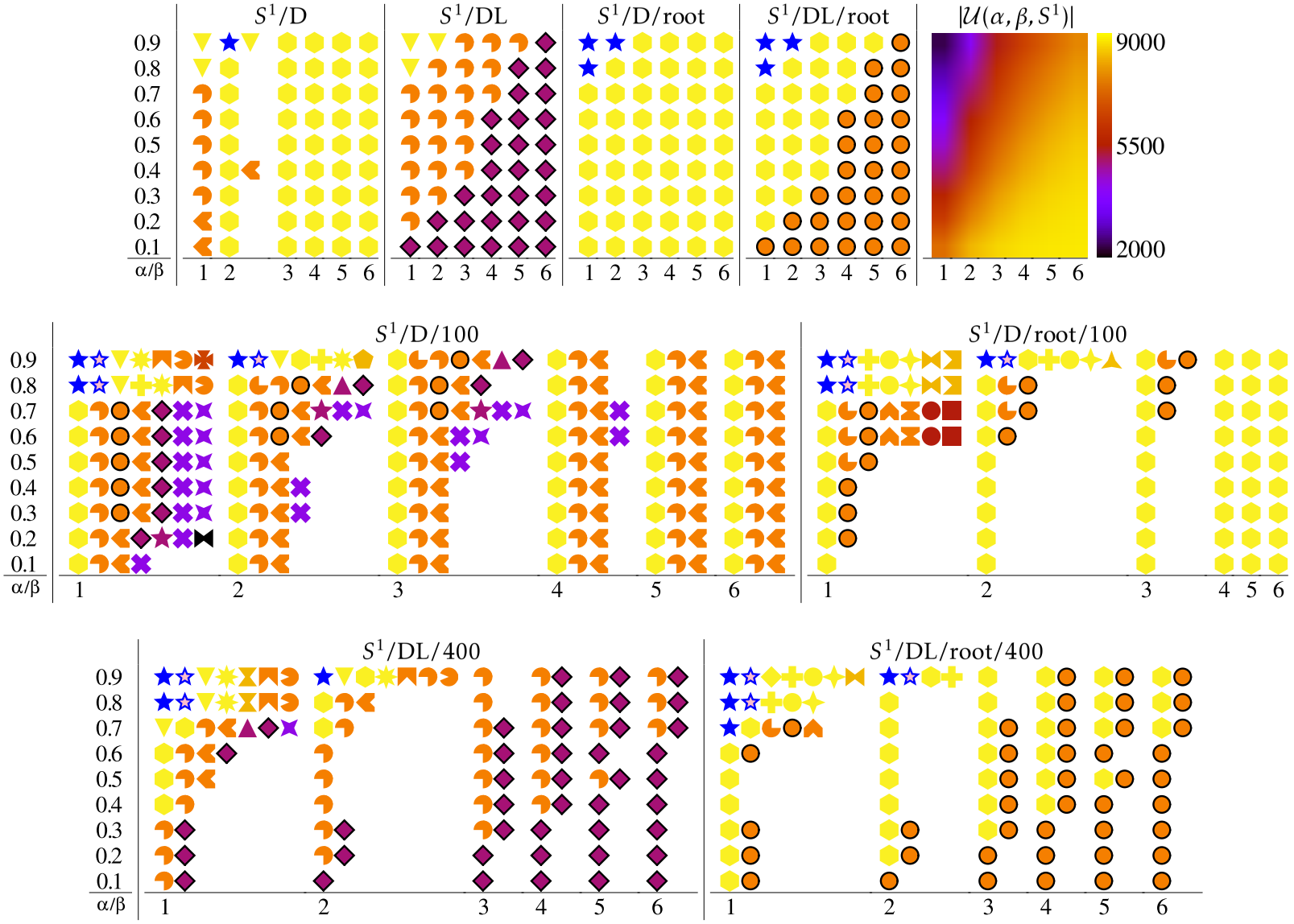


Figure 3.7: Results for the model tree S^1.

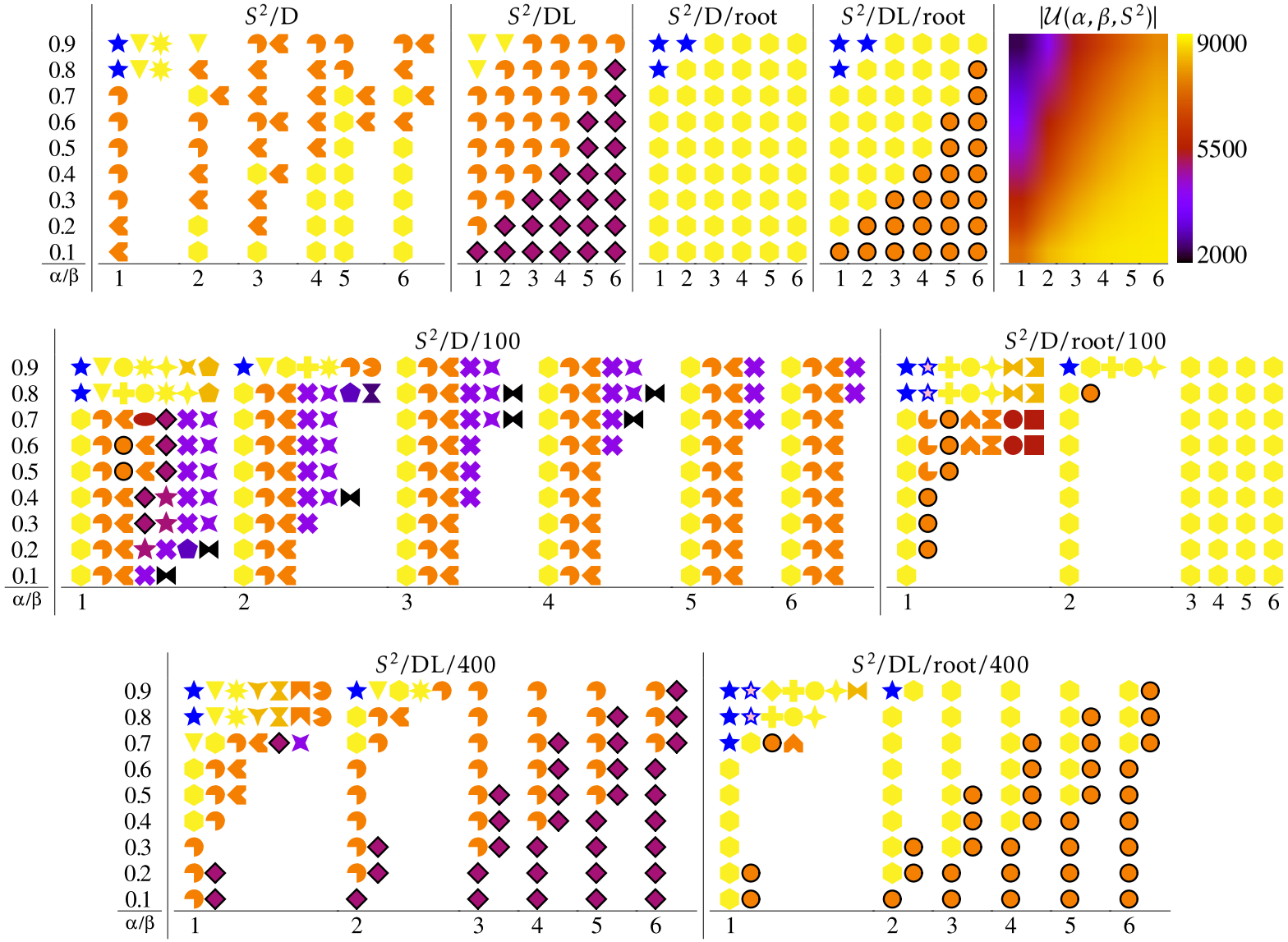


Figure 3.8: Results for the model tree S^2.

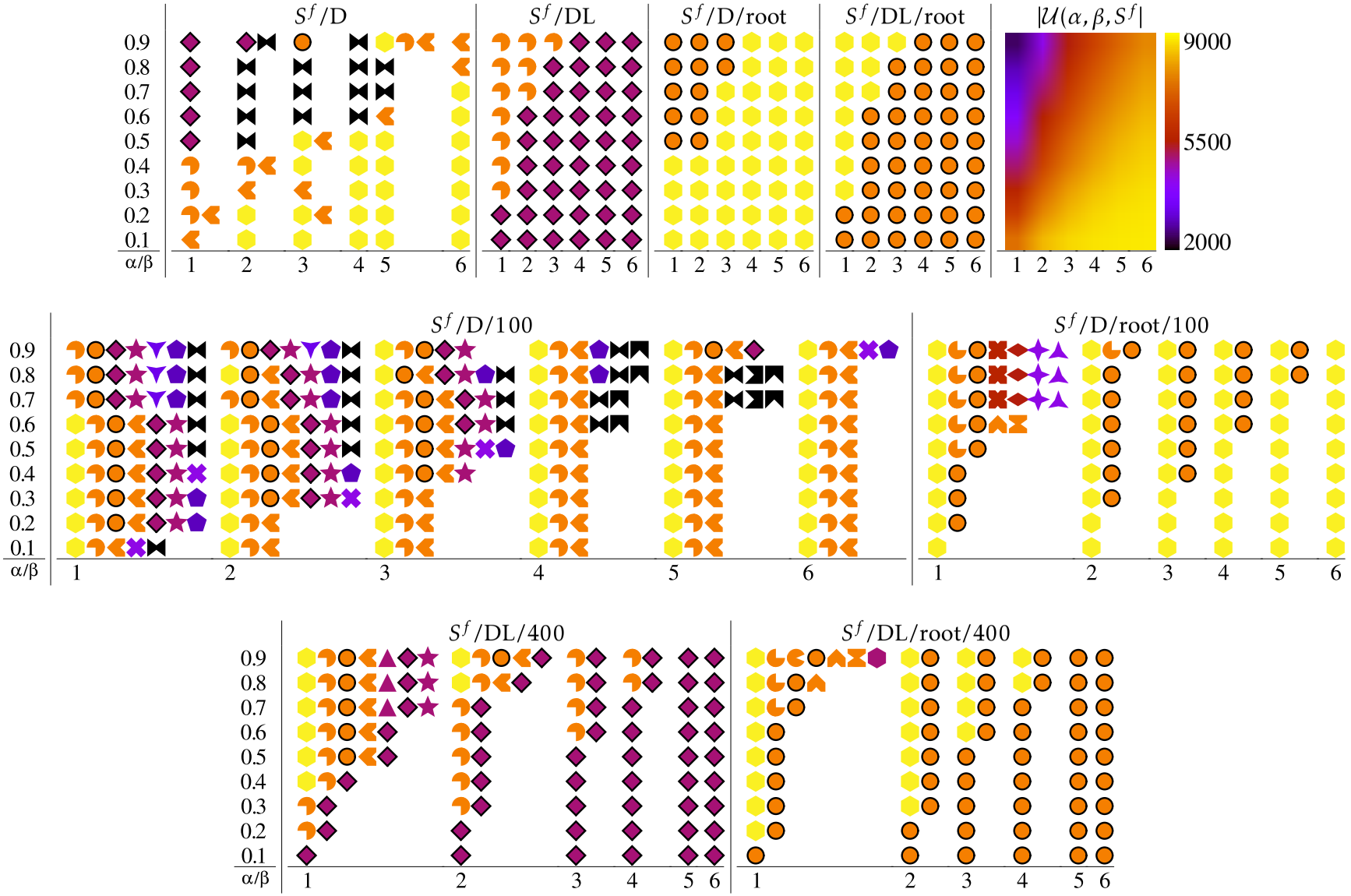


Figure 3.9: Results for the model tree S^f.

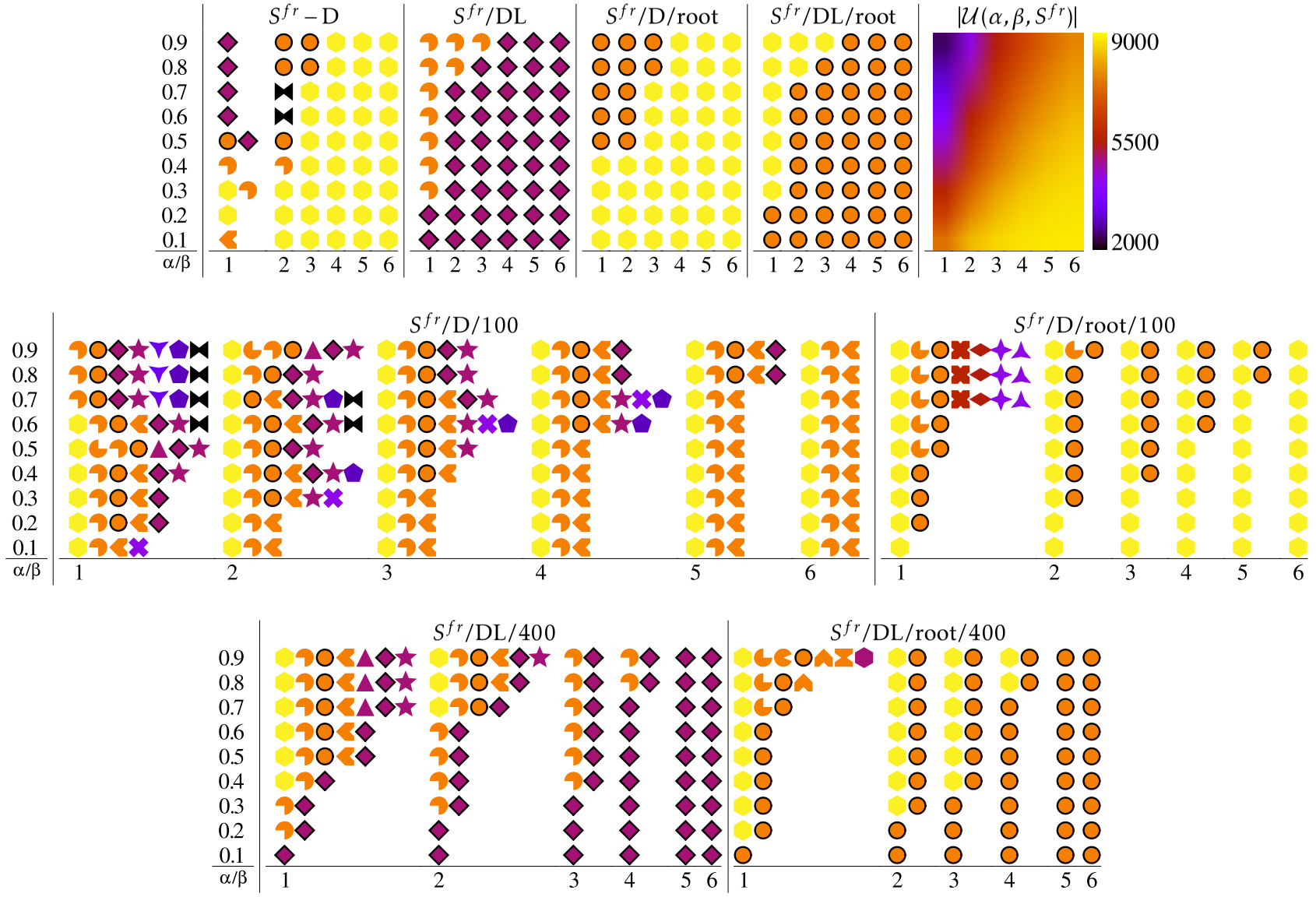


Figure 3.10: Results for the model tree S^{fr}.

3.3. Conclusions and Future Work

In this Chapter, we introduced a novel approach to assess credibility of gene duplication and speciation events in rooted and unrooted gene trees. We proposed a concept of the support values for evolutionary events based on the extended tree reconciliation and non-parametric bootstrap. While this approach can be used to annotate orthology and paralogy in unrooted trees, we also showed how it could be used to verify the reliability of tree reconciliation with applications to the rooting and supertree problem. We provided several theoretical and algorithmic results, in particular, we showed the correspondence between our method and the classical non-parametric bootstrapping. We also showed that species trees inferred from gene trees having highly supported events are more biologically consistent.

In future, we plan to extend this approach to the case when the support is evaluated for subtrees rather than clusters. Such modification would allow to capture more detailed relationships between the gene trees, although the bootstrap values modified this way would be lower than the bootstrap values of the corresponding clusters.

4

Inference of Credible and Time-consistent Horizontal Gene Transfer Events

Almost all places on Earth are inhabited by microorganisms, which are an important element of the Earth's ecosystem. Reactions carried out by bacteria affect the chemical composition and pH of the environment, while the products of these reactions are often used by other organisms. In animal and human organisms, bacteria support many metabolic processes such, as digestion, vitamin synthesis and degradation of toxins. In addition, they strengthen immunity, and perform many other functions necessary for life. Apart from environmental, bacteria are also of great industrial importance. They are used to produce of antibiotics, fertilizers, food that require fermentation, and to remove pollutants from industrial wastewater. Despite the widespread prevalence in nature and their importance to the environment our knowledge of bacteria is limited. So far, research has been limited to bacteria that can be grown in laboratories. That was until the emergence of DNA sequencing methods, which enabled new approaches to biological experiments. With these modern methods, the increasing number of genes and whole genomes can be characterized.

The evolution of a single gene family or a group of species can usually be represented in the form of a phylogenetic tree. To address microbiological studies, the model had to be extended by horizontal gene transfer (HGT). However, models extended by HGT are more complex and computationally demanding [Nguyen *et al.* (2013); Scornavacca *et al.* (2014); Sjöstrand *et al.* (2014); Szöllősi *et al.* (2013a); Tofigh *et al.* (2011)].

Several strategies have been proposed to overcome these limitations. For example, allowing transfers to form cycles results in reducing the time complexity of the problem to a polynomial. It can be done by dynamic programming [Bansal *et al.* (2012); Mykowiecka *et al.* (2017); Tofigh *et al.* (2011)], though, there is no guarantee that the inferred scenario is biologically valid. Another approach is to assign a divergence time to some or all of the nodes of the species tree, which creates an additional requirement of temporal ordering in the scenario [Ranwez *et al.* (2015)]. With such constraint, the problem has a polynomial time solution [Bansal *et al.* (2012); Doyon *et al.* (2010)]. However, such dated species trees are rarely available, especially for bacterial species where horizontal gene transfers are frequent. A different way to address the problem is to insert candidate transfer edges directly to the species tree. As a result, we obtain a directed graph, called a *species graph*, representing the evolution of species with a set of horizontal edges that can be used by gene lineages as horizontal transfers. If the candidate transfers do not form cycles, such a model has a polynomial time complexity [Górecki (2004a); Scornavacca *et al.* (2017)].

In this Chapter we present a new efficient iterative method for the inference of well-supported and time-consistent horizontal gene transfer events. We introduce the concept of transfer support values and evolutionary scenarios with HGT events along with the problem of finding scenarios with the minimal cost. Having this, we propose a general algorithmic framework for iterative insertion of horizontal transfers based on the reconciliation cost gene duplication, losses, horizontal transfer events and transfer support values. Our method starts from an initial species tree, i.e., a species graph without transfers, and at each step ensures that the inferred transfer scenarios are acyclic and well-supported by transfer support values. Last sections are devoted to the description of three experiments conducted with our algorithm. On two empirical examples from the literature [Druzhinina *et al.* (2018); Eme *et al.* (2017)], we show that our method can be used to support known transfer hypotheses between distantly and more closely related species. In the last experiment using simulated data sets, we demonstrate high accuracy of our method by presenting a high percentage of correctly inferred transfer scenarios reached by the algorithm.

4.1. Structure definitions

In the following section, we introduce the key notions of a species graph, and an extended species tree. Please refer to Section 2.1 on basic definitions of species and related terms. In this chapter, we assume that gene trees are rooted and each non-leaf node of a species tree has out-degree two (for modeling classical vertical evolution) or one (nodes of HGTs).

4.1.1. Species graph definition

A species graph is a structure that models a species tree with additional horizontal edges that form a time-consistant (acyclic) graph. Now, we recall several definitions from [Górecki (2004a); Górecki (2010); Górecki and Tiuryn (2012)]. A *species graph* S is an ordered triple $S = \langle V, E, H \rangle$, such that $B = \langle V, E \rangle$ is a species tree and $H \subseteq V \times V$ is a set of all transfer edges present in S which satisfies the following conditions:

- for every $\gamma \in H$, nodes of γ are not on a path in B,
- for every $\gamma \in H$, both nodes of γ have out-degree 1 in B,
- no two edges in H have a node in common,
- every node of V with out-degree 1 is contained in an edge from H,
- and the relation $\{\langle \gamma_1, \gamma_2 \rangle : \gamma_1, \gamma_2 \in H$ and there exists a path in B from a node of γ_1 to a node of $\gamma_2\}$ is a partial order on H.

It follows from the last condition that every species graph is a directed acyclic graph. Every node in S of out-degree one will be called a *transfer node*. By $\overrightarrow{H}$ we denote the set of transfer start nodes, i.e., $\overrightarrow{H} = \{v : \langle v, w \rangle \in H\}$.

The node in B, whose out-degree is two, we call a *speciation*. The set of all speciation nodes in S we denote by Σ.

4.1.2. Extended species tree

Below, we introduce a concept of an extended species tree. Please recall, that by $\mathcal{L}_T$ we denote the set of all leaves in a tree T, by L_T the set of all leaf labels in T and by $\widehat{v}$ the set of all children of v The labelling of a species tree is the identity function. For a species graph S, the *extended species tree*, denoted S', is the tree obtained from S by a sequence of unfolding HGT operations defined as follows: For the *lowest transfer* $\langle v, w \rangle$ replace $\langle v, w \rangle$ with the edge, called *transfer edge*, connecting v with a new copy of the subtree whose root is the only child of w, and contract w, i.e., replace edges

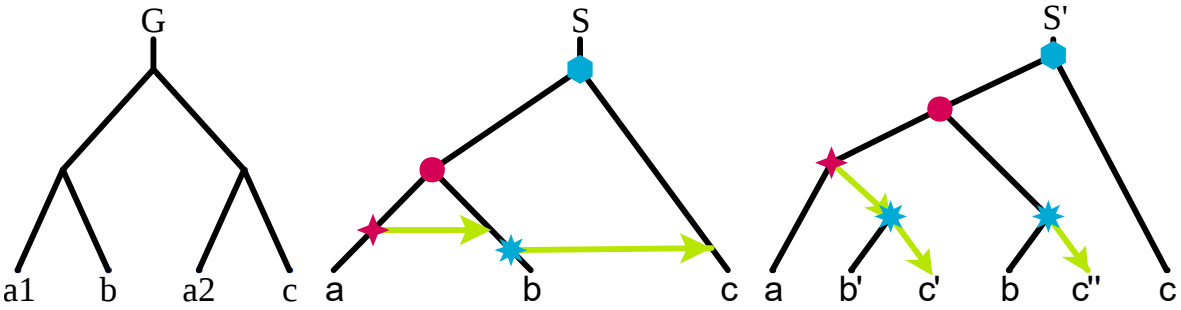


Figure 4.1: An example of a gene tree G, a species graph S and the extended species tree S' with four gene sequences from three species a, b and c. Note that genes a1 and a2 sampled from the same species a. The decoration in S and S' indicates the mapping $\mathsf{m}\colon S' \to S$ (omitted for the leaves). G has 3 non-singleton clusters: {a1,b}, {a2,c} and {a1,a2,b,c}.

$\langle \mathsf{par}(w), w\rangle$ and $\langle w, c\rangle$ with $\langle \mathsf{par}(w), c\rangle$, where c is the child of w. In this construction, every node $s' \in V_{S'}$ uniquely corresponds to its source node in S. Such a mapping we denote by m. An example is depicted in Figure 4.1.

The correctness and uniqueness of the definition follow from the fact that transfers do not form cycles. The labeling of leaves in S' is inherited from S. Similarly to H and $\overrightarrow{H}$ in S, we define H' in S' as the set of all transfer edges in S' and $\overrightarrow{H'}$ as the set of transfer start nodes in S'. We say that v is a speciation in S' if $\mathsf{m}(v)$ is a speciation in S.

4.2. HGT-Scenario

Here we introduce the concept of evolutionary scenarios with horizontal gene transfer events[1]. Given a gene tree G and a species graph S, a *HGT-scenario*[2] $\xi\colon \mathcal{L}_G \to \mathcal{L}_{S'}$ is a function that preserves the leaf labeling, i.e., for every leaf $g \in G$, species (labels) of g and $\xi(g)$ are equal. For a non-leaf node g, by g' and g'' we denote the children of g. By M_ξ we denote the lca-mapping between G and S' that extends ξ such that $\mathsf{M}_\xi(g) = \mathsf{M}_\xi(g') \oplus \mathsf{M}_\xi(g'')$. See an example in Figure 4.2. A cluster of a node v in a gene tree is the set of leaves (gene sequences) reachable from v. We say that a transfer $h \in S$ *transfers the cluster of v in a HGT-scenario* ξ if the shortest path containing nodes $\mathsf{m}(p_1), \ldots, \mathsf{m}(p_k)$ in S contains h, where p is the path connecting $\mathsf{M}_\xi(v)$ and $\mathsf{M}_\xi(\mathsf{par}(v))$ in S'.

[1] In Chapter 5, we introduce the notion of DTL-scenarios, which also use transfer events. For a discussion on the differences between the two models, please refer to Section 5.1.

[2] In our article [Mykowiecka *et al.* (2018)], where we introduced HGT-scenarios, we simply used the notion *scenario*, but here we have changed the name due to a notation conflict with Chapter 5 and for better readability.

4.3. DTL cost

Below we define duplication, loss and horizontal gene transfer costs along with the formulas for their calculation. Let g be an internal node of G. It is said that g is a *duplication* for a HGT-scenario ξ, if $\mathsf{M}_\xi(g) = \mathsf{M}_\xi(g')$ or $\mathsf{M}_\xi(g) = \mathsf{M}_\xi(g'')$. We say, that the total number of duplications $\mathsf{dup}_\xi(G,S)$ in G defines the *duplication cost* for a scenario ξ.

The number of horizontal gene transfers for the HGT-scenario ξ is given by

$$\mathsf{hgt}_\xi(G,S) = \sum_{g \in G} \mathsf{hgt}'(\mathsf{M}_\xi(g), \mathsf{M}_\xi(\mathsf{par}(g))),$$

where $\mathsf{hgt}'(v,w) =\mid \{(x,y) \in H_{S'} : v \succeq x \succ y \succeq w\} \mid$ is the number of transfer edges on the path connecting v and w in S'. Let $\mathsf{spec}'(v,w)$ be the number of speciation nodes on the path form v to w in S' excluding w, i.e., $\mathsf{spec}'(v,w) =\mid \{x : v \succeq x \succ w$ and x is a speciation $\} \mid$ and let $\mathsf{loss}'(v,w) = \mathsf{hgt}'(v,w) + \mathsf{spec}'(v,w)$.

Finally, the number of gene losses for a HGT-scenario ξ is

$$\mathsf{loss}_\xi(G,S) = \sum_{g \in G} \mathsf{loss}_{\xi,g},$$

where

(L1) $\mathsf{loss}_{\xi,g} = \mathsf{loss}'(\mathsf{M}_\xi(g'), \mathsf{M}_\xi(g))$ if g is an internal node and $\mathsf{M}_\xi(g') \prec \mathsf{M}_\xi(g) = \mathsf{M}_\xi(g'')$,

(L2) $\mathsf{loss}_{\xi,g} = \mathsf{loss}'(\mathsf{M}_\xi(g'), v) + \mathsf{loss}'(\mathsf{M}_\xi(g''), w)$, where g is an internal node of G, v and w are the children of $\mathsf{M}_\xi(g)$ such that $\mathsf{M}_\xi(g') \preceq v \prec \mathsf{M}_\xi(g) \succ w \succeq \mathsf{M}_\xi(g'')$,

(L3) and $\mathsf{loss}_{\xi,g} = 0$, otherwise.

Let $c(G,S',\xi)$ be a reconciliation cost for the HGT-scenario ξ. The DTL cost is a weighted sum of the number of evolutionary events and it is equal

$$c(G,S,\xi) = \mathbf{DUP} \cdot \mathsf{dup}_\xi(G,S) + \mathbf{LOSS} \cdot \mathsf{loss}_\xi(G,S) + \mathbf{HGT} \cdot \mathsf{hgt}_\xi(G,S),$$

where **DUP**, **LOSS** and **HGT** are non-negative event weights for duplication, loss and horizontal gene transfer events, respectively.

Problem 3. *Given a gene tree G and a species graph S'. Find the minimal cost $c(G,S',\xi)$ in the set of all HGT-scenarios ξ between G and S.*

The minimal cost we denote by $\mathsf{c}_{\mathsf{HGT}}(G,S)$ and the set of all HGT-scenarios that yield the minimal cost we denote by $\Xi_{\mathsf{HGT}}(G,S)$. See an example in Figure 4.3.

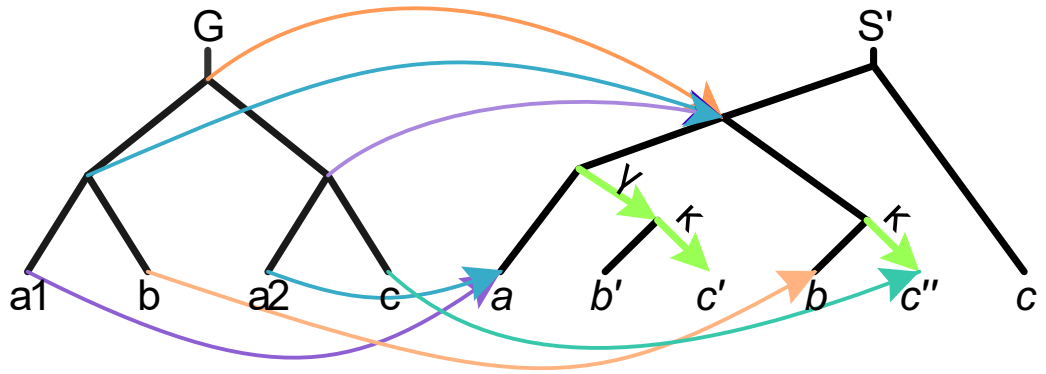


Figure 4.2: An example mapping $M_\xi \colon V_G \to V_S'$ for the trees from Figure 4.1 and for the HGT-scenario ξ in which $\xi(\text{a1}) = \xi(\text{a2})$, $\xi(\text{b}) = b$ and $\xi(\text{c}) = c'$. The HGT-scenario induces one loss, one duplication, and one transfer. The HGT-scenario corresponds to the embedding D1L1T1 from Figure 4.3. Note that the transfer κ in this HGT-scenario transfers the cluster {c}.

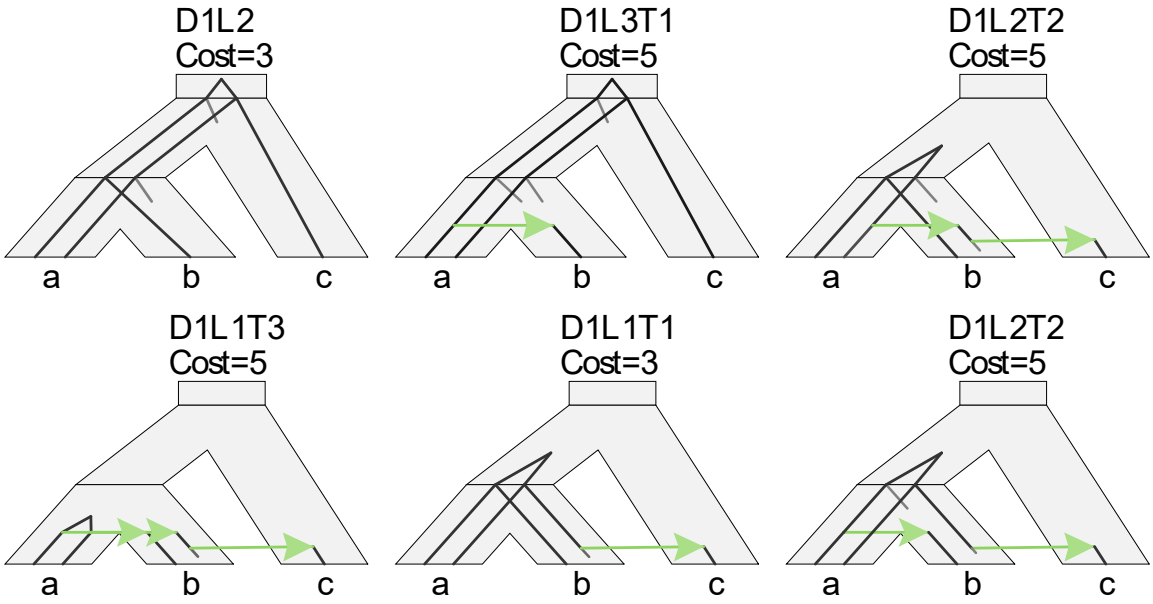


Figure 4.3: All HGT-scenarios for G and S from Figure 4.1 visualized in the form of embeddings [Górecki and Tiuryn (2006)]. DLT costs are computed for the following weights **DUP** = **HGT** = **LOSS** = 1. Here, the minimal cost is 3 and it is reached by two HGT-scenarios (see also the HGT-scenario ξ from Figure 4.2).

4.4. Minimal cost and HGT-scenario inference

Below, we show our solution of the problem of finding the minimal cost in the set of all HGT-scenarios ξ between G and S and proofs of deduced mathematical formulas.

Problem 3 can be solved by a dynamic programming algorithm in polynomial time [Górecki (2004a); Scornavacca *et al.* (2017)]. In this section, we describe a revised $O(|G||S|)$ time solution based on [Scornavacca *et al.* (2017)] that we used in our approach to infer optimal acyclic HGT-scenarios.

In the formalization below, we simplify the edge relation in S, by identifying each HGT-transfer termination node with its first non transfer end descendant. Formally if $s \in V_S$ then $s^* := c^*$ if s is a transfer end node with the child c, and $s^* := s$ otherwise. Note that for every $s \in V_S$, s^* is a leaf or s^* has two children.

By $G|g$ we denote the subtree of G rooted at g. The dynamic programming formula has two components δ and $\delta^{\uparrow}$ that denote the minimal cost of HGT-scenarios for $G|g$ and S where for $g \in V_G$ and $s \in V_S$ we have additional conditions:

H1 $\delta(g,s)$ is the minimal cost in the set of all HGT-scenarios ξ for $G|g$ and S' such that $\mathsf{M}_\xi(g) \in \mathsf{m}^{-1}(s)$.

H2 $\delta^{\uparrow}(g,s)$ is the minimal cost in the set of all HGT-scenarios ξ for $G|g$ and S such that there exists a node $v \in \mathsf{m}^{-1}(s^*) \in S'$ and $\mathsf{M}_\xi(g) \preceq v$ including the cost of the path π connecting v and $\mathsf{M}_\xi(g)$. Formally the cost of the path is calculated as:

1. the cost of the HGT-scenario ξ,
2. plus the cost of HGT events counted as the weighted number of HGT edges on the path π connecting v with $\mathsf{M}_\xi(g)$, i.e., $\mathsf{hgt}'(\mathsf{M}_\xi(g), v) \cdot \mathbf{HGT}$,
3. plus the cost of gene losses on π counted as the number of HGT events on the path π and the number of speciation nodes on π excluding the last node, i.e., $\mathsf{loss}'(\mathsf{M}_\xi(g), v) \cdot \mathbf{LOSS}$.

For δ we have the following formulas:

$$\delta(g,s) = \begin{cases} 0 & \text{if } g \text{ and } s \text{ are leaves and } g \text{ is labelled by } s, \quad (1) \\ \min\{\alpha, \beta\} & \text{if } s \text{ and } g \text{ are not leaves,} \quad (2) \\ \beta & \text{if } s \text{ is a leaf and } g \text{ is not a leaf,} \quad (3) \\ +\infty & \text{otherwise,} \quad (4) \end{cases}$$

where, s' and s'' denote the children of s, and

$$\alpha = \mathbf{HGT} \cdot \mathbb{1}[s \in \overrightarrow{H}] + \min_{c \in \widehat{g}} \delta^{\uparrow}(c, s') + \delta^{\uparrow}(\mathsf{sib}(c), s'')$$

$$\beta = \mathbf{DUP} + \min_{c \in \widehat{g}} (\delta(c, s) + \delta^{\uparrow}(\mathsf{sib}(c), s))$$

Here, α represents either speciation or transfer event, while β is a duplication. The correctness of the above formula follows from the fact that optimal solutions are generated by some lca-scenarios [Górecki (2010); Górecki and Tiuryn (2012)].

The formula for $\delta^{\uparrow}$ can be expressed as:

$$\delta^{\uparrow}(g,s) = \begin{cases} \delta^{\uparrow}(g, s^*) & \text{if } s \neq s^*, \quad (5) \\ \delta(g,s) & \text{if } s \text{ is a leaf}, \quad (6) \\ \min\Big(\delta(g,s), \min_{x \in \widehat{s}}\{\delta^{\uparrow}(g,x) + \\ \quad + \mathbf{HGT} \cdot \mathbb{1}[\langle s, x\rangle \in H] + \\ \quad + \mathbf{LOSS} \cdot \mathbb{1}[s \in \Sigma \ \vee \ \langle s, x\rangle \in H]\}\Big) & \text{otherwise.} \quad (7) \end{cases}$$

Note that if $s \neq s^*$ then $\delta(g,s) = +\infty$ and $\delta^{\uparrow}(g,s) = \delta^{\uparrow}(g,s^*)$.

Theorem 6 (Solution to Problem 1)**.** *For a binary gene tree G and a species tree S we have*

$$\mathsf{c}_{\mathsf{HGT}}(G,S) = \min_{s \in S} \delta(\mathsf{root}(G), s).$$

Proof. The proof is by induction on the structure of G and S. In the induction step, it is sufficient to show that the properties H1-H2 hold for δ and $\delta^{\uparrow}$.

Base step of induction:

If s is a leaf labelled x, then $\mathsf{m}^{-1}(s)$ contains all leaves labelled x in S'. Thus, the set of HGT-scenarios from condition H1 is not empty if all leaves in $G|g$ are labelled by x. In such a case the cost equals $(|G|g| - 1) \cdot \mathbf{DUP}$ which is jointly modelled by cases (1) and (3) with β. If the set of HGT-scenarios is empty then the cost is $+\infty$ (see case (4)) and β if $|G|g| > 1$. This completes the base step for H1. Note that if s is a leaf H2 becomes H1 which is expressed in (6).

Main step of induction:

Assume that s is an internal node and g is a gene tree node. Assume that H1 and H2 hold for every pair of nodes $(v,w) \neq (g,s)$ such that there are directed paths from s to w in S and from q to v in G.

We show that H1 and H2 hold for g and s. We may assume that g is internal, otherwise there is no HGT-scenario where a leaf is mapped to an internal node (4).

We start with the proof for H1 in which there are three cases for HGT-scenarios:

H1-I (Duplication)

If g is a duplication in a HGT-scenario ξ then for one child of g, say c, we have $\mathsf{M}_\xi(c) = v$ and $\mathsf{M}_\xi(\mathsf{sib}(c)) \preceq v$.

Based on the inductive assumption, it follows that the cost of such a HGT-scenario is modelled by β.

Now we show correctness of the cost calculation in $\delta(g,s)$. We will show that $\delta(g,s)$ is the cost of a HGT-scenario, in this case, by showing that all events are correctly identified. Here we have one duplication event at s and the remaining duplications are present in the subtrees of g' and g'', where g' and g'' are the children of g. Clearly, all such duplication events are identified by the inductive assumption. For gene losses, we consider two cases: $\mathsf{M}_\xi(g'') = v$ and $\mathsf{M}_\xi(g'') \prec v$. If $\mathsf{M}_\xi(g'') = v$ then $\delta^\uparrow(g'',s) = \delta(g'',s)$ and no losses are present on the path π defined in H2 for $\delta^\uparrow(g'',s)$. Thus, by the inductive assumption, we obtained (L3) for $\mathsf{M}_\xi(g') = \mathsf{M}_\xi(g'') = v$.

In the second case, there are $\mathsf{loss}'(\mathsf{M}_\xi(g'), v)$ losses by the inductive assumption for H2. Since $v = \mathsf{M}_\xi(g)$, the number of losses is $\mathsf{loss}'(\mathsf{M}_\xi(g'), \mathsf{M}_\xi(g))$ which corresponds to (L1). Similarly, we obtain the number of HGT events. We omit the details.

H1-II (Speciation)

Assume that g is not a duplication and s is not a transfer start, i.e., s is a speciation. Then both children of g maps to or below distinct children of s. Thus the formula for the cost is

$$\min\{\delta^\uparrow(g',s') + \delta^\uparrow(g'',s''), \delta^\uparrow(g',s'') + \delta^\uparrow(g'',s')\},$$

where s' and s'' are the children of s. See (α), where there is no HGT contribution.

The correctness of the cost computation follows similarly to the previous case. Here, we show only the case of gene losses. Other cases are similar.

Now, we have that $\mathsf{M}_\xi(g') \prec v$ and $\mathsf{M}_\xi(g'') \prec v$. This, the number of losses for g' equals $\mathsf{loss}'(\mathsf{M}_\xi(g'), v)$, where $v \in \mathsf{m}^{-1}(s^{*\prime})$ by the inductive assumption for H2 plus $\mathsf{loss}'(\mathsf{M}_\xi(g''), w)$, where $w \in \mathsf{m}^{-1}(s^{*\prime\prime})$ by the inductive assumption for H2. Since $v \succeq \mathsf{M}_\xi(g')$ and $w \succeq \mathsf{M}_\xi(g'')$, we conclude that v and w are the children of $\mathsf{M}_\xi(g)$ in S'. Thus, we obtained the cost of losses matching the case (L3).

H1-III (HGT)

The case is similar to the previous one. The only difference is the HGT cost.

This completes the proof of H1 for g and s.

Now, we prove H2. Under the notation from condition H2, we have two cases:

A If $\mathsf{M}_\xi(g') = v$ for some HGT-scenario satisfying H2 then its cost is $\delta(g,s)$, since no additional cost of **HGT** and **LOSS** is needed.

B If $\mathsf{M}_\xi(g) \prec v$, then we have several cases. Let $s \in \Sigma$.

B1 one gene loss is needed, since the path from $\mathsf{M}_\xi(g)$ to v visits a node from $\mathsf{m}^{-1}(s')$ or $\mathsf{m}^{-1}(s'')$. Thus, the minimal cost is achieved by $\min(\delta(g,s), \mathbf{LOSS} + \delta^{\uparrow}(g,s'))$.

B2 If s is a transfer start node, then the path connecting $\mathsf{M}_\xi(g)$ with v goes either by the transfer edge $(s,x) \in H$ which start is s ($v \in S'$) or its sibling non-transfer edge. In the first case we have one **HGT** and one **LOSS**, while in the second one no events are present. This leads to the following formula for the minimal cost of a HGT-scenario in this case: $\min(\delta(g,s), \mathbf{HGT} + \mathbf{LOSS} + \delta^{\uparrow}(g,x), \delta^{\uparrow}(g,\mathsf{sib}(x)))$.

All the above cases are incorporated in (6).

This completes the proof for property H2.

Now to complete the proof of H2, we need to show that the number of losses on the path π is $\mathsf{loss}'(\mathsf{M}_\xi(g), v)$ and the number of HGT events is $\mathsf{hgt}'(\mathsf{M}_\xi(g), v)$ in a HGT-scenario ξ satisfy condition for H2. Computation of the cost follows by a sequence of $\delta^{\uparrow}$ calls, say $\delta^{\uparrow}(g,s_0), \delta^{\uparrow}(g,s_1), \cdots, \delta^{\uparrow}(g,s_{k-1})$ in (5) and (6) that terminates with the call of $\delta(g,s_k)$ either in (6) or (5), where $k \geq 0$ and $s_0 = s$. Now, the path $\pi = \pi_0, \cdots, \pi_m$ in S' with $\pi_0 = v$ and $\pi_m = \mathsf{M}_\xi(s)$ corresponds to the path $p = s_{j_1}, s_{j_2}, \cdots, s_{j_m}$ in S where $0 \leq j_1 < \cdots < j_m = k$ such that p is obtained from $s_0, \cdots, s_k$ by removing all transfer end nodes, and, for each i, $\pi_i \in \mathsf{m}^{-1}(s_{ji})$. Note that there could be more paths π that satisfy the property, however, all subtrees of S' rooted at nodes from $m^{-1}(s)$ are isomorphic (see the unfolding operation). Therefore, the obtained cost calculation is independent of v. In other words we showed that the computation is composed of:

1. the cost of HGT-scenario ξ for $G|g$ and S where g is mapped to $\pi_m \in \mathsf{m}^{-1}(s_k)$, i.e., $\delta(g,s_k)$ from the inductive assumption,

2. the cost of path π calculated in the sequence of $\delta^{\uparrow}$ calls from which we have to show that the number of loss and HGT events equal $\mathsf{loss}'(\pi_0, \pi_k) = \mathsf{loss}'(v, \mathsf{M}_\xi(g))$ and $\mathsf{hgt}'(v, \mathsf{M}_\xi(g))$, respectively.

The case of transfer follows directly from the fact that we count HGT events only when the whole transfer edge is present in π (see (6)). Similarly, loss events are counted when π contains a transfer edge, i.e., $\mathsf{hgt}'(\mathsf{M}_\xi(g), v)$ times, and when a specification is present in π, i.e., $\mathsf{spec}'(\mathsf{M}_\xi(g), v)$ times. This completes the proof ■

This algorithm can be naturally extended to infer *optimal HGT-scenarios*, i.e., the HGT-scenarios with the minimal cost, by using standard backtracking method.

Theorem 7 (Time Complexity)**.** *For a binary gene tree G and a species tree S the minimal cost can be computed in $O(|G||S|)$ time and space.*

Proof. It follows immediately from the fact, that each δ and $\delta^{\uparrow}$ computing requires a constant number of steps and two arrays of size $|G| \cdot |S|$ are required to store values of δ and $\delta^{\uparrow}$. ■

4.5. Transfer support values

In the following Section we introduce the notion of *transfer support* by merging the concepts of non-parametric bootstrap described in Section 2.3 and optimal evolutionary scenarios. Our approach is based on the bootstrap method, and therefore the preparation phase is required to infer sample trees. Starting from the multiple alignment of gene sequences, we create a set of *bootstrap alignments*. Each bootstrap alignments is created based on random sampling by drawing with replacement columns of the original alignment. Then, from each bootstrap alignment, we infer a sample tree by some standard tree-building tool, e.g., PhyML [Guindon *et al.* (2009)]. Finally, each sample tree has to be rooted by using out-group, median or other rooting methods, e.g., Urec [Górecki and Tiuryn (2007b)].

Before defining our transfer support value, we need one more notion. We say that a transfer $h \in H$ is *used* by a HGT-scenario $\xi : \mathcal{L}_G \to \mathcal{L}_{S'}$ (when reconciling G and S), if there is at least one cluster transferred by h in ξ. For instance, the HGT-scenario from Figure 4.2 uses only transfer δ. Now, we can define the notion of *transfer support value*.

Definition 2 (Transfer Support)**.** *Given a species graph S with a transfer $h \in H$ and the set of sample trees $\mathcal{U}$ obtained from the same input alignment. The support of a transfer*

h is defined as follows:

$$\mathsf{support}(h, S, \mathcal{U}) = \frac{1}{|\mathcal{U}|} \sum_{G \in \mathcal{U}} \frac{|\{\xi \in \Xi_{\mathsf{HGT}}(G, S) \colon h \textit{ is used by } \xi\}|}{|\Xi_{\mathsf{HGT}}(G, S)|}.$$

Note that transfer support is a value representing the ratio of transfer usage in the set of sample trees. For example, it is zero if no optimal HGT-scenario (when reconciling sample trees with the species graph) uses a given transfer h and it is one if all optimal HGT-scenarios use h.

4.6. Algorithm

In this Section, we present the details of our algorithmic solution to infer low scoring species graphs with well-supported transfers.

Algorithm 2 is an iterative method, which in each step of the main loop inserts a new transfer into candidate species graphs. First, we formalize this operation. Given a species graph S, and a pair of non-transfer edges $\langle e, e' \rangle$ from S, we can create a new graph, denoted $S_{e,e'}$ by inserting new nodes v and v' in the middle of e and e', respectively, and by inserting a new transfer edge from v to v'. Note that if such a graph is acyclic, it is a species graph. A pair of edges $\langle e, e' \rangle$, such that $S_{e,e'}$ is a species graph, will be called *valid* for S.

In Figure 4.4, we show an artificial example of Algorithm 2 execution.

From a computational point of view, Algorithm 2 is a heuristic whose complexity depends on the applied stopping conditions (given here in a general way), parameters (e.g., how many iterations), the input trees, and the number of optimal HGT-scenarios inferred by the DP algorithm in each step of the main loop. In practice, our tests indicated that Algorithm 2 performs well on empirical data (see the next section). Also, we observed that in the majority of cases, the set of optimal HGT-scenarios was small and usually consisted of one HGT-scenario.

4.7. Experiments

We conducted three experiments on biological and simulated data sets. In the first experiment, we re-analyzed data from inter-kingdom transfers to *Pezizomycotina* fungi [Druzhinina *et al.* (2018)]. In the second analysis, we re-analyzed results from a recently published *Blastocystis* spp. genome annotation which revealed the presence of diverse laterally transferred genes from distant taxa [Eme *et al.* (2017)]. Simulated gene and species trees were used in the third experiment in which we examined the accuracy of the algorithm.

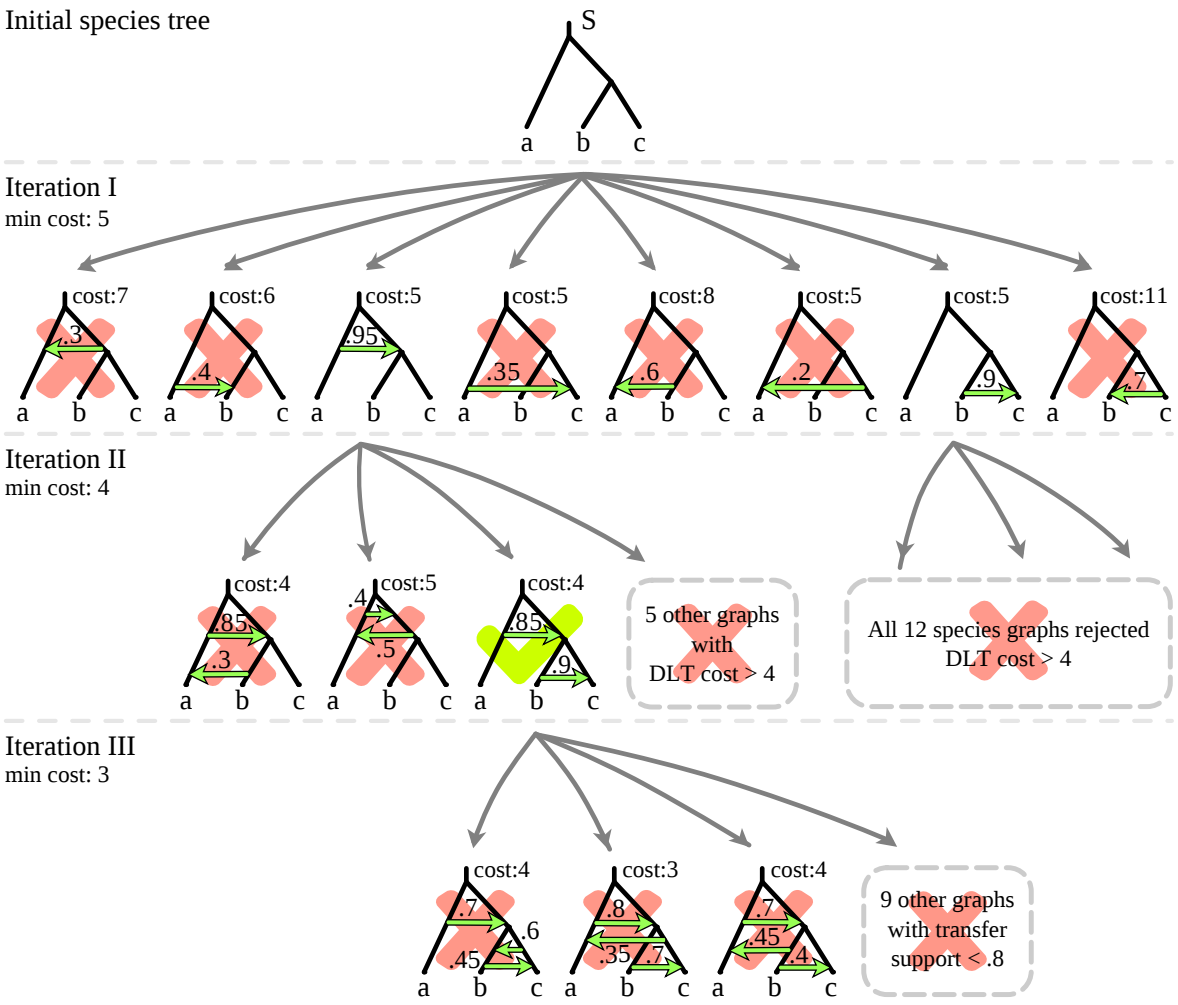


Figure 4.4: Example of execution of Algorithm 2. *From the top:* the initial species tree $(a,(b,c))$ and three iterations of the main loop with candidate species graphs. For each species graph, DLT cost and transfer support values are indicated on the right side of the rooting edge and the transfer edges, respectively. Here, a species graph has well-supported transfers if the support of each transfer is greater than .8. Rejected graphs are marked by red crosses. Under these criteria our algorithm returns the green-marked graph having cost 4 from the 2-nd iteration.

Algorithm 2 Inferring low scoring species graphs with well-supported transfers

1: **Input** T - a binary species tree, A - an alignment of sequences sampled from the species present in T.
2: Infer a rooted gene tree G from A.
3: Infer a set of sample trees $\mathcal{U}$ from A.
4: $\mathcal{R} := \{T\}$ # *Initialize the set of low scoring species graphs with well-supported transfers*
5: **Repeat** lines 6-11 until stopping condition is met # *Main loop:*
6: Let $X := \{S_{e,e'} : S \in \mathcal{R}$ and $\langle e, e' \rangle$ is valid for $S\}$.
7: $Y := \arg\min_{S \in X} \mathsf{c}_{\mathsf{HGT}}(S, G)$
8: For each $S \in Y$ compute $\mathsf{support}(h, S, \mathcal{U})$ for each transfer $h \in S$.
9: Remove species graphs from Y with low transfer support.
10: **If** Y is empty **Return** $\mathcal{R}$.
11: $\mathcal{R} := Y$
12: **Additional stopping conditions:** Return $\mathcal{R}$ if

- a sufficient number of transfers is inserted
- new transfers are not used by HGT-scenarios (equivalently, the cost is not changing)
- or the average support is below some threshold.

4.7.1. Analysis of inter-kingdom horizontal gene transfers

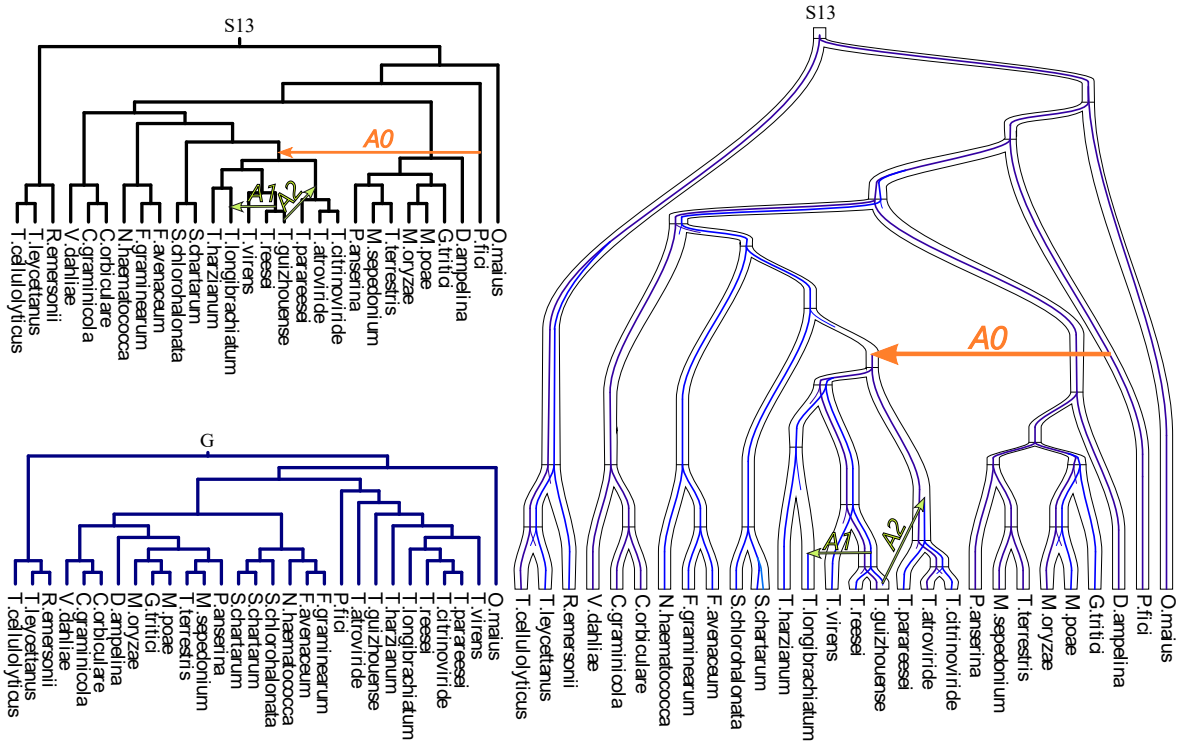


Figure 4.5: *Experiment I:* The results of the analysis of inter-kingdom horizontal gene transfers in the CaZy GH6 family of gene sequences from 29 fungus species [Druzhinina *et al.* (2018)]. *Left*: A species tree (S13) with transfers A0, A1 and A2 inferred after three iterations of our algorithm and the gene family tree (G). *Right*: Embedding of the gene tree into the species graph.

In this study, we wanted to check whether our algorithm will be able to find transfers within a group of relatively closely related organisms. We used an example of transfers detected between plant-associated filamentous fungi and members of genus *Pezizomycotina* [Druzhinina *et al.* (2018)].

Data preparation: First we inferred a binary species tree based on the NCBI taxonomy by using *phyloT* web-service. After downloading the gene sequences [Druzhinina *et al.* (2018)] from the NCBI database, we aligned them with MUSCLE [Edgar (2004)], and then we inferred a gene tree with *PhyML* program [Guindon *et al.* (2009)]. Next, we created a set of 100 bootstrap alignments from the original alignment with *seqboot* from PHYLIP package [Felsenstein]. Finally, we inferred sample gene trees from sample alignments by *PhyML*. In this experiment, all trees were rooted by *Urec* program [Górecki and Tiuryn (2007b)].

Experimental setting: For our study, we chose a gene tree of CaZy GH6 family groups cellobiohydrolases present in several plant-related fungi. In [Druzhinina *et al.* (2018)] authors claim that one transfer occurred in the evolution of this family. Our goal is to verify whether our algorithm can infer HGT-scenarios congruous

with the transfers proposed in [Druzhinina *et al.* (2018)]. We conducted several repetitions of the experiment with different sets of events weights $\langle$**DUP**, **HGT**, **LOSS**$\rangle$: $\langle 1,2,1\rangle$, $\langle 1,3,1\rangle$, $\langle 3,3,1\rangle$, $\langle 2,3,1\rangle$, and $\langle 1,2,0\rangle$. Our algorithm was set to perform three iterations. The total runtime for a single set of weights was approximately 4 hours with the DP algorithm executed about 35000 times on a standard PC workstation with Linux operating system.

Results: After performing tests we observed that the results for our choice of event weights are similar. Therefore, in Table 4.1 we present the summary of only one experiment for the weights $\langle 2,3,1\rangle$. At each step 5 species graphs were inferred reaching the lowest score of 37 for the graphs with 3 horizontal gene transfers. Almost every species graphs had exactly one optimal HGT-scenario.

Iteration	Sp. graph	A0	A1	A2	A3	A4	A5	A6	A7	A8	DLT Cost
	S1							.52			47
	S2				.5						47
1	S3			.5							47
	S4								.53		47
	S5									.68	47
	S6		.72	.72							41
	S7				.71	.71					41
2	S8				.74		.74				41
	S9		.71		.71						41
	S10		.73					.73			41
	S11	.65	.45		.55						37
	S12	.67			.57	.47					37
3	S13	.7	.49	.58							37
	S14	.67			.57	.48					37
	S15	.66	.46					.58			37

Table 4.1: *Experiment I:* Support values of HGTs and DLT costs calculated in three iterations of the iterative algorithm for the CaZy GH6 family of gene sequences from 29 fungus species. The transfer A0, that is highly supported in the third iteration, was proposed in [Druzhinina *et al.* (2018)]. The transfers A0, A1 and A2 are depicted in Figure 4.5, while the remaining HGTs are as follows - A3: *T. reesei*→((*T. citrinoviride*, *T. atroviride*), *T. parateesei*), A4: *T. reesei*→*T. longibrachiatum*, A5: *T. reesei*→*T. longibrachiatum*, A6: (*T. reesei*, *T. guizhouense*)→((*T. citrinoviride*, *T. atroviride*), *T. parateesei*), A7: *T. virens*→((*T. citrinoviride*, *T. atroviride*), *T. parateesei*), A8: ((*T. virens*, *T. reesei*), *T. guizhouense*)→((*T. citrinoviride*, *T. atroviride*).

Discussion: Obtained results for a cellobiohydrolase are congruous with results from [Druzhinina *et al.* (2018)] showing that there was a transfer from the pathogen of *Ficus carica* and tea endophyte *Pestalotiopsis fici* (order Amphisphaeriales) to the mycoparasitic *Trichoderma* clade (order Hypocreales). They also show that subsequent transfers of the locus within the *Trichoderma* clade are probable.

4.7.2. Analysis of horizontal transfers gene between distantly related species

In the second experiment, we tested our algorithm with a data set containing distantly related species separated in the species tree.

Data preparation: The dataset for this experiment was processed analogously to the dataset from the previous study except for the rooting step. Since in this case, we have a group of *Blastocystis* species that is outside of the main species group, our rooting method based on reconciliation will force the gene tree to have the root placed between the main group and the outgroup. As a result, the topology of the gene tree would not reflect the real evolutionary history. This is because it is more likely that the outgroup evolution is explained in a parsimonious way using vertical evolution without HGTs after the initial speciation event placed in the root of the gene tree. To avoid this problem, we decided to root the gene tree and bootstrap trees after removing the outgroup species and restoring them after the rooting is located. In that way, genes originated from *Blastocystis* were placed in the proper place in the gene family tree near their homologs and not extracted as an outgroup.

Experiment: In this study, we analyzed the family tree of Choline/sodium solute transporter genes from *Blastocystis* sp. ATCC 50177/Nand II and their closest homologs in Metazoa. The goal was to test whether we will be able to detect the same HGT that was identified in [Eme *et al.* (2017)]. We used the same sets of events weights as in the previous experiment. Having a similar experimental setting, the total runtime for the test with a single set of weights was approx. 25 minutes, as each iteration yielded only one species graph. In total, the DP algorithm was executed approx. 4000 times.

Results: In Figure 4.6 we depict the results of the first iteration, which appeared to be sufficient to infer the horizontal gene transfer, suggested in [Eme *et al.* (2017)], leading to the *Blastocystis* clade with the very high support of 91%. The next iterations also highly supported the transfer.

Discussion: Our algorithm shows that choline/sodium solute transporter which is a membrane protein was transferred from the springtail (*Folsomia candida*) lineage to a clade formed by several Stramenopile sequences. The Stramenopiles group unicellular or multicellular algae and other flagellated eukaryotic microorganisms among others parasitic *Blastocystis*.

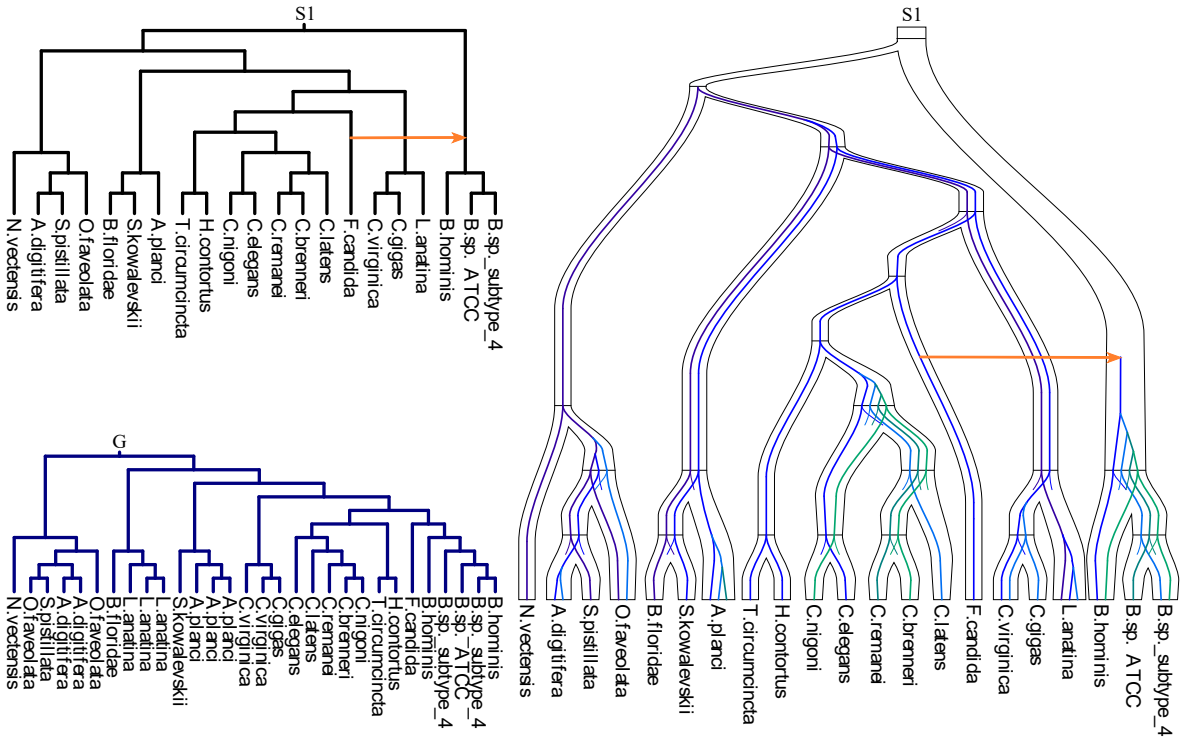


Figure 4.6: *Experiment II:* Results of the analysis of horizontal transfers gene between distantly related species. *Left*: Species tree (S1) of *Blastocystis* spp. and selected Metazoa species with one horizontal gene transfer best supported in the first iteration of the algorithm and the gene tree of choline/sodium solute transporter genes (G). *Right*: Embedding of the gene tree G1 into the species tree S1 representing the HGT-scenario with one well-supported HGT. * B.sp. ATCC 50177/NandII

4.7.3. Inference of the simulated transfers

In this experiment, we checked whether Algorithm 2 is able to infer known transfers. To perform this test, we simulated artificial data sets consisting of trees with known number and positions of HGTs.

Data preparation: For the simulation of gene and species trees, we used tools from *JPrIME-GenPhyloData* program [Sjöstrand *et al.* (2013)]. First, we simulated a set of species trees with *HostTreeGen* tool. The trees were generated over 80 time units with birth and death rate set to 0.03 and 0.0006, respectively. We accepted only trees with 10 leaves and with branches not shorter than 3. Then, for the simulation of gene trees, we used *GuestTreeGen* tool. The rates for duplications, losses and HGTs were 0, 0, and 0.006, respectively. The minimal and maximal number of leaves were set to 6 and 15. In summary, we created 100 species trees and for each species tree we simulated two gene trees – one with only one transfer (1-HGT) and the other with two transfers (2-HGT).

Next, we simulated a set of DNA sequences for each gene tree by using *Seq-Gen* [Rambaut and Grass (1997)]. To obtain different levels of the similarity of simulated sequences, we performed simulations with branch scaling parameter $s \in \{0.001, 0.004, 0.01, 0.05, 0.5\}$. Identity levels and the alignment evaluation score TCS [Chang *et al.* (2014)] of the simulated alignments are given in Table 4.2. Then, for the inference of the bootstrap trees, we used the *PhyML* program. Inferred trees were rooted with *Urec* [Górecki and Tiuryn (2007b)] using the original gene tree instead of the species tree to preserve the possible HGT signal in the bootstrap trees. The statistics concerning simulated sequences and inferred trees are presented in Table 4.2.

	1-HGT			2-HGT		
s	TCS	AIC	ACD	TCS	AIC	ACD
0.001	99.04	71%	22.36	99.04	69%	49.43
0.004	97.31	26%	21.74	97.29	23%	48.68
0.01	53.27	4.3%	23.43	53.45	3.6%	51.01
0.05	7.55	0.21%	71.04	6.86	0.14%	104.04
0.5	6.08	0.12%	84.41	5.54	0.06%	119.24

Table 4.2: The results of the simulation of the trees with transfers depending on the scaling parameter s and the number of simulated transfers (one and two). TCS and AIC columns present the alignment evaluation score [Chang *et al.* (2014)] and the average percentage of identical columns in the simulated alignments. ACD is the average cophenetic distance between the simulated gene tree and its sample trees.

Experiment: In the first step of the experiment, we checked how well simulated

transfers are supported. To conduct this test, we added transfers from the simulated gene trees to their proper occurrence locations in the species trees. For each species tree, we inferred two species graphs – one for the gene tree with a single transfer and another one for the gene tree with two transfers. Then, using simulated sets of the sample trees, we calculated the transfer support values for each species graph as we described in Section 4.5. The results are shown in the diagrams in Fig 4.7.

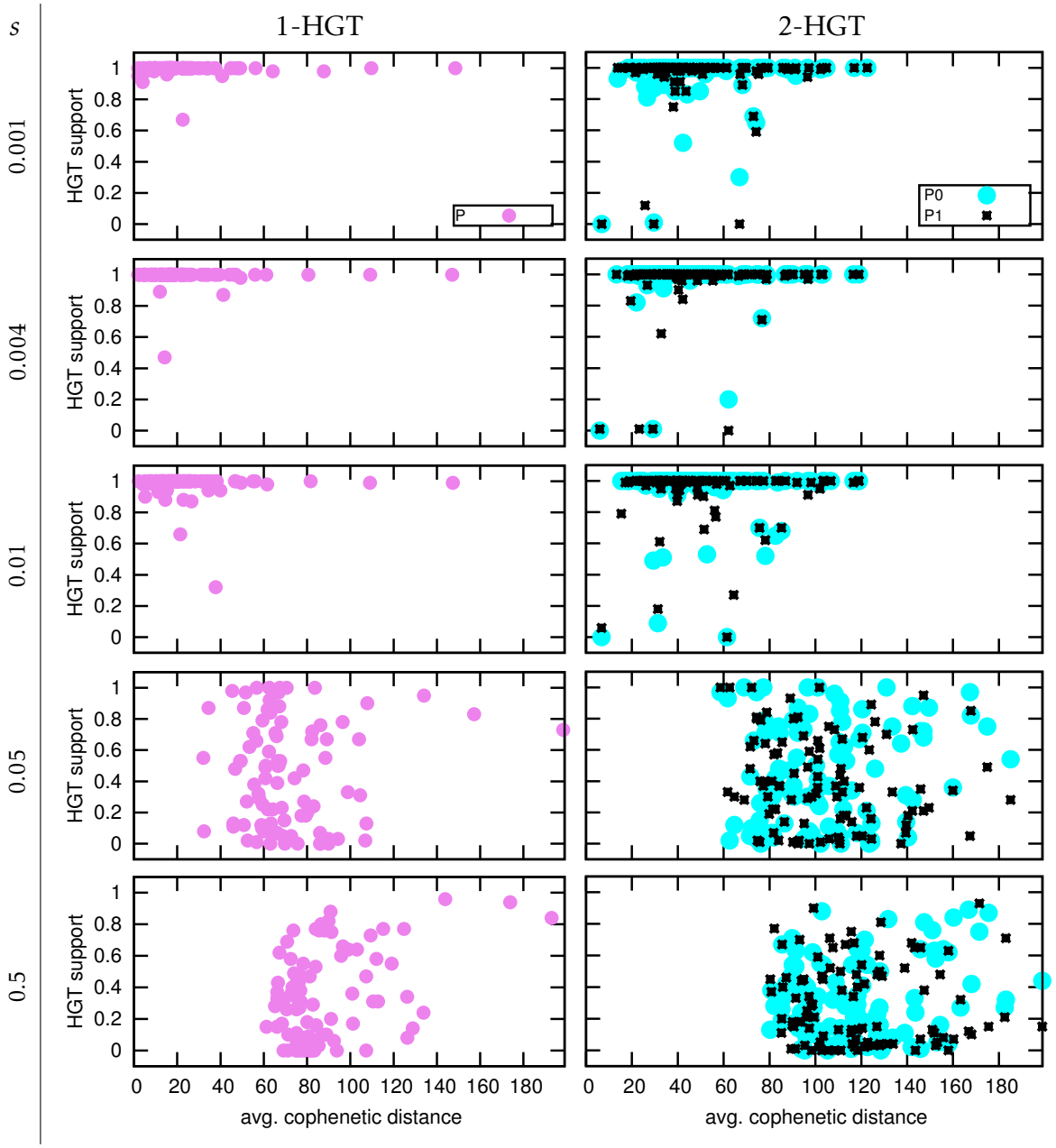


Figure 4.7: The results of the simulation of the trees with transfers. The diagrams show support values for the simulated transfers P for trees with single transfer (1-HGT) and $P0$, $P1$ for trees with two transfers (2-HGT), depending on the average cophenetic distance between the simulated gene tree and its sample trees.

	1- HGT		2- HGT		
s	CIS	avg. P supp.	CIS	avg. $P0$ supp.	avg. $P1$ supp.
0.001	91%	0.99	79% (53%)	0.95	0.94
0.004	91%	0.99	79% (53%)	0.96	0.94
0.01	91%	0.98	79% (53%)	0.94	0.93
0.05	85%	0.47	73% (52%)	0.46	0.39
0.5	79%	0.35	73% (54%)	0.33	0.29

Table 4.3: The results of the inference of simulated HGTs depending on the scaling parameter s and the number of simulated transfers (1-HGT, 2-HGT). CIS columns show the percentage of accepted inferred HGT-scenarios. For 1-HGT trees the results for basic and more restrictive conditions were the same, and for 2-HGT the results for the restrictive conditions are presented in brackets. The table also presents the average transfer support values for inferred transfers.

In the second step, we tried to infer simulated transfers using Algorithm 2 and the simulated gene and species trees. For each species tree, we run four iterations of the algorithm for the gene tree with one and two HGTs. Then, we checked whether the inferred transfers are the same as the simulated ones and if the HGT-scenario with a proper number of the transfers was optimal. If the algorithm found the simulated HGT-scenario and if it had the lowest cost, we counted a success. We also evaluated the results using more restrictive conditions. We accepted the HGT-scenario only if it satisfied the above conditions and if the cost of the HGT-scenario was lower than the cost from the previous iteration. The results are summarized in Table. 4.3, where we also present the average support values calculated for the inferred transfers P and $P0$, $P1$ for 1-HGT and 2-HGT, respectively.

Discussion: The results depicted in Table 4.3 show that the greater branch scaling parameter s was, the more variable was the alignment and the greater average cophenetic distance [Sokal and Rohlf (1962)] between the gene tree and its sample trees. Support values for the transfers decrease for less similar sample trees which is caused by more diversified topology. For branch scaling factors $s \leq 0.01$, support values were high for almost every set of the simulated trees and for $s \geq 0.05$ they gradually decreased. The significant difference between the results for $s = 0.01$ and $s = 0.05$ is due to the fact that for $s = 0.05$ the alignments are much more diverse and the average cophenetic distance is two times higher than for $s = 0.01$.

4.8. Discussion

In this chapter, we investigated the problem of the inference of well-supported horizontal gene transfers from multiple sequence alignments. To address the issue,

based on non-parametric bootstrap for phylogenetic trees we proposed a new measure, called transfer support, to verify the credibility of inferred transfers. We applied this approach to design a new iterative algorithm for inferring acyclic well-supported HGT-scenarios.

We tested the performance of our solution on two empirical datasets containing relatively closely and distantly related species groups. The results of both experiments showed that the method can be used to support known transfer hypotheses. However, it must be used with awareness considering the rooting problem. To examine the accuracy of the algorithm, we conducted experiments on simulated data. The results show that Algorithm 2 reached a high percentage of correctly inferred transfers both for trees with one and two HGTs. In particular, for alignments with good quality scores our proposed method can infer the correct HGT-scenario with high accuracy.

In the future, we plan to extend the definition of transfer support by incorporating alternative scoring schemas, e.g., based on cluster/clade contents rather than on the usage of transfers.

5

Inference of gene-species relationships

The complexity of microbial physiology and biochemistry makes it often impossible to create a clear picture of a microbial community one species at a time [Handelsman *et al.* (1998)]. Most frequently a shotgun metagenomic approach is used instead, where all microorganisms are sequenced together. As a result, the community is seen through the lens of its functional capabilities, not through the lens of the species that co-exist in this particular environment, as the latter picture is blurry due to limitations of species assignments. An approach used in parallel to shotgun metagenomics, where amplicons of marker genes are being sequenced, provides more accurate depiction of biodiversity [Weisburg *et al.* (1991)]. However, only handful of genes have a phylogeny similar to a species tree, such as ribosomal gene 16S or protein gene RecA [Thompson *et al.* (2004)]. Others are rarely used as species markers as neither LCA nor phylogeny based methods can be relied on, when no highly similar sequence to our query is available in reference database [Mande *et al.* (2012)].

In the classical reconciliation model, any incongruence between gene and species trees is explained as biologically consistent scenario having the minimal number of gene duplication and losses, called *duplication-loss* cost [Bonizzoni *et al.* (2005); Górecki and Tiuryn (2006)]. One of the most important extensions of the duplication-loss model [Bansal *et al.* (2015); Charleston (1998); Górecki (2004b); Górecki (2010); Hallet and Lagergren (2001); Lafond *et al.* (2012); Stolzer *et al.* (2012); Tofigh *et al.*

(2011); Szöllősi *et al.* (2013b)] is horizontal gene transfer, or HGT. Then the problem of reconciling two trees is defined as follows: *given a gene tree and a species tree, find a reconciliation scenario that minimizes the total number of gene duplication, gene loss and horizontal gene transfer events.* However, by postulating gene transfers reconciliation becomes usually computationally complex. For instance, reconstructing transfer scenarios under biologically consistent models, i.e., assuming that transfers cannot form cycles, is NP-complete [Tofigh *et al.* (2011)]. There are several approaches to deal with the complexity. By ignoring the acyclicity condition the problem becomes solvable in polynomial time by dynamic programming [Tofigh *et al.* (2011); Bansal *et al.* (2012)], however, there is no guarantee that the inferred scenario is biologically valid. Another approach is to assign a divergence time for the nodes of the species tree. With an additional requirement that a scenario respects the temporal ordering induced by the speciation times [Ranwez *et al.* (2015)] the problem of a scenario reconstruction has a polynomial time solution [Doyon *et al.* (2010); Bansal *et al.* (2012)].

In this Chapter we study recent applications [Betkier *et al.* (2015); Zhang and Cui (2010)] of reconciliation to the problem of *gene-species assignment*, that can be generally formulated as follows: *given a gene tree with partial leaf labelling and a species tree, resolve all missing labels in a gene tree such that the total reconciliation score is minimized*. An example is illustrated in Figure 5.3. According to our knowledge the gene-species assignment problem has been never studied before under the HGT extension. The first heuristic algorithm for a similar problem without transfers, defined for the deep coalescence [Maddison (1997)] cost and a special case of binary gene trees with bijective leaf labellings, was proposed in [O'Meara (2010)]. In [Zhang and Cui (2010)] $O(n^3)$ time algorithm was developed for the deep coalescence and gene duplication-loss cost functions and the analogous reconstruction problem under general leaf labellings. In different biological context $O(n^2)$ time algorithm was developed for the simplest duplication cost [Bafna *et al.* (2000)]. The optimal unified algorithm for all non-transfer reconciliation costs, that runs in $O(n^2)$ time, was recently developed by Betkier et al. [Betkier *et al.* (2015)].

We propose the first reconciliation-based formulation of the gene-species assignment problem for a model of evolutionary scenario with gene duplication, gene loss and horizontal gene transfer events for the case of a gene tree G and a species trees S with possible multifurcations and two tractable models with transfer events: time consistent (tcDTL) and general scenarios (DTL).

First, we present the definition of a DTL-scenario[1] and formulas for scenarios cost calculation with proofs. We follow by explaining the reconstruction of gene-species assignments and showing our algorithm for the assignment inference. We developed an algorithm for the DTL model that runs in $O(|G||S|)$ time if both trees are binary, and in $O(|G||S|\Delta S)$ time, where ΔS is the maximal out-degree of nodes from S and the gene tree is binary. For the time consistent model, we describe an $O(|G||S|^2)$ time algorithm and propose improvement that runs in $O(|G||S|\log|S|)$ based on data structures from [Bansal *et al.* (2012)]. We also propose a Monte-Carlo approach to approximate the distribution of gene-species mappings by sampling the space of optimal reconstructions. Having this, we provide a comparative study of reconstructions for empirical and simulated datasets using a prototype implementation of our algorithms.

5.1. General model of DTL-scenarios

The following definition of a DTL-scenario is adopted from [Bansal *et al.* (2012); Tofigh *et al.* (2011)], except here we focus more on event-based conditions.

Important note on HGT and DTL-scenarios: It is essential to distinguish the DTL model presented here from the HGT model described in Chapter 4. In HGT-scenarios, transfer events were inserted arbitrarily using specific rules. These rules ensured that the species graphs considered in the algorithm satisfied the acyclicity condition. In the DTL model, only a species tree is given, and transfer events are inferred by the algorithm based on the gene tree topology.

Recall that two nodes v and w from a rooted tree are incomparable if neither $v \preceq w$ nor $w \preceq v$ holds.

Definition 3 (DTL-scenario)**.** *A DTL-scenario for a binary tree G, and a tree S and a labelling $\Lambda\colon cL_G \to cL_S$ is a tuple $\langle \mathsf{M}, \Sigma, \Delta, \Theta, \xi\rangle$ such that Λ is the leaf labelling function, $\mathsf{M}\colon V_G \to V_S$ is a mapping that extends Λ, $\{\Sigma, \Delta, \Theta\}$ is a partition of I_G into speciation, duplication and transfer nodes, respectively, and $\xi\colon \Theta \to V_G$ determines the termination node of a transfer in G, subject the following conditions. For any internal node $g \in G$ such that c_1 and c_2 are the children of g let $s = \mathsf{M}(c_1) \oplus \mathsf{M}(c_2)$, then*

- *We have $g \in \Sigma$ if and only if the mappings of the children of g are incomparable, and $s = \mathsf{M}(g)$.*
- *If $g \in \Delta$ then $s \preceq \mathsf{M}(g)$.*

[1] In our work [Mykowiecka *et al.* (2017)], the DTL-scenario notion is simply called *scenario*. Here the name has been changed for better readability and due to a notation conflict.

- *If* $g \in \Theta$ *then* $\xi(g)$ *is a child of* g, $\mathsf{M}(\mathsf{sib}(\xi(g))) \preceq \mathsf{M}(g)$, *and* $\mathsf{M}(g)$ *and* $\mathsf{M}(\xi(g))$ *are incomparable. The edge* $\langle g, \xi(g)\rangle \in E_G$ *is called a transfer edge.*

The three conditions denote the cases of speciation, duplication and horizontal gene transfers events, respectively. In DTL-scenarios, *vertical*, i.e. tree-like, transfer is modelled by the condition that the mapping of a child is below or equal to the mapping of its parent. The condition holds for the children of speciation and duplication nodes, that are modeled in the classical vertical way [Górecki and Tiuryn (2006)]. For the transfer node g, its *horizontal* destination is defined by $\xi(g)$, therefore we require that both the mapping of g and the mapping of $\xi(g)$ are incomparable. On the other hand, the sibling of $\xi(g)$ is transfered vertically (see the last condition).

While it is clear how to interpret gene duplication and loss events in binary trees, it is generally difficult to model these events when multifurcations are present in input trees [Chang and Eulenstein (2006); Maddison (1989); Vernot *et al.* (2008); Zheng and Zhang (2014)]. Here we propose a computationally tractable solution in which a multifurcation in a species tree is a "true" speciation, where missing gene lineages are counted as gene loss events. In such models, however, a species tree with many multifurcations might induce optimal DTL-scenarios that prefer HGTs rather than gene losses as indicated in Figure 5.1 (see also discussion on the gene loss model in the second experiment with multifurcated species tree in Section 5.4). In practice, we suggest to set appropriate event weights **DUP**, **LOSS** and **HGT** depending on the biological context. Note that setting **HGT** $= +\infty$ yields a solution for the duplication-loss model similar to the algorithm from [Betkier *et al.* (2015)], however, in the latter article the number of losses is underestimated due to a different model applied in the loss formula.

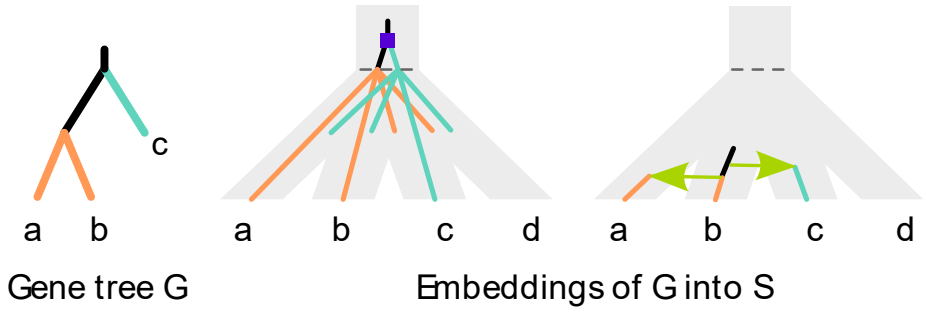


Figure 5.1: Two evolutionary scenarios for the non-binary species tree (a, b, c, d). *Left:* a gene tree G. *Middle:* scenario without HGTs having 1 gene duplication and 5 gene losses. *Right:* optimal DTL-scenario with 2 HGTs.

Since the species tree is multifurcated, the number of gene losses must incorporate degrees of the nodes. If a gene lineage passes through a speciation node in S

without branching, then $|\widehat{s}|-1$ gene losses are generated, i.e., one loss per each omitted child of s. See example in Figure 5.1. Such losses we call intermediate losses at S. Let $\mathsf{loss}(v,w)$ be the total number of intermediate losses at node s for each $v \prec s \prec w$. Similarly we define $\overline{\mathsf{loss}}(v,w)$ for nodes satisfying $v \prec s \preceq w$. Formally,

$$\mathsf{loss}(v,w) = \sum_{v\prec s\prec w} (|\widehat{s}|-1),$$

$$\overline{\mathsf{loss}}(v,w) = |\widehat{w}|-1+\mathsf{loss}(v,w) = \sum_{v\prec s\preceq w} (|\widehat{s}|-1).$$

Then the number of losses assigned to g is denoted by $\mathsf{L}(g)$ and defined as follows:

L1 If g is a speciation, then there are:

- $|\mathsf{M}(\widehat{g})|-|\widehat{g}|$ losses at $\mathsf{M}(g)$,
- $\mathsf{loss}(\mathsf{M}(c),\mathsf{M}(g))$ intermediate losses for each child c of g.

L2 If g is a duplication, then there are $\overline{\mathsf{loss}}(\mathsf{M}(c),\mathsf{M}(g))$ intermediate losses for each child c of g.

L3 If g is a transfer node, then there are $\overline{\mathsf{loss}}(\mathsf{M}(c),\mathsf{M}(g))$ intermediate losses for each child c of g such that $c \neq \xi(g)$.

L4 Finally, there are no losses at g if g is a leaf.

This yield the following formula for $\mathsf{L}(g)$:

(L1) If $g \in \Sigma$, then $\mathsf{L}(g) = |\widehat{\mathsf{M}(g)}|-|\widehat{g}|+\sum_{c\in\widehat{g}}\mathsf{loss}(\mathsf{M}(c),\mathsf{M}(g))$,

(L2) If $g \in \Delta$, then $\mathsf{L}(g) = \sum_{c\in\widehat{g}}\overline{\mathsf{loss}}(\mathsf{M}(c),\mathsf{M}(g))$,

(L3) If $g \in \Theta$, then $\mathsf{L}(g) = \sum_{\xi(g)\neq c\in\widehat{g}}\overline{\mathsf{loss}}(\mathsf{M}(c),\mathsf{M}(g))$,

(L4) $\mathsf{L}(g) = 0$ if g is a leaf.

Examples showing cases L1, L2 and L3 are depicted in Figure 5.2. Note that the formula becomes $\|\mathsf{M}(g),\mathsf{M}(\mathsf{par}(g))\| - \mathbb{1}[g \in \Sigma]$[2] when the trees are binary.

Let **DUP**, **LOSS** and **HGT** be non-negative event weights for duplication, loss and transfer events, respectively. The weighted *cost* of a DTL-scenario ϵ, denoted by $|\epsilon|$ is defined as the weighted total number of gene duplication, transfer and loss events present in ϵ. Formally, $|\epsilon| = \mathbf{HGT}\cdot|\Theta| + \mathbf{DUP}\cdot|\Delta| + \mathbf{LOSS}\cdot\sum_g \mathsf{L}(g)$. For given

[2]Here $\mathbb{1}$ is the indicator function, that is, $\mathbb{1}[p]$ is 1 if p is satisfied and 0 otherwise.

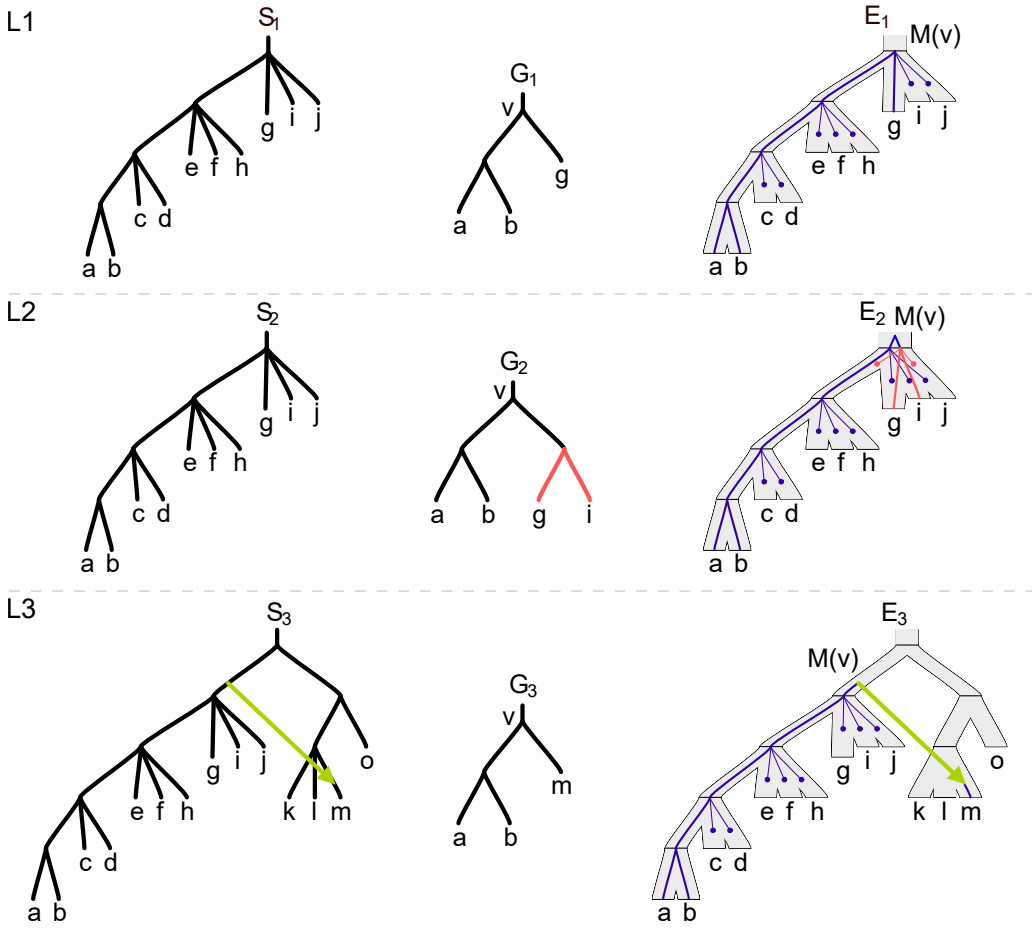


Figure 5.2: Examples of the tree reconciliation showing losses assigned to v in cases L1, L2 and L3. The numbers of losses L(v) for cases L1, L2 and L3 are 7, 8 and 8, respectively.

trees G, S and a labelling Λ a DTL-scenario is *optimal* if its cost is minimal in the set of all DTL-scenarios for G, S and Λ.

DTL-scenarios may be biologically incorrect in the sense that the transfers may form cycles. To capture only valid DTL-scenarios, [Tofigh *et al.* (2011)] introduced the notion of an *acyclic* DTL-scenario, called tcDTL-scenario, by using dated species trees. In the case of general DTL-scenarios, the minimal cost for a given gene tree and a species tree can be computed in $O(|G||S|)$ time [Tofigh *et al.* (2011); Bansal *et al.* (2012)], while the problem for acyclic DTL-scenarios is NP-hard [Tofigh *et al.* (2011)]. If there are no transfers in the scenario, the DTL model is equivalent to the DL model.

5.1.1. Inferring Gene-Species Assignments

We present two main problems for the reconstruction of gene-species mappings. For modeling undefined labels in gene trees we use partial functions. We express the problem of reconstruction of gene-species assignment in terms of converting a

partial function into a total one. For example, if $(a,(\perp,\perp))$ is a gene tree with two undefined labels denoted by $\perp$ and $(a,(b,c))$ is a species tree, then the problem is to replace all occurrences of $\perp$ by a, b or c such that the total cost is minimized (in this case the minimal cost would be 0). Another example of an inferred gene-species assignment is shown in Figure 5.3.

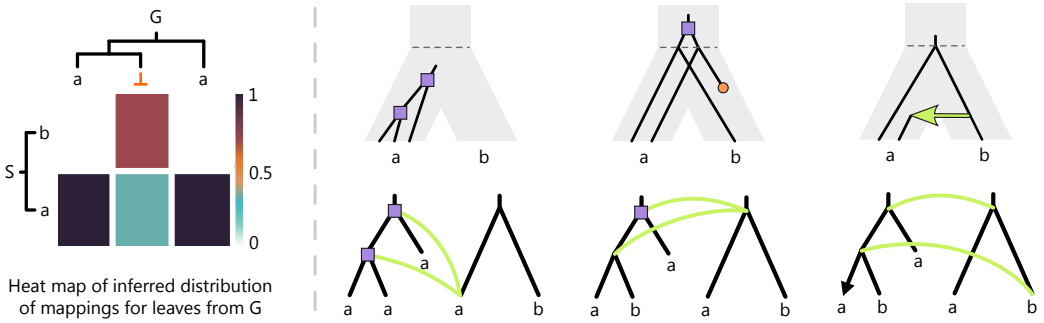


Figure 5.3: *Gene-species assignment inferred for the example gene tree G and species tree S. Left:* Heatmap showing inferred distributions of leaf mappings, i.e., the frequency of mapping of a given gene to each species. Trees S and G are placed on the sides of the heatmap. Missing leaf assignment in the gene tree G is denoted by "$\perp$". *Right:* Optimal evolutionary scenarios. Under the assumption that HGT event has cost 2 times that of duplications and losses, there are three optimal evolutionary scenarios.There are two duplications in the first scenario, one duplication and one loss in the second scenario, and HGT in the third scenario. If every optimal scenario is equally probable, the probability that $\perp$ is a is $\frac{1}{3}$, while for b it is $\frac{2}{3}$.

Let G and S be trees. Any partial function $\phi\colon \mathcal{L}_G \to \mathcal{L}_S$ will be called a *partial (leaf) labelling* from G into S. We write $\phi(x) = \perp$ if ϕ is undefined for x. Now, we present the problem of the reconstruction of leaf labellings, i.e., total functions, from partial labellings.

Problem 4. *Given trees G, S and a partial labeling ϕ between G and S. Find a DTL-scenario ϵ for G, S and Λ such that (1) Λ (a total function) extends ϕ, and (2) $|\epsilon|$ is minimal in the set of DTL-scenarios for G, S and Λ' such that Λ' extends ϕ. Such a DTL-scenario is called* optimal.

For given trees G, S and a partial labeling ϕ we denote the minimal cost introduced in Problem 4 by $c_{DTL}(G,S,\phi)$.

5.1.2. Dynamic programming formula for optimal DTL-scenarios

In this section we propose a polynomial time algorithm for the computation of minimal costs. The algorithm is an extension of Algorithm 1 from [Betkier *et al.* (2015)].

Given a gene tree G, a species tree S and $\phi\colon \mathcal{L}_G \to \mathcal{L}_S$ a partial labeling we show how to compute $c_{DTL}(G,S,\phi)$. By $G|g$ we denote the subtree of G rooted at g.

The dynamic programming formula has several components δ, $\delta^{\triangle}$, $\delta^{\uparrow}$ and $\delta^{\rightarrow}$ that denote the minimal cost of DTL-scenarios for $G|g$, S and Λ where for $g \in V_G$ and $s \in V_S$ we have additional conditions:

D1 $\delta(g,s)$ is the minimal cost in the set of all DTL-scenarios for $G|g$, S and Λ such that Λ extends ϕ (in the set $\mathcal{L}_{G|g}$) and g is mapped into s.

D2 $\delta^{\triangle}(g,s)$ is the minimal cost for DTL-scenarios for $G|g$, S and Λ such that Λ extends ϕ and g is mapped into node form $S|s$.

D3 $\delta^{\uparrow}(g,s)$ as above but the cost has additional gene loss events on the path between $\mathsf{M}(g)$ and s. Formally, the cost is the minimal value defined as:

- the cost of a DTL-scenario for $G|g$, S and Λ such that Λ extends ϕ and g is mapped to a node v from $S|s$, i.e., $\delta(g,s)$,
- plus the cost of intermediate gene losses for each node on the path from s to v excluding v, i.e., $\overline{\mathsf{loss}}(s,v) \cdot \mathbf{LOSS}$.

D4 $\delta^{\rightarrow}(g,s)$ is the minimal cost in the set of all DTL-scenarios for $G|g$, $S|s'$ and Λ such that Λ extends ϕ, s and s' are incomparable, and g is mapped into s'.

For δ we have the following formulas:

$$\delta(g,s) = \begin{cases} 0 & \text{if } g \text{ and } s \text{ are leaves and } \phi(g) \in \{s, \bot\}, \quad (1) \\ \min\{\alpha, \beta, \gamma\} & \text{if } g \text{ is not a leaf}, \quad (2) \\ +\infty & \text{otherwise}, \quad (3) \end{cases}$$

where,

$$\alpha = (|\hat{s}| - |\hat{g}|) \cdot \mathbf{LOSS} + \min_{\substack{p\colon \widehat{g} \to \widehat{s} \\ p \text{ is a "one-to-one" function}}} \sum_{c \in \widehat{g}} \delta^{\uparrow}(c, p(c)),$$

$$\beta = \mathbf{DUP} + \min_{\substack{p\colon \widehat{g} \to \widehat{s} \cup \{s\} \\ p(x)=s \text{ for some } x}} \sum_{c \in \widehat{g}} \begin{cases} \delta(c,s) & \text{if } p(c) = s, \quad (4) \\ \delta^{\uparrow}(c, p(c)) + (|\hat{s}| - 1) \cdot \mathbf{LOSS} & \text{if } p(c) \in \widehat{s}, \quad (5) \end{cases}$$

$$\gamma = \mathbf{HGT} + \min_{c \in \widehat{g}} \delta^{\rightarrow}(c,s) + \delta^{\uparrow}(\mathsf{sib}(c), s).$$

Functions p in above definitions denote all valid mapping assignments for the children of g. In particular, α represents the case when g is a speciation node [Górecki and Tiuryn (2006)], i.e., all children of g are mapped below s, β represents the case when g is a duplication node, i.e., at least on child of g is mapped to s. Finally, γ is the case of transfer, where the child c of g is transfered.

The formulas for $\delta^{\uparrow}$ and $\delta^{\rightarrow}$ can be expressed as:

$$\delta^{\uparrow}(g,s) = \begin{cases} \delta(g,s) & \text{if } s \text{ is a leaf,} \quad (6) \\ \min\{\delta(g,s), (|\hat{s}|-1)\cdot \mathbf{LOSS} + \min_{x\in\widehat{s}} \delta^{\uparrow}(g,x)\} & \text{otherwise,} \quad (7) \end{cases}$$

$$\delta^{\rightarrow}(g,s) = \begin{cases} +\infty & \text{if } s \text{ is the root of } S, \quad (8) \\ \min\{\delta^{\rightarrow}(g,\mathsf{par}(s)), \min_{q \text{ is sibling of } s}(\delta^{\triangle}(g,q))\} & \text{otherwise,} \quad (9) \end{cases}$$

where

$$\delta^{\triangle}(g,s) = \min\{\delta(g,s), \min_{c\in\widehat{s}} \delta^{\triangle}(g,c)\}.$$

Theorem 8 (Correctness)**.** *For a binary gene tree G, a species tree S, a partial labeling ϕ we have $c_{DTL}(G,S,\phi) = \delta^{\triangle}(\operatorname{root} G, \operatorname{root} S)$.*

Proof. The proof is by induction on the structure of G and S. In the induction step it is sufficient to show that the properties (I)-(IV) hold for δ functions.

The base step is when g is a leaf. Then, DTL-scenario:

- For D1 exists only when g has label s or $\perp$ which yields cost 0 (case (1)). Otherwise, the cost is $+\infty$ (case (3)).
- For D2, we have $\delta^{\triangle}(g,s) = 0$ if and only if $\phi(g) \in \mathcal{L}(S|s)$ or $\phi(g) = \perp$. Otherwise, $\delta^{\triangle}(g,s) = +\infty$
- For D3, $\delta^{\uparrow}(g,s) = \overline{\mathsf{loss}}(s,s^*)$ if and only if s^* is a leaf in $S|s$ sich that either $\phi(g) = s^*$ or $\phi(g) = \perp$ and s^* has the minimum edge distance to s in $S|s$. Otherwise the value of $\delta^{\uparrow}(g,s)$ is $+\infty$.
- For D4 observe that if $\pi(g) \notin \mathcal{L}(S|s)$, then $\delta^{\rightarrow}(g,s) = 0$ and $\delta^{\rightarrow}(g,s) = +\infty$, otherwise.

We omit easy verification of cases D2, D3 and D4.

Inductive assumption: For a non leaf $g \in V_G$ and $s \in V_S$ cases D1, D2, D3 and D4 hold for every node $c \prec g$ and every node from V_S.

D1 If g is mapped to s in a DTL-scenario, then we have three cases depending on the type of event:

Speciation If g is a speciation, then its all children map below $\mathsf{M}(g)$. Therefore, the cost of such scenarios has $|\widehat{s}| - |\widehat{g}|$ losses assigned to g at S. Next, each child c of g must map to or below a unique child s' of s. The cost contribution of a child c is, by the inductive assumption for D3, $\delta^{\uparrow}(c,s')$, which is $\mathbf{LOSS}\cdot\overline{\mathsf{loss}}(\mathsf{M}(c),s') + \delta(c,\mathsf{M}(c)) = \mathbf{LOSS}\cdot\mathsf{loss}(\mathsf{M}(c),s)) + \delta(c,\mathsf{M}(c))$. Based on L1 we see that all gene losses at g are included in (α), where p is the "one-to-one" function that assigns children of g to the children of s.

Duplication In this case at least one child has to map to s. Then, each child of g contributes to the cost with $\overline{\mathsf{loss}}(s, \mathsf{M}(c))$ losses (from (L2)) plus the cost $\delta(c, \mathsf{M}(c))$.

Now, let $p : \widehat{g} \to \widehat{s} \cup \{s\}$ be the funtion assigning $p(c) = s$ if c maps to s (at least one c satisfies the property) and $p(c) = s'$, otherwise, where s' is the child of s such that $\mathsf{M}(c) \preceq s'$. Note that p does not have to be one-to-one. Then, if $p(c) = s'$, for $s' \in \widehat{s}$, then $\overline{\mathsf{loss}}(c, s') = |\widehat{s}| - 1 + \mathsf{loss}(c, s')$. By incorporating one duplication at g, we obtain the formula (β).

HGT If g is a transfer node such that $c = \xi(g)$, then c must be mapped to a node s' that is not comparable with s. Next, the sibling of c requires $\overline{\mathsf{loss}}(\mathsf{M}(g), s)$ intermediate losses (L3). Thus, the cost of such a HGT-scenario is $\mathbf{HGT} + \delta^{\rightarrow}(c, s) + \delta^{\uparrow}(\mathsf{sib}(c), s)$. Minimizing over all choices of transfer node yields (γ).

This completes the proof of D1.

D2 The proof follows easily from $\delta^{\triangle}(g, s) = \min_{s' \preceq s} \delta(g, s') = \min(\delta(g, s), \min_{c \in \widehat{s}} \delta^{\triangle}(g, c))$.

D3 If the mapping of g is s in the optimal scenario, then the cost is $\delta(g, s)$ by the already proved condition D1. Otherwise, there is a child s' of s such that $\mathsf{M}(g) \preceq s'$. Now, by inductive assumption for D3, we have that $\delta^{\uparrow}(g, s') = \mathbf{LOSS} \cdot \overline{\mathsf{loss}}(\mathsf{M}(g), s') + \delta(g, \mathsf{M}(g))$. Since at s we have $|\widehat{s}| - 1$ intermediate losses, we obtain $\delta^{\uparrow}(g, s') + \mathbf{LOSS} \cdot (|\widehat{s}| - 1) = \overline{\mathsf{loss}}(\mathsf{M}(g), s) \cdot \mathbf{LOSS} + \delta(g, \mathsf{M}(g))$, and the above value equals $\delta^{\uparrow}(g, s)$ if the minimal cost is obtained for a HGT-scenario with $\mathsf{M}(g, s) \preceq s'$.

D4 If s is the root of S, then there is no node incomparable with s. Thus, $\delta^{\rightarrow}(g, s) = +\infty$ in such a case. Otherwise, let $p_1, p_2, \cdots, p_k$ be the path connecting s and the root of S. Then a node $v \in S$ is incomparable with s if and only if there is $i \in \{1, \ldots, k\}$ and a sibling q of p_i such that $v \preceq q$.

Easy proof of the above property follows from the fact that q is a child of the least common ancestor of s and v. Now,

$$\delta^{\rightarrow}(g, s) = \min_{\substack{i \in \{1,\ldots,k\} \\ q \text{ is a sibling of } p_i, \\ \text{and } v \preceq q}} \delta(g, v) = \min_{\substack{i \in \{1,\ldots,k\} \\ q \text{ is a sibling of } p_i}} \delta^{\triangle}(g, s) =$$

$$= \min(\delta^{\rightarrow}(g, \mathsf{par}(s)), \min_{q \text{ is a sibling of } s} \delta^{\triangle}(g, q)).$$

This completes the proof of D4.

Now the main formula from Theorem 8 follows easily from D2. This completes the proof. ■

This algorithm can be naturally extended to infer an optimal DTL-scenario by using standard backtracking method.

Theorem 9 (Complexity of multifurcated variant)**.** *For a binary gene tree G and a species tree S. The minimal cost can computed in $O(|G||S|\Delta S)$ time, where ΔS is the maximal out-degree of nodes from S.*

Proof. It is sufficient to show that all δ's can be computed in $O(\Delta S)$ time. Note that only α and β are more difficult to compute. For α, we have to find a "one-to-one" function p having the minimal value $\delta^{\uparrow}(c, p(c)) + \delta^{\uparrow}(c', p(c'))$, where $\hat{g} = \{c, c'\}$. This can be done in $O(\Delta S)$ time by finding the two mappings with the minimal and the second minimal value of $\delta^{\uparrow}$ for both children and choosing the optimal pair among four possibilities. For the second case, β is the minimal value of $\delta^{\uparrow}(c, s) + \min_{s' \in \hat{s}}(\delta^{\uparrow}(c, s') + \mathbf{LOSS}, \delta(c, s))$ for $c \in \hat{g}$, which can be computed in $O(\Delta S)$ time. ■

5.2. Time consistent DTL-scenarios

The most standard way of modelling tractable acyclic scenarios, is by introducing a time stamp for the nodes of a species tree and defining consistency conditions [Górecki (2004b)] or by introducing an ordering based on transfers mappings. Probably the simplest acyclicity condition is given in [Tofigh *et al.* (2011)]. A scenario is *acyclic* if there is a total order $\prec$ on V_S such that:

1. (1) if $(x, y) \in E_S$ then $x \prec y$,
2. (2) if $u, v \in \Theta$ and $\xi(u) \succeq \xi(v)$ then $\mathsf{par}(\mathsf{M}(u)) \prec \mathsf{M}(\xi(u))$ (see [Tofigh *et al.* (2011)]).

Our model of HGT-reconciliation is based on *dated species trees* in which each node has a divergence time defined as a function $\tau\colon V_S \to \mathbf{R}^+$ such that $\tau(s) \succ \tau(s')$ if $s \prec s'$. We say that two distinct species represented by nodes s and r *coexisted* if the time intervals $(\tau(\mathsf{par}(s)), \tau(s))$ and $(\tau(\mathsf{par}(r)), \tau(r))$ have a non-empty intersection. We write that s is *transferable* to r, if s and r or an ancestor of r coexisted. Note that transferability implies incomparability. For genes in a DTL-scenario with dated trees, the notion is expressed in terms of mappings as follows.

Definition 4. *A time consistent DTL-scenario, or tcDTL-scenario, for a gene tree G, a dated species tree S, and a labelling $\Lambda\colon \mathcal{L}_G \to \mathcal{L}_S$ is a DTL-scenario for G, S and Λ such that for every $g \in \Theta$, the node $\mathsf{M}(g)$ is transferable to $\mathsf{M}(\xi(g))$.*

[Tofigh *et al.* (2011)] proved that tcDTL-scenarios are acyclic. While the definition of a DTL-scenario cost (ϵ) can be rewritten from DTL-scenarios, for proper computation of gene losses we need to modify the definition of $\mathsf{L}(g)$ given in the previous section. For the case when g is a transfer node, there are additional intermediate gene loss nodes on the path connecting $\mathsf{M}(\xi(g))$ and s, including the losses at $\mathsf{M}(\xi(g))$, where s is the lowest ancestor of ancestor of $\mathsf{M}(\xi(g))$ that coexisted with $M(g)$. Thus, the formula L3 becomes:

(L3)

$$\mathsf{L}(g) = \overline{\mathsf{loss}}(\mathsf{M}(\xi(g)), s') + \sum_{\substack{c \in \widehat{g} \\ c \neq \xi(g)}} \overline{\mathsf{loss}}(\mathsf{M}(c), \mathsf{M}(g))$$

Cases L1, L2 and L4 remain unchanged. Example showing case L3 is depicted in Figure 5.4. Similarly, the definition of the tcDTL-scenario cost is analogous to the definition of the cost for the DTL-scenario. We omit straightforward details. Then, the minimal cost for a given gene tree, i.e., when all labells are defined in G, and a dated species tree can be computed by $O(|G||S|\log|S|)$ time [Tofigh *et al.* (2011); Bansal *et al.* (2012)].

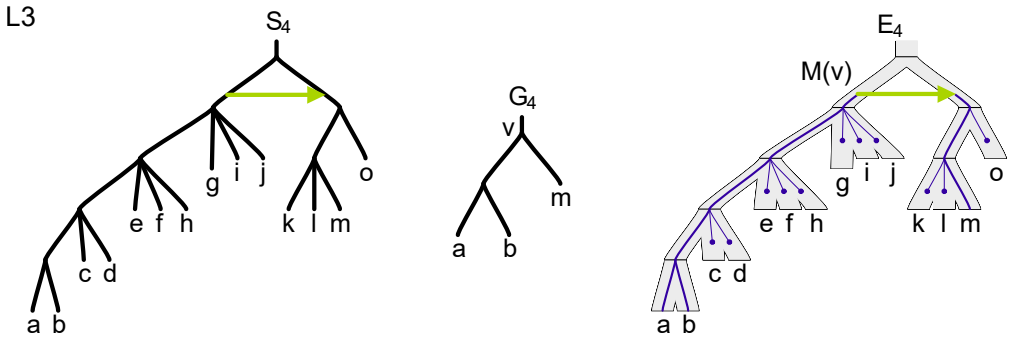


Figure 5.4: Examples of the tree reconciliation showing losses assigned to v in case L3. The number of losses $\mathsf{L}(v)$ is 11.

5.2.1. Inference of gene-species association under the tcDTL model

We show a polynomial time algorithm for the problem of gene-species assignment under the time consistent model of tcDTL-scenarios. To avoid repetitions we skip straightforward definitions and problem formulation. The algorithm is defined similarly: we take the definitions and the formulas from the previous Section and modify the definition of $\delta^{\rightarrow}$ as follows. Now, $\delta^{\rightarrow}(g,s)$ is the minimal cost in the set of all tcDTL-scenarios for $G|g$, $S|s$ and Λ such that Λ extends $\phi|_{L_{G(g)}}$, s is transferable to s', and g is mapped into s' or its descendant. Formally,

$$\delta^{\rightarrow}(g,s) = \min_{s \text{ coexists with } s'} \delta^{\uparrow}(g,s').$$

Due to the above change, computing a single value of $\delta^{\rightarrow}$ requires traversal of $O(|S|)$ nodes in S. Hence, the time complexity of computing the minimal cost of under time consistent model for a given binary gene tree and a dated species tree S is $O(|G||S|^2\Delta S)$. This complexity can by further improved to $O(|G||S|\Delta S \log|S|)$ by applying data structures proposed in [Bansal *et al.* (2012)].

5.3. Extensions

Inferring gene-species assignments in unrooted gene trees can be formulated by searching for the rooting that minimizes the optimal (rooted) cost [Górecki *et al.* (2013)]. In such a case the algorithm consists of an additional loop that iterates over all possible edges of the input gene tree. In each step, a formula similar to δ^{Δ} is evaluated for the artificial root with children being the end nodes of the current edge. We omit easy details for brevity. The complexity of this algorithm is the same as for the rooted case.

Another extension is to extend the problem to non-binary gene trees. Assuming the simplest tractable model of non-binary reconciliation it can be shown that the time complexity of computing minimal cost under the DTL model is $O(|G||S|\max(\Delta G, \Delta S)^3)$.

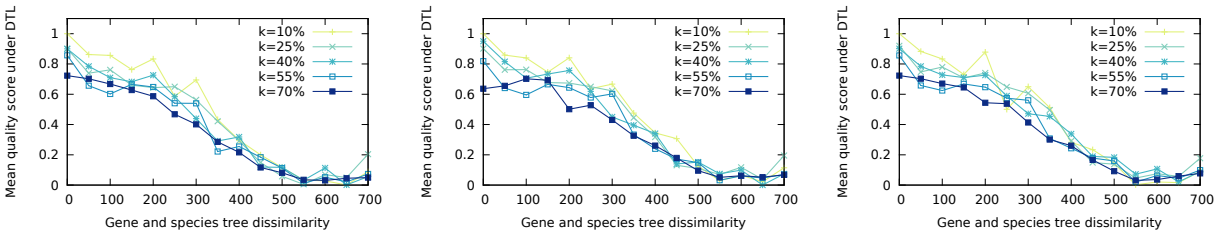


Figure 5.5: Mean quality score for the reconstructions of gene species mappings under DTL cost function. The quality score on the Y axis represents the correctness of gene-species assignment, e.g., the quality score equal 1 means, that every unknown label was correctly assigned. The parameter k denotes what percentage of labels were set to be unknown in the input labeling of a gene tree. From the left side the diagrams depict results for the following weights: D1 L1 T1, D1 L1 T2 and D3 L1 T3.

5.4. Experimental Results

We conducted two computational experiments by using our prototype implementation of the algorithm described in the previous sections with additional procedures

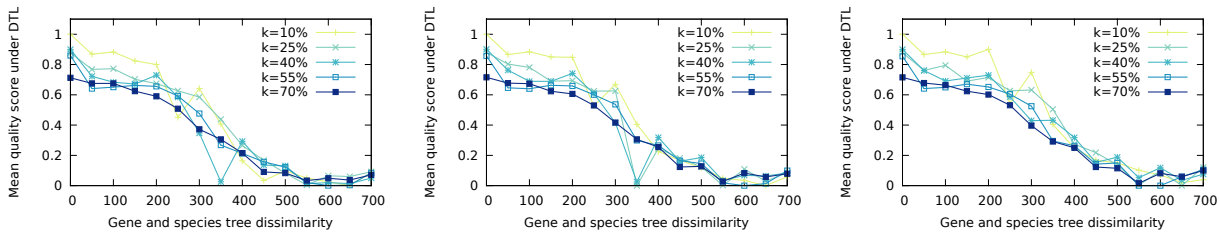


Figure 5.6: (Continued from Figure 5.5) Mean quality score for the reconstructions of gene species mappings under the tcDTL cost

for optimal scenarios and leaf mappings inference. In the first experiment, we investigated the quality of the inferred gene-species assignments using simulated data. In the second one, we studied the dataset consisting of 100 proteins from *Methanobrevibacter ruminantium*.

In the experiments, the triplets notation, such as D10 L1 T10, denotes the set of event weights. In this example we the weight of gene duplication (D) and transfer (T) equals 10 while for the gene loss (L) it is 1.

5.4.1. Reconstruction quality

First, we performed a set of tests on simulated binary trees to check the quality of the inferred reconstructions. We performed computational experiments for both, the DTL and tcDTL models.

Data preparation: We used a simulated dataset which consists of 350 pairs of gene and species trees having 100 bijectively labelled leaves simulated by using the procedure form [Betkier *et al.* (2015)] according to the Yule-Harding model. For the time consistent model we also simulated branch lengths – for every species tree there where 100 repeats of branch lengths drawings from uniform distribution. Gene and species trees were generated in a way, that the dissimilarity measure based on the *deep coalescence* score [Maddison (1997)] (DC) in our dataset is uniformly distributed over the interval $[0, 700]$, where the score 0 indicates identity of trees, while the score close to 700 indicates high level of topological incongruence.

To validate if unknown genes are assigned to proper species, we removed $k \in \{10\%, 25\%, 40\%, 55\%, 70\%\}$ leaf labels from each input gene tree. The procedure of leaf removal was repeated 10 times for each k.

Evaluation: For all pairs of trees, we reconciled a gene tree with its species tree. Since more than one reconstruction can have the minimal DTL cost, we drew 1000 optimal scenarios by using Monte-Carlo method and inferred the distribution of

gene-species assignments. Then, we checked the probability that leaves with unknown labels are assigned to proper species. The inferred quality score is the ratio of proper gene-species assignments divided by number of leaves with unknown labels. In the time consistent model, the score for 10 variants of tree dating was calculated as an average from all sampled trees. Diagrams with species assignment quality for the DTL model are presented in Figure 5.5 and results for the tcDTL model are depicted in Fig 5.6. For both models we present results for three sets of events weights. To make diagrams more smooth and readable, tree pairs are binned into groups of the size 50 having the dissimilarity score in interval 0-49, 50-99, and so on.

Discussion: Results for the DTL model, depicted in Figure 5.5, show that for trees with low DC score, the mean quality of gene-species assignment is relatively high. In the diagram we can see, that the best quality is obtained when reconciled trees have similar topology and the parameter k is low, although the differences between the lowest and the highest k value are relatively small. The quality of reconstructions obtained for the tcDTL model (see Figure 5.6) are similar to the DTL model, which is slightly surprising, given more constrained model of scenarios and random model of speciation time generation.

5.4.2. Real dataset evaluation: multifurcated species tree vs. binary gene tree

In the following experiment, we studied the performance of our algorithm on the dataset containing phylogenetic trees inferred from real sequences.

Data preparation: Our algorithm for the DTL model was tested with the dataset from [Betkier *et al.* (2015)], representing a typical scenario of amplicon analysis. The dataset consists of 100 proteins from *Methanobrevibacter ruminantium* similar to mcrA gene that has been proposed as a marker gene in the phylogenetic analysis of archeal methanogen populations [Luton *et al.* (2002)]. The list contained genes from uncultured archeons. The unrooted gene tree was built using program `proml` from the phylip package. Original species tree containing over 1400 known *Euryarchaeota* species from SILVA database [Quast *et al.* (2013)] was contracted to the set of species reconstructed in [Betkier *et al.* (2015)]. We have attempted to resolve mappings of 9 unknown sequences out of 100.

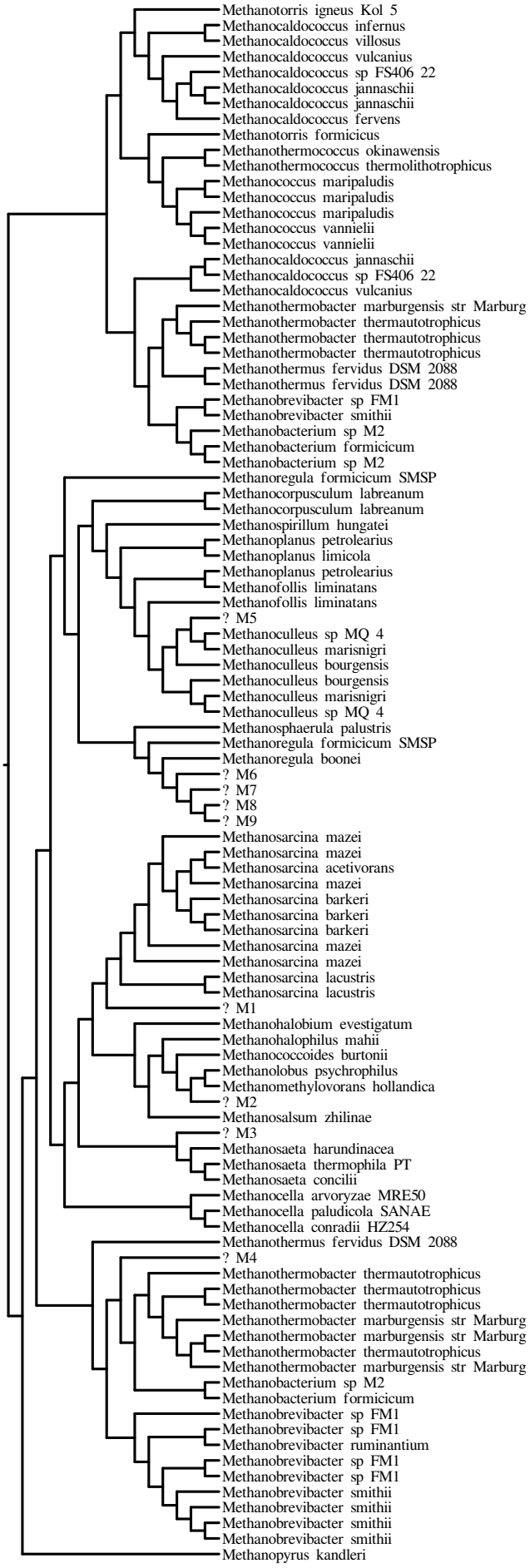


Figure 5.7: A joint optimal rooting of a gene tree from [Betkier *et al.* (2015)] of 100 proteins similar to mcrA from *Methanobrevibacter ruminantium* with 9 unknown gene-species labels M1-M9.

Evaluation: For the mcrA dataset we carried out six rounds with different sets of weights **DUP**, **LOSS** and **HGT**. Our algorithm requires a rooted input, therefore we first rooted the unrooted gene tree depicted in Figure 5.7, by choosing one rooting that minimizes the weighed cost for every set of weights. Next, as the number of optimal DTL-scenarios may be large, we conducted 1000 draws by using Monte-Carlo method. DTL-scenarios are drawn according to the uniform distribution, i.e., each path is selected proportionally to the number of optimal solutions which are represented by the subtree it points to. Having a sample of 1000 random optimal DTL-scenarios for a single set of weights, we inferred the distribution of gene leaf mappings in these scenarios. Summary of results is depicted in Figure 5.8.

Discussion: The first analysis of the resulting distribution of gene-species mappings indicated that in many cases the missing assignments were uniquely reconstructed despite the results from [Betkier *et al.* (2015)] for the DL cost. A closer analysis of the corresponding DTL-scenarios especially in the first two columns representing cases of similar weights, showed that some assignments, e.g. M1, are unique, however, is is not difficult to find a possible better assignment based on the known mappings of some neighboring genes. Moreover, given high level of multifurcation in the species tree one could expect more uniform distribution of mappings to species whose parent represent a speciation spanning many species (see for example the unique reconstruction for M2 in D1 L1 T1). This phenomenon can be partially explained by the model of gene loss events (see also Figure 5.1) that is probably overestimating the number gene losses when reconciling in a non-transfer way in the presence of a multifurcation in the species tree. In other words, when searching for the optimal DTL-scenario, it is "cheaper" to embed a binary part of a gene tree into a binary part of a species tree and then transfer gene lineages to its proper location, rather than forcing expensive embedding with a large multifurcation that usually requires many gene losses. This observation can be also confirmed by the highly ambiguous distributions for D1 L0 T1, where gene losses are ignored, and partially in D10 L1 T10, where gene losses are significantly cheaper than duplications and transfers.

5.5. Conclusions

In this Chapter we proposed the first HGT reconciliation-based approach for inferring gene-species mappings. We developed efficient algorithms for optimal cost computation and inference of gene-species assignment under weighted cost functions with gene duplication, gene loss and HGT events. Our prototype implemen-

tation of the algorithm indicated that this approach is capable of enhancing the taxonomic assignment of metagenomic sequences.

In future we plan to test in more detail the impact of event weights and reconciliation models with possible multifurcations on the reconstruction quality. We also plan to further evaluate the method on large empirical and simulated datasets. Further extensions include methods for the analysis of whole metagenomic samples that may contain sequences from many gene families.

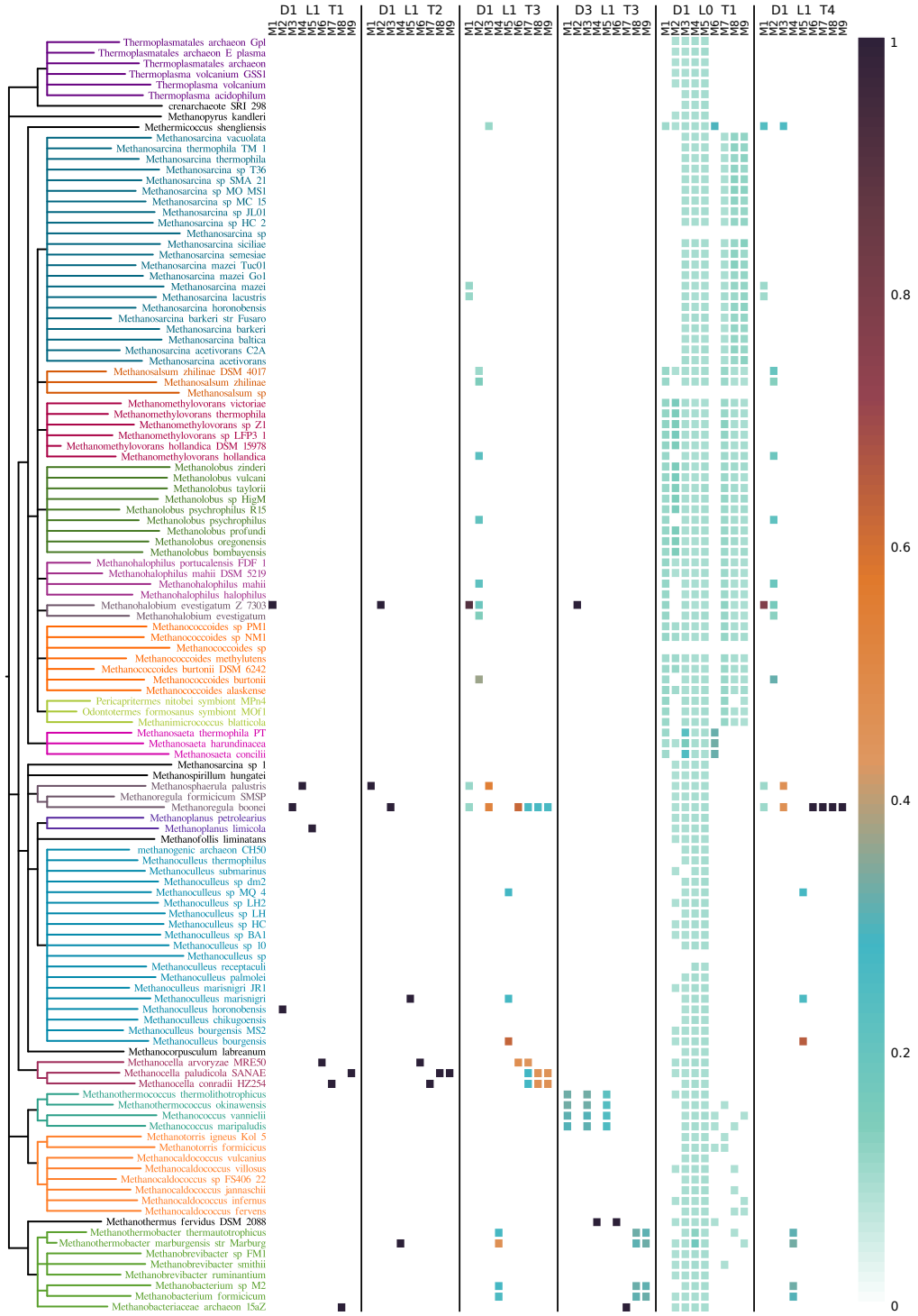


Figure 5.8: *Inferrence of gene-species assignment for mcrA dataset. Left:* A part of the SILVA species tree with species present in the reconstructed gene-species distributions. *Right:* Six heat maps of inferred distributions of mappings for leaves M1-M9 from the gene tree. Each heat map corresponds to one parameter set. Weights of gene duplication (D), loss (L) and HGT (T) events are depicted at the top of the figure. Minimal costs for the six experiments were 194, 220, 240, 325, 59 and 249, respectively.

6

Beyond Phylogenetic Trees: Network-based Methods

Tree structure seems to be the most natural way to show the evolutionary history of species. However, while they are sufficient in many cases, evolutionary relationships cannot always be represented by a tree-like structure. In case of reticulation events such as recombination and hybridization or horizontal gene transfers, additional branching and new types of nodes are needed to show new and more complex relationships. These needs are met by phylogenetic networks, which are increasingly being used in phylogenetic studies. Another advantage of phylogenetic networks lies in the ability of showing many possible evolutionary paths. When examining very closely related sequences originating from microorganisms, single cells or animal breeds, the bootstrap support values for inferred trees are typically very low due to the difficulty in determining which sequences are most closely related. Therefore, there are many equally probable topologies representing their evolution and there is no method to select the correct one. All possible paths, however, can be seen in the network, which makes the analysis of the data more complete and insightful.

A phylogenetic network is a graph, which like phylogenetic tree, represents evolutionary relationships between genes, proteins, genomes or species, but with the addition of hybrid nodes, i.e, nodes with two parents. We can divide networks into

rooted and unrooted and distinguish several types, such as median, consensus, recombination or hybridization networks. There are also other types of networks besides the well-defined types, and we can call a phylogenetic network any graph that represents evolutionary relationships [Huson *et al.* (2010); Gusfield (2014); Linder *et al.* (2004)].

In this Chapter, we show the application of phylogenetic networks to visualization and analysis of the evolution of B-cells in follicular lymphoma patients. We describe the issue, and present problems and challenges arising from the nature of the data under study. Finally, we show examples and results obtained so far on the data collected from patients. Datasets and biological knowledge in this project were provided by members of Professor H. Veelken's team at Leiden University Medical Center. Our project is still in progress and early results were published in [van Bergen *et al.* (2019)].

6.1. Key concepts

In the following section, we provide some background information on B cells and their role in the human immune system, describing their structure and the processes involved in their maturation and function. We also describe characteristic of the follicular lymphoma, a tumor that affects B-cells and their functions.

6.1.1. The role and characteristics of B-cells

B-cells, also known as B-lymphocytes, are an important part of human immune system. They are a type of white blood cells that produce antibodies, i.e., antigen-specific immunoglobulins (Ig) directed against invasive pathogens. This role makes B-cells a part of the adaptive humoral immune system [Murphy and Weaver (2016)]. B-cells originate in the bone marrow and differentiate into plasma cells and memory cells in the peripheral lymphoid organs during the course of the immune response. An essential element for B-lymphocyte function is the B-lymphocyte receptor (BCR). During the B-cell development, the BCR genetic sequence undergoes the V(D)J recombination process in which variable (V), joining (J), and in some cases, diversity (D) gene segments are randomly rearranged resulting in the highly diverse receptor sequences [Tonegawa (1983)]. This process guarantees the creation of B-cells with receptors capable of recognizing antigens that the cell has never encountered before, including nearly all pathogens like bacteria, viruses, parasites, worms or even cancer cells. To avoid the risk of the body's own cells being recognized as antigens, B lymphocytes undergo very strict selection during the maturation pro-

cess [LeBien and Tedder (2008); Pelanda and Torres (2012)]. The BCR receptor consist of a signal transduction moiety and an antigen-binding subunit called membrane immunoglobulin, build from two B-cell membrane-bound immunoglobulin heavy chains and two light chains. Activation of B-cells occurs when their BCR binds to an encountered antigen. After the activation, cells proliferate and begin to secrete antibodies, which can neutralize foreign object or mark it for attack by other cells of the immune system.

6.1.2. Follicural lymphoma

Follicular lymphoma (FL) is a cancer that involves indolent B-cells and is characterized by a follicular growth pattern of clonal B-cells that accumulate in germinal centers, i.e., transient structures located in lymph nodes, ileal Peyer's patches, and spleen, where mature B cells can be activated, proliferate, and mutate their antibody genes [Natkunam (2007); Xerri *et al.* (2016)]. Follicular lymphoma has many morphological variants and a broad spectrum of symptoms. Among the most common are swelling of lymph nodes in the neck, armpits, and groin. Spleen and bone marrow can also be affected which leads to low certain blood cells levels. The cancer is usually characterized by a slow progression but is essentially incurable. The exact underlying causes of FL are not yet fully understood but it appears to be related to the accumulation of genetic mutations in B-cell precursors [Fischer *et al.* (2018)]. Recent studies allowed to introduce new treatments that have improved overall survival time for patients. Better understanding of the causes of the disease may lead to further improvements in treating protocols.

6.2. WILLOW protocol

Below, we explain our motivations for using phylogenetic networks and present our approach to model the subclone evolution of follicular lymphoma cells.

6.2.1. First steps and motivations

Our first attempt to observe the evolution of sampled B-cells was to infer a phylogenetic tree. It seemed that since the sample contains cells from a certain time interval, i.e., those closer evolutionarily to the common ancestor of all BCR sequences, and those more divergent from it, the tree-like evolution should be observable. However, since the inferred trees had very low support values, we incorporated in our visualizations edges between nodes representing sequences that differed by only

one position. The results made it readily apparent that the sequences are too similar and the tree structure will show only one of the similarly probable evolutionary paths. To see the complex relationships between studied B-cells, we decided to use networks.

6.2.2. Network inference

In the preprocessing all identical BCR sequences are aggregated into distinct subclones and the primordial sequence (PO) inference is conducted. The PO sequence simulates an ancestor of all sequences in the sample and is inferred as a combination of non-mutated VDJ segments from CDR sequence libraries with the highest homology to the studied sequences.

Each node of the network represents a subclone and is labelled by unique id and the sequence counter, i.e, number of identical sequences found in the sample. The structure of the network is based on the neighborhood of subclones that differ by only one position, which we call neighbours, and on the distance of each subclone from the PO, measured as the number of sequence differences. All neighbours are connected by the edges and all nodes are placed at levels corresponding to their distance from the PO. Depending on whether the mutation was non-synonymous or silent the edges are black or gray, respectively. Additionally, connected groups of subclones are visually grouped into subgraphs to help distinguish groups that split early in evolution.

The color and the size of each node depend on whether the node has children (green) or not (red) and the sequence count, respectively. For each subclone, the predicted protein sequence is checked for the presence of acquired N-linked glycosylation motifs (aNGM), which are protein modifications frequently acquired by mutations of the BCR genes in the FL cells [Koning *et al.* (2019)]. The aNGMs are defined as Asn-X-Ser or Asn-X-Thr where X can be any amino acid except proline.

Since some PCR errors may occur during sequencing, singleton nodes with small counters are removed from the network and nodes with no children and only one parent are removed and their sequence count is added to their parent. Due to the sizes of the networks we decided to limit the number of input sequences to 1000 randomly picked sequences. For better readability, nodes are arranged in the levels using a heuristic algorithm to minimize the number of intersecting edges between levels. Resulting networks are shown in Figures 6.3-6.8.

Inferred networks can be compared by the complexity level. Networks for healthy donor samples are expected to be less complex and more tree-like than FL networks. To asses that, we used a diversity score defined as:

$$D = 2(\frac{e+1}{n} - 1),$$

where e is the number of edges and n is the number of nodes in the network. The more tree-like the network, the closer the value of D is to 0. In addition, we calculate the intraclonal accumulation of mutations as the difference between the most mutated node relative to the PO and the least mutated one.

6.2.3. Further development of WILLOW

During research we added more information to our networks. Initially, networks were inferred exclusively for light or heavy chain sequences. New networks contain combined sequences and the color of the edge between neighbours informs whether the mutation was detected in the heavy (blue) or the light (orange) chain. The thickness of the edge depends on the size of the node, so edges between bigger and possibly more important nodes are thicker than others. Edges between small nodes were removed. In addition to the existing color coding, the nodes also contain pie charts with additional information about the distribution of gene expression. Examples of the new network are presented in Figures 6.1 and 6.2.

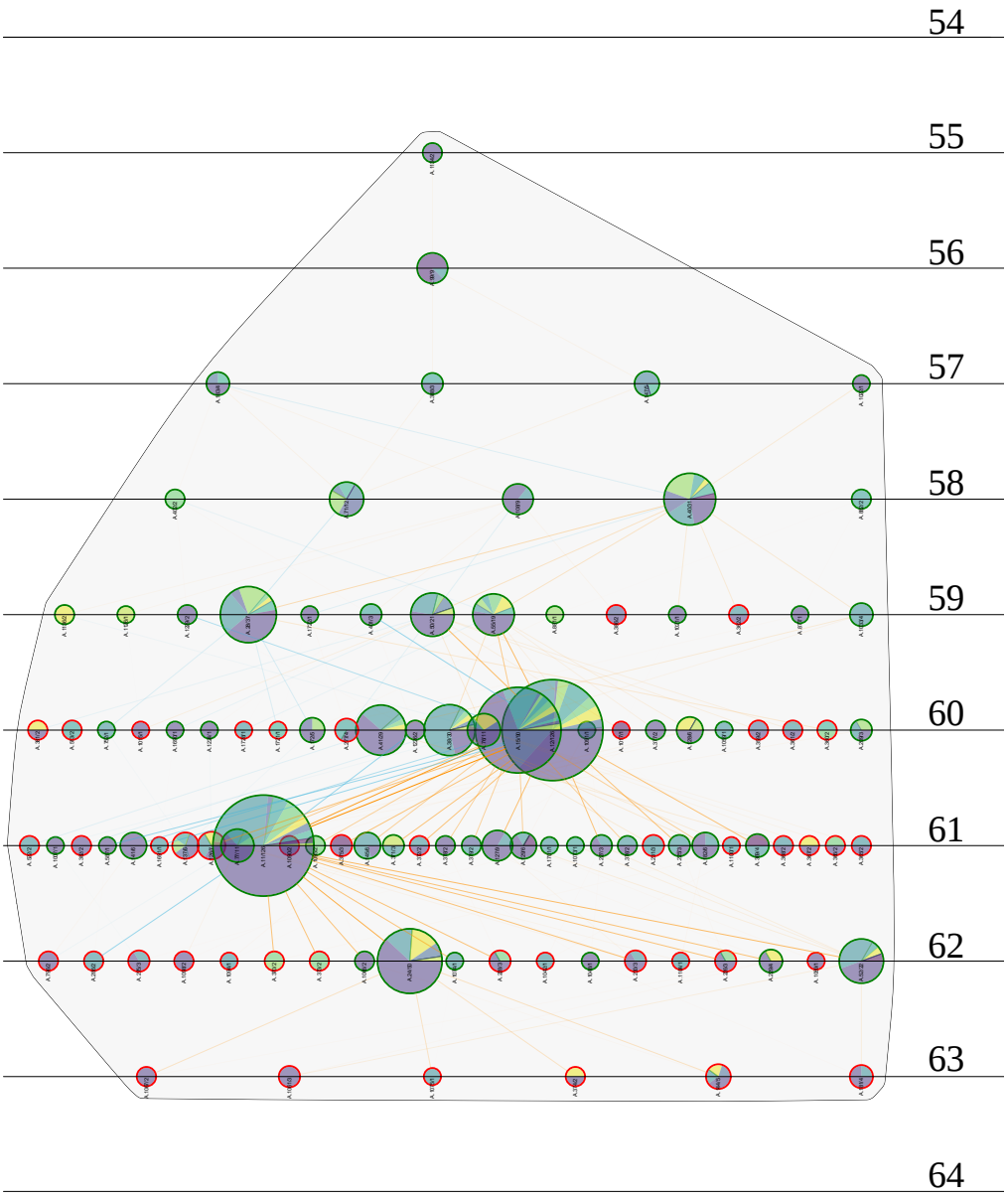


Figure 6.1: One of the subgraphs from the exemplary new WILLOW network. Subclone sequences are the combined heavy and light chains of the BCR variable region. The network shows relations between subclones that differ at only one position in the sequence. The color of the edge depend on whether the mutation was detected in the heavy (blue) or the light (orange) chain. Edges between bigger and more relevant nodes are thicker than other and edges between small nodes are removed. Coloured borders of the nodes indicate that the node is a leaf (red) or an internal node (green). Each node contains a pie chart showing the distribution of gene expression for a given subclone.

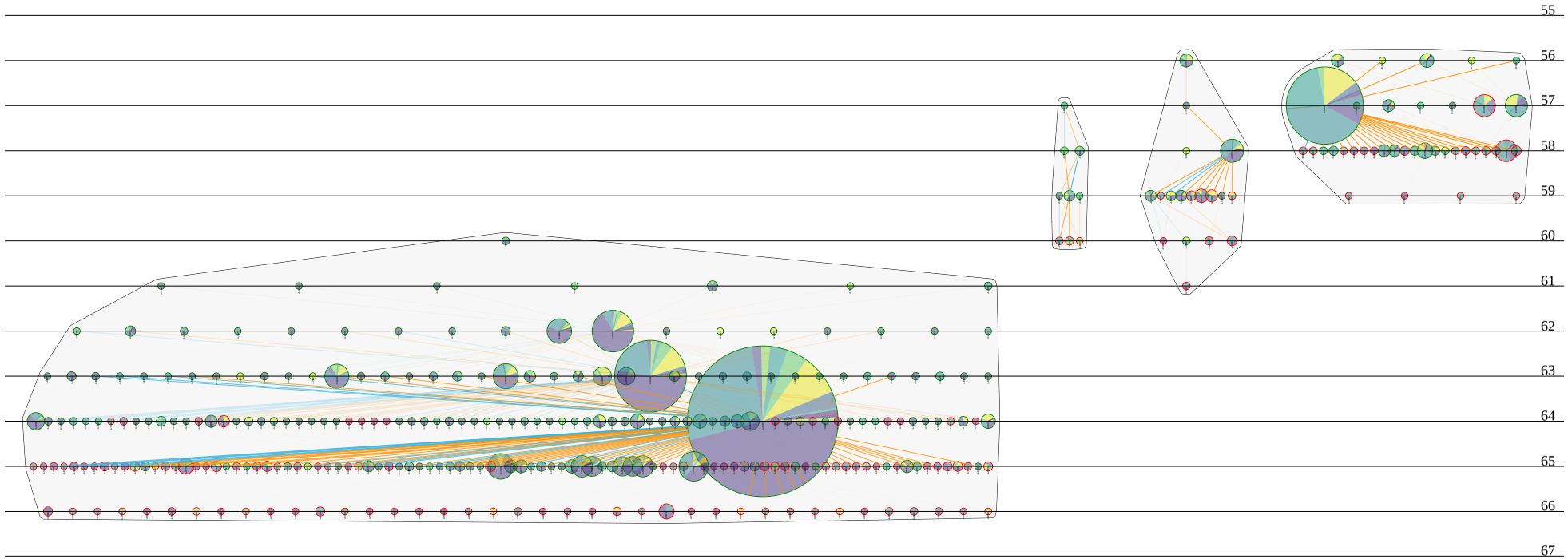


Figure 6.2: Larger part of the network presented in Figure 6.1 showing four subgraphs, i.e., disjointed parts of the network.

6.3. Experiment

Here we present the results of applying the WILLOW protocol to a dataset containing samples obtained from 13 patients by our colleagues from the Leiden University Medical Center. Yielded results were compared with analogous data from healthy donors (HD).

Dataset

Studied samples consisted of 23 biopsies from from diagnostic lymph nodes and bone marrow of 13 patients, with some patients having biopsies taken more than once. Two of the examined biopsies showed transformation to aggressive lymphoma. Complete heavy and light chain BCR variable regions were sequenced by Sanger method or PacBio full-length deep sequencing. The sequencing yielded a median library size of 744 sequences (range: 63-12782) per sample. Within each library, identical sequences were compiled into unique subclones resulting in a median of 200 unique subclones per BCR chain type and case (range: 15-3301). To assess the extent of subclonal diversification in FL, Shannon's diversity index was calculated. As expected, larger sequencing library sizes generally yielded higher Shannon scores. From 2000 sequences per network and upwards, score remained stable, suggesting near-complete sequencing of these networks.

Results

As expected, we observed the presence of aNGM in a very high percentage ($> 77\%$) of the subclones in the majority of samples. We also found no instances of losses of aNGM, which is consistent with the theory, that aNGM are important for FL survival. The sequences were also checked for the presence of stop codons, whose appearance was found in a very small percentage of cells. Networks inferred for FL data were more complex than HD networks resulting in higher D scores of average 1.8 and 0.5, respectively.

In case of the samples from different time points, the networks showed progressively higher number of mutations with time, indicating continuing recombination process. An additional observation is the fact, that samples from bone marrow had lower diversity than the samples from the lymph node and there was no significant progression. This may indicate a higher frequency of mutations occurring in B-cells located in lymph nodes. Selected networks inferred for bone marrow and lymph node B-cells are presented in Figures 6.3-6.8.

Discussion

The results showed that FL networks have a more complex and branching structure in contrast to HD networks which are more tree-like. This may imply the influence of factors other than under normal conditions on the evolution of FL cells. The absence or very low percentage of stop codons in subclone sequences indicates the need for the FL cell to maintain a fully functioning BCR. Networks topologies show, that some BCR mutations are advantageous over others, as some of the subclones has significantly higher sequence counts. They tend appear in connected groups/subgraphs of closely related subclones apparently containing some mutations essential for the survival, while other subclones appear in much smaller groups.

Overall results confirm an important role of the BCR mutations in follicural lymphoma, however, further studies are needed to investigate the factors that may influence mutations.

6.4. Conclusions

Although phylogenetic trees are sufficient for many studies of species or gene evolution, they cannot always be used to represent complex evolutionary relationships. Despite possible extensions like HGT edges, the tree-like structure may be too simple or too strict for datasets consisting of closely related genes or proteins and consequently very similar sequences. In such cases, it is impossible to infer a credible phylogenetic tree as it will only represent one of many convergent evolutionary paths, with no way to determine whether it is the correct one. To show such complex or challenging to interpret relations, we can use phylogenetic networks.

In this Chapter, we showed the application of the phylogenetic networks to cancer data. We analyzed the datasets containing BCR receptor sequences from B-cells, which are a core part of the humoral immune system, and compared networks inferred for healthy donors and follicular lymphoma patients. Our novel approach allows to model tumor evolution and observe subclonal selection driven by BCR mutations. We believe that better understanding of the process will influence treatment methods for FL patients. In future we plan to incorporate information about mutations from outside of the BCR region, to get a more complete picture of the impact of BCR mutations on FL progression.

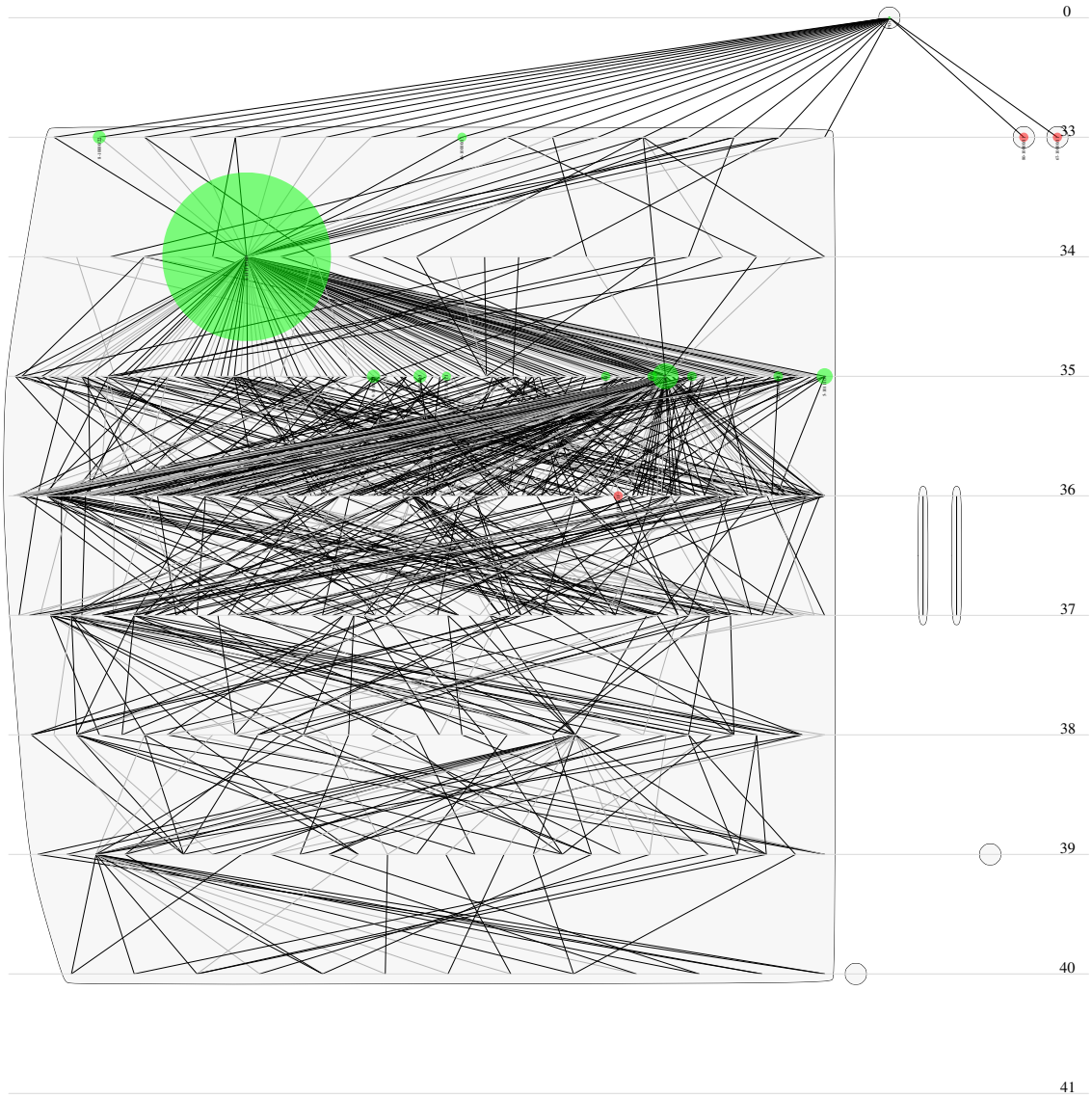


Figure 6.3: Network for heavy chain BCR variable region sequences obtained from bone marrow B-cells of follicular lymphoma patients (February 2012). At the top of the network, at level 0, there is the PO sequence inferred for the sample. Subsequent levels contain subclones whose sequences differ from the PO at the number of positions corresponding to the level number. Nodes visible in the network are from a given sample, while the edges are compiled from all samples to show changes between samples and the relative distribution of nodes. Red colored nodes are leaves and all other nodes are green. The edges are black or gray, respectively, depending on whether the mutation was non-synonymous or silent. Bone marrow samples are from two time points showing changes in the distribution of subclones.

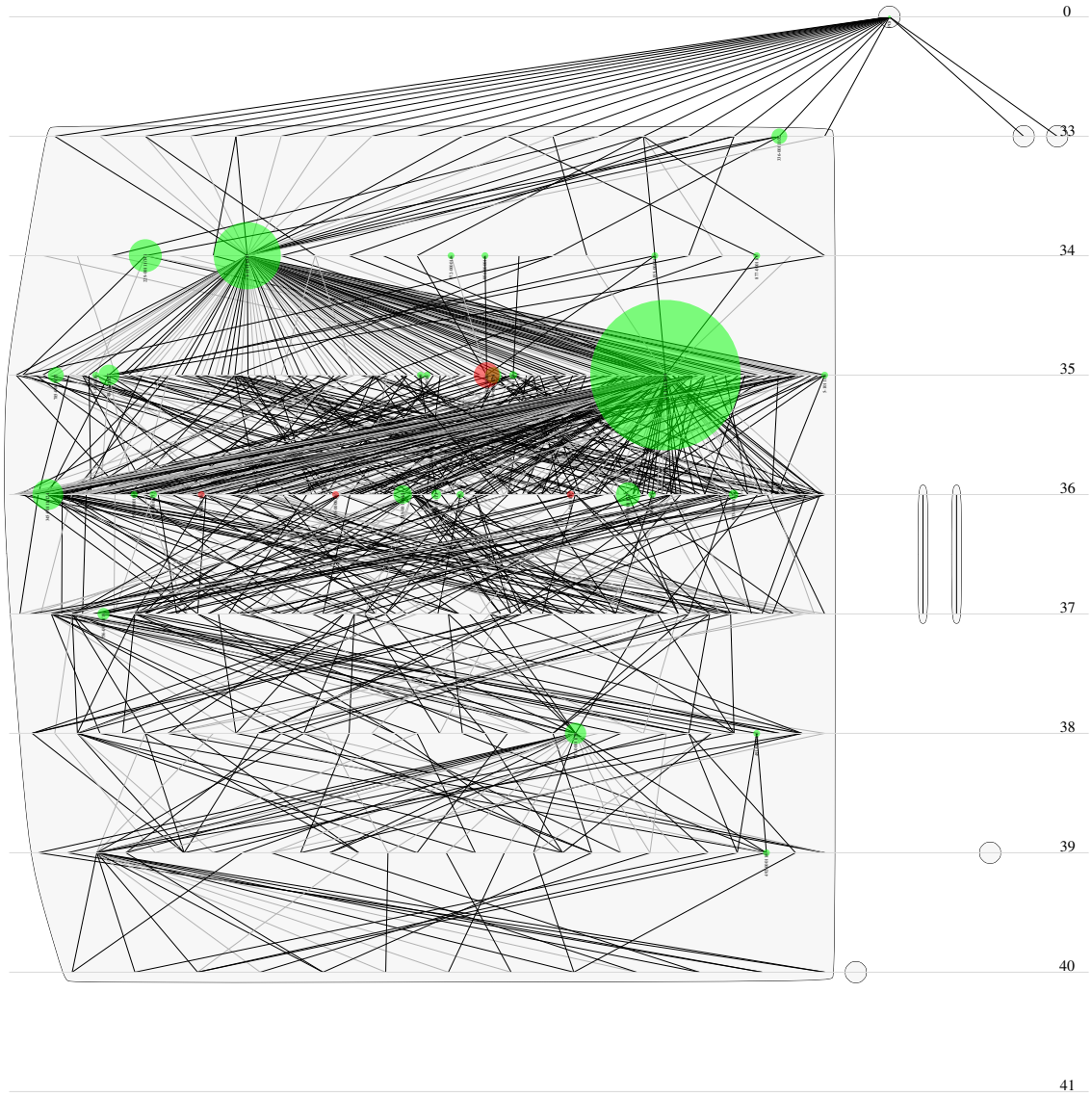


Figure 6.4: Continued from Figure 6.3. Network for heavy chain BCR variable region sequences obtained from bone marrow B-cells of follicular lymphoma patients (January 2015).

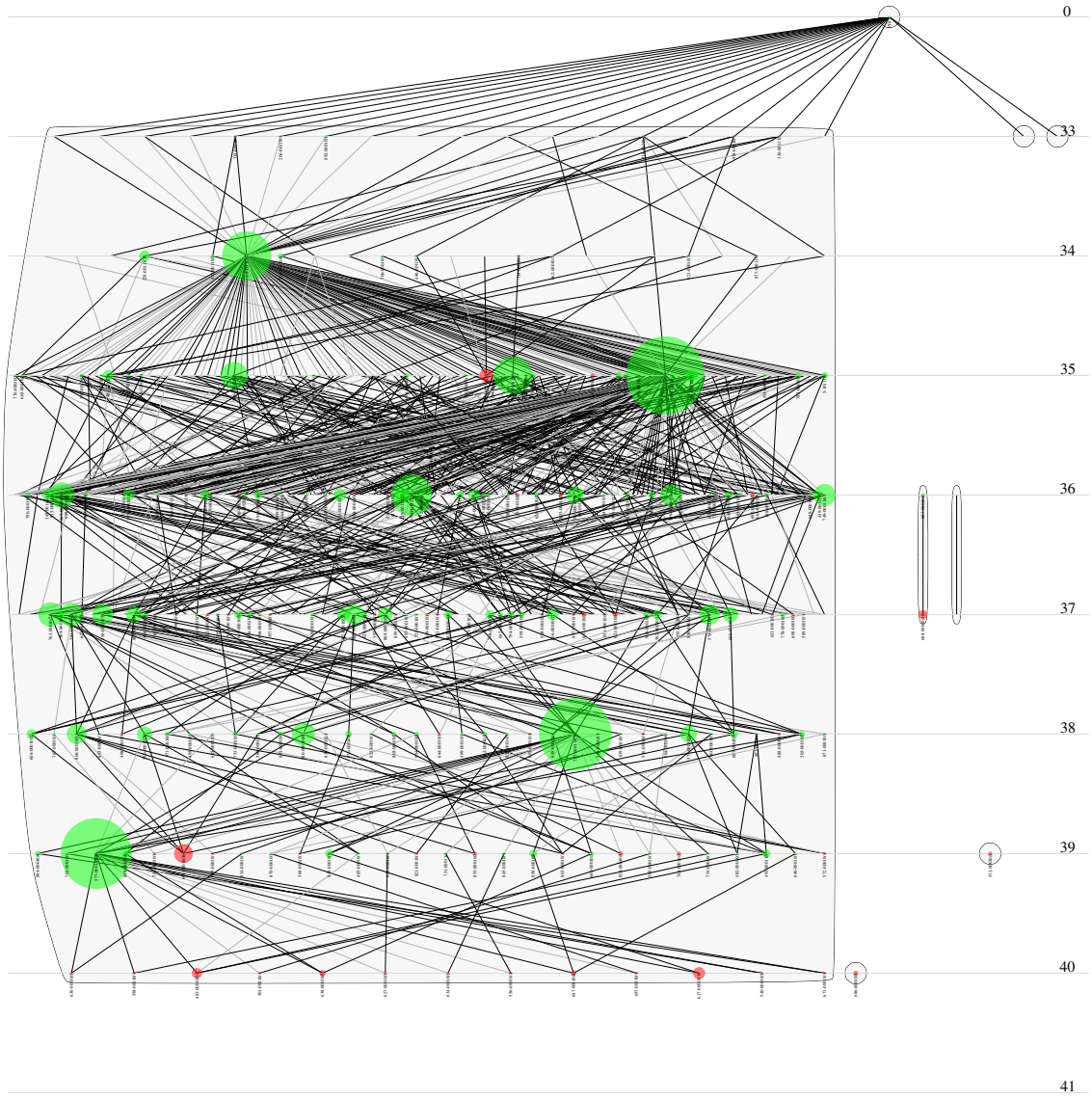


Figure 6.5: Continued from Figure 6.3. Network for heavy chain BCR variable region sequences obtained from lymph node B-cells of follicular lymphoma patients (January 2015).

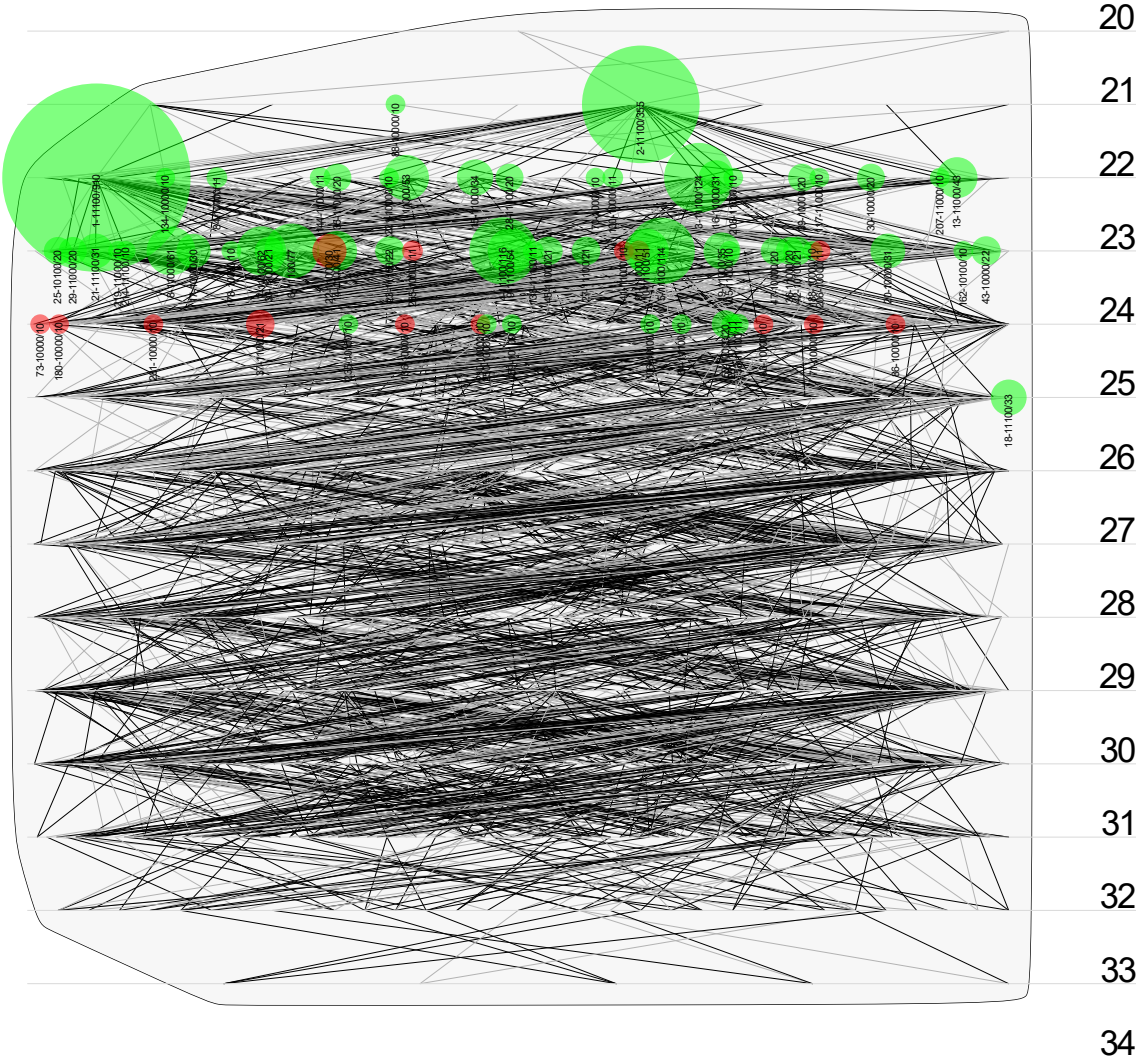


Figure 6.6: Network corresponding to the network depicted in Figure 6.3, inferred for light chain of the BCR variable region. All markings are the same. Due to the size of the sample, we present only their most relevant parts of the networks omitting PO sequence and other subgraphs.

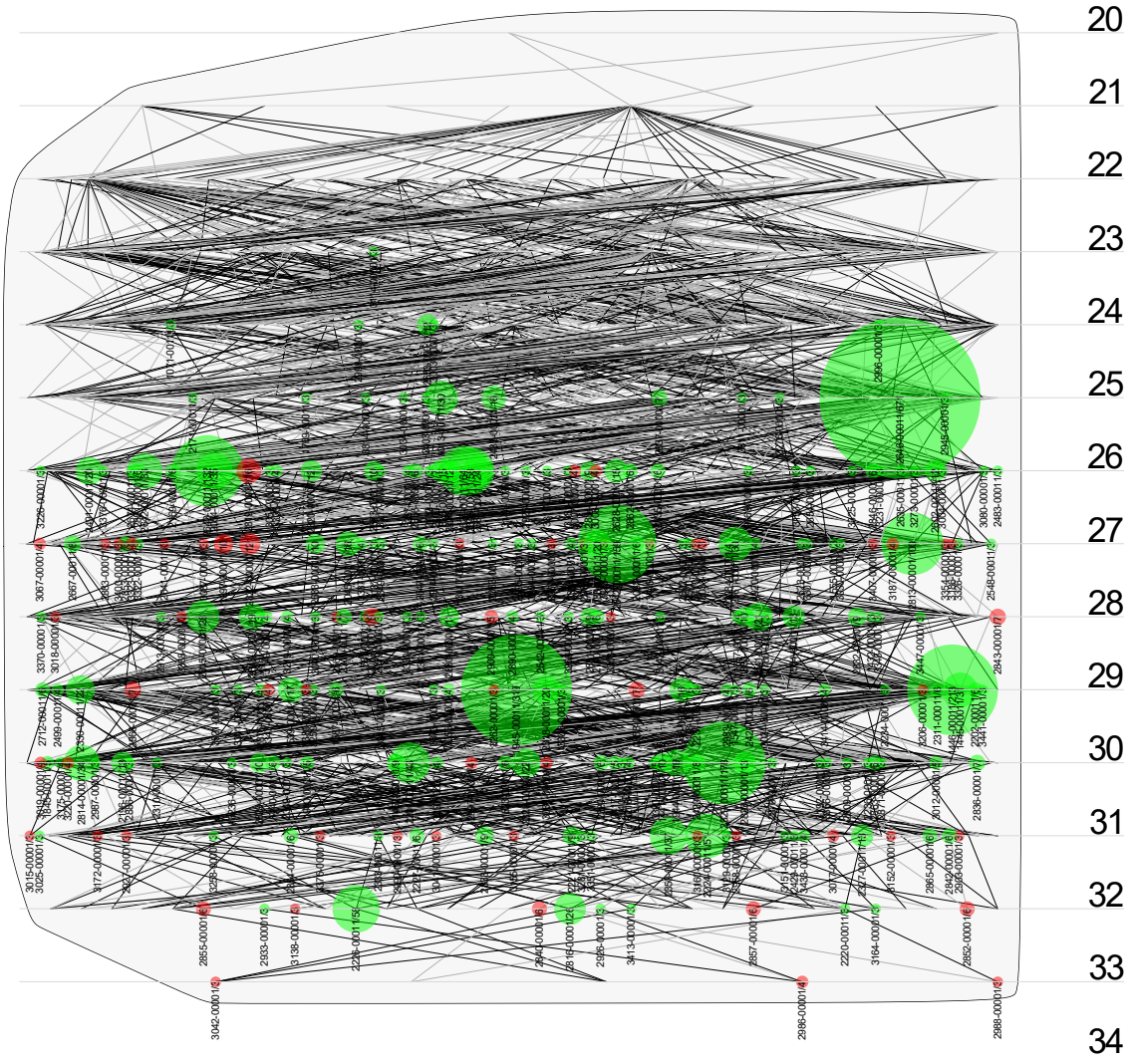


Figure 6.7: Network corresponding to the network depicted in Figure 6.4, inferred for light chain of the BCR variable region.

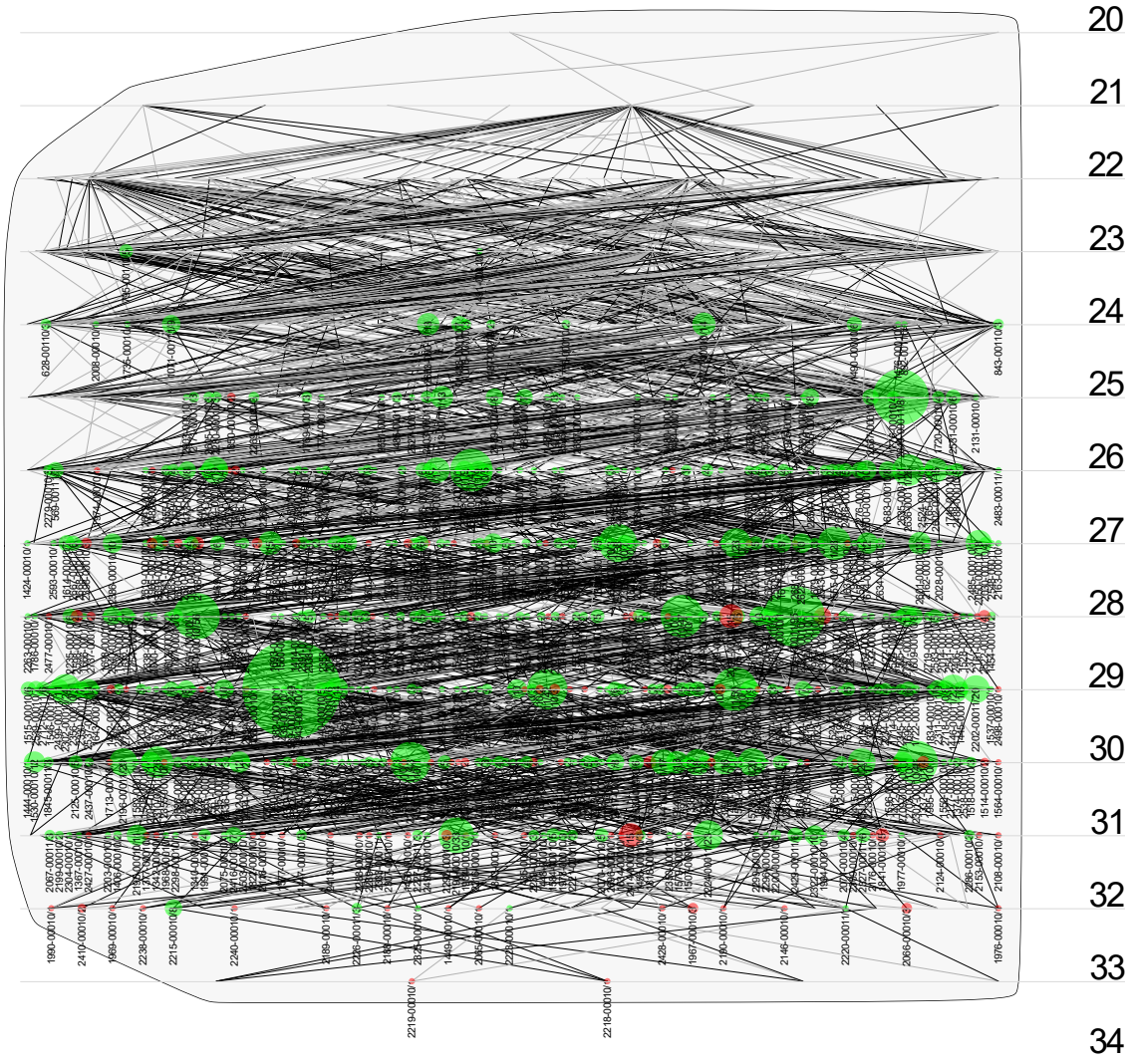


Figure 6.8: Network corresponding to the network depicted in Figure 6.5, inferred for light chain of the BCR variable region.

7 Conclusions

Presented work addresses issues in phylogenetics and graph theory with particular emphasis on finding the most credible evolutionary events in gene trees. We present both theoretical and experimental results including new theorems, lemmas and algorithms along with proofs of correctness, properties of the presented structures, and results showing on real and simulated data the potential use and the performance of our methods and algorithms.

We introduced basic notions from the field of phylogenetics such as concepts of a gene and species tree and the tree reconciliation method based on the lca-mapping and the duplication-loss cost. Then, we showed extensions including unrooted gene trees and models that incorporate horizontal gene transfer events in addition to duplications and losses. Presented tools served as an apparatus for creating structures and methods for finding the most reliable evolutionary events in gene trees.

In Chapter 3, we presented results from our work [Mykowiecka and Górecki (2018)] in which, we proposed a new approach to assess credibility of gene duplication and speciation events in rooted and unrooted gene trees. Our method is based on the unrooted tree reconciliation and non-parametric bootstrap. The concept of the support values for gene duplications and speciations, that we introduced, allows to verify the reliability of tree reconciliation with applications to the rooting and supertree problem. While there are several commonly used rooting methods, our results show, that the majority of them may return incorrectly rooted trees. An-

other possible application of the evolutionary events support values is a supertree problem. By removing gene trees with poorly supported events, the quality of the inferred supertree can be improved. Our approach can also, in very straightforward way, be used to annotate orthology and paralogy in unrooted trees.

We provided several theoretical and algorithmic results, in particular, we showed the correspondence between our method and the classical non-parametric bootstrapping. We also showed that by using gene trees having highly supported events we can infer species trees that are more biologically consistent.

Our approach can be extended to the case when the support is evaluated for subtrees rather than clusters, i.e, not only the presence of the leaves is taken into account, but also their topological arrangement. This modification will allow us to capture more detailed relationships between gene trees, although the bootstrap values modified in this way will be lower than the bootstrap values of the corresponding clusters.

In the next Chapter, we continue to focus on the credibility of evolutionary events in phylogenetic trees, however, we introduce another type of the event. Horizontal gene transfers are a significant factor introducing genetic variability especially in microorganisms. Here, we investigated the problem of the inference of well-supported horizontal gene transfers from multiple sequence alignments. To verify the credibility of inferred transfers, we proposed a new measure based on non-parametric bootstrap, called transfer support. We used this measure to design a new iterative algorithm for inferring acyclic well-supported transfer scenarios.

To test the performance of our method, we conducted experiments on two empirical datasets containing relatively closely and distantly related species groups. Both experiments showed that this approach can be used to support known transfer hypotheses, although, it must be used with awareness of the rooting problem. The accuracy of the algorithm was verified in experiments using simulated data. The results show that Algorithm 2 reached a high percentage of correctly inferred transfers both for trees with one and two HGTs. In particular, for alignments with good quality scores our proposed method can infer the correct HGT scenario with high accuracy.

As in the previous Chapter, the definition of transfer support can be extended. In future we plan to incorporate alternative scoring schemas, e.g., based on the leaves present in clusters rather than on the usage of transfers.

In Chapter 5 we further address the topic of HGT, but we approached the subject from a different angle. We focused on the problem of the gene-species assignment in metagenomic studies, i.e, the situation in which we know the origin of only part

of the genes in the gene tree. The problem was defined as follows: *given a gene tree with partial leaf labelling and a species tree, resolve all missing labels in a gene tree such that the total reconciliation score is minimized*.

To address this problem, we proposed the first HGT-reconciliation based approach to infer gene-species mappings. We proposed efficient algorithms for the optimal cost computation and gene-species assignment inference under weighted cost functions with gene duplication, gene loss and HGT events. Conducted experiments indicated that this approach is able to strengthen the taxonomic assignment of metagenomic sequences. We plan to investigate the influence of event weights and models of reconciliation with possible multifurcations on the quality of inferred assignments. We also plan to apply our method to larger empirical and simulated datasets. Possible extensions worth exploring include methods for analyzing whole metagenomic samples that may contain sequences from multiple gene families.

In previous Chapters, we showed the need to extend phylogenetic trees to include processes such as HGT that occur in nature but do not fit into a simple tree structure. More complex structures are also required when recombination or hybridization events are believed to be involved. Moreover, in case of studying closely related sequences, their high similarity makes it impossible to infer a credible phylogenetic tree. To represent less evident and more complex relationships, we can use a phylogenetic network.

In Chapter 6, we proposed the network-based approach for the cancer data analysis. Studied data contained BCR receptor sequences from B-cells, a part of the humoral immune system, obtained from follicular lymphoma patients. Inferred networks allow to model tumor evolution and observe subclonal selection driven by BCR mutations. Better understanding and knowledge about processes involved in follicular lymphoma progression should improve treatment methods and help discover the causes of the cancer. We plan to further develop our method and extend it by sequences from outside of the BCR region. Additional information may shed new light on FL development and its potential causes.

Bibliography

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. *et al.* (2003). Molecular biology of the cell. *Scandinavian Journal of Rheumatology*, **32** (2), 125–125.

Arvestad, L., Berglund, A., Lagergren, J. and Sennblad, B. (2004). Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *RECOMB*, pp. 326–335.

Bafna, V., Hannenhalli, S., Rice, K. and Vawter, L. (2000). Ligand-Receptor pairing via tree comparison. *J Comput Biol*, **7**, 59–70.

Bansal, M. S., Alm, E. J. and Kellis, M. (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28** (12), i283–i291.

—, Wu, Y.-C., Alm, E. J. and Kellis, M. (2015). Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, **31** (8), 1211–1218.

Bateson, W. *et al.* (1906). The progress of genetic research. In *Report of the Third International Conference 1906 on Genetics*, London, England: Royal Horticultural Society, p. 91.

Bender, M. A. and Farach-Colton, M. (2000). The lca problem revisited. *LATIN*, pp. 88–94.

Beretta, S. and Dondi, R. (2014). Gene tree correction by leaf removal and modification: Tractability and approximability. *LNCS*, **8493**, 42–52.

Berg, J. M., Tymoczko, J. L. and Stryer, L. (2007). *Biochemistry (Loose-Leaf)*. Macmillan.

Betkier, A., Szczęsny, P. and Górecki, P. (2015). Fast algorithms for inferring gene-species associations. *LNCS*, **9096**, 36–47.

Bonizzoni, P., Della Vedova, G. and Dondi, R. (2005). Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical Computer Science*, **347** (1-2), 36–53.

Boykin, L. M., Kubatko, L. S. and Lowrey, T. K. (2010). Comparison of methods for rooting phylogenetic trees: A case study using orcuttieae (poaceae: Chloridoideae). *Molecular phylogenetics and evolution*, **54** (3), 687–700.

Chang, J.-M., Di Tommaso, P. and Notredame, C. (2014). Tcs: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Molecular Biology and Evolution*, **31** (6), 1625–1637.

Chang, W. and Eulenstein, O. (2006). Reconciling gene trees with apparent polytomies. In *Computing and Combinatorics, 12th Annual International Conference, COCOON 2006, Proceedings*, Springer, *Lecture Notes in Computer Science*, vol. 4112, pp. 235–244.

Charleston, M. A. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, **149** (2), 191–223.

Chaudhary, R., Boussau, B., Burleigh, J. G. and Fernández-Baca, D. (2014). Assessing approaches for inferring species trees from multi-copy genes. *Syst Biol*.

—, Burleigh, J. G. and Eulenstein, O. (2012). Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. *BMC Bioinformatics*, **13 Suppl 10**, S11.

Cohen, J. S. and Portugal, F. H. (1974). The search for the chemical structure of dna. *Connecticut Medicine*, **38** (10), 551–2.

Demuth, J. P., Bie, T. D., Stajich, J. E., Cristianini, N. and Hahn, M. W. (2006). The evolution of mammalian gene families. *PloS one*, **1** (1), e85.

Dondi, R., El-Mabrouk, N. and Swenson, K. M. (2014). Gene tree correction for reconciliation and species tree inference: Complexity and algorithms. *Journal of Discrete Algorithms*, **25**, 51–65.

Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllősi, G. J., Ranwez, V. and Berry, V. (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *LNCS*, pp. 93–108.

Drake, J. W., Charlesworth, B., Charlesworth, D. and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, **148** (4), 1667–1686.

Druzhinina, I. S., Chenthamara, K., Zhang, J., Atanasova, L., Yang, D., Miao, Y., Rahimi, M. J., Grujic, M., Cai, F., Pourmehdi, S. *et al.* (2018). Massive lateral transfer of genes encoding plant cell wall-degrading enzymes to the mycoparasitic fungus trichoderma from its plant-associated hosts. *PLoS genetics*, **14** (4), e1007322.

Durand, D., Halldórsson, B. V. and Vernot, B. (2006). A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, **13** (2), 320–335.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *NAR*, **32**, 1792–1797.

Eme, L., Gentekaki, E., Curtis, B., Archibald, J. M. and Roger, A. J. (2017). Lateral gene transfer in the adaptation of the anaerobic parasite blastocystis to the gut. *Current Biology*, **27** (6), 807–820.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *The American Naturalist*, **106** (951), 645–668.

Felsenstein, J. (). PHYLIP. `http://evolution.genetics.washington.edu/phylip.html`.

— (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology*, **27** (4), 401–410.

— (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.

— and Felenstein, J. (2004). *Inferring phylogenies*, vol. 2. Sinauer associates Sunderland, MA.

Fischer, T., Zing, N. P. C., Chiattone, C. S., Federico, M. and Luminari, S. (2018). Transformed follicular lymphoma. *Annals of Hematology*, **97** (1), 17–29.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, **20** (4), 406–416.

Gollery, M. (2005). Bioinformatics: sequence and genome analysis. *Clinical Chemistry*, **51** (11), 2219–2220.

Gontier, N. (2015). Reticulate evolution everywhere. In *Reticulate evolution*, Springer, pp. 1–40.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E. and Matsuda, G. (1979). Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, **28** (2), 132–163.

Górecki, P. (2004a). Reconciliation problems for duplication, loss and horizontal gene transfer. In *Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, RECOMB '04, New York, NY, USA: ACM, pp. 316–325.

Górecki, P. (2004b). Reconciliation problems for duplication, loss and horizontal gene transfer. In *Proceedings of RECOMB 2004*, pp. 316–325.

Górecki, P. (2010). H-trees: a model of evolutionary scenarios with horizontal gene transfer. *Fundam. Inform.*, **103** (1-4), 105–128.

Górecki, P. and Eulenstein, O. (2012a). Algorithms: simultaneous error-correction and rooting for gene tree reconciliation and the gene duplication problem. *BMC Bioinformatics*, **13** (Suppl 10), S14.

— and Eulenstein, O. (2012b). GTP supertrees from unrooted gene trees: linear time algorithms for nni based local searches. *LNCS*, **7292**, 83–105.

—, Eulenstein, O. and Tiuryn, J. (2013). Unrooted tree reconciliation: A unified approach. *IEEE/ACM Trans Comput Biol Bioinform*, **10** (2), 522–536.

— and Tiuryn, J. (2006). DLS-trees: A model of evolutionary scenarios. *Theor Comput Sci*, **359** (1-3), 378–399.

— and Tiuryn, J. (2007a). Inferring phylogeny from whole genomes. *Bioinformatics*, **23** (2), e116–e122.

— and Tiuryn, J. (2007b). Urec: a system for unrooted reconciliation. *Bioinformatics*, **23** (4), 511–512.

— and Tiuryn, J. (2012). Inferring evolutionary scenarios in the duplication, loss and horizontal gene transfer model. *Lecture Notes in Computer Science*, pp. 83–105.

Guindon, S., Delsuc, F., Dufayard, J. and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Methods in Molecular Biology*, **537**, 113–37.

— and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52** (5), 696–704.

Gusfield, D. (2014). *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT press.

Hallet, M. and Lagergren, J. (2001). Efficient algorithms for horizontal gene transfer problems. In *RECOMB*.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, **5** (10), R245–R249.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M. and Ira, G. (2009). Mechanisms of change in gene copy number. *Nature Reviews Genetics*, **10** (8), 551–564.

Henig, R. M. (2000). *The monk in the garden: the lost and found genius of Gregor Mendel, the father of genetics*. Houghton Mifflin Harcourt.

Holland, B., Penny, D. and Hendy, M. (2003). Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. *Syst Biol*, **52**, 229–238.

Huelsenbeck, J. P., Bollback, J. P. and Levine, A. M. (2002). Inferring the Root of a Phylogenetic Tree. *Syst Biol*, **51** (1), 32–43.

Huson, D. H., Rupp, R. and Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press.

Johannsen, W. (1909). *Elemente der exakten Erblichkeitslehre*. Fischer.

Koning, M. T., Quinten, E., Zoutman, W. H., Kiełbasa, S. M., Mei, H., van Bergen, C. A., Jansen, P., Vergroesen, R. D., Willemze, R., Vermeer, M. H. *et al.* (2019). Acquired n-linked glycosylation motifs in b-cell receptors of primary cutaneous b-cell lymphoma and the normal b-cell repertoire. *Journal of Investigative Dermatology*, **139** (10), 2195–2203.

Lafond, M., Chauve, C., Dondi, R. and El-Mabrouk, N. (2014). Polytomy refinement for the correction of dubious duplications in gene trees. *Bioinformatics*, **30** (17), i519–i526.

—, Swenson, K. M. and El-Mabrouk, N. (2012). An optimal reconciliation algorithm for gene trees with polytomies. *LNCS*, **7534**, 106–122.

LeBien, T. W. and Tedder, T. F. (2008). B lymphocytes: how they develop and function. *Blood, The Journal of the American Society of Hematology*, **112** (5), 1570–1580.

Linder, C. R., Moret, B. M., Nakhleh, L. and Warnow, T. (2004). Network (reticulate) evolution: biology, models, and algorithms. In *The Ninth Pacific Symposium on Biocomputing (PSB)*.

Luton, P. E., Wayne, J. M., Sharp, R. J. and Riley, P. W. (2002). The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology*, **148** (11), 3521–3530.

Ma, B., Li, M. and Zhang, L. (2000a). From gene trees to species trees. *SIAM Journal on Computing*, **30** (3), 729–752.

—, — and — (2000b). From gene trees to species trees. *SIAM Journal on Computing*, **30** (3), 729–752.

Maddison, W. (1997). Gene trees in species trees.

Maddison, W. P. (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics – The International Journal of the Willi Hennig Society*, **5** (4), 365–377.

— and Maddison, D. (2015). Mesquite: a modular system for evolutionary analysis.

Mande, S. S., Mohammed, M. H. and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in bioinformatics*, p. bbs054.

Maston, G. A., Evans, S. K. and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.

Morange, M. (2000). *A history of molecular biology*. Harvard University Press.

Murphy, K. and Weaver, C. (2016). *Janeway's immunobiology*. Garland science.

Mykowiecka, A. and Górecki, P. (2018). Credibility of evolutionary events in gene trees (accepted). *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

—, Muszewska, A. and Górecki, P. (2018). Inferring time-consistent and well-supported horizontal gene transfers. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, pp. 79–83.

—, Szczęsny, P. and Górecki, P. (2017). Inferring gene-species assignments in the presence of horizontal gene transfer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **15** (5), 1571–1578.

Natkunam, Y. (2007). The biology of the germinal center. *ASH Education Program Book*, **2007** (1), 210–215.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48** (3), 443–453.

Nguyen, T. H., Ranwez, V., Pointet, S., Chifolleau, A.-M. A., Doyon, J.-P. and Berry, V. (2013). Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms for Molecular Biology*, **8** (1), 12.

Noutahi, E., Semeria, M., Lafond, M., Seguin, J., Boussau, B., Guéguen, L., El-Mabrouk, N. and Tannier, E. (2016). Efficient gene tree correction guided by genome evolution. *PloS one*, **11** (8), e0159559.

O'Meara, B. C. (2010). New heuristic methods for joint species delimitation and species tree inference. *Syst Biol*, **59**, 59–73.

Page, R. (1997). From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/Species Tree Problem. *Mol Phylogenet Evol*, **7** (2), 231–240.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, **43** (1), 58–77.

Park, H. J., Jin, G. and Nakhleh, L. (2010). Bootstrap-based support of hgt inferred by maximum parsimony. *BMC Evol Biol*, **10** (1), 1–11.

Pelanda, R. and Torres, R. M. (2012). Central b-cell tolerance: where selection begins. *Cold Spring Harbor perspectives in biology*, **4** (4), a007146.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, **41** (D1), D590–D596.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, **13** (3), 235–238.

— and Grassly, N. C. (1997). Seq-Gen: An application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences*, **13**, 235–238.

Ranwez, V., Scornavacca, C., Doyon, J.-P. and Berry, V. (2015). Inferring gene duplications, transfers and losses can be done in a discrete framework. *Journal of Mathematical Biology*, pp. 1–34.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res*, **22** (4), 755–765.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J. and Durbin, R. (2008). TreeFam: 2008 Update. *Nucleic Acids Res*, **36**, D735–40.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, **4** (4), 406–425.

Scornavacca, C., Jacox, E. and Szöllősi, G. J. (2014). Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31** (6), 841–848.

—, Mayol, J. C. P. and Cardona, G. (2017). Fast algorithm for the reconciliation of gene trees and lgt networks. *Journal of theoretical biology*, **418**, 129–137.

Sjöstrand, J., Arvestad, L., Lagergren, J. and Sennblad, B. (2013). Genphylodata: realistic simulation of gene family evolution. *BMC bioinformatics*, **14** (1), 209.

—, Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B. and Lagergren, J. (2014). A bayesian method for analyzing lateral gene transfer. *Systematic Biology*, **63** (3), 409–420.

Smith, T. F., Waterman, M. S. *et al.* (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147** (1), 195–197.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, pp. 33–40.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B. and Durand, D. (2012). Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28** (18), i409–i415.

Swenson, K., Doroftei, A. and El-Mabrouk, N. (2012). Gene tree correction for reconciliation and species tree inference. *Algorithm Mol Biol*, **7** (1), 31.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. and Daubin, V. (2013a). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, **62** (6), 901–912.

—, Tannier, E., Lartillot, N. and Daubin, V. (2013b). Lateral gene transfer from the dead. *Systematic biology*, p. syt003.

Taylor, J. S. and Raes, J. (2004). Duplication and divergence: The evolution. *Annual Review of Genetics*, **38**, 615–43.

The Génolevures Consortium (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res*, **37** (suppl 1), D550–D554.

Thompson, C. C., Thompson, F. L., Vandemeulebroecke, K., Hoste, B., Dawyndt, P. and Swings, J. (2004). Use of recA as an alternative phylogenetic marker in the family Vibrionaceae. *Int J Syst Evol Micr*, **54** (3), 919–924.

Tofigh, A., Hallett, M. and Lagergren, J. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **8** (2), 517–535.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, **302** (5909), 575–581.

van Bergen, C. A., Koning, M. T., Quinten, E., Mykowiecka, A., Sepulveda, J., Monajemi, R., De Groen, R. A., Vermaat, J., Kloet, S. L., Kielbasa, S. M. *et al.* (2019). High-throughput bcr sequencing and single-cell transcriptomics reveal distinct transcriptional profiles associated with subclonal evolution of follicular lymphoma. *Blood*, **134**, 298.

Vernot, B., Stolzer, M., Goldman, A. and Durand, D. (2008). Reconciliation with non-binary species trees. *Journal of Computational Biology*, **15** (8), 981–1006.

Watson, J. and Crick, F. (1958). On protein synthesis. In *The Symposia of the Society for Experimental Biology*, vol. 12, pp. 138–163.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, **171** (4356), 737–738.

Weisburg, W. G., Barns, S. M., Pelletier, D. A. and Lane, D. J. (1991). 16s ribosomal dna amplification for phylogenetic study. *J Bacteriol*, **173** (2), 697–703.

Wu, Y., Rasmussen, D., M., Bansal, S., M. and Kellis, M. (2012). Treefix: Statistically informed gene tree error correction using species trees. *Syst Biol*.

Xerri, L., Dirnhofer, S., Quintanilla-Martinez, L., Sander, B., Chan, J. K., Campo, E., Swerdlow, S. H. and Ott, G. (2016). The heterogeneity of follicular lymphomas: from early development to transformation. *Virchows Archiv*, **468** (2), 127–139.

Yu, Y., Warnow, T. and Nakhleh, L. (2011). Algorithms for MDC-based multi-locus phylogeny inference. *Research in Computational Molecular Biology*, pp. 531–545.

Zhang, L. and Cui, Y. (2010). An efficient method for DNA-based species assignment via gene tree and species tree reconciliation. *LNCS*, **6293**, 300–311.

Zheng, Y. and Zhang, L. (2014). Reconciliation with non-binary gene trees revisited. In *Proceedings of RECOMB 2014*, pp. 418–432.

Zmasek, C. and Eddy, S. (2002). Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, **3** (1), 14.