University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Jarosław Paszek
Student no. 209217

Inferring genomic duplication events

PhD's dissertation
in COMPUTER SCIENCE

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

May 2018

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfills 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

Inferring genomic duplication events

One of evolutionary molecular biology fundamental problems is to discover genomic duplication events and their locations in the species tree. Such events can be reconstructed by clustering single gene duplications inferred by reconciling a set of gene trees with a species tree. Existing reconciliation based approaches vary in the two fundamental aspects: (a) the choice of evolutionary scenarios that model allowed locations of duplications in the species tree, and (b) the rules of clustering gene duplications from gene trees into a single multiple duplication event, i.e., **episode clustering (EC)** or **minimum episodes (ME)** methods. There are several models in the literature that specify how gene duplications from gene families can be interpreted as one duplication episode. However, in all duplication episode problems gene trees are rooted. This restriction limits the applicability, since unrooted gene family trees are frequently inferred by phylogenetic methods.

In this dissertation, we propose a model of evolutionary scenarios that preserves the minimal number of gene duplications. We study the RME problem, that is, ME method of clustering when input gene trees are rooted. Our analysis concerns several models of allowed evolutionary scenarios with a focus on interval models in which every gene duplication has an interval consisting of allowed locations in the species tree and fulfills some additional requirements like monotonicity. We present mathematical foundations for general genomic duplication problems. Next, we propose the first linear time and space algorithm for RME jointly for any interval model and the algorithm for the most general model in which every evolutionary scenario is allowed. We also present a comparative study of different models of genomic duplication based on simulated and empirical datasets. We provide algorithms and tools that can be applied to solve efficiently RME problems for various models. Our comparative study helps to identify which model is the most reasonable choice in inferring genomic duplication events.

This dissertation proposes the first solutions to the open problems of UEC (unrooted episode clustering) and UME (unrooted minimum episodes) in which every reconciliation with the minimal number of single gene duplications is allowed and under the assumption that input gene trees are unrooted. In particular, we show new theoretical properties of unrooted reconciliation for the duplication cost and apply them to design several exact and heuristic algorithms for solving the genomic duplication problems. Our comparative study shows that we can improve known results on genomic duplication inference from rooted trees. Moreover, our evaluation on empirical datasets confirms several genomic duplication events from the literature and demonstrates that the proposed algorithms can be successfully applied in practice.

Streszczenie

Studium zdarzeń duplikacji w genomie

Jednym z fundamentalnych zagadnień w molekularnej biologii obliczeniowej jest wykrywanie zdarzeń duplikacji w genomie oraz określenie ich położenia w drzewie gatunków. Rekonstrukcja tych zdarzeń jest możliwa poprzez klastrowanie pojedynczych duplikacji genu, wyznaczonych przez uzgadnianie zbioru drzew genów ze zbiorem gatunków. Istniejące metody różnią się w dwóch zasadniczych kwestiach: (a) wyboru scenariuszy ewolucyjnych, które modelują dopuszczalne lokalizacje duplikacji w drzewie gatunków oraz (b) określenia zasad klastrowania duplikacji genów z drzew genów w jedno zdarzenie wielokrotnej duplikacji, metod jak np. **episode clustering (EC)** lub **minimum episodes (ME)**. Analizując literaturę można wyróżnić kilka modeli opisujących jak duplikacje genów z drzew rodzin genów interpretować jako jedno zdarzenie, jednak wszystkie one dotyczą przypadku, gdy drzewa genów są ukorzenione. Warunek ten ogranicza możliwości zastosowań, gdyż to nieukorzenione drzewa genów są wynikiem popularnych metod filogenetycznych.

W niniejszej rozprawie, proponujemy model scenariuszy ewolucyjnych, który zachowuje minimalną liczbę duplikacji genów. Badamy problem RME, czyli klastrowania metodą ME w przypadku, gdy wejściowe drzewa genów są ukorzenione. Przeanalizowaliśmy kilka modeli dopuszczalnych scenariuszy ewolucyjnych, ze szczególnym uwzględnieniem modeli interwałowych, w których każda duplikacja genu ma przypisany interwał dopuszczalnych lokalizacji w drzewie gatunków, oraz nakładających dodatkowe ograniczenia jak monotoniczność. Przedstawiamy matematyczne podstawy dla ogólnych problemów duplikacji genomowych. Następnie, dla problemu RME proponujemy pierwszy liniowy algorytm uniwersalny dla modeli interwałowych oraz algorytm dla najbardziej ogólnego modelu, w którym każdy scenariusz ewolucyjny jest dopuszczalny. Dodatkowo przedstawiamy studium porównawcze dla różnych modeli duplikacji genomowych, które bazuje na danych symulowanych i empirycznych. Dostarczamy algorytmów i narzędzi do efektywnego rozwiązywania problemów RME dla różnych modeli. Nasze studium porównawcze pozwala na zidentyfikowanie, który model stanowi najrozsądniejszy wybór przy wnioskowaniu zdarzeń duplikacji genomu.

Niniejsza rozprawa przedstawia pierwsze rozwiązania dla otwartych problemów UEC (episode clustering) i UME (minimum episodes), w których każdy scenariusz implikujący minimalną liczbę pojedynczych duplikacji genu jest dopuszczalny oraz przyjmujemy założenie, że wejściowe drzewa genów są nieukorzenione. W szczególności prezentujemy nowe teoretyczne własności nieukorzenionego uzgadniania dla kosztu duplikacyjnego i wykorzystujemy je do zaprojektowania dokładnych i heurystycznych algorytmów rozwiązujących problemy duplikacji genomowych. Nasza ewaluacja eksperymentalna pokazuje, że potrafimy ulepszyć znane wyniki dla wnioskowania duplikacji genomowych z ukorzenionych drzew. Dodatkowo nasza analiza na empirycznych danych potwierdziła kilka zdarzeń duplikacji genomowych z literatury demonstrując, że proponowane algorytmy można z sukcesem wykorzystywać w praktyce.

Keywords

genomic duplication, duplication episode, reconciliation, minimum episodes problem, episode clustering problem, unrooted gene tree, gene tree, species tree

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

Studium zdarzeń duplikacji w genomie

List of publications of major results from the thesis:

Paszek, J. and Górecki, P. (2016). Genomic duplication problems for unrooted gene trees. *BMC Genomics*, 17(1):165–175. doi: 10.1186/s12864-015-2308-4.

Paszek, J. and Górecki, P. (2017a). Efficient algorithms for genomic duplication models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* doi: 10.1109/TCBB.2017.2706679.

Paszek, J. and Górecki, P. (2017b). New algorithms for the genomic duplication problem. *In: Meidanis J., Nakhleh L. (eds) Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science*, 10562:101–115. Springer, Cham.

Paszek, J. and Górecki, P. (2018). Inferring duplication episodes from unrooted gene trees. *BMC Genomics*, 19(5):288. doi: 10.1186/s12864-018-4623-z.

List of selected publications in phylogenetics:

Górecki, P., Paszek, J., and Eulenstein, O. (2017). Unconstrained Diameters for Deep Coalescence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5):1002–1012. doi: 10.1109/TCBB.2016.2520937.

Górecki, P., Markin, A., Mykowiecka, A., Paszek, J., and Eulenstein, O. (2017). Phylogenetic tree reconciliation: Mean values for fixed gene trees. *LNCS*, 10330:234–245.

Górecki, P., Paszek, J., and Mykowiecka, A. (2016). Mean values of gene duplication and loss cost functions. *LNCS*, 9683:189–199.

Górecki, P., Paszek, J., and Eulenstein, O. (2014). Duplication Cost Diameters. *LNCS*, 8492:212–223.

Górecki, P., Paszek, J., and Eulenstein, O. (2014). Unconstrained gene tree diameters for deep coalescence. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, pages 114–121. (ACM SIGBio best paper award in 2014)

Selected other publications:

Gambin, T., Startek, M., Walczak, K., Paszek, J., Grzebelus, D., and Gambin, A. (2013). Tirfinder: A web tool for mining class ii transposons carrying terminal inverted repeats. *Evolutionary Bioinformatics*, 9:17–27.

Huczko, A., Lange, H., Paszek, J., Bystrzejewski, M., Cudziło, S., Gachet, S., Monthioux, M., Zhu, Y. Q., Kroto, H. W., and Walton, D. R. M. (2005). A simple route to new 1d nanostructures. *In: Intelligence in a Small Materials World. Selected Papers from IPMM-2003 The Fourth International Conference on Intelligent Processing and Manufacturing of Materials*, pages 721–729. DEStech Publications, Inc., Lancaster, Pennsylvania, USA.

Acknowledgements:

I would like to express my gratitude to my supervisor Paweł Górecki. I am deeply grateful especially for teaching me to focus on main goal and to not get lost. Paweł, you are the person that I can always rely on and your advice has great value to me. Thank you for your patience in teaching me the precision of expressing ideas, formulas and to present them in a flow. I really enjoyed the search for examples and counterexamples to invent and then test the theories and algorithms. Thank you for all the time, football games, and despite we haven't climbed the Mt.Blanc (yet), the sunrise pictures of Huangshan have special place in my memories.

I would also like to thank the institutions that supported me with funding:

- the National Science Centre for entrusting me with PRELUDIUM grant: NCN # 2015/19/N/ST6/01193
- the Institute of Informatics, University of Warsaw for the scholarships for PhD students

Let me thank for the financial support of the National Science Centre grants NCN # 2011/01/B/ST6/02777 and NCN # 2015/19/B/ST6/00726.

I would like to thank my dear collaborators in, concurrent to my PhD studies, research in phylogenetics: Paweł Górecki, Agnieszka Mykowiecka, Alexey Markin and Oliver Eulenstein.

I would like to express my gratitude foremost to God, and to my parents Gerard and Irena for the gift of life and unconditional love. I would like to thank my sister Dorota, and my family, Michał Szynwelski, and my friends for their invaluable support. I am deeply grateful to Antonina Borzęcka. Tosia, thank you for your love, support and understanding.

Contents

List of Figures

List of Tables

List of Algorithms

CHAPTER 1

General Introduction to Biology

In this Chapter we present the overview on the fascinating branch of science, which is biology. During the years, biology becomes more and more interspersed with human sciences, chemistry, physics and now with computer science. On the other hand, it still elusive and full of exceptions in cautiously defined rules.

Section 1.1 consists of brief description of the evolution of thinking in biology. The first goal is to provide a motivation and show that scientific breakthroughs in that area are deeply connected to other branches of science. Therefore, they may influence the world in the way of perceiving of fundamental ideas, as well as in common everyday life. The second goal is to show that elusiveness. Despite the significant development, some fundamental questions remain still the subject of debate.

Section 1.2 contains selected papers that were crucial to understand the mechanism in genetics, whereas Section 1.3 concludes mentioned research and mainly basing on [Brown, 2002, Brown, 2009] provide biological background and definitions used throughout this work.

The motivation for the choice of the topic of the dissertation is presented in Section 1.4 where we describe selected whole-genome duplication studies. In this dissertation we propose novel solutions and algorithms, and the experimental evaluation of our methods on biological datasets will be compared to results presented in Section 1.4.

1.1. Brief description of the evolution of thinking in biology

Since ancient history of the mankind there are evidences of the curiosity to answer the following questions: What is the purpose of my existence? Why am I here? How the cosmos and its inhabitants were created? What is life? About 9,000-10,000 years ago in the Middle East humans instead of hunting and gathering started to cultivate crops such as wheat and barley [Brown, 2002]. There is a probability that primitive cultures recognized the influence of heredity for domestic animals breeding and early civilization has applied its principles to the improvement of cultivated crops. One of eldest examples may be a Babylonian tablet that is more than 6,000 years old and shows pedigrees of horses and indicates possible inherited characteristics [Winchester, 2018]. In 4th century BCE ancient Greek science reached a climax with polyhistors like Aristotle, who is an author (among many others) of "History of Animals", "Metaphysica", "Nicomachean Ethics" and his successor in the Lyceum and an author of "Inquiry into Plants" and "Growth of Plants" Theophrastus, who also was interested in ethics, metaphysics and more [Rogers et al., 2018].

1.1.1 Early concepts of heredity and evolution that remain actual

Aristotle formulated principles that all organisms are structurally and functionally adapted to their habits and habitats and that Nature is parsimonious [Rogers et al., 2018] which are essential for modern science. The words genus and species are translations of the Greek genos and eidos used by Aristotle [Rogers et al., 2018].

1.1.2 Elucidation of the theory of origin of life

However, many other ideas of crucial importance that were formulated, needed to be extended and elucidated by modern science. Aristotle identified four means of reproduction including the abiogenetic origin of life [Rogers et al., 2018].

On the one hand, spontaneous generation theory explain the origins of life similarly by a hypothetical process in which living organisms develop from nonliving matter [The Editors of Encyclopaedia Britannica, 2018c]. However, it explained such occurrences as the appearance of maggots on decaying meat. Only in 1668 Francesco Redi in Experiments on the Generation of Insects described one of the first biological experiments using sealed and open flasks with different types of meats that produced the evidence against spontaneous generation of maggots [Hawgood, 2003]. In XVIII century there was still a debate. Spallanzani in his experiment boiled chicken broth in sealed flask and showed that cloudiness did not appear, however, he drew off the air in the process and the antagonists stated that he proved only that spontaneous generation need air [VanMeter and Hubert, 2015]. Finally in 1859 Pasteur conducted a refined experiment with beef broth sterilized by boiling in a flask with curved long neck. Only re-exposition to air cause cloudiness which indicates microbial contamination [Ullmann, 2018].

On the other hand, in the 1920s Haldane and Oparin independently presented similar ideas concerning the conditions required for the origin of life on Earth. Both suggested that abiogenic materials in the presence of an external energy source could form organic molecules [Rogers, 2018a]. In 1953 Urey and Miller tested this theory and successfully produced organic molecules from some of the inorganic components thought to have been present on prebiotic Earth [Rogers, 2018a](see Section 1.3.3).

1.1.3 Debate on the origin of changes

Anaximander, who lived in VII and VI century BCE, proposed a theory that is still prevailing. Life arose spontaneously in mud and the first animals to emerge had been fishes, whose descendants eventually left water and moved to dry land and gave rise to other animals. In this early evolutionary theory those changes were described as a transmutation which is a conversion of one form into another [Rogers et al., 2018].

Explanation of the process of heredity, that is how to pass instruction necessary to create new life, was also the subject of considerations. Aristotle believed that the instructions were constant and inherent in gametes, while Hippocrates imagined that instructional particles shaped by experience are gathered by adult body [Lander and Weinberg, 2000].

However, Aristotle rejected any suggestion of **natural selection** [Rogers et al., 2018], a process *that results in the adaptation of an organism to its environment by means of selectively reproducing changes in its genotype, or genetic constitution* [The Editors of Encyclopaedia Britannica, 2018b], which was firstly postulated by reading Alfred Russel Wallace and Charles Darwin work at the Linnean Society on July 1,

1858 [Desmond, 2018]. In 1859 famous Charles Darwin book "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life" was published [Desmond, 2018]. However, Darwin believed in the theory [Desmond, 2018] that organisms acquire traits that allows them to adapt to the environment and such traits can be further inherited, the theory was postulated by Jean-Baptiste Lamarck in his Philosophie zoologique from 1809 [Brown, 2002]. Lamarck was the first who proposed that this basic idea originated from Hippocrates could result in species change [Burkhardt, 2018]. Only August Weismann in 1880s performed an experiment in which he removed tails from mice and observed five generations of progeny to conclude that mutilations were not inherited [Beale, 2018].

Existence of DNA, discovered in 1869 by Johann Friedrich Miescher [Brown, 2002], supports Aristotle concept of constant instructions. Hermann Joseph Muller argued for the gene as the basis of life, and therefore of evolution [Crow, 2005]. His research published in 1927 [Muller, 1927] shown that X-irradiation induces genetic mutations, for which he won the Nobel Prize for Physiology or Medicine in 1946 [The Editors of Encyclopaedia Britannica, 2018a]. Muller noted that new gene functions, and therefore greater complexity, could arise from gene duplication and emphasized mutation as the basic element of evolution [Crow, 2005]. Mutations are random with respect to phenotypic effect, and therefore most are harmful [Crow, 2005]. Random character of mutations in bacteria was confirmed by an experiment described by Luria and Delbrück [Luria and Delbruck, 1943] (who won the 1969 Nobel Prize in Physiology or Medicine in part for this work). They grew cultures of Escherichia coli and then add T1 bacteriophage which kill most of the bacterias except for the mutants. Cultures had different numbers of resistant bacterias and therefore mutations occurred at random [Brown, 2009].

Today after more than 2,300 years, our knowledge is significantly greater, however, answers lead to new discoveries and the debate is still ongoing. The phenomena of Hypermutation and programmed mutations appear to contradict randomness of mutations [Brown, 2009]. Hypermutation occurs when there is an increase of the rate at which mutations occur in cell genome [Brown, 2002]. At V-gene segments of immunoglobulin genes, the repair system changes the nucleotide in the parent strand, and so stabilizes the mutation rather than correcting it [Cascalho et al., 1998]. The authors [Brown, 2009] argue that increased mutation rate do not contradict randomness of the process, however, review [Steele and Lloyd, 2015] from 2015 reports non-random mutation patterns.

There is ongoing debate about program mutations started in 1988 in [Cairns et al., 1988]. Authors proposed that E. coli is able to direct mutations towards genes whose mutation would be advantageous under the environmental conditions that the bacterium is encountering [Brown, 2002]. The original proposal was discarded even by the author [Chicurel, 2001]. In 1998 a model was proposed [Andersson et al., 1998] by which gene amplification during selective growth can give the appearance of adaptive mutability without requiring any change in mutability. A report from 2017 [Bragonzi et al., 2017] suggests that Burkholderia cenocepacia strain increased its capacity to cause a chronic lung infection after serial passages in mice, adapting to the local environmental conditions of murine lung tissues and establishing chronic infection. However, answer to the question how environmental bacteria adapt in the complex and variable environment of the host is still largely unknown.

The cells in a multicellular organism have nominally identical DNA sequences, yet derived from one fertilized egg become specialized, for example as brain cells, or skin

cells. *The lack of identified genetic determinants that fully explain the heritability of complex traits, and the inability to pinpoint causative genetic effects in some complex diseases, suggest possible epigenetic explanations for this missing information. An epigenetic system should be heritable, self-perpetuating, and reversible.* [Riddihough and Zahn, 2010] *It is clear that at least some epigenetic modifications are heritable, passed from parents to offspring in a phenomenon that is generally referred to as epigenetic inheritance. The mechanism by which epigenetic information is inherited is unclear.* [Rogers and Fridovich-Keil, 2018] Epigenetic inheritance might be conceptually closer to Hippocrates idea.

A good summary of the debate may be the conclusion from Koonin and Wolf paper [Koonin and Wolf, 2009] that *Both Darwinian and Lamarckian modalities of evolution appear to be important, and reflect different aspects of the interaction between populations and the environment.*

1.2. Selected milestones in genetics

In 1865 Gregor Mendel describes for the first time principles of heredity based on his breeding experiments with peas [Mendel, 1866]. Johann Friedrich Miescher discovered DNA in 1869 in the extract from human white blood cells (that also contained proteins) and obtained pure DNA sample from salmon sperm the following year [Brown, 2002]. The behaviour of chromosomes in the cell nucleus during normal cell division (mitosis) was observed in salamander larvae cells [Lander and Weinberg, 2000] and described by Walther Flemming in 1882 [Flemming, 1882]. From 1885 to 1901 Albrecht Kossel with his students chemically analyzed the nucleic acids using hydrolysis and other techniques, and in result discovered their component compounds: adenine, cytosine, guanine, thymine, and uracil [The Editors of Encyclopædia Britannica, 2018a]. For his contributions to understanding the chemistry of nucleic acids and proteins he was awarded the Nobel Prize for Physiology or Medicine in 1910. Walter S. Sutton described in 1903 that heredity of genes in a cell corresponds to chromosome behavior during mitosis [Sutton, 1903]. Observations that genes are material beings, placed on chromosomes, linearly, on well defined positions were published in 1915 [Morgan, 1915]. Thomas Hunt Morgan as a result of this research was awarded the Nobel Prize for Physiology or Medicine in 1933 for the discovery of "hereditary transmission mechanisms in Drosophila" [Allen, 2018]. Only in 1928 Frederick Griffith observed that extract prepared from virulent, disease-causing Streptococcus pneumoniae bacteria killed by heat treatment, when added to avirulent strain cause conversion of living benign bacteria into deadly ones [Griffith, 1928, Brown, 2002]. This process was called transformation and in 1944 Oswald Avery, Colin McLeod and Maclyn McCarty [Avery et al., 1944] revealed that DNA is the factor responsible for this conversion [Lander and Weinberg, 2000].

However, the explanation how heredity instructions are stored and encoded remain an open question [Lander and Weinberg, 2000]. Erwin Schrödinger in 1944 in his book *What is Life?* [Schrödinger, 1944] proposed that genes should be constructed from small number of isomeric elements whose sequence form a code for heredity instructions in similar way as Morse code is based on [Lander and Weinberg, 2000].

Meanwhile, in 1941 George W. Beadle and Edward L. Tatum proposed that genes affect heredity by determining enzyme structure [Beadle and Tatum, 1941]. That discovery lead to the 1958 Nobel Prize for Physiology or Medicine award [The

Editors of Encyclopædia Britannica, 2018f]. Also in 1941 A.J.P. Martin and R.L.M. Synge whilst working on the separation of amino acids published a paper [Martin and Synge, 1941] that led to development of partition chromatography and later a paper chromatography. Martin and Synge were awarded the Nobel Prize for Chemistry in 1952.

In 1946 [Lederberg and Tatum, 1946] paper was shown, that two different strains of a bacterium when mixed, result in genetic recombination between them and thus a new, crossbred strain of the bacterium is created. Joshua Lederberg was awarded 1958 Nobel Prize for Physiology or Medicine for discovering the mechanisms of genetic recombination in bacteria [The Editors of Encyclopædia Britannica, 2018g].

Barbara McClintock not only confirmed Morgan's ideas about the role played by the chromosome in heredity [Ravindran, 2012] by showing that genetic recombination involved the physical exchange of chromosome segments in 1931 [Creighton and McClintock, 1931], but also in 1950 paper [McClintock, 1950] she described that some genes are mobile and can move from one position to another in a chromosome. The concept of transposition did not fit the current understanding and it took decades to be widely recognized, but finally Barbara McClintock was awarded the Nobel Prize in Physiology or Medicine in 1983 for her discovery [Ravindran, 2012].

Due to Erwin Chargaff experiments, especially those in 1949/50 [Vischer et al., 1949, Chargaff et al., 1950, Zamenhof et al., 1952], following observations were made: (1) between cytosine and guanine and between adenine and thymine there is one to one ratio, (2) total amount of bases as well as other ratios vary among species. Alfred D. Hershey, who was awarded in 1969 the Nobel Prize in Physiology or Medicine for discoveries concerning the replication mechanism and the genetic structure of viruses, and Martha Chase in their experiment in 1952 observed that when bacteriophages infect bacteria cell then DNA is the main component that enter the cell [Hershey and Chase, 1952]. In result, scientists were aware of the fact that DNA might be the genetic material and was therefore worth studying [Brown, 2002].

In 1953 in Nature three consecutive articles introduced DNA double-helix structure [Watson and Crick, 1953, Wilkins et al., 1953, Franklin and Gosling, 1953]. Rosalind Franklin X-ray diffraction analysis provided the experimental data in support of the double helix [Brown, 2002] and Raymond Gosling has taken the famous *photograph 51* [No authors listed, 2013]. The 1962 Nobel Prize for Physiology or Medicine for the determination of the molecular structure of DNA was awarded to Francis Crick with James Watson and Maurice Wilkins [The Editors of Encyclopædia Britannica, 2018e]. Matthew Meselson and Franklin Stahl in 1958 [Meselson and Stahl, 1958] used isotope of nitrogen to label DNA in order to distinguish newly synthesized DNA from the parental polynucleotides Escherichia coli. They showed that **DNA replication**, which is a synthesis of a new copy of the genome, in living cells follows the **semiconservative** scheme, in which each daughter double helix is made up of one polynucleotide from the parent and one newly synthesized polynucleotide [Brown, 2002].

Vincent du Vigneaud in 1954 reported [du Vigneaud et al., 1954] of the first synthesis of a polypeptide hormone and already in 1955 he was awarded the Nobel Prize for Chemistry. Polynucleotide phosphorylase (PNPase) which is an enzyme that degrades RNA was discovered by Severo Ochoa in 1955 [Grunberg-Manago et al., 1955] in conditions that enabled to use that enzyme to synthesize RNA [The Editors of Encyclopædia Britannica, 2018h]. In 1956 Arthur Kornberg presented evidence of an enzyme-catalyzed polymerization reaction [Bessman et al., 1956] and

isolated the first DNA polymerizing enzyme (DNA polymerase I). Work of Severo Ochoa and Arthur Kornberg won them the Nobel Prize for Physiology or Medicine in 1959 [The Editors of Encyclopædia Britannica, 2018b].

In 1958 William H. Stein and Stanford Moore helped in developing the first automatic amino-acid analyzer [Spackman et al., 1958] and reporting the complete sequence of Ribonuclease A [Smyth et al., 1963] and polypeptide chain structure of Deoxyribonuclease I [Salnikow et al., 1973]. Christian B. Anfinsen results from 1961 paper [Anfinsen et al., 1961] postulated that amino-acid sequence is sufficient for protein to adopt its final conformation. He described that ribonuclease after denaturation could be refolded and still preserve enzyme activity. In 1972 Anfinsen, Stein and Moore were awarded the Nobel Prize for Chemistry.

The first operon, which is a group of genes that are located adjacent to one another in the genome, was described by François Jacob and Jacques Monod in 1961 [Jacob and Monod, 1961]. They were awarded the Nobel Prize for Physiology or Medicine in 1965 and their work is the foundation of our understanding of regulatory control over transcription initiation in bacteria [Brown, 2002]. Moreover, in [Jacob and Monod, 1961] Jacob and Monod proposed the existence of a messenger ribonucleic acid (mRNA) which sequence is translated into a polypeptide when protein is synthesized. Three nucleotides in such a sequence form a codon and all codons except for nonsense ones ultimately are responsible for the incorporation of a specific amino acid into a cell protein. The secret code of life, which was highlighted by Schrödingers book, was revealed in Marshall Nirenberg work mainly in the article [Nirenberg and Leder, 1964] from 1964. Robert William Holley, Har Gobind Khorana and Marshall Warren Nirenberg for interpretation of the genetic code received the Nobel Prize for Physiology or Medicine in 1968.

Hamilton Othanel Smith in 1970 [Smith and Wilcox, 1970, Kelly and Smith, 1970] described first type II restriction enzyme, an enzyme that not only recognizes a specific region in a DNA sequence, but also always cuts the DNA at that very site. Werner Arber, Daniel Nathans and Hamilton O. Smith for the discovery of restriction enzymes and their application were awarded the Nobel Prize for Physiology or Medicine in 1978.

In 1974 [Kornberg, 1974] Roger D. Kornberg (Nobel Prize in Chemistry in 2006) described nucleosome thus explaining how DNA is packed in the cell.

The fact that fragments of DNA that code proteins are interrupted by parts that do not contain genetic information was reported independently in 1977 [Berget et al., 1977, Chow et al., 1977] by two teams. Richard J. Roberts and Phillip A. Sharp for that discovery received the 1993 Nobel Prize for Physiology or Medicine. The coding segments are called exons and the noncoding ones - introns. In 1977 Sanger's group used the method of DNA sequencing to deduce most of the first complete genome to be sequenced, bacteriophage ΦX174. Frederic Sanger for his determination of base sequences in nucleic acids won a Nobel Prize in Chemistry 1980 (he was also awarded in 1958 for determining the structure of insulin molecule) [Jeffers, 2018].

Edward B. Lewis in 1978 [Lewis, 1978] showed that genes order on the chromosome is generally the same as their corresponding body segments but genetic regulatory functions may overlap [The Editors of Encyclopædia Britannica, 2018d]. Eric F. Wieschaus and Christiane Nüsslein-Volhard divided genes responsible for Drosophila melanogaster embryonic development into three categories (see [Nusslein-Volhard and Wieschaus, 1980] published in 1980): gap genes, a scheme for general body development, pair-rule genes, which determine body segmentation, and segment-polarity

genes, which organize repeating structures within each segment [The Editors of Encyclopædia Britannica, 2018c]. Edward B. Lewis, Eric F. Wieschaus and Christiane Nüsslein-Volhard won the Nobel Prize for Physiology or Medicine in 1995.

Stanley B. Prusiner published a paper [Prusiner, 1982] in 1982 about disease-causing proteins called prions and was awarded in 1997 for his discovery the Nobel Prize for Physiology or Medicine. Thomas Robert Cech [Kruger et al., 1982] (in 1982) and Sidney Altman [Guerrier-Takada et al., 1983] (in 1983) independently discredited a belief that enzymatic activity is an exclusive domain of protein molecules, by showing that RNA is also capable of triggering and acceleration of chemical reactions in living cells. Cech and Altman share the 1989 Nobel Prize for Chemistry.

Tonegawa Susumu answer the question how antibody diversity is generated, that is, how from a limited number of genes obtain much more antibodies. According to Tonegawa's research [Tonegawa, 1983], published in 1983 and awarded the Nobel Prize for Physiology or Medicine in 1987, antibodies are constructed from gene fragments that are rearranged randomly to generate different antibody molecules [The Editors of Encyclopædia Britannica, 2018i]. The same year genetic markers on chromosome 4 for Huntington's disease were discovered [Gusella et al., 1983] and thus enabling scientists having the ability to screen people for a disease without being able to cure it.

In 1984 Michael W. Young [Bargiello et al., 1984] independently with Jeffrey C. Hall and Michael Rosbash [Reddy et al., 1984] discovered the period gene which protein product is called PER. Further studies lead Young to discovery of another gene responsible for creating protein TIM [Rogers, 2018c]. When TIM bind to PER it enables PER to enter the cell nucleus and to inhibit its own transcription (synthesis of RNA from DNA). Therefore, PER accumulates in the cell nucleus at night, while during the day its levels decline, when the TIM protein degrades in a light-dependent mechanism. Rhythm-regulating genes responding to light and other factors influence circadian rhythm, which is the self-regulating 24-hour biological clock that drives daily behavioral patterns and physiological processes including metabolism and sleep. Young, Rosbash and Hall received the 2017 Nobel Prize for Physiology or Medicine.

Alec Jeffreys work [Jeffreys et al., 1985, Gill et al., 1985] was the foundation for genetic profiling which is well known technique in forensic science. The combination of microsatellite length variants is unique to every human (genetically identical twins are exceptions) and thus can be used as a genetic profile [Brown, 2002].

Telomerase was discovered by Elizabeth H. Blackburn and Carol W. Greider (see work published in 1985 [Greider and Blackburn, 1985]). This enzyme plays a fundamental role in maintaining chromosomes by adding DNA to telomeres which shorten following cell division and are essential determinants of cell life span. The Nobel Prize for Physiology or Medicine in 2009 was granted Elizabeth H. Blackburn, Carol W. Greider and Jack W. Szostak [Rogers, 2018b]. In 1986 H. Robert Horvitz reported the first two genes that trigger self-destruction process in a cell [Ellis and Horvitz, 1986]. For the discovery that tissue and organ development is regulated by genes responsible for *programmed cell death* mechanism he won the Nobel Prize for Physiology or Medicine in 2002. Yoshinori Ohsumi in 1992 paper [Takeshige et al., 1992] was the first to demonstrate **autophagy** in yeasts, which is a mechanism of degradation and recycling proteins in cells. He identified genes essential for this process and for his studies Ohsumi was awarded the 2016 Nobel Prize for Physiology or Medicine.

The problem of revealing unknown (ex. human) gene functions is being dealt

by inactivating the equivalent genes in the mouse. Method called **gene targeting** uses **homologous recombination** (the exchange of genetic material between two strands of DNA) to change an endogenous gene. The difficulty that not only one mutated cell is needed, but a whole mutant mouse, was overcome by creating such mouse in 1989 (see article [Capecchi, 1989] by Mario R. Capecchi). The solution was to use a special type of a mouse cell, an embryonic stem or ES cell [Brown, 2002] which were discovered by Martin J. Evans in 1981 [Evans and Kaufman, 1981]. Martin J. Evans, Mario R. Capecchi and Oliver Smithies won the 2007 Nobel Prize for Physiology or Medicine for developing gene targeting. A new method of gene inactivation, that rather than disrupt the gene itself destroys its mRNA in RNA degradation process called RNA inference, was suggested in 1998 [Fire et al., 1998] and two of the authors Andrew Z. Fire and Craig C. Mello received the Nobel Prize for Physiology or Medicine in 2006 for the discovery.

Human Genome Project launched in 1990 and resulted in first complete genome DNA sequences of: a free living organism - bacteria Haemophilus influenzae in 1995, an eucaryote - yeast Saccharomyces cerevisiae in 1996, a multicellular organism - roundworm Caenorhabditis elegans in 1998, and finally in 2000 first draft of a human genome [Lander and Weinberg, 2000].

Shinya Yamanaka described in 2006 how to generate stem cells from existing cells of the body. Process of reversion of an adult state to a pluripotent state is triggered by inserting specific genes into the nuclei [Takahashi and Yamanaka, 2006]. In result he obtained **induced pluripotent stem (iPS) cells** that regained the capacity to differentiate into any cell type of the body. This discovery, its importance to regenerative medicine, led him to the 2012 Nobel Prize in Physiology or Medicine award.

The possibilities of modern science seems to be unlimited, able to fulfill dreams of generations and the prospects are surely breathtaking. 26.01.2000 at the ceremony of announcing the success of Human Genome Project in sequencing human genome the president of the USA Clinton said: *Today we are learning the language in which God created life. We are gaining ever more awe for the complexity, the beauty and the wonder of God's most divine and sacred gift.* And the director managing the project Francis Collins added: *It is humbling for me, and awe-inspiring to realize that we have caught the first glimpse of our own instruction book, previously known only to God* [Collins, 2006]. Collins in his 2006 book [Collins, 2006] points to theistic evolution, which holds that God used the elegant mechanism of evolution to create all of life, including human beings (and which is a view espoused by Asa Grey or saint Pope John Paul II), as an explanation that reconciles faith and science [Marroquin, 2007]. He tried to offer a model for a dialogue and points out that: *In 1916, researchers asked biologists, physicists and mathematicians whether they believed in God who actively communicates with humankind and to whom one may pray in expectation of receiving the answer. About 40 percent answered in the affirmative. In 1997 [...] the percentage remained very nearly the same* [Collins, 2006]. Seemingly unlimited progress should come with caution and respect. New ethical issues arises particularly regarding genetic discrimination. In example, shall identification of genes responsible for an inherited disease (like Huntington leading to early death) enable the possibility of the denial of medical insurance? [Ruse, 2018]. Only in 1932 Aldous Huxley, a friend of J.B.S. Haldane, inspired by his friend's work [Hughes, 2008] publish a greatly renowned and acclaimed book entitled *Brave New World* [Huxley, 1932] that describes *future of controlled reproduction, genetic engineering, neurotechnology and*

a world socialist state as an alienated hell [Hughes, 2008]. In [Lander and Weinberg, 2000] Eric Lander warns that some may conclude that *the human spirit and human potential are shackled by double-helical chains.* And in the last sentence ends with a warning: *Meeting these challenges, some quite insidious, will require our constant vigilance, lest we lose sight of why we are here, who we are, and what we wish to become.* That are the very questions which were the starting point.

1.3. Introduction to genetics

Life is defined by **genomes** which are entities that contain all of the biological information essential for a creation of a given organism and a preservation of its life.

Genomes are made of **DNA** (deoxyribonucleic acid) in cellular life forms while a few viruses have **RNA** (ribonucleic acid) genomes.[1] DNA and RNA are polymeric molecules which consists of chains of monomeric subunits called **nucleotides**. Each nucleotide consist of a five-carbon sugar to which a phosphate group and nitrogenous base are attached. The chemical link between adjacent monomers is a phosphodiester bond between sugar molecules, which are deoxyribose molecules (in DNA) or ribose molecules (in RNA). Nitrogenous bases in DNA can be divided into two purines (adenine and guanine) and two pyrimidines (cytosine and thymine) whereas in RNA there is uracil instead of thymine. The full chemical name of a nucleotide 2′-deoxyadenosine 5′-triphosphate can be abbreviated into dATP but in the context of DNA sequence we use simply A, and C, G, T for nucleotides with other bases, respectively. Similarly, adenosine 5′-triphosphate with abbreviation ATP in context of RNA sequence is simply denoted by A, and C, G, U for other nucleotides, respectively. Therefore, we can think of both DNA and RNA as words over four letter alphabet.

Similarly to DNA, **proteins** are linear, unbranched polymers, called also polypeptides. Monomers, called **amino acids**, are linked by a peptide bond and their sequence forms primary structure of a protein. Secondary structure describes the conformations taken up by a polypeptide, while tertiary structure results from folding the secondary structural units and sometimes embeds a knot thus providing interesting topological aspect of study. Quaternary structure defines the association of two or more polypeptides.

Genome is a repository of genetic information. One can think of it as the container for the **genes**, which are DNA segments containing biological information and hence coding for an RNA and/or polypeptide molecule. Complex biochemical reactions controlled by coordinated activity of enzymes and other proteins lead to gene expression. The process of copying of a single gene into a RNA molecule is called **transcription**. The first product of gene expression, **transcriptome**, is a set of RNA molecules containing currently needed biological information in the cell, which are called mRNA (messenger RNA). The next step is **proteome**, which is a repertoire of proteins in a cell that define the scope of biochemical reactions for that cell. The chemical substances produced as a result of metabolism form endogenous **metabolome**.

[1] Information in this section is based on [Brown, 2002, Brown, 2009] and publications from Section 1.2.

The RNA in the cell can be divided into two groups. The first group consists of non-coding, or functional, RNA which are: ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), short interfering RNA (siRNA), microRNA (miRNA) and piwi-interacting RNA (piRNA). The second group, coding RNA, consists of mRNA. The process of creating proteins from mRNA is called **translation**. There are 20 amino acids, so the question is, how does a genetic code on 4 letter nucleotide alphabet look like. Minimal number of nucleotides needed to encode all amino acids is 3 ($min_{x\in\mathbb{N}} 4^x > 20$). In fact, a **codon** consists of three nucleotides and is either a punctuation codon that indicate the points within an mRNA sequence where translation should start and finish or encodes an amino acid. There is one termination codon, three initiation codons, and there are one, two, three, four or six codons corresponding to an amino acid. This feature of the genetic code is called degeneracy.

The first idea of an information flow in a cell, named the central dogma of molecular biology, was focused on translation and transcription and was refined by Crick in [Crick, 1970]. However, the discovery of prions start the debate how to interpret processes in which they are involved [Koonin, 2012]. Figure 1.2 aim is to present the complexity of the processes in the cell and is based on many breakthrough researches which lead to at least 6 Nobel prizes.

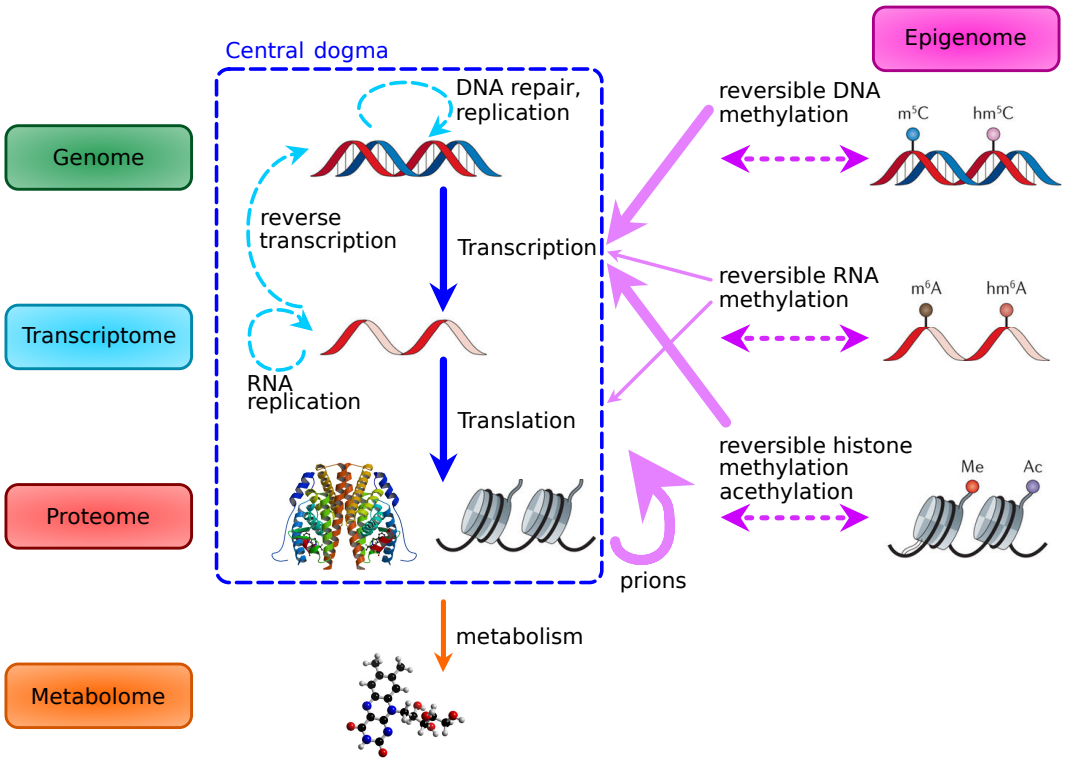

Figure 1.1. The flow of information in a cell. The scheme is based on [Katrib et al., 2016, Fu et al., 2014, Crick, 1970, Koonin, 2012]. Vitamin B_2 as metabolome example from [Harrison, 2017]. Estrogen protein (ESC1) from [Rose and Hildebrand, 2015, Norman et al., 2015].

DNA consists of two strands of nucleotide sequences between which hydrogen bonds are created in two **base-pair** combinations: A with T, and G with C. Those connected strands form double helix structure which in eukaryotes in nucleus is

packed with histone proteins and forms chromatin structures in chromosomes. Eight histone proteins form a barrel-shaped core octamer which is called nucleosome and the DNA wound it twice around the outside. Eukaryotes have at least two chromosomes which during metaphase are in compact form that consist of two chromatids held together by a centromere and their ends we call telomeres. The structure has crucial impact on gene expression which is greater when DNA is between nucleosomes. Moreover, it was thought that centromere were genetically inactive and the discovery that DNA in that region contains genes was a surprise [Brown, 2009]. Histone modifications like methylation play essential roles in many epigenetic phenomena [Cheung and Lau, 2005].

1.3.1 Genes and gene expression

To analyze the gene we can study the nucleotide sequence of an organism. An **open reading frame (ORF)** is a part of a protein-coding gene that is translated into protein and consists of codons starting with an initiation codon and ending with a termination codon. The sequence positioned upstream of a gene, that is a place of binding of RNA polymerase in order to initiate transcription we call a promoter region. An operon is a set of adjacent genes subject to the same regulatory regime and transcribed from a single promoter. The nucleotide sequence to which a repressor protein binds in order to prevent transcription of a gene or operon is called operator. Comparative genomics can be used to identify groups of proteins with functional relationship. One way is to study bacterial operons, while the other focus on protein sequences. For instance, two proteins in *Escherichia coli his2* and *his10* together are similar to *HIS2* protein in yeasts. From the analysis of human proteome researches made a hypothesis that there are about 80,000-100,000 human genes (popular in 1990s see [Pertea and Salzberg, 2010]). However, after genome sequencing the estimate is set to 35,000. Similarly, *Drosophila melanogaster* has only about 13,600 genes while *Caenorhabditis elegans* around 19,000. Most human genes are discontinuous genes that consist of coding regions exons and non-coding regions called introns. Alternative splicing is a process that by joining together different combinations of exons from a single pre-mRNA result in the production of two or more mRNAs. Around 35% of human genes are subject to that process and one of them is *slo*. This gene encode membrane protein that regulates the flow of potassium ions into and out of cells and consists of 35 exons from which only 8 are optional. That gives $8! = 40320$ possible membrane proteins with different properties, while in fact there are over 500 found in human cells. Different hair cells on the basilar membrane of the cochlea in the inner ear respond to different sound frequencies. *Slo* protein properties by partially determining their individual capabilities are responsible for the auditory range of humans. In *Drosophila melanogaster* alternative splicing in gene expression influence the determination of the sex. Males for *slx* gene has all exons, thus do not synthesize *SXL* protein while females do. For *dsx* gene males and females produce different proteins. The number of different immunoglobulin and T-cell receptor proteins, which humans can make approximately 10^8, exceeds by a few levels of magnitude the number of genes. Creation of either immunoglobulin D or immunoglobulin M also depends on alternative splicing but also on genome rearrangements. The genes may be assigned into multigene families, or **gene families** according to the level of similarity of their sequences. Further division is possible. A classical gene family like rRNA genes conserves nearly identical sequence during evolution, whereas a complex family like mammalian globin genes produce proteins

with distinctive biochemical properties and its genes are expressed at different stages in human development. The classification of into families may base either on amino acid sequence of a protein encoded by a gene, or on the nucleotide sequence of DNA of a gene.

1.3.2 Mechanism of DNA sequence modifications

The evolution of DNA sequence over time is a result of structural rearrangements resulting from recombination and transposition and the accumulation of mutations.

Structural rearrangements

First example of major scale DNA sequence modifications is a chromosomal translocation in which the fragment of one chromosome is merging with other chromosome. In human genome such a translocation between 9 and 22 chromosomes is a frequent cause of chronic myelogenous leukemia.

Interesting fact which refer to chromosomes is that cells can do mathematical calculations. Genomic imprinting is a process of inactivating of a gene on one of a pair of homologous chromosomes that is done by methylation. There exist mechanisms that in a female nucleus can count the number of X chromosomes and the number of autosomes and compare those values. If the cell has a diploid set of autosomes and four X chromosomes, in three of them most of the genes will be methylated. However, for the same number of four X chromosomes, if the cell is tetraploid, only two X chromosomes will be inactivated.

DNA replication is the synthesis of a new copy of the genome. Replication errors and the effects of mutagenic agents are dealt with process called DNA repair. The breast and ovarian cancer susceptibility gene BRCA1 product functions in pathway that maintain DNA damage repair [Wu et al., 2010].[2] Homologous recombination which occurs between two double-stranded DNA molecules that share extensive nucleotide sequence similarity. That recombination is responsible for crossing-over during meiosis but currently it is believed that its major role is DNA repair. When two double-stranded DNA sequences share only short regions of nucleotide similarity still site-specific recombination may occur. That include processes of integration and cutting of bacteriophage λ genome which are typical strategies applied in genetic engineering. Recombination is used to create genetically modified crops or knockout mice, which is a mice with inactive gene to discover this gene potential function in human organism. Selectable marker is a gene that enables to distinguish transformed cells during cloning. One of them, frequently used kan^R, encode neomycin phosphotransferase II. It is a selective marker that may lead to antibiotic resistance (like kanamycin). Thus, during a debate researchers argue about the possibility of passing resistance to bacteria in digestive tract. Nowadays, there is a second step in which *Cre* recombinase cut out kan^R gene from the DNA of a crop.

Finally, the last structural DNA modification is done by **transposons**, genetic elements which are able to move in a DNA molecule by copy-and-paste mechanism (replicative transposition) or cut-and-paste process (conservative transposition). Transposition may have adverse effect to the genome like silencing a gene by cutting in its coding region, and influencing expression of neighboring genes when they include promotor sequence. Hemophilia, a disease responsible for a coagulation disorder, may be caused by a transposition of *LINE-1* transposon into a sequence of

[2] The article [Gowen et al., 1998] cited in [Brown, 2002, Brown, 2009] was retracted from Science in 2003 [Gowen et al., 2003].

Factor VIII gene that encode an essential blood-clotting protein. Cells try to minimize those potential adverse effects by methylation, which is a common mechanism of silencing genome areas and makes transposition impossible.

Mutations

A **mutation** is an alteration in the nucleotide sequence of a DNA molecule. Point mutation is a single nucleotide change and can be either a transition when a purine is replaced another purine or a pyrimidine with another pyrimidine or a transversion, otherwise. We distinguish also insertions and deletions of one or more nucleotides.

Most of mutations do not influence the function or the expression of genes and are called silent mutations. In human a mutation in around 98,5% of the genome will cause no effect. However, mutation in coding region of a gene are of utmost importance. Synonymous mutation changes a codon into a second codon that specifies the same amino acid and thus is also a silent mutation. Nucleotide change may result in converting a codon for one amino acid into a codon for another amino acid. and is called missense or **non-synonymous mutation**. A protein with one different amino acid may maintain its functions. Nonsense mutation changes amino acid codon into termination codon, whereas readthrough mutation changes a termination codon into one that specify an amino acid. Both shortening and elongation of proteins may have negative impact on their functions.

Insertion or deletion of a number of nucleotides that is not divided by 3 changes the grouping of nucleotides into codons and is called a frameshift mutation. DNA sequence contains microsatellites comprised of tandem copies of 1-, 2-, 3- or 4-nucleotide repeat units. Replication slippage is an error in DNA replication that leads to an increase or decrease in the number of that units. In result, no two humans alive today, except for identical twins, have exactly the same combination of microsatellite length variants. Therefore, for every person by examining enough microsatellites we can provide an unique pattern called the genetic profile. Nowadays it is used in forensic medicine in crime solving or to determine relationship. Genetic profile is inherited partly from the mother and partly from the father so it can be used to study the populations of humans, animals or even plants. Replication slippage is probably also responsible Huntington's disease and other diseases caused by the expansion of trinucleotide repeats in or near to a gene.

All that changes influence the evolution of genes and result in a loss or rarely in a gain of a function in an organism. An example is the acquisition of streptomycin resistance in *Escherichia coli*. The change in targeted by an inhibitor structure of ribosomal protein *S12* result in fact that antibiotic no longer bind to the protein and interfere with its function.

Drosophila melanogaster wild type has DNA transposon called *P element* in inactive state, while laboratory type do not contain that transposone. Laboratory and wild type when mating produce an offspring with an active *P element*. Transposition activity influence genes causing infertility and other aberrations. In result, those population cannot interbreed and we may state the hypothesis that speciation may occur due to the activity of transposons.

1.3.3 How genes evolve?

The universe started existing about 14×10^9 years ago according to current knowledge [Brown, 2009]. 10×10^9 years ago the galaxies were formed, and then our

solar system was created around $4,6 \times 10^9$ years ago. Finally, life in the form of the cells arose $3,5 \times 10^9$ years ago [Brown, 2009].

The atmosphere of Earth back then was composed of mostly methane, ammonia, water, and carbon dioxide [Brown, 2009, Michaelian, 2011]. Only in 1924 Oparin suggested the material origin of life [Oparin, 1924, Michaelian, 2011, Rogers, 2018a] which inspired famous experiment by Miller and Urey in 1953 [Miller, 1953, Miller and Urey, 1959, Rogers, 2018a]. To simulate primitive ocean and the prebiotic atmosphere they combined warm water with a mixture of four gases: water vapor, methane, ammonia, and molecular hydrogen. Moreover, in order to simulate lightnings [Brown, 2009] they pulsed that atmosphere with electric discharges to produce at least 11 of the 20 then known amino acids [Michaelian, 2011]. The polymerization may happen in a pure geochemical process. The greater challenge is to explain how to obtain an organized system from random biomolecules [Brown, 2009]. One of hypothesis is that early life was based on RNA and its catalytic activity [Bartel and Unrau, 1999, Brown, 2009]. Many such systems may evolve independently, however, the clear analogy of most fundamental biochemical mechanisms in bacterial, archaeal and eucaryotic cells (like the genetic code) imply the existence of common ancestor [Brown, 2009].

The uneventful progress of evolution was interrupted by periods of rapid changes, called explosions, when many new organisms emerged. One of such periods took place around $1,4 \times 10^9$ years ago when Eucaryotes appeared. Eucaryotes have two times more genes than Procaryotes, estimating 10,000 in comparison to 5,000. The second period of drastic increase of organism diversity took place about $0,541 \times 10^9$ years ago in Cambrian explosion, when Vertebrates emerged. Again, basing on today's Vertebrates we can estimate that the number of genes in emerging organisms were three times greater than in antecedent ones [Brown, 2009].

There are two ways of acquiring new genes in a genome: by the duplication of selected or all genes or by obtaining genes from other species [Brown, 2009].

Gene duplication plays crucial role in genome evolution

The history of the hypothesis that gene duplication is a driving force in the evolution is well presented in [Taylor and Raes, 2004]. I would like to present only a few milestones. The beginning we can date back to Darwin questioning sudden leaps in evolution [Friedman, 2009]. The sudden rise and rapid diversification of angiosperm plants he called an *abominable mystery* [Dittmar and Liberles, 2011]. In 1911, Kuwada reported chromosome duplication event in *Zea mays* (maize). He recognized two sets of paralogous chromosomes in karyotype of maize and concluded that maize was an ancient tetraploid [Taylor and Raes, 2004]. Haldane, in 1932, described the possibility that duplication events are favorable because they produce genes that could be altered without disadvantage to the organism [Taylor and Raes, 2004, Taylor and Raes, 1932]. P. Golik in translator's note in [Brown, 2009] also indicated J.B.S. Haldane as the creator of first theories that emphasize the role of duplication in evolution. Haldane was a friend of A. Huxley and his studies inspired his friends' famous book (see Section 1.2). Moreover, he independently to Oparin's work formulated theories of forming the organic molecules from abiogenic materials in the presence of an external energy source [Rogers, 2018a]. In 1933, Haldane proposed an idea of preservation of a duplicate gene through positive selection by the gain of a new beneficial function, which is a process called neofunctionalization [Haldane, 1933, Dittmar and Liberles, 2011].

Susumu Ohno in an author of a book, published in 1970, which is entitled *Evolution by Gene Duplication* [Ohno, 1970]. Ohno suggested that large-scale duplication events are essential for increasing biological complexity and that one or two genome duplications facilitate the evolution of vertebrates [Dittmar and Liberles, 2011]. In his opinion neofunctionalization is crucial for preserving duplicate genes. In the book *Evolution after Gene Duplication* we can read about methods to analyze duplication retention mechanism [Dittmar and Liberles, 2011].

The authors divide genomic duplications into two categories: small-scale duplications (SSDs) and **whole genome duplications (WGDs)** [Dittmar and Liberles, 2011]. SSDs may occur by several mechanisms described in Section 1.3.2 like recombination (unequal crossing-over, unequal sister chromatid exchange, DNA amplification) or replication slippage. All those mechanisms lead to **tandem duplications**, in which duplicated areas are near each other in the genome [Brown, 2009]. Such arrangement can be observed in several gene families, for example, α-globin family on chromosome 16 and β-globin family on chromosome 11 in human genome. SSDs may also arise from duplicative retrotransposition and some mechanisms are yet uncharacterized [Dittmar and Liberles, 2011]. The copy created by transposition do not have promotor sequence, which is absent in mRNA. However, such pseudogene can be inserted near the promotor of existing gene and become active. We call them **retrogenes**. Retrogenes do not have introns that existed in ancestral copy. On the other hand, transcription of antisense RNA may enable the existence of full length functional gene copy inserted in any position in the genome [Brown, 2002].

The first category concerns duplication of short sequences. The duplication of the whole chromosome result in human in a state called trisomy, when cells have three copies of one chromosome and two copies of all of the rest chromosomes. This condition is either lethal, or causes genetic diseases like the Down syndrome in which an extra chromosome 21 is present. However, these pernicious effect do not concern the duplication of the whole set of chromosomes. Whole genome duplications may occur in an error during meiosis. Due to an error gametes may be diploid rather then haploid. In the event that two of them fuse, a new autopolyploid is created, which in this case is tetraploid. Autopolyploidy is not uncommon, especially in plants. New organisms are able to live because every chromosome has a homologous partner. Moreover, they form new species because tetraploid and diploid when interbreed will produce a triploid descendants that cannot breed. The observation that autopoliploidy is the cause of speciation was made for example by Hugo de Vries, one of the rediscoverers of Mendel's experiments. For *Oenothera lamarckiana* a diploid plant he isolated a tetraploid version, which he named *Oenothera gigas* [Brown, 2002]. Although there is an impression that stable polyploidy is maintained more frequently in plants than in animals, recent studies showed that it is not as exceptional as was originally assumed. In fact, evidence of polyploidy was found in all major taxonomic animal groups, especially in fish and amphibians. However, present existence of polyploid mammals is equivocal [Wertheim et al., 2013]. Tetraploidy in *Tympanoctomys barrerae* (red vizcacha rat) is reported in [Gallardo et al., 1999, Gallardo et al., 2006, Brown, 2009], while authors of [Svartman et al., 2005] discard that statement and argue that *polyploidy in mammals remains as unlikely as it has always been.* Finally, in [Suarez-Villota et al., 2012] reason to a hybrid nature of the *Tympanoctomys barrerae* karyotype, which is a result of a hybridization event in the origin of this species.

The second possible way of acquisition of new genes by the genome is an adap-

tation of genes from other species. In plants not only autoploidy is not uncommon but also **allopolyploidy**, which is a fusion between gametes from different species that lead to a polyploid nucleus. *Triticum aestivum*, which is a common crop bread wheat, is a hexaploid that originated from allopolyploidization between tetraploid *Triticum turgidum*, which is a cultivated emmer wheat, and a diploid wild grass, *Aegilops squarrosa* [Brown, 2002]. We can read in [Brown, 2002]: *The wild-grass nucleus contained novel alleles for the high-molecular-weight glutenin genes which, when combined with the glutenin alleles already present in emmer wheat, resulted in the superior properties for breadmaking displayed by the hexaploid wheats.* However, in [Ray, 2015] we can find: *Gluten, the protein that helps provide the 'glue' that binds certain foods also divides us into those who cannot eat it, those who can, and those who choose not to. Awareness and understanding of gluten-related disorders has improved over the past few decades as we began to understand the full spectrum of gluten-induced disease: from sensitivity, to allergy, to autoimmune response.* In fact, our food may be a trigger to a diseases like Coeliac disease, which is a chronic, immune-based enteropathy triggered by gluten [Ray, 2015]. This is briefly mentioned positive and negative impact of allopolyploidization, which can be described as a combination of genome duplication and interspecies gene transfer [Brown, 2002].

Horizontal gene transfer (HGT) is a transfer of a gene from one species to another. Among animals HGTs are possible in result of the activities of retroviruses or transposones, while in bacteria there are several mechanisms for HGT like for example conjugation and transformation [Brown, 2002].

Genome evolution by gene rearrangements

Novel protein functions can be produced not only by duplication followed by mutation that lead to new genes but also by rearranging existing genes. Rearranging the sequences that encode a domain may lead to creation of a protein with novel functions. DNA sequences may be duplicated by unequal crossing-over, replication slippage and other methods that lead to gene duplication. The case when only the gene segment coding for a structural domain is duplicated we call the domain duplication. Domain shuffling occurs when coding sequences for structural domains from completely different genes merge and form a new encoding for a hybrid or mosaic protein [Brown, 2002]. Exons are separate gene fragments, thus are good candidates to code structural domains that can be duplicated and shuffled. On the one hand, $\alpha 2$ Type I collagen gene of vertebrates have repetition of the structural domains that clearly might evolved from exon duplications. Moreover, tissue plasminogen activator (TPA) gene responsible of blood clotting regulatory in vertebrates consist of four exons: one could be derived from fibronectin (responsible for the module that enables the TPA protein to bind to fibrin), one from the epidermal growth factor gene (may enable TPA to stimulate cell proliferation), and two from plasminogen gene (code for structures which TPA uses to bind to fibrin clots) [Brown, 2002]. These are good examples, however, in general duplication and shuffling of domenes is less precise.

Gene family examples

The remarkable example of the role of SSDs or WGDs in gene evolution can be provided by homeotic selector genes that have the key developmental function and

are responsible for specification of the body plans of animals[3]. Eight homeotic selector genes in the genome of *Drosophila* fly form a cluster called HOM-C. The more complex organism in terms of variations of the basic body plan have greater number of Hox clusters. Among chordates the amphioxus has two Hox clusters, and the most vertebrates have four Hox gene clusters, but ray-finned fishes have seven Hox clusters.

In the evolutionary process gene duplication is not always followed by sequence divergence and creation of a family of genes with different functions. When it results in the members of a multigene family retaining the same or similar sequences we call it concerted evolution. The prime examples are the rRNA genes. Their copy numbers varies among species, in *Mycoplasma genitalium* there are two, whereas in *Xenopus laevis* more then 500, but all of the copies preserves almost the same sequence. Special mechanisms facilitates that preservation of functional sequence by preventing individual copies from accumulating mutations.

Another example present the situation where evolution produce two genes with complementary functions in place of one. The trypsin and chymotrypsin are two genes derived from a common ancestor that have complementary protein functions in the vertebrate digestive tract. Both encode proteases that are involved in protein breakdown and are cutting proteins at different amino acids. Trypsin is cutting at arginine and lysine, while chymotrypsin at phenylalanines, tryptophans and tyrosines.

In conclusion, genes may be obtained by the genome in several different processes. Then, they may lost its function and vanish, preserve virtually the same sequence or evolve and obtain new function. Moreover, there are susceptible to gene rearrangements. The picture of a gene family due to all that mechanisms could be convoluted.

Human genome

One of the mysteries concerns the origin of the introns. The self-splicing introns are believed to remain almost unchanged after evolving in the RNA world. There are two hypothesis that concern the GU-AG introns in eukaryotic genomes. *Introns early* theory states that they are ancient and being lost, however, *introns late* idea states that they are recent and are being accumulated [Brown, 2002].

Humans closest relative among the primates is the chimpanzee. The next mystery is to answer the question what is determining the difference between species. We can answer $1,73\%$ of nucleic sequence, less than $1,5\%$ in coding regions and rarely more then 3% is some noncoding regions. About 29% of all genes in human genome encode proteins with identical amino acid sequences to chimpanzee equivalents. Nevertheless, the centromere DNA sequences are significantly different. The greatest dissimilarity is in number of chromosomes. Human chromosome 2 was created probably by merging two chimpanzee chromosomes. Moreover, chromosomes 5, 6, 9 and 12 had undergone a few noticeable rearrangements. However, remaining 18 chromosomes appear to be almost identical. The differences in human genes may result from diet more rich in meat, or resistance to diseases like tuberculosis. The study of genes involved in brain and neural activity revealed that *FOXP2* sequences in human and chimpanzee differs by two amino acids [Brown, 2009]. Mutations in FOXP2 cause developmental verbal dyspraxia [Feuk et al., 2006], therefore, this gene

[3]This part is based on [Brown, 2002]

may decide about human ability to use speech. In conclusion, most or even all crucial differences between human and chimpanzee are not in genome sequence but result from regulatory of gene expression [Brown, 2009].

The recent study shows that in biology one should be cautious to make general assumptions. Even the statement of the double helix structure of the DNA has an exception. In the nuclei of human cells i-motifs structures can be formed, i.e., in regulatory regions of the human genome, including promoters and telomeric regions. Those structures are cell-cycle and pH dependent and could provide key regulatory roles [Zeraati et al., 2018].

1.4. Motivation

In this Section we present brief review of the research to highlight the importance of genomic duplications. In particular, the development of the study of genomic duplications has broad spectrum of potential applications that can lead to an economical and societal impact. First, in Section 1.4.1 we present the examples of general applications of phylogenetics. Then, in Section 1.4.2 we focus on whole-genome duplications. Detection and the determination of time of occurrence of such events is a desired research goal for biologists. Thus, this Section presents our motivation and our justification for the choice of the topic of the dissertation.

1.4.1 The selected applications of Molecular Phylogenetics

To depict the vast spectrum of impact areas let us emphasize that evolutionary trees were used to study the dynamic range of patients' cancer progressions, in order to tailor corresponding treatments [Nik-Zainal et al., 2012] and to predict outbreaks of infectious diseases like meticillin-resistant *Staphylococcus aureus* (MRSA) [Harris et al., 2013].

Another usage of evolutionary trees was supportive to the determination of the origin of AIDS, that was the question, which public opinion was intensely interested to answer. The analysis of the topology of phylogenetic trees shows that the AIDS epidemic may began when virus cross from chimpanzees to humans (SIV in chimpanzees is the closest relative to HIV-1 among primates). Initially, there was a very small number of viruses, perhaps just one, which have spread and diversified since entering the human population [Brown, 2002]. An example of the popularity of the topic, can be the book [Warszewski, 2014] from the cycle of historical battles. Warszewski suggests and argues for the hypothesis that for the spread of the AIDS epidemic might be responsible the Che Guevara partisan unit after their return from Democratic Republic of the Congo. The article from Science [Korber et al., 2000] present the estimated date 1931, to be the date of the last common ancestor of the main group of HIV-1, basing on a comprehensive full-length envelope sequence alignment.

Furthermore, the research in phylogenetics settle the argument between molecular biologist and paleontologists in determining the time of the speciation of humans and chimpanzees. Paleontologists, from studies of fossils, had concluded that speciation event took place some 15 million years ago. The evaluation of hypotheses using gene tree-species tree mismatch probabilities in a likelihood ratio test, favors the

phylogeny with humans and chimpanzees clade [Ruvolo, 1997]. The split of those two lineages is estimated to occur 4,6 – 5 million years ago [Brown, 2002].

Another problem to debate was the question of the place of origin of humans. The paleontological evidence indicates that *Homo erectus* first moved outside of Africa over 1 million years ago. The hypothesis of multiregional evolution was rejected after mitochondrial DNA analysis which revealed that the ancestors of modern humans still lived in Africa 200,000 years ago [Cann et al., 1987]. The individual who carried the ancestral mitochondrial DNA, called *mitochondrial Eve*, could accompanied the individual, called *Y chromosome Adam*, who also lived in Africa some 200 000 years ago according to studies of Y chromosome [Pääbo, 1999]. However, the analysis of β-globin sequences estimates the common ancestor to have lived 800,000 years ago [Harding et al., 1997], whereas the examination of the PDHA1 gene of an X chromosome suggest that time to be 1,900,000 years ago [Harris and Hey, 1999]. Neanderthals are the descendants of *Homo erectus* that lived in Europe and extincted. Current evidence reveals that the human and Neanderthal lineages diverged before the emergence of modern humans [Noonan, 2010]. The analysis of mitochondrial DNA haplotypes in [Richards et al., 2000] shows that farming spread in Europe was done by a small group of 'pioneers' who interbred with the existing pre-farming communities rather than displacing them [Brown, 2002]. There is no evidence for the presence of *Homo erectus* in the Americas. The hypothesis of the origins of humans in Americas is that humans cross the Bering Strait and came from Asia about 22,500 years ago [Brown, 2002]. Recent studies show that western Eurasian genetic signatures in modern-day Native Americans originate not only from post-Columbian admixture, but also from a mixed ancestry of the First Americans [Raghavan et al., 2014]. The mix of ancient population is likely to have occurred after the divergence of Native American ancestors from east Asian ancestors, but before the diversification of Native American populations in the New World [Raghavan et al., 2014]. Moreover, the presence of related population in Siberia during the period of the Last Glacial Maximum suggests that route of migration was through the Bering Strait [Raghavan et al., 2014].

The big question is also to how to provide the food for a rapidly growing population. The development of pesticides and the study of its influence can be tackled with the help of phylogenetics. The authors of [Hammond et al., 2012] report a significant phylogenetic signal that shows the sensitivity to the insecticide and the existence of time lag effects on tadpole mortality. The genomic duplications are crucial for the evolution of crops like bread wheat *Triticum aestivum* (see Section 1.3.3.1). Autopolyploidy and allopolyploidy are a common phenomenon in plants. In next Section we focus on multiple gene duplications.

1.4.2 Biological study of multiple gene duplications

Here, we show the significance of tackling the scientific problem of multiple genomic duplications.

Genomic duplication events play crucial role in evolution of life on Earth. The elemental goal in evolutionary biology is to unravel intricate histories of how the gene families and genomes evolve. The research in phylogenetics focus on the ways of detecting and classifying such events extensively studying various plant, bacterial, fungus and animal genomes [Kellis et al., 2004, Guyot and Keller, 2004, Vision et al., 2000, Costantino et al., 2014, Aury et al., 2006, Cui et al., 2006, Van de Peer et al., 2009].

The phenomenon of whole-genome duplication has impacted the evolutionary history of plants, yeasts and vertebrates. In particular, WGD was common in the history of flowering plants, and therefore, this phenomenon has crucial influence on the evolution of crops [Aury et al., 2006, Van de Peer et al., 2009, Vandepoele et al., 2003, Sato et al., 2012].

The studies of whole-genome duplication focus on detecting its occurrences. This phenomenon occurred before the differentiation of species in cereals [Vandepoele et al., 2003]. Moreover, WGDs affected the evolution of rice, maize [Gaut, 2001] and soya bean [Schmutz et al., 2010]. The cause of similarities between tomato and potato is an event of a whole-genome triplication followed by widespread gene loss that occurred in their common ancestor [Sato et al., 2012].

Important is also to analyze plants that are undesirable in agriculture. The study of knapweed in [Blair et al., 2012] mention that currently in America coexist two incompatible types which are a diploid *Centaurea diffusa* and a tetraploid *Centaurea stoebe subsp. micranthos*. The authors suggest that if diploid *Centaurea stoebe subsp. stoebe* from Europe is introduced to North America, interspecific hybridization has the potential to result in even more aggressive invaders [Blair et al., 2012].

Additionally, to mention very recent practical research, WGD was studied in the context of mechanisms underlying metabolic diversity within plant species and the potential strategies (and barriers) to introgress novel metabolic traits [Scossa et al., 2016]. Furthermore, numerous plant WGDs seem associated with periods of increased environmental stress and/or fluctuations [Vanneste et al., 2014]. In conclusion, the research on phenomenon of whole-genome duplication not only can explain ancient evolution but also recent studies concern influence of WGDs on the quality of crops and the association between WGD and environment. Research on multiple gene duplications has a great potential for further practical applications.

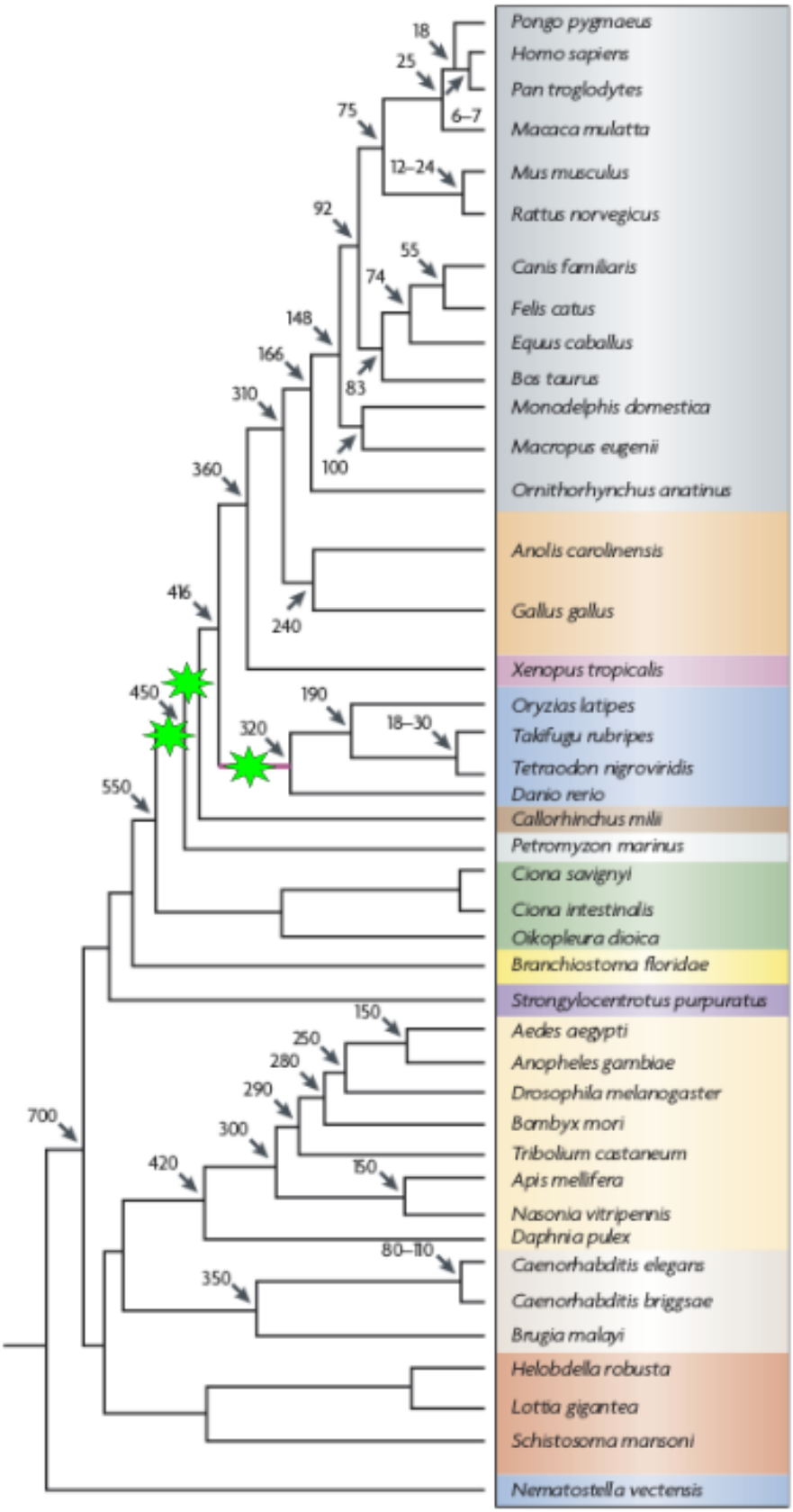

Figure 1.2. The whole-genome duplication events, potential locations denoted by stars, in vertebrates. Figure from [Ponting, 2008].

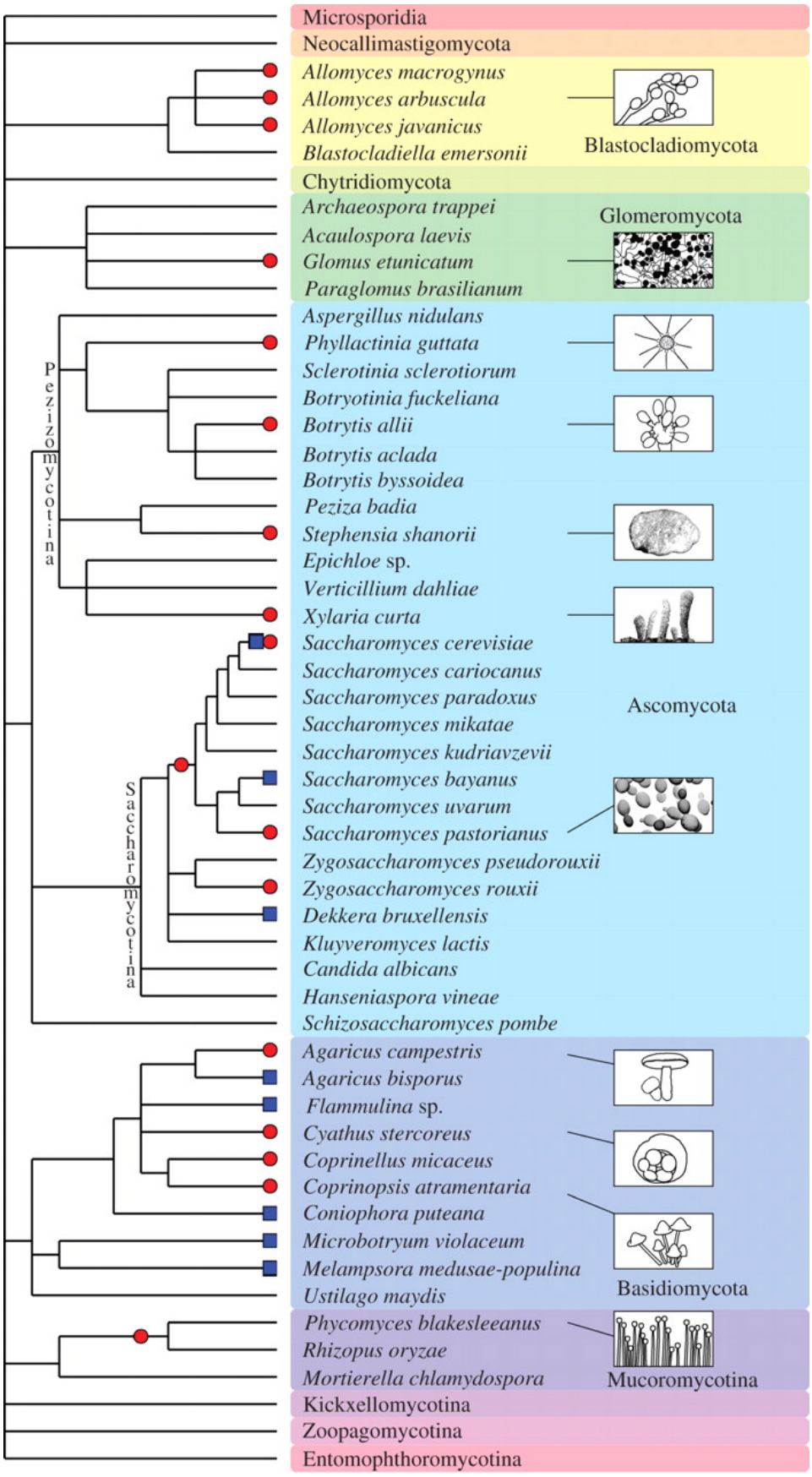

Figure 1.3. The whole-genome duplication events in fungi. Red circles mark locations of suspected polyploidy, blue squares indicate lineages with individuals having hybrid origin. Figure from [Albertin and Marullo, 2012].

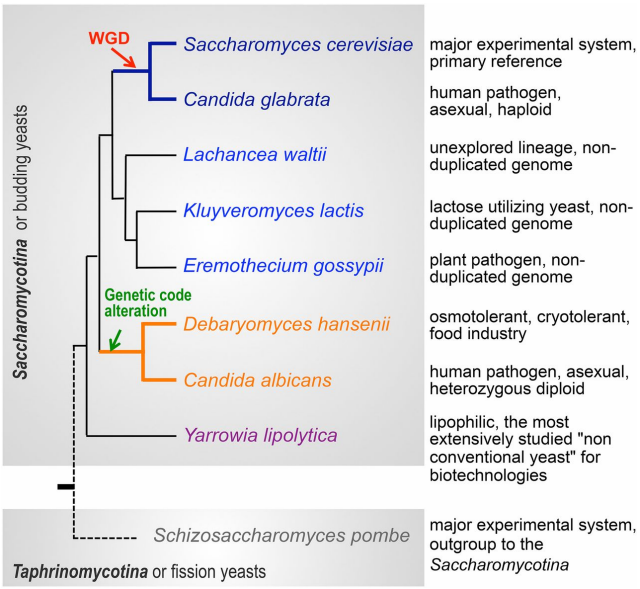

Figure 1.4. The whole-genome duplication event in Saccharomycotina (subset of first fully sequenced genomes in 2004). The event was reported already in [Wolfe and Shields, 1997]. To find branch of *Zygosaccharomyces rouxii*, which parent speciation is an ancestor of WGD event, please refer to Figure 1.3. Figure from [Dujon and Louis, 2017].

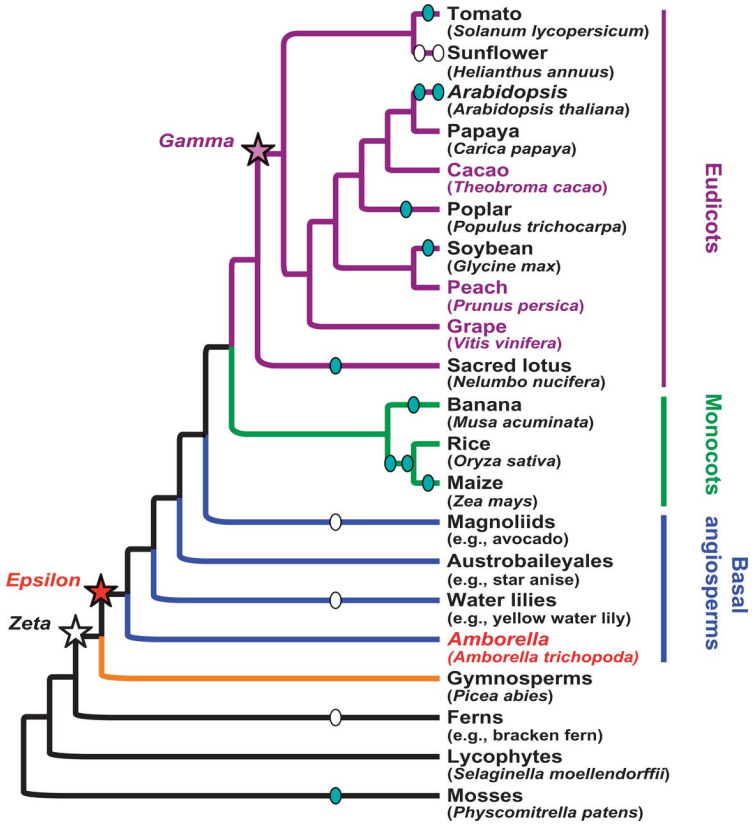

Figure 1.5. The whole-genome duplication events in plants (potential locations denoted by various marks). Figure from [Albert et al., 2013].

CHAPTER 2

Introduction to reconciliation

In this Chapter we start with the explanation and definition of basic concepts. Section 2.1 contains the mathematical description of the relations between species and between genes. Next, there is an introduction to the reconciliation, a process that from biological point of view explains the evolution by an introduction of events such as gene duplications, gene losses or speciations. Throughout this thesis, we do not consider such events as horizontal gene transfers. Section 2.2 presents the concept of an evolutionary scenario that explains the evolution of genes within species. The reconciliation and evolution are illustrated in Sections 2.1 and 2.2 for the case when both gene tree and species tree are rooted. Then, in Section 2.3 we introduce the background definitions for the reconciliation when gene trees are unrooted. In such a case, there is a need of the inference of a rooted gene tree that is derived from unrooted gene tree. The process of selecting a rooted gene tree we call a rooting of unrooted gene tree and the results of this process we call rootings. In Section 2.3, we present the properties of rootings and their connection to unrooted gene tree topology.

2.1. Classical rooted reconciliation

In this thesis we propose to use the model of the reconciliation in which a rooted gene tree is reconciled with its rooted species tree. The work of Goodman [Goodman et al., 1979] introduced the concept of reconciliation that is using *mapping* between trees to explain the differences between a gene and a species tree. That idea was formalized by Page in [Page, 1994, Page and Charleston, 1997a], where potential incongruences between trees were explained by introducing evolutionary events such as **gene duplications**, **gene losses**, and **speciation events**. The events of gene losses are frequent in both prokaryotes and eukaryotes [Koonin and Galperin, 2003, Sebat et al., 2004, Demuth et al., 2006], while duplications are recognized as a driving force of the evolution of eukaryotes [Maere et al., 2005, Lynch and Conery, 2000, Lynch and Conery, 2003, Fischer et al., 2014].

The reconstruction of evolutionary history of individual genes is generally well established. The theoretical properties and practical aspects of tree reconciliation have been extensively studied [Koonin and Galperin, 2003, Bourque and El-Mabrouk, 2006, Nakhleh, 2013], e.g., in the context of the introduction to gene, species tree phylogeny [Goodman et al., 1979, Page and Holmes, 1998, Slowinski and Page, 1999], the introduction to mappings and reconciliation [Page, 1994, Mirkin et al., 1995, Guigó et al., 1996], modeling evolutionary scenarios [Maddison, 1997, Bonizzoni et al.,

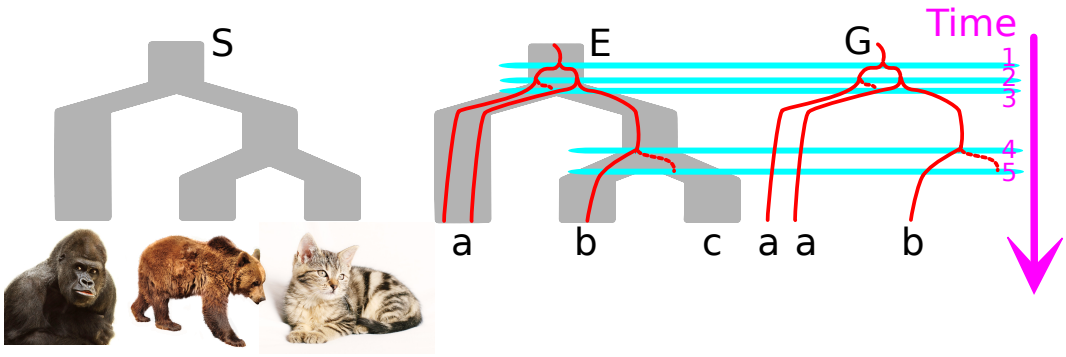

Figure 2.1. An example of hypothetical reconciliation for a species tree S and a rooted gene family tree G. Labels are a for an ape, b for a bear and c for a cat. The gene tree consists of three genes: two from ape and one from bear. The evolutionary time interval in this scenario consist of events: 1-gene duplication, 2-speciation, 3-gene loss, 4-speciation and 5-gene loss. The scenario E is depicted as an embedding of the gene tree G into the species tree S according to the rule that "species are containers for genes" generalized to trees.

2005, Arvestad et al., 2009, Doyon et al., 2009, Górecki et al., 2011, Górecki and Eulenstein, 2014a], application of Bayesian methods [Arvestad et al., 2003, Rasmussen and Kellis, 2011, Sjostrand et al., 2014], probabilistic framework for exploring space of reconciliations [Doyon et al., 2012], gene tree bootstrapping [Felsenstein, 1985, Behzadi and Vingron, 2006, Durand et al., 2006, Mykowiecka and Górecki, 2018], supertree inference [Hallett and Lagergren, 2000, Ma et al., 2000, Page, 2000, Chen et al., 2006, Górecki and Tiuryn, 2007b, Bansal and Shamir, 2011, Burleigh et al., 2011, Górecki and Eulenstein, 2012b, Lafond et al., 2015], error correction in gene trees [Górecki and Eulenstein, 2012a, Wu et al., 2013, Noutahi et al., 2016], non-binary species tree [Stolzer et al., 2012], horizontal gene transfer detection [Hallett and Lagergren, 2001, Górecki, 2010], incomplete lineage sorting influence on detecting duplications [Zhang, 2011, Zheng and Zhang, 2014], maximum parsimony in case when each event has assigned a cost [Wu et al., 2014], NP-hardness of duplication-loss-coalescence model [Bork et al., 2017], gene tree reconstruction [Scornavacca et al., 2014, Dondi et al., 2017], gene order [Holloway et al., 2013, Duchemin et al., 2017], metagenomics [Betkier et al., 2015, Mykowiecka et al., 2017], relations to comparison functions [Górecki et al., 2013], and theoretical results related to the mathematical properties of tree comparison functions in general [Górecki and Eulenstein, 2014b, Górecki and Eulenstein, 2015, Górecki et al., 2014a, Górecki et al., 2014b, Górecki et al., 2016, Górecki et al., 2017a, Górecki et al., 2017b]. It has become clear that molecular evolution cannot only be studied based on the analysis on the nucleotide level.

2.1.1 Gene and species trees - basic definitions

A **species tree** is a rooted binary tree with leaves uniquely labeled by the names of species. Throughout this work, we usually use S to denote a species tree.

A **rooted gene tree** is a rooted binary tree with leaves labeled by the names of species. The set of species present in T is denoted by $\mathcal{L}(T)$.

In this thesis we always assume that $\mathcal{L}(G) \subseteq \mathcal{L}(S)$ for any gene tree G and species tree S.

The rooted tree (T_1, T_2) has two subtrees T_1 and T_2 whose roots are children of the tree root. Additionally, for nodes a and b, we introduce a partial order and we write $a \preceq b$ when a and b are on the same path from the root, with b being closer to the root than a. Notation $a \prec b$ means that $a \preceq b$ and $a \neq b$. In other words if $a \prec b$ then a is said to be a **descendant** of b while b is an **ancestor** of a. The root of a tree T we denote by $\mathsf{root}(T)$. The **height** of a tree is the maximal number of edges on the path from a leaf to the root of the tree.

By $T(v)$ we denote the subtree of T rooted at v. An **cluster** for a node v is the set of all species present in $T(v)$. An **interval** is a path of nodes in T connecting two comparable nodes s and s' such that $s \preceq s'$.

2.1.2 LCA-based reconciliation

Here, we present the notation from graph theory that is applied to describe evolutionary processes.

Let $T = \langle V_T, E_T \rangle$ be a rooted gene tree. **The least common ancestor (lca) mapping**, $\mathsf{M}_T : V_T \to V_S$, is defined as follows. If v is a leaf in T then $\mathsf{M}_T(v)$ is the leaf in S labeled by the label of v. When v is an internal node in T having two children a and b, then $\mathsf{M}_T(v)$ is the least common ancestor of $\mathsf{M}_T(a)$ and $\mathsf{M}_T(b)$ in S. An internal node $g \in V_T$ is called a **duplication** if $\mathsf{M}_T(g) = \mathsf{M}_T(a)$ for a child a of g.

The duplication cost, denoted by $\mathsf{D}(T, S)$, is the total number of duplications in T. Each non-duplication node of T we call a **speciation** (including all the leaves). The total number of **gene losses** required to reconcile T and S can be defined by:

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

where $\|a, b\|$ is the number of edges on the path connecting a and b in S. Finally, we can define the **duplication-loss cost** of reconciling a rooted gene tree T and a species tree S as follows: $\mathsf{DL}(T, S) = \mathsf{D}(T, S) + \mathsf{L}(T, S)$. Examples of the reconciliation are depicted in Figure 2.2.

We denote by $\mathsf{Dup}(T)$, the set of all duplication nodes in T.

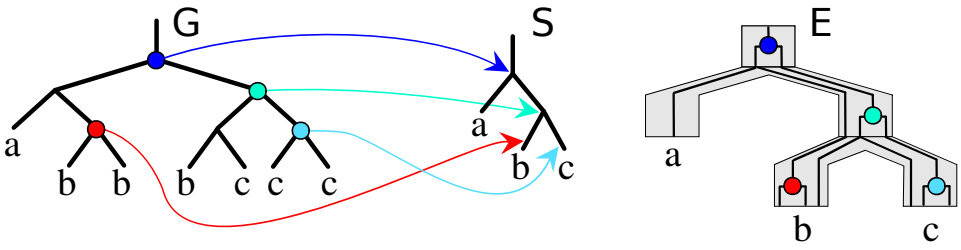

Figure 2.2. An example of reconciliation for a gene tree G and a species tree S with the least common ancestor mapping (LCA-mapping), LCA-mappings, inferred from **leaf labelings** (here depicted with thin arrows), play a crucial role in interpreting macro-evolutionary events located in gene and species trees.

2.2. Evolutionary scenarios

In this thesis, we use the idea of interpreting the reconciliation as the model a biologically consistent scenarios which are embeddings of a gene tree into a species

tree that induce the location of evolutionary events in the species tree [Górecki and Tiuryn, 2006]. Identification of such a scenario is made by a function called the **duplication mapping** that assigns a gene tree node, interpreted as an event called a **single gene duplication**, to a node of a species tree [Guigó et al., 1996, Page and Cotton, 2002, Bansal and Eulenstein, 2008, Burleigh et al., 2008, Mettanant and Fakcharoenphol, 2008, Burleigh et al., 2010, Luo et al., 2011, Paszek and Górecki, 2016, Nøjgaard et al., 2017].

In the classical LCA-based reconciliation (see Section 2.1.2), a **duplication** is a node g of a gene tree such that g and at least one of its children are mapped to the same node in the species tree. The example of the joint evolution of genes and species is depicted on Figure 2.2. The gene duplications are marked, in example, we have a gene duplication at node $((b, c), (c, c))$ in the gene tree G and mapped to the node (b, c) in the species tree S. In other words, the tree reconciliation identified a single gene duplication event including its **location** in the species tree. Reconciliation explains incongruence between a species tree S and a gene tree G (a tree inferred from gene sequences) by using the minimal number of duplication and loss events, called the **duplication-loss cost** (DL). A typical interpretation of reconciliation is an embedding of a gene tree into the species tree as shown in Figure 2.2.

The most fundamental properties of the model for a given rooted gene tree and its corresponding rooted species tree are:

- reconciliation is linear in time and space [Page, 1994, Ma et al., 2000],
- there exists exactly one scenario based on LCA reconciliation (see Section 2.1.2) that minimizes the total number of gene duplication and loss events [Bonizzoni et al., 2005, Górecki and Tiuryn, 2006],
- however, the scenario is non-unique for the duplication cost only [Górecki and Tiuryn, 2006].
- the number of evolutionary scenarios that are compatible with these trees is infinite [Górecki and Tiuryn, 2006].

2.2.1 DLS-trees: a model of evolutionary scenarios

Now we present a description of the model of DLS-trees [Górecki and Tiuryn, 2006] that will be used to represent evolutionary scenarios. A **DLS-tree** is a binary tree having two types of internal nodes, denoting gene **duplications** and **speciations**, and two types of leaves denoting **gene losses** and sampled **gene sequences**. By using the standard nested parenthesis notation, we define DLS-trees (introduced in [Górecki and Tiuryn, 2006], below we use the formula from [Górecki and Eulenstein, 2014a], in form from [Paszek and Górecki, 2017a]):

1. a is a single-noded DLS-tree denoting a **gene sequence** from species a,
2. $A-$ is a single-noded DLS-tree denoting a **lost gene** lineage, where A is a non-empty set of species,
3. $(R_1, R_2)+$ is a DLS-tree whose root is a duplication node and its children are DLS-trees R_1 and R_2 such that $\mathcal{L}(R_1) = \mathcal{L}(R_2)$,
4. $(R_1, R_2)\sim$ is a DLS-tree whose root represents a speciation and its children are DLS-trees R_1 and R_2 such that $\mathcal{L}(R_1) \cap \mathcal{L}(R_2) = \emptyset$.

For instance, $T = ((a,b)\sim, ab-)+$ is a DLS-tree in which one copy of a gene is immediately lost after a duplication event (for simplicity we write $ab-$ instead of $\{a,b\}-$). We say that a DLS-tree is **lost** if its every leaf is lost. A gene tree can be extracted from a non-lost DLS-tree T by removing all lost subtrees from T and suppressing nodes of degree 2. Such operation will be denoted by $\mathtt{gt}(T)$. For example, $\mathtt{gt}((((a,b)\sim, ab-)+, c)\sim) = ((a,b),c)$.

We say that T is **compatible** with a species tree S if every cluster of T is present in S. We write that a DLS-tree T is a **scenario** for a gene tree G and a species tree S if $\mathtt{gt}(T) = G$ and T is compatible with S. In such a case, every node g in G uniquely corresponds to a node in T denoted by $\xi(g)$. Therefore, we can define mappings $\xi\colon G \to T$ and $\phi_T\colon G \to S$, such that $\phi_T(g)$ is the node in S whose cluster equals the cluster of $\xi(g)$ (see example in Figure 2.3).

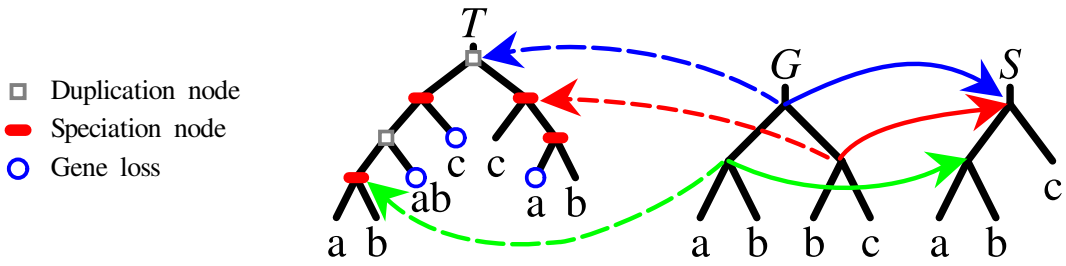

Figure 2.3. An example of scenario T for a gene tree G and a species tree S and two corresponding mappings: $\xi\colon V_G \to V_T$ and $\phi_T\colon V_G \to V_S$ shown for internal nodes of G (Figure from [Paszek and Górecki, 2017a]).

A scenario in which all internal nodes are mapped into the root of the species tree we call **fat** (see [Górecki and Tiuryn, 2006]).

2.2.2 LCA model

LCA model consists of the single most parsimonious scenario that has minimal number of gene duplication and gene loss events. We presented the formal definition of this model in [Paszek and Górecki, 2017a].

Let us fix a gene tree G and a species tree S such that $\mathcal{L}(G) \subseteq \mathcal{L}(S)$. Then the smallest size scenario for G and S induces the minimal number of gene duplications [Bonizzoni et al., 2005, Górecki and Tiuryn, 2006]. Such scenarios, which are equivalent to reconciled trees [Page, 1994], are defined by the lca-mapping between G and S as follows.

Definition 1 (LCA model of allowed scenarios)**.** *The* ***lca-scenario*** *for G and S is a DLS-tree $R^*(G,S) = \rho(\mathsf{root}(G), \mathsf{M}(\mathsf{root}(G)))$, such that $\rho(g,s) = s$ is when g and s are leaves. Otherwise,*

$$\rho(g,s) = \begin{cases} (\rho(g,u), \mathcal{L}(S(v))-)\sim & s \succ u \succeq \mathsf{M}(g), \quad (2.1) \\ (\rho(p,u), \rho(q,v))\sim & \mathsf{M}(p) \preceq u \prec \mathsf{M}(g) = s \succ v \succeq \mathsf{M}(q), \quad (2.2) \\ (\rho(p,s), \rho(q,s))+ & \mathsf{M}(g) = \mathsf{M}(p) = s. \quad (2.3) \end{cases}$$

where u, v are the children of s in (2.1)-(2.2) and p, q are the children of g in (2.2)-(2.3).

Having the above definition, we can divide the set of internal nodes of G into two parts: **lca-duplications** satisfying condition (2.3) and the remaining elements (i.e., satisfying (2.2)) called **lca-speciations**.

Now, the classical duplication cost between a G and S is defined as the total number of gene duplication nodes in $R^*(G,S)$. Note that $\mathsf{M} = \phi_{R^*(G,S)}$ and $\mathsf{LCA}(G,S) = \{R^*(G,S)\}$.

2.3. Unrooted Reconciliation

While the classical reconciliation model has been introduced with rooted trees as evolution is a time-directed process, in computational practice, most standard phylogenetic inference methods from molecular sequences, like maximum likelihood, maximum parsimony or neighbor joining, infer unrooted gene family trees, and it is often difficult, to identify credible rootings [Farris, 1970, Fitch and Markowitz, 1970, Felsenstein, 1981, Saitou and Nei, 1987, Ronquist and Huelsenbeck, 2003].

For example, outgroup rooting can result in incorrect rootings when evolutionary events cause heterogeneity in the gene trees, and rooting gene trees under the molecular clock assumption, or similarly by using midpoint rooting, also can result in error when there is a molecular rate variation throughout the tree [Holland et al., 2003, Huelsenbeck et al., 2002].

The approach that selects the rooting of an unrooted gene tree basing on minimization of a cost function has been addressed in [Chen et al., 2000]. The idea analyzed in [Górecki and Tiuryn, 2007b, Górecki and Tiuryn, 2007a, Górecki and Eulenstein, 2012a, Górecki et al., 2013] is to seek the rooting that in the context of a given species tree indicates the minimum number of evolutionary events such as gene duplications or gene duplications and losses. Here, we introduce this idea by presenting background definitions and selected theoretical results that were fundamental to this thesis.

The unrooted gene tree is an undirected acyclic connected graph in which each node has degree 1 (leaves) or 3 (internal nodes), and the leaves are labeled by the names of species. For an unrooted gene tree $U = \langle V_U, E_U \rangle$ and an edge $e \in E_U$, by U_e, we denote the rooting of U obtained from U by placing the root on e. Such a rooting induces the duplication cost $\mathsf{D}(U_e, S)$. We call **D-minimal**, the rooting or edges having the minimal duplication cost in the set of all rootings of U. It follows from the theory of unrooted reconciliation [Górecki and Tiuryn, 2007a, Górecki et al., 2013] that the set of D-minimal edges, called **D-plateau**, is a full subtree of U (see Figure 2.4).

The same property holds for the DL-plateau, that is, the set of edges with the minimal duplication-loss cost. We use a similar notation for DL-minimal edges, rootings and so on. The most important property of these plateaus is below.

Theorem 1 (From [Górecki et al., 2013])**.** *DL-plateau is a subgraph of D-plateau.*

In this work, the subtree induced by the set of all D-minimal edges will be denoted by U^*. For X, the set of edges of unrooted tree U, by $U|_X$ we denote the smallest subgraph of U containing all edges from X.

Without loss of generality we assume that every root of a gene tree is mapped into the root of S, denoted by $\mathsf{root}(S)$, and both trees are non-trivial. An edge $e = \langle v, w \rangle$ of U is **empty** if the root of U_e is a speciation, i.e., $\mathsf{M}_{U_e}(v) \neq \mathsf{root}(S) \neq \mathsf{M}_{U_e}(w)$. We call e **double** if $\mathsf{M}_{U_e}(v) = \mathsf{root}(S) = \mathsf{M}_{U_e}(w)$. Otherwise, e is called **single**. A single edge e is called v**-incoming** or w**-outgoing** if $\mathsf{M}_{U_e}(v) \neq \mathsf{root}(S) = \mathsf{M}_{U_e}(w)$.

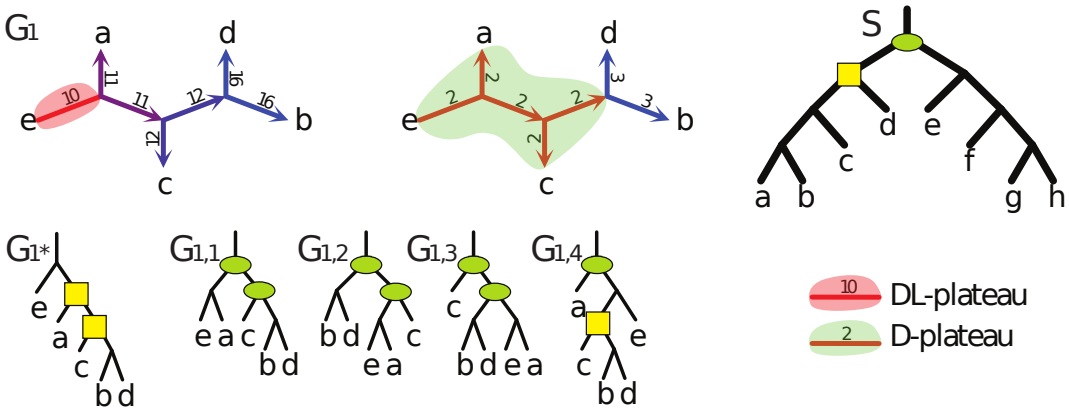

Figure 2.4. An example of plateaus. **Top:** Unrooted gene tree G_1 with DL-plateau, and with D-plateau subtrees marked. The corresponding species tree S with two locations of duplications shown. The number placed above the edge in unrooted gene tree denotes the duplication-loss cost when DL-plateau is marked (and duplication cost for D-plateau) of the reconciliation of the rooting at that edge with S. **Bottom:** All D-plateau rootings of G_1.

Let v be an internal node of U, then a **star** with a **center** v consists of three edges, denoted by e_a, e_b and e_c, sharing v and incident to nodes a, b and c, respectively (see Figure 2.5). The are several types of possible star topology based on the above classification of edges: the $S1$ star has one v-incoming edge and two v-outgoing edges, the $S2$ star has exactly two v-outgoing edges and one empty edge, the $S3$ star has two v-outgoing edges and one double edge, the $S4$ star all 3 edges are double, and the $S5$ star has one v-outgoing edge and two double edges. The star topologies are depicted in Figure 2.5.

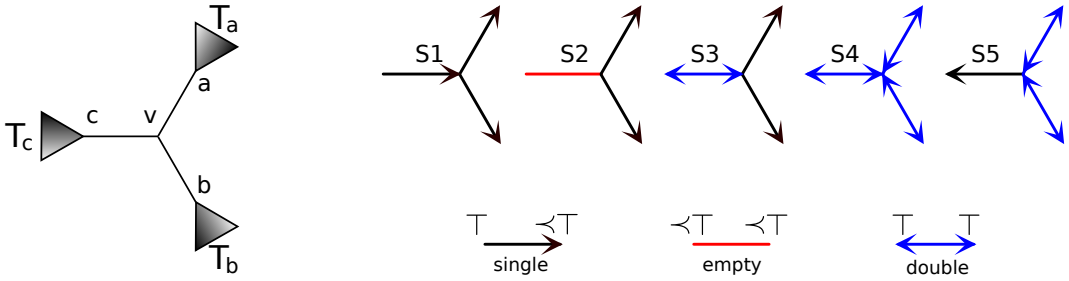

Figure 2.5. Types of stars. Star topology with the center v, types of edges and stars, where $\top$ denotes $\mathsf{root}(S)$. (Figure from [Paszek and Górecki, 2016]).

Theorem 2 (Adopted from [Górecki and Tiuryn, 2007a]). *For a given unrooted gene tree U, we have*

- *either U has exactly one empty edge or G has at least one double edge,*
- *if the DL-plateau of U consists of exactly one edge, then this edge is either empty or double, and all other edges are single.*
- *if the DL-plateau of U has more than one edge, then it contains all edges present in stars $S4$ and $S5$, and all other edges are single.*

Note that if a gene has an empty edge, then it has at most two stars S2 (see examples in Figure 5.2).

The presence of the star called $S2$ having one empty edge in an unrooted gene tree will be of major interest in our analysis. In such a case the remaining edges are single, and by using the notation from Figure 6.1, for $x \in \{a, b\}$ we have that $\mathsf{M}_{U_{\langle v,x\rangle}}(x) \neq \mathsf{root}(S) = \mathsf{M}_{U_{\langle v,x\rangle}}(v)$.

CHAPTER 3

Genomic duplications

This Chapter presents the concept of genomic duplication and introduce the formal definitions to describe it. Moreover, it contains our theoretical results in the area of duplication models from [Paszek and Górecki, 2016, Paszek and Górecki, 2017a].

The desired research goal is to discover the locations of gene duplications. In Section 3.1 we explain various approaches of detecting whole-genome duplications. The conclusion from that review is the fact of the absence of a method which is based on phylogenetics and reconciliation of gene trees with a species tree. Probabilistic methods that use phylogenetic information exist but this information is used to refine main approach.

Section 3.2 is an introduction to the concept of multiple gene duplication.

The reconciliation model described in Chapter 2 is defined for a single rooted gene tree. In such a case scenario of the evolution defined by the lca-mapping is the most parsimonious, that is, it requires the minimal number of gene duplications and gene losses events [Górecki and Tiuryn, 2006]. However, to properly depict evolution we need to incorporate information from numerous gene trees. In that case, the problem becomes more complex, i.e., the phenomena of whole-genome duplication and hybridization result in creation of organisms with multiple copies of genes. This phenomena is not unusual, polyploids are common among plants, as well as among certain groups of fish and amphibians. Moreover, the mechanisms that lead to tandem duplications also may result in a duplication of multiple number of genes. Therefore, to obtain models more suitable for multiple gene trees, we may relax the reconciliation based on lca-mapping for a single gene tree. In Section 3.3 we present description of such models. In particular, we introduce the model from [Paszek and Górecki, 2016] that preserves the minimal number of single gene duplications. Moreover, we propose mathematical description, classify and compare such models (see [Paszek and Górecki, 2017a]).

Section 3.4 presents existing studies that focus on analyzing multiple gene duplications and are based on reconciliation. The research described in this thesis concerns the unsolved problems named in Section 3.4. Solutions to that problems are a milestone in the development of successful method of detecting whole-genome duplications and the approach of this methods is novel in comparison to existing ones described in Section 3.1.

Finally, this Chapter presents related research in order to show the context of our work that enables the evaluation of pioneering nature of the project and its impact on the development of the research field. Moreover, the review of related research reveal interesting algorithmical and mathematical (topological) problems and the opportunity to challenge them was also our motivation.

3.1. Whole genome duplications

In this dissertation we study problems that concern multiple gene duplication events. The special case of such events are whole-genome duplication (WGD) events. The phenomenon of WGD and its impact on the theory of evolution is described in Chapter 1. In Section 1.4.2 we present arguments for the great potential of practical applications of multiple gene duplications studies.

Section describes existing methods of detection of whole-genome duplications. Identifying genomic duplications is a challenging task as duplicated genome fragments can be either lost or retained, in each organism independently leading to a patchy distribution of duplicated copies (see Figure 3.1).

3.1.1 Methods of detection

The methods of detecting whole-genome duplications can be divided into three categories: based on synteny and colinearity comparison of genomes [Kellis et al., 2004, Tang et al., 2008, Holloway et al., 2013], the estimation of the age distribution of paralogous gene pairs [Vision et al., 2000, Lynch and Conery, 2000, Blanc and Wolfe, 2004], and phylogenetic tree inference [Bowers et al., 2003, Jiao et al., 2011, Rabier et al., 2014].

Synteny uses the genome sequence of given species to infer relatively recent whole genome duplications. WGD events have characteristic signature of matching pairs of synteny blocks. The process of detection of whole genome duplication is based on interspecies comparison of genomes [Kellis et al., 2004, Tang et al., 2008, Lyons et al., 2008]. Limitation of this approach is the requirement of data on whole genomes with synteny locations. Moreover, extensive rearrangements of the genome and loss of copies of genes over time reduce the size of synteny blocks and hinder the identification of ancient WGD events.

Another method, called K_S, is the estimation of the age distribution of paralogous gene pairs through the average number of synonymous substitutions per synonymous site [Vision et al., 2000, Lynch and Conery, 2000, Blanc and Wolfe, 2004]. However, negative impact on K_S approach have factors like: excessive gene loss, molecular rate heterogeneity among lineages, gene families or even genes, concentration of duplicate pair estimates on more recent nodes, and saturation of Ks between older paralogue pairs. For example, WGD event inferred by synteny method was not evident in K_S plot for paralogue pairs of **Arabidopsis thaliana** [Blanc and Wolfe, 2004, Jiao et al., 2011]. Hence, K_S method is the most effective in detecting recent WGD events. Moreover, both synteny and K_S method do not directly estimate the timing of whole genome duplications.

The mapping of the paralogs created by given WGD event onto phylogenetic trees can be used to determine whether the paralogs resulted from a duplication event before or after a given speciation event [Bowers et al., 2003]. Similar phylogenetic strategy that use gene family trees of many species was applied to detect and locate WGD events [Jiao et al., 2011, Jiao et al., 2012, McKain et al., 2012]. First, duplication nodes are selected from those estimated to occur on a specific branch of species phylogeny. Then, the age distribution of these duplications is analyzed in a similar way to K_S values. The advantages of this method are: the potential to detect much older WGD events than synteny or K_S based methods, and the ability of estimating

the time and the phylogenetic location of the WGD event. The limitation of this method is the selection of particular duplications that have to occur in given time interval. The authors of [Rabier et al., 2014] proposed a probabilistic model to refine this method. The WGD inference that bases on the mutation rate and incorporates the protein-protein interaction network perspective is presented in [Zhu et al., 2013]. In conclusion, phylogenetic methods are based on probabilistic approach (bootstrap, Bayesian Information Criterion).

The methods proposed in this dissertation can be classified as phylogenetic. In the next sections, we focus on the phylogenetic concepts needed to understand the problem of genomic duplication.

3.2. Multiple gene duplications - reconciliation approach

The problem of discovering locations of gene duplications and multiple gene duplications is fundamental to understand the way gene families and genomes evolve. That lead to the studies on the inference of large **genomic duplications**, called also **multiple gene duplication events**, that can span through thousands of genes families, in which parts of a genome are duplicated. In fact, it is known that a large duplication event is usually followed by many gene losses and gene rearrangements (see Chapter 1). In consequence, the reconstruction of such events may be difficult (see Figure 3.1). The reconciliation of a single gene family tree with a species tree is relatively simple from computational point of view (see Section 2.1). However, when focusing on multiple gene duplications, the problem becomes more complex.

We can apply the reconciliation in order to identify the location of gene duplications in the species tree. Then, we are able to infer the event of genomic duplication by **grouping** single duplication events located at the same node of a species tree. Now we can formulate the general concept of the genomic duplication problem as the problem of clustering as follows:

Given a collection of gene trees and a species tree.
Find the minimal size clustering of all single gene duplications.

The formulation of the problem defined as above for the classical LCA-based reconciliation (as depicted in Figure 2.2), requires perfect trees with complete data without errors. In practice, errors in sequencing, computational limitations or biological processes such as gene loss or horizontal gene transfer make obtaining perfect trees an impossible task. Therefore, in general, the LCA-reconciliation that locates a single duplication event on the lowest possible node in the species tree is not appropriate to model multiple duplication events (see Figure 3.1).

3.3. Models of allowed evolutionary scenarios

In the pioneering article [Guigó et al., 1996], Guigó et al. in their approach to detect multiple gene duplication episodes proposed to relax the LCA-reconciliation, by allowing some additional locations for single duplication events. In result, this method may lead to the increase of the reconciliation cost (see examples in Figure 3.2).

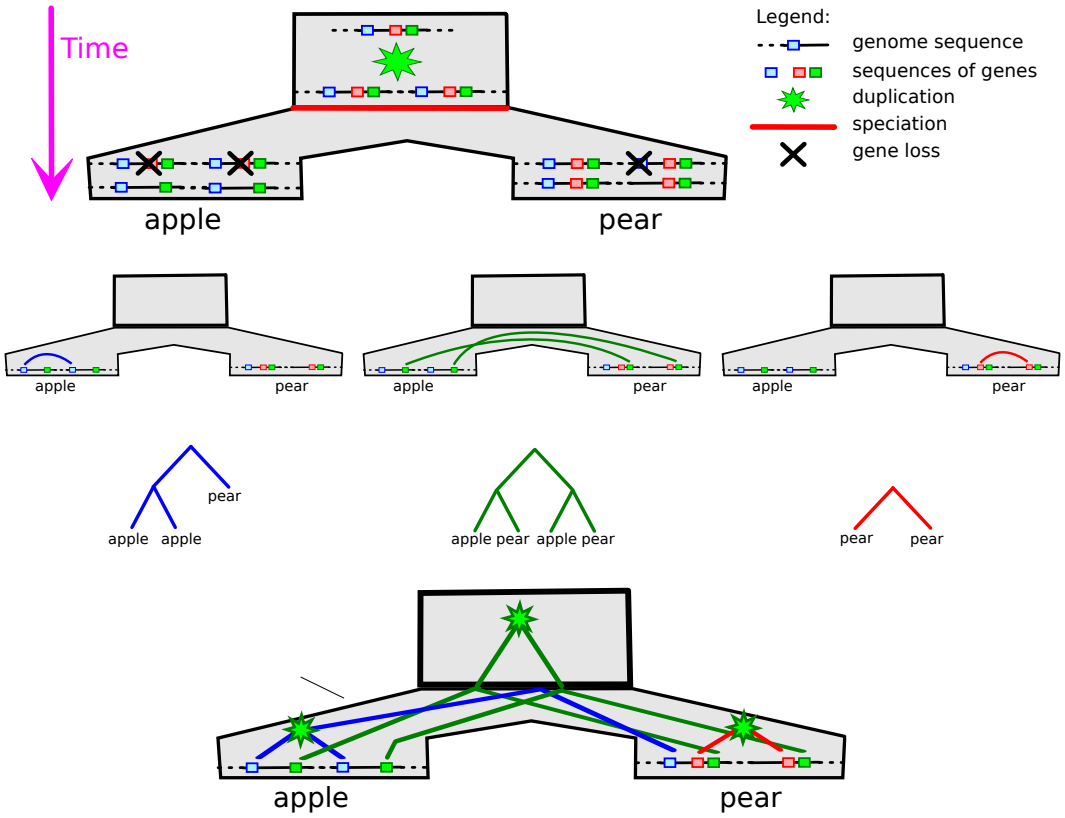

Figure 3.1. An example of evolution with an occurrence of the duplication of multiple genes. **Top:** A species tree in which every node represents a species, depicts the relation between three species: a pear, an apple and an ancestor of pear and apple. Inside a node, there is a selected fragment of a sequence of the corresponding species. Due to a multiple duplication (marked by a green star in the root of the tree), the sequence of three genes (denoted by colors red, green, blue) were duplicated in the ancient ancestor of pear and apple. However, as a result of processes that cause gene losses, current sequences of pear and apple genomes contain different combinations of genes. **Center:** An example of the inference of gene trees from species. **Up:** A species tree in which a node represents a species. Inside a leaf node, there is a selected fragment of a sequence of the corresponding species. The lines connect the sequences of homologous genes, that is sequences identified to share a common ancestor. **Below:** Inferred rooted gene trees that are based on the similarity of sequences. **Bottom:** An example of an evolutionary scenario in the LCA reconciliation. The gene trees inferred from sequences suggest three events of single gene duplication instead of one multiple duplication event as indicated in top part of the picture.

In this Section we describe models that for a given gene tree and a species tree induce the set of the best evolutionary scenarios, called **allowed scenarios**, evaluated by criteria defined by model (please refer to Section 2.2 for the definition of an evolutionary scenario).

For a gene tree G and a species tree S, a **model of allowed scenarios**, or a **model**, denoted by $\mathcal{A}(G, S)$ is a set of scenarios, called allowed scenarios, for G and S.

In conclusion, we can reformulate the problem in the following way:

Given a collection of gene trees, a species tree and a model of allowed scenarios. Find the minimal size clustering of all single gene duplications.

We distinguish various models of allowed scenarios that we describe in the following sections.

The LCA model that induces only one allowed scenario which is a product of reconciliation of a gene tree with the species tree is described in Section 2.2.

3.3.1 GMS model

Let us describe the model from the pioneering work of Guigó et al., which was proposed in the original paper that introduce multiple gene duplications [Guigó et al., 1996]. This model was used in the majority of the following studies of the subject [Page and Cotton, 2002, Bansal and Eulenstein, 2008, Burleigh et al., 2008, Mettanant and Fakcharoenphol, 2008, Burleigh et al., 2010, Luo et al., 2011].

Definition 2 (GMS model of allowed scenarios)**.** $\mathsf{GMS}(G,S)$ *consists of all scenarios* T *having the minimal number of duplications such that for any lca-duplication node* d *in* G*:*

- $F_T(d) \succeq \mathsf{M}(d)$*, if* $d = \mathsf{root}(G)$*,*
- $F_T(d) = \mathsf{M}(d)$*, if* $\mathsf{M}(d) = \mathsf{M}(\mathsf{par}(d))$*,*
- $\mathsf{M}(d) \preceq F_T(d) \prec \mathsf{M}(\mathsf{par}(d))$*, otherwise.*

The restrictiveness of the model of allowed scenarios GMS was the reason for us to seek for more general model definition.

3.3.2 PG model - a parsimonious model that preserves minimal number of single gene duplications

Now, we describe our model introduced in [Paszek and Górecki, 2016]. To model gene duplication episodes we allow to relocate a gene duplication from its lca-mapping location to one of its ancestors. In other words, we introduce mappings representing evolutionary scenarios that can differ from the scenario defined by the lca-mapping. Additionally, we require that the total number of gene duplications is minimal. To ensure biological correctness of such mappings, we introduce several conditions, e.g., time order preservation. The restriction to preserve the minimal number of single gene duplications is motivated by the parsimony. The increase in number of gene losses is justified by the desire to represent the theory of evolution, in which duplication of numerous genes is followed by many gene losses [Ohno, 1970]. The commonness of polyploidy strongly supports that theory (see Chapter 1 for more details). The formal definition of the model is:

Definition 3 (PG model of allowed scenarios)**.** PG *model can be defined by a mapping* $\mathsf{F}_G\colon V_G \to V_S$*, which is called* ***valid*** *if the following conditions are satisfied:*

- $\mathsf{F}_G(a) \preceq \mathsf{F}_G(b)$ *if* $a \preceq b$ *(time consistency),*
- $\mathsf{F}_G(a) = \mathsf{M}_G(a)$ *for any speciation node* a *(fixed speciations),*
- $\mathsf{F}_G(a) \succeq \mathsf{M}_G(a)$ *for any duplication node* a *(duplication can be raised),*

- $\mathsf{F}_G(a) \prec \mathsf{M}_G(b)$ *for any speciation node* b *such that* $a \prec b$ *(fixed number of gene duplications).*

It can be shown that every valid mapping uniquely defines an evolutionary scenario represented by a DLS-tree [Górecki and Tiuryn, 2006], which will be called an allowed scenario for PG model. Additionally, every DLS-tree obtained from a valid mapping can be transformed into the optimal evolutionary scenario (i.e., lca-based scenario), by a sequence of TMOVE (i.e., lowering duplication) transformations. Please refer to [Górecki and Tiuryn, 2006] for more details on formal modeling of evolutionary scenarios. $\mathsf{PG}(G, S)$ is the set of all scenarios for G and S having the minimal number of gene duplications. Observe, that the above model is more general than the GMS model [Bansal and Eulenstein, 2008].

3.3.3 FHS model

Th most general model was proposed in [Fellows et al., 1998].

Definition 4 (FHS model of allowed scenarios)**.** *In this model, we call it the* FHS *model, any scenario for* G *and* S *is allowed.*

Depending on the rules of how to cluster single gene duplications, one problem for this model is NP-hard, whereas the solution to other problem is trivially a fat scenario. Please refer to Table 3.1 for more details. The definition of fat scenario is in Section 2.2, the definition of genomic duplication problems in Section 3.4.

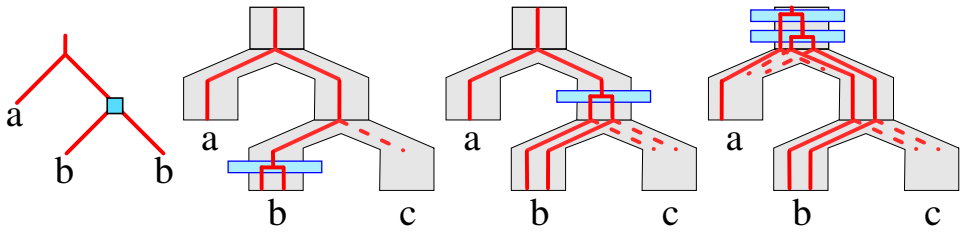

Figure 3.2. An example of evolutionary scenarios for a gene tree $G = (a, (b, b))$ with single duplication and a species tree $S = (a, (b, c))$. **Left:** A gene tree G with single duplication marked by blue rectangle. **Right:** Three embeddings of a gene tree G into species tree S that represent evolutionary scenarios. Observe, that first scenario defined by LCA mapping has minimal location of the duplication in the species tree. It infers the minimal number of duplication and loss events. The formal relation between scenarios described in [Górecki and Tiuryn, 2006] we can treat as the movement of the duplication up in the species tree. Note, that there are two kinds of moves (see [Górecki and Tiuryn, 2006]). First kind is like the move between the first and the second scenario, it does not increase the number of duplications, whereas the second kind, like move between the second and the third scenario, does increase the number of duplications. The last (third) scenario is an example of a fat scenario. The first scenario is allowed in LCA model. The first and second scenarios are possible variants in the GMS model. All three scenarios are allowed in FHS model.

3.3.4 Interval models

In this Section we present the idea of **interval models**, the models in which allowed duplication locations can be defined by a path in the species tree. The concept

of intervals was introduced in [Czabarka et al., 2012] in a more general framework without requirement that the intervals induce a biologically consistent evolutionary scenarios. Below, we present a refined definition that introduce restrictions like preserving the monotonicity of nodes in gene tree.

Definition 5 (Interval model of allowed scenarios)**.** *A model $\mathcal{A}$ is an **interval model** if for every gene tree G, a species tree S and $T \in \mathcal{A}(G,S)$, we have:*

- *for every lca-speciation g in G, $\xi(g)$ is a speciation node in T,*
- *for every lca-duplication d, $F_T(d) \in \mathsf{Int}(d)$, where $\mathsf{Int}(d)$ is an interval in S, connecting two comparable nodes s and s'; by $\mathsf{I}(d)$ we denote the pair $\langle s, s' \rangle$,*
- *for any duplications $d \preceq d'$, $\min \mathsf{I}(d) \preceq \min \mathsf{I}(d')$ and $\max \mathsf{I}(d) \preceq \max \mathsf{I}(d')$.*

The intervals for LCA are trivially defined by $\mathsf{I}(d) = \langle \mathsf{M}(d), \mathsf{M}(d) \rangle$, while intervals in GMS are given above in the definition of models. In PG the interval $\mathsf{Int}(d)$ is the maximal interval of nodes $\succeq \mathsf{M}(d)$ such that there is no lca-speciation g satisfying $F_T(g) \in \mathsf{Int}(d)$. It should be clear that LCA, GMS and PG satisfy the above conditions, while FHS does not. Directly from this property we have that every allowed scenario in the interval model has the minimal number of duplications, therefore, PG is the most general interval model for biologically consistent evolutionary scenarios (see Figure 4.4 for example).

3.4. Genomic duplication problems

3.4.1 Rules of clustering single gene duplications

Two fundamental issues arise when dealing with genomic duplication problems: (1) a model of allowed evolutionary scenarios and (2) the rules of clustering gene duplications from gene trees into a single multiple duplication event. The model is described in previous section, here we show the definitions of clustering rules.

To provide an accurate model of many simultaneous genomic duplications, we need additional rules that determine when two single duplications can be clustered. Given a model of allowed scenarios, we can distinguish three variants of **multiple gene duplication problems** (see Figure 3.3), that differ in such rules:

- **Episode Clustering** - gene duplications can be clustered if they can be mapped to the same node of the species tree [Guigó et al., 1996, Page and Cotton, 2002, Bansal and Eulenstein, 2008],
- **Minimum Episodes Clustering** - duplications from the same gene tree can be clustered if they are not comparable and both can be mapped to the same node of the species tree [Guigó et al., 1996, Bansal and Eulenstein, 2008],
- **Gene Duplication Clustering** - gene duplications cannot be clustered if they occur in the same gene tree [Fellows et al., 1998].

In this thesis we consider two variants of the above problems: for rooted gene trees and unrooted gene trees. Therefore, we use abbreviations REC and RME to denote rooted variants of Episode Clustering and Minimum Episodes, respectively, and UEC

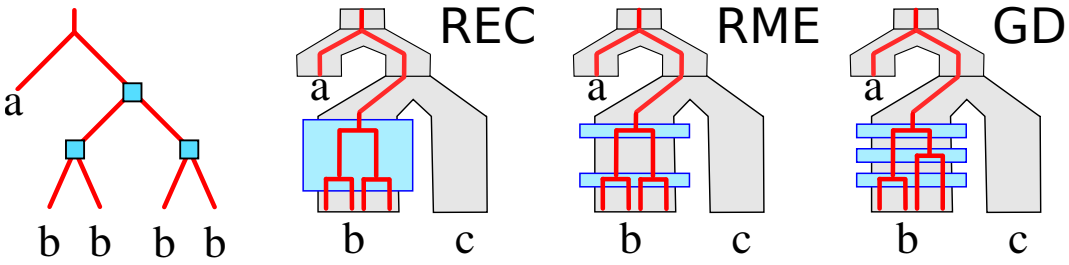

Figure 3.3. An example of REC, RME and GD duplication clustering with one gene tree, where GD denotes Gene Duplication Clustering. **Left:** a gene tree with three gene duplications mapped to b. **Right:** embeddings (scenarios) of the gene tree into a species tree showing a solution to REC, RME, and GD problems, respectively. In REC all duplications are clustered together, while in RME the top duplication cannot be clustered with its children. Hence, the solution to RME consists of two clusters represented by rectangles. GD clustering has 3 clusters as duplications from the same tree cannot be clustered together.

and UME to denote Episode Clustering and Minimum Episodes for unrooted gene trees.

REC Problem can be treated as a simplified version of the general genomic duplication problem in which the goal is to find only the minimal number of nodes of the species tree at which the multiple duplication event occurred. In other words, solutions to REC rather provide a rough estimation of the genomic duplication events. For example, if two WGD occurred between two consecutive speciation events, they will be clustered as one multiple duplication event. From the biological point of view, the most desired are solutions to RME. See also Figure 3.3.

3.4.2 Related work

REC Problem is to find evolutionary scenarios with the minimal number of locations of duplication episodes in a species tree. In other words, two duplications can be clustered if they have the same location in the species tree. The problem was introduced by Guigó et al. [Guigó et al., 1996] with the GMS-model and a heuristic solution. Page and Cotton [Page and Cotton, 2002] formulated the problem of locating episodes of gene duplication as a set cover problem and proposed a heuristic. Bansal and Eulenstein [Bansal and Eulenstein, 2008] introduced the polynomial time algorithm to solve REC Problem under GMS-model which is a special case of Tree Interval Cover Problem (TIC) [Burleigh et al., 2008]. Burleigh et al. [Burleigh et al., 2008] presented polynomial time solution to TIC Problem. Finally, Luo et al. [Luo et al., 2011] proposed a linear time and space algorithm for TIC Problem that applies to REC Problem under every interval model. REC Problem for FHS-model has a trivial outcome with one cluster.

Gene Duplication Clustering Problem is similar to REC Problem with the difference that a cluster cannot have two gene duplications from the same tree. Gene Duplication Clustering Problem for the FHS-model is NP-hard [Fellows et al., 1998].

Clustering for the REC Problem could be refined by excluding cases in which a duplication and its ancestor duplication (from the same gene tree) are clustered together. Such a formulation of multiple gene duplication problem is called RME Problem [Guigó et al., 1996, Bansal and Eulenstein, 2008] (see Figure 3.4).

The first polynomial time algorithm for RME Problem under GMS-model was pro-

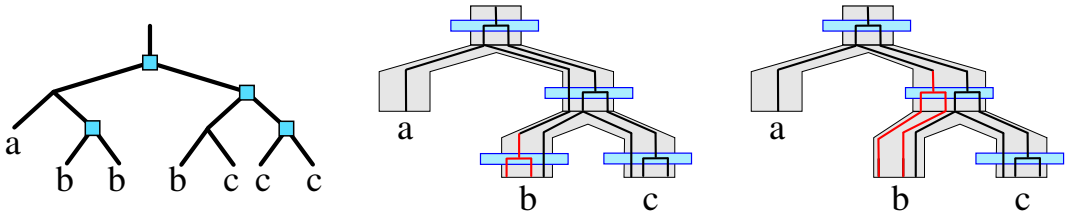

Figure 3.4. An example of a solution to RME Problem for the gene tree G and species tree S from Figure 2.2. **Left:** The gene tree with duplication marked. **Right:** The minimal scenario defined by LCA mapping and the optimal scenario for RME clustering under GMS model.

posed in [Bansal and Eulenstein, 2008], whereas the optimal linear time and space algorithm in [Mettanant and Fakcharoenphol, 2008, Luo et al., 2011]. The concept of intervals was introduced in [Czabarka et al., 2012] in a more general framework without requirement that the intervals induce a biologically consistent evolutionary scenarios. The iterative algorithm from [Czabarka et al., 2012] implemented in straightforward way has $O(|S|^2|G|)$ complexity ([Czabarka et al., 2012] suggests that the algorithm from [Bansal and Eulenstein, 2008] can solve instances for every interval model, however, it is designed for GMS-model and cannot be generalized). In summary, there was a need of a general algorithm that solves RME. In Chapter 4 we present a solution for a variety of models, in particular, the linear time algorithm applicable for any interval model.

Moreover, there were no solutions to the variants of the problem in which input gene trees are unrooted.

Table 3.1. Summary of genomic duplication problems for rooted gene trees. Here, we assume that an instance consists of a set of rooted trees and a species tree all having n leaves in total.

Model Problem	LCA	GMS	PG	FHS
Episode Clustering REC	$O(n)$ time trivial solution	$O(n)$ by Alg. from [Luo et al., 2011]	$O(n)$ we adapted in [Paszek and Górecki, 2016] Alg. from [Luo et al., 2011]	trivial solution one location the root
Minimum Episodes RME	$O(n)$ time easy	$O(n)$ by Alg. from [Luo et al., 2011]	$O(n)$ our interval Alg. in [Paszek and Górecki, 2017a]	complexity unknown exponential Alg. in [Paszek and Górecki, 2017a]
Gene Duplication GD	$O(n)$ time easy	complexity unknown	complexity unknown	NP-hard [Fellows et al., 1998]

In this Chapter we presented mathematical foundations for general genomic duplication problems introduced in [Paszek and Górecki, 2017a]. In summary, often problems in computational biology are defined in two variants: for rooted and unrooted gene trees independently. There are algorithms for classical computational biology problems defined for unrooted trees like [Górecki and Eulenstein, 2011, Górecki and Eulenstein, 2012b, Górecki and Eulenstein, 2012a, Chang et al., 2013, Górecki and Eulenstein, 2014c, Betkier et al., 2015]. Finding a solution for a particular problem in the unrooted variant is sometimes more desired as rooting a gene tree might be difficult. In Chapter 5 and in Chapter 6 we present the solutions to **Unrooted Episode Clustering** (UEC), and **Unrooted Minimum Episodes** (UME) problems, respectively (see [Paszek and Górecki, 2016, Paszek and Górecki, 2017b, Paszek

and Górecki, 2018]). The fundamental result that enabled the creation efficient algorithms for UME was the linear time solution to RME universal for any interval model described in Chapter 4 (see [Paszek and Górecki, 2017a]).

Table 3.1 summarizes the genomic duplication problems for rooted gene trees. The corresponding problems for unrooted gene trees were open. In this dissertation, efficient algorithmic solutions to UEC and UME under PG model, proposed in [Paszek and Górecki, 2016, Paszek and Górecki, 2018], are described in Chapter 5 and Chapter 6.

CHAPTER 4

Minimum Episodes Problem for rooted gene trees

In this Chapter we focus an the problem of multiple genomic duplications for the case where all gene trees are rooted. The description of the problem is presented in Section 4.1. The solution to that problem is described in Section 4.2, which presents description of the algorithms. In particular, we propose a linear time and space algorithm for solving RME Problem jointly for any interval model including GMS and PG models. Next, we describe how it can be applied to solve RME for the most computationally demanding FHS model.

Section 4.5 describes the datasets which were selected by us for experimental evaluation of our algorithms on real biological data. Section 4.6 contains a comparative study for RME Problem under four models of allowed scenarios for simulated and biological datasets. This chapter contains the main results published in [Paszek and Górecki, 2017a].

4.1. Multiple gene duplications

Detecting the events of multiple gene duplications is both interesting from the mathematical and algorithmic point of view and desired by biologists in their studies. This statement is supported by the fact that novel methods and their application to detect the phenomenon of whole-genome duplication are published in top-rated journals like Nature or Science [Vision et al., 2000, Lynch and Conery, 2000, Bowers et al., 2003, Kellis et al., 2004, Tang et al., 2008, Jiao et al., 2011].

Two fundamental aspect are crucial to define a multiple gene duplication problem. First, is to choose a model of allowed evolutionary scenarios, which is responsible for determining the allowed locations in the species tree for gene duplications. Next, is to define the rules of clustering gene duplications from gene trees, which are assigned to the same location, into a single multiple duplication event. Then, the problem is to find the clustering of the minimal size.

4.1.1 The definitions of RME problems

First, we introduce the cost for determining the number of multiple gene duplication episodes for a collection of evolutionary scenarios.

Let $\mathcal{R}$ be a collection of scenarios compatible with a species tree S. We say that duplication nodes d and d' from $\mathcal{R}$ are **clusterable**, denoted $d \sim_c d'$, iff

- (1) d and d' have the same cluster and

- (2) if d and d' are present in the same DLS-tree then either d and d' are incomparable or equal.

Then, the minimum number of duplication episodes for a collection of scenarios, denoted $\mathsf{MES}(\mathcal{R}, S)$, is the minimal size of the partition of the set of all duplication nodes present in scenarios such that every two duplications from the same partition set are clusterable. The elements of the partition we call **(multiple duplication) episodes**. Formally,

$$\mathsf{MES}(\mathcal{R}, S) = \min_{\bigcup P = \mathsf{Dup}(R)} \{|P| \colon \forall_{A \in P} \forall_{d,d' \in A}\ d \sim_c d'\},$$

where $\mathsf{Dup}(\mathcal{R})$ is the set of all duplication nodes present in $\mathcal{R}$.

The minimum number of episodes for a collection of scenarios can be obtained as follows.

Lemma 1. *For a collection $\mathcal{R}$ of scenarios compatible with a species tree S,*

$$\mathsf{MES}(\mathcal{R}, S) = \sum_{v \in V_S} \max_{T \in \mathcal{R}} \mathsf{duppath}(T, v),$$

where $\mathsf{duppath}(T, v)$ is the maximal (node) length of path in T that consists of comparable duplication nodes whose cluster equals the cluster of v.

Proof. It follows from the fact that every $v \in S$ requires at least $\max_{T \in \mathcal{R}} \mathsf{duppath}(T, v)$ episodes in order to satisfy the condition from the MES score definition. □

Originally the problem of multiple gene duplications was introduced by [Guigó et al., 1996] and formalized in [Fellows et al., 1998, Bansal and Eulenstein, 2008], however, despite similar concept of multiple gene duplication these problems are not equivalent due to the differences in the model of allowed scenarios.

Multiple duplication studies generate broad class of problems that vary in properties that depend on the method of clustering and the model of allowed scenarios. Let choose minimum episodes as the clustering method. Now, we obtain the following meta-problem parameterized by a model.

Problem 1 (Minimum Episodes under $\mathcal{A}$)**.** *Given: a collection of gene trees G_1, $G_2, \ldots, G_n$ a species tree S and a model of allowed scenarios $\mathcal{A}$. Compute* ***minimum episodes score****, or* RME *score, as:*

$$\mathsf{RME}_{\mathcal{A}}(G_1, G_2, \ldots, G_n, S) = \min_{\forall_i R_i \in \mathcal{A}(G_i, S)} \mathsf{MES}(\{R_i\}_{i=1,2,\ldots,n}, S).$$

Examples of RME clustering are depicted in Figure 4.1 and in Figure 4.2. Note that the definition of $\mathsf{RME}_{\mathsf{GMS}}$ score is equivalent to definitions from [Bansal and Eulenstein, 2008, Luo et al., 2011].

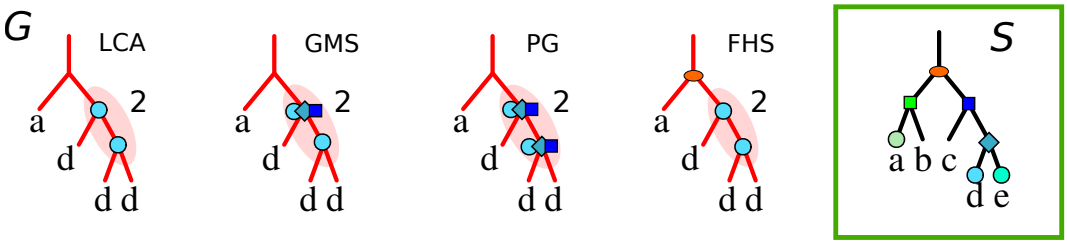

Figure 4.1. An example of solutions to RME Problem under LCA, GMS, PG and FHS for a species tree S and gene tree G (Figure based on Figure from [Paszek and Górecki, 2017a]). For interval models (LCA, GMS, PG) the duplication nodes are marked by species tree nodes from the corresponding interval. For FHS the duplication and speciation nodes which are ancestors of some duplication node are marked according to their lca-mapping. Each highlighted region in a gene tree denotes nodes mapped into the same node in the species tree in an RME clustering. The numbers near regions denote the contribution of each region to the overall optimal RME score. The RME score is 2 in every model.

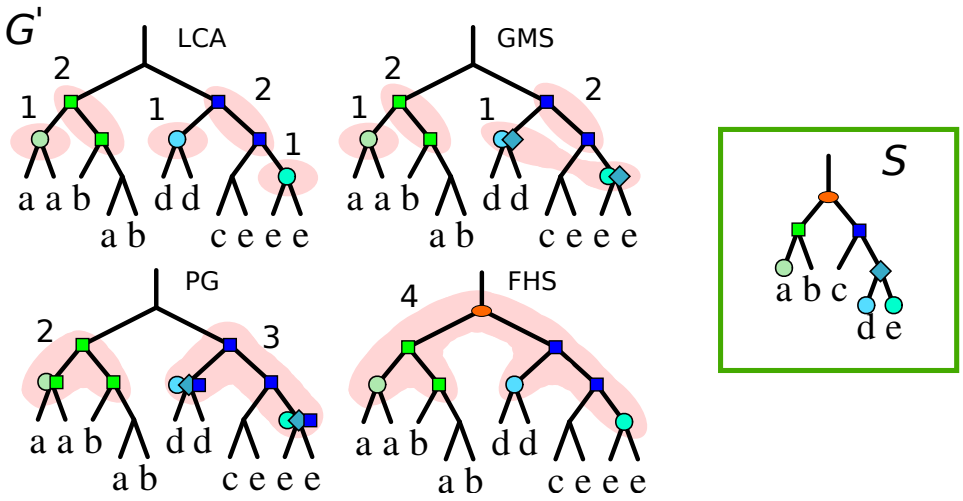

Figure 4.2. An example of solutions to RME Problem under LCA, GMS, PG and FHS for a species tree S and gene tree G' (Figure based on Figure from [Paszek and Górecki, 2017a]). For the description of markings please refer to Figure 4.1. The numbers near regions denote the contribution of each region to the overall optimal RME score. $\mathsf{RME}_{\mathsf{LCA}}$, $\mathsf{RME}_{\mathsf{GMS}}$ and $\mathsf{RME}_{\mathsf{PG}}$ score for G and S is 7, 6 and 5, respectively. In order to obtain $\mathsf{RME}_{\mathsf{FHS}}$ score the speciation node at the root of G has to be converted into a duplication. In consequence, the optimal RME clustering consists of 4 episodes, where all duplications are mapped to the root of S.

Theorem 3. *For any collection of gene trees $\mathcal{G}$ and a species tree S we have*

$$\mathsf{RME}_{\mathsf{FHS}}(\mathcal{G},S) \leq \mathsf{RME}_{\mathsf{PG}}(\mathcal{G},S) \leq \mathsf{RME}_{\mathsf{GMS}}(\mathcal{G},S) \leq \mathsf{RME}_{\mathsf{LCA}}(\mathcal{G},S).$$

Proof. It follows from inclusions of the corresponding allowed scenarios sets. □

See also Figure 4.3 for a more complex example of episode clustering.

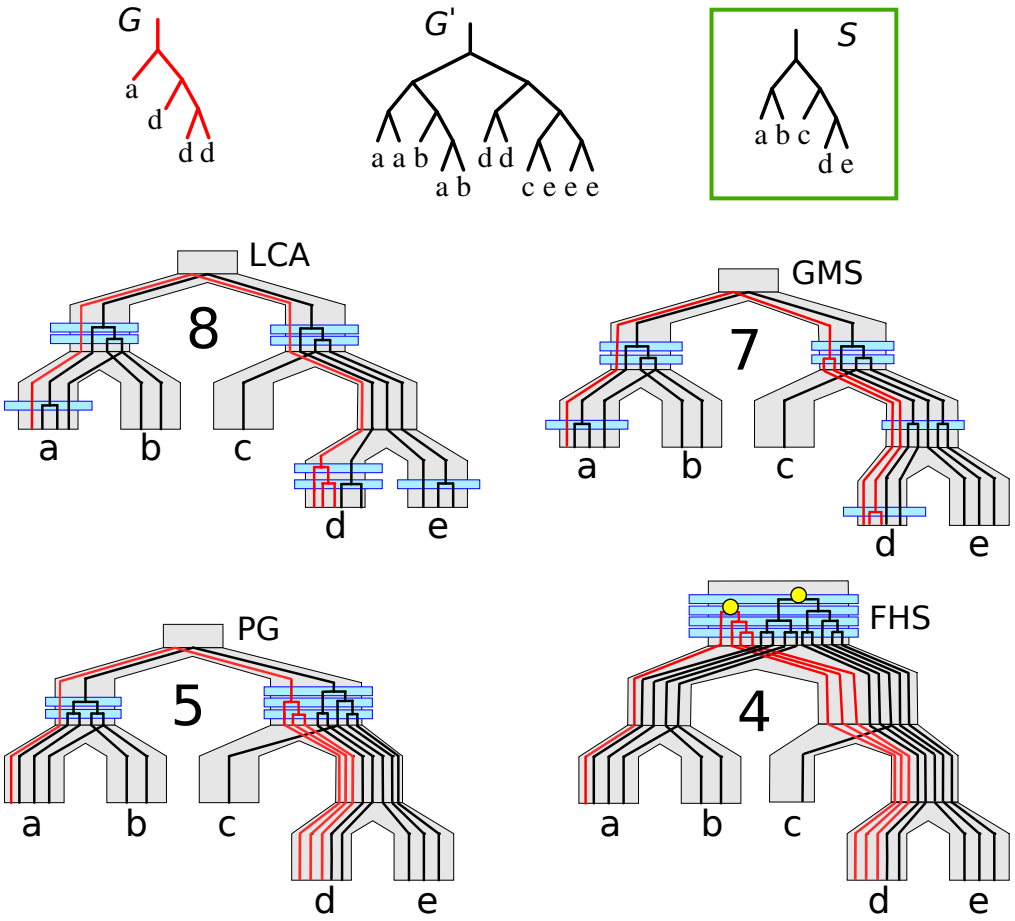

Figure 4.3. Selected optimal solutions to RME Problem under all models for two input gene trees $\{G, G'\}$ and S from Figure 4.1 and Figure 4.2 (Figure based on Figure from [Paszek and Górecki, 2017a]). **Top:** The gene trees and the species tree. **Bottom:** The solutions to RME Problem is presented as an embedding of G and G' into S. Rectangles denote episodes, while circles in FHS scenario denote the speciation nodes converted into duplications. Embeddings are decorated with RME scores (in large font size).

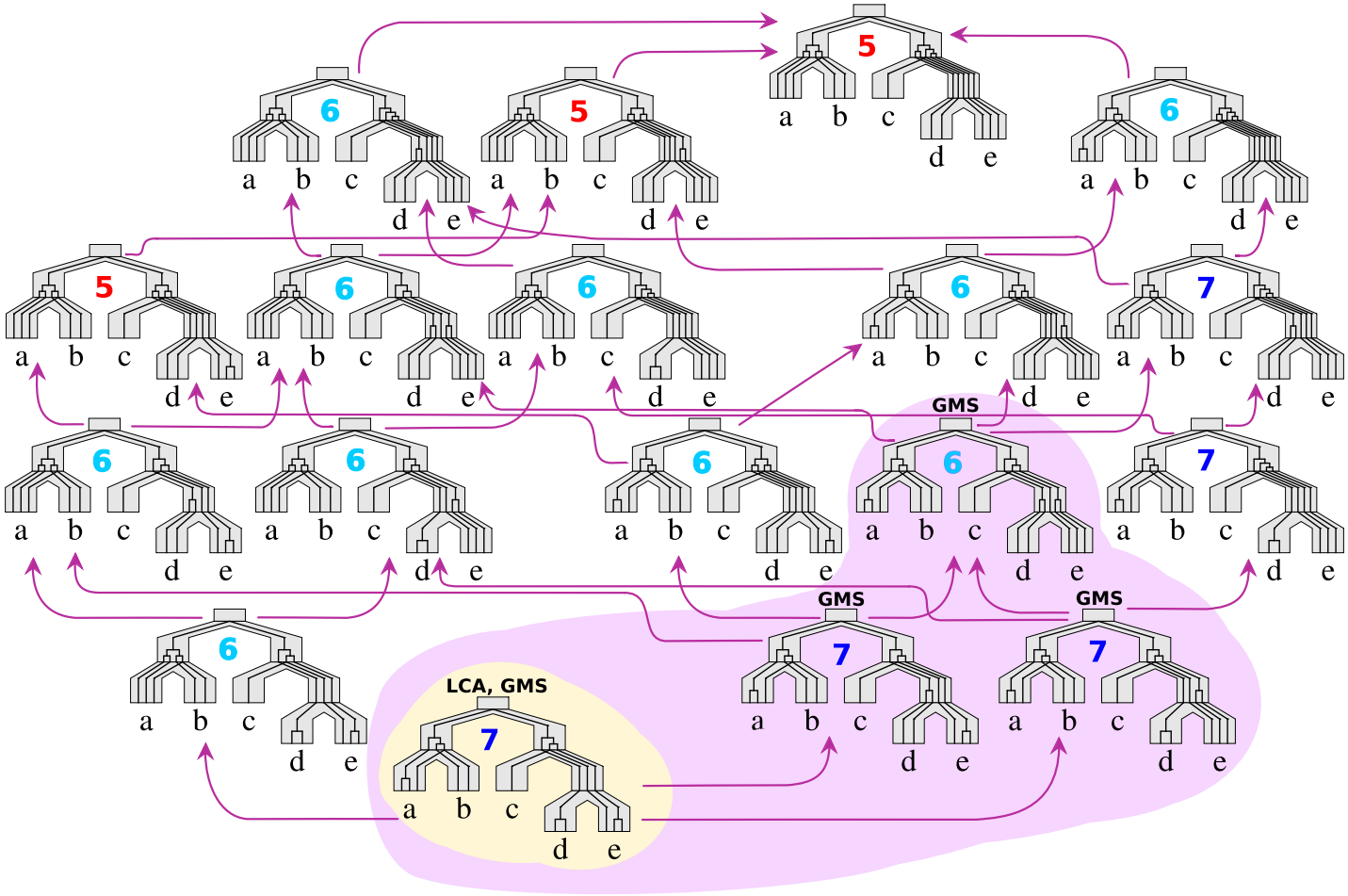

Figure 4.4. (Cont. from Figure 4.2) All evolutionary scenarios for the gene tree G and species tree S with the minimal number of gene duplications shown with RME scores (Figure from [Paszek and Górecki, 2017a]). The relations between DLS-trees are shown by arrows (this is the reverse TMOVE rule defined in [Górecki and Tiuryn, 2006]). All these scenarios are allowed in PG model. The set of allowed scenarios under GMS model consist of 4 scenarios only. Note that more than one scenario could be optimal to RME Problem. Algorithm 2 for PG model returns score 5 based on the top scenario.

4.2. Solution to the RME problem

In [Czabarka et al., 2012] we can find a study of the solution to a genomic duplication problem, however, the proposed model of intervals was used without the requirement that the intervals can model properly evolutionary scenarios. The authors presented a general iterative algorithm to compute the RME score given an interval model (depicted here in Algorithm 1). [Czabarka et al., 2012] provided a proof of correctness and claimed that the algorithm from [Bansal and Eulenstein, 2008] can be used to solve efficiently the problem. However, the latter algorithm is designed to solve RME Problem under GMS and cannot be generalized due to a simple interval model (e.g., intervals for comparable duplications in GMS intersect in at most one node). Starting with a naive implementation of Algorithm 1, we need $O(|S|)$ steps for the main loop and for the most expensive line 5 at least $O(|G|^2)$. Hence, the time complexity of Algorithm 1 is $O(|S||G|^2)$.

In summary, we have the following complexity results for RME Problem: the complexity for FHS is open, GMS can be solved in linear time [Luo et al., 2011], while PG can be solved in $O(|S|^2|G|)$ by naive implementation of Algorithm 1. In the next section, we propose an efficient linear time algorithm for RME score inference for any interval model.

Algorithm 1 RME score under an interval model (adopted from [Czabarka et al., 2012])

1: **Input:** A collection of gene trees $G_1, G_2, \ldots, G_n$, a species tree S and a set of duplication nodes $\mathsf{Dup} \subset \bigcup_i V_{G_i}$ such that for every $d \in \mathsf{Dup}$, $\mathsf{I}(d) = \langle s, s' \rangle$, where $s \preceq s'$ are the ends of the interval associated with d.
2: **Output:** $\mathsf{RME}(G_1, G_2, \ldots, G_n, S)$.
3: Let t be the lowest among top nodes of intervals, i.e., $t = \min_d \max \mathsf{I}(d)$.
4: Let k be the maximal length of the t-chain, where t-chain is a path consisting of duplication nodes d such that $\max \mathsf{I}(d) = t$.
5: For every t' mark a duplication d such that $t \in \mathsf{Int}(d)$ if there is no t'-chain having at least k nodes below d
6: Remove all marked duplication intervals, add k to the score and repeat steps 3-6 until there is no interval left.

4.3. Linear-time solution to RME under interval models

In this section, we propose Algorithm 2 and we show that it is a linear time variant of Algorithm 1. Having this, there is no need to provide a separate proof of correctness of Algorithm 2, because it is already given in [Czabarka et al., 2012] for Algorithm 1. The following Lemmas 2, 3 and 4 study the correspondence between Algorithm 2 and Algorithm 1.

Lemma 2. *The sequence of nodes t visited in the main loop of Algorithm 1 is equal to the sequence of nodes t (from T) visited in the main loop of Algorithm 2.*

Proof. It follows easily from the construction of tree T from Algorithm 2. □

Lemma 3. *For every fixed t, visited in the main loops of Algorithm 1 and Algorithm 2, values of k from line 4 of Algorithm 1 and line 8 of Algorithm 2 are equal.*

Algorithm 2 Solution to RME under an interval model

1: **Input/output:** see Algorithm 1.
2: Let T be a subtree of S induced by $\{\max \mathsf{I}(d)\}_{d\in\mathsf{Dup}}$;
Prepare lca data structures for T.
3: **For** $d \in \mathsf{Dup}$: $\mathsf{lowest}(d) := \min \mathsf{Int}(d) \cap V_T$,
where $\mathsf{Int}(d) = \{v : \min \mathsf{I}(d) \le v \le \max \mathsf{I}(d)\}$.
4: **For** $t \in T$: $\mathcal{U}(t) := \{d \in \mathsf{Dup}:\ \max \mathsf{I}(d) = t$ and $\big(\exists_i d = \mathsf{root}(G_i)$ or $\mathsf{par}(d) \notin \mathsf{Dup}$ or $\mathsf{par}(d) \in \mathsf{Dup}$ and $\max \mathsf{I}(d) \neq \max \mathsf{I}(\mathsf{par}(d))\big)\}$
5: **For** $t \in T$: $\mathcal{B}(t) := \{d \in \mathsf{Dup}:\ \mathsf{lowest}(d) = t$ and $\mathsf{ch}(d) \cap \mathsf{Dup} = \emptyset\}$,
where $\mathsf{ch}(d)$ is the set of children of d.
6: RME $:= 0$;
For every gene tree node g: $g.\mathsf{active} := g \in \mathsf{Dup}$.
7: **For** $t \in T$ in postorder // **Main loop, lines 7-29**
8: Let $k := \max_{r\in\mathcal{U}(t)} \mathsf{epi}(r)$, where under assumption that $\max \emptyset = -\infty$,

$$\mathsf{epi}(r) = \begin{cases} -\infty & \text{if not } r.\mathsf{active} \\ 1 + \max(0, \max_{c\in\mathsf{ch}(r)} \mathsf{epi}(c)) & \text{otherwise.} \end{cases}$$

9: **If** $k = -\infty$ **Then** $\mathcal{B}(\mathsf{par}_T(t)) := \mathcal{B}(\mathsf{par}_T(t)) \cup \mathcal{B}(t)$
Else {
10: RME $:=$ RME $+k$; cand $:= \emptyset$
11: **For** $d \in \mathcal{B}(t)$ // **Loop A, lines 11-22**
12: $h := 1$; newcand $:=$ **null**
13: **While** $h \le k$ and not newcand // **Loop A', lines 13-21**
14: $d.\mathsf{visited} := t$
15: **If** d is the root of a gene tree **Then Break**
16: $\sigma := \mathsf{sibling}(d)$; $d.h := h$; $d := \mathsf{par}(d)$
17: **If** not $d.\mathsf{active}$ **Then Break**
18: **If** $\mathsf{lca}_T(\mathsf{lowest}(d), t) \neq t$ or $h = k$ **Then** newcand $:= d$
19: **Elif** not $\sigma.\mathsf{active}$ **Then** $h := h + 1$
20: **Elif** $\sigma.\mathsf{visited} = t$ **Then** $h := 1 + \max(k, \sigma.h)$
21: **Else Break**
22: **If** newcand **Then** cand $:=$ cand $\cup\{$newcand$\}$
23: **For** $d \in \mathcal{B}(t)$ // **Loop B, lines 23-24**
24: **While** $d.\mathsf{active}$ and $d.\mathsf{visited} = t$: { $d.\mathsf{active} :=$ **false**; $d := \mathsf{par}(d)$ }
25: **For** d in cand // **Loop C, lines 25-28**
26: **If** not $\mathsf{left}(d.\mathsf{active})$ and not $\mathsf{right}(d.\mathsf{active})$ **Then**
27: **If** $\mathsf{lca}_T(\mathsf{lowest}(d), t) = t$ **Then** $\mathcal{B}(\mathsf{par}_T(t)) := \mathcal{B}(\mathsf{par}_T(t)) \cup \{d\}$
28: **Else** $\mathcal{B}(\mathsf{lowest}(d)) := \mathcal{B}(\mathsf{lowest}(d)) \cup \{d\}$
29: }
30: **Return** RME

Lemma 4. *At the beginning of main loops, d is a duplication with a non-removed interval in Algorithm 1 if and only if d is an active (i.e., $d.active$=True) duplication in Algorithm 2.*

Proof. The proof is by induction, where Loop A corresponds to line 5 from Algorithm 1. The key observation is that marked nodes in line 5 (Algorithm 1) are visited in Loop A (line 14, Algorithm 2).

The next lemma describes the property of the main structure for efficient marking and removal of intervals.

Lemma 5. *For every fixed t, visited in the main loop of Algorithm 2, let A_t be the set of all active duplications at the beginning of the main loop. Let F_t be the subgraph (a forest) of G induced by A_t. Then, $\mathcal{B}(t)$ consists of all leaves d from F_t such that $t \in \mathsf{Int}(d)$.*

Proof. The proof is by induction. The property should be clear when visiting the first node (see line 5). Assume that, t is visited and in line 8 we have a node d in $\mathcal{B}(t)$ that is not a leaf in F_t. Then, there is a node $t' \prec t$, such that $d \in \mathcal{B}(t')$ was inserted into $\mathcal{B}(t)$ either in line 9 or in lines 27/28 of the step of the main loop processing t'. In the first case, t is a parent of t', and there was no active duplication below d, which is a contradiction. The second case is similar: see the condition in line 26. Finally, every leaf from F_t is present in $\mathcal{B}(t)$, which is a consequence of the traversal in Loop A/A', where the parent of the last visited/marked node becomes a new leaf candidate (see line 18). □

Theorem 4. *Given a collection of gene trees $\mathcal{R}$ and a species tree S, Algorithm 2 computes* $\mathsf{RME}(\mathcal{R}, S)$.

Proof. By Lemmas 2-5, Algorithm 2 implements Algorithm 1. The proof of correctness of Algorithm 1 is given in [Czabarka et al., 2012]. □

Theorem 5. *Algorithm 2 has $O(|S| + \sum |G_i|)$ time and space complexity.*

Proof. We assume that trees are implemented in a standard pointer-like structure. For every $t \in T$, the set of **duplication leaves** $\mathcal{B}(t)$ is a list that uses pointers. For computation of lca_T (e.g., in line 18) in constant time, we need the lca-structure preprocessing in $O(|T|)$ time present in line 2 [Bender and Farach-Colton, 2000]. Dup is not stored as a separate structure, it is sufficient to have an attribute in every node of gene trees. Having this, it is not difficult to see that lines 2-6 require several traversals of input trees and a tree T (being smaller than S).

In line 8, only duplication intervals whose top is t are processed, therefore, in total, every duplication node is visited at most once when computing epi.

Loop A, traverses up to k levels of all active duplication trees starting from the current set of duplication leaves. All visited duplications are then set to be non-active, therefore, they will never be visited again. Hence, in all runs of the main loop, Loop A requires time proportional to $|\mathsf{Dup}|$, which is $O(\sum |G_i|)$. The same applies to Loop B. Finally, Loop C consists of the update of duplication leaves. Every new candidate located in line 18 has to be appended to the list of a node $\succ t$ from T. However, every candidate requires the removal of at least one unique duplication node in Loop A' (i.e., setting its active status to False). Therefore, in the total execution of Algorithm 2, the number of steps in line 25 is limited by $|\mathsf{Dup}|$. We conclude that in the total execution of Algorithm 2, the steps of Loops A-C require $O(|\mathsf{Dup}|)$ time. We conclude that the time and space complexity is linear. □

4.4. Algorithms for RME Problem under FHS model

It is unknown whether RME Problem under FHS is tractable, however, there are several properties of $\mathsf{RME}_{\mathsf{FHS}}$ that can be used to solve or approximate hard instances of this problem in practice.

Recall that by height of a tree, we define the maximal number of edges on the path from a leaf to the root of the tree.

Lemma 6 ($\mathsf{RME}_{\mathsf{FHS}}$ upper bound)**.** *The maximal height of a gene tree from the input is an upper bound of* $\mathsf{RME}_{\mathsf{FHS}}$.

Proof. The bound is obtained by converting all speciation nodes into duplications. Then, all internal nodes are duplications which can be mapped to the root. In such a case the number of episodes is equal to the maximal height of input gene trees. □

We conclude from the proof that the upper bound is reached by a trivial mapping where all internal nodes are mapped into the root of the species tree (a fat scenario, see Section 2.2).

The number of lca-duplications present in an interval of a gene tree we call **duplication index**.

Algorithm 3 RME score under the FHS model

1: **Input/output:** see Algorithm 1.
2: Let Spec be a set of all lca-speciation nodes (for $G_1, G_2, \ldots, G_n$) which are ancestor of some duplication.
3: Calculate lower and upper bounds for RME score (see Lemma 6, 7).
4: **Return** RME score if the bounds are equal
5: **For** every subset X of Spec
6: Convert every speciation from X into a duplication
7: Set intervals for all duplication nodes by using PG model
8: Apply Algorithm 2 to obtain RME score
9: **Return** minimal RME score

Lemma 7 ($\mathsf{RME}_{\mathsf{FHS}}$ lower bound)**.** *The lower bound of* $\mathsf{RME}_{\mathsf{FHS}}$ *is the maximal duplication index among all intervals from all input gene trees.*

Proof. RME clustering rules determine that comparable duplications have to be in different episodes. Therefore, the score cannot be lower than the maximal duplication index. □

Examples of bounds can be found in Figure 4.8.

The complexity of Algorithm 3 is $O(2^k)$, where k is the size of Spec (see line 5 of Algorithm 3).

4.5. Datasets for experimental evaluation

In our analysis we performed our experiments both on simulated and biological datasets. The simulated trees were generated by urec [Górecki and Tiuryn, 2007b]

according to Yule model of random trees and the choice of different parameters (like size, labelling) determined the relation between gene trees and the corresponding species tree. The evaluation of our algorithms focused on three biological datasets: Guigó dataset [Guigó et al., 1996], Génolevures [Sherman et al., 2009] and TreeFam [Ruan et al., 2008].

To properly define a dataset we need the description of the collection of gene trees and a species tree to reconcile with. In our study, for a given collection of gene trees, we sometimes performed multiple tests for different choices of species trees.

The smallest dataset originates from [Guigó et al., 1996], called here **Guigó dataset**, that introduced the concept of multiple duplications, therefore it is our first choice to compare our results to. This set is a collection of 53 rooted gene trees from 16 Eucaryotes [Guigó et al., 1996]. In our study we focused on: (a) the species tree from [Page and Charleston, 1997b] described as the most biologically reasonable, (b) the original species tree introduced with gene trees in [Guigó et al., 1996] paper and with (c) 71 species trees from [Chang et al., 2013], known to have the total minimal duplication cost.

Génolevures consists of 4144 unrooted gene families from nine yeast genomes [Sherman et al., 2009]. We used the corresponding gene family trees inferred by the authors of [Górecki and Eulenstein, 2012a] using tools from Phylip [Felsenstein, 1989]. Our analysis included species tree: (a) from [Dujon, 2006], (b) from [Shen et al., 2016] and (c) the one having the lowest duplication-loss cost computed by Fasturec [Górecki and Eulenstein, 2012b].

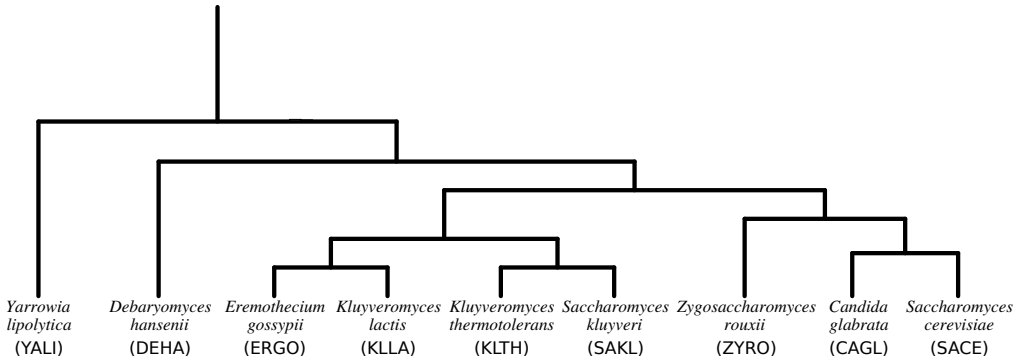

Figure 4.5. Génolevures dataset. Species tree topology from [Shen et al., 2016] with full and abbreviated species names. Please refer to Figure 1.3 and Figure 1.4 to find reference genomic duplications. To find genomic duplications inferred by our tools see Figure 6.5.

TreeFam consists of 1274 unrooted gene family trees [Ruan et al., 2008] sampled from mostly animal species. For this dataset we also produced a set of rooted gene trees obtained by urec [Górecki and Tiuryn, 2007b] by choosing the rooting having the minimal number of gene duplications. The species tree used was based on NCBI taxonomy [Wheeler et al., 2007] (we used two variants one that consists of 28 species and one that span over 25 species).

4.6. Experimental evaluation of RME

We performed several computational experiments on simulated and biological datasets in order to compare the solutions to RME Problem under all four models of evolutionary scenarios.

In the first simulation 100 pairs of bijectively labeled random gene and species trees of the size $n \in \{5, \ldots, 50\}$ were generated according to Yule model (see [Harding, 1971, Steel and McKenzie, 2001]) by urec [Górecki and Tiuryn, 2007b]. In the second experiment we generated 100 pairs consisting of a random species tree of the size n and a random gene tree of the size $2n$, for $n \in \{5, \ldots, 25\}$.

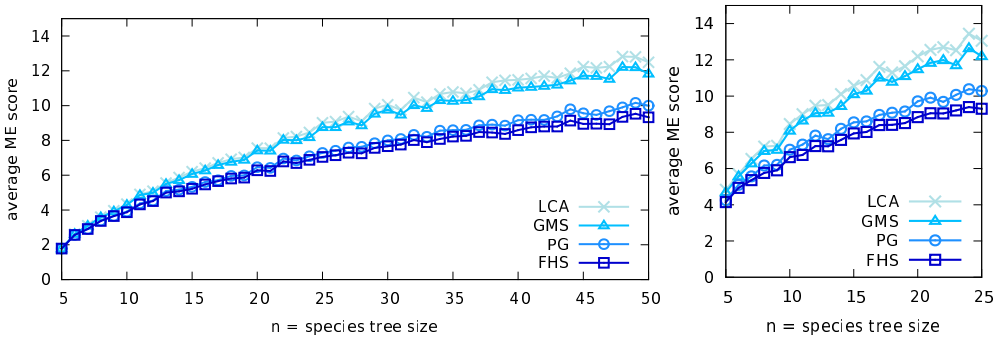

Figure 4.6. The average of optimal RME scores under LCA, GMS, PG and FHS models for simulated datasets (Figure from [Paszek and Górecki, 2017a]). In both experiments for every size n we generated 100 random pairs of gene and species trees. Then, we calculated an average RME score. **Left:** gene trees and species trees were bijectively labeled. **Right:** result for gene trees of the size $2n$.

Experimental evaluation on simulated datasets show that GMS provide similar results to LCA, and RME scores for PG and FHS are comparable (see Figure 4.6).

Guigó dataset. We inferred multiple gene duplication events for the dataset for two species trees: one from [Guigó et al., 1996] and the second tree from [Page and Charleston, 1997b]. Algorithm 3 for FHS reported 15 and 14 lca-speciations in Spec for S_1 and S_2, respectively, therefore, we were able to compute the score. For the first species tree we observe the same multiple gene duplication scenarios for GMS and PG models (see Figure 4.7).

This confirms the results obtained in [Guigó et al., 1996, Bansal and Eulenstein, 2008]. Moreover, the same location of episodes were obtained by [Page and Cotton, 2002] for REC Problem. The species tree from [Page and Charleston, 1997b], considered biologically more meaningful than the first tree, has more diverse scores. Under PG model 6 episodes are placed on 4 nodes. The same number of locations was obtained in the solution to the unrooted variant of EC clustering UEC under PG model (see Section 3.3), however, the results are incomparable as input gene trees were unrooted (see Chapter 5 and [Paszek and Górecki, 2016]).

TreeFam dataset. Algorithm 3 cannot be directly applied as the number of lca-speciations in Spec is 12263. Firstly, we calculated the lower and the upper bounds, which equals 18 and 23 respectively and both are reached by a single unique tree G^* depicted in Figure 4.8. In G^* the number of lca-speciations after preprocessing was equal to 40, which is still too large. Next, we show that 18 of these speciations have

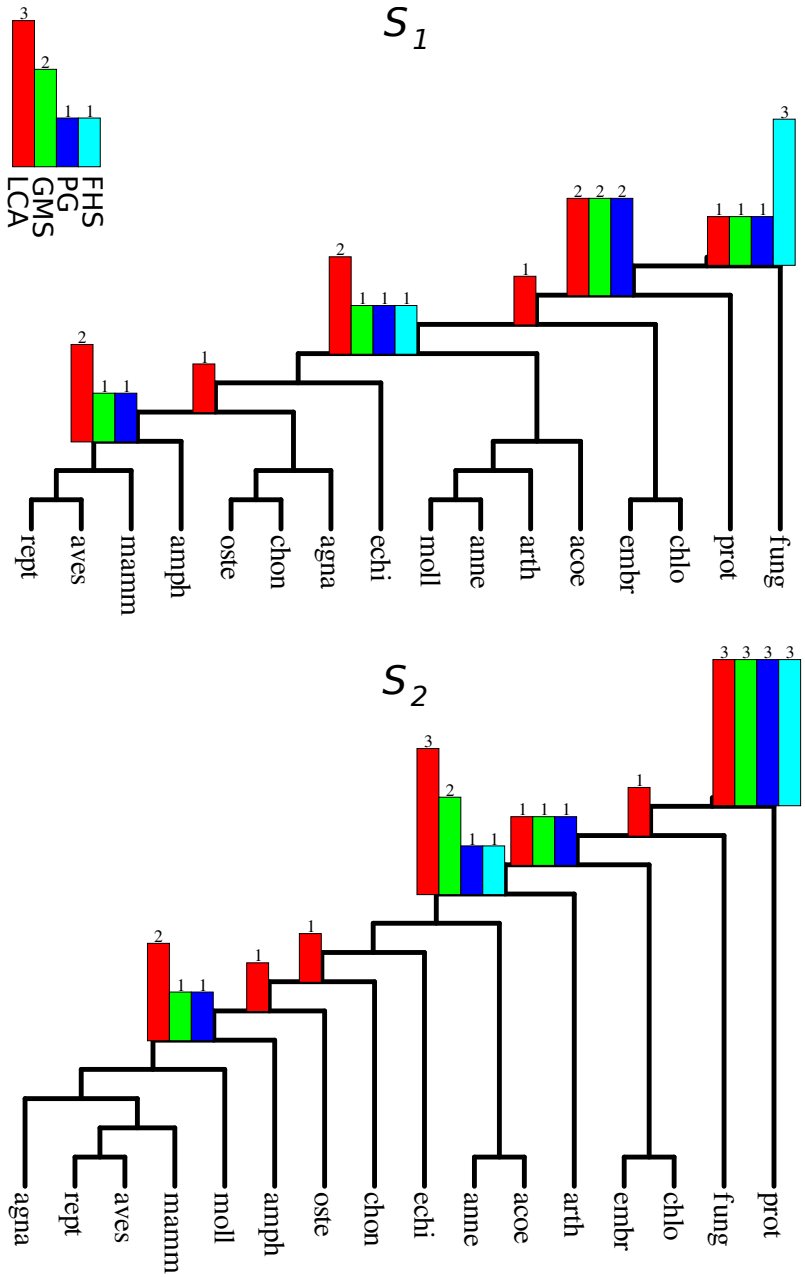

Figure 4.7. The location and the number of episodes (multiple gene duplication events) corresponding to the optimal RME score under LCA, GMS, PG and FHS models in two species trees for Guigó dataset [Guigó et al., 1996] (Figure from [Paszek and Górecki, 2017a]). **Left:** a species tree from [Guigó et al., 1996]. $\mathsf{RME}_{\mathsf{LCA}}$ is 9, $\mathsf{RME}_{\mathsf{GMS}}$ and $\mathsf{RME}_{\mathsf{PG}}$ equals 5 and $\mathsf{RME}_{\mathsf{FHS}}$ is 4. **Right:** a species tree from [Page and Charleston, 1997b]. $\mathsf{RME}_{\mathsf{LCA}}$ equals 12, $\mathsf{RME}_{\mathsf{GMS}}$ - 7, $\mathsf{RME}_{\mathsf{PG}}$ - 6 and $\mathsf{RME}_{\mathsf{FHS}}$ - 4.

to be converted into duplication (otherwise there is no better score than the upper bound of 23). Having this Algorithm 3 reported the best score of 23 and the same result is for the whole input (see Lemma 6).

The results for TreeFam (see Figure 4.9) suggest that RME Problem under FHS for large empirical instances have a simple solution induced by fat scenarios (see

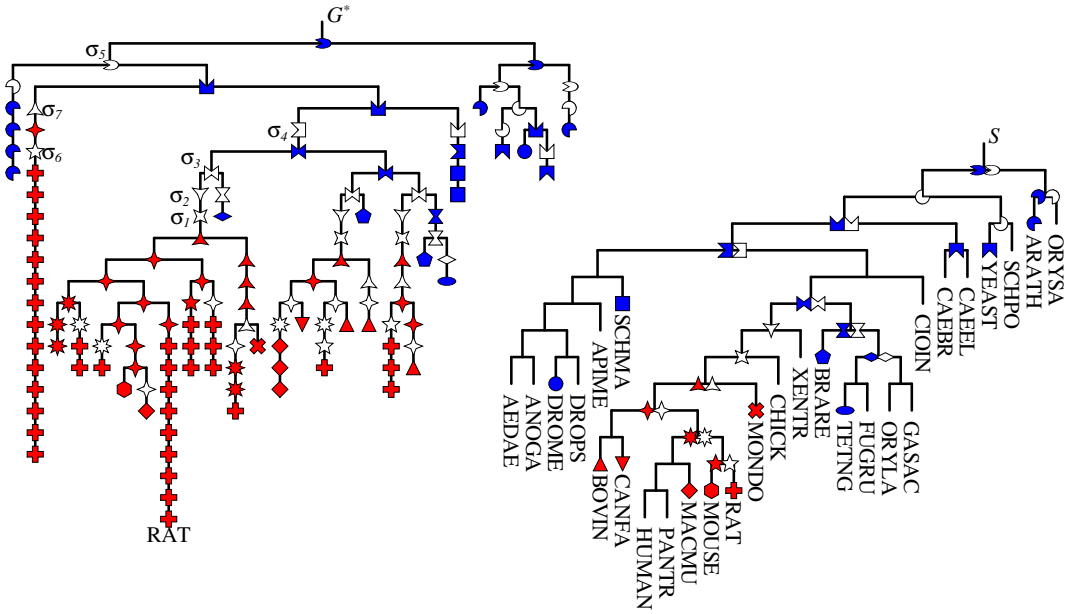

Figure 4.8. Analysis of a Spec speciation set from a gene tree from TreeFam (Figure from [Paszek and Górecki, 2017a]). **Left:** the unique highest tree G^* from TreeFam dataset contracted and compressed to represent all lca-speciations from Spec (see Algorithm 3). Subtrees composed of duplications mapped into the same leaf are replaced by its longest path (see e.g. RAT). Leaves are removed. **Right:** TreeFam species tree. Empty symbols in both trees denote speciation nodes. Red and blue symbols denote duplications. The longest path of G^* has the length 23 and starts in the leaf labeled RAT. It contains five speciation nodes: $\sigma_1, \sigma_2, \ldots, \sigma_5$, therefore the lower bound for RME score is $23 - 5 = 18$. It is not difficult to prove that if σ_6 is not converted into duplication then the RME score is at least 23. Similarly, if σ_7 is not converted then the cost is at least 24. In both cases the score does not improve the upper bound (23), therefore, when searching for the optimal solution, all descendants of red nodes including σ_7 can be converted into duplications, which limits the search space in Algorithm 3 to a tractable size, i.e., $O(2^{22})$, where 22 is the number of the remaining lca-speciation nodes.

Lemma 6), which corresponds to the trivial solution to EC problem under FHS. Moreover, it seems that the upper bound is usually a good approximation of $\mathsf{RME}_{\mathsf{FHS}}$ score.

Runtime. The experiments on simulated datasets were performed on 64 core server for two days. Analysis of the Guigó and TreeFam datasets were done on a standard workstation in 30 seconds and 10 hours, respectively.

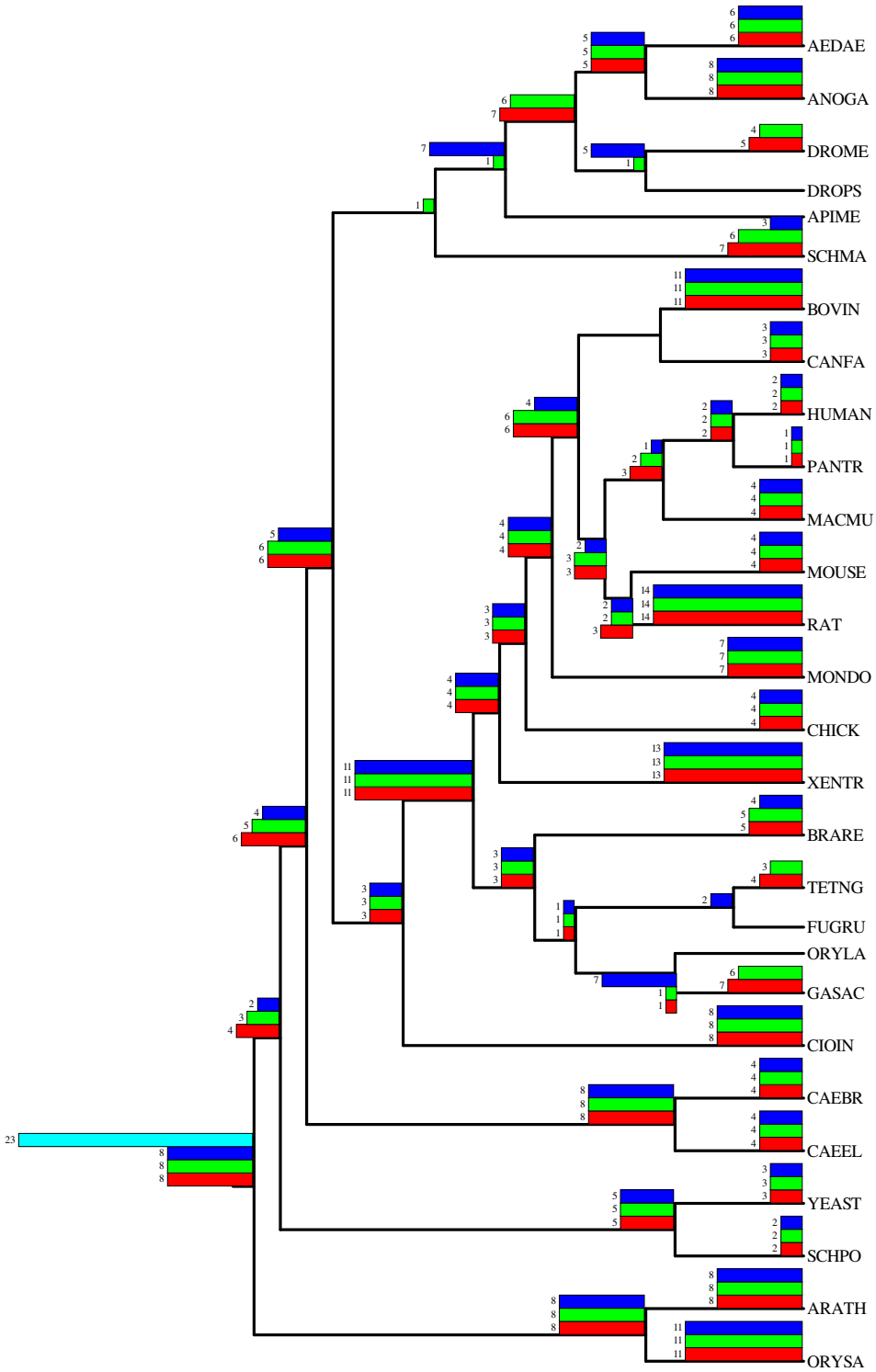

Figure 4.9. The solution to RME Problem under LCA, GMS, PG and FHS for TreeFam [Ruan et al., 2008]. The scores are 249, 243, 230 and 23, respectively.

4.7. Discussion

In this chapter we proposed the first linear time algorithm for solving RME Problem for any interval model of allowed evolutionary scenarios. Then, we applied the algorithm to conduct experiments on biological and simulated datasets in order to compare RME Problem solutions under four models of allowed scenarios. We showed that GMS model infers results comparable to the simplest model LCA, while the most general FHS on a large biological dataset induced a biologically disputable solution, where all duplications are placed at the root of a species tree. Therefore, among all known models, PG model seems to be a reasonable choice for modeling genomic duplications. However, further research is needed especially for the assessment of credibility of inferred genomic duplication events.

There are several further directions. The first is to consider optimal solution that have the lowest number of speciations converted into gene duplications. Next, the complexity of RME Problem under FHS model is an open problem. Finally, multiple gene duplication problems are more complex when the input trees are unrooted. Consideration of unrooted gene trees increases the applicability since such trees are frequently inferred by phylogenetic methods. Algorithm 2 is the first step towards solving UME,the unrooted variant of RME Problem. Our solution to UME is described in Chapter 6. Our implementation of Algorithm 2 and Algorithm 3 is publicly available at http://www.mimuw.edu.pl/jpaszek/rme.html.

CHAPTER 5

Unrooted Episode Clustering

The reconciliation becomes more complex when we consider unrooted gene trees instead of rooted gene trees (see Chapter 2). Similarly, REC and RME problems are defined for rooted trees. Therefore, in order to solve an unrooted variant of, i.e. UEC, we need to choose an edge from the unrooted gene tree, obtain the rooting, and then solve the REC Problem for that rooting. In this Chapter we study UEC Problem in which PG is the model of allowed scenarios and the rules that define how to group duplications defined by episode clustering and input gene trees are unrooted. Section 5.1 provides the formal description of the problem.

In Section 5.2 we present new results for unrooted reconciliation that are applicable to the solution of the problem proposed in Section 5.3. We present the first solution to the open problem UEC enunciated in [Burleigh et al., 2008]. We show that for a given set of unrooted gene trees and a species tree we can solve the UEC by reducing it to the rooted episode clustering problem that has a linear time complexity. Our solutions require a linear time preprocessing and a creation of at most $1 + 2^k$ collections of rooted gene trees, that is, instances of REC, where k is the number of input gene trees having a special topology located in the plateau of the duplication cost (formally, the condition requires two stars S2 [Górecki and Tiuryn, 2007a]). Usually k represents a small fraction of the whole input, thus, this condition significantly reduces the complexity. In other words, we show that the problem of UEC is fixed parameter tractable.

Section 5.6 contains experimental analysis of the implementation of our algorithms. In a number of empirical computational experiments we show that despite the exponential worst case complexity our algorithm is able to resolve instances of the problem after the verification of at most two rooted datasets. In consequence, our solution can be efficiently applied to locate duplication clusters in collections of unrooted gene trees.

This Chapter is mostly based on results published in [Paszek and Górecki, 2016].

5.1. Episode Clustering Problems

In this Section we provide the definition of genomic duplication problems in which we cluster all gene duplications mapped into the same location in a species tree. We call it **episode clustering**.

Recall that by $\mathsf{Dup}(T)$, we define the set of all duplication nodes in T. Let $G_1, G_2, \ldots, G_n$ be a collection of rooted gene trees. Assume that, for every $i \in \{1, 2, \ldots, n\}$, F_i is a valid mapping between G_i and the species tree S (see Definition 3

of valid mapping in Section 3.3). Every element $s \in \bigcup_i \mathsf{F}_i(\mathsf{Dup}(G_i))$ denotes the location of multiple gene duplication events in S. Such locations will be called **duplication episodes**. In this Chapter we use the PG model to determine allowed scenarios which infer those locations (the description of PG model is in Section 3.3).

A **duplication cluster** for s is the set of all gene duplications present in G_i's that are mapped to s. By $\top$-cluster we denote the duplication cluster whose elements are mapped to $\mathsf{root}(S)$.

Problem 2 (Rooted Episode Clustering, REC)**.** *Given a collection of rooted gene trees $G_1, G_2, \ldots, G_n$ and a species tree S. Compute the minimal number of duplication episodes, denoted by $\mathsf{REC}(G_1, G_2, \ldots, G_n, S)$, in the set of all valid mappings $\mathsf{F}_1, \mathsf{F}_2, \ldots, \mathsf{F}_n$ such that $\mathsf{F}_i \colon V_{G_i} \to V_S$.*

The linear-time and space solution to REC for GMS in [Luo et al., 2011] can be applied also to PG model. In this Chapter we solve the following problem for PG model (see Section 3.3.2 for model description).

Problem 3 (Unrooted Episode Clustering, UEC)**.** *Given a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$ and a species tree S. Compute the minimal $\mathsf{REC}(T_1, T_2, \ldots, T_n, S)$ in the set of rooted gene trees $\{T_1, T_2, \ldots, T_n\}$ such that T_i is a rooting obtained from G_i by placing the root on the edge from the D-plateau.*

Observe, that we allow rootings only in the D-plateau (see definition in Section 2.3). Otherwise, the total number of gene duplications is not minimal. By SINGLE-UEC we denote the problem UEC for a single unrooted gene tree, i.e., when $n = 1$. Every edge in an unrooted gene tree that induces the optimal solution for SINGLE-UEC will be called **optimal** (for SINGLE-UEC). For convenience, we assume that S is fixed and use $\mathsf{REC}(T_1, T_2, \ldots, T_n)$ instead of $\mathsf{REC}(T_1, T_2, \ldots, T_n, S)$.

5.2. Novel properties of D-plateau nodes

In this Section we describe new theoretical results in unrooted reconciliation, that is, new properties of the nodes that are inside D-plateau. We start with a technical lemma that shows correspondence between D-plateau and DL-plateau for a case when unrooted gene tree has double edge.

Lemma 8. *If the DL-plateau consists of exactly one double edge then the D-plateau and the DL-plateau are equal.*

Proof. Let $\langle v, a \rangle$ be the DL-plateau edge (see Figure 2.5). It follows from the property of star S3 that both v and a are mapped to $\mathsf{root}(S)$ in the DL-minimal rooting and their children (if present) are mapped below $\mathsf{root}(S)$. Hence, the root is a duplication, while v and a are speciation nodes. Now, it is easy to show that rooting on edge $\langle v, b \rangle$ (or $\langle v, c \rangle$) induces one additional gene duplication at v. We conclude that the only edge with the minimal duplication cost is $\langle v, a \rangle$. □

We say that a node is a **super-duplication** (respectively, a **super-speciation**) if it is a duplication (respectively, a speciation) in every rooting with the minimal duplication cost.

Please recall, that the plateau is a subtree of a gene tree, thus a leaf of the D-plateau may refer to an internal node of a gene tree. For example, in Figure 5.2, the D-plateau of G_1 has four leaves: one is an internal node of G_1 and others, labeled a, c, e, are leaves of G_1.

Lemma 9. *Assume that an unrooted tree has a double edge. Then*

- *every leaf of the D-plateau is a super-speciation*
- *and every internal node of the D-plateau is a super-duplication.*

Proof. For the first part of the proof, let us assume that v is a leaf of the D-plateau. By using the notation from Figure 2.5, let v be a center of a star such that $\langle v, a\rangle$ belongs to the D-plateau. Assume that v is a duplication in every D-minimal rooting. Then, the D-minimal rooting $G_{\langle v,a\rangle}$ has one duplication in v. The edge $\langle v, b\rangle$ does not belong to D-plateau, therefore, the rooting $G_{\langle v,b\rangle}$ has at least one more duplication than $G_{\langle v,a\rangle}$. Hence, $G_{\langle v,b\rangle}$ has two duplications in v and in the root. Moreover, the root of $G_{\langle v,a\rangle}$ is not a duplication. However, this is possible only when $\mathsf{root}(T(a))$ and $\mathsf{root}(T(v))$ are mapped below $\mathsf{root}(S)$, thus the $\langle v, a\rangle$ is an empty edge, which is a contradiction with Theorem 2. This completes the first part of the proof.

Next, if the DL-plateau consists of exactly one double edge, then, by Lemma 8 the property holds trivially. Now, we assume that the DL-plateau has more than one edge. We show that every internal node v of the DL-plateau is a super-duplication. From Theorem 2 we know that v is incident to at least two double edges. Hence, in any rooting at least one of its children is mapped to $\mathsf{root}(S)$. We conclude that v is a duplication mapped to $\mathsf{root}(S)$.

Let us consider a path $p = v_1, v_2, \ldots, v_n$ $(n > 1)$ connecting an internal node v_1 from the DL-plateau with a leaf v_n from the D-plateau. We show that the first $n-1$ nodes on p are duplications for every rooting placed on this path. It follows from the first part of this proof that v_1 is a super-duplication mapped to $\mathsf{root}(S)$. Hence, when rooting at $\langle v_{n-1}, v_n\rangle$, we have n gene duplications: for $v_1, v_2, \ldots, v_{n-1}$ and one for the root. All edges from p are elements of the D-plateau, thus moving the root to other edges on p will preserve the total number of gene duplications.

It should be clear that the same holds when choosing other root positions even outside of the D-plateau (see [Paszek and Górecki, 2016], note that in [Paszek and Górecki, 2016] the super-duplication definition differs from the definition which is used both in [Paszek and Górecki, 2018] and in this dissertation). We omit the details. □

5.3. Solution to SINGLE-UEC under PG

First, we analyze the case when unrooted tree has an empty edge. Then, we focus on a complementary case when double edge is present. Next, we present the solution to UEC for a single unrooted gene tree. Finally, we describe algorithms for solving UEC for multiple input trees.

5.3.1 Episodes in a gene tree with an empty edge

In this Section we solve SINGLE-UEC problem for the case when the input gene tree has one empty edge.

Let v be a center of the star that contains the only DL-plateau edge in a gene tree G. This star induces three rooted subtrees T_a, T_b and T_c rooted at neighbors a, b and c, respectively, as indicated in Figure 2.5. Let $\mathbb{1}$ be the indicator function, that is, $\mathbb{1}(p)$ is 1 if p is satisfied and 0 otherwise.

Lemma 10. *Let $a_0, a_1, a_2, \ldots, a_{n+1}$ (for $n \geq 0$) be the path of D-plateau nodes connecting $v = a_0$ and $a_{n+1} \in T_a$ in G. Let G_n be the D-minimal rooting induced by the edge $\langle a_n, a_{n+1}\rangle$. If $e_* = \langle v, c\rangle$ is empty then*

$$\mathsf{REC}(G_n) = \mathsf{REC}(T_1, T_2, \ldots, T_{n+1}, T_b, T_c) + \mathbb{1}(\mathsf{root}(T_i) \notin \mathsf{Dup}(G_n) \text{ for all } i),$$

where $T_1, T_2, \ldots, T_{n+1}$ are subtrees of T_a such that $T_a = (T_1, (T_2, \ldots, (T_n, T_{n+1}) \ldots))$ and the root of T_{n+1} is a_{n+1} (see Figure 2.5 and Figure 5.1).

Proof. First we show that v is a speciation node in G_n. It follows from the fact that v is a center of S2 star and $\langle v, b\rangle$ is single. Thus, $\mathsf{M}_n(v) = \mathsf{root}(S)$, $\mathsf{M}_n(c) \prec \mathsf{root}(S)$ and $\mathsf{M}_n(b) \prec \mathsf{root}(S)$, where M_n is the lca-mapping for G_n. From the fact that $\mathsf{M}_n(v) = \mathsf{root}(S)$ we conclude that all nodes on the path connecting the parent of v with the root in G_n are mapped to $\mathsf{root}(S)$, therefore, they are duplications.

Lets consider the number of duplication clusters in G_n. We have the $\top$-cluster composed of the duplication nodes $a_1, a_2, \ldots, a_n, \mathsf{root}(G_n)$ mapped to $\mathsf{root}(S)$. Both T_c and T_b in G_n are under speciation node v so their clusters are disjoint with the $\top$-cluster. Finally, if the root of some T_i is a duplication then its cluster can be merged with the $\top$-cluster. Therefore, the $\top$-cluster contributes to $\mathsf{REC}(G_n)$ only if the root of T_i is a speciation for every i. Now, it is easy to conclude the final formula. □

Lemma 11. *Under the assumptions from the previous lemma, we have*

$$\mathsf{REC}(G_n) = \mathsf{REC}(G_*) + \mathbb{1}(b \in \mathsf{Dup}(G_*) \text{ and } \mathsf{root}(T_i) \notin \mathsf{Dup}(G_*) \text{ for all } i),$$

where G_* is the rooting induced the empty edge $e_* = \langle v, c\rangle$ (see Figure 5.1).

Proof. Both rootings G_n and G_* are D-minimal. Hence, $\mathsf{D}(G_*, S) = \mathsf{D}(G_n, S)$ and, in consequence, the number of duplication nodes in $A = \{a_1, a_2, \ldots, a_n, v, \mathsf{root}(G_*)\}$ in G_* and $B = \{a_1, a_2, \ldots, a_n, v, \mathsf{root}(G_n)\}$ in G_n are equal. It follows from the properties of star S2, that in G_n node v is a speciation mapped to $\mathsf{root}(S)$. Hence, all predecessors of v are duplications in G_n. Thus, we have exactly $n+1$ duplications in B. On the other hand, by star S2, $\mathsf{root}(G_*)$ is a speciation, therefore all remaining nodes in A are duplications.

We conclude that G_n has the $\top$-cluster containing duplications from A, and G_* has a cluster (mapped below $\mathsf{root}(S)$) containing duplications from B, respectively. These two clusters we call **high clusters**. If the root of one of T_i's is a duplication, then it can be merged with the high cluster in both rootings. Otherwise, if every root of these subtrees is a speciation then the high cluster is disjoint with clusters from $T_1, T_2, \ldots, T_{n+1}$. Moreover, if b is a duplication then the high cluster contains b in G_*. However, in G_n the cluster of b will be disjoint with the $\top$-cluster due to the speciation node v. Combining the above observations we obtain our formula. □

Lemma 10 and Lemma 11 complete the case of empty rootings. We proved that rooting on empty edge has the best REC.

5.3.2 Episodes in a gene tree with a double edge

In the next lemma we show that rootings at edges of the D-plateau induce the same REC cost.

Lemma 12. *If an unrooted gene tree G has no empty edge then for any D-minimal rooting of G denoted by G_**

$$\mathsf{REC}(G_*) = \mathsf{REC}(T_1, T_2, \ldots, T_n) + 1,$$

where $T_1, T_2, \ldots, T_n$ are the rooted subtrees of G obtained from G by removing all internal nodes of the D-plateau.

Proof. It follows from Lemma 9 and its proof that all internal nodes of the D-plateau are present in the $\top$-cluster in the clustering with minimal number of clusters. This cluster is separated from other duplication clusters by speciation nodes located on the border of the D-plateau. Thus, the clusters induced by optimal solution to REC for G_* are the clusters induced by optimal solution to REC of $T_1, T_2, \ldots, T_n$ plus the $\top$-cluster. □

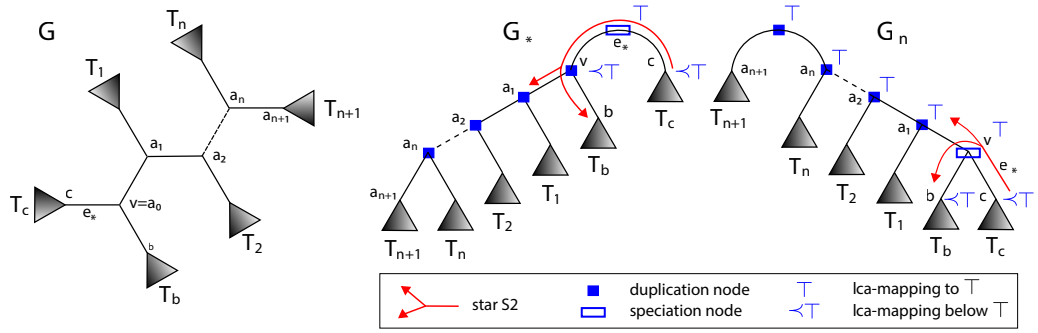

Figure 5.1. Trees from Lemma 10 and 11 (Figure from [Paszek and Górecki, 2016]). A gene tree G (left) and the rootings of G (right) from Lemma 10 and Lemma 11.

Theorem 6 (Solution to SINGLE-UEC)**.** *For any gene tree G, an edge e is optimal for* SINGLE-UEC*, if either e is empty or e is in the D-plateau and G has a double edge.*

Proof. The first part of the proof follows immediately from Lemma 11 and the second part from Lemma 12. □

5.4. Solution to UEC under PG

Now we present solutions to our unrooted episode clustering problem.

Theorem 7. *For a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$, if every gene tree has a double edge then rooting every gene tree on an edge from the D-plateau yields the optimal solution for UEC.*

Proof. Assume that $n = 2$ and let G'_1 and G'_2 be two D-plateau rootings of G_1 and G_2, respectively. It should be clear that $\mathsf{REC}(G'_1, G'_2) = \mathsf{REC}(T)$, where $T = (G'_1, G'_2)$. Next, by Lemma 12, $\mathsf{REC}(T)$ is independent on the choice of rooting of G_1 and G_2, as long as the rootings are in the D-plateau. Therefore, we conclude that $\mathsf{REC}(T)$ is the solution to UEC Problem for G_1 and G_2. This observation can be easily generalized by induction to any n. □

Note that we cannot generalize the property stated in Theorem 7 to gene trees with empty edges. The example is shown in Figure 5.2. Consider the dataset $\{G_1, G_2\}$. G_1 has five D-minimal rootings, while G_2 has exactly one. In G_{2*} we have one $\top$-cluster, therefore G_{2*} with G_{1*}, i.e., the empty edge rooting of G_1, have two duplication clusters. However, the best clusterings for $\{G_1, G_2\}$ having exactly one cluster are obtained for $G_{1,1}$, $G_{1,2}$ or $G_{1,3}$. On the other hand, the best clusterings can be also obtained for empty edge rootings, e.g. $\{G_{1,*}, G_{4,*}\}$ with cost 2 for the input $\{G_1, G_4\}$. From these examples, we see that the empty edges have different properties than double edges in the context of UEC, and we cannot generalize Theorem 7 to empty edges.

Theorem 8 (Candidate rootings for UEC)**.** *For a collection of unrooted gene trees $\mathcal{G}$, the solution to UEC is induced by a rooting edge e of $G \in \mathcal{G}$ satisfying:*

(U1) if G has a double edge, then e is any D-minimal edge in G,

(U2) if G has an empty edge, then e is an element of star S2.

Proof. If some $G \in \mathcal{G}$ has a double edge then the property follows from Theorem 7 and Lemma 12. For gene trees with an empty edge e_* we show that any D-minimal rooting of the edge that is not adjacent to e_* can be equivalently replaced by a rooting adjacent to e_*. By using the notation from Figure 2.5, let $T_a = (T_{a'}, T_{a''})$ such that a' and a'' are the roots of $T_{a'}$ and $T_{a''}$, respectively. We show that the rooting $G_{\langle v,a\rangle}$ denoted by G_a (see Figure 5.3) has the same duplication episodes as the rooting $G_{a'}$ obtained for the edge $\langle a, a'\rangle$. In both rootings v is a speciation, therefore the structure of clusters present in T_b and T_c is the same in both rootings. The edge $\langle v, a\rangle$ is a-incoming, thus the roots are duplications mapped to $\mathsf{root}(S)$. From the fact that $\langle a, a'\rangle$ is in the D-plateau we have that a is a duplication. Thus, every root and a induce the $\top$-cluster. Finally, if a'' is a duplication node, then in both rootings it will be a member of the $\top$-cluster. We proved these two adjacent rootings have the structure of clusters. Therefore, it is sufficient to choose the rooting G_a instead of $G_{a'}$. This proof can be naturally extended by induction to any edge from the D-plateau. □

We conclude that for a gene tree G we have at most 5 candidates for rootings. For instance, G_4 has two stars S2 in the D-plateau, therefore we have 5 candidate rootings: the empty edge rooting $G_{4,*}$ and the rootings of adjacent edges $G_{4,1}$, $G_{4,4}$, $G_{4,7}$ and $G_{4,10}$. Note that the clusters from $G_{4,1}$ are equivalent to clusters from $G_{4,2}$ and $G_{4,3}$. Similar property holds for other candidates.

Next, we show that the condition U2 can be improved.

Lemma 13. *Under the assumptions from Theorem 8. Let the set of clusters induced by the solution to UEC contains $\top$-cluster. Then, the condition (U2) from Theorem 8 can be refined as follows:*

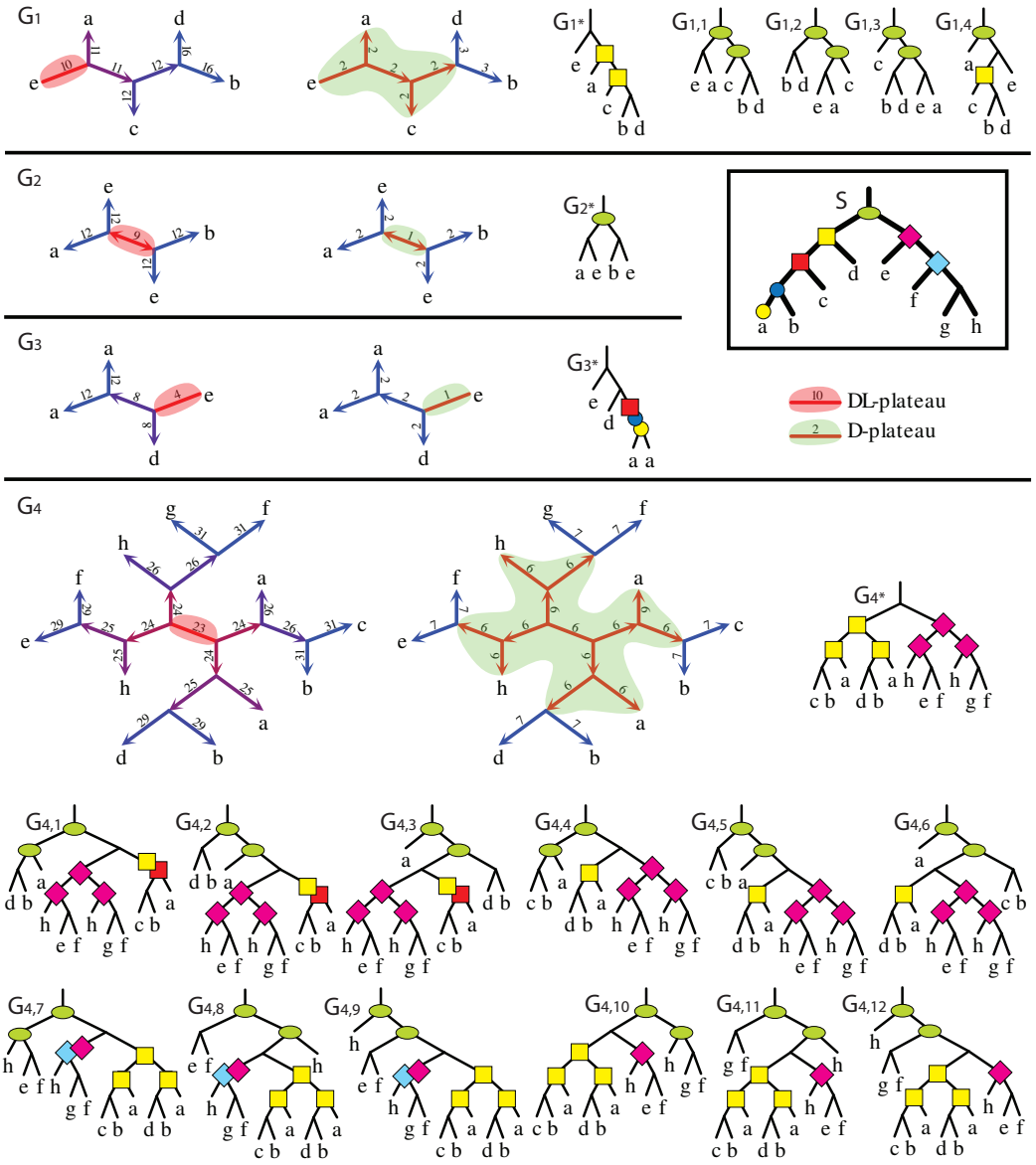

Figure 5.2. An example of unrooted episode clustering (Figure from [Paszek and Górecki, 2016]). A species tree S and four unrooted gene trees G_1, G_2, G_3, G_4 with all D-minimal rootings. For every gene tree two star topologies are shown: one for the duplication-loss cost (left) and one for the duplications cost (right). Every edge of a gene tree is decorated with the corresponding cost of rooting. Every duplication node in rootings of gene trees is decorated by all possible locations (i.e., valid mappings) of its duplication cluster from optimal solutions of SINGLE-UEC. Note that the rooting G_{4*}, whose lca-mappings are shown in Figure 2.2, has two duplications at $(c,(b,a))$ and $(h,(f,g))$ that are raised (here) to create two duplications clusters. Let $\{G_2, G_4\}$ be an instance of UEC Problem. Then, the $\top$-cluster, that is present in G_{2*}, contributes to the optimal solution. In such a case, the solution is induced by one of the two instances of REC Problem: $\{G_{2*}, G_{4,1}\}$ or $\{G_{2*}, G_{4,7}\}$. This property is proved in Theorem 8 and in Lemma 13.

(U2') *if e_* is the empty edge in G, then e is one among at most two non-adjacent edges such that $e = \langle x, y\rangle$ is adjacent to e_* and $\mathsf{M}_*(x) = \mathsf{M}_*(y)$, where M_* is the lca-mapping for G_*.*

Proof. Let G be a gene tree with an empty edge. Let e_a be that edge from (U2'). By using the notation from Figure 5.3, we compare the rooting G_* and $G_{\langle v,a\rangle}$, denoted here by G_a. We have the following clusters in G_*: the cluster C that contains c (if c is a duplication) and the cluster X that contains v (it follows from the proof of Lemma 11 that v is a duplication node). Thus, $X = \{v\} \cup A \cup B$ where A and B denote duplications from T_a and T_b, respectively. Note that C has the same contribution to EC in both rootings, which follows from the property that valid mappings of C are the same in both rootings. In G_a, A is a subset of the $\top$-cluster whose contribution to EC is already incorporated (by the assumption). The node v is a duplication in G_*. Hence, without loss of generality we assume that $M_*(a) = M_*(v)$, i.e., the rooting edge $\langle v, a\rangle$ satisfies the condition from (U2').

We have two cases depending on whether B is empty. If B is empty then G_a has "better" composition of clusters than in G_*, i.e., one cluster less then in G_* and other clusters has the same valid mappings. Otherwise, both rootings are equivalent if $M_*(b) = M_*(v)$ (B in G_a has the same valid mappings as X in G_*), or again G_a has a better structure of clusters than G_* if $M_*(b) \prec M_*(v)$ (valid mappings of X in G_* are included in valid mappings of B in G_a). Similarly, we show that G_a is also better than $G_{\langle v,b\rangle}$ (see also rootings of G_4 in Figure 5.2).

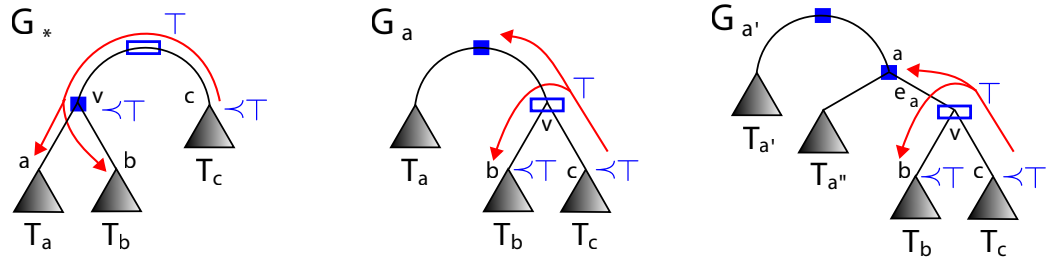

Figure 5.3. Trees from Theorem 8 and Lemma 13 (Figure from [Paszek and Górecki, 2016]). The rootings of G from Theorem 8 and Lemma 13. We use the notation G_a instead of $G_{\langle v,a\rangle}$. See Figure 5.1 for a legend of the symbols used.

We proved that among three rootings from the star S2 we can choose one candidate. The second edge is obtained from the second star S2 (sharing the empty edge) if it is present in the gene tree (see Theorem 2). □

From the last lemma we have at most two candidates for any gene tree from the input collection. For example, the candidate rooting $G_{4,1}$ has more flexible valid mappings than $G_{4,4}$, e.g. the duplication cluster of $((c, b), a)$ in $G_{4,1}$ has larger range of possible mappings than the duplication cluster of $((d, b), a)$ in $G_{4,4}$, while the remaining two clusters have the same locations in the species tree. Hence, for the dataset $\{G_3, G_4\}$, if the $\top$-cluster is present in the solution to UEC, we have two candidates $G_{4,1}$ and $G_{4,7}$ (which is more flexible than $G_{4,10}$). Note, that the clustering cost 3 is obtained by rootings $G_{3,*}$ and $G_{4,1}$ (or $G_{4,2}$, $G_{4,3}$).

5.5. Algorithms for UEC under PG

Algorithm 4 presents the solution to UEC Problem. The correctness of this algorithm follows from Theorem 8 and Lemma 13. Algorithm 4 has two phases. In the first phase for every gene tree a set of candidate rootings is prepared with respect to the conditions (U1) and (U2'). To find optimal rootings we use a linear time algorithm (procedure FindOptEdge) based on greedy descent method that search a double or an empty edge in a gene tree [Górecki and Tiuryn, 2007a]. Based on condition U2', we divide possible solutions into two categories depending on the presence of $\top$-cluster in an optimal clustering. If the $\top$-cluster is not present then every gene tree has an empty edge (in line 10). Otherwise, we check every possible variant of rooting candidates. Note that from Lemma 13, a gene tree has two candidates if and only if the gene tree has two stars S2 that are included in the D-plateau. Thus, the overall time complexity depends on the presence of such trees in the input. From this observation we conclude the following result.

Theorem 9. *The time complexity of Algorithm 4 is $O(2^k(\sum_i |G_i| + |S|))$, where k is the number of input gene trees having two stars S2 that are included in the D-plateau.*

Algorithm 4 Exact solution to UEC

1: **Input** A binary species tree S, a collection of unrooted gene trees $G_1, G_2, \ldots, G_n$.
2: **Output** Minimal $\mathsf{REC}(T_1, T_2, \ldots, T_n, S)$ in the set of all rootings T_i of G_i such that T_i is a rooting obtained from G_i by placing the root on the edge from the D-plateau.
3: **For** every i compute the set of candidate rooting edges R_i:
4: e_*:=FindOptEdge(G_i)
5: **If** e_* is double: $R_i := \{e_*\}$
6: **If** e_* is empty **Then For** $x \in e_*$ such that x is not a leaf.
7: **Let** c be a child of x in G_* such that $\langle x, c\rangle$ is D-minimal and not adjacent to any edge from R_i and $M_*(c) = M_*(x)$
8: $R_i := R_i \cup \{\langle x, c\rangle\}$.
9: **If** every G_i has an empty edge **Then** $\alpha := \mathsf{REC}(T_1, T_2, \ldots, T_n)$, where T_i is the empty edge rooting of G_i **Else** $\alpha := +\infty$.
10: $\beta = \min_{e_i \in R_i} \mathsf{REC}(G_{e_1}, G_{e_2}, \ldots, G_{e_n})$.
11: **Return** $\min\{\alpha, \beta\}$.
12: **Function FindOptEdge(G)**
13: **Let** $m_{x,y} = M_{G_{\langle x,y\rangle}}(x)$
// can be computed in $O(|G|)$ steps [Górecki and Tiuryn, 2007a].
14: **Let** v be a node from V_G and **Let** $\mathsf{root}(S)$ be the lca-mapping of some rooting of G.
15: **While** there exists a node w adjacent with v such that $m_{w,v} = \mathsf{root}(S) \neq m_{v,w}$
16: **Do:** set $v := w$ (star S1).
17: **Return** $\langle v, w\rangle$ such that $\langle v, w\rangle$ an empty or double edge i.e., $m_{v,w} = \mathsf{root}(S) = m_{w,v}$ or $m_{v,w} \neq \mathsf{root}(S) \neq m_{w,v}$.

Thus, from theoretical point of view UEC is fixed parameter tractable. Later we show that k usually represents a small fraction (up to 5%) of the whole input. For

the cases when 2^k is still too large for efficient computation, we propose Algorithm 5, in which we first solve the instance of UEC for the collection of gene trees that have a unique candidate. Clearly, if there are rootings of the whole input that have the same cost, then this cost is optimal. The overall complexity of Algorithm 5 is the same as Algorithm 4, however, for large datasets this strategy appeared to be successful after checking just one additional candidate set (in lines 2-4).

Algorithm 5 Exact solution to UEC (Two step approach)

1: **Input/output** The same as in Algorithm 4.
2: **Let** $\delta = \mathsf{REC}(G'_1, G'_2, \ldots, G'_{n'})$ computed by Algorithm 4, such that $\{G'_1, G'_2, \ldots, G'_{n'}\}$ is the set of all input gene trees having a unique candidate rooting edge (i.e., $|R_i| = 1$).
3: **If** $\delta = \alpha$ **Return** α, where α is from the 9th line of Algorithm 4 computed for the whole input.
4: **For** every $e_1 \in R_1, e_2 \in R_2, \ldots, e_n \in R_n$
(i.e. candidate rootings of the whole input)
5: **If** $\mathsf{REC}(G_{e_1}, G_{e_2}, \ldots, G_{e_n}) = \delta$ **Then Return** δ.
6: **Return** the minimal REC value computed in lines 3 and 6.

5.6. Experimental evaluation

We performed several computational experiments on three empirical datasets (see Section 4.5).

Guigó dataset was evaluated with 71 species trees from [Chang et al., 2013], known to have the total minimal duplication cost. **Génolevures** was paired with two species trees: one from [Dujon, 2006] and the second one having the lowest duplication-loss cost computed by Fasturec [Górecki and Eulenstein, 2012b].

We implemented our algorithms and the algorithms for the REC variant of the Problem (based on [Luo et al., 2011]). In our experiments the rooting candidates were used to compare the results for UEC with the model of mappings (for rooted gene trees) proposed in [Bansal and Eulenstein, 2008].

We performed two series of 74 computational experiments, one for our model and one with the model described in [Bansal and Eulenstein, 2008]. The total running time of our program was about 7 minutes on a standard PC workstation. For every dataset we were able to find solutions to UEC by testing at most two rooted instances of input gene trees (see Algorithm 5). The summary of experiments is depicted in Table 5.1.

For the Guigó dataset we found four duplication clusters, while for the rooted model from [Bansal and Eulenstein, 2008] we located five clusters. The difference can be explained by the properties of our model that is more flexible: the input trees are unrooted and the model of valid mappings is more generic. Observe that this dataset has unique rooting candidates ($k = 0$).

Génolevures is the most complex dataset due to its size and potentially large parameter k. Despite these properties, Algorithm 5 located 17 clusters for the filtered input with all unique rooting candidates. In other words, in this filtered dataset a duplication cluster is present in every node of the species tree. Obviously, the whole

input dataset has the same property. The same holds for the model from [Bansal and Eulenstein, 2008].

In TreeFam we located 45 clusters for the filtered dataset with unique rooting candidates. Then, Algorithm 5 found the solution having the same cost for the whole dataset (see Figure 5.4). The same result was obtained for the model from [Bansal and Eulenstein, 2008] (see Table 5.1).

Table 5.1. The experimental results of UEC evaluation

Set	# species trees	# leaves	# gene trees	k	PG Model		GMSModel	
					UEC	% locations	UEC	% locations
Guigó	71	16	53	0	4	12,9 %	5	16,1 %
Génolevures	1 [1]	9	4144	55	17	100 %	17	100 %
	1 [2]	9	4144	156	17	100 %	17	100 %
TreeFam	1	28	1274	67	45	81,8 %	45	81,8 %

5.7. Discussion

In this Chapter we presented the first solution to the open problem of the duplication episode clustering for case when the input collection is composed of unrooted gene trees. By using theoretical properties of the unrooted reconciliation we proved that the problem has nice mathematical and computational properties. From practical point of view, we were able to provide efficient algorithms and tools that were successfully applied to locate duplication clusters in real datasets.

From the computational point of view the complexity of our algorithms depends on the parameter k, i.e., in the worst case REC Problem has to be solved 2^k times in order to find a solution to UEC. Even if k usually represents a small fraction of the whole input it can be still large, e.g. $k > 100$ for the yeast dataset, which may prohibit computation of all possible variants. Here we proposed a solution, that is based on the observation that the clustering induced from the input gene trees having unique candidates (that is, without k gene trees with non-unique variants), usually represents an optimal solution for the whole input. Thus, the strategy that we applied in Algorithm 5, i.e., first cluster easy part and then try to incorporate the hard one by using already identified clusters, appeared to be successful even for potentially complex datasets.

Our computational experiments show that the duplication clusters are usually located in large parts of the species tree especially when the input dataset consists of thousands of gene trees. To provide more detailed information on the duplication clusters, we studied minimal episode problem (UME) which is a natural extension of the episode clustering problem described in Chapter 6.

Our software for solving unrooted episode clustering problem is publicly available at http://www.mimuw.edu.pl/~jpaszek/uec.php.

[1] Tree from [Dujon, 2006]
[2] Tree from [Górecki and Eulenstein, 2012b]

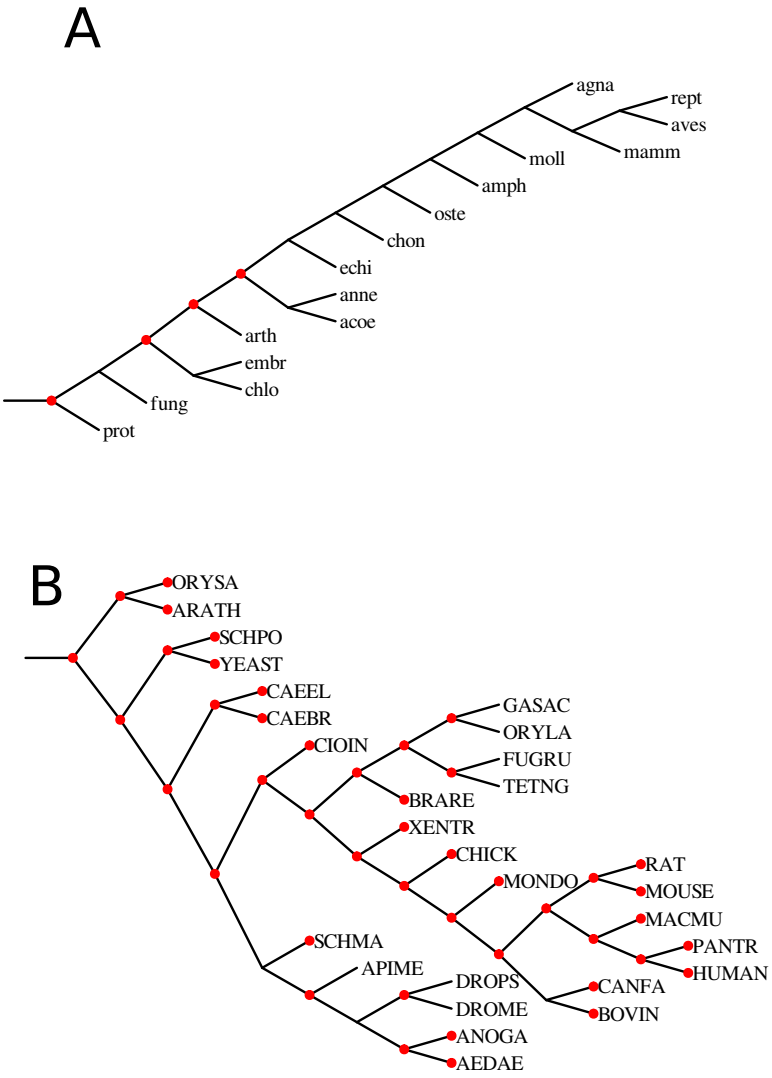

Figure 5.4. Duplication clusters in empirical datasets (Figure from [Paszek and Górecki, 2016]). Duplication clusters (marked by red circles) inferred from experiments. (A) Guigó species tree (chosen from 71 species trees from [Chang et al., 2013] as the most biologically reasonable [Page and Charleston, 1997b]). (B) TreeFam species tree based on NCBI taxonomy.

CHAPTER 6

Unrooted Minimum Episodes

In this chapter we study UME, that is,the multiple genomic problem for unrooted gene trees, clustering called minimum episodes and the model of allowed scenarios that preserves minimal number of single gene duplications. The solution to this problem for the case when input trees are rooted is described in Chapter 4. Section 6.1 presents the definition of the unrooted and general variants of the problem.

New results in the theory of unrooted reconciliation are described in Section 6.2. Here, we expanded the theory of unrooted reconciliation by presenting new properties of the plateau which is the subtree of an unrooted gene tree containing edges whose rootings have the minimal duplication cost. Next, we show that these properties lead to a decomposition of an unrooted gene tree that allows limiting the possible search space significantly.

The solution to the problem is presented in Section 6.4. According to our knowledge, the complexity of UME is unknown. We show that every instance of UME can be transformed into at most 5^k "simpler" instances that can be solved in linear time, where k is bounded above by special cases of S2 stars [Górecki and Tiuryn, 2007a] in input trees. Next, we propose two linear time algorithms for computing bounds of the score. Finally, for the case when k is large, we propose an efficient heuristic algorithm, which in practice allows solving exactly empirical instances consisting of thousands of unrooted gene trees.

The results from experimental evaluation of the implementation of algorithms is in Section 6.5. Our evaluation study on empirical dataset confirmed several genomic duplication events from the literature and demonstrate that algorithms can be successfully applied.

This Chapter is based on [Paszek and Górecki, 2018] and [Paszek and Górecki, 2017b].

6.1. Minimum Episodes Problems

This section contains the formulation of minimum episodes problems. The variant of this problem for rooted gene trees is analyzed in Chapter 4. Here, we define unrooted and general variants.

Let $\mathcal{A}(G, S)$ be the set of all scenarios allowed in PG model, which are scenarios that for a rooted gene tree G and a species tree S have the minimal number of gene duplications (please refer to Section 3.3.2 for model description). In this Chapter every element of $\mathcal{A}(G, S)$ will be referred to as an **allowed scenario**. Now, we formulate the general problem in which the input consists of mixed types of gene trees: rooted and unrooted.

Problem 4 (General Minimum Episodes, GME, under $\mathcal{A}$)**.**
Given a collection of gene trees (rooted or not) $\mathcal{U} = \{U^1, U^2, \ldots, U^n\}$ and a species tree S. Compute ***minimum episodes score*** $\mathsf{RME}(\mathcal{U}, S)$*, or* RME *score, as the minimal value of* $\mathsf{MES}(\{R_i\}_{i=1,2,\ldots,n}, S)$ *in the sets of scenarios R_i such that $R_i \in \mathcal{A}(U^i, S)$ if U^i is rooted or $R_i \in \mathcal{A}(U^i_e, S)$ if U^i is unrooted, where e is a D-minimal edge.*

Observe that we allow only scenarios that preserve the minimal number of gene duplications. We distinguish two variants of GME Problem:

- unrooted minimum episodes (UME) and
- rooted minimum episodes (RME)

in which the instances consist entirely of unrooted and rooted gene trees, respectively. RME Problem has a linear time and space solution described in Chapter 4 and in [Paszek and Górecki, 2017a]. See also [Bansal and Eulenstein, 2008, Luo et al., 2011] for more details on RME Problem.

6.2. New properties of D-plateau nodes

Section 5.2 introduces the definition of super-duplication and super-speciation and analyze unrooted gene tree with double edge. Here, we start with the case when there is an empty edge in an unrooted tree. Recall that, U^* denotes the set of all D-minimal edges (see Section 2.3). We have:

Lemma 14. *Let U be an unrooted gene tree with an empty edge e. A node incident to e is a speciation in U_e if and only if it is a leaf of the D-plateau.*

Proof. We use the notation from Figure 6.1 where e is $\langle v, c\rangle$. We may assume that c is an internal node of U, otherwise, we have a trivial case where c is a leaf in the rooting of U which is a speciation. Thus, we have two S2 stars sharing the empty edge. ($\Leftarrow$) Without loss of generality, we may assume that v is a leaf of U^*. If v is not a speciation in $U_{\langle v,c\rangle}$ then it is a duplication. From the definition of the empty edge the root of $U_{\langle v,c\rangle}$ and node v in $U_{\langle v,a\rangle}$ are speciations. Moreover, the node v in $U_{\langle v,a\rangle}$ is mapped to $\mathsf{root}(S)$ thus the root of $U_{\langle v,a\rangle}$ is a duplication. Both rootings $U_{\langle v,c\rangle}$ and $U_{\langle v,a\rangle}$, have the same number of duplications having the same setting of duplications in subtrees T_a, T_b and T_c as indicated in Figure 6.1. Hence, $\langle v, a\rangle$ is a U^* edge, a contradiction. ($\Rightarrow$) The proof is similar to the first case. □

The conclusion from the above Lemma 14 is that either only empty edge or the whole S2 star is included in the D-plateau (see Section 2.3). Moreover, we can describe the D-plateau having an empty edge by the following lemma:

Lemma 15. *If the unrooted gene tree has an empty edge then every leaf of the D-plateau is a super-speciation, and every internal node of the D-plateau not incident to an empty edge is a super-duplication.*

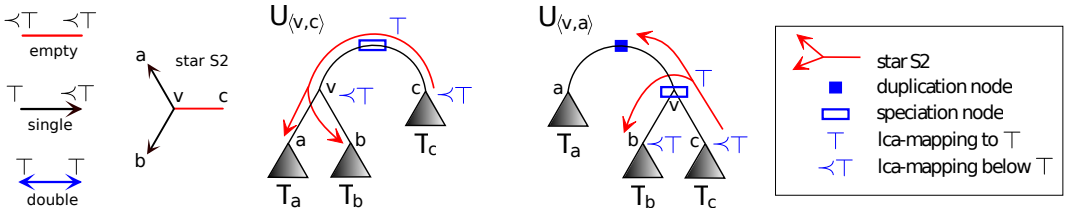

Figure 6.1. Types of edges, star S2, and two rootings of an unrooted gene tree U: on the empty edge $\langle v, c\rangle$ and on the single edge $\langle v, a\rangle$. Here, $\top$ denotes the root of S (Figure from [Paszek and Górecki, 2018]).

Proof. For the first part of the proof, let assume that v is a leaf of U^* which consists of $\langle v, c\rangle$ edge. Assume that v is a duplication in some D-plateau rooting. Then, the subtree T_v in this rooting is also a subtree in all D-plateau rootings because v is a leaf of U^*. Hence, v is a super-duplication. If $\langle v, c\rangle$ is an empty edge we have a contradiction from Lemma 14. Assume that $\langle v, c\rangle$ is non-empty. The edge $\langle v, a\rangle$ does not belong to U^*, therefore, the rooting $U_{\langle v,a\rangle}$ has more duplications than $U_{\langle v,c\rangle}$. Hence, $U_{\langle v,a\rangle}$ has two duplications in v and in the root. Therefore, the root of $U_{\langle v,c\rangle}$ is not a duplication. However, this is possible only when T_a and T_v are mapped below the $\mathsf{root}(S)$, thus the $\langle v, c\rangle$ is an empty edge, a contradiction. For the next part of the proof, if the U^* consists of exactly one empty edge then the property holds trivially. Let assume that the U^* has more than one edge. We show that every internal node v of U^*, that is, not incident to an empty edge is a super-duplication. Let us consider a path $p = v_1, v_2, \ldots, v_n$ $(n > 1)$ consisting of nodes not incident with the empty edge connecting $v = v_1$ with a leaf v_n of U^*. Hence, when rooting on p, v is mapped to $\mathsf{root}(S)$ as it is the ancestor of nodes incident with the empty edge. Moreover, when rooting on $\langle v_{n-1}, v_n\rangle$ we have n gene duplications: for $v_1, v_2, \ldots, v_{n-1}$ and one for the root. All edges from p are elements of U^*, thus moving the root to other edges on p will preserve the total number of gene duplications. We showed that the first $n - 1$ nodes on p are duplications for every rooting placed on this path. If v is incident to an empty edge it is a speciation mapped to the $\mathsf{root}(S)$ when rooting on p. When rooting on an empty edge the root is a speciation. Moreover, from Lemma 14 a child of the root is a duplication if it is an internal node of U^*. Hence, all D-plateau rootings have the same number of duplications equalling the number of internal nodes of U^*. When rooting on an empty edge, the root is a speciation and all internal nodes of U^* are duplications. Otherwise, if we place the root on the edge from U^*, the root is a duplication node and the only speciation is that node among nodes incident to an empty edge which is an ancestor to the other. □

6.3. Unrooted tree decomposition

Now, we show that every unrooted gene tree can be decomposed into a set of trees having at most one unrooted tree with a simplified structure allowing to solve UME in a more efficient way. Please recall that $\mathsf{M}(v)$ is lca-mapping function (see Section 2.1.2), by $S(v)$ we denote the subtree of S rooted at v (see Section 2.1.1), and the function ϕ is defined in Section 2.2.1. We start with the following observation.

Lemma 16. *Let U be an unrooted gene tree and T be a rooted subtree of U rooted at v. Let $X \subseteq U^*$ such that*

- X *is disjoint with* $V_T \setminus \{v\}$,
- v *is a speciation in every scenario from* $\mathcal{A}(U_e, S)$ *for all* $e \in E_X$.

Then, for any set of scenarios $\mathcal{X}$*:*

$$\min_{R \in \mathcal{A}(U_e,S), e \in E_X} \mathsf{MES}(\mathcal{X} \cup \{R\}, S) = \min_{\substack{R' \in \mathcal{A}(U'_e,S), e \in E_X, \\ R'' \in \mathcal{A}(T,S)}} \mathsf{MES}(\mathcal{X} \cup \{R', R''\}, S), \quad (6.1)$$

where U'_e *is the unrooted tree obtained from* U *by replacing* T *with* $S(M(v))$.

Proof. In every allowed scenario R from the left side, $\phi_{U_e}(v)$ is a speciation node. Thus, scenarios R' and R'' can be obtained from R as follows: R'' is the subtree rooted at $\phi_{U_e}(v)$ in R, while R' is obtained from R by replacing the subtree with the copy of $S(\mathsf{M}(v))$, where every internal node is a speciation. Such a transformation is a bijection that preserves the clusterability of duplication nodes. We omit technical details. □

Given a species tree S and a rooted tree G by $\widetilde{G}$ we denote the set of all $\preceq$-maximal elements in the set of all non-root speciation nodes from G. Lets $\sim$ be a relation on edges of U^* for an unrooted gene tree U such that $e \sim e'$ if $\widetilde{U}_e = \widetilde{U}_{e'}$. It should be clear that $\sim$ is an equivalence relation. The set of equivalence classes of this relation we denote by $U^*/\!\sim$. An example is depicted in Figure 6.2.

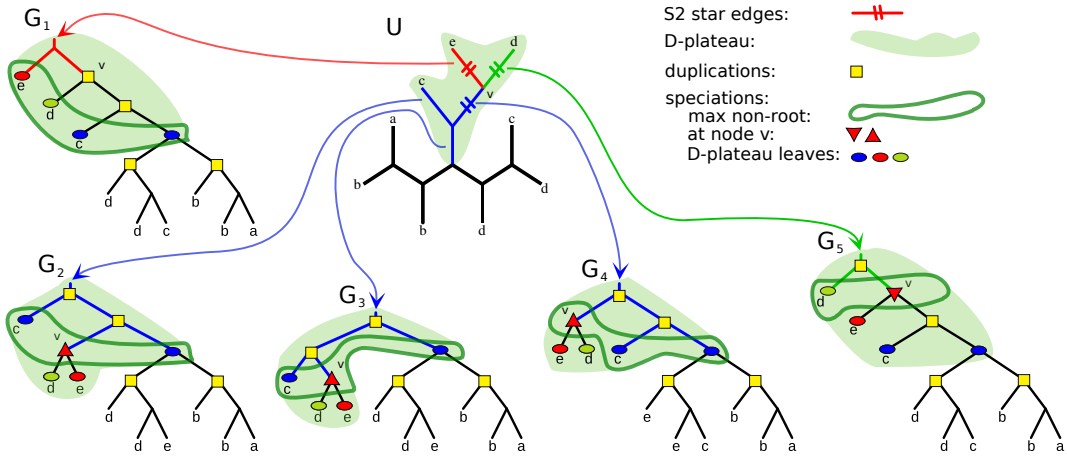

Figure 6.2. **Equivalence relation** $\sim$ (Figure from [Paszek and Górecki, 2018]). An example of an unrooted gene tree U with one S2 star and all D-plateau rootings reconciled with a species tree S=(((a,b),(c,d)),e). U^* contains five edges and induces three $\sim$-equivalence classes. The first consists of an empty edge $\langle e, v\rangle$, the second of $\langle d, v\rangle$ while the last class consists of the remaining three edges. These three classes induce rootings $\{G_1\}$, $\{G_5\}$ and $\{G_2, G_3, G_4\}$, respectively. Observe, that $\widetilde{G}_2 = \widetilde{G}_3 = \widetilde{G}_4$ consist of a subset of U^* leaves and a speciation (different for each class) at node v which is a center of S2 star.

Lemma 17. *If an empty edge is present in an unrooted gene tree then every D-plateau edge present in S2 star uniquely defines one* $\sim$*-equivalence class. Otherwise, the tree has exactly one* $\sim$*-equivalence class.*

Proof. Let U be an unrooted gene tree. We have two cases: (a) either U has a double edge or (b) U has an empty edge. In the case (a), it follows from Lemma 9, that $\widetilde{U}_e$ consists of all U^* leaves for every e from U^*. Thus, we have one equivalence class consisting of all U^* edges. Let use the notation from Figure 6.1. For the case (b), from the proof of Lemma 15 we conclude that for the empty edge $\langle v, c\rangle$ the set $\widetilde{U}_{\langle v,c\rangle}$ consists of all U^* leaves. Moreover, from the conclusion from the proof of Lemma 14, there are $0, 2$ or 4 single edges in U^* present in S2 stars. Let $\langle v, a\rangle$ be such an edge. The set $\widetilde{U}_{\langle v,a\rangle}$ consists of: (a) v which is the root of the subtree $T(v) = (T(b), T(c))$ and thus it is a speciation (it maps to $\mathsf{root}(S)$ and both its children map below the $\mathsf{root}(S)$) and (b) all leaves of U^* present in $T(a)$. From Lemma 15 for every edge e of U^* present in $T(a)$, we have $\widetilde{U}_e = \widetilde{U}_{\langle v,a\rangle}$. Summing up there can be $1, 3$ or 5 $\sim$-equivalence classes uniquely defined by every edge of U^* present in S2 star (see Figure 6.2). □

If an empty edge is an element of a class $X \in U^*/_\sim$, X will be called **plain**. Otherwise, we call X **complex**. Recall, that for the set X of edges of unrooted tree U, by $U|_X$ is the smallest subgraph of U containing all edges from X (see Section 2.3).

Lemma 18. *If $X \in U^*/_\sim$ is complex then the leaves from $U|_X$ are speciations in every tree U_e for every e in X.*

Proof. U has either an empty or a double edge. The leaves of U^* are super-speciations from Lemma 9 and Lemma 15. If U has a double edge, then there is only one $\sim$-equivalence class (Lemma 17) and every leaf v of $U|_X$ is also a leaf in U^*. If U has an empty edge, say e, then there are 0, 2 or 4 classes X disjoint with $\{e\}$. For all of them the set of the leaves of $U|_X$ consists of a subset of the leaves of U^* (disjoint with subsets corresponding to other classes see Figure 6.2) and a node v which is the center of a star S2 and a speciation when rooting on edges from X (see the proof of Lemma 17). □

Definition 6 (Unrooted Decomposition)**.** *Let U be an unrooted gene tree, and $X \in U^*/_\sim$, then:*

- *If X has an empty edge e then $\Delta(\mathrm{U}, \mathrm{X}) = \{U_e\}$.*
- *Otherwise, $\Delta(\mathrm{U}, \mathrm{X})$ is the set of all maximal subtrees T_v of U such that v is a leaf of $U|_X$ and $T_v \cap U|_X = \{v\}$.*

For a complex class X, U^{X} denotes a tree obtained from $U|_X$ by replacing every leaf v with the subtree $S(\mathsf{M}(\mathsf{root}(T_v)))$. For example, for the largest class X from Figure 6.2, we have: $\Delta(\mathrm{U}, \mathrm{X}) = \{c,\ (d, e),\ ((a, b), b), ((c, d), d))\}$ and $\mathrm{U}^{\mathrm{X}} = ((((a, b), (c, d)), e), ((a, b), (c, d)), c)$.

The intuition is that $\Delta(\mathrm{U}, \mathrm{X})$ is the set of rooted trees T induced by X with the following properties: (a) the root of T is a speciation, and (b) T is a subtree present in all rootings induced by X. For example, when we consider an empty class there is only one possible rooting U_e. Hence, $\Delta(\mathrm{U}, \mathrm{X}) = \{U_e\}$. Lemma 18 describes the properties of $\Delta(\mathrm{U}, \mathrm{X})$ for a complex class X. Finally, for an unrooted tree U we have the following formula:

Lemma 19 (Decomposition Lemma)**.** *For a given set of input gene trees $\mathcal{G}$, an input unrooted gene tree U and a species tree S we have,* $\mathsf{GME}(\mathcal{G} \cup \{U\}, S) =$

$$= \min_{X \in U^*/_\sim} \begin{cases} \mathsf{GME}(\mathcal{G} \cup \{U_e\}) & \text{if } X = \{e\} \text{ and } e \text{ is empty,} \\ \min_{e \in X} \mathsf{GME}(\mathcal{G} \cup \{U_e^X\} \cup \Delta(\mathrm{U}, \mathrm{X}), S) & \text{otherwise.} \end{cases}$$

Proof. Let us consider the set of allowed DLS scenarios induced by rootings of edges from each $X \in U^*/\sim$. If X is plain, then the set is $\mathcal{A}(U_e, S)$. If X is complex, then by Lemma 18, X and every leaf v from $U|_X$, satisfies assumptions from Lemma 16. Thus, the subtree of U disjoint with $X \backslash \{v\}$ can be detached and replaced by $S(M(v))$ in U. By Lemma 16 the MES score is preserved. The rest follows by induction on the set of leaves v, where we show that the unrooted tree after all transformations is U^X and the set of detached subtrees is $\Delta(U, X)$. □

6.4. Solution to UME under PG

The linear time algorithm for RME from [Paszek and Górecki, 2017a], which is described in Chapter 4 in Section 4.2 as Algorithm 2, is an essential part of the solution of UME.

Recall that for the input consisting of rooted gene trees, every duplication d is associated with the interval consisting of all possible locations of d in the species tree.

Algorithm 2 is a greedy bottom-up algorithm that iteratively assigns duplications to the top-end of intervals. In every step, it finds the lowest top node s of available intervals and assigns to s all duplications d having $\max \mathsf{I}(d)$ equal to s. Additionally, the algorithm assigns other duplications to s but only if the RME score is not increased, which is controlled by $\lambda(s)$. For details please refer to [Paszek and Górecki, 2017a] or Section 4.2.

6.4.1 Exact solution to UME under PG

A naïve solution to UME is to run RME algorithm for every combination of D-plateau rootings from input gene trees. In many cases the D-plateau can be large, hence, the time complexity of such a solution is $O(\prod_i |U_i|(\sum_i |U_i| + |S|))$. Here, we propose an algorithm based on Lemma 19 to limit the cases that have to be checked to the number of classes of $\sim$ relation.

Lemma 20 (Correctness of gnaw)**.** *Let U be an unrooted gene tree and X be a complex class. Let $\mathcal{X}_\mathrm{r}$ be a set of rooted gene trees T such that the root of every T is a speciation. Let $\mathsf{me}(u, v) = \langle s, n \rangle$, in a call of* gnaw *with U^X and $\mathcal{X}_\mathrm{r}$, such that v is internal in X. Then,*

- *for every rooting U_e^X such that $e \in X$, and having v below the root, if Algorithm 2 (RME) is executed for $\mathcal{X}_\mathrm{r} \cup \{U_e^X\}$, then v is assigned to a node s and $n = \mathsf{level}_s(d)$,*
- *the call of* gnaw *returns $\min_{e \in X} ME(\mathcal{X}_\mathrm{r} \cup \{U_e^X\})$.*

Proof. First, observe that every call of gnaw satisfies the assumptions (see Definition 6). Assume that $e \in X$. Then, by the properties of a complex class X, we have in U_e^X that the root and all internal nodes of X, are duplications, while all leaves of X are speciations. Let X'_e be the set of duplication nodes from X including the root. Thus, for every $d \in X'_e$, we have $\mathsf{I}(d) = \langle M_e(v), \mathsf{root}(S) \rangle$, where M_e is the lca-mapping from U_e^X to S. Hence, all duplications from $\mathcal{X}_\mathrm{r}$ have the top interval node below the root, therefore, if Algorithm 2 (RME) would be called with

the input consisting of $\mathcal{X}_{\mathrm{r}} \cup \{U_e^X\}$, then, for v being the root of S (in line 2 of Algorithm 2), all $\mathcal{X}_{\mathrm{r}}$ duplications are already processed. Additionally, a duplication d from X'_e can be assigned earlier to a node $v \succeq M_e(d)$ only in step 5, if the condition is satisfied. Thus, we can separate the process of RME computation for $\mathcal{X}_{\mathrm{r}}$ (line 7 of Algorithm 6) and the rootings of U^X. Furthermore, processing U^X can be done collectively for all rootings from X, by using a dynamic programming that jointly executes the assignment operation. Note, that in line 11 the first elements of $\mathsf{me}(x,u)$ and $\mathsf{me}(y,u)$ are comparable (i.e., u is a duplication), therefore, max is well defined by using lexicographical order. The proof of the first part follows by induction, in which a node in a rooted subtree of U^X is assigned to the first next free "slot" in a species node. Such a slot can be located by using next. When all slots of non-root nodes are occupied then duplications have to be assigned to the root. Such assignments create new episode events. Thus, the score of every rooting placed on $e = \{u, v\}$ can be easily computed by verifying if such additional episodes were created. This information is stored for the two subtrees of the root in $\mathsf{me}(u,v)$ and $\mathsf{me}(v,u)$, respectively, i.e., if $\mathsf{me}(u,v) = \langle \mathsf{root}(S), n\rangle$, then n additional episodes are required. This value for both subtrees is stored in m_e. Note that, max in line 12 is well defined, otherwise, X cannot be complex. Additionally, the root of U_e^X creates one more episode. Therefore, the score returned by gnaw consists of r (from rooted trees), the minimal value of m_e (the contribution of X) and 1 (the root duplication). An example is depicted in Figure 6.3. □

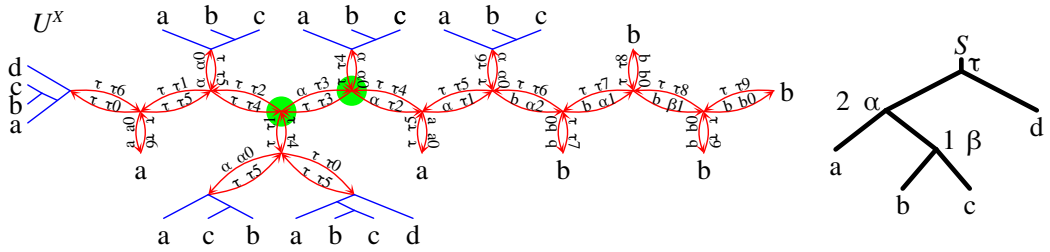

Figure 6.3. Illustration of gnaw for U with a double edge (Figure from [Paszek and Górecki, 2018]). Here, τ denotes the root of S. Assume that S has two positive lambda's computed in line 7 of Algorithm 6: $\lambda(\alpha) = 2$ and $\lambda(\beta) = 1$. Every edge $e = \langle u, v\rangle$ of the D-plateau is split into two directed edges: $\langle u, v\rangle$ and $\langle v, u\rangle$. Each directed edge $\langle u, v\rangle$ is decorated with the lca-mapping $M_{\langle u,v\rangle}(v)$ and $\mathsf{me}(u,v)$. For example, $\tau 6$ denotes the lca-mapping to τ and $\mathsf{me}(u,v) = \langle \tau, 6\rangle$. Here, gnaw returns $3 + 1 + 3$ induced by the marked edge.

Algorithm 6 Exact solution to UME

1: **Input:** Unrooted gene trees $U_1, U_2, \ldots, U_n$, a species tree S.
Output: $\mathsf{UME}(\{U_1, U_2, \ldots, U_n\}, S)$.
2: **For** every sequence $X_1, X_2, \ldots, X_n$ of classes
from the product $U_1^*/_\sim \times U_2^*/_\sim \times \cdots \times U_n^*/_\sim$:
3: $\mathcal{X}_\mathrm{r} := \bigcup_i \Delta(U_i, X_i)$ and
$\mathcal{X}_\mathrm{u} := \bigcup_i \{U^{X_i} \colon X_i \text{ has no empty edge}\}$
4: $\mathsf{mex} := \max_{U^X \in \mathcal{X}_\mathrm{u}} \mathsf{gnaw}(U^X, \mathcal{X}_\mathrm{r}, S)$
5: **Return** the minimal value of mex computed in the above loop,
where gnaw is defined below:
6: **Function** $\mathsf{gnaw}(U^X, \mathcal{X}_\mathrm{r}, S)$:
7: Compute $r = \mathsf{RME}(\mathcal{X}_\mathrm{r}, S)$ and $\lambda(v)$ for every $v \in S$
by Algorithm 2. // *Solve an instance of* RME
8: **Let** $\lambda(\mathsf{root}(S)) = +\infty$ and $\lambda(v) = 0$ for every
$v \neq \mathsf{root}(S)$ not visited in Algorithm 1 in line 3.
9: **For** every $s \in S$,

$$\textbf{Let } \phi(s) = \begin{cases} \mathsf{root}(S) & \text{if } s = \mathsf{root}(S), \\ \mathsf{par}(s), & \text{if } \lambda(\mathsf{par}(s)) > 0, \\ \phi(\mathsf{par}(s)) & \text{otherwise.} \end{cases}$$

10: **For** every ordered pair $\langle u, v\rangle$ of adjacent nodes in X:
11:

$$\mathsf{me}(u,v) = \begin{cases} \langle M_{\langle u,v\rangle}(v), 0\rangle & u \text{ is a leaf in } X, \\ \mathsf{next}(\max(\mathsf{me}(x,u), \mathsf{me}(y,u))) & u \text{ is internal in X and } \{\text{x,y,v}\} \\ & \text{are all nodes adjacent to } u, \end{cases}$$

$$\text{where } \mathsf{next}(s,n) = \begin{cases} (s, n+1) & \text{if } n < \lambda(s), \\ (\phi(s), 1) & \text{otherwise.} \end{cases}$$

12: **For** $e = \{u, v\} \in X$,
$m_e := \max\{n \colon \text{for } \langle s, n\rangle \in \{\mathsf{me}(u,v), \mathsf{me}(v,u)\} \text{ such that } s = \mathsf{root}(S)\}$
13: **Return** $r + 1 + \min_{e \in X} m_e$ // *End of* gnaw

Lemma 21 (Correctness)**.** *Given a collection of unrooted gene trees $\mathcal{U}$ and a species tree S, Algorithm 6 returns* $\mathsf{UME}(\mathcal{U}, S)$.

Proof. The proof follows from Decomposition Lemma 16 and Lemma 20. □

Lemma 22 (Complexity of Exact UME)**.** *Algorithm 6 requires $O((|S| + \sum_i |U_i|)5^k)$ time and $O(\sum_i |U_i| + |S|)$ space, where k is the number of gene trees with S2 star having more than one class of $U^*/_\sim$.*

Proof. Time: The number of iterations of the main loop is bounded above by 5^k. Locating classes of $\sim$ and transforming trees can be done in linear time. Each call of function gnaw requires $O(\sum_{T \in \mathcal{X}_\mathrm{r}} |T| + |U^X|)$ time. *Space:* It follows from the complexity of Algorithm 2 and gnaw. □

6.4.2 Heuristics for UME under PG

In this section, we propose several alternative solutions to our problem designed to cope with hard instances of ME Problem. For example, when the input consists of thousands of trees, it is more likely that k is large enough (e.g., for $k \geq 20$) to prohibit applications of Algorithm 6.

The first approach, presented in Algorithm 7 and Algorithm 8, is to decrease the search space by introducing the lower and upper bounds on the optimal solution in a similar way that we proposed in [Paszek and Górecki, 2017a]. In these algorithms we define function gnawrooting, being a variant of gnaw from Algorithm 6, that instead of the minimal score it returns the corresponding D-minimal rooting of the input gene tree.

Algorithm 7 Lower Bound of UME score

1: **Input:** see Algorithm 6.
   Output: a lower bound of $\mathsf{UME}(\{U_1, U_2, \ldots, U_n\}, S)$.
2: **Function** gnawrooting(U^X, S):
   // Assumption: λ and ϕ are computed.
3: Execute lines 10-12 from Algorithm 6.
4: **Return** one element from $\arg\min_{e \in X} m_e$
5: **End of Function**
6: $\mathcal{X}_\mathrm{r} := \emptyset$
7: **For** U in $\{U_1, U_2, \ldots, U_n\}$:
8: **If** $U^*/_\sim$ consist of a single class X **Then**
   $\mathcal{X}_\mathrm{r} := \mathcal{X}_\mathrm{r} \cup \Delta(U, X)$
   If X is not an empty class **Then** $\mathcal{X}_\mathrm{u} := \mathcal{X}_\mathrm{u} \cup \{U^X\}$
9: **Else**
   Add to $\mathcal{X}_\mathrm{r}$ all maximal rooted subtrees obtained
   from U by removing all internal nodes of U^*
10: Given $\mathcal{X}_\mathrm{r}$ and S compute λ and ϕ (the lines 7-9 of Algorithm 6).
11: **For** U in $\mathcal{X}_\mathrm{u}$:
   $e =$ gnawrooting(U, S)
   $\mathcal{X}_\mathrm{r} := \mathcal{X}_\mathrm{r} \cup \{U_e\}$
12: **Return** $\mathsf{RME}(\mathcal{X}_\mathrm{r}, S)$ *// Solve an instance of* RME *by Algorithm 2*

Lemma 23. *Algorithm 7 computes the lower bound of ME score in $O(|S| + \sum_i |U_i|)$ time and space.*

Proof. Algorithm 7 computes the score from a set of input gene trees as follows. For each gene tree U:

- If $U^*/_\sim$ contains exactly one class then decompose the tree similarly to Algorithm 6, i.e., incorporate all duplications from U into the clustering space.
- Otherwise, ignore every duplication located in the D-plateau. In other words, to preserve all non-D-plateau duplications, it is sufficient to extract all (rooted) subtrees of U obtained from U by removing all internal nodes of the D-plateau.

Having this, we conclude that the size of the clustering computed by Algorithm 7 is less or equal to the size of the clustering from Algorithm 6.

The function gnawrooting processes all edges of the input tree in linear time, thus, the time complexity of the loop from line 11 is equal to $O(\sum_i |U_i|)$. A similar property has the decomposition from lines 7-9. The ME score for rooted trees is computed by Algorithm 2 two times: in line 10 and in line 12. Hence, the time and space complexity of Algorithm 6 is $O(|S| + \sum_i |U_i|)$. □

Lemma 24. *Algorithm 8 computes the upper bound of ME score in $O(|S| + \sum_i |U_i|)$ time and space.*

Proof. Algorithm 8 returns the number of episodes computed for exactly one set of rootings that uniquely corresponds to an element from the product of classes $U_1^*/_\sim \times U_2^*/_\sim \times \cdots \times U_n^*/_\sim$. Hence, this number of episodes is evaluated in max-formula in line 4 of Algorithm 2. Therefore, the ME score computed by Algorithm 2 is bounded above by output of Algorithm 8. The class of the maximal size for a gene tree G can be found in $O(|G|)$ time, therefore, the complexity of the decomposition from lines 3-6 is $O(\sum_i |U_i|)$. □

Algorithm 8 Upper Bound of UME score

1: **Input:** see Algorithm 6.
   Output: an upper bound of $\mathsf{UME}(\{U_1, U_2, \ldots, U_n\}, S)$.
2: $\mathcal{X}_\mathrm{r} := \emptyset$
3: **For** U in $\{U_1, U_2, \ldots, U_n\}$:
4: Let $X \in U^*/_\sim$ be the class having the maximal size
5: $\mathcal{X}_\mathrm{r} := \mathcal{X}_\mathrm{r} \cup \Delta(U, X)$
6: **If** X is not an empty class **Then** $\mathcal{X}_\mathrm{u} := \mathcal{X}_\mathrm{u} \cup \{U^X\}$
7: Execute lines 10-12 from Algorithm 7.

Algorithm 8 is a greedy heuristic in which the method of class selection can be replaced in several ways, e.g., by using a random class, the minimal size class or the class with the minimal value of gnaw. Moreover, it could be further refined to obtain a feasible algorithm similar to one presented in [Paszek and Górecki, 2016].

Finally, we present Algorithm 9. It is a heuristic solution to UME Problem having a quadratic time complexity. Algorithm 9 is designed to utilize the following property: if the input consists of thousands of trees, then it is more likely that clustering of duplications from all non-D-plateau rooted subtrees is sufficient to approximate, or even to provide, the exact ME score. Therefore, Algorithm 9 first solves computationally simple instances of RME extracted from the input gene trees and, then if the solution is not found, it proceeds to complex unrooted parts. In the next Section (see Table 6.1), we observe a surprising performance of Algorithm 9 allowing to solve exactly hard instances containing a large number of complex classes with runtimes counted in seconds. Also, when the 'rooted' part of an instance is small (see the Guigó dataset with 53 trees), the runtime could be much worse than for the large and potentially hard datasets (e.g., Génolevures with 4144 trees).

Lemma 25. *Algorithm 9 is a heuristic solution to UME that runs in* $O((|S| + \sum_i |U_i|)^2)$ *time and* $O(|S| + \sum_i |U_i|)$ *space.*

Proof. The first part of Algorithm 9 consists of two phases. The first phase (lines 10-11) has a linear time complexity (see Lemma 24 and Lemma 23). In the second phase (lines 12-24) it may provide an exact solution in quadratic time due to the calls of gnaw.

In the second part of Algorithm 9, depending on the size of $\mathcal{E}$ it is either computing an exact solution by applying Algorithm 6, or it returns a heuristic solution that has quadratic worst-time complexity. This part of the heuristic is similar to Algorithm 8, however, instead of selecting the largest class we choose the class with the minimal RME score of gnaw output (see line 20).

Observe, that some duplications, which are included in Algorithm 9 in line 12 and corresponding to Algorithm 7 line 9 in Algorithm 9 are included for the second time.

Note, the UME score will remain the same, because all of them have a D-plateau leaf ancestor. □

Algorithm 9 UME Heuristic

1: **Input/Output:** see Algorithm 6.
2: **Function** mixedUME(U, R, S):
3: // *U - unrooted gene trees, R - rooted gene trees*
4: **For** every sequence $X \in U_1^*/_\sim \times U_2^*/_\sim \times \cdots \times U_n^*/_\sim$:
5: $\mathcal{X}_\mathrm{r} := R \cup \bigcup_i \Delta(U_i, X_i)$
6: $\mathcal{X}_\mathrm{u} := \bigcup_i \{U^{X_i} : X_i \text{ has no empty edge}\}$
7: mex $:= \max_{U^X \in \mathcal{X}_\mathrm{u}}$ gnaw$(U^X, \mathcal{X}_\mathrm{r}, S)$
8: **Return** the minimal value of mex computed in the above loop.
9: **End of Function**
10: Compute lower bound (α) and upper bound (β) by Algorithm 7 and Algorithm 8, respectively, for the input U and S.
11: **If** $\alpha = \beta$ **Then Return** α // *Exact solution*
12: Let $\mathcal{X}_\mathrm{r}$ be the set of rooted trees for which ME score is returned by Algorithm 7 when computing α.
13: $\mathcal{E} = \emptyset$ // *a set of unprocessed trees for the exact solution*
14: $\mathcal{H} = \emptyset$ // *a set of pairs (tree, abstract class) for a heuristic*
15: **For** U in $\{U_1, U_2, \ldots, U_n\}$: // *Loop A*
16: **If** $|\ U^*/_\sim\ | > 1$ **Then**
17: $m := -1$ // *minimal* gnaw *value for chosen class*
18: **For** $X \in U^*/_\sim$:
19: $p :=$ gnaw$(U^X, \mathcal{X}_\mathrm{r} \cup \Delta(U, X), S)$;
20: **If** $m = -1$ or $p < m$ **Then** $m := p; Y := X$;
21: **If** $m = \alpha$ **Then** break;
22: **If** $m = \beta$ **Then Return** β // *Exact solution found*
23: **Elif** $m > \alpha$ **Then** $\mathcal{E} := \mathcal{E} \cup \{U\}$; $\mathcal{H} := \mathcal{H} \cup \{(U, Y)\}$
24: **If** $|\mathcal{E}|$ is empty **Then Return** α // *Exact solution found*
25: **If** $|\mathcal{E}| < q$, where q is a small constant (e.g. $q = 10$) **Then**
26: **Return** mixedUME$(\mathcal{E}, \mathcal{X}_\mathrm{r}, S)$ // *Compute exact solution*
27: // *Heuristic solution*
28: **For** every pair (U, X) from $\mathcal{H}$
$\mathcal{X}_\mathrm{r} := \mathcal{X}_\mathrm{r} \cup \Delta(U, X)$
If X is not an empty class **Then** $\mathcal{X}_\mathrm{u} := \mathcal{X}_\mathrm{u} \cup \{U^X\}$
Execute lines 10-12 from Algorithm 7.

Our algorithms are implemented in a prototype computer program written in C++ and python. Additionally, for a more detailed output, all score computing algorithms are extended with a routine for the reconstruction of gene duplication clusters (episodes) with their location in the species tree.

6.5. Experimental evaluation of UME

Datasets properties including the size of classes are depicted in Table 6.1 (see datasets description in Section 4.5).

Table 6.1. Decomposition properties of selected datasets

Dataset	Size	Species tree		1 class		3 classes	5 classes
		name	reference	double edge	empty edge	empty edge	empty edge
Guigó	53	S_1	[Page and Charleston, 1997b]	0	41	12	0
		S_2	[Guigó et al., 1996]	3	38	12	0
TreeFam	1274	NCBI	[Wheeler et al., 2007]	133	611	463	67
Génolevures	4144		[Dujon, 2006]	589	2226	1274	55
			[Page and Charleston, 1997b]	673	2250	1079	142

Note, that hard instances for UEC and UME consist of trees that have two S2 stars. Number of such trees is in the column k in Table 5.1 and the column 5 *classes* in Table 6.1. Values match for the same input sets, which are NCBI species tree with TreeFam gene trees and species tree from [Dujon, 2006] with Génolevures gene trees.

In Table 6.2 we present the UME Problem solutions for selected datasets, that are provided by our tool, which implements algorithms presented in this Chapter. Note, that due to small implementation error, results in partial results of upper bound differ slightly to those presented in [Paszek and Górecki, 2018]. The runtime of implementations of Algorithm 7 and Algorithm 8 is lower than 30 seconds on empirical datasets on standard portable computer. Observe, that *Loop A* in Algorithm 9 can be processed in parallel, therefore, on multiprocessor dedicated servers even for Génolevures calculations might take less than an hour. Note, that in *Loop A* there is no need to check all classes from $U^*/_{\sim}$. The column, labeled *# trees > lower bound*, denotes the trees for which we have to check all classes. The $|\mathcal{E}|$ equal to 0 means that those trees do not increase the lower bound. In summary, to obtain results linear processing of trees in *Loop A* (every step executes linear-time Algorithm 2) is enough.

Table 6.2. UME scores for selected datasets

Dataset	Size	Species tree reference	Lower bound by Alg.7	Upper bound by Alg.8	Loop A in Alg.9 # trees > lower bound	$\lvert\mathcal{E}\rvert$	UME score (exact) by Alg.9
Guigó	53	[Page and Charleston, 1997b]	3	7			5
		[Guigó et al., 1996]	3	6			5
TreeFam	1274	[Wheeler et al., 2007]	227	228	1	0	227
Génolevures	4144	[Dujon, 2006]	100	104	3		
		[Page and Charleston, 1997b]	91	92	1		

Guigó dataset. Multiple gene duplication events were inferred for two species trees: S_1 from [Page and Charleston, 1997b] and S_2 from [Guigó et al., 1996]. The comparison of the results for RME [Paszek and Górecki, 2017a] and Algorithm 6 is shown in Figure 6.4, where the original rooting of each gene tree was ignored in UME.

Génolevures. The gene trees were reconciled with the species trees from [Dujon, 2006] and [Shen et al., 2016]. The summary of duplication episodes found by our algorithms is depicted in Figure 6.5.

TreeFam. The species tree is based on NCBI taxonomy [Wheeler et al., 2007]. The summary of duplication episodes found by our algorithms is depicted in Figure 6.6.

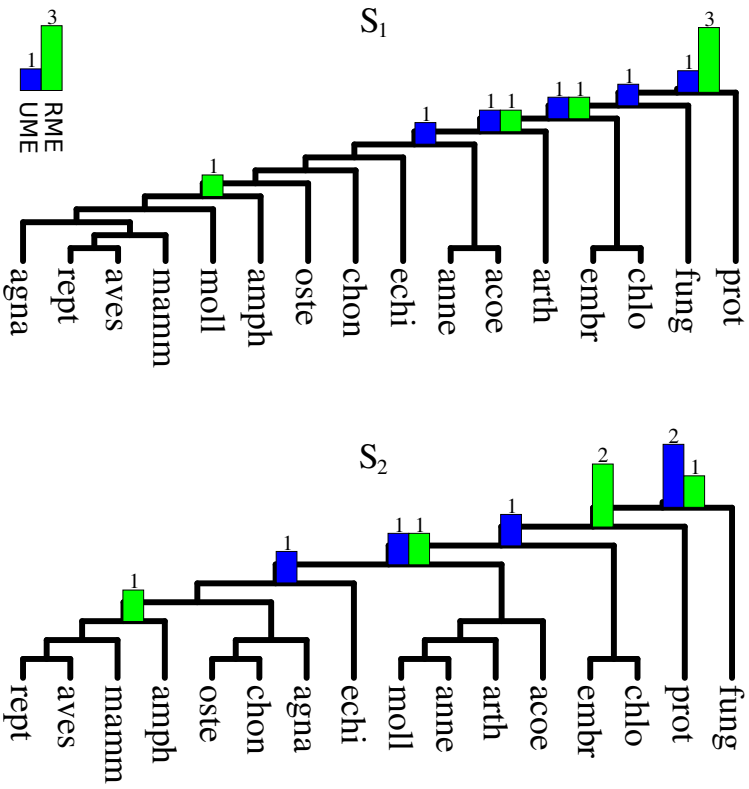

Figure 6.4. Duplication episodes in Guigó dataset [Guigó et al., 1996] inferred by RME [Paszek and Górecki, 2017a] and UME algorithms for the species trees S_1 [Page and Charleston, 1997b] and S_2 [Guigó et al., 1996] (Figure from [Paszek and Górecki, 2018]).

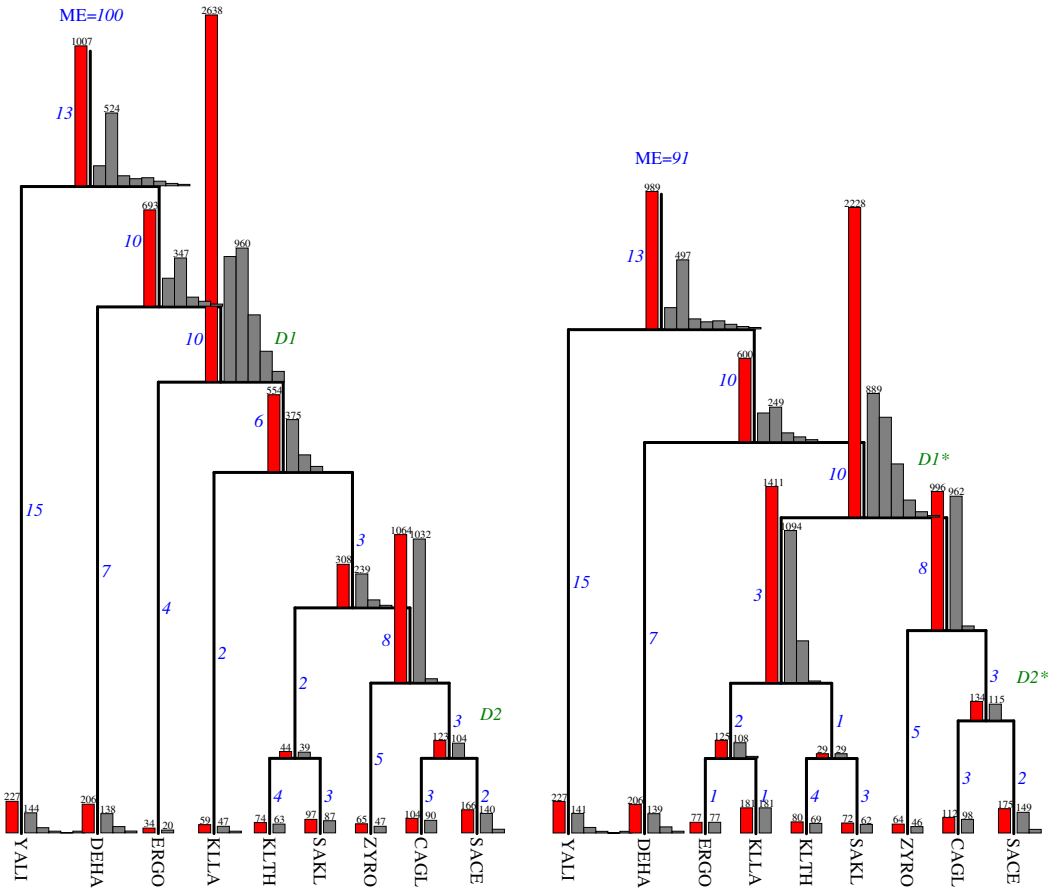

Figure 6.5. Duplication episodes found in Génolevures (Figure from [Paszek and Górecki, 2018]). **Left:** a summary of 100 duplication episodes found in Génolevures dataset [Sherman et al., 2009] by Algorithm 9 for the species trees from [Dujon, 2006]. **Right:** 91 duplication episodes found in the species tree from [Shen et al., 2016]. $D2$ and $D2^*$ denote one whole genome duplication (WGD) event suggested in [Capra et al., 2010, Hudson and Conant, 2012], while $D1$ and $D1^*$ denote one WGD event from [Marcet-Houben and Gabaldón, 2015]. The number of episodes assigned to a single edge is presented on the side (blue italic), for example, our algorithm found 13 duplication episodes in the rooting edge in both trees. A gray histogram (the right side of a node) denotes the frequency of gene trees (Y axis) being involved into exactly x (X-axis starting from 1) episodes located on the corresponding node. The number above the highest bar denotes the maximal number of such gene trees. For example, the gray histogram in the left tree with the second bar of the size 960 denotes that there are 960 gene trees contributing to exactly 2 episodes at the current node. Bars of frequency lower that 10 are not shown. A red bar on the left of a node denotes the number of gene trees having at least one duplication event mapped to this node, i.e., the sum of bars of the corresponding gray histogram.

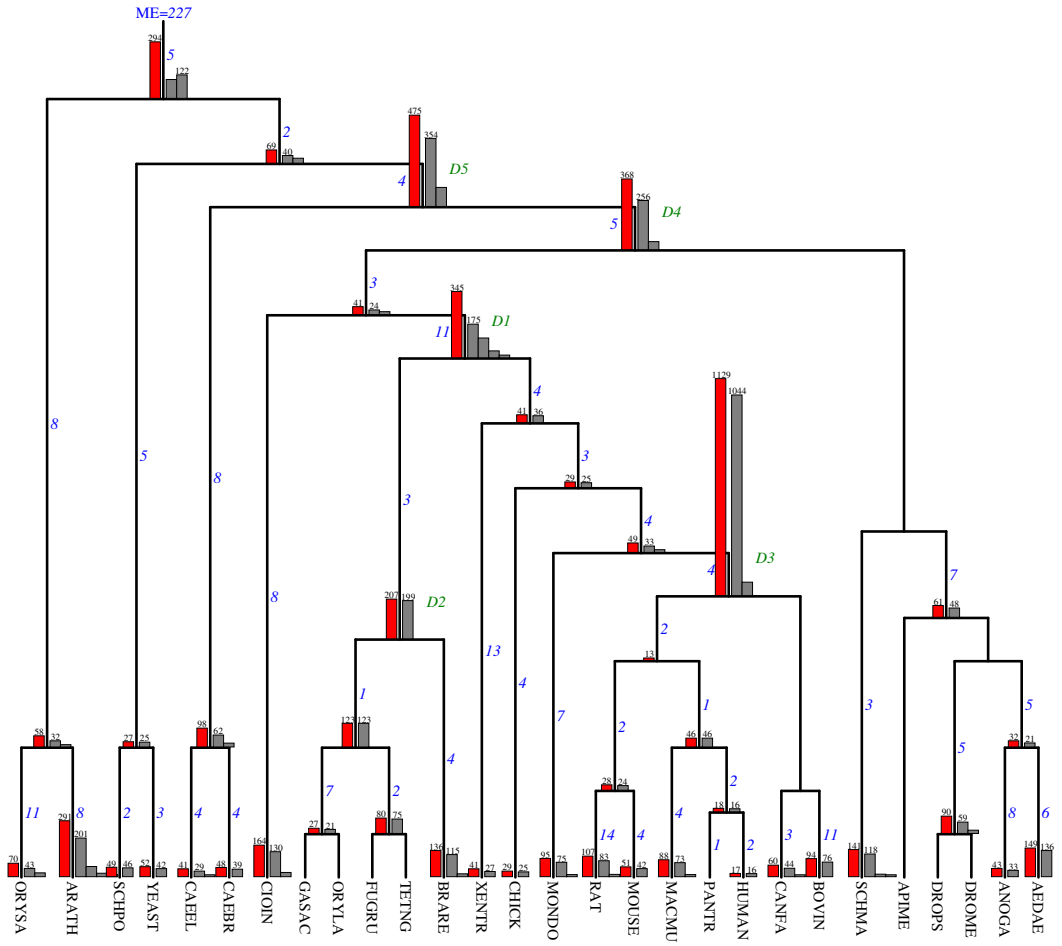

Figure 6.6. 227 duplication episodes found by Algorithm 9 for the TreeFam dataset (Figure from [Paszek and Górecki, 2018]). The upper and lower bounds returned by our algorithms are the same, therefore, 227 is the exact solution. Please refer to Figure 6.5 for the description of numbers and histograms. Two consecutive WGD events at $D1$ and one WGD event at $D2$ are reported in [Hufton et al., 2008, Inoue et al., 2015, Braasch and Postlethwait, 2012].

6.6. Discussion

Guigó dataset: The clustering for the species tree S_1 indicates that UME algorithm found a better scenario than RME, i.e., 5 episodes vs. 6. Additionally, the duplication locations are generally in agreement with the solution to the unrooted variant of episode clustering (see more in [Paszek and Górecki, 2016]). Next, the result of RME for S_2 is consistent with [Guigó et al., 1996, Bansal and Eulenstein, 2008]. However, in [Page and Cotton, 2002] authors suggested a different evolutionary scenario having more duplication episodes. The results differ, i.e., for the gene tree for β-nerve growth factor precursor (NGF) of topology $(rept, (mamm, (amph, aves)))$ in the placement of two duplications inferred by that gene tree and S_2. In the optimal solution from UME algorithm, the rooting of NGF gene tree is $(aves, ((mamm, rept), amph))$ and it infers one duplication with S_2.

Génolevures: We locate two genomic duplication events spanning a large number of gene trees in the left species tree: one situated at $D1$ (2638 trees) and the other above $D2$ (1064 trees). While in the right tree, we have three such events: at $D1^*$ (2228 trees) and the children of $D1^*$. There is a definite correspondence between the events located above $D2$ and $D2^*$. Next, we observe at least 960 trees participating in two duplication clusters at $D1$. Therefore, we postulate that $D1$ has at least two large genomic duplications. Also, they seem to correspond to two events from the right tree located at $D1^*$ and the left child of $D1^*$.

In comparison to the literature, we claim that the peaks at $D1$ and $D1^*$ match the whole genome duplication that was a direct consequence of ancient interspecies hybridization [Marcet-Houben and Gabaldón, 2015]. The location of a WGD event at $D2$ and $D2^*$ [Capra et al., 2010, Hudson and Conant, 2012] is not supported by our analysis. Based on UME clustering, the most likely location of such an event is their parent, i.e., the root of $(ZYRO, (CAGL, SACE))$.

TreeFam: The episode clustering (see Figure 6.6) indicates several genomic duplications located at $D1$, $D2$, $D3$, $D4$ and $D5$. The dataset have only two plant genomes so it is inadequate to study WGD in plants. The same applies to yeasts (2 species), worms (2 species) and insects (6 species). The major part of TreeFam consists of Chordates, for which various studies [Hufton et al., 2008, Inoue et al., 2015, Braasch and Postlethwait, 2012] suggest the existence of two consecutive WGDs located at $D1$ as well as one WGD event at $D2$. Both are partially supported by our analysis by the presence of relatively large number of gene trees contributing to gene duplication events at these two nodes. The genomic duplication at D3 spans almost every tree from the dataset suggesting one WGD event, however, we did not find any evidence of such an event in the literature.

6.7. Conclusions

In this Chapter, we proposed the first solution to the problem of minimum episodes clustering for the case when input gene trees are unrooted. We showed new properties of unrooted reconciliation for the duplication cost. Then, we proposed a decomposition of an unrooted gene tree that allows transforming a gene tree into a set of rooted trees and a simplified unrooted tree. Based on the tree decomposition, we designed several exact and heuristic algorithms for solving the problem. From the application point of view, the most important is an efficient heuristic algorithm, which in practice allows solving exactly empirical instances consisting of thousands of unrooted gene trees. Our evaluation on empirical datasets confirmed several genomic duplication events from the literature.

Future work will focus on the open question of the complexity of UME (we conjecture that UME is intractable). Moreover, we plan to research on the applications of the developed theory to infer genomic duplication events from simulated and empirical datasets of unrooted gene trees including a comparative study of other models of duplication intervals [Paszek and Górecki, 2016].

CHAPTER 7

Conclusions

In this dissertation we present new theoretical results for unrooted reconciliation and apply them to develop solutions to several algorithmic problems.

Moreover, we propose a model of allowed evolutionary scenarios that preserves the minimal number of single gene duplications. We show the biological motivation for the model and present a comparative study with existing models.

We propose the first linear time and space algorithm for RME jointly for any interval model, and the solutions to open problems UME and UEC for unrooted gene trees and under our model.

Popular phylogenetic methods infer unrooted gene family trees, hence, we provide broader applicability of methods that cluster duplications. Moreover, we show that unrooted approach might improve known results on genomic duplication inference from rooted trees.

Our experimental evaluation on biological dataset indicate that we can provide new insights into the genomic duplication inference.

In future, we plan further testing of models of allowed evolutionary scenarios. Our goal is to, with the collaboration of biologists, use the implementations of created algorithms to study multiple genomic duplications like whole-genome duplications. Currently, we can detect whole-genome duplications as well hybridizations.

Bibliography

[Albert et al., 2013] Albert, V. A., Barbazuk, W. B., dePamphilis, C. W., Der, J. P., Leebens-Mack, J., Ma, H., Palmer, J. D., Rounsley, S., Sankoff, D., Schuster, S. C., Soltis, D. E., Soltis, P. S., Wessler, S. R., Wing, R. A., Albert, V. A., Ammiraju, J. S., Barbazuk, W. B., Chamala, S., Chanderbali, A. S., dePamphilis, C. W., Der, J. P., Determann, R., Leebens-Mack, J., Ma, H., Ralph, P., Rounsley, S., Schuster, S. C., Soltis, D. E., Soltis, P. S., Talag, J., Tomsho, L., Walts, B., Wanke, S., Wing, R. A., Albert, V. A., Barbazuk, W. B., Chamala, S., Chanderbali, A. S., Chang, T. H., Determann, R., Lan, T., Soltis, D. E., Soltis, P. S., Arikit, S., Axtell, M. J., Ayyampalayam, S., Barbazuk, W. B., Burnette, J. M., Chamala, S., De Paoli, E., dePamphilis, C. W., Der, J. P., Estill, J. C., Farrell, N. P., Harkess, A., Jiao, Y., Leebens-Mack, J., Liu, K., Mei, W., Meyers, B. C., Shahid, S., Wafula, E., Walts, B., Wessler, S. R., Zhai, J., Zhang, X., Albert, V. A., Carretero-Paulet, L., dePamphilis, C. W., Der, J. P., Jiao, Y., Leebens-Mack, J., Lyons, E., Sankoff, D., Tang, H., Wafula, E., Zheng, C., Albert, V. A., Altman, N. S., Barbazuk, W. B., Carretero-Paulet, L., dePamphilis, C. W., Der, J. P., Estill, J. C., Jiao, Y., Leebens-Mack, J., Liu, K., Mei, W., Wafula, E., Altman, N. S., Arikit, S., Axtell, M. J., Chamala, S., Chanderbali, A. S., Chen, F., Chen, J. Q., Chiang, V., De Paoli, E., dePamphilis, C. W., Der, J. P., Determann, R., Fogliani, B., Guo, C., Harholt, J., Harkess, A., Job, C., Job, D., Kim, S., Kong, H., Leebens-Mack, J., Li, G., Li, L., Liu, J., Ma, H., Meyers, B. C., Park, J., Qi, X., Rajjou, L., Burtet-Sarramegna, V., Sederoff, R., Shahid, S., Soltis, D. E., Soltis, P. S., Sun, Y. H., Ulvskov, P., Villegente, M., Xue, J. Y., Yeh, T. F., Yu, X., Zhai, J., Acosta, J. J., Albert, V. A., Barbazuk, W. B., Bruenn, R. A., Chamala, S., de Kochko, A., dePamphilis, C. W., Der, J. P., Herrera-Estrella, L. R., Ibarra-Laclette, E., Kirst, M., Leebens-Mack, J., Pissis, S. P., Poncet, V., Schuster, S. C., Soltis, D. E., Soltis, P. S., and Tomsho, L. (2013). The Amborella genome and the evolution of flowering plants. *Science*, 342(6165):1241089.

[Albertin and Marullo, 2012] Albertin, W. and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. *Proc. Biol. Sci.*, 279(1738):2497–2509.

[Allen, 2018] Allen, G. E. (2018). *Thomas Hunt Morgan.* In Encyclopaedia Britannica.

[Andersson et al., 1998] Andersson, D. I., Slechta, E. S., and Roth, J. R. (1998). Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. *Science*, 282(5391):1133–1135.

[Anfinsen et al., 1961] Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.*, 47:1309–1314.

[Arvestad et al., 2003] Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, 19 Suppl 1:i7–15.

[Arvestad et al., 2009] Arvestad, L., Lagergren, J., and Sennblad, B. (2009). The gene evolution model and computing its associated probabilities. *J ACM*, 56(2).

[Aury et al., 2006] Aury, J. M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Camara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A. M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepere, G., Malinsky, S., Nowacki, M., Nowak, J. K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Betermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J., and Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, 444(7116):171–178.

[Avery et al., 1944] Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine*, 79(2):137–158.

[Bansal and Eulenstein, 2008] Bansal, M. S. and Eulenstein, O. (2008). The multiple gene duplication problem revisited. *Bioinformatics*, 24(13):i132–8.

[Bansal and Shamir, 2011] Bansal, M. S. and Shamir, R. (2011). A note on the fixed parameter tractability of the gene-duplication problem. *IEEE/ACM Trans Comput Biol Bioinform*, 8(3):848–50.

[Bargiello et al., 1984] Bargiello, T. A., Jackson, F. R., and Young, M. W. (1984). Restoration of circadian behavioural rhythms by gene transfer in Drosophila. *Nature*, 312(5996):752–754.

[Bartel and Unrau, 1999] Bartel, D. P. and Unrau, P. J. (1999). Constructing an RNA world. *Trends Cell Biol.*, 9(12):M9–M13.

[Beadle and Tatum, 1941] Beadle, G. W. and Tatum, E. L. (1941). Genetic Control of Biochemical Reactions in Neurospora. *Proc. Natl. Acad. Sci. U.S.A.*, 27(11):499–506.

[Beale, 2018] Beale, G. H. (2018). *August Weismann.* In Encyclopaedia Britannica.

[Behzadi and Vingron, 2006] Behzadi, B. and Vingron, M. (2006). An improved algorithm for the macro-evolutionary phylogeny problem. In *CPM*, pages 177–187.

[Bender and Farach-Colton, 2000] Bender, M. A. and Farach-Colton, M. (2000). The lca problem revisited. In *Proceedings of the 4th Latin American Symposium on Theoretical Informatics*, LATIN '00, pages 88–94, London, UK, UK. Springer-Verlag.

[Berget et al., 1977] Berget, S. M., Moore, C., and Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mrna. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175.

[Bessman et al., 1956] Bessman, M. J., Kornberg, A., Lehman, I. R., and Simms, E. S. (1956). Enzymic synthesis of deoxyribonucleic acid. *Biochim. Biophys. Acta*, 21(1):197–198.

[Betkier et al., 2015] Betkier, A., Szczęsny, P., and Górecki, P. (2015). Fast algorithms for inferring gene-species associations. *Lecture Notes in Computer Science*, 9096:36–47.

[Blair et al., 2012] Blair, A. C., Blumenthal, D., and Hufbauer, R. A. (2012). Hybridization and invasion: an experimental test with diffuse knapweed (Centaurea diffusa Lam.). *Evolutionary Applications*, 5(1):17–28.

[Blanc and Wolfe, 2004] Blanc, G. and Wolfe, K. H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant cell*, 16(7):1667–78.

[Bonizzoni et al., 2005] Bonizzoni, P., Della Vedova, G., and Dondi, R. (2005). Reconciling a gene tree to a species tree under the duplication cost model. *Theor Comput Sci*, 347(1-2):36–53.

[Bork et al., 2017] Bork, D., Cheng, R., Wang, J., Sung, J., and Libeskind-Hadas, R. (2017). On the computational complexity of the maximum parsimony reconciliation problem in the duplication-loss-coalescence model. *Algorithms Mol Biol*, 12:6.

[Bourque and El-Mabrouk, 2006] Bourque, G. and El-Mabrouk, N., editors (2006). *Comparative Genomics, RECOMB 2006 International Workshop, RCG 2006, Montreal, Canada, September 24-26, 2006, Proceedings*, volume 4205 of *Lect Notes Comput Sc*, Berlin, Germany. Springer.

[Bowers et al., 2003] Bowers, J. E., Chapman, B. A., Rong, J., and Paterson, A. H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–8.

[Braasch and Postlethwait, 2012] Braasch, I. and Postlethwait, J. H. (2012). *Polyploidy in Fish and the Teleost Genome Duplication*, pages 341–383. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Bragonzi et al., 2017] Bragonzi, A., Paroni, M., Pirone, L., Coladarci, I., Ascenzioni, F., and Bevivino, A. (2017). Environmental Burkholderia cenocepacia Strain Enhances Fitness by Serial Passages during Long-Term Chronic Airways Infection in Mice. *Int J Mol Sci*, 18(11).

[Brown, 2002] Brown, T. (2002). *Genomes 2nd edition.* Oxford, United Kingdom: Wiley-Liss. available at https://www.ncbi.nlm.nih.gov/books/NBK21134/.

[Brown, 2009] Brown, T. (2009). *Genomes 3rd edition.* Warszawa, Wydawnictwo Naukowe PWN.

[Burkhardt, 2018] Burkhardt, R. W. (2018). *Jean Baptiste Lamarck*. In Encyclopaedia Britannica.

[Burleigh et al., 2011] Burleigh, J. G., Bansal, M. S., Eulenstein, O., Hartmann, S., Wehe, A., and Vision, T. J. (2011). Genome-scale phylogenetics: inferring the plant tree of life from 18,896 discordant gene trees. *Systematic Biology*, 60:117–125.

[Burleigh et al., 2010] Burleigh, J. G., Bansal, M. S., Eulenstein, O., and Vision, T. J. (2010). Inferring species trees from gene duplication episodes. *ACM BCB*, pages 198–203.

[Burleigh et al., 2008] Burleigh, J. G., Bansal, M. S., Wehe, A., and Eulenstein, O. (2008). Locating multiple gene duplications through reconciled trees. In Vingron, M. and Wong, L., editors, *RECOMB*, volume 4955 of *Lect Notes Comput Sc*, pages 273–284, Berlin, Germany. Springer.

[Cairns et al., 1988] Cairns, J., Overbaugh, J., and Miller, S. (1988). The origin of mutants. *Nature*, 335(6186):142–145.

[Cann et al., 1987] Cann, R. L., Stoneking, M., and Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36.

[Capecchi, 1989] Capecchi, M. R. (1989). Altering the genome by homologous recombination. *Science*, 244(4910):1288–1292.

[Capra et al., 2010] Capra, J. A., Pollard, K. S., and Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biol.*, 11(12):R127.

[Cascalho et al., 1998] Cascalho, M., Wong, J., Steinberg, C., and Wabl, M. (1998). Mismatch repair co-opted by hypermutation. *Science*, 279(5354):1207–1210.

[Chang et al., 2013] Chang, W., Górecki, P., and Eulenstein, O. (2013). Exact solutions for species tree inference from discordant gene trees. *J Bioinform Comput Bio*, 11(5).

[Chargaff et al., 1950] Chargaff, E., Zamenhof, S., and Green, C. (1950). Composition of human desoxypentose nucleic acid. *Nature*, 165(4202):756–757.

[Chen et al., 2006] Chen, D., Eulenstein, O., Fernández-Baca, D., and Burleigh, J. G. (2006). Improved heuristics for minimum-flip supertree construction. *Evolutionary Bioinformatics*, 2:347–356.

[Chen et al., 2000] Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.*, 7(3-4):429–447.

[Cheung and Lau, 2005] Cheung, P. and Lau, P. (2005). Epigenetic regulation by histone methylation and histone variants. *Mol. Endocrinol.*, 19(3):563–573.

[Chicurel, 2001] Chicurel, M. (2001). Genetics. Can organisms speed their own evolution? *Science*, 292(5523):1824–1827.

[Chow et al., 1977] Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8.

[Collins, 2006] Collins, F. (2006). *The Language of God: A Scientist Presents Evidence for Belief*. Free Press.

[Costantino et al., 2014] Costantino, L., Sotiriou, S. K., Rantala, J. K., Magin, S., Mladenov, E., Helleday, T., Haber, J. E., Iliakis, G., Kallioniemi, O. P., and Halazonetis, T. D. (2014). Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science*, 343(6166):88–91.

[Creighton and McClintock, 1931] Creighton, H. B. and McClintock, B. (1931). A correlation of cytological and genetical crossing-over in zea mays. *Proceedings of the National Academy of Sciences*, 17(8):492–497.

[Crick, 1970] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

[Crow, 2005] Crow, J. F. (2005). Timeline: Hermann Joseph Muller, evolutionist. *Nat. Rev. Genet.*, 6(12):941–945.

[Cui et al., 2006] Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., and dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res*, 16(6):738–749.

[Czabarka et al., 2012] Czabarka, E., Székely, L., and Vision, T. (2012). Minimizing the number of episodes and gallai's theorem on intervals. *arXiv:1209.5699*.

[Demuth et al., 2006] 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:e85.

[Desmond, 2018] Desmond, A. J. (2018). *Charles Darwin*. In Encyclopaedia Britannica.

[Dittmar and Liberles, 2011] Dittmar, K. and Liberles, D. (2011). *Evolution after Gene Duplication*. Wiley.

[Dondi et al., 2017] Dondi, R., Mauri, G., and Zoppis, I. (2017). Orthology correction for gene tree reconstruction: Theoretical and experimental results. *Procedia Computer Science*, 108:1115 – 1124.

[Doyon et al., 2009] Doyon, J.-P., Chauve, C., and Hamel, S. (2009). Space of gene/species tree reconciliations and parsimonious models. *J Comput Biol*, 16.

[Doyon et al., 2012] Doyon, J.-P., Hamel, S., and Chauve, C. (2012). An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *IEEE/ACM Trans Comput Biol Bioinform*, 9(1):26–39.

[du Vigneaud et al., 1954] du Vigneaud, V., Ressler, C., Swan, J. M., Roberts, C. W., and Katsoyannis, P. G. (1954). The synthesis of oxytocin. *Journal of the American Chemical Society*, 76(12):3115–3121.

[Duchemin et al., 2017] Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., and Tannier, E. (2017). Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome biology and evolution*, 9(5):1312–1319.

[Dujon, 2006] Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet*, 22(7):375 – 387.

[Dujon and Louis, 2017] Dujon, B. A. and Louis, E. J. (2017). Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina). *Genetics*, 206(2):717–750.

[Durand et al., 2006] 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.

[Ellis and Horvitz, 1986] Ellis, H. M. and Horvitz, H. R. (1986). Genetic control of programmed cell death in the nematode C. elegans. *Cell*, 44(6):817–829.

[Evans and Kaufman, 1981] Evans, M. J. and Kaufman, M. H. (1981). Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156.

[Farris, 1970] Farris, J. S. (1970). Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92.

[Fellows et al., 1998] Fellows, M., Hallet, M., and Stege, U. (1998). On the multiple gene duplication problem. In *9th International Symposium on Algorithms and Computation (ISAAC'98), Lecture Notes in Computer Science 1533*, pages 347–356, Taejon, Korea.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

[Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791.

[Felsenstein, 1989] Felsenstein, J. (1989). PHYLIP. *Cladistics*, 5:164–166. `http://evolution.genetics.washington.edu/phylip.html`.

[Feuk et al., 2006] Feuk, L., Kalervo, A., Lipsanen-Nyman, M., Skaug, J., Nakabayashi, K., Finucane, B., Hartung, D., Innes, M., Kerem, B., Nowaczyk, M. J., Rivlin, J., Roberts, W., Senman, L., Summers, A., Szatmari, P., Wong, V., Vincent, J. B., Zeesman, S., Osborne, L. R., Cardy, J. O., Kere, J., Scherer, S. W., and Hannula-Jouppi, K. (2006). Absence of a paternally inherited FOXP2 gene in developmental verbal dyspraxia. *Am. J. Hum. Genet.*, 79(5):965–972.

[Fire et al., 1998] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811.

[Fischer et al., 2014] Fischer, I., Dainat, J., Ranwez, V., Glémin, S., Dufayard, J.-F., and Chantret, N. (2014). Impact of recurrent gene duplication on adaptation of plant genomes. *BMC Plant Biol*, 14:151.

[Fitch and Markowitz, 1970] Fitch, W. M. and Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem Genet*, 4(5):579–593.

[Flemming, 1882] Flemming, W. (1882). *Zellsubstanz, Kern und Zelltheilung*. Leipzig, F. C. W. Vogel.

[Franklin and Gosling, 1953] Franklin, R. E. and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741.

[Friedman, 2009] Friedman, W. E. (2009). The meaning of Darwin's 'abominable mystery'. *Am. J. Bot.*, 96(1):5–21.

[Fu et al., 2014] Fu, Y., Dominissini, D., Rechavi, G., and He, C. (2014). Gene expression regulation mediated through reversible m^6A RNA methylation. *Nat. Rev. Genet.*, 15(5):293–306.

[Gallardo et al., 1999] Gallardo, M. H., Bickham, J. W., Honeycutt, R. L., Ojeda, R. A., and Kohler, N. (1999). Discovery of tetraploidy in a mammal. *Nature*, 401(6751):341.

[Gallardo et al., 2006] Gallardo, M. H., Gonzalez, C. A., and Cebrian, I. (2006). Molecular cytogenetics and allotetraploidy in the red vizcacha rat, Tympanoctomys barrerae (Rodentia, Octodontidae). *Genomics*, 88(2):214–221.

[Gaut, 2001] Gaut, B. S. (2001). Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.*, 11(1):55–66.

[Gill et al., 1985] Gill, P., Jeffreys, A. J., and Werrett, D. J. (1985). Forensic application of DNA 'fingerprints'. *Nature*, 318(6046):577–579.

[Goodman et al., 1979] 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. *Syst Zool*, 28(2):132–163.

[Gowen et al., 1998] Gowen, L. C., Avrutskaya, A. V., Latour, A. M., Koller, B. H., and Leadon, S. A. (1998). BRCA1 required for transcription-coupled repair of oxidative DNA damage. *Science*, 281(5379):1009–1012.

[Gowen et al., 2003] Gowen, L. C., Avrutskaya, A. V., Latour, A. M., Koller, B. H., and Leadon, S. A. (2003). Retraction. *Science*, 300(5626):1657.

[Greider and Blackburn, 1985] Greider, C. W. and Blackburn, E. H. (1985). Identification of a specific telomere terminal transferase activity in Tetrahymena extracts. *Cell*, 43(2 Pt 1):405–413.

[Griffith, 1928] Griffith, F. (1928). The Significance of Pneumococcal Types. *J Hyg (Lond)*, 27(2):113–159.

[Grunberg-Manago et al., 1955] Grunberg-Manago, M., Oritz, P. J., and Ochoa, S. (1955). Enzymatic synthesis of nucleic acidlike polynucleotides. *Science*, 122(3176):907–910.

[Guerrier-Takada et al., 1983] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–857.

[Guigó et al., 1996] Guigó, R., Muchnik, I. B., and Smith, T. F. (1996). Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol*, 6(2):189–213.

[Gusella et al., 1983] Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., and Sakaguchi, A. Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–238.

[Guyot and Keller, 2004] Guyot, R. and Keller, B. (2004). Ancestral genome duplication in rice. *Genome*, 47(3):610–614.

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

[Górecki et al., 2011] Górecki, P., Burleigh, G. J., and Eulenstein, O. (2011). Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics*, 12 Suppl 1:S15.

[Górecki and Eulenstein, 2011] Górecki, P. and Eulenstein, O. (2011). A linear time algorithm for error-corrected reconciliation of unrooted gene trees. In *ISBRA*, pages 148–159.

[Górecki and Eulenstein, 2012a] 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.

[Górecki and Eulenstein, 2012b] Górecki, P. and Eulenstein, O. (2012b). GTP supertrees from unrooted gene trees: linear time algorithms for nni based local searches. *Lect Notes Comput Sc*, 7292:83–105.

[Górecki and Eulenstein, 2014a] Górecki, P. and Eulenstein, O. (2014a). Drml: probabilistic modeling of gene duplications. *Journal of Computational Biology*, 21(1):89–98.

[Górecki and Eulenstein, 2014b] Górecki, P. and Eulenstein, O. (2014b). Maximizing deep coalescence cost. *IEEE/ACM Trans Comput Biol Bioinform*, 11(1):231–242.

[Górecki and Eulenstein, 2014c] Górecki, P. and Eulenstein, O. (2014c). Refining discordant gene trees. *BMC Bioinformatics*, 15(13):S3.

[Górecki and Eulenstein, 2015] Górecki, P. and Eulenstein, O. (2015). Gene tree diameter for deep coalescence. *IEEE/ACM Trans Comput Biol Bioinform*, 1:155–165.

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

[Górecki et al., 2017a] Górecki, P., Markin, A., Mykowiecka, A., Paszek, J., and Eulenstein, O. (2017a). Phylogenetic tree reconciliation: Mean values for fixed gene trees. *LNCS*, 10330:234–245.

[Górecki et al., 2014a] Górecki, P., Paszek, J., and Eulenstein, O. (2014a). Duplication Cost Diameters. *LNCS*, 8492:212–223.

[Górecki et al., 2014b] Górecki, P., Paszek, J., and Eulenstein, O. (2014b). Unconstrained gene tree diameters for deep coalescence. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 114–121.

[Górecki et al., 2017b] Górecki, P., Paszek, J., and Eulenstein, O. (2017b). Unconstrained Diameters for Deep Coalescence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(5):1002–1012. doi: 10.1109/TCBB.2016.2520937.

[Górecki et al., 2016] Górecki, P., Paszek, J., and Mykowiecka, A. (2016). Mean values of gene duplication and loss cost functions. *Lecture Notes in Computer Science*, 9683:189–199.

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

[Górecki and Tiuryn, 2007a] Górecki, P. and Tiuryn, J. (2007a). Inferring phylogeny from whole genomes. *Bioinformatics*, 23(2):e116–e122.

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

[Haldane, 1933] Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *The American Naturalist*, 67(708):5–19.

[Hallett and Lagergren, 2000] Hallett, M. T. and Lagergren, J. (2000). New algorithms for the duplication-loss model. In *RECOMB*, pages 138–146.

[Hallett and Lagergren, 2001] Hallett, M. T. and Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, RECOMB '01, pages 149–156, New York, NY, USA. ACM.

[Hammond et al., 2012] Hammond, J. I., Jones, D. K., Stephens, P. R., and Relyea, R. A. (2012). Phylogeny meets ecotoxicology: evolutionary patterns of sensitivity to a common insecticide. *Evolutionary Applications*, 5(6):593–606.

[Harding, 1971] Harding, E. F. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):44–77.

[Harding et al., 1997] Harding, R. M., Fullerton, S. M., Griffiths, R. C., Bond, J., Cox, M. J., Schneider, J. A., Moulin, D. S., and Clegg, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.*, 60(4):772–789.

[Harris and Hey, 1999] Harris, E. E. and Hey, J. (1999). X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):3320–3324.

[Harris et al., 2013] Harris, S. R., Cartwright, E. J. P., Török, M. E., Holden, M. T. G., Brown, N. M., Ogilvy-Stuart, A. L., Ellington, M. J., Quail, M. A., Bentley, S. D., Parkhill, J., and Peacock, S. J. (2013). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant staphylococcus aureus: a descriptive study. *Lancet Infect Dis*, 13(2):130–6.

[Harrison, 2017] Harrison, K. (2017). 3dchem.com. http://www.3dchem.com/ vitaminb2.asp.

[Hawgood, 2003] Hawgood, B. J. (2003). Francesco Redi (1626-1697): Tuscan philosopher, physician and poet. *J Med Biogr*, 11(1):28–34.

[Hershey and Chase, 1952] Hershey, A. D. and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, 36(1):39–56.

[Holland et al., 2003] 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.

[Holloway et al., 2013] Holloway, P., Swenson, K., Ardell, D., and El-Mabrouk, N. (2013). Ancestral genome organization: An alignment approach. *Journal of Computational Biology*, 20(4):280–295.

[Hudson and Conant, 2012] Hudson, C. M. and Conant, G. C. (2012). *Yeast as a Window into Changes in Genome Complexity Due to Polyploidization*, pages 293–308. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

[Hufton et al., 2008] Hufton, A. L., Groth, D., Vingron, M., Lehrach, H., Poustka, A. J., and Panopoulou, G. (2008). Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res.*, 18(10):1582–1591.

[Hughes, 2008] Hughes, J. J. (2008). Back to the future. Contemporary biopolitics in 1920s' British futurism. *EMBO Rep.*, 9 Suppl 1:59–63.

[Huxley, 1932] Huxley, A. (1932). *Brave New World*. New York, NY, USA: Harper & Row.

[Inoue et al., 2015] Inoue, J., Sato, Y., Sinclair, R., Tsukamoto, K., and Nishida, M. (2015). Rapid genome reshaping by multiple-gene loss after whole-genome duplication in teleost fish suggested by mathematical modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 112(48):14918–14923.

[Jacob and Monod, 1961] Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356.

[Jeffers, 2018] Jeffers, J. S. (2018). *Frederick Sanger*. In Encyclopaedia Britannica.

[Jeffreys et al., 1985] Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*, 314(6006):67–73.

[Jiao et al., 2012] Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J. E., McKain, M. R., McNeal, J., Rolf, M., Ruzicka, D. R., Wafula, E., Wickett, N. J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter, E. J., Deyholos, M. K., Kutchan, T. M., Chanderbali, A. S., Soltis, P. S., Stevenson, D. W., McCombie, R., Pires, J. C., Wong, G. K.-S., Soltis, D. E., and Depamphilis, C. W. (2012). A genome triplication associated with early diversification of the core eudicots. *Genome biology*, 13(1):R3.

[Jiao et al., 2011] Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., and dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345):97–100.

[Katrib et al., 2016] Katrib, A., Hsu, W., Bui, A., and Xing, Y. (2016). "RADIO-TRANSCRIPTOMICS": A synergy of imaging and transcriptomics in clinical assessment. *Quant Biol*, 4(1):1–12.

[Kellis et al., 2004] Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428:617–624.

[Kelly and Smith, 1970] Kelly, T. J. and Smith, H. O. (1970). A restriction enzyme from Hemophilus influenzae. II. *J. Mol. Biol.*, 51(2):393–409.

[Koonin and Galperin, 2003] Koonin, E. and Galperin, M. (2003). *Sequence - evolution - function: computational approaches in comparative genomics*. Kluwer Academic.

[Koonin, 2012] Koonin, E. V. (2012). Does the central dogma still stand? *Biol. Direct*, 7:27.

[Koonin and Wolf, 2009] Koonin, E. V. and Wolf, Y. I. (2009). Is evolution Darwinian or/and Lamarckian? *Biol. Direct*, 4:42.

[Korber et al., 2000] Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the hiv-1 pandemic strains. *Science*, 288(5472):1789–1796.

[Kornberg, 1974] Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871.

[Kruger et al., 1982] Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E., and Cech, T. R. (1982). Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*, 31(1):147–157.

[Lafond et al., 2015] Lafond, M., Ouangraoua, A., and El-Mabrouk, N. (2015). Reconstructing a supergenetree minimizing reconciliation. *BMC Bioinformatics*, 16(14):S4.

[Lander and Weinberg, 2000] Lander, E. S. and Weinberg, R. A. (2000). Genomics: journey to the center of biology. *Science*, 287(5459):1777–1782.

[Lederberg and Tatum, 1946] Lederberg, J. and Tatum, E. L. (1946). Gene recombination in Escherichia coli. *Nature*, 158(4016):558.

[Lewis, 1978] Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature*, 276(5688):565–570.

[Luo et al., 2011] Luo, C.-W., Chen, M.-C., Chen, Y.-C., Yang, R. W. L., Liu, H.-F., and Chao, K.-M. (2011). Linear-time algorithms for the multiple gene duplication problems. *IEEE/ACM Trans Comput Biol Bioinform*, 8(1):260–265.

[Luria and Delbruck, 1943] Luria, S. E. and Delbruck, M. (1943). Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 28(6):491–511.

[Lynch and Conery, 2000] Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155.

[Lynch and Conery, 2003] Lynch, M. and Conery, J. S. (2003). The evolutionary demography of duplicate genes. *Journal of structural and functional genomics*, 3(1-4):35–44.

[Lyons et al., 2008] Lyons, E., Pedersen, B., Kane, J., and Freeling, M. (2008). The value of nonmodel genomes and an example using synmap within coge to dissect the hexaploidy that predates the rosids. *Tropical Plant Biology*, 1(3):181–190.

[Ma et al., 2000] Ma, B., Li, M., and Zhang, L. (2000). From gene trees to species trees. *SIAM J Comput*, 30(3):729–752.

[Maddison, 1997] Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

[Maere et al., 2005] Maere, S., Bodt, S. D., Raes, J., Casneuf, T., Montagu, M. V., Kuiper, M., and de Peer, Y. V. (2005). Modeling gene and genome duplications in eukaryotes. *PNAS*, 102(15):5454–5459.

[Marcet-Houben and Gabaldón, 2015] Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLOS Biology*, 13(8):1–26.

[Marroquin, 2007] Marroquin, J. (2007). The Language of God: A Scientist Presents Evidence for Belief by Francis Collins. *Proceedings (Baylor University. Medical Center)*, 20(2):198–199.

[Martin and Synge, 1941] Martin, A. J. and Synge, R. L. (1941). A new form of chromatogram employing two liquid phases: A theory of chromatography. 2. Application to the micro-determination of the higher monoamino-acids in proteins. *Biochem. J.*, 35(12):1358–1368.

[McClintock, 1950] McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355.

[McKain et al., 2012] McKain, M. R., Wickett, N., Zhang, Y., Ayyampalayam, S., McCombie, W. R., Chase, M. W., Pires, J. C., dePamphilis, C. W., and Leebens-Mack, J. (2012). Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in agavoideae (asparagaceae). *American journal of botany*, 99(2):397–406.

[Mendel, 1866] Mendel, G. (1866). Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, Bd. IV für das Jahr 1865:Abhandlungen, 3–47.

[Meselson and Stahl, 1958] Meselson, M. and Stahl, F. W. (1958). THE REPLICATION OF DNA IN ESCHERICHIA COLI. *Proc. Natl. Acad. Sci. U.S.A.*, 44(7):671–682.

[Mettanant and Fakcharoenphol, 2008] Mettanant, V. and Fakcharoenphol, J. (2008). A linear-time algorithm for the multiple gene duplication problem. *NCSEC*, pages 198–203.

[Michaelian, 2011] Michaelian, K. (2011). Entropy Production and the Origin of Life. *Journal of Modern Physics*, 2:595–601.

[Miller, 1953] Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529.

[Miller and Urey, 1959] Miller, S. L. and Urey, H. C. (1959). Organic compound synthesis on the primitive earth. *Science*, 130(3370):245–251.

[Mirkin et al., 1995] Mirkin, B., Muchnik, I., and Smith, T. F. (1995). A biologically consistent model for comparing molecular phylogenies. *J Comput Biol*, 2(4):493–507.

[Morgan, 1915] Morgan, T. H. (1915). *The Mechanism of Mendelian Heredity*. New York: H. Holt and company,.

[Muller, 1927] Muller, H. J. (1927). ARTIFICIAL TRANSMUTATION OF THE GENE. *Science*, 66(1699):84–87.

[Mykowiecka and Górecki, 2018] Mykowiecka, A. and Górecki, P. (2018). Credibility of evolutionary events in gene trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. DOI: 10.1109/TCBB.2017.2788888.

[Mykowiecka et al., 2017] Mykowiecka, A., Szczesny, P., and Gorecki, P. (2017). Inferring gene-species assignments in the presence of horizontal gene transfer. *IEEE/ACM Trans Comput Biol Bioinform*.

[Nakhleh, 2013] Nakhleh, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in ecology & evolution*, 28(12):719–728.

[Nik-Zainal et al., 2012] Nik-Zainal, S. et al. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.

[Nirenberg and Leder, 1964] Nirenberg, M. and Leder, P. (1964). RNA CODEWORDS AND PROTEIN SYNTHESIS. THE EFFECT OF TRINUCLEOTIDES UPON THE BINDING OF SRNA TO RIBOSOMES. *Science*, 145(3639):1399–1407.

[No authors listed, 2013] No authors listed (2013). Due credit. *Nature*, 496(7445):270.

[Nøjgaard et al., 2017] Nøjgaard, N., Geiß, M., Merkle, D., Stadler, P. F., Wieseke, N., and Hellmuth, M. (2017). Forbidden time travel: Characterization of time-consistent tree reconciliation maps. In Schwartz, R. and Reinert, K., editors, *17th International Workshop on Algorithms in Bioinformatics, WABI 2017, August 21-23, 2017, Boston, MA, USA*, volume 88 of *LIPIcs*, pages 17:1–17:12. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik.

[Noonan, 2010] Noonan, J. P. (2010). Neanderthal genomics and the evolution of modern humans. *Genome Research*, 20(5):547–553.

[Norman et al., 2015] Norman, R., Weir, H., Bradbury, R., Lawson, M., Rabow, A., Buttar, D., Callis, R., Curwen, J., de Almeida, C., Ballard, P., Hulse, M., Donald, C., Feron, L., Gingell, H., Karoutchi, G., MacFaul, P., Moss, T., Pearson, S., Tonge, M., Davies, G., Walker, G., Wilson, Z., Rowlinson, R., Powell, S., Hemsley, P., Linney, E., Campbell, H., Ghazoui, Z., Sadler, C., Richmond, G., Pazolli, E., Mazzola, A., DCruz, C., and De Savi, C. (2015). A Novel Oral Selective Estrogen Receptor Down-regulator, AZD9496, drives Tumour Growth Inhibition in Estrogen Receptor positive and ESR1 Mutant Models. *PDB ID: 5ACC*.

[Noutahi et al., 2016] 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):1–22.

[Nusslein-Volhard and Wieschaus, 1980] Nusslein-Volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature*, 287(5785):795–801.

[Ohno, 1970] Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag.

[Oparin, 1924] Oparin, A. I. (1924). The Origin of Life. *Moscow Worker Publisher, Moscow*. A. I. Oparin, Translate, "The Origin and Development of Life," NASA TTF-488, Washington DC, 1968.

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

[Page, 2000] Page, R. D. M. (2000). Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Mol Phylogenet Evol*, 14:89–106.

[Page and Charleston, 1997a] Page, R. D. M. and Charleston, M. A. (1997a). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7:231–240.

[Page and Charleston, 1997b] Page, R. D. M. and Charleston, M. A. (1997b). Reconciled trees and incongruent gene and species trees. *DIMACS 96, Mathematical Hierarchies and Biology*, (37):57–70.

[Page and Cotton, 2002] Page, R. D. M. and Cotton, J. A. (2002). Vertebrate phylogenomics: reconciled trees and gene duplications. *Pacific Symposium on Biocomputing*, pages 536–547.

[Page and Holmes, 1998] Page, R. D. M. and Holmes, E. C. (1998). *Molecular evolution: a phylogenetic approach*. Blackwell Science.

[Paszek and Górecki, 2016] Paszek, J. and Górecki, P. (2016). Genomic duplication problems for unrooted gene trees. *BMC Genomics*, 17(1):165–175. doi: 10.1186/s12864-015-2308-4.

[Paszek and Górecki, 2017a] Paszek, J. and Górecki, P. (2017a). Efficient algorithms for genomic duplication models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. doi: 10.1109/TCBB.2017.2706679.

[Paszek and Górecki, 2017b] Paszek, J. and Górecki, P. (2017b). New algorithms for the genomic duplication problem. *In: Meidanis J., Nakhleh L. (eds) Comparative Genomics. RECOMB-CG 2017. Lecture Notes in Computer Science*, 10562:101–115. Springer, Cham.

[Paszek and Górecki, 2018] Paszek, J. and Górecki, P. (2018). Inferring duplication episodes from unrooted gene trees. *BMC Genomics*, 19(5):288. doi: 10.1186/s12864-018-4623-z.

[Pertea and Salzberg, 2010] Pertea, M. and Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biol.*, 11(5):206.

[Ponting, 2008] Ponting, C. P. (2008). The functional repertoires of metazoan genomes. *Nature Reviews Genetics*, 9(9):689–698.

[Prusiner, 1982] Prusiner, S. B. (1982). Novel proteinaceous infectious particles cause scrapie. *Science*, 216(4542):136–144.

[Pääbo, 1999] Pääbo, S. (1999). Human evolution. *Trends in Cell Biology*, 9(12):M13–16.

[Rabier et al., 2014] Rabier, C.-E., Ta, T., and Ané, C. (2014). Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach. *Molecular biology and evolution*, 31(3):750–62.

[Raghavan et al., 2014] Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T. W., Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Magi, R., Campos, P. F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., Osipova, L. P., Fedorova, S. A., Voevoda, M. I., DeGiorgio, M., Sicheritz-Ponten, T., Brunak, S., Demeshchenko, S., Kivisild, T., Villems, R., Nielsen, R., Jakobsson, M., and Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, 505(7481):87–91.

[Rasmussen and Kellis, 2011] Rasmussen, M. D. and Kellis, M. (2011). A bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution*, 28:273–290.

[Ravindran, 2012] Ravindran, S. (2012). Barbara mcclintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences*, 109(50):20198–20199.

[Ray, 2015] Ray, K. (2015). Going against the grain. *Nat Rev Gastroenterol Hepatol*, 12(10):547.

[Reddy et al., 1984] Reddy, P., Zehring, W. A., Wheeler, D. A., Pirrotta, V., Hadfield, C., Hall, J. C., and Rosbash, M. (1984). Molecular analysis of the period locus in Drosophila melanogaster and identification of a transcript involved in biological rhythms. *Cell*, 38(3):701–710.

[Richards et al., 2000] Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Golge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., N?rby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Scozari, R., Torroni, A., and Bandelt, H. J. (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.*, 67(5):1251–1276.

[Riddihough and Zahn, 2010] Riddihough, G. and Zahn, L. M. (2010). Epigenetics. What is epigenetics? Introduction. *Science*, 330(6004):611.

[Rogers, 2018a] Rogers, K. (2018a). *Abiogenesis*. In Encyclopaedia Britannica.

[Rogers, 2018b] Rogers, K. (2018b). *Elizabeth Blackburn*. In Encyclopaedia Britannica.

[Rogers, 2018c] Rogers, K. (2018c). *Michael W Young*. In Encyclopaedia Britannica.

[Rogers and Fridovich-Keil, 2018] Rogers, K. and Fridovich-Keil, J. L. (2018). *Epigenetics*. In Encyclopaedia Britannica.

[Rogers et al., 2018] Rogers, K., Joshi, S. H., and Green, E. R. (2018). *Biology*. In Encyclopaedia Britannica.

[Ronquist and Huelsenbeck, 2003] Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

[Rose and Hildebrand, 2015] Rose, A. S. and Hildebrand, P. W. (2015). NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, 43(W1):W576–579.

[Ruan et al., 2008] 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.

[Ruse, 2018] Ruse, M. (2018). *philosophy of biology*. In Encyclopaedia Britannica.

[Ruvolo, 1997] Ruvolo, M. (1997). Molecular phylogeny of the hominoids: inferences from multiple independent dna sequence data sets. *Molecular Biology and Evolution*, 14(3):248–265.

[Saitou and Nei, 1987] Saitou, N. and Nei, N. (1987). The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Journal of Molecular Biology and Evolution*, 4:406–425.

[Salnikow et al., 1973] Salnikow, J., Liao, T. H., Moore, S., and Stein, W. H. (1973). Bovine pancreatic deoxyribonuclease A. Isolation, composition, and amino acid sequences of the tryptic and chymotryptic peptides. *J. Biol. Chem.*, 248(4):1480–1488.

[Sato et al., 2012] Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C., Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., Ren, Y., Wang, J., Albiero, A., Dal Pero, F., Todesco, S., Van Eck, J., Buels, R. M., Bombarely, A., Gosselin, J. R., Huang, M., Leto, J. A., Menda, N., Strickler, S., Mao, L., Gao, S., Tecle, I. Y., York, T., Zheng, Y., Vrebalov, J. T., Lee, J., Zhong, S., Mueller, L. A., Stiekema, W. J., Ribeca, P., Alioto, T., Yang, W., Huang, S., Du, Y., Zhang, Z., Gao, J., Guo, Y., Wang, X., Li, Y., He, J., Li, C., Cheng, Z., Zuo, J., Ren, J., Zhao, J., Yan, L., Jiang, H., Wang, B., Li, H., Li, Z., Fu, F., Chen, B., Feng, Q., Fan, D., Wang, Y., Ling, H., Xue, Y., Ware, D., McCombie, W. R., Lippman, Z. B., Chia, J. M., Jiang, K., Pasternak, S., Gelley, L., Kramer, M., Anderson, L. K., Chang, S. B., Royer, S. M., Shearer, L. A., Stack, S. M., Rose, J. K., Xu, Y., Eannetta, N., Matas, A. J., McQuinn, R., Tanksley, S. D., Camara, F., Guigo, R., Rombauts, S., Fawcett, J., Van de Peer, Y., Zamir, D., Liang, C., Spannagl, M., Gundlach, H., Bruggmann, R., Mayer, K., Jia, Z., Zhang, J., Ye, Z., Bishop, G. J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, R., Singh, M., Singh, A., Pal, J. K., Pandit, A., Singh, P. K., Mahato, A. K., Gaikwad, V. D., Sharma, R. R., Mohapatra, T., Singh, N. K., Causse, M., Rothan, C., Schiex, T., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P., Vautrin, S., Zouine, M., Latche, A., Rousseau, C., Regad, F., Pech, J. C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., Fernandez del Carmen, A., Monforte, A., Granell, A., Fernandez-Munoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S., Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perotta, G., Daddiego, L., Bryan, G., Orozco, M., Pastor, X., Torrents, D., van Schriek, M. G., Feron, R. M., van Oeveren, J., de Heer, P., daPonte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., van Eijk, M. J., Janssen, A., van Haaren, M. J., Jungeun Kim, S. H., Kwon, S. Y., Kim, S., Koo, D. H., Lee, S., Hur, C. G., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jocker, A., Warfsmann, J., Gobel, U., Kawamura, S., Yano, K., Sherman, J. D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E. G., van de Belt, J., van Haarst, J. C., Peters, S. A., van Staveren, M. A., Henkens, M. H., Mooyman, P. J., Hesselink, T., van Ham, R. C., Jiang, G., Droege, M., Choi, D., Kang, B. C., Kim, B. D., Park, M., Kim, S., Yeom, S. I., Lee, Y. H., Choi, Y. D., Li, G., Gao, J., Liu, Y., Huang, S., Fernandez-Pedrosa, V., Collado, C., Zuniga, S., Wang, G., Cade, R., Dietrich, R. A., Rogers, J., Knapp, S., Fei, Z., White, R. A., Thannhauser, T. W., Giovannoni, J. J., Botella, M. A., Gilbert, L., Gonzalez, R., Goicoechea, J. L., Yu, Y., Kudrna, D., Collura, K., Wissotski, M., Wing, R., Meyers, B. C., Gurazada, A. B., Green, P. J., Vyas, S. M., Solanke, A. U., Kumar, R., Gupta, V., Sharma, A. K., Khurana, P., Khurana, J. P., Tyagi, A. K., Dalmay, T., Mohorianu, I., Walts, B., Chamala, S., Barbazuk, W. B., Li, J., Guo, H., Lee, T. H., Wang, Y., Zhang, D., Paterson, A. H., Wang, X., Tang, H., Barone, A., Chiusano, M. L., Ercolano, M. R., D'Agostino, N., Di Filippo, M., Traini, A.,

Sanseverino, W., Frusciante, L., Seymour, G. B., Elharam, M., Fu, Y., Hua, A., Kenton, S., Lewis, J., Lin, S., Najar, F., Lai, H., Qin, B., Qu, C., Shi, R., White, D., White, J., Xing, Y., Yang, K., Yi, J., Yao, Z., Zhou, L., Roe, B. A., Vezzi, A., D'Angelo, M., Zimbello, R., Schiavon, R., Caniato, E., Rigobello, C., Campagna, D., Vitulo, N., Valle, G., Nelson, D. R., De Paoli, E., Szinay, D., de Jong, H. H., Bai, Y., Visser, R. G., Klein, R., Beasley, H., McLaren, K., Nicholson, C., Riddle, C., and Gianese, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641.

[Schmutz et al., 2010] Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X. C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–183.

[Schrödinger, 1944] Schrödinger, E. (1944). *What is Life?: The Physical Aspect of the Living Cell*. What is Life?: The Physical Aspect of the Living Cell. The University Press.

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

[Scossa et al., 2016] Scossa, F., Brotman, Y., de Abreu E Lima, F., Willmitzer, L., Nikoloski, Z., Tohge, T., and Fernie, A. R. (2016). Genomics-based strategies for the use of natural variation in the improvement of crop metabolism. *Plant Sci.*, 242:47–64.

[Sebat et al., 2004] Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., Zetterberg, A., and Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528.

[Shen et al., 2016] Shen, X. X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., and Rokas, A. (2016). Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3 (Bethesda)*, 6(12):3927–3939.

[Sherman et al., 2009] Sherman, D. J., Martin, T., Nikolski, M., Cayla, C., Souciet, J.-L., and Durrens, P. (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Res*, 37(suppl 1):D550–D554.

[Sjostrand et al., 2014] Sjostrand, J., Tofigh, A., Daubin, V., Arvestad, L., Sennblad, B., and Lagergren, J. (2014). A Bayesian method for analyzing lateral gene transfer. *Syst Biol*, 63(3):409–420.

[Slowinski and Page, 1999] Slowinski, J. and Page, R. D. M. (1999). How should species phylogenies be inferred from sequence data? *Systematic Biology*, 105:147–158.

[Smith and Wilcox, 1970] Smith, H. O. and Wilcox, K. W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J. Mol. Biol.*, 51(2):379–391.

[Smyth et al., 1963] Smyth, D. G., Stein, W. H., and Moore, S. (1963). The sequence of amino acid residues in bovine pancreatic ribonuclease: revisions and confirmations. *J. Biol. Chem.*, 238:227–234.

[Spackman et al., 1958] Spackman, D. H., Stein, W. H., and Moore, S. (1958). Automatic recording apparatus for use in chromatography of amino acids. *Analytical Chemistry*, 30(7):1190–1206.

[Steel and McKenzie, 2001] Steel, M. and McKenzie, A. (2001). Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170(1):91–112.

[Steele and Lloyd, 2015] Steele, E. J. and Lloyd, S. S. (2015). Soma-to-germline feedback is implied by the extreme polymorphism at IGHV relative to MHC: The manifest polymorphism of the MHC appears greatly exceeded at Immunoglobulin loci, suggesting antigen-selected somatic V mutants penetrate Weismann's Barrier. *Bioessays*, 37(5):557–569.

[Stolzer et al., 2012] 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.

[Suarez-Villota et al., 2012] Suarez-Villota, E. Y., Vargas, R. A., Marchant, C. L., Torres, J. E., Kohler, N., Nunez, J. J., de la Fuente, R., Page, J., and Gallardo, M. H. (2012). Distribution of repetitive DNAs and the hybrid origin of the red vizcacha rat (Octodontidae). *Genome*, 55(2):105–117.

[Sutton, 1903] Sutton, W. S. (1903). The chromosomes in heredity. *The Biological Bulletin*, 4(5):231–250.

[Svartman et al., 2005] Svartman, M., Stone, G., and Stanyon, R. (2005). Molecular cytogenetics discards polyploidy in mammals. *Genomics*, 85(4):425–430.

[Takahashi and Yamanaka, 2006] Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676.

[Takeshige et al., 1992] Takeshige, K., Baba, M., Tsuboi, S., Noda, T., and Ohsumi, Y. (1992). Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. *J. Cell Biol.*, 119(2):301–311.

[Tang et al., 2008] Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, 320(5875):486–488.

[Taylor and Raes, 1932] Taylor, J. S. and Raes, J. (1932). The Causes of Evolution. *Ithaca, NY: Cornell Univ. Press.*

[Taylor and Raes, 2004] Taylor, J. S. and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.*, 38:615–643.

[The Editors of Encyclopaedia Britannica, 2018a] The Editors of Encyclopaedia Britannica (2018a). *Hermann Joseph Muller.* In Encyclopaedia Britannica.

[The Editors of Encyclopaedia Britannica, 2018b] The Editors of Encyclopaedia Britannica (2018b). *Natural selection.* In Encyclopaedia Britannica.

[The Editors of Encyclopaedia Britannica, 2018c] The Editors of Encyclopaedia Britannica (2018c). *Spontaneous generation.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018a] The Editors of Encyclopædia Britannica (2018a). *Albrecht Kossel.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018b] The Editors of Encyclopædia Britannica (2018b). *Arthur Kornberg.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018c] The Editors of Encyclopædia Britannica (2018c). *Christiane Nusslein Volhard.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018d] The Editors of Encyclopædia Britannica (2018d). *Edward Lewis.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018e] The Editors of Encyclopædia Britannica (2018e). *Francis Crick.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018f] The Editors of Encyclopædia Britannica (2018f). *George Wells Beadle.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018g] The Editors of Encyclopædia Britannica (2018g). *Joshua Lederberg.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018h] The Editors of Encyclopædia Britannica (2018h). *Severo Ochoa.* In Encyclopaedia Britannica.

[The Editors of Encyclopædia Britannica, 2018i] The Editors of Encyclopædia Britannica (2018i). *Tonegawa Susumu.* In Encyclopaedia Britannica.

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

[Ullmann, 2018] Ullmann, A. (2018). *Louis Pasteur.* In Encyclopaedia Britannica.

[Van de Peer et al., 2009] Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nat Rev Genet*, 10(10):725–732.

[Vandepoele et al., 2003] Vandepoele, K., Simillion, C., and Van de Peer, Y. (2003). Evidence that rice and other cereals are ancient aneuploids. *Plant Cell*, 15(9):2192–2202.

[VanMeter and Hubert, 2015] VanMeter, K. and Hubert, R. (2015). *Microbiology for the Healthcare Professional - E-Book*. Elsevier Health Sciences.

[Vanneste et al., 2014] Vanneste, K., Maere, S., and Van de Peer, Y. (2014). Tangled up in two: a burst of genome duplications at the end of the cretaceous and the consequences for plant evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1648).

[Vischer et al., 1949] Vischer, E., Zamenhof, S., and Chargaff, E. (1949). Microbial nucleic acids; the desoxypentose nucleic acids of avian tubercle bacilli and yeast. *J. Biol. Chem.*, 177(1):429–438.

[Vision et al., 2000] Vision, T. J., Brown, D. G., and Tanksley, S. D. (2000). The origins of genomic duplications in Arabidopsis. *Science*, 290(5499):2114–2117.

[Warszewski, 2014] Warszewski, R. (2014). *Kongo 1965*. Bellona, Warszawa.

[Watson and Crick, 1953] Watson, J. D. and Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

[Wertheim et al., 2013] Wertheim, B., Beukeboom, L. W., and van de Zande, L. (2013). Polyploidy in animals: effects of gene expression on sex determination, evolution and ecology. *Cytogenet. Genome Res.*, 140(2-4):256–269.

[Wheeler et al., 2007] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 35(Database issue):5–12.

[Wilkins et al., 1953] Wilkins, M. H., Stokes, A. R., and Wilson, H. R. (1953). Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–740.

[Winchester, 2018] Winchester, A. (2018). *Genetics*. In Encyclopaedia Britannica.

[Wolfe and Shields, 1997] Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–713.

[Wu et al., 2010] Wu, J., Lu, L. Y., and Yu, X. (2010). The role of BRCA1 in DNA damage response. *Protein Cell*, 1(2):117–123.

[Wu et al., 2013] Wu, Y. C., Rasmussen, M. D., Bansal, M. S., and Kellis, M. (2013). TreeFix: statistically informed gene tree error correction using species trees. *Systematic biology*, 62(1):110–120.

[Wu et al., 2014] Wu, Y. C., Rasmussen, M. D., Bansal, M. S., and Kellis, M. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Res.*, 24(3):475–486.

[Zamenhof et al., 1952] Zamenhof, S., Brawerman, G., and Chargaff, E. (1952). On the desoxypentose nucleic acids from several microorganisms. *Biochim. Biophys. Acta*, 9(4):402–405.

[Zeraati et al., 2018] Zeraati, M., Langley, D. B., Schofield, P., Moye, A. L., Rouet, R., Hughes, W. E., Bryan, T. M., Dinger, M. E., and Christ, D. (2018). I-motif DNA structures are formed in the nuclei of human cells. *Nature Chemistry*.

[Zhang, 2011] Zhang, L. (2011). From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Trans Comput Biol Bioinform*, 8(6):1685–1691.

[Zheng and Zhang, 2014] Zheng, Y. and Zhang, L. (2014). Effect of incomplete lineage sorting on tree-reconciliation-based inference of gene duplication. *IEEE/ACM Trans Comput Biol Bioinform*, 11(3):477–485.

[Zhu et al., 2013] Zhu, Y., Lin, Z., and Nakhleh, L. (2013). Evolution after whole-genome duplication: a network perspective. *G3: Genes, Genomes, Genetics*, 3(11):2049–2057.