Probabilistic Models for Species Tree Inference and Orthology Analysis

IKRAM ULLAH

Doctoral Thesis
Stockholm, Sweden 2015
Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologidoktorsexamen i datalogi fredagen den 12 juni 2015, klockan 13.00 i conference room Air, Scilifelab, Solna.

© Ikram Ullah, June 2015

Tryck: Universitetsservice US AB
To my family
Abstract

A phylogenetic tree is used to model gene evolution and species evolution using molecular sequence data. For artificial and biological reasons, a gene tree may differ from a species tree, a phenomenon known as gene tree-species tree incongruence. Assuming the presence of one or more evolutionary events, e.g., gene duplication, gene loss, and lateral gene transfer (LGT), the incongruence may be explained using a reconciliation of a gene tree inside a species tree. Such information has biological utilities, e.g., inference of orthologous relationship between genes.

In this thesis, we present probabilistic models and methods for orthology analysis and species tree inference, while accounting for evolutionary factors such as gene duplication, gene loss, and sequence evolution. Furthermore, we use a probabilistic LGT-aware model for inferring gene trees having temporal information for duplication and LGT events.

In the first project, we present a Bayesian method, called DLRSOrthology, for estimating orthology probabilities using the DLRS model: a probabilistic model integrating gene evolution, a relaxed molecular clock for substitution rates, and sequence evolution. We devise a dynamic programming algorithm for efficiently summing orthology probabilities over all reconciliations of a gene tree inside a species tree. Furthermore, we present heuristics based on receiver operating characteristics (ROC) curve to estimate suitable thresholds for deciding orthology events. Our method, as demonstrated by synthetic and biological results, outperforms existing probabilistic approaches in accuracy and is robust to incomplete taxon sampling artifacts.

In the second project, we present a probabilistic method, based on a mixture model, for species tree inference. The method employs a two-phase approach, where in the first phase, a structural expectation maximization algorithm, based on a mixture model, is used to reconstruct a maximum likelihood set of candidate species trees. In the second phase, in order to select the best species tree, each of the candidate species tree is evaluated using PrIME-DLRS: a method based on the DLRS model. The method is accurate, efficient, and scalable when compared to a recent probabilistic species tree inference method called PHYLDOG. We observe that, in most cases, the analysis constituted only by the first phase may also be used for selecting the target species tree, yielding a fast and accurate method for larger datasets.

Finally, we devise a probabilistic method based on the DLTRS model: an extension of the DLRS model to include LGT events, for sampling reconciliations of a gene tree inside a species tree. The method enables us to estimate gene trees having temporal information for duplication and LGT events. To the best of our knowledge, this is the first probabilistic method that takes gene sequence data directly into account for sampling reconciliations that contains information about LGT events. Based on the synthetic data analysis, we believe that the method has the potential to identify LGT highways.
Contents

List of Publications 1

Acknowledgments 3

1 Introduction 7
  1.1 Outline of the Thesis ......................... 9

2 The biology of evolution 11
  2.1 From DNA to genome .......................... 11
  2.2 Evolution of gene families .................... 12
      Gene duplication ............................. 12
      Lateral gene transfer ........................ 14
      Natural selection and genetic drift .......... 14
  2.3 Homology: Finding equivalence among genes .... 15
  2.4 Orthology: evolutionary and functional perspective 16
      Phylogenetic inference and the Tree of Life ... 16
      The Ortholog conjecture and functional annotation 16

3 Computational Techniques 19
  3.1 Bayesian Modeling and Inference ................ 20
      Bayesian vs Frequentist: Two sides of a coin ... 20
      Bayesian modeling ............................ 20
      Inference in Bayesian networks ................ 21
  3.2 Mixture model and EM .......................... 22
      Maximum Likelihood Estimation ................. 22
      Mixture model ................................ 23
      Expectation Maximization ...................... 23
  3.3 Markov Chain Monte Carlo ........................ 27
      Motivation .................................. 27
      Mechanism .................................. 28
  3.4 Stochastic models for evolution ................. 31
The Birth-death model ........................................... 31
The Coalescent model ........................................... 32

4 Models and Methods for Gene Tree, Species Tree and Orthology
Inference ................................................................. 33
4.1 Phylogenetic terminology ..................................... 33
Gene tree and species tree ......................................... 33
Reconciliation and realization ..................................... 34
Gene tree-species tree incongruence ............................... 35
4.2 Computational approaches for gene tree and species tree inference ......................................................... 36
Parsimony ................................................................. 36
Probabilistic ............................................................ 38
Miscellaneous .......................................................... 40
Latest Trends ........................................................... 40
4.3 Computational approaches for orthology inference .......... 40
Heuristic-based sequence similarity methods .................. 40
Phylogeny-based methods ......................................... 42
Comparison of phylogeny-based and sequence similarity-based methods ................................................... 42
Future directions ....................................................... 43

5 Overview of Papers in this Thesis ................................ 45

Bibliography ............................................................. 47
List of Publications

Paper I: Integrating Sequence Evolution into Probabilistic Orthology Analysis
Ikram Ullah, Joel Sjöstrand, Peter Andersson, Bengt Sennblad and Jens Lagergren
Manuscript, under review in Systematic Biology

Paper II: Species tree inference using a mixture model
Ikram Ullah, Pekka Parviainen and Jens Lagergren

Paper III: Probabilistic inference of lateral gene transfer events
Mahmood Alam Khan, Owais Mahmudi, Ikram Ullah, Lars Arvestad and Jens Lagergren
Manuscript
Time really flies by! I vividly remember my first day at AlbaNova when I was meeting Jens in his office. Today, when I am writing these lines, it feels like that meeting was yesterday. Between yesterday and today, lies a chapter of my life full of fun, surprises, deadlines, and some achievements. This chapter is special because of all those friends who directly, and indirectly, helped me achieve my goal.

First, and foremost, I would like to thank my supervisor Jens Lagergren, for his full support during my PhD tenure. He is one the smartest guys, I have worked with in my career. Apart from the scientific ingenuity, his great sense of humor made it easy for me to cope with few stressful periods of my research. I would like to thank both of my co-supervisors for their academic and administrative assistance. I am grateful to Lars Arvestad for sharing his scientific advice and knowledge, and for many insightful discussions and suggestions. I am also grateful to Bengt Sennblad for assisting me in my research. He truly knows the art of explaining complex scientific concepts in an easy way.

I would like to thank Owais, my twin brother at Scilifelab (Some people have actually asked me if we are real brothers). Owais! You have been such a nice fellow right from our undergraduate days. I will always cherish your cheerful personality and all the home-baked food that I took for granted.

I would like to thank all my Gamma 6 friends for making these years so special. Pekka! If someone asks me to introduce you in two words, I will say intelligent and cool. I have the opportunity to learn so much from you. Joel! If I am able to write half as good code as you do, I will consider myself the best programmer in town. Hussein! I guess you still keep that super-encrypted backup hard disk containing your research. I know you are preparing yourself for the knighthood sometime soon. Mehmood! Still remember the day when we missed our flight to Paris, then, in AlbaNova, acted like nothing happened, while hiding from Jens' funny comments. Hashim! I will always envy your punctuality. Auwn! Remember your first day in Stockholm when I was your guide and, after multiple mishaps, we reached your room late in the night, without one of your bag. Mattias! Amazing, how easily you can solve all those hard puzzles. Kristofer! You are jolly good, both in science and gym. Lumi! Always striving for excellence and, no doubt, you deserve it. Viktor! Love to hear your funny sci-fi stories at fika time. Erik! Although I can not compete you in playing those different musical instruments that
you play so good, I will keep on challenging you for a tennis match.

I would like to thank all the friends whom I have met at Scilifelab, during these memorable years of PhD. Not in any particular order, I would like to thank Linus, Ino, Jose, Stefania, Lukas, Maria, Arman, Samuel, Matthew, Yrin, and Wenjing.

I would also like to thank my friends outside Scilifelab whose company has always been a pleasure. These include, but are not limited to, Irshad, Akbar, Kaleem, Sami, Izhar, Ibrar, Naved, Waseem, Arif, Usman, Laeeq, Mansoor, Salman, Daniel, and Johan.

Last, but not the least, I would like to thank my family for their support and love during my PhD. Mom! I know you are feeling so proud up there in the Heavens. I will always miss you. Sidra! Thanks for your support during these years while taking care of Anas. Both of you have made my life so colorful.
# List of Notations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary deoxyribonucleic acid</td>
</tr>
<tr>
<td>LCA</td>
<td>Last common ancestor</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>LGT</td>
<td>Lateral gene transfer</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation maximization</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum likelihood</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimation</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>NNI</td>
<td>Nearest neighbour interchange</td>
</tr>
<tr>
<td>SPR</td>
<td>Subtree pruning and regrafting</td>
</tr>
<tr>
<td>DL</td>
<td>Duplication loss</td>
</tr>
<tr>
<td>DLT</td>
<td>Duplication loss transfer</td>
</tr>
<tr>
<td>MPR</td>
<td>Most parsimonious reconciliation</td>
</tr>
<tr>
<td>GTP</td>
<td>Gene tree parsimony</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

“When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.”

— John Muir

Earth has been home to an amazing variety of species since life started almost 3.7 billion years ago [157]. Even more amazing is our ability to infer this history using a combination of scientific disciplines as tools and the repertoire of existing species and fossil records as data. While the tools and the data are indispensable in this investigation, an important lead is provided by the theory of evolution which posits that each and every form of life on Earth is an evolved copy of the Universal Common Ancestor [207]. This theory, coupled with the invention of molecular sequencing technology during the previous century, has revolutionized Life Sciences research.

In many aspects, the second half of twentieth century was remarkable in biomedical research. The discovery of the structure of deoxyribonucleic acid (DNA) in 1953, by James Watson and Francis Crick [218], paved the way for molecular genetics. At the same time, the world experienced tremendous technological advancements, both in computer hardware and software, which facilitated efficient analysis of molecular data. The culmination was perhaps the advent of the Internet and the rise of the Information age, which made a pool of genomic data available to researchers worldwide. On the computational side, during the same time, Bayesian inference made its way into mainstream mathematical statistics [62], and sampling methods like Markov chain Monte Carlo facilitated its implementation [14]. Last, but not the least, next generation sequencing (NGS) platforms dramatically reduced the sequencing costs [191]. All of these developments greatly facilitated the field of phylogenomics: a fairly new discipline at the intersection of evolution and genomics that uses genomic data to infer phylogenetic relationships, infer putative functions for DNA or protein sequences, and gain insights into the mechanisms of molecular evolution [171].

In phylogenomics, a tree is a popular structure to represent the evolutionary relationship among species and genes. Thus, a gene tree represents evolutionary
CHAPTER 1. INTRODUCTION

Figure 1.1: Tree of Life divided into three domains namely Bacteria, Eukaryote, and Archaea.

history among a set of homologous genes and a species tree represents the evolutionary history among a set of species. Species tree inference often uses a set of gene trees, while gene tree inference may be more robust when guided by a species tree [97, 34, 176, 4]. Due to artifactual or biological reasons, gene trees may actually be distinct from each other as well as from the species tree, a phenomenon known as gene tree-species tree incongruence [77]. Various models and methods have been developed over the years to resolve the incongruence, which will be described in Chapter 4.

Species tree inference is one of the central problem in evolutionary biology due to its applications in all major biological disciplines. Organization and inference of the biodiversity among species use a species tree either directly or indirectly. This includes, but is not limited to, estimating orthologous relationship among genes, investigating protein domain evolution, identifying genetic markers, and automatic annotation of genes/proteins. Species tree inference is a challenging problem in phylogenetics, both in terms of modeling and computation. From a modeling perspective, the challenge is to reconstruct a biologically realistic tree accommodating the effects of evolutionary events causing gene tree-species tree incongruence. From a computational perspective, the challenge is to reconstruct “Tree of Life”, illustrated at higher taxonomic level in Figure 1.1, for millions of living and extinct species. Computational scalability is a balance between speed and accuracy; the faster maximum parsimony and distance-based methods lack the statistical rigour offered by probabilistic approaches like maximum likelihood and Bayesian methods (see further Chapter 4). Regarding modeling, probabilistic approaches, due to their explicit modeling of incongruence between gene trees and species tree, promises robustness and reliability, but are computationally demanding. Several models and methods for species tree inference are discussed in Chapter 4.
Fitch introduced the concept of orthology and defined two genes as orthologs of each other, if their last common ancestor in a phylogenetic tree is a speciation event [71]. Experimental results have since shown that orthologs are more probable to perform similar functions, while paralogs are more probable to perform novel functions; this is popularly known as the ortholog conjecture [115]. Since each gene in all of the sequenced genomes can hardly be studied experimentally, transfer of functional annotation from model organisms to other genomes is, in most cases, the best means of detailed functional characterization. Orthology analysis, thus, has an important role to play in the post-genomic era. Chapter 4 contains a description of various models and methods for orthology analysis.

A significant proportion of genetic diversity in prokaryotes has been attributed to the transfer of DNA material from distantly related species: a phenomenon known as lateral gene transfer (LGT) or horizontal gene transfer [200, 154]. Today, it has been established as an important mode of evolution not only in prokaryotes, but also in some eukaryotes like protists [13]. A reconciliation-based approach is well-suited to explain LGT events, but is challenging from an algorithmic point of view. While parsimony-based methods have been proposed [27, 210], probabilistic methods have the potential to identify not only the existence, but also relative timing of LGT events. Paper III proposes a probabilistic method, based on DLTRS model [195], to sample reconciliations of the gene tree inside a species tree in the presence of gene duplication, gene loss, and LGT events.

1.1 Outline of the Thesis

Chapter 2 briefly describes the biological background for the computational models presented in the papers. For an in-depth biological insight, the reader is referred to [125, 166]. Chapter 3 describes different computational techniques used in the papers included in this thesis. For further details, the reader is referred to [38, 30, 148]. Chapter 4 describes popular models and methods for inference of species tree and orthology analysis. Chapter 5 includes a concise description of the papers included in the thesis.
Chapter 2

The biology of evolution

This chapter briefly describes the biological perspective on molecular evolution, within and across genomes. In particular, we will discuss several evolutionary events that are pivotal in shaping gene families; ranging from single point mutations to whole genome duplications. Insights into gene family evolution may be helpful, for instance, in orthology analysis and ancestral genome reconstruction.

The outline of the present chapter is as follows. Section 2.1 describes genetic material at different levels of granularity. Section 2.2 describes various evolutionary events shaping gene families. Section 2.3 describes homology and its types, while Section 2.4 describes how orthologous information among the genes can be beneficial for a biologist.

2.1 From DNA to genome

Deoxyribonucleic acid (DNA) is the fundamental unit of coding genetic information in all known living organisms and viruses. It has a double helix structure consisting of a sugar group, a phosphate group, and a base, where the base may be Adenine, Thymine, Guanine, or Cytosine. A stretch of DNA, known as gene, is the molecular unit of heredity of an organism. Genes are packaged into a tightly wound structure called a chromosome. Each chromosome is made up of two sister chromatids joined at a specialized region called centromere. Different organisms have different number of chromosomes. For instance, human (Homo sapiens) has 46 chromosomes, dog (Canis lupus familiaris) has 78 chromosomes, wheat (Triticum aestivum) has 42 chromosomes, and rice (Oryza sativa) has 24 chromosomes. The collection of all chromosomes in an organism constitute its genome. In short, a genome is divided into chromosomes, chromosomes contains genes and genes are made up of DNA.

While genes are the coding parts of a genome, a major portion of a genome consists of non-coding sequences which do not code for proteins. For instance, almost 98% of the human genome consists of non-coding region, which may partly be attributed to extensive gene duplication events followed by loss of function and
degredation of duplicate copies [64]. Non-coding sequences also consist of introns, non-coding functional RNA, repeat sequences and \textit{cis}-regulatory elements.

2.2 Evolution of gene families

A gene family is a group of homologous genes that are likely to have similar functions [49]. It may either be monocopy, i.e., there is one gene per species in the family, or multigene, i.e., some of the species have more than one gene in the family. In case of multigene families, the set of genes from the same species encode proteins with similar sequences. Chromosomal rearrangements, illustrated in Figure 2.1, disperse the multigene families throughout the genome. Genomic analyses of model organisms have shown that more than one-third of all protein-coding genes are members of multigene families [185, 109] and there is also a considerable percentage of monocopy gene families across various clades of “Tree of Life” [113, 216, 46]. Below, I describe some of the main evolutionary forces shaping the gene families across genomes.

Gene duplication

Gene duplication mechanism can be categorized based either on the scale or on the basis of whether it is RNA-mediated or not [67]. The former can be divided into tandem gene duplication, segmental gene duplication, and whole-genome duplication (also known as polyploidization). The latter can be divided into retroposition and transposition.

Tandem duplication refers to the duplication of a small segment of a chromosome. The duplicates usually exist in proximity of one another. Paralogs created as a result of tandem duplication tends to share the same regulatory element and are likely to perform similar biological functions. Comparative genomic analyses of closely related species have revealed that tandem duplication is one of the major mechanisms creating new genes, in particular genes clustered into a gene family [12, 63, 192].

Segmental duplications are large copies of genomic DNA, which can be more than 200 kilobases in size [81]. An estimated 5% of Human genetic material is composed of segmental duplications that have emerged during the past 35 million years of our species’ evolution [118]. They have been identified to be involved in genome’s evolution and organization, and new gene functions.

Polyploidization is defined as acquisition of one or more additional set of chromosomes by a cell. The phenomenon has contributed to the formation of many gene families [156], especially in plants [169]. In fact, it is believed to have occurred many times during the evolutionary history of all angiosperms [104]. In animals, it occurs rarely but there are some animals that exist as polyploids [158]. There are two types of polyploidization namely autopolyploidization and allopolyploidization. In former, the whole of a single genome is duplicated, whereas in latter, two closely
2.2. EVOLUTION OF GENE FAMILIES

Figure 2.1: Types of chromosomal re-arrangements. In the left sub-figure, an extra copy of a chromosomal region (for instance a gene) is inserted in the chromosome. In the central sub-figure, a chromosomal region is deleted, while in the right sub-figure, a region is reversed end-to-end within the chromosome.

related genomes are duplicated through hybridization followed by duplication of whole set of chromosomes.

Transposons, also known as transposable elements, are mobile genetic elements that move around in the genome and, in the process, often make duplicate copies of themselves. They make up a significant part of species’ genomes. For instance, around 45% of human genome [118] and more than 80% of maize genome [187] consists of transposons. There are two types of transpositions namely Class II transposons and Class I transposons (retrotransposons). Class II transposons are DNA segments which generally move by a cut-and-paste mechanism in which the transposon is excised from one location and reintegrated elsewhere. They are common in bacteria. Retrotransposition is the process responsible for Class II transposons during which functional (paralogs lacking introns) or non-functional (processed pseudogenes) copies from messenger RNA are derived. The process is mediated by L1 retrotransposons where L1-derived enzymes reverse-transcribe mRNAs and subsequently insert the resulting complementary DNA (cDNA) into the genomes, hence creating gene copies lacking introns [127]. Since retroposed genes usually lack regulatory regions, the general belief is that most retrogenes are non-functional, and that retroposition plays a very minor role in the expansion of gene families [84]. However, in case of plants, studies have shown that retroposition can be one of the main contributors to the expansion of gene families [112, 215].
Lateral gene transfer

Although Eukaryotes evolve principally via evolution of existing genetic material, a significant proportion of genetic diversity in prokaryotes comes from direct acquisition of genetic material from distantly related species, a phenomenon known as lateral gene transfer (LGT) or horizontal gene transfer (HGT) [154]. LGT can occur within prokaryotes, for instance, between two bacteria [150], and between prokaryotes and eukaryotes, for instance, between bacteria and plants [96, 78].

In bacteria, LGT compensates for lack of sexual recombination and is thus also referred to as ‘bacterial sex’ [150]. There are three modes of LGT in prokaryotes namely transformation, transduction and conjugation. Transformation is the uptake of exogenous DNA from the environment. Transduction is the injection of foreign DNA by a bacteriophage into the host bacterium. Conjugation is the transfer of genetic material between two organisms using plasmids, where such a transfer may occur either using a direct cell-to-cell contact or using a bridge like connection between them [152]. LGT has a central role in shaping prokaryotic life on Earth and it is now evident that a vast majority of expansions of protein families, even among large genome, are due to LGT [120, 212].

There is a higher variation of LGT frequency within eukaryotes [13]. For instance, animals and fungi are less exposed to LGT as compared to phagotrophic protists. Two modes of LGT in eukaryotes have been suggested. The first one, known as endosymbiotic gene transfer, is the transfer of genes from mitochondria and plastids (organelles with an endosymbiotic origin) to the nucleus of the eukaryotic cell [13]. The second one is the transfer of genetic material between unrelated species. Rosewich et al [183] argues that the first type has played an important role in the evolution of fungi, in spite of the fact that the mechanisms by which it may occur are unknown and the plausibility of alternative explanations for an LGT event.

Natural selection and genetic drift

Natural selection is one of the driving forces shaping genetic evolution. It can be broadly categorized into four types, namely stabilizing selection, directional selection, disruptive selection, and balancing selection [126]. Stabilizing selection, as the name implies, stabilizes the mean of a trait in the population around a stable optimal value. Directional selection drives the frequency of a trait towards an optimum that is away from the mean. Disruptive selection favours extreme phenotypes over intermediate ones by increasing the frequency of small and large values of a trait. Balancing selection, on the other hand, favors the optimal compromise between several constraints thereby maintaining polymorphism at a locus within a population. Purifying selection, i.e., removing the unfavorable mutations, plays important role in stabilizing biological structures by removing deleterious mutations. Directly after a duplication event, the purifying selection for the resulting paralogous genes is temporarily relaxed until one of the copy either becomes non-functional or per-
forms a novel function. In general, natural selection plays an important role in shaping the evolution of protein-coding genes in animals and plants [186, 45].

2.3 Homology: Finding equivalence among genes

Classically, phylogenetics has been studied using a tree-based structure, which is still the dominant paradigm in the literature [70]. The concept of paleontological chart, which is an early representation of a phylogenetic tree, may be found in the mid-19th century literature [98]. One of the important concepts in evolutionary trees is that of shared ancestry, which, in phylogenetic jargon, is referred to as homology. Homology refers to the conjecture that two (or more) genes have descended from a common ancestral gene. Since, for an arbitrary set of genes, we cannot directly explore their common ancestor and all intermediate forms due to the unavailability of fossil records, identification of homology is technically an inference problem [114]. Homologous genes may exhibit considerable sequence similarity due to their common ancestry. Although not always true, sequence similarity often suggests structural and functional similarity between the homologous genes.

Based on the type of the evolutionary event at the last common ancestor (LCA) of the given genes, homology may refer to orthology, paralogy or xenology. In

---

Figure 2.2: Illustration of different homology relations among genes, in three species A, B, and C. The red circle represents a speciation, the blue square represents a duplication, and the pink diamond represents an LGT.

---

[^1]: With the passage of time, finer classifications of homology are appearing in the literature like ohnology [221], partial homology, gametology and synology [144], but we will restrict ourselves to the three relatively established types, namely orthology, paralogy and xenology, in this thesis.
ORTHOLGY, the type of evolutionary event is speciation, while in paralogy and xenology, the event is duplication and LGT respectively. Paralogous genes may be further classified as either in-paralogs (duplication postdating the speciation) or out-paralogs (duplication predating the speciation). In-paralogs are also called co-orthologs or super-paralogs [234]. Figure 2.2 illustrates different types of homologous relationships. For instance, $a_1$ and $a_2$ are paralogs since there is a duplication event at their LCA. In contrast, $b_2$ and $c_2$ are orthologs since there is a speciation event at their LCA, while $a_3$ and $b_1$ are xenologs since there is an LGT event at their LCA. Moreover, $a_1$ and $a_2$ are in-paralogs since the duplication occurred after the speciation while, e.g., $c_1$ and $c_2$ are out-paralogs since the duplication occurred before the speciation. Although faster heuristics methods exist for detecting orthologs and in-paralogs from two species [179, 121], they may be detected more reliably using phylogenetic methods, for any number of species [190]. Paper I in this thesis uses the DLRS model and a dynamic programming algorithm for this purpose.

2.4 Orthology: evolutionary and functional perspective

Walter Fitch introduced the notion of orthology in an evolutionary perspective to find genes linked by a speciation event in their last common ancestor gene. With the passage of time, orthology has found numerous applications in a vast array of research areas [50].

Phylogenetic inference and the Tree of Life

Orthology inference plays a central role in investigating phylogenetic relationships in genes and species. There is considerable interest in understanding evolutionary history of species [31, 118, 80], in order to find patterns of evolutionary conservation, which are helpful in, for instance, finding mutations that are deleterious to protein functions and identifying non-coding sequences under negative selection. One of the ultimate goals in evolutionary biology is to reconstruct 'Tree of Life', i.e., the phylogenetic relationship among all the living and extinct species. Existing research in species tree inference, which utilizes orthologous information among genes, is an effort to reach that goal.

The Ortholog conjecture and functional annotation

The ortholog conjecture posits that orthologs tend to retain similar molecular and biological functions, while paralogs tend to perform novel functions using the criteria of subfunctionalization and neofunctionalization [155, 170]. In this regard, the impact of orthology has been studied both in terms of functional similarity [204, 101] and structural similarity [170]. Moreover, as one of the foundations of molecular biology, it is assumed that a protein’s sequence determines its structure,
which in turn determines its functions [95, 220]. Together, these phenomena assist us in inferring evolutionary relationships between proteins (and between the organisms having these proteins). In comparative genomics, the same relationship is used for gene annotation of uncharacterized genes using functional information of well-studied genes. Finally, in drug design, they are used to direct the design of novel protein functions.
Chapter 3

Computational Techniques

“The scientific imagination always restrains itself within the limits of probability.”

— Thomas Huxley

This chapter briefly describes the computational components on which the current thesis is based. Although evolutionary histories may be modeled either as a tree or as a network, the current thesis only deals with the tree-based models. For lateral gene transfer (LGT) events, where the reticulation events are often modeled using a network [147], the current thesis models them by relating the gene tree to the species tree [195], which can be viewed as a network. For network-based models and algorithms, the reader is referred to [146, 103, 147].

The outline of this chapter is as follows. Section 3.1 compares the salient features of the Bayesian paradigm vis-à-vis the Frequentist paradigm and the Bayesian perspective on probabilistic modeling and inference. Section 3.2 describes mixture models, the expectation maximization (EM) algorithm, and how maximum likelihood estimation (MLE) of parameters of a mixture model can be performed using an EM algorithm. A special case of EM, known as Structural EM, is also described where parameters related to the structure are optimized in each iteration of the EM. These models and methods constitute the main computational part of Paper II. Section 3.3 describes the mechanism and some algorithms based on MCMC framework. Paper I utilizes an MCMC framework and a dynamic programming algorithm for probabilistic orthology analysis. The chapter concludes with a brief discussion of two popular stochastic models of evolution: the birth-death model and the coalescent model. Although coalescent model is not used in the current thesis, it has been used in a variety of gene tree and species tree reconstruction methods, as described briefly in Chapter 4.
CHAPTER 3. COMPUTATIONAL TECHNIQUES

3.1 Bayesian Modeling and Inference

Bayesian vs Frequentist: Two sides of a coin

There is a basic difference between the interpretation of probability in Bayesian and frequentist paradigms which leads to a different perspective on the way inference is performed.

Intuitively, the frequentists interpret probability to be *frequency*. In other words, they assume that there is infinite sampling and the parameters of the process, called hypotheses, remains constant during its lifetime. On the other hand, Bayesians assume that the state of the process can always be updated and thus the underlying parameters are treated probabilistically. Thus frequentists play with data in presence of fixed parameters while Bayesians play with parameters in presence of fixed data.

Another point of difference between the two paradigms is the idea of *subjectivity*. In Bayesian statistics, in addition to the given data, a statistician may use existing domain knowledge in the form of a *prior distribution* when estimating the model parameters (see Equation 3.1 and 3.2), something a frequentist can not use because the parameters are fixed. By integrating over model parameters, Bayesian inference is less prone to over-fitting as compared to its counterpart.

There is a fundamental difference between the way statistical inference is performed in Bayesian and frequentist regimes where the former estimates the conditional probability of parameters given the data (a.k.a the posterior probability), while the latter estimates the conditional probability of data given the parameters or hypotheses. In this regard, Bayesian inference for complicated models is possible via Markov Chain Monte Carlo (MCMC) sampling, as we use in Paper I. Fortunately, such an inference offers theoretical guarantees for convergence if the Markov chain is ergodic, i.e., it is both irreducible and aperiodic (see Section 3.3 for details).

The statistical advantage of Bayesian methods comes at the cost of increased computational complexity [40]. Although this used to be a major bottleneck before, the ever-increasing available computing power is facilitating, and encouraging, the use of such methods.

Bayesian modeling

Computations using the Bayesian paradigm are based on an apparently simple but powerful equation formulated in 18th century by the English Reverend Thomas Bayes [26].

\[
P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}
\]  

(3.1)

where \(D\) denotes the data and \(\theta\) denotes the parameters. Moreover, \(P(D|\theta)\) denotes the likelihood, while \(P(\theta)\) and \(P(\theta|D)\) denote the prior probability and the
posterior probability of the parameters, respectively. The posterior distribution is a probability distribution that represents our updated beliefs about the parameters after having seen the data. The denominator $P(D)$ is often intractable and can be factored out as a normalizing constant in many applications, yielding the formulation

$$P(\theta|D) \propto P(D|\theta)P(\theta).$$  \hspace{1cm} (3.2)

For instance, in an MCMC framework, the denominator cancels out when we take a ratio between successive posterior probabilities. The parameters of the prior distribution may either be fixed, or, in case of a hierarchical Bayesian model, they may be taken from a distribution indexed by another set of parameters, called hyperparameters. For instance, in Eq 3.1, if we assume that $\theta$ is modeled using a normal distribution, then the mean $\mu_\theta$ and the variance $\sigma_\theta^2$ of this distribution are called hyperparameters.

In most of the real world models, there may be complex dependencies between different variables of the system. Such dependencies can be easily expressed using a graphical model, which may either be a directed network, also called a Bayesian network, or an undirected network, also called a Markov network. Graphical models can be learned using both Bayesian and frequentist methods, however they are more popular within the Bayesian paradigm [105]. Figure 3.1 shows a Bayesian network for the Student model as given in Koller and Friedman [111]. It is a toy example where the final letter grade of a student is modeled given course-related and student-related variables. The nodes in the network represent random variables, while the arrows in the network represent the dependencies between the random variables. An arrow from node $A$ to node $B$ denotes that $B$ is dependent on $A$. This is written as conditional probability $P(B|A)$ conveying the fact that any change in the distribution of $A$ will effect the distribution of $B$. For instance in Figure 3.1, SAT score of a student is dependent on his/her Intelligence, i.e. the more intelligent a student is, the more he/she will score in SAT. Using the chain rule for Bayesian networks [111, p 62], the joint distribution for Student network may be written as

$$P(I, D, G, S, L) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)$$

where $I, D, G, S,$ and $L$ denotes Intelligence, Difficulty, Grade, SAT score, and Letter respectively. While the Student network models five variables, Bayesian networks have been successively used for modeling complex biological phenomena involving tens and hundreds of variables [73, 141, 180, 99, 230].

**Inference in Bayesian networks**

Given a model of the phenomenon under study, probabilistic inference in Bayesian networks may be performed using either an exact method like Variable Elimination [47], or an approximate method like Metropolis Hasting algorithm [94] and Gibbs sampling [79]. Variable elimination, as the name implies, eliminates a subset of the
variables by marginalizing them out, using dynamic programming, until the variables of interest are left. Approximate methods, on the other hand, approximates the posterior distribution, often expressed as a multi-dimensional integral, using a heuristics, like moving towards high density regions [142, 94], or optimizing one variable at a time [79]. Although exact methods are desirable in most cases, computational issues limit their applicability for larger problems. Thus, approximate methods have gained considerable recognition in machine learning due to their wide applicability [16, 17, 5, 4, 18, 97, 175, 34, 195].

3.2 Mixture model and EM

A statistical model may consist of a single or a mixture of distributions. While models consisting of a single distribution often have lower complexity, in most of the cases, real world phenomena can be explained better using a mixture model consisting of more than one distribution. Mixture models appear as fundamental models in many areas of statistics such as statistical pattern recognition, classification, and clustering [57, 92]. Expectation maximization is a popular technique for maximum likelihood estimation in a mixture model.

Maximum Likelihood Estimation

Maximum likelihood estimation is perhaps the most widely used parameter estimation method in statistical studies. Given data $X$ and the set of parameters $\Theta$, the likelihood of $\Theta$ given $X$ is defined as the conditional probability of $X$ given $\Theta$, i.e., $L(\Theta|X) = P(X|\Theta)$, where $L(.)$ denotes the likelihood. The maximum likelihood estimate of $\Theta$, denoted $\Theta^*$, are the parameter values that maximizes the data likelihood. From a generative modeling perspective, $\Theta^*$ are the parameter values that,
3.2. MIXTURE MODEL AND EM

with the highest probability, has generated $X$. Mathematically, it is written as

$$
\Theta^* = \arg\max_{\Theta} L(\Theta|X).
$$

MLE has nice asymptotic properties including consistency in approaching the true value as data size increases, unbiasedness, and being statistically efficient, i.e., they have the smallest variance among unbiased estimates [228]. The log-likelihood, $\ell(\Theta|X) = \log(L(\Theta|X))$, is often computed for analytical convenience.

**Mixture model**

A mixture model is a generative model in which the data is generated from a convex combination of $K$ base distributions. Although not always true, the base distributions are usually from the same parametric family, but with different parameters. From a clustering perspective, the data $Y$ consists of $K$ clusters, one cluster per distribution. Associated with each data point $y_i \in Y$ is a random variable $z_i \in Z$ that takes a value $k$ when the data point is generated from the $k$th distribution. In other words, $z_k$ represents the membership probability of $k$th cluster for $y_i$. Mathematically, it is expressed as

$$
P(Y|Z, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} P(y_i|z_i = k, \theta)
$$

and the likelihood is written as

$$
L(\theta|Y) = \sum_{Z} P(Y|Z, \theta)P(Z|\theta).
$$

Parameter estimation using MLE takes the following form

$$
\Theta^* = \arg\max_{\Theta} \ell(\Theta|Y)
$$

where $\Theta = (\theta, Z)$ and $\ell(\Theta|Y) = \log(L(\Theta|Y))$. Computing analytically $\ell(\Theta^*|Y)$ is often intractable and a method of choice in such cases is the EM algorithm. Although EM was popularized by Dempster et al [48], the idea was around for few decades, as described briefly in [139].

**Expectation Maximization**

Suppose the data $X$ is generated according to a Gaussian distribution with mean $\mu$ and standard deviation $\sigma$, and let $\theta = (\mu, \sigma^2)$. For estimating the parameters of the distribution, we can analytically solve for $\mu$ and $\sigma^2$ by calculating the derivative of $\ell(\theta|X)$ with respect to each of the parameter, setting it to zero, and solving the resulting equation. However, in many practical problems, analytical derivation may be non-trivial due to the presence of hidden variables in the model. Expectation
maximization is a general method for finding the maximum likelihood estimate of the parameters of a distribution when, apart from observed variables, the data is dependent on some hidden variables. EM reduces the task of calculating the likelihood function, which may potentially be multimodal and have no closed form solution, to a sequence of simpler subproblems. The subproblems are chosen in a way that guarantees their corresponding solution \( \theta^1, \theta^2, \ldots \) and will converge to a local optimum of the likelihood function [52].

Assume that the complete data \( X = (Y, H) \), consisting of observed data \( Y \) and unobserved data \( H \), is governed by the set of parameters \( \Theta = (\theta, Z) \). For instance, in a phylogenetic tree reconstruction problem, \( Y \) consists of the observed sequence data for the leaves of the tree representing the extant species, \( H \) consists of the unobserved sequence data for the internal vertices representing the ancestral species, and \( \Theta \) consists of the sequence evolution parameters. The conditional probability of \( X \) may be written as

\[
P(X|\Theta) = P(H|Y, \Theta)P(Y|\Theta)
\]

and

\[
P(Y|\Theta) = \sum_H P(Y, H|\Theta) = L(\Theta|Y)
\]

Computing \( P(Y, H|\Theta) \) is usually easy for any fixed \( H \); intractability comes from the fact that usually there are at least exponentially many possible instantiations of \( H \). Therefore, we are interested in approximating \( L(\Theta|Y) \) using an EM formulation. Below, we briefly describe the steps in a standard EM algorithm, followed by its application for parameter estimation in a mixture model. Being a widely used parameter estimation method across a variety of domains, most machine learning textbooks include the description of EM [148, 57, 92, 30]. Parameter estimation in mixture models using EM is explained in [29].

**Standard EM**

EM is an iterative procedure consisting of two steps: the expectation step (E-step) and the maximization step (M-step). In the \( i \)th iteration, the steps are performed as follows.

1. In the E-step, the expected complete data log-likelihood is computed as

\[
Q(\theta, \theta^{i-1}) = E_{H|Y, \theta^{i-1}} \left[ \log p(Y, H|\theta) \right]
\]  

(3.3)

where \( \theta \) are the parameters we want to optimize, \( \theta^{i-1} \) are parameter estimates from \((i-1)\)th iteration, and the expectation is taken over \( P(H|Y, \theta^{i-1}) \): the distribution induced on \( H \) by \( \theta^{i-1} \) and \( Y \). \( Q(\theta, \theta^{i-1}) \) is usually referred to as Q-term.
2. In the M-step, the value of \( \theta \) that maximizes the Q-term is selected, i.e.,

\[
\theta^i = \arg\max_{\theta} Q(\theta, \theta^{i-1}).
\]  

(3.4)

The log-likelihood improves in each iteration until the local maxima is reached. The termination criterion is to test if

\[
\ell(\theta^i) > \ell(\theta^{i-1})
\]

and continue to the next iteration, otherwise we have converged to the local maxima and the process is completed.

**EM algorithm for mixture model**

For the sake of simplicity, we describe the EM algorithm for parameter estimation of a mixture of 2 distributions for one-dimensional Gaussian data. The case of mixture of \( k \) distributions for \( D \)-dimensional data can be explained similarly. The number of distributions is given apriori in a mixture model.

Suppose \( Y = \{Y_1, Y_2, \cdots, Y_N\}^T, Y_i \in R^1 \) is generated using a mixture of two Gaussian distribution \( K_1 = N(\mu_1, \sigma^2_1) \) and \( K_2 = N(\mu_2, \sigma^2_2) \), and \( K = \{K_1, K_2\} \). The task is to infer the parameter \( \Theta = (\mu_1, \sigma^2_1, \mu_2, \sigma^2_2, \pi) \) of the mixture model, i.e., mean and variance of each distribution as well as the distribution membership variable. Let \( Z_i \) be a random variable that takes value \( j \) when \( Y_i \) is generated by \( K_j \). After initializing each \( \Theta \in \Theta \), the following steps are repeated until convergence.

1. In the E-step, expected sufficient statistics of the model are computed. In this case, responsibility score \( \gamma_{ij} \) is computed for each \( Y_i \in Y \), and each \( K_j \in K \) as

\[
\gamma_{ij} = \frac{P(Z_i = j|Y_i, K)}{P(Y_i|K)} = \frac{P(Y_i, Z_i = j|K)}{P(Y_i|K)} = \frac{P(Y_i|Z_i = j, K)P(Z_i = j|K)}{\sum_{k=1}^{2} P(Y_i|Z_i = k, K)P(Z_i = k|K)}
\]

Intuitively, \( \gamma_{ij} \) is the responsibility that the distribution \( K_j \) takes for ‘explaining’ the observation \( y_i \). [30].

2. In the M-step, the maximum likelihood estimate of the expected sufficient statistics is computed, i.e., each parameter \( \Theta \in \Theta \) is maximized using newly computed responsibilities.
\[ \mu_j^* = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij} Y_i \]
\[ \sigma_j^2 = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij} (Y_i - \mu_j^*)^2 \]
\[ \pi_j^* = \frac{\sum_{i=1}^{N} \gamma_{ij}}{\sum_{i=1}^{N} \sum_{k=1}^{2} \gamma_{ik}} \]

By comparing the new values of the parameters to their corresponding previous values, convergence is tested, i.e., if
\[ \ell(\Theta^* | \mathbf{Y}) > \ell(\Theta | \mathbf{Y}) \]

then the parameter values are updated and the EM steps are repeated, otherwise the algorithm is terminated.

In Paper II, as the first phase of a two-phase algorithm MixTreEM-DLRS, we have devised MixTreEM: an EM algorithm based on a mixture model for reconstructing a maximum likelihood set of \( K \) species trees. The input consists of a set of \( N \) gene families \( \mathbf{F} \) and a set of \( K \) initial species trees \( \mathbf{S} \). The responsibility score \( \gamma_{ij} \) is defined as the log-likelihood of \( S_j \in \mathbf{S} \) given \( F_i \in \mathbf{F} \). In the \( k \)th iteration of the EM, in the E-step, the algorithm computes \( \gamma_{ij} \) for each combination of gene family \( F_i \) and species tree \( S_{j}^{k} \), which quantifies the ‘influence’ of a gene family \( F_i \) on the reconstruction of \( S_j^{(k+1)} \), i.e., \( S_j \) for \( (k+1) \)th iteration. In the M-step, the responsibility score of \( S_{j}^{k} \) for each family is used to reconstruct a weighted supersequence, which is used to reconstruct \( S_j^{k+1} \). The second phase of the algorithm uses DLRS model [4] to assess each \( S_j \in \mathbf{S} \) with respect to a larger set of gene families and selects the best species tree \( S_* \).

**Structural EM**

Likelihood-based tree reconstruction methods are quite popular in phylogenetics since Felsenstein’s pruning algorithm [69]. A computational bottleneck of Felsenstein’s algorithm is that, since it operates on a fixed tree, each time the tree is rearranged, say using a subtree pruning and regrafting (SPR) or a nearest neighbour interchange (NNI) move, some or all of the edge lengths have to be re-estimated.

Structural EM algorithm, discovered by Nir Friedman [72], is a variant of the standard EM algorithm which performs a structure search in successive iterations of the EM. The structure search is guided by a scoring function that evaluates how well a structure matches the data. An extension of the algorithm, known as SEMPHY, has been devised by Friedman et al [75], for phylogenetic inference. SEMPHY significantly improves the convergence rate due to simultaneous optimization of
3.3. MARKOV CHAIN MONTE CARLO

Bayesian statistics often deals with high-dimensional problems for which, in contrast to traditional numerical techniques, Monte Carlo (MC) methods typically provide a good approximation strategy. At the heart of MC methods is the idea of approximating a computationally hard problem using statistical sampling. So being able to sample from complex multidimensional distributions is central to the success of this approach; in Markov Chain Monte Carlo (MCMC) [14] this task is performed by sampling from a Markov Chain whose stationary distribution is the target distribution. In Bayesian analysis, the goal is to estimate a posterior distribution, which poses the problem of constructing a Markov Chain having the posterior distribution as its stationary distribution.

Motivation

Integration and optimization are two of the fundamental challenges when solving problems in high-dimensional spaces. These problems arise in almost all scientific disciplines, e.g., bioinformatics, machine learning, Bayesian statistics, combinatorics and econometrics. For instance, in Bayesian statistics, while computing the posterior distribution \( p(\theta | x) \) of the data \( x \in X \) and parameters \( \theta \in \Theta \), i.e.,

\[
p(\theta | x) = \frac{p(x | \theta) p(\theta)}{\int_{\Theta} p(x | \theta') p(\theta') d\theta'}
\]

computing explicitly the normalization factor in the denominator (by integrating over all possible \( \theta \)) can be prohibitively expensive. Similarly, computing the
marginalization factor, given the joint posterior of \( (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2 \), i.e.,

\[
p(\theta_1|x) = \int_{\theta_2 \in \Theta_2} p(\theta_1, \theta_2|x)d\theta_2
\]
is again computationally intractable.

Monte Carlo methods offer a nice approximation for intractable problems. Although they have considerable computational burden necessitated by large number of samples for better approximation, for high-dimensional problems, they may outperform deterministic methods. A candid analysis of a Monte Carlo-based method is perhaps given by Alan Sokal [197]. While explaining the method, he starts by warning that Monte Carlo is a bad method and it should be used only when all alternative methods are worse. However, by comparing its performance with Simpson’s rule in coping with the curse of dimensionality, he shows that, for reasonably high-dimensional problems, comparative methods can perform much worse than Monte Carlo. Thus Monte Carlo is a method of choice in high dimensional problems.

**Mechanism**

A Markov chain is a stochastic process in which the next state, irrespective of the past states, depends only on the current state. Mathematically, this is expressed as

\[
P(x^{i+1}|x^i, x^{i-1}, x^{i-2}, \ldots x^1) = P(x^{i+1}|x^i)
\]

where, for \( i > 0 \), \( x^i \) is the state of the system at time step \( i \). For all pairs of states \( (a, b) \) in the system, the probability \( T_{ab} \) of going from state \( a \) to state \( b \) is given in a transition matrix \( T \), with the condition that \( \sum_b T_{ab} = 1 \) (in discrete case) or \( \int_b T_{ab}dj = 1 \) (in continuous case). A Markov chain is time homogeneous if \( P(x^j = a|x^{j-1} = a') = T_{aa'} \) for all time points \( j \), i.e., the transition probabilities are independent of time. Below, the discrete time Markov chain is explained followed by the continuous case.

Consider the transition graph \( X = \{x_1, x_2, x_3\} \) illustrated in Figure 3.2 and taken from [14], consisting of 3 states. Transition matrix for \( X \) is

\[
T = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0.1 & 0.9 \\
0.6 & 0.4 & 0
\end{bmatrix}.
\]

Some example transition probabilities in \( T \) are \( T_{12} = 1, T_{23} = 0.9 \) and \( T_{32} = 0.4 \). Taking an arbitrary value for the initial state of the Markov chain, say \( \pi_0 = (0.5, 0.2, 0.3) \), we can calculate the distribution over the states in successive iterations as
where we observe that, for \( m \geq n \), \( \pi_m \) remains invariant. In other words, after \( n \) steps, the chain has converged to the stationary distribution, denoted as \( \mu \). Every irreducible and aperiodic finite state Markov chain has a limiting distribution, which is its unique stationary distribution [148]. A Markov chain is irreducible if any state \( j \) is reachable from any state \( i \) in finite time. For aperiodicity, we first define the period of a state \( i \) as

\[
d(i) = \gcd\{t : T_{ii}(t) > 0\}
\]

where \( \gcd \) denotes greatest common divisor and \( T_{ii}(t) \) denotes the probability that, starting at state \( i \), we return to the same state in \( t \) steps. A state \( i \) is aperiodic if \( d(i) = 1 \) and a Markov chain is aperiodic if all its states are aperiodic [148]. For example, every state of a complete bipartite graph has period 2 since we can return to the same state in \( 2, 4, 6 \cdots \) steps. A sufficient, but not necessary, condition to ensure that a particular distribution for a Markov chain is a stationary distribution is via the detailed balance condition, stated as

\[
\pi_i T_{ij} = \pi_j T_{ji}.
\]
Markov chain having rapid mixing time, i.e., where the chain efficiently converges to a stationary distribution, are preferred. The relation \( \pi T = \pi \), at convergence implies that \( \pi \) is the left eigenvector of \( T \) with an eigenvalue of 1. According to Perron-Frobenius theorem, each of the remaining eigenvalues have absolute value less than 1. The second largest eigenvalue determines the rate of convergence of the chain and is desired to be as small as possible [14]. Apart from heuristics algorithms like Metropolis-Hastings [142, 94], Boyd et al [35] formulated the problem, of assigning transition probabilities in \( T \) in such a way that the second largest eigenvalue is minimized, as a convex optimization problem, in principal computing the fastest mixing Markov chain. They demonstrated their method on Markov chains for graphs having up to 1000 edges.

Exact methods to obtain the stationary distribution are not feasible for modeling real world scenarios where, owing to high-dimensionality of the problem, the search space is very large. The method of choice in such a situation is the Metropolis-Hastings algorithm which performs a random walk on a graph representing the search space. Given the current value \( x^i \), the algorithm proposes a new value \( x^j \) using a proposal distribution \( q(x^j|x^i) \). The new value is accepted with an acceptance probability \( A(x^i, x^j) \) according to the following formula

\[
A(x^i, x^j) = \min \left( 1, \frac{P(x^j)q(x^i|x^j)}{P(x^i)q(x^j|x^i)} \right).
\]

For an \( n \)-dimensional state space \( X = \{X_1, X_2, \ldots, X_n\} \) where we have \( n \) proposal distributions \( Q_k(x_k|x) \), \( 1 \leq k \leq n \), one for each state variable, the steps are as follows. Given the vector representing the current state \( x = \{x_1, x_2, \ldots, x_n\} \), \( x_k \in X_k \), \( 1 \leq k \leq n \), the next state \( x' \) may be proposed according to the proposal distribution \( Q_j(x'|x) \), \( j \leq n \) where the proposed vector \( x' \) differs from \( x \) only in the \( j \)th component. The proposed state is accepted with an acceptance probability

\[
A(x, x') = \min \left( 1, \frac{P(x')Q_j(x|x')}{P(x)Q_j(x'|x)} \right).
\]

The component to perturb is chosen at random in each step. When sampling from conditional distribution is feasible, a special case of Metropolis-Hastings algorithm named Gibbs sampling [79] is often used. The Gibbs sampling algorithm samples from the distribution of each variable in turn, conditional on the current values of the other variables. A nice property of Gibbs sampling is that the proposed state is always accepted [148].

An important factor in the design of MCMC algorithms is the choice of proposal distribution, which determines the rate at which proposed states are accepted. Movement in the state space with reasonable acceptance rate is desirable, i.e., the rate should neither be too low nor too high. If the acceptance rate is too high, the chain is perhaps not mixing well; if the acceptance rate is too low, the performance of the algorithm is not efficient due to rejecting too many proposed states. While parameters of the proposal distribution are fixed in Metropolis-Hasting algorithm
3.4. STOCHASTIC MODELS FOR EVOLUTION

and Gibbs sampling, they are learnt ‘on the fly’ in adaptive MCMC algorithms
like Adaptive Metropolis algorithm and Single Component Adaptive Metropolis
algorithm [86, 87].

For an introduction to MCMC algorithms in machine learning, see for instance
[14]. For detailed discussion of Markov chain Monte Carlo methods and their vari-
ants like sequential Monte Carlo, slice sampling and perfect sampling, see [38, 123].

3.4 Stochastic models for evolution

In 1874, in order to investigate the extinction of family names, Francis Galton
and Henry Watson derived a stochastic branching process named Galton-Watson
process [217]. The Galton-Watson process is a discrete time Markov process in
which a population evolves in generations. Let the number of individuals in \( k \)
generation be \( Z_k \) and let \( Z_0 = 1 \). Each member of the \( k \)th generation gives birth
to a family of \( N \geq 0 \) offspring, constituting the \( (k + 1) \)th generation, according
to a common probability distribution \( p_k \). In each generation, the offspring count
\( C_\alpha, C_\beta, C_\gamma, \cdots \) for distinct individuals \( \alpha, \beta, \gamma, \cdots \) are independent mutually as well
as independent of the offspring count of previous generations.

The stochastic models of evolution derived during the first half of the 20th cen-
tury extended the Galton-Watson model using the landmark work on continuous-
time Markov process by Andrey Markov in 1906 [134]. Such extensions included
the pure birth model [229], and birth and death model [68, 107].

The Birth-death model

In 1924, George Udny Yule derived a model for the distribution of biological taxa
and subtaxa using a single-parameter stochastic process known as Yule-Furry pro-
cess [229, 153]. It is a pure-birth process which has been reinvented and renamed
several times. For instance, it has been used to study the distribution of wealth
[193], the phenomenon of rich getting richer, called Matthew effect [140], the evo-
lution of citation networks, called cumulative advantage [172], and more recently,
the number of links to a web page, called preferential attachment [25]. In the Yule
process, a population starts at time 0 with one individual. As time progresses,
individuals may give birth to new individuals according to a birth rate \( \lambda \), i.e., for
any individual, the probability of a birth during the time window \( [t, t + dt] \) is \( \lambda dt \).

The Yule model was extended by William Feller who considered a birth and
death model with constant birth rate \( \lambda_0 \) and constant death rate \( \mu_0 \) [68]. Feller’s
model was later generalized by David Kendall to allow the birth rate \( \lambda(t) \) and death
rate \( \mu(t) \) be any specified function of the time \( t \).

The generalized birth-death process is a continuous time, discrete space Markov
process, where the birth and death rate at any time depends on the existing number
of extant particles (species, genes, lineages, customers etc.). Given \( k \) particles, a
CHAPTER 3. COMPUTATIONAL TECHNIQUES

birth and a death event occurs at an instantaneous rate of $\lambda_k$ and $\mu_k$, respectively; both of these parameters may be any function of $k$ but are time-homogeneous [107, 108, 41]. The birth-death model has a vast array of applications in modeling of biological processes including population dynamics [93], speciation [229], genome evolution [178], growth of paralogous gene families, horizontal gene transfer and somatic evolution in cancer [136, 16, 17, 18, 4, 195, 143, 211].

The Coalescent model

One of the important phenomena in population genetics is genetic drift: the change in frequency of an allele due to random sampling, introduced by unconstrained mating in the population [137]. Just like natural selection, mutation, and recombination, it is also responsible for defining geneologies in a population [182].

One of the first model of genetic drift is the discrete time Wright-Fisher model [222]. In its basic form, it is a Markov model which assumes a constant diploid population of size $N$. The generations are non-overlapping and each copy of the gene in the succeeding generation (of size $N$) is drawn uniformly at random from the $2N$ copies of the gene with replacement. In 1958, Patrick Moran presented an improved model called Moran model. The model assumes overlapping generations with constant population size $N$. A new generation is created by sampling randomly such that the birth of gene (by sampling uniformly from previous generation genes) is followed by the death of a gene.

The coalescent model provides a probabilistic description of the genealogical tree and the coalescent times for a sample of individuals taken from a population, and runs backwards in time. J.F.C. Kingman presented his landmark model [110] to explain the coalescent process using an approximation of the Wright-Fisher model. He assumed that population size is large and the sample size is much smaller than the population size. Kingman’s model can be viewed as a reverse counterpart to the diffusion approximation of the Wright–Fisher model [88, ch 7]. Coalescent model has been used in phylogenetics for reconstructing a species tree, and for estimating ancestral population size and species divergence times [174, 123, 97].
Chapter 4

Models and Methods for Gene Tree, Species Tree and Orthology Inference

“If one has really technically penetrated a subject, things that previously seemed in complete contrast, might be purely mathematical transformations of each other.”

— John von Neumann

This chapter describes the state-of-the-art models and methods for phylogenetic inference in gene trees and species trees. Apart from other applications, these are used for inferring orthology relationships among genes residing in different species. Section 4.1 introduces phylogenetic terminology used in the rest of the chapter, and the factors that can cause a gene tree to differ from the corresponding species tree. Such factors play a role in designing algorithms for reconciling a gene tree within the species tree. Section 4.2 describes different models and methods for gene tree and species tree inference. Finally, Section 4.3 describes computational tools for inference of orthologous relationship among genes.

4.1 Phylogenetic terminology

In this section, we define some phylogenetic terminologies used later in the chapter.

Gene tree and species tree

In phylogenetics, a species tree is used to represent the evolutionary history of a set of species. The leaves of a species tree represent extant species, while the internal vertices represent speciations. Similarly, a gene tree is used to represent the evolutionary history of a set of homologous genes. The leaves of a gene tree represent extant genes, while an internal vertex represent a speciation, a duplication or an LGT event. Figure 4.1 shows some examples of gene trees and species tree.
CHAPTER 4. MODELS AND METHODS FOR GENE TREE, SPECIES TREE AND ORTHOLOGY INFERENCE

In all subfigures, the outer tree (colored light blue in 4.1a and 4.1b, and multicolored in 4.1c and 4.1d) represents a species tree, while the inner tree colored black represents a gene tree. In each species tree, the internal vertices (colored dark blue in all subfigures) represent speciation. In each gene tree, the internal vertices colored green represent duplication while those colored red represent speciation. Each of the capital letter, A, B, C, D, represents a species name, while each small letter, a₁, a₂, b₂, etc., represents a gene name. For instance, in Figure 4.1a, species A has three genes namely a₁, a₂ and a₃, while the other species B, C, and D has one gene namely b₁, c₁, and d₁, respectively.

Reconciliation and realization

In a general setting, a reconciliation models the historical association between some host entity and a guest entity that tracks the evolution of the host entity. Examples of such associations of host and guest entities include the following: the vicariance in biogeography where organisms track geographical areas; the host-parasite co-speciation in parasitology where parasites track hosts; and the gene tree evolution inside a species tree in phylogenomics where genes track host species [160]. Ignoring reticulate phenomena like hybridization, in each case, we can visualize the process as a guest tree evolving inside a host tree, as shown in Figure 4.2. In phylogenomics, a reconciliation is used to explain the evolution of a gene tree inside a species tree [82, 162, 145]. Each leaf in the reconciled tree, representing a gene, is placed on a leaf in the species tree, representing the species to which the gene belongs. Each internal vertex in the reconciled tree is placed either on an internal vertex of the species tree, implying a speciation, or along an edge of the species tree, implying a duplication or an LGT [82, 162, 33]. Examples of a reconciliation is given in Figure 4.1a, and in Figure 4.1b. Note that, in a reconciliation, a duplication or an LGT may occur at any point along a species tree edge.

A realization is a constrained reconciliation, in which every gene tree vertex mapped to a species tree edge, implying a duplication or an LGT, is also pinpointed.
4.1. PHYLOGENETIC TERMINOLOGY

Figure 4.2: Illustration of the common pattern among three different evolutionary phenomena. In a), a gene tree is evolving inside a host tree; in b), a parasite is coevolving along side its host; in c), a clade is diverging in relation to the geological splits in an area. Note that, across all three phenomena, we can visualize the outer tree as a host tree, while the inner tree as a guest tree.

to a time on that edge. The DLRS and DLTRS model (see below), use discretized realization of a gene tree inside a species tree, where a duplication event or an LGT event is allowed to occur in an interval on a species tree edge. Figure 4.1c and 4.1d illustrates two examples of discretized realizations of the reconciliation in Figure 4.1a; the different colours on the species tree in Figure 4.1c and 4.1d indicates different discretization intervals. Species tree-aware gene tree inference based on realization, rather than reconciliation, of a gene tree inside a species tree provide more insight into the evolutionary process. For instance, it can be used to infer gene age, and to investigate the relative timing of pseudogenization events [131].

Gene tree-species tree incongruence

Evolutionary events like gene duplication, gene loss, and LGT often affect the gene tree but not the species tree, resulting in incongruence between the two trees. Figure 4.3a illustrates the gene tree-species tree incongruence, introduced as a result of gene duplication and gene loss, while Figure 4.3b illustrates the incongruence, introduced as a result of lateral gene transfer. Assuming the presence of one or more such events, different models have been proposed in the literature as described in Section 4.2. In essence, one of the main modeling challenges in reconstructing Tree of Life is to reliably address the gene tree-species tree incongruence in the presence of such confounding evolutionary events.
CHAPTER 4. MODELS AND METHODS FOR GENE TREE, SPECIES TREE AND ORTHOLOGY INFERENCE

Figure 4.3: Illustration of gene tree-species tree incongruence, due to duplication-loss in 4.3a, and lateral gene transfer in 4.3b. In both sub-figures, on the left, gene tree evolution within the species tree is shown while, on the right, the resulting gene tree is shown for comparison with the species tree.

4.2 Computational approaches for gene tree and species tree inference

The problem of gene tree and species tree inference has been studied using different computational paradigms. They may broadly be classified into parsimony and probabilistic approaches [53]. Parsimony methods are based on the philosophy of “less is more” where, given the gene sequences, the tree that allows evolution of the sequences with the fewest changes, is selected. Probabilistic approaches, based on a maximum likelihood or a Bayesian Markov chain Monte Carlo framework, select the tree that has the highest probability according to the parameters of the evolutionary model. Parsimony methods are more efficient but less realistic than probabilistic methods and are prone to perform poorly in certain cases, for instance, in the presence of long branch attraction [159]. Probabilistic methods are often biologically more realistic, and hence more accurate, since they explicitly model the evolutionary process. Apart from the methods using either the parsimony or probabilistic approach, there are methods that combines both of them.

We will now describe general reconciliation-based approaches, and their applications for gene tree and species tree inference.

Parsimony

Parsimony, originally developed for analysing discrete morphological characters [228], is intended to minimize the reconciliation cost, which usually refers to the total number of evolutionary events required for reconciling a gene tree and a species tree [66, 51]. A reconciliation cost may be defined as duplication cost, i.e., the number of duplication events; a duplication-loss (DL) cost, i.e., the sum of duplication and loss events; duplication-loss-transfer (DLT) cost, i.e., sum of duplication, loss and transfer events; or deep coalescence cost, i.e., the number of ILS events [206]. The deep coalescence cost is related to the problem of minimizing deep coa-
4.2. COMPUTATIONAL APPROACHES FOR GENE TREE AND SPECIES TREE INFERENCE

lescence (MDC) that seeks a species tree that reconciles the given gene trees with the minimum number of deep coalescence events [129, 130].

Gene tree inference

Goodman et al [82] introduced the concept of a reconciliation and provided an algorithm for computing the most parsimonious reconciliation (MPR) of a gene tree inside a species tree based on the DL cost, which was later formalized by Guigó et al [85]. Polynomial time algorithms for the DL model are given in [231, 65, 233].

Page et al [165] introduced the term “episode” for gene duplications in different gene trees, explainable by a single event. Episode-based gene duplication problems are concerned with identifying and locating episodes of gene duplications on the species tree, given a collection of gene trees and a comparable species tree. There are different flavors of the problem depending on whether we want to minimize the number of episodes, or the number of species tree locations. One variant, called the minimum episode (ME) problem [22], assigns duplication events to nodes in a species tree such that the total number of episodes is minimized. Bansal et al [22] presented an exact algorithm for the ME problem. Another variant, the episode clustering (EC) problem, determines a minimum number of locations for gene duplication events from the gene trees on the species tree, given a collection of rooted, binary gene trees and a rooted, binary species tree. Burleigh et al [37] presented an exact algorithm for the EC problem.

There has been considerable development in parsimonious methods for modeling LGT events [1, 44, 54, 151, 19, 2, 20]. Hallett and Lagergren [90] developed a reconciliation model in the presence of LGT events when the number of LGT events and the activity level are fixed. Hallett et al [91] formulated a parsimonious model for simultaneous identification of duplication and LGT events. David and Alm [44] devised a polynomial time algorithm, while accounting for duplication, loss and LGT events, called AnGST (Analyzer of Gene and Species Trees). Recently Scornavacca et al [188] presented TERA - a parsimonious method based on a scoring scheme combining DTL events with an estimate of the sequence likelihood.

Species tree inference

The species tree inference problem is NP-hard for different reconciliation costs [128, 24, 232, 23] necessitating the use of heuristic algorithms. Analogous to gene tree inference, there is a rich set of reconciliation-based parsimony methods for species tree inference. Guigo et al [85] devised a heuristic based on the DL cost involving neighbor-swapping. Page and Charleston [164, 163] showed that a combination of, so-called, nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) moves to alter the tree topology are effective in finding the most parsimonious species trees. They proposed a method using the duplication cost for selecting the optimum species tree which is implemented in GeneTree [161]. Hallett and Lagergren [89] devised a parsimony-based method using the DL cost
with the condition that, for some gene trees, no more than a bounded number $k$ of simultaneously active genes are located in any species at any time. Bansal et al [21] proposed an optimized version of the parsimony method of Page [164]. Their method has been implemented in Duptree [219] and iGTP [39]. Zhang [232] showed that, although deep coalescence and gene duplication are two different mechanisms responsible for the discord of gene trees and species trees, the deep coalescence cost is equal to the number of gene losses minus two times the duplication cost in the reconciliation of a uniquely leaf-labeled gene tree and a species tree. Maddison [129] proposed gene tree parsimony (GTP) criteria for inferring species trees from gene trees by minimizing deep coalescence. Maddison and Knowles [130] implemented a heuristic-based approach using this criteria. Than and Nakhleh [206] devised exact algorithms for this problem which was extended to include the case for sampling multiple alleles in their subsequent paper [205].

Probabilistic

Probabilistic models and methods are mainly categorized into those using a likelihood-based approach and those using a Bayesian approach. Likelihood methods made its way into mainstream phylogenetics with Felsenstein’s likelihood algorithm [69], while later MCMC methodologies made Bayesian methods computationally viable [173, 138].

Gene tree inference

Bayesian methods based on MCMC sampling for gene tree inference has progressed quite rapidly, especially during the last three decades, due to their power to resolve complex relationships, e.g., in evolutionary biology [122, 227, 138, 100, 55, 97]. For instance, MrBayes [100], BEAST [55] and PhyloBayes [119] have been successfully used for phylogenetic inference, coalescent analysis, divergence time dating, phylogeography and related evolutionary analyses [56]. Arvestad et al. presented the first probabilistic model based on the DL cost, known as gene evolution model or the DL model [16, 18], where, over any species tree edge, the gene duplication and loss was modeled by a linear birth-death process [107]. The DL model was extended to the gene sequence evolution model [17] by incorporating any standard models of sequence evolution and an accompanying MCMC framework to estimate the posterior distribution was also developed. Finally, Åkerborg et al [4] extended the gene sequence evolution model to DLRS (previously GSR) model along with an MCMC-based implementation, by integrating models for gene duplication and loss, a relaxed molecular clock for substitution rates, and sequence evolution. The software implementation of these algorithms are available in the PrIME [189] and JPrIME [194] packages. In order to mitigate the computational overhead of such models, the reconciliation space may be restricted to be in the vicinity of MPR [176, 175, 34]. Maximum likelihood models and accompanying algorithms for DL cost were devised in [83]. Since the histories of genes and species are tightly linked
4.2. COMPUTATIONAL APPROACHES FOR GENE TREE AND SPECIES TREE INFERENCE

but they are seldom identical due to gene tree-species tree incongruence, modeling gene tree inference in light of a species tree is biologically more realistic and accurate [175, 202].

Sjöstrand et al [209, 195] extended the DLRS model to DLTRS model, by modeling the LGT events, in addition to the previously modeled evolutionary events, and implemented an MCMC framework for estimating the posterior distribution. The authors concluded that the common practice of disregarding a species tree while performing gene tree inference overestimates the number of LGT and duplication events. Szöllösi et al [201, 203] extended the model to reconstruct chronologically ordered species phylogenies with the possibility that some of the gene transfers may be from the species outside the given species tree. Although network-based reconciliation approaches in the presence of LGT are common [102, 28, 214, 224, 167, 168, 226], they are better suited for modeling hybridization events [32].

Species tree inference

Probabilistically, a species tree may be reconstructed by taking gene trees [15] or molecular sequences as a starting point [56]. Since gene tree inference is improved by using a species tree and species tree inference is often based on one or more gene trees, there is a circular dependency between the methods for the reconstruction of gene tree and species tree. One possible approach, thus, is to use a DLRS-based framework to jointly reconstruct the species phylogeny and the histories of all gene families present in their genomes [34]. Coalescent-based methods are also common as they are a natural framework for modeling ILS [149].

Boussau et al [34] presented the first probabilistic method, accounting for gene duplication and gene loss using an extension of the DLRS model, to jointly infer rooted gene trees and species tree. Earlier, Heled and Drummond [97] presented a similar method under a multispecies coalescent model. While the method of Liu and Pearl [124] is also coalescent-based like Heled and Drummond [97], there are few differences: the former requires an outgroup, and estimate each gene tree individually followed by species tree inference; the latter generates rooted species tree and co-estimate gene trees and species tree in a single Bayesian MCMC analysis. Bryant et al [36] presented a coalescent-based model and an algorithm, called SNAPP, that can be used to reconstruct a species tree assuming unlinked biallelic data. STEM [117] and STELLS [225] are two maximum likelihood-based methods for inferring a species trees under the coalescent model, where the former requires gene tree with branch lengths while the latter only requires the gene tree topology. Ané et al [15] devised a Bayesian consensus-based approach, based on concordance among gene trees, by estimating the proportion of the sampled genes for which any given clade in the species tree is true. Their method is more efficient than the method by Suchard et al [199], since the latter considers both gene tree and branch length, and is thus computationally expensive, while the former considers only gene trees, and is thus scalable to larger datasets.
CHAPTER 4. MODELS AND METHODS FOR GENE TREE, SPECIES TREE AND ORTHOLOGY INFERENCE

Miscellaneous

Concatenation is a fairly popular approach in which a supergene, constructed by combining the sequences from multiple loci, is used to infer a species tree. Another method, called the consensus method, infers gene trees separately and then infers the species tree as a consensus of those gene trees. According to Gadagkar et al [76], concatenation often performs better than consensus, both for parsimony and likelihood methods. Concatenation has been used, for instance, by Rokas et al [181] to infer a species tree for a set of yeast species, and by Amrine et al [10] to identify a novel phylogenetic marker representing all extant orders of Placentalia.

Latest Trends

With the possibility of denser taxonomic sampling, investigation of hybrid models, combining the birth-death model and the multispecies coalescent model, is one of the interesting research avenues [198, 177, 23, 135, 223]. Another interesting area is devising efficient methods for co-estimation of gene trees and species tree, due to the inherent dependency between gene tree reconstruction and species tree reconstruction methods. Although such methods, separately, under the DL cost [34] and the ILS cost [97] have been proposed, it will be interesting to see hybrid methods combining different evolutionary models in this context. Last, but not the least, the rapid increase in volumes of NGS data has, in general, added a ‘big data’ dimension to phylogenomics models, thereby questioning the computational viability of some of the existing algorithms for gene tree and species tree reconstruction. A new set of algorithms having lower complexity is thus desirable, by investigating, for instance, effective divide-and-conquer approaches.

4.3 Computational approaches for orthology inference

Methods for orthology inference can be broadly classified into heuristics-based sequence similarity methods and phylogeny-based methods. The former is computationally efficient but prone to relatively higher errors, while the latter, thanks to explicit models of gene evolution and species evolution, is expected to deliver more reliable orthology estimates, although at the expense of higher computational requirements. Figure 4.4 shows examples of both type of ortholog detection schemes where, for the example in Figure 4.4a, the corresponding phylogeny-based representation is given in Figure 4.4b.

Heuristic-based sequence similarity methods

Methods using direct sequence similarity, based on, so-called, Reciprocal Blast Hit (RBH) measure, aims at forming clusters of orthologs. Such methods often use a local alignment algorithm, and a particular scoring scheme for pairwise alignments. Example of an alignment algorithm is BLAST [9] and Smith-Waterman [196], while
4.3. COMPUTATIONAL APPROACHES FOR ORTHOLOGY INFERENCE

Example of a scoring scheme is the percentage of the trusted columns in the reference that have been correctly aligned [208]. Synteny, i.e., conservation of gene order, may also be used to optimize the inference in heuristics-based methods [7, 3].

We can identify two broad categories of sequence similarity methods, termed here as homolog-to-ortholog and ortholog-to-ortholog, where in both categories, orthology inference is a two step process. In homolog-to-ortholog methods, the first step consists of obtaining homologs using all-vs-all BLAST search. The second step may consist of using a clustering scheme, like Markov clustering [213], to identify clusters of orthologs based on the relative connectivity score among genes. Examples of homolog-to-ortholog methods are InParanoid [179] and OrthoMCL [121]. In ortholog-to-ortholog methods, the first step consists of querying a set of precomputed high-quality orthologs (a.k.a core-orthologs) against candidate sequences to construct core ortholog groups. In the second step, the resulting hits are queried against all genes in the reference taxa to construct final ortholog groups. One such method is HaMStR [59] where, in the first step, core-orthologs from a set of completely sequenced reference taxa, with a reliable phylogeny, are used to construct a profile Hidden Markov Model (pHMM) [58]. A faster version of the hmmssearch algorithm [60] is then used to search candidate sequences from a set of query taxa. In the second step, each candidate ortholog, identified by the pHMM, is verified using reciprocal BLASTP back to the reference taxa genomes before adding it to an appropriate ortholog group.

Heuristic methods are generally faster, but, in their basic form, are hard to generalize to more than two species. Recently, although, Altenhoff et al [8] used the idea of orthology graphs in the GETHOG algorithm to compute multiple species...
orthologies. Moreover, RBH misses many orthologs in duplication-rich clades such as plants and animals. In fact, a recent study shows that it may miss as much as 60% of orthologous relations [43]. One of the reason for such a significant number of misses may be attributed to the direct approximation of orthology relationships from sequence similarity measures without taking the species tree into account.

Phylogeny-based methods

Phylogeny-based methods use sequence information and a statistical model to infer a gene tree or, in a probabilistic model like the one described in Paper I, a set of gene trees, reconcile the gene tree with a species tree, and then, using Fitch definition [71], classify its vertices into speciations and duplications. Since in a gene tree, a speciation vertex, or a duplication vertex, may be an LCA for more than two leaves, Fitch definition allows us to define orthology relationship for two sets of genes as well, where each set can contain one or more genes. Phylogeny-based methods can be classified into parsimony and probabilistic methods.

Parsimony-based approaches use the most parsimonious reconciliation (MPR) of the gene tree and the species tree to identify speciation events, resulting in orthology pairs, and duplication events, resulting in paralogy pairs [82, 162]. Such methods are faster than probabilistic methods due to the possibility of using faster deterministic algorithms. However, MPR is biologically less realistic since it fails to account for different duplication rates. Moreover, it ignores the possibility that long evolutionary time-spans may imply that more than the minimal number of duplications have occurred. The inference may thus be far from the correct explanation of the evolutionary history [132], making MPR-based methods prone to incorrect orthology estimates [190].

Probabilistic methods infer orthology relationships by considering all possible reconciliations of a gene tree inside a species according to a mathematical model describing an evolutionary process. Sennblad et al [190] presented the first probabilistic method, named primeGEM, for orthology analysis, based on the DL model of Arvestad et al [18]. In an MCMC-based framework, all possible reconciliations were considered and weighted by their individual probabilities according to the gene evolution model, thus inferring biologically realistic orthology relationships. Paper I in the current thesis extends Sennblad et al [190] work by using the DLRS model [5], which additionally models the sequence evolution, thus inferring biologically more realistic orthology relationships.

Comparison of phylogeny-based and sequence similarity-based methods

Phylogeny-based methods have several advantages over sequence similarity-based methods. Firstly, by using gene tree-species tree reconciliations, they provide inference that is much closer to the definition of orthology and paralogy [71]. Secondly,
by their design, phylogeny-based methods can easily detect in-paralogs and group-orthologs, i.e., the orthologous relationship among two sets of genes. Thirdly, by using an explicit model of evolution, they can model phenomena like unequal rate of evolution and differential gene loss. Finally, given a ‘true’ reconciliation, we are guaranteed to obtain true orthology and paralogy estimates, while given the correct pairwise distances, reciprocal blast hit may still deliver incorrect orthology estimates [190].

There are, however, some shortcomings of phylogeny-based methods. Perhaps the major bottleneck is their computational complexity, especially for probabilistic methods. Moreover, since such methods utilize gene trees and a species tree, their performance is dependent on the performance of the alignment algorithm and the accuracy of species tree, respectively. Finally, their performance may also be affected by problems occurring in phylogenetic tree reconstruction, e.g., long branch attraction [159].

Future directions

Current methods infer orthology using a duplication-loss model by assuming that the genetic content is propagated solely by vertical descent from pre-existing genes either via speciation, which results in orthologs, or via duplication, which results in paralogs. However, it is known that prokaryotes and the microbial groups of eukaryotes, particularly protists, often exchange genetic material using LGT [13, 106, 133, 11]. A recent comparative study by Dalquen et al shows that current orthology inference methods perform poorly in the presence of LGT [42]. Orthology inference in presence of LGT is thus a promising research direction. Previously the DL and DLRS model has been used to devise probabilistic orthology methods namely primeGEM [190] and DLRSOrthology (Paper I in the current thesis) respectively. In the same vein, the DLTRS model may be used to devise an orthology inference algorithm, which can infer orthologous genes in the presence of LGT events.
Chapter 5

Overview of Papers in this Thesis

**Paper I:** We propose DLRSOrthology: a sound, comprehensive Bayesian MCMC-based method, based on the DLRS model [4], to compute orthology probabilities. It sums over the possible gene trees and jointly takes into account the current gene tree, all possible reconciliations to the species tree, and the, typically strong, signal conveyed by the sequences. For each gene tree in the set of possible gene trees, we use a dynamic programming algorithm to efficiently sum orthology probabilities over all reconciliations. By testing DLRSOrthology on synthetic and biological (primate and vertebrate) datasets, we show that our method provides quality predictions that surpasses PrIME-GEM: one of our previous probabilistic orthology method [190], and MrBayesMPR: a traditional approach based on MPR analysis of MrBayes-generated gene trees. We also present heuristics, based on receiver operating characteristics (ROC) curves, to estimate suitable thresholds for deciding orthology events.

**Paper II:** We present a two-phase approach, called MixTreEM-DLRS, for reconstructing a species tree in the presence of gene duplications and losses. In the first phase, MixTreEM, a novel structural EM algorithm based on a mixture model is used to reconstruct a set of candidate species trees, given sequence data for monocopy gene families from the genomes under study. In the second phase, PrIME-DLRS, a method based on the DLRS model [4], is used for selecting the best species tree. PrIME-DLRS can handle multicopy gene families since DLRS, apart from modeling sequence evolution, models gene duplication and loss using a gene evolution model [18]. We evaluate MixTreEM-DLRS using synthetic and biological data, and compare its performance to a recent genome-scale species tree reconstruction method PHYLDOG [34] and a fast parsimony-based algorithm Duptree [219]. Our method outperform PHYLDOG both in terms of accuracy and runtime, and Duptree in terms of accuracy. We observe that, in most cases, the analysis constituted only by MixTreEM may also be used for selecting the target species tree, yielding a fast and accurate method for larger datasets.
Paper III: We propose a probabilistic method to sample reconciliations of a gene tree with the species tree and to compute maximum a-posteriori probabilities. It uses an MCMC-based implementation of the DLTRS model [195], which is an extension of the DLRS model that additionally models LGT events. The method enables us to estimate gene trees having temporal information for duplication and LGT events. We evaluate the method on synthetic data to demonstrate good performance in terms of identifying gene tree edges corresponding to LGT events and the species tree edges between which LGT events have occurred. Based on its performance, we believe that the method has the potential to identify LGT highways.
Bibliography


BIBLIOGRAPHY


