Methods and analysis of genome-scale gene family evolution across multiple species

by

Matthew D. Rasmussen

B.S. Mathematics and Computer Science, University of Minnesota

M.S. Computer Science, Massachusetts Institute of Technology

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2010

Author_____
Department of Electrical Engineering and Computer Science
July 30, 2010

Certified by_____
Manolis Kellis
Associate Professor of Computer Science
Thesis Supervisor

Accepted by_____
Professor Terry P. Orlando
Chair, Department Committee on Graduate Students.

Methods and analysis of genome-scale gene family evolution across multiple species

by

Matthew D. Rasmussen

Submitted to the Department of Electrical Engineering and Computer Science on
July 30, 2010 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

ABSTRACT

The fields of genomics and evolution have continually benefited from one another in their common goal of understanding the biological world. This partnership has been accelerated by ever increasing sequencing and high-throughput technologies. Although the future of genomic and evolutionary studies is bright, new models and methods will be needed to address the growing and changing challenges of large-scale datasets.

In this work, I explore how evolution generates the diversity of life we see in modern species, specifically the evolution of new genes and functions. By reconstructing the history of the diverse sequences present in modern species, we can improve our understanding of their function and evolutionary importance. Performing such an analysis requires a principled and efficient means of computing the most probable evolutionary scenarios.

To address these challenges, I introduce a new model of gene family evolution as well as a new method SPIMAP, an efficient Bayesian method for reconstructing gene trees in the presence of a known species tree. We observe many improvements in reconstruction accuracy, achieved by modeling multiple aspects of evolution, including gene duplication and loss rates, speciation times, and correlated substitution rate variation across both species and loci. I have implemented and applied this method on two clades of fully-sequenced species, 12 *Drosophila* and 16 fungal genomes as well as simulated phylogenies, and find dramatic improvements in reconstruction accuracy as compared to the most popular existing methods, including those that take the species tree into account.

Lastly, I use the SPIMAP method to reconstruct the evolutionary history of all gene families in 16 fungal species including several relatives of the pathogenic species *C. albicans*. From these reconstructions, we identify several families enriched with duplications and positive selection in pathogenic lineages. Theses reconstructions shed light on the evolution of these species as well as a better understanding of the genes involved in pathogenicity.

Thesis Supervisor: Manolis Kellis
Title: Associate Professor of Computer Science

Contents

Acknowledgments

To my friend and the love of my life: *Xin*.

To my loving family for their support and guidance: *Amy, Dave, Megan, Jacob, George, Irene, Vivian, Ricky, Winston, and Dillon.*

Thanks to *Manolis* for his advice and guidance throughout my graduate career. In addition, thanks to the whole CompBio lab for their advice and insights.

Lastly, thanks to my thesis committee *Scott Edwards*, *Eric Alm*, and *Tommi Jaakkola* for fruitful discussions at various stages of this work.

> *Practice your form and the shot will follow... and, finish on your toes.*
>
> — Basketball Coach

Chapter 1

Introduction

There has never been a better time for studying genomes and evolution. This decade began with the publishing in 2001 of the first drafts of the human genome [88, 141]. Although the project began in 1990, over a majority of the sequencing was completed in only the last two years of the project, an outcome of the continually increasing improvements in sequencing technology and computational techniques [19]. Like Moore's law for the increase in microprocessor speed, DNA sequencing appears to have its own law. Throughout the project, the amount of sequenced DNA increased by 100 fold while matched by a 100 fold decrease in costs [19]. Now, as the decade comes to a close, we have witnessed the sequencing of several dozen animal, plant, and fungal genomes, along with thousands of bacterial and viral genomes. This data has completely changed our understanding of life at the molecular level and has enabled entirely new techniques and fields of study.

Going forward, even more ambitious goals are on the horizon. The 1000 Human Genomes Project [117] has set out to sequence the genomes of 1000 human individuals in order to better capture the true genetic variability present in the human population, with the hope that this will elucidate the genetic basis of many human diseases and the evolutionary history of the human population. Beyond human evolution, the Genome 10K Project has set its sights on sequencing over 10,000 vertebrate genomes in order to detail the full diversity of animal genetics [111]. This wealth of information will open the way to many new kinds of research questions and ultimately to a greater understanding of biology within ourselves and across our planet.

A key motivation for sequencing the genomes of so many diverse species has been to compare them. This has led to the development of a fruitful track of research called *comparative genomics*, where the genomes of multiple species are compared in order better understand their function. By identifying specific

patterns of similarity and differences between genomes, one can obtain many clues about the information they contain. This is because, by comparing genomes of modern species, we are really studying the effects of millions of years of evolution. Since evolution is the guiding force responsible for the structure and function of genomes as they exist today, learning about genomes and evolution goes hand in hand.

The field of *phylogenetics* has provided indispensable tools for tackling questions related to genome evolution. The basic problem the field addresses is the following: given a set of characters in modern species reconstruct the evolutionary history relating the species. The evolutionary history most commonly takes the form of a tree, called a *phylogeny*. The leaves or tips of the tree represent modern taxa, whether they be species, genes, or individuals, and the internal nodes of the tree represent ancestral states of the taxa. The branching patterns and branch lengths of the tree represent when and how the taxa have diversified over evolutionary time. Thus the phylogeny provides a concise summary of many evolutionary events and can provide a framework for posing and answering many questions about evolution. Given their importance, reconstructing phylogenies has been a primary concern for the field. Methods for reconstructing phylogenies have a long history [46, 127, 119] and new methods are continually developed to address a wide range of evolutionary questions.

Lately, there has been great interest in developing phylogenetic methods that are applicable for genome-scale data and questions. These methods have been very successful and take many forms [64, 89, 69, 145, 10]. One particular combination of phylogenetics and genomics has been the research program of *phylogenomics*, the study of all the *gene families* from multiple fully sequenced genomes [41]. A gene family is a set of genes that although appear in many different species, share a common ancestry. Through several mechanisms such as gene duplication and loss, gene families can expand and contract in copy number as well as diversify in sequence and function. By comparing genomes with the aid of phylogenetic trees describing each gene family, these duplication and loss events can be reconstructed and studied. In addition, the trees provide a way of transferring knowledge about a gene's function in one well-studied species to less well understood species or genes. By taking a phylogenetic view of comparative genomics, we can infer when new functions and classes of genes appear in evolutionary time.

The purpose of this work is to better understand how evolution can create and alter genes and their functions. Our focus will be on studying gene families in a phylogenomic framework. Our dataset will consist of gene sequences as they exist in dozens of modern day species, and from them we will infer the likely evolutionary scenarios that explain the diversity of species and genes we see today. By tracking the evolution of genes within a clade of species, we can correlate the changing genotypes with the changes observed in phenotype to better infer the function of genes and to learn how new functions arise. The

research questions posed here require a principled means of integrating many disparate types of information. This work introduces several new mathematical models and computational methods designed specifically to study the many forms of gene evolution.

1.1 Thesis contributions

In this thesis, I will discuss my work on developing new models and methods for understanding gene family evolution as well as specific discoveries made by their application. The thesis is organized as follows:

- I will begin with a review of studies in gene family evolution, and give examples of what kinds of lessons we can learn from gene families (Chapter 2).
- I will then review the relevant concepts from phylogenetics (Chapter 3) and phylogenomics (Chapter 4). I will also present our own study of the challenges in applying phylogenomics to *Drosophila* evolution (Chapter 4.4).
- From this case study, we can draw several insights for how to improve the reconstruction of gene evolution. In Chapter 5, I will present a new probabilistic model of gene family evolution that incorporates these ideas. We have implemented a method called SPIMAP, which uses this model to efficiently reconstruct the history of gene families.
- In Chapter 6, I will present a learning framework for how to learn the parameters of our model.
- From an analysis of 12 *Drosophila* genomes, we justify the assumptions of our substitution rates model (Chapter 7).
- Armed with the SPIMAP method, we apply it to the gene families of 16 fungal genomes to understand the evolution of pathogenicity (Chapter 8).
- In Chapter 9, I demonstrate the accuracy of our method using an extensive set of benchmarks from both real and simulated data.
- Lastly, I will discuss this implications of this work to the field, and possibilities for future directions (Chapter 10).

Chapter 2

Examples of gene evolution

2.1 The role of duplication in the creation of new genes

The question of where new genes come from has been an old and exciting question throughout the study of evolutionary biology. As early as 1939, well before the discovery of DNA, While Muller speculated that new genes may come from other genes, writing, "every gene from a pre-existing gene" [107]. As molecular biology began to develop thirty years later, Susumu Ohno more clearly advanced the idea of gene duplication as providing the major mechanism for the creation of new genes [112]. Now in the age of genomics, it is possible to thoroughly investigate this hypothesis and to truly track down the changing gene content of the genomes of thousands of species.

As DNA is replicated during cell division and meiosis, there is an opportunity for the daughter cells to inherit an altered copy of the DNA simply by random errors in the replication process. Some of these errors, can be as simple as single changes in DNA sequences (substitutions), or can effect larger segments of DNA by deleting or inserting long stretches of sequence. In extreme cases, entire genes, chromosomes, or genomes can be deleted or duplicated. Duplicated regions can appear in tandem to their original copy or be inserted on other chromosomes.

In the case of gene duplication, there are several theories about the fate of the new duplicate. Initially, the new duplicate will begin to acquire mutations independently of the original copy, thus diverging in sequence similarity and possibly function. Many of these mutations may be tolerated, given that the original gene is still present to fulfil its function. Eventually, one of three scenarios will occur: (1) the duplicate acquires too many mutations to properly encode a functioning protein and thus becomes a *pseudogene* (non-functionalization), (2) it will stumble upon a new function for which it will be selected (neofunctionalization), or (3) both gene copies acquire mutations such that they specialize in restricted forms of their

original function (subfunctionalization) [51, 97]. In either case, if the new duplicate acquires a function that becomes selected, the duplicate can be retained within the genome for long periods of time.

In a similar way, large scale changes in the genome can also occur due to gene loss. During processes such as DNA replication or repair, a segment of DNA may be deleted from the genome. If this deletion contains a gene, it can have a large impact on the organism or have a minor effect if the products of the gene are no longer necessary.

In the following sections, we will briefly review what is known about the rates of these events, and give a few examples of how evolution can use them to shape the genomes of many species.

2.1.1 Rates of gene duplication

It is estimated that gene duplication and loss is just as important for evolutionary change as sequence substitution. In terms of frequency, duplication and loss has been estimated to occur at a rate that is 10-40% that of substitutions in many species [29]. In mammalian genomes, it has been estimated from dog, mouse, and rat genomes that genes are gained and lost at a rate of 0.0014 events per gene per million years (my) [30]. Interestingly within the primates, the rate appears to have accelerated to 0.0024 events/gene/my, even while the rate of substitution has slowed [65]. However, these rates only describe the rate of gene duplications that are retained and not lost to pseudogenization. By counting the number of young duplicates, those with less than 1% silent site difference with the original copy, it is estimated that the underlying duplication rate is much higher at around 0.001-0.016 per gene per my within eukaryotic species [96, 61].

2.2 The role of gene duplication and loss in shaping gene families

By successively duplicating genes, large groups of similarly functioning genes can be created, called *gene families*. These families can grow and contract over time depending on the forces of selection, adaptation, and drift. By studying genomic sequences, it is now estimated that about 1.6-3% of gene families in flies, mammals, and yeasts undergo unusually high rates of change in gene copy number [29]. Studying these families can shed light on the possibilities of evolution for new gene innovations as well as to better understand the varying evolutionary pressures that different species experience.

2.2.1 The evolution of receptor gene families

An interesting and well-studied example of gene family evolution are the receptor gene families. Their copy number appears to be especially variable across species in both the vertebrate and fly clades. They are also a

good illustration of how only a small number of changes are needed to produce a new function, in this case the binding of a changing variety of ligands.

Olfactory receptors

Olfactory receptors (OR) represent a large class of gene families present throughout the mammalian species. The size of these families are quite variable, reaching as low as 400 functional genes in the primates and as many as 800-1200 in other mammals [109]. These families are often present in genomic clusters along several chromosomes. Primates (Human and Macaque) appear to have lost many OR receptors as nearly half of their OR sequences are pseudogenes. In contrast, the mouse and rat genomes appear to be continually expanding in OR families as determined by phylogenetic analysis [109]. Interestingly, all OR sequences in toothed whales have been pseduogenized, suggesting that although they have been inherited by their terrestrial mammalian ancestor, OR genes may not be needed for their current aquatic environment [104]. It is thought that OR gene families may be especially variable due to changing environmental demands of species.

Vision - opsin receptors

In vertebrates, vision is enabled by opsin genes that encode photoreceptors in the retina. Early identification and sequencing of four of these genes revealed their ancient evolutionary history [108]. Through gene duplication, the Rhodopsin (RHO) gene, which is expressed in rod cells and is very sensitive in low lighting, was duplicated several times to create several cone cell expressed pigment genes, which are sensitive to a range of wavelengths.

In most vertebrates, there exists four cone expressed opsins each of which is sensitive to a different wavelength: LW (long wavelength, red), MW (medium wavelength green), SW1 (short wave length, violet), and SW2 (short wave length, blue). The actual wavelength to which each opsin gene is sensitive varies across species and may be subject to natural selection [72]. In most mammals, only the LW and SW1 opsins are present and thus for these species only dichromatic vision is possible [139]. However in the human genome, the LW opsin has undergone a more recent gene duplication. This has been followed by additional sequence divergence, such that now a new opsin sensitive to green wavelengths has been recreated. This duplication likely occurred within old world monkeys after their divergence from new world monkeys, as the duplication is found only in old world monkeys, apes, and humans [72].

Lastly, there is some evidence that the gain of color vision may have been coordinated with the loss of olfactory receptors in primates [54]. However, whether these two processes actually influenced one another

is still debated [102].

2.3 Whole genome duplication

One of the more dramatic cases of genome evolution is that of *whole genome duplication* (WGD), where a daughter cell inherits two full copies of the entire genome. These events are traumatic for the cell, and lead to significant genomic instability [103]. Consequently, it is thought that only a handful of such events have occurred in the evolutionary history of modern species. Still, such events present the opportunity for significant gene innovation [78, 73, 118], even allowing entire pathways to duplicate and specialize [9].

2.3.1 Vertebrate 2R and 3R whole-genome duplications

In vertebrates, it has long been suggested that two rounds (2R) of whole genome duplication occurred before the radiation of the clade [112, 95]. This theory has more recently gained stronger support by a comparison of the human and mouse genomes with the *Ciona intestinalis* genome, a species which out-groups the vertebrates [27, 26]. By looking for regions of conserved gene order between *Ciona* and the vertebrate genomes, ancient blocks of duplicated chromosomes can be identified. The genome sequencing of additional important out-grouping species, such as the amphioxus, have also contributed greater support to the 2R hypothesis [118].

A common hypothesis is that the 2R duplications played an important role in providing raw genetic material for the development of increased morphological complexity in the early vertebrates. Recent genomic studies have found an enriched retention of gene duplicates involved in signal transduction, transcriptional regulation, neuronal activities, and developmental processes [118]. Two gene clusters in particular that have been studied in detailed with respect to the 2R are the HOX and MHC gene families [70, 77]. These families to this day continue to exist in four large duplicated blocks within the genomes of many vertebrates.

Lastly, a more recent third whole genome duplication (3R) has been identified at the base of the teleost fish clade [73]. Many genes in the descendant fish genomes have up to eight copies, even before considering individual family expansions.

2.3.2 Fungal whole-genome duplication

In fungi, genome sequencing has also revealed a WGD event in the ancestry of baker's yeast, *S. cerevisiae* [148, 78]. Earlier studies had questioned the existence of WGD in yeasts, due to the fact that a small percentage of gene families exist in a 2:1 ratio between *S. cerevisiae* and several hypothesized out-grouping

(pre-WGD) species [94, 93]. Instead, it was suggested that these numerous duplicated regions could be explained by many independent segmental duplication events that occurred over a period of time.

However, the sequencing of the *K. waltii* genome provided strong evidence for WGD and an explanation for the gene count anomaly. By comparing the two genomes, it became clear that *K. waltii* out-grouped the WGD event and thus for each region in *K. waltii*, exactly two pairs of duplicated regions in *S. cerevisiae* could be aligned. Still, many genes showed 1:1 orthology between the species, a signal previous studies had used to argue against WGD [94, 93]. Kellis *et al.* explained this by invoking massive gene loss after the WGD event. In fact, when genome alignment maps were made, clear sequence similarity could be seen surrounding many of these 1:1 orthologs indicating that a 2:1 relationship previously existed and was followed by gene loss. Further studies compared the genomes of many more species that proceeded the WGD (post-WGD species) [129]. In doing so, a clear picture of rapid gene loss could be determined, and individual events could be dated.

Many cases of neofunctionalization were also identified in the gene pairs . This was done by identifying gene families with "asymmetric divergence", that is one duplicate acquiring mutations at a significantly faster rate than the other [78, 130]. Genes that showed such a pattern were significantly enriched in protein kinases and regulatory proteins, such as cell-cycle transcription factors Swi5 and Ace2, and the filamentation factors Phd1 and Sok2 [78]. Some particular examples of novel gene functions include Sir3, a silencer in telomeres and mating type cassettes, which has been acquiring substitutions more than twice as fast as its duplicate Orc1, which performs origin-of-replication binding. Another example is Ski7 that now recognizes and represses non-polyadenylated messenger RNA for anti-viral defense, but has been derived from the translation-elongation function of Hbs1.

2.4 Future studies of gene family evolution

The evolutionary discoveries reviewed in this chapter are only a few examples of the many possible stories we can learn from evolutionary analysis of genomes. As more diverse and closely related genomes are sequenced, the opportunities for such discoveries will only increase. However, to truly exploit the power of such increasingly large datasets, we will need principled and scalable methods for performing systematic analysis of gene family evolution. In the following chapter, we will review the computational tools of phylogenetics and see how they can be applied to understanding gene family evolution.

Chapter 3

Phylogenetics

The field of phylogenetics provides an important theoretical foundation for studying evolution. In order to understand gene family evolution, we make use of several concepts from the field, which we review in this chapter.

3.1 The phylogeny

In phylogenetics, the goal is to understand the evolutionary history of a set of taxa, which can be species, populations, individuals, or even genes. The most commonly used representation of an evolutionary history is a tree call a *phylogeny* (Figure 3.1). The phylogeny depicts when and how a set of taxa have diversified. The tree is said to have a *topology* T, which is a graph that describes how the taxa are connected. This graph has a set of vertices $V(T)$ also known as "nodes" and it has a set of connecting edges $E(T)$ which are commonly called "branches". The leaves of the tree $L(T) \subset V(T)$ represent the modern or *extant* taxa, and the internal nodes $I(T) = V(T) \setminus L(T)$ represent the ancestral taxa. If the oldest node in the tree is known, it is called the *root* and the tree is said to be *rooted*. A rooting imposes a directionality on every branch in tree, such that there is a directed path from the root to every leaf. If the oldest node is not known, which may be the case in many analyses, the tree is said to be *unrooted*, and the branches do not have any particular directionality. Throughout this thesis we will most often work with rooted trees.

We will use several functions to discuss how nodes are related to one another within a rooted tree. For example, for a node $v \in V(T)$, we use $child(v)$ to represent the set of children of v. In most cases we will work with binary trees, and thus we will use $left(v)$ and $right(v)$ to represent the left and right children. We use $parent(v)$ to represent its parental node. Lastly, we use $b(v)$ to denote the branch $(v, parent(v))$.

The branches of tree can also be labeled with a measure of relatedness between the connected nodes.

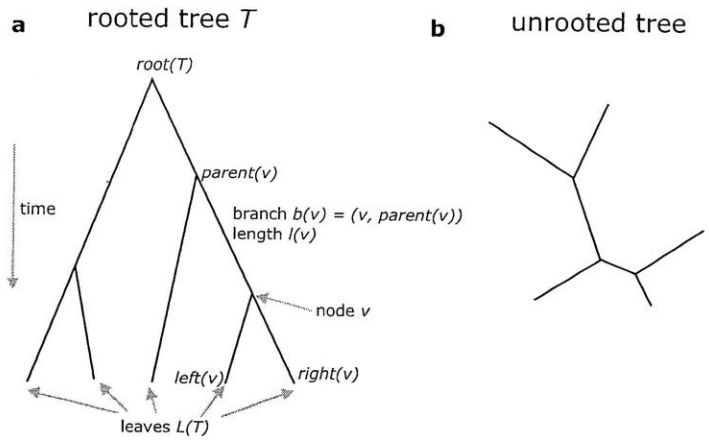


Figure 3.1: **Rooted and unrooted phylogenetic trees.** (a) In a rooted tree, the progression of time is known for each branch. An example use of our notation is given for a particular node v and its neighborhood in the tree T. (b) For an unrooted tree, it is not stated which node represented the oldest point in time.

This is often depicted in diagrams by using the length of branch. For a branch $b(v)$ its *branch length* $l(v)$ can either represent a duration of time (e.g. millions of years) or it can represent a measurement of divergence, such as nucleotide substitutions per site. All the branch lengths of a tree can be represented by a vector $\mathbf{l}$ such that $\mathbf{l} = (l(v_1), ..., l(v_N))$. Therefore, a phylogeny can be represented by the tuple $(T, \mathbf{l})$, which describes both its topology and branch lengths.

3.2 Gene trees and species trees

There are two kinds of phylogenies that we will use in this thesis. One is a *species tree* which describes how a set of species are related and the other is a *gene tree* which describes how a set of genes are related.

In a species tree, each leaf represents an *extant* (i.e. modern) species population and the internal nodes represent ancestral species populations. The bifurcations in the tree represent *speciations*, that is points in time when the ancestral population was divided into two or more populations that ceased to interbreed, and thus begin to evolve into a distinct species.

A gene tree is similar to a species tree, except that it describes how gene sequences are related: the leaves represent extant genes and the internal nodes represent the ancestral states of the genes. The bifurcations in a gene tree represent the act of replicating the DNA into separate sequences that then each evolve independently, and there are several different ways DNA can be replicated. To understand them it is best to think of a gene tree as evolving "inside" of the species tree (Figure 3.2).

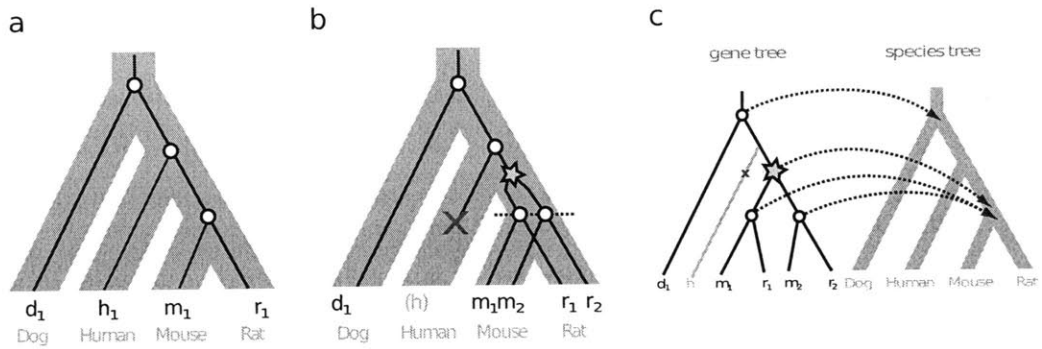


Figure 3.2: **Gene trees, species trees, and reconciliations.** (a) A gene tree (black lines) evolving inside of a species tree (blue lines). In the simplest case, the two trees are congruent and every bifurcation in the gene is a speciation event (white circles). (b) A more complicate scenario that depicts a gene family with one duplication event (star) and one loss event (red "X"). (c) A reconciliation (dashed arrows) maps gene nodes to species nodes, allowing one to infer gene duplications, gene losses, and speciations.

In the simplest case, the two trees are identical, indicating a single gene in the common ancestral species (the root) has been inherited as a single copy present in all modern species (Figure 3.2a). When a species population speciates, each of the descend populations inherit a copy of the gene. Thus, when a species tree bifurcates, it imposes a bifurcation in the gene tree called a *speciation node* (white circles in Figure 3.2).

In Figure 3.2b, a more complex evolutionary scenario is illustrated. Throughout this thesis, we will be mainly interested in modeling two kinds of evolutionary events, *gene duplications* and *gene losses*. A gene duplication creates an additional copy of a gene within the genome, which can then begin to acquire mutations independently. In the diagram, we represent duplications (stars) as a bifurcation in the gene tree that happens in the middle of a species branch. Notice, that once a gene duplicates, all descendant species will also inherit an additional copy. The second event we will model is a gene loss (red "X" in Figure 3.2b), where a gene is deleted from the genome at some point along a species branch and then is absent in all descendant species. Lastly, these two events can combine in multiple ways producing a great variety of scenarios for gene evolution. In this work, we call the set of genes that are descendant from a single gene in the common ancestral species the *gene family*.

3.3 Reconciliation

The relationship between a gene tree and species tree is captured formally using a particular mapping R called a *reconciliation*. Several definitions of the reconciliation have been defined [5, 59, 58, 32], but the simplest is a mapping from gene nodes to species nodes that defines the species to which each extant and

ancestral gene belongs [57, 114] (Figure 3.2c). In this setting, a gene tree is *congruent* if R is an isomorphic mapping between T and S (Figure 3.2a), and *incongruent* otherwise (Figure 3.2b).

For a given gene tree and species tree, there are several possible ways to reconcile them, and each implies a different set of duplication and loss events. Thus, several lines of research have been made in exploring these possibilities and developing algorithms for determining the "optimal" reconciliation. The earliest optimization algorithm, called Maximum Parsimonious Reconciliation (MPR), found the reconciliation for a given gene tree and species tree that minimized the number of implied duplications [114]. The method uses the concept of the Least Common Ancestor (LCA, also known as most recent common ancestor) to recursively define the mapping R. Given a mapping R of the extant genes $v \in L(T)$ to their known species $u \in L(S)$, the reconciliation can be defined as

$$R(v) = \begin{cases} R(v) & \text{if } v \in L(T) \\ R(LCA(left(v), right(v))) & \text{if } v \in I(T) \end{cases} \tag{3.1}$$

Given a reconciliation R, each internal node in the gene tree can be classified as a duplication or a speciation. In the parsimony model, there is a simple rule for determining duplication nodes. Specifically, a node is duplication if it reconciles to the same species as one of its children

$$dup = \{v : v \in I(T), R(v) = R(right(v)) \vee R(v) = R(left(v))\}.$$

The reconciliation may also imply loss events, the minimum number of which can be computed as follows. Losses appear in a gene tree as bifurcations that should have been present along a branch in the gene tree, but are not present because one of the lineages has been lost. These branches $(v, parent(v))$ occur whenever $R(v) \neq R(parent(v))$. If we let $p(u_1, u_2)$ represent the set of vertices in the path between vertices u_1 and u_2 (excluding u_1 and u_2), then the number of losses that occur across the gene branch $b(v)$, is $|p(R(v), R(parent(v)))|$. Thus, the set of losses is

$$loss = \{(v,s) : v \in I(T), s \in p(R(v), R(parent(v)))\}.$$

Algorithms for reconciliation

It has been shown that the LCA mapping can be computed in linear-time [44], although a worst-case quadratic algorithm works efficiently in practice [150]. The LCA mapping finds the unique reconciliation that minimizes the sum of the number of duplications and losses (also call the *mutation cost*) [98].

The LCA mapping also minimizes the number of duplications, however, it is not unique [98]. Lastly, other reconciliation algorithms have been developed that efficiently minimize the number of implied gene losses [13].

More recent work has explored several variations of the reconciliation problem. For example, most algorithms assume binary gene trees and species trees, however, non-binary formulations have also recently been considered [12, 142]. Although, traditionally most attention has been focused on modeling gene duplications and losses, more recent work has begun considering reconciliations in the presence of horizontal transfer events [84, 59, 67]. Lastly, many of these formulations are defined in a parsimony framework, however, probabilistic formulations are now being developed and are more frequently used [5, 33]. Going forward, additional work will be needed to combine these ideas into a unified framework that can describe all of the various evolutionary events that we observed in gene families.

3.4 Orthologs and paralogs

In addition to determining duplication and loss events, the reconciliation can also be used to determine several useful relationships between extant genes. Orthology and paralogy are two very popular ways of describing the evolutionary relationship between extant genes within and between species [49]. Two genes are *orthologs* if their most recent common ancestor represents a speciation (e.g. genes m_1, r_1 from Figure 3.2b) and two genes are called *paralogs* if their most recent common ancestor is a gene duplication (e.g. m_1, m_2 or m_1, r_2). If two genes are thought to be closely related, but it is unknown whether they are either orthologs or paralogs, then the term *homology* or *homologs* is used.

From these definitions, we can see that the reconciliation provides the necessary information for determining orthology and paralogy relationships. And in turn, to perform the reconciliation we need to reconstruct the gene tree for the given set of genes and we need a known species tree. Methods for systematically determining orthologs and paralogs using this approach are reviewed in Chapter 4.

In general, orthology is the most sought after relationship, because orthologs are commonly thought to maintain similar function after speciation, although not always. Therefore, orthology can be a good predictor of gene function between a well studied gene in one species and a poorly understood gene in another species. However, this form of functional prediction can be complicated if a gene has a paralog. As seen in the review of gene evolution (Chapter 2), duplications can lead to neofunctionalization and subfunctionalization of paralogs. Because of these possibilities, it is important distinguish between orthology and paralogy.

3.4.1 Determining orthology by sequence clustering

Although, reconciling gene trees and species trees is the most principled way to determine orthology and paralogy, it may become too computationally expensive for handling extremely large numbers of species or genes. Even though this thesis does not use such techniques for determining orthologs, these approaches are useful for determining larger sets of genes such as putative gene families.

The BLAST algorithm is often used to search a database of known sequences in order to find sequences that are similar to a query sequence [4]. Ideally, given a gene in one species we could find its ortholog in another species by using BLAST to find the most similar match or "BLAST hit". However, due to changing mutation rates over evolutionary time and the approximate nature of BLAST, the top BLAST hit can very often not be the correct ortholog [85, 89]. Yet, BLAST and other sequence similarity measures are still very useful for building up more sophisticated methods of orthology determination. Among these strategies are gene cluster databases such as Clusters of Orthologous Genes (COGs) [138], clustering methods such as OrthoMCL [90], and paralog determination methods such as In-paranoid [122].

3.5 Phylogenetic methods

Many algorithms for constructing phylogenies exist. The input to these algorithms is either a gene alignment or a pair-wise distance matrix, and the output is a phylogenetic tree. The problem has been posed in terms of distances (Nieghbor-joining[127], Least Squared Error[11]), maximum parsimony[50], maximum likelihood[46], and Bayesian methods[119].

3.5.1 Complexity of phylogenetic methods

The computational complexity of these methods and the theoretical limits of their accuracy have been important areas of research within theoretical computer science. Neighbor-joining has been shown to have cubic run-time [135], although faster approximations exist [134]. A quadratic distance-based method called "Fast Neighbor Joining" has also been developed, that while optimizes a slight variant of the Neighbor-joining criteria, has been shown to perform equally well [43]. In the 1980s, various definitions of the parsimony and compatibility problems were shown to be NP-complete [24, 23]. However, interestingly it was only recently determined that optimizing the maximum likelihood criteria is NP-hard [15]. The works of Addario-Berry *et al.*[2] and Tuffle and Steel [140] have supplied important concepts towards understanding the complexity of maximum likelihood phylogenetics.

3.5.2 Probabilistic model of evolution

In this thesis, we will build upon the probabilistic approach to phylogenetics. We review here the basic concepts that are most frequently used in phylogenetic probabilistic models.

The most popular probabilistic model for representing the evolution of a set of molecular sequences on a phylogeny was initially developed by Felsenstein for his Maximum Likelihood method for phylogenetic reconstruction [46]. In the model, we have a binary tree T, where the leaves of a tree are numbered $1,...,n$ and the ancestral nodes are numbered $n+1,...,2n-1$. The branches of the tree are numbered by the most recent of the two nodes it touches (e.g. branch i connects node i and $parent(i)$). For a tree, we have its topology T and the branch times $t_1,...,t_{2n-2}$, where t_i is the time between nodes i and $parent(i)$.

Our sequence data can be represented as a matrix $\boldsymbol{x}$ (n rows, m columns), such that $x_{i,j}$ is the j^{th} character of the i^{th} sequence and each sequence has length m. In most applications, we will be given sequence data for only the extant sequences $\boldsymbol{x}_1,...,\boldsymbol{x}_n$. The ancestral sequences $\boldsymbol{x}_{n+1},...,\boldsymbol{x}_{2n-1}$, on the other hand, will not be directly observed and thus we will have integrate over all their possible states.

With these definitions, the maximum likelihood method seeks to compute the maximum likelihood tree $(\hat{T},\hat{\boldsymbol{t}})$, where

$$\hat{T},\hat{\boldsymbol{t}} = \underset{T,\boldsymbol{t}}{\operatorname{argmax}} P(\boldsymbol{x}_1,...,\boldsymbol{x}_n|T,\boldsymbol{t}). \tag{3.2}$$

Thus, in order to develop this method, we must define a probability distribution for term on the right hand side of equation 3.2. Once a distribution is defined, we can determine a method for efficiently computing the argmax.

Defining the distributions: factoring by sites and branches

Before we can tackle the equation above, we must first define the distributions of our variables. We will make two assumptions about the process of sequence evolution in order to make the math and algorithm tractable.

First, assume *sites evolve independently*. Thus we can factor the probability along the sites j,

$$P(\boldsymbol{x}_1,...,\boldsymbol{x}_n|T,\boldsymbol{t}) = \prod_j P(\boldsymbol{x}_{1,j},...,\boldsymbol{x}_{n,j}|T,\boldsymbol{t}). \tag{3.3}$$

Now we can focus on defining a model for a single site as it evolves down the tree. To aid in this task, it is easier if all sequences (both extant and ancestral) are specified. To do this, we can express the right hand

side of equation 3.3 as a marginal of the joint distribution over all sequences

$$P(\boldsymbol{x}_{1,j},...,\boldsymbol{x}_{n,j}|T,\boldsymbol{t}) = \sum_{\boldsymbol{x}_{n+1,j},...,\boldsymbol{x}_{2n-1,j}} P(\boldsymbol{x}_{1,j},...,\boldsymbol{x}_{2n-1,j}|T,\boldsymbol{t}). \tag{3.4}$$

The next assumption we will make is that *branches evolve independently*. This means that a base $\boldsymbol{x}_{i,j}$ only depends on the sequence of its parent $\boldsymbol{x}_{parent(i),j}$ and the time along its branch t_i. This assumption allows us to factor the term in equation 3.4 as follows

$$P(\boldsymbol{x}_{1,j},...,\boldsymbol{x}_{2n-1,j}|T,\boldsymbol{t}) = P(\boldsymbol{x}_{1,j}|\boldsymbol{x}_{2,j},...,\boldsymbol{x}_{2n-1,j},T,\boldsymbol{t})P(\boldsymbol{x}_{2,j}|\boldsymbol{x}_{3,j},...,\boldsymbol{x}_{2n-1,j},T,\boldsymbol{t})...P(\boldsymbol{x}_{2n-1,j}|T,\boldsymbol{t}) \tag{3.5}$$

$$= P(\boldsymbol{x}_{1,j}|\boldsymbol{x}_{parent(1),j},t_1)P(\boldsymbol{x}_{2,j}|\boldsymbol{x}_{parent(2),j},t_2)...P(\boldsymbol{x}_{2n-1,j}) \tag{3.6}$$

$$= P(\boldsymbol{x}_{2n-1,j}) \prod_{i=1}^{2n-2} P(\boldsymbol{x}_{i,j}|\boldsymbol{x}_{parent(i),j},t_i). \tag{3.7}$$

Notice that equation 3.7 requires us to define only two things, a prior distribution of the root base $P(\boldsymbol{x}_{2n-1,j})$ and the distribution of how single bases evolve over a branch $P(\boldsymbol{x}_i, j|\boldsymbol{x}_{parent(i),j},\boldsymbol{t})$. The following sections will define these terms.

Defining the distribution: evolving a single branch

We now consider the probability distribution for a single site j evolving along a single branch i,

$$P(x_{i,j}|x_{parent(i),j},t_i).$$

There are many models for defining this distribution. They range from the simplest, Jukes Cantor (JC)[75], towards more complex ones such as Kimura's 2 parameters (K2P)[82], HKY[68], and the fully general GTR model [87]. All of these models make the same basic set of assumptions, and only differ in how many free parameters they have.

First, they all assume that the evolutionary process is *time reversible*, that is

$$P(b)P(a|b,t) = P(a)P(b|a,t). \tag{3.8}$$

You can think of $P(a|b,t)$ as a 4x4 matrix $S(t)$ with indices a,b. They also assume that this probability

matrix is *multiplicative*, such that

$$P(c|a,t_1+t_2) = \sum_b P(b|a,t_1)P(c|b,t_2) \tag{3.9}$$

$$S(t_1+t_2) = S(t_1)S(t_2). \tag{3.10}$$

In the Jukes Cantor model [75], every base substitutes into every other base with an equal rate α. From these properties, our substitution probability matrix S is then

$$P(a|b,t) = S(t) = \begin{pmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{pmatrix}, \tag{3.11}$$

where

$$r_t = \frac{1}{4}(1+3e^{-4\alpha t})$$
$$s_t = \frac{1}{4}(1-e^{-4\alpha t}).$$

The derivation of these equations are given in [37]. Notice that when t is zero we have the identity matrix and when t tends to infinity the sequence adopts the equilibrium frequency.

In Figure 3.3 we illustrate more examples of nucleotide substitution models and how they relate to one another. For example, the K2P model allows another parameter to model the different rates of transversions (purines to pyrimidines and *vice-versa*) and transitions (purine to purine, pyrimidine to pyrimidine). The HKY model allows three additional parameters on top of K2P in order to model a different equilibrium distribution than 25% for each base. The TrN model generalizes HKY by giving two parameters for the two kinds of transitions. And lastly there is the fully parametrized model General Time Reversible (GTR).

Defining the distributions: Modeling the root sequence

The last distribution that needs to be defined is the probability of the root base

$$P(x_{2n-1,j})$$

Often this is just the background distribution of bases used in the sequence model.

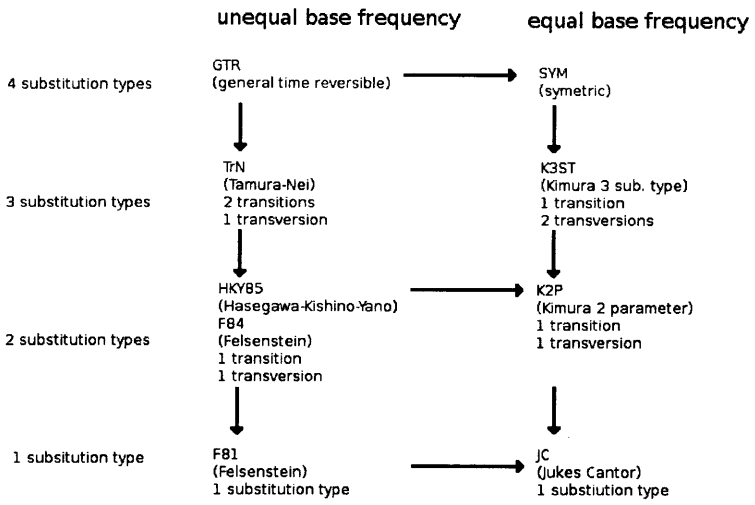


Figure 3.3: **The hierarchy of sequence substitution models.** Arrows point from general models to special cases. The main features are whether or not models assume equal background nucleotide frequencies and how many substitution classes they distinguish.

3.5.3 The Maximum Likelihood (ML) Algorithm

Now that we have fully defined our model of sequence evolution, we can estimate the maximum likelihood phylogenetic tree topology $\hat{T}$ and branch lengths $\hat{t}$.

Unfortunately, this is a difficult problem to solve. To appreciate the difficultly, consider how large the space of possible phylogenetic trees is. For example, for n leaves the number of unrooted topologies is

$$N_u = 3*5*7*...*(2n-5) = (2n-5)!!. \tag{3.12}$$

For rooted rooted topologies the number is even higher

$$N_r = (2n-3)*N_u = (2n-3)*(2n-5)!!. \tag{3.13}$$

Thus, computing the likelihood for each of these possible trees would be intractable. One might try to find ways of efficiently reducing the space of trees to consider, however, given that the ML phylogenetic problem is NP-hard [15] in turns out there exists no way of reducing this space while also finding the maximum likelihood tree exactly. Therefore, current approaches must use heuristic methods that search only a fraction of the total tree-space. These search methods traverse the space of possible trees by taking a

```
while searching:
    propose new T, t using local rearrangements (e.g. NNI, SPR)
    Calculate likelihood P(x_1, ..., x_n | T, t)
return T, t that achieved max likelihood
```

Figure 3.4: **Pseudo-code for ML algorithm.** The algorithm heuristically searches over the space of tree topologies and branch lengths by using a tree rearrangement operation such as Nearest Neighbor Interchange (NNI). After searching for a specified amount of time, the tree (T, t) with the highest likelihood seen thus far is returned.

proposed tree and performing local rearrangements to propose other possible trees (e.g. Nearest Neighbor Interchange (NNI)[147] and Subtree Pruning and Regrafting (SPR)).

In his paper, Felsenstein presented how one could combine a heuristic tree search with an efficient method for computing the likelihood of proposed tree [46]. See Figure 3.4 for an overview of the algorithm. Felsenstein showed that the likelihood of a tree can be efficiently computed using a technique now known as dynamic programming, which works by eliminating many redundant computations. This technique is also a special case of the sum-product algorithm. Recall, that the full factoring of the likelihood term is

$$P(\boldsymbol{x}_1,...,\boldsymbol{x}_n|T,\boldsymbol{t}) = \prod_j P(x_{1,j},...,x_{n,j}|T,\boldsymbol{t}) \tag{3.14a}$$

$$= \prod_j \sum_{\boldsymbol{x}_{n+1,j},...,\boldsymbol{x}_{2n-1,j}} P(\boldsymbol{x}_{1,j},...,\boldsymbol{x}_{2n-1,j}|T,\boldsymbol{t}) \tag{3.14b}$$

$$= \prod_j \sum_{x_{2n-1,j}} ... \sum_{x_{n+1,j}} P(x_{2n-1,j}) \prod_{i=1}^{2n-2} P(x_{i,j}|x_{parent(i),j},t_i). \tag{3.14c}$$

If we let $L_{i,j,a}$ represent the likelihood of the subtree rooted at node i with base a present at site j, we have the following recursive expression

$$P(\boldsymbol{x}_1,...,\boldsymbol{x}_n|T,\boldsymbol{t}) = \prod_j \sum_a L_{2n-1,j,a} P(a) \tag{3.15a}$$

$$L_{i,j,a} = \begin{cases} 1 & \text{if } x_{i,j} = a, i \leq n \\ 0 & \text{if } x_{i,j} \neq a, i \leq n \\ \sum_{b,c} P(b|a,t_{left(i)}) L_{left(i),j,b} & \text{if } i > n \\ P(c|a,t_{right(i)}) L_{right(i),j,c} & \end{cases} \tag{3.15b}$$

The values of $L_{i,j,a}$ can be thought of as a table, and its entries can be computed in $O(nm)$ runtime by using a postorder traversal of the tree.

Chapter 4

Phylogenomics

4.1 The phylogenomics problem

In the last chapter, we reviewed how the field of phylogenetics has had a long history of producing many powerful techniques for analyzing molecular sequences. With the continued advances in DNA sequencing technology, we now have dozens of whole-genome sequences from several clades of species, which provide a new opportunity for studying evolution on a large scale. Just as importantly, evolution provides a framework for better understanding the content and function of genomes. Consequently, phylogenetic analysis has had a growing presence in genome studies. Because of the scale of the data and the new challenges whole genomes present, many new phylogenetic techniques have been developed specifically for genomic applications.

This general approach of combining phylogenetics and genomics has been be called *phylogenomics* [41, 42] and its ideas play a central role in achieving our goal of understanding gene family evolution. Our goal is to reconstruct how each family of genes has expanded and contracted over evolutionary time in a clade of related species. By comparing these expansions and contractions to gene function and to the phenotypes of the species, we can characterize how evolution creates and changes new genes and functions. In a phylogenomics framework, our input data are fully sequenced genomes that are annotated with gene models. Using these gene annotations, we can cluster them by sequence similarity into families, and for each family, we can reconstruct a gene tree. Each gene tree is then reconciled to a common species tree, which allows us to infer orthologs, paralogs, and all evolutionary events, including gene duplications, losses, and horizontal transfers.

Although this growing field has made several successful advances, many computational challenges remain. Here, we review the previous work in the field, and by detailing its challenges, we will motivate our

own approach to reconstructing gene family evolution.

4.2 Exploiting genome-wide information for greater accuracy

As with any computational approach, the quality of the conclusions of a phylogenetic analysis heavily depends on the accuracy of the underlying methodology. Accordingly, there has been much recent work on measuring and improving methods for phylogenetic reconstruction for both species trees and individual gene family trees. Advances have come from increased sequencing data for both additional taxa and loci, as well as from new methods for leveraging that data.

4.2.1 The species-tree problem

For the problem of *species tree reconstruction*, many advances have been made by combining data across loci either by concatenating multiple aligned loci into a "supermatrix" [124, 16], combining multiple gene trees into a "supertree" [21], or using a model for how such loci are correlated and coordinated in their evolution [99, 91]. For example, in the BEST model [91], the correlated evolution of loci is captured by modeling a common species tree that constrains the evolution of each locus while still allowing some topological differences at each locus to occur via a coalescent process [144]. This is a hierarchical model, where a species tree defines the distribution for gene tree topologies and branch lengths, and those gene trees in turn define the distribution of sequences evolving down each gene tree. A probabilistic approach such as this allows one to use sequence alignments from multiple loci to estimate the posterior distribution of the species tree.

4.2.2 The gene-tree problem

We believe the problem of *gene tree reconstruction* will need a similar strategy for exploiting the abundant sequence data. Many recent efforts to reconstruct gene families in isolation (i.e. not accounting for their shared species tree or correlated evolution) have met many challenges. For example, the TreeFam project [89] had found that automatic methods of reconstruction (such as ML [46], MAP [119], NJ [127], and parsimony [47]) were not sufficiently accurate for systematic use, and thus relied on human curators to adjust trees using additional information from the species tree, syntenic alignments, and the relevant literature. In a study by Hahn *et al.* [63], simulations were used to study how errors in gene tree reconstruction propagate into later inference of gene duplication and loss events. In particular, the study showed that methods such as

Neighbor-Joining frequently make reconstruction errors that lead to a biased inference of many erroneous duplications in ancestral lineages followed by numerous compensating losses in recent lineages.

4.3 Modeling gene trees and species trees

Our work fits within a growing body of literature addressing the simultaneous modeling of gene and species evolution. In one branch of this field, the primary concern is to model orthologous loci whose phylogeny may become incongruent with the species phylogeny due to incomplete lineage sorting [99, 91]. In that case, the rate at which alleles propagate in a population is commonly modeled by the coalescent process [144], which defines how gene tree topologies and branch lengths are distributed across loci [120], and has been used to reconstruct both gene trees [69, 38] and species trees [99, 91], as well as many population related statistics, such as ancestral population sizes and recombination rates.

In another branch of the field, the loci of interest are those whose phylogeny is incongruent because of evolutionary events such gene duplication, loss, and horizontal transfer, and several models have been developed for each of these events. In the specific case of modeling duplication and loss, both probabilistic approaches [60, 5, 64] and non-probabilistic or parsimony-based methods have been developed [57, 114, 14, 145] to improve the reconstruction of either gene trees [5, 121, 145] or species trees [113]. Our focus will be in this part of the field and specifically on the goal of the probabilistic reconstruction of gene trees in the context of a common and previously determined species tree.

4.3.1 Existing work on gene tree reconstruction

For studying gene trees, Hahn *et al.* used the birth-death process to track changes in the number of paralogs in a gene family across several clades of species [64]. While it provides a way to look for significantly changing paralog copy counts, the method lacks a way of incorporating information from DNA or peptide sequences.

A method for incorporating such sequences was later developed by Wapinski *et al.* and was implemented in their SYNERGY gene tree reconstruction program [145]. The method makes use of peptide sequences by combining a species-aware Neighbor-Joining algorithm along with an optimization for minimizing duplications and losses while maximizing synteny (i.e. conserved gene order) between orthologs. However, this combination is *ad hoc* and non-probabilistic, making it difficult to determine the best way to weigh conflicting information [3]. For example, in the cases where synteny information can be misleading, such as cases of gene conversions, SYNERGY shows significantly reduced reconstruction accuracy, suggesting

that the primary sequence information is not sufficiently incorporated into the reconstruction (Figure 9.5).

A fully Bayesian model was proposed by Arvestad *et al.* that combined a model for gene duplications and losses with sequence evolution [5]. This was done by defining a prior for gene tree topologies and branch lengths using a birth-death process, which when combined with a sequence substitution model (e.g. JC69 [75]) produced a Bayesian method for gene tree reconstruction and reconciliation. One disadvantage of this approach was the assumption of a clock model for substitution (i.e. constant substitution rates).

In 2007, we introduced a distance-based maximum likelihood method for gene tree reconstruction that incorporates information from the species tree, but avoids the clock model assumption [121]. Our model decomposes substitution rates into gene-specific and species-specific components, which was motivated by our observation of substitution rate correlations across the genomes of 12 *Drosophila* and 9 fungal species (Chapter 4.4). By first learning parameters for gene and species-specific rate distributions from genome-wide information and then using that model to reconstruct gene trees, SPIDIR showed significantly increased reconstruction accuracy compared to several other popular phylogenetic algorithms at the time. However, despite these improvements, the approach was distance-based and thus did not fully utilize all of the information available in sequence data.

Recently, Åkerborg *et al.* have introduced PRIME-GSR, an extension of their previous work [5], which relaxes the clock assumption by using identical independent gamma distributions to model rate variation [3], however, no species-specific rate variation is learned or modeled. In our evaluations (Chapter 9) we find that modeling these rates can provide a significant benefit in gene tree reconstruction.

In summary, while much progress has been made in gene tree reconstruction, what remains missing is a principled, fast, and accurate method that incorporates all of these various models. In addition, freely available software is needed to facilitate further analyses in this field. In Chapter 5, I present a novel method that addresses these issues.

4.4 Drosophila challenges case study

In a recent paper of ours [121], we reviewed the accuracy of several existing phylogenetic methods for reconstructing gene trees. Phylogenetic accuracy has been extensively investigated on simulated data [126, 86, 137, 116], however for real datasets accuracy is difficult to assess since a ground truth is rarely known. The recently sequenced 12 *Drosophila* and 9 fungal species provided a unique opportunity to assess phylogenetic accuracy on real datasets, because the species are close enough evolutionarily to have a large fraction of genes in conserved gene order, known as *synteny*. For the 12 *Drosophila*, we found 5154 one-to-one

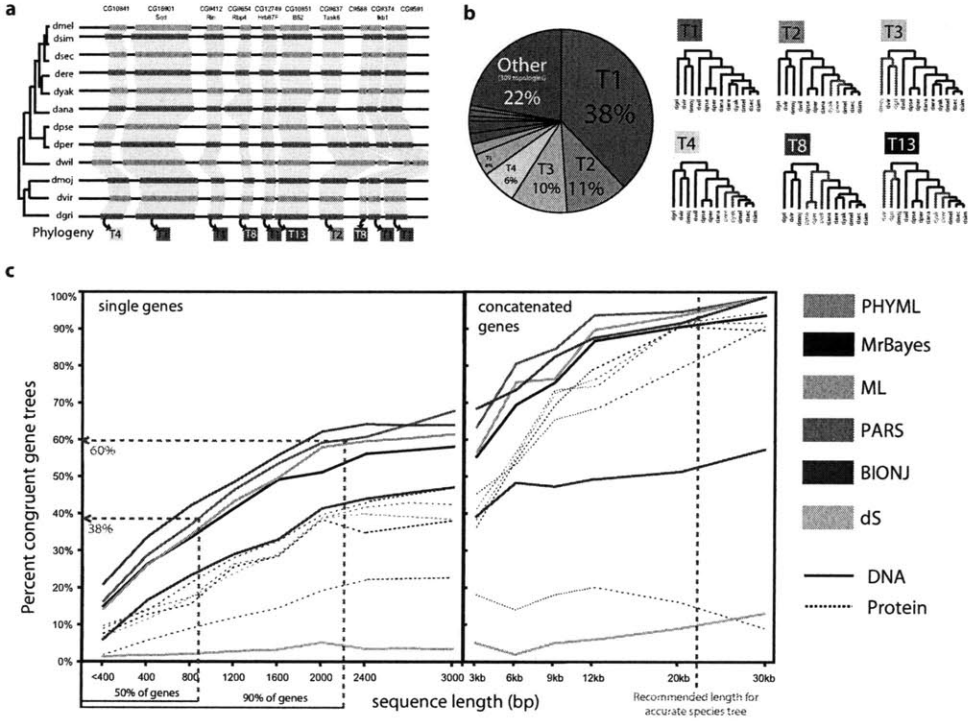


Figure 4.1: *Phylogenetic accuracy increases with sequence length.* (a) Maximum likelihood trees were reconstructed (using PHYML) for each one-to-one syntenic gene alignment in a particular section of the *Drosophila* genome. Although it is expected that all gene trees should match the species tree (topology T1), several different topologies (labeled T1-T13) were found. (b) Across the genome, 5154 one-to-one syntenic gene families have up to 310 different maximum likelihood gene tree topologies. The topology congruent with the species tree (T1) appears the most frequently at 38%. The next most frequent topologies T2-T4 are similar but differ by one or two branches (red branches). (c) When gene alignments are grouped by the number of ungapped sites, a clear trend in gene tree congruency and gene length is seen. This indicates that without using additional information, a typical gene family is too short in sequence length to reliable reconstruct its gene tree.

syntenic gene families or about one third of the fly genome. Synteny can be used as a independent line of evidence for orthology between genes. By identifying clusters of genes that are one-to-one syntenic across all 12 fly species (Figure 4.1a), we can obtain clusters of real genes where we expect the phylogenetic tree to contain no duplications or losses (i.e. be congruent to the species tree).

Our analysis showed that existing phylogenetic methods have a high level of inaccuracy for reconstructing gene families in the 12 *Drosophila* and 9 fungal species. For example, in the recently sequenced 12 *Drosophila* species, we reconstructed Maximum Likelihood gene trees of 5154 clusters of syntenically orthologous genes using the PHYML program [62]. We found a great variety of gene tree topologies (Figure 4.1b). Although the species tree topology did occur the most frequent, it only appeared for 38% of the gene alignments. We then grouped gene families into eight groups based on their gene length and asked how the congruence rate varied with gene length (Figure 4.1c). We found congruence to increase steadily with gene length, indicating that incongruence at smaller gene lengths were due to a lack of phylogenetic information (fewer characters). Most importantly, the median gene length of about 900 gapless alignment sites had only 38% reconstruction accuracy. Lastly, this trend was true regardless of whether we used nucleotide or peptide sequences or which phylogenetic method we used: Maximum Likelihood (PHYML [62]), Parsimony (PHYLIP [47]), Neighbor-joining (BIONJ [53]), and Maximum A Posteriori (MrBayes [125]).

We also studied the alignments that supported incongruent topologies. We found that while alternate topologies T2-T5 were reconstructed for as much as 4%-11% of the alignments, those rates fell to 1%-5% when we required all phylogenetic methods to agree. We also found that the alternate topologies showed significantly lower bootstrap support. Lastly, for the 62% of alignments that supported an incongruent Maximum Likelihood (ML) gene tree topology, only 5.7% did so with sufficient statistical significance ($P<.01$; SH-test [131]).

4.4.1 Overcoming low information within individual loci

This evidence along with several other measures of information content indicated that most loci lack enough information to confidently support one gene tree topology over the many other competing alternatives. Therefore, in order to make progress in gene tree reconstruction, we must look elsewhere for additional information. Fortunately, in the phylogenomic setting, where thousands of gene trees evolve within only a relatively small number of species, there is a large amount of shared information between gene trees that could be learned and applied to reconstruct gene trees more accurately. In the following chapter, we develop this idea further and demonstrate how the species tree and information about substitution rates can be greatly

informative and can be used to improve reconstruction accuracy.

Chapter 5

SPIMAP reconstruction method

5.1 Introducing the SPIMAP method

Here, we present SPIMAP, a Bayesian gene tree reconstruction method that incorporates within a unified framework models for gene duplication and loss, gene- and species-specific rate variation, and sequence substitution. We model gene duplication and loss using the birth-death process [5]. Similar to the other methods, we do not attempt to model incomplete lineage sorting or horizontal transfers, although approaches for doing so in the future could be useful. We have implemented a relaxed clock, defined using the rate variation model we have previously developed [121]. A key distinction of our method is that we employ an Empirical Bayes approach, where the parameters of the rate model are learned using a novel Expectation-Maximization (EM) training algorithm that incorporates sequence data across many loci. Once these parameters are estimated, we use them along with the species tree to reconstruct gene trees for thousands of sequence alignments from across the genome. Our method also achieves significant speed increases by using a novel tree search strategy derived from our gene tree topology prior. The SPIMAP software is written in C++ and is available for download at `http://compbio.mit.edu/spimap/`.

5.2 The phylogenomic pipeline

The reconstruction of gene trees for every gene family in several genomes typically requires a computational pipeline similar to the one shown in Figure 5.1a. Databases that have followed this general outline include TreeFam [89], Ensembl [143], and many others [71, 22], while other methods such as SYNERGY [145] perform similar tasks but not necessarily as separate consecutive steps. The general pipeline goes as follows: the input (blue boxes in Figure 5.1a) consists of nucleotide or peptide sequences for all genes in all genomes

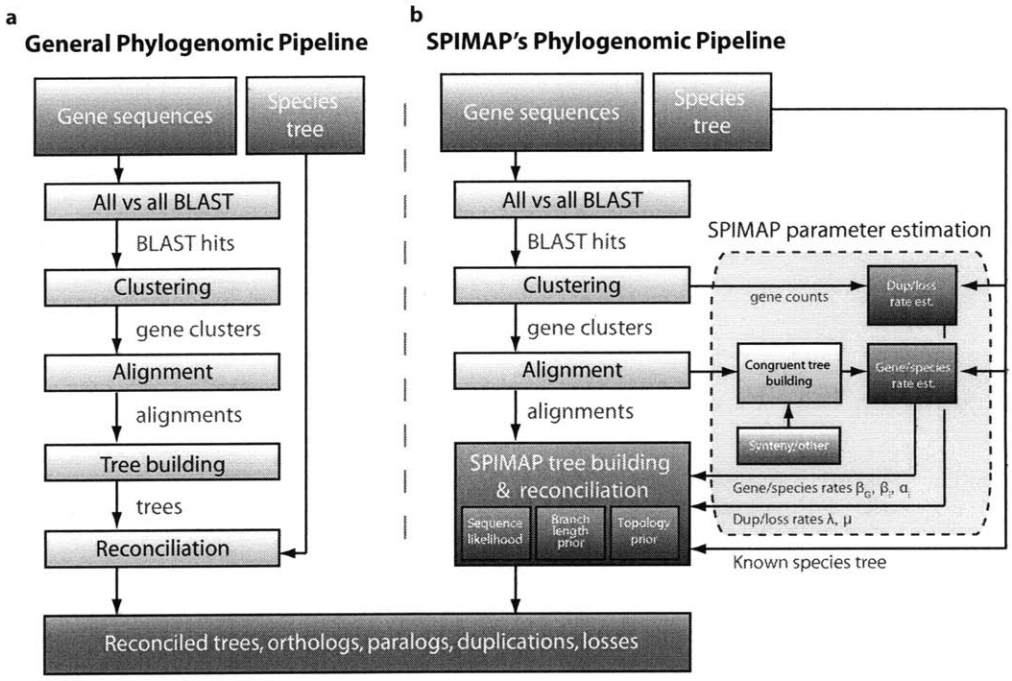


Figure 5.1: **Overview of the phylogenomic pipeline.** (a) The typical phylogenomic pipeline consists of several common steps, although particular implementations may vary. The pipeline input is the set of all gene sequences across several species and the known species tree relating the species (blue boxes). Gene sequences are then compared across species and clustered according to their sequence similarity, resulting in a set of homologous gene families. A multiple sequence alignment is then constructed for each gene family, followed by phylogenetic reconstruction of each aligned family to produce gene trees. Each gene tree is then reconciled to the known species tree in order to infer orthologs, paralogs, and gene duplications and loss events, which are the pipeline outputs (orange box). (b) Our phylogenomic pipeline follows similar steps, except that SPIMAP includes a model parameter estimation step (dashed light green box) for duplication and loss rates (learned from the per-species gene counts in the gene families resulting from the clustering step), and gene- and species-specific substitution rates (learned from a subset of trusted orthologous alignments supported by synteny or other information and congruent to the species trees). These learned evolutionary parameters are then used in a joint tree building and reconciliation step (dark green box), specifically informing our topology prior (duplication/loss model) and our branch length prior (gene/species-specific substitution model). The joint step also enables us to use the known species tree and duplication/loss model to rapidly score topology proposals and speed up tree search, in contrast to the traditional pipeline that only uses the known species tree in the reconciliation step.

under consideration as well as a species tree estimated prior to the pipeline computation using any method or information desired. Next, the sequences are compared with each other using a method such as an all-vs-all BLAST search [4] or HMMER [39]. The BLAST hits are then clustered using a method such as OrthoMCL [90] or a method like that of PHIGs [28] in order to form clusters of highly similar genes that are likely to represent gene families. For each cluster, a multiple sequence alignment is then constructed (e.g. MUSCLE [40]) followed by gene tree reconstruction using a phylogenetic algorithm (e.g. PHYML [62], BIONJ [53], or MrBayes [125]). Lastly, a *reconciliation* algorithm is used to compare each gene tree to the species tree in order to infer all duplication and loss events, as well as all ortholog and paralog relationships. Reconciliation methods include Maximum Parsimony Reconciliation (MPR) [114, 150], RAP [34], and Notung [14], each of which take different approaches to inferring gene duplication and loss events in presence of possibly uncertain gene trees. The duplications, losses, orthology, paralogy, and the gene trees themselves typically constitute the outputs of a phylogenomic pipeline (orange box; Figure 5.1).

The pipeline we have constructed for SPIMAP follows the same general structure (Figure 5.1b). For clustering, we have implemented our own method [10] similar to that of PHIGs. For multiple sequence alignment, we have used the MUSCLE [40] program. In contrast to other methods, however, ours takes an Empirical Bayes approach by including a "training" step (dashed green box; Figure 5.1b) which supplies several species-level evolutionary parameters to SPIMAP's gene tree reconstruction step. In the training step, we estimate the average genome-wide gene duplication and loss rates $\theta_t = (\lambda, \mu)$ based on gene counts within each gene family cluster using a method similar to that of Hahn *et al.* [64] (Chapter 6.2). We also estimate substitution rate parameters $\theta_b = (\alpha_G, \beta_G, \alpha, \beta)$ based on a subset of the alignments using a novel EM method for (Chapter 6.3). These parameters are then used in a combined gene tree reconstruction and reconciliation step (dark green box; Figure 5.1b) performed simultaneously within a single probabilistic model. From this model, we compute the the maximum *a posteriori* (MAP) gene tree using a novel rapid gene tree search that incorporates information from the species tree and from duplication and loss rates. In the following sections, we will discuss how we compute the posterior probability of a gene tree and describe the details of our rapid tree search.

5.3 Gene tree and species tree definitions

We define a *gene family* as the set of all genes descended from a single gene in the last common ancestor (LCA) of all species in consideration. We represent the rooted phylogenetic tree of n genes by a tree with topology $T = (V, E)$, which describes the set of nodes (vertices) $V(T)$ and a set of branches (edges) $E(T)$

of the tree. The leaves $L(T) \subset V(T)$ of a gene tree represent observed genes from extant species while the internal nodes $I(T) = V(T) \setminus L(T)$ represent ancestral genes from ancestral species. We will use several functions to discuss how nodes are related to one another. For example, we use $child(v)$ to represent the set of children of v, $left(v)$ and $right(v)$ to represent the left and right children, and $parent(v)$ to represent its parental node. For any node v, we use $b(v)$ to denote the branch $(v, parent(v))$ and $l(v)$ to be the length of that branch, measured in substitutions per site. Lastly, we use $\mathbf{l}$ to denote the vector of all branch lengths of a tree, namely $\mathbf{l} = (l(v_1), ..., l(v_{(2n-2)}))$. Thus, a *gene tree* is represented by the tuple $(T, \mathbf{l})$.

In addition, we will also consider a phylogeny S relating species, called a *species tree*. The branch lengths $\boldsymbol{t}$ of S are expressed in units of time (e.g. millions of years) and are thus typically ultrametric. For a node $u \in V(S)$ we express its length as time $t(u)$. We will assume all trees are rooted and all nodes have at most two children.

Each gene tree can be viewed as evolving "inside" the species tree (Figure 5.2a). A *reconciliation* R is a mapping from gene nodes to species nodes that defines the species to which each extant and ancestral gene belongs [57] (Figure 5.3a). In this setting, a gene tree is *congruent* if R is an isomorphic mapping between T and S, and *incongruent* otherwise. Also, all internal nodes of a gene tree represent either *gene duplication* or *speciation* events (represented as stars and white circles, respectively; Figure 5.2a).

5.4 Generative model of gene family evolution

In our generative model, gene trees are generated in three steps: given a species tree with specified topology and speciation times, (1) we first generate a gene tree topology and duplication times by repeated use of a birth-death process, (2) we then generate substitution rates from gene and species-specific distributions, and (3) lastly, we use these rates to generate molecular sequences according to a continuous-time Markov process (Figure 5.2).

The parameters of our model are $\theta = (S, \boldsymbol{t}, \theta_t, \theta_b)$, where S and $\boldsymbol{t}$ are the species tree topology and branch lengths, θ_t are the topology parameters λ and μ, and θ_b are the branch length parameters α_G, β_G, α, and β, the details of which are given below.

(1) Generating topology and divergence times. We use the gene duplication and loss model first developed by Arvestad *et al.* [7] which is based on a repeated use of the birth-death (BD) process [45] to define the topologies and branch lengths (in units of time) of a gene tree evolving "inside" a species tree (Figure 5.2a).

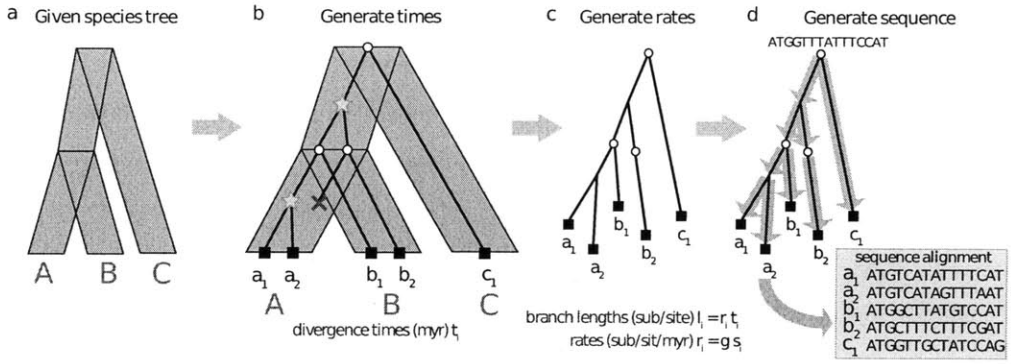


Figure 5.2: SPIMAP's **generative model.** (a) Species tree S and divergence times $\boldsymbol{t}$ are given as input. (b) A gene tree T (black lines and labels) is evolved inside the known species tree according to a duplication-loss model. The gene tree bifurcates either at speciation events (white circles) at species tree nodes, or at duplication events (stars) along species tree branches. Gene tree lineages can also terminate within a species branch at gene loss events (red "X"). (c) Substitution rates are generated according to our relaxed clock rates model of species-specific and gene-specific substitution rates. (d) Lastly, sequences are evolved down the resulting gene tree according to a continuous-time Markov process to produce a sequence alignment (yellow box).

The BD process is a continuous-time process that generates a binary tree according to a constant rate λ of lineage bifurcation (representing gene duplication) and rate μ of lineage termination (representing gene loss). After running a birth-death process for a time t, all lineages that exist at time t are called *surviving*, while all others are called *extinct*. A node is *doomed* if it has no surviving descendants. The BD process has been used widely in phylogenetics [60, 5, 64], although typically for defining priors for species trees [119].

The gene duplication and loss (DL) model is defined by repeatedly using the BD process to generate a gene tree. To initialize, we begin with a single gene node v reconciled to the root of S (i.e. $R(v) = root(S)$) and mark it as a speciation node. We then recursively apply the following: (1) for each speciation node v at the top of a species branch $b(u)$ of length $t(u)$, we generate a tree according to the birth-death process for $t(u)$ units of time. (2) For each newly created node w, we record its reconciliation as $R(w) = u$. (3) For each w that survives across that species branch, we mark it as an *extant gene* if u is a leaf species, otherwise mark it as a speciation. (4) We recursively apply steps 1-3 until all speciation nodes have been processed. (5) We mark all nodes in the gene tree not marked as *extant genes* or *speciations* as *duplications*. (6) As a post-processing step, we prune all doomed lineages, namely lineages with no extant descendants.

(2) Generating substitution rates. We use a relaxed clock model where substitution rates are allowed to vary between lineages (Figure 5.2b). Each branch has a length $l(v)$ (measured in substitutions/site)

which is the product of a duration of time $t(v)$ and a substitution rate $r(v)$. The times are given by the DL model. The substitution rates indicate the number of substitutions per site per unit time and are described by a rates model. Previously [121], we developed a rates model that captured the substitution rate $r(v)$ as the production of two components, a gene-specific rate and a species-specific rate. Here, we define these components with the following distributions:

(a) For each gene family j, the *gene-specific rate* g_j scales all rates in a tree. We represent the gene rate as a random variable G_j that is distributed across families as an inverse-gamma distribution with shape and scale parameters, α_G and β_G. Without loss of generality, we constrain G_j to have a mean value of one across all gene families (i.e. $\alpha_G = \beta_G + 1, \alpha_G > 1$). Thus we have,

$$P(G_j = g_j|\beta_G) = InvGamma(g_j|\alpha_G = \beta_G + 1, \beta_G). \tag{5.1}$$

(b) For each branch $b(v_k)$, the *species-specific rate* s_k defines a rate specific to that branch in the gene tree. It is represented by a random variable S_k that has a gamma distribution whose scale and shape parameters (α_i, β_i) depend on the species $u_i = R(v_k)$. This allows one to model rate accelerations and decelerations that are specific to a species i and exists across all genes of that species. Thus,

$$P(S_k = s_k|\alpha_i, \beta_i) = Gamma(s_k|\alpha_i, \beta_i), \text{where } u_i = R(v_k). \tag{5.2}$$

We also assume that each S_k is independent of the others and of the gene rate G. Given these definitions for the substitution rate, we can then express the branch length $l(v_k)$ of a gene tree j as

$$l(v_k) = r(v_k) \times t(v_k) = g_j \times s_k \times t(v_k). \tag{5.3}$$

In total, our rate model has parameters $\theta_b = (\beta_G, \alpha, \beta)$, where $\alpha = (\alpha_1, ..., \alpha_m)$, $\beta = (\beta_1, ..., \beta_m)$, and m is the number of species branches $|E(S)|$.

(3) Generating sequence. After generating a gene tree with a topology, divergence times, and substitution rates, we finally evolve a molecular sequence down the tree using a continuous-time Markov chain to model sequence substitution. Specifically, we have implemented HKY [68] to generating nucleotide sequences. See Chapter A.1 for an overview of the model. The HKY process uses the branch lengths $l(v_k) = r(v_k)t(v_k)$ as parameters for sampling derived sequences. Only sequences on the leaves of the tree are emitted, whereas ancestral sequences are hidden (Figure 5.2c). In our current formulation, sequence insertion and deletion

(indels) are not modeled. Instead, gaps in the sequence alignment are treated as missing data.

5.5 Maximum *a posteriori* reconstruction of gene family evolution

In our current implementation of the algorithm, we compute the maximum *a posteriori* (MAP) gene tree according to our model. Thus, we seek to calculate

$$\hat{\mathbf{l}}, \hat{T}, \hat{R} = \underset{\mathbf{l},T,R}{\operatorname{argmax}} P(\mathbf{l}, T, R|\mathbf{D}, \theta) \tag{5.4a}$$

$$= \underset{\mathbf{l},T,R}{\operatorname{argmax}} P(\mathbf{D}|\mathbf{l}, T, R, \theta) P(\mathbf{l}|T, R, \theta) P(T, R|\theta) / P(\mathbf{D}|\theta) \tag{5.4b}$$

$$= \underset{\mathbf{l},T,R}{\operatorname{argmax}} P(\mathbf{D}|\mathbf{l}, T) P(\mathbf{l}|T, R, \theta) P(T, R|\theta). \tag{5.4c}$$

The first term in equation 5.4c is the likelihood of a gene tree with branch lengths $\mathbf{l}$ and topology T given the sequence data $\mathbf{D}$. The probability is defined by the sequence evolution model (e.g. HKY) and can be computed efficiently using the pruning algorithm [46], which we have implemented for SPIMAP. Because this model only depends on the topology and branch lengths of the gene tree, the likelihood term is conditionally independent of the reconciliation R and parameters θ.

The prior of our model is factored into two terms: the prior of the topology and the prior of the branch lengths. The topology prior $P(T, R|\theta)$ is defined based on the duplication-loss model and it can be computed efficiently (Chapter 5.6) [6]. We show that factoring out the topology prior from the branch lengths also provides a unique advantage for fast tree search (Chapter 5.7.4).

Lastly, the branch length prior $P(\mathbf{l}|T, R, \theta)$ represents the probability of observing of gene tree branch lengths $\mathbf{l}$. This prior incorporates both divergence times of duplications in the birth-death process as well as the distribution of substitution rates. We present how to compute this term numerically (Chapter 5.7).

5.6 Computing the topology prior

The topology prior $P(T, R|\theta)$ (from equation 5.4c) helps SPIMAP reconstruct gene trees that have plausible patterns of gene duplication and loss. For completeness, we describe how to compute this term.

According to the duplication-loss (DL) model [5, 6], the birth-death (BD) process is repeatedly used to generate the gene tree topology T as it evolves from the root of the species tree S to the leaves. Therefore, T can be viewed as a union of several subtrees, each of which was generated by one BD process. Since these processes are independent of one another, we can view the topology prior $P(T, R|\theta)$ of gene tree T as

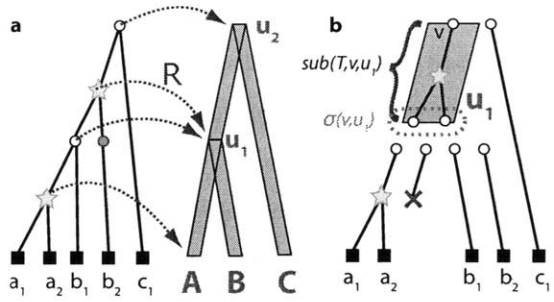


Figure 5.3: **Reconciliation and duplication subtrees.** (a) A reconciliation R maps gene nodes to species nodes for both speciation events (white circles) and duplication events (stars). *Implied speciation nodes* (gray circle) are then inferred based on the reconciliation. (b) Our algorithm breaks the gene tree T into subtrees $sub(T, v, u_1)$ where the subtree root v is a speciation and the subtree leaves $\sigma(v, u_1)$ are the next speciation nodes below v that reconcile to species u_1.

a product of the probabilities of the BD process generating each of the subtrees. Performing this factoring is the key step in computing the topology prior, but, there are two additional caveats to consider: (1) how to account for lineages in the gene tree that are hidden from observation due to extinction and (2) how to account for labeled and unlabeled nodes in the gene tree. By combining these ideas, we can compute the prior of a gene tree topology.

5.6.1 Factoring the gene tree

Given a gene tree topology T, we first decompose it into the subtrees that were generated from each individual BD process (Figure 5.3). We call each of these subtrees *duplication subtrees* since all of their internal nodes consist of duplication nodes. To identify these subtrees, first notice that each speciation node v is the root of two such subtrees. If v has reconciliation $R(v) = w$ and $w \in V(S)$, then the two subtrees perfectly reconcile within the child species branches $left(w)$ and $right(w)$. Also notice, that the leaves of each duplication subtree are either speciation nodes or extant genes.

Some speciation nodes (e.g the grey node in Figure 5.3a) may be initially hidden in a gene tree due to gene losses. We call such nodes *implicit speciation nodes* and they can be added to a gene tree by identifying gene tree branches that span multiple branches in the species tree (e.g. branch b_2 in Figure 5.3a). If a given gene tree T lacks implied speciation nodes, we can add them by locating each v and $w = parent(v)$ where $parent(R(v)) \neq R(w)$. Next, the edge (v, w) is replaced by a new speciation node x and two new edges (v, x) and (x, w), while setting $R(x) = parent(R(v))$. This procedure can applied repeatedly until all implied speciation nodes are identified.

When all speciation nodes are explicit, we can identify duplication subtrees by partitioning the gene tree at all speciation nodes $spec(T)$ (Figure 5.3). We denote a particular subtree as $sub(T,v,u)$, where $v \in spec(T)$ is the root of the subtree and $u \in child(R(v))$ is the species to which the leaves $L(sub(T,v,u))$ reconcile. The leaves are defined by the set

$$\sigma(v,u) = \{w : w \in spec(T) \cup L(T), R(w) = u, w \in V(T_v)\}, \tag{5.5}$$

where T_v is a subtree of T containing node v and all of its descendants.

For each duplication subtree, we can derive its probability from the BD process [119]. First, for a BD process with a birth rate λ and death rate μ, the probability that 1 lineage will leave s survivors after time t is

$$p(s,t) = (\lambda/\mu)^s p(1,t)(p(0,t))^{s-1}, \tag{5.6}$$

where

$$p(0,t) = \frac{\mu(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}} \qquad p(1,t) = \frac{(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\lambda-\mu e^{-(\lambda-\mu)t})^2}. \tag{5.7}$$

Second, for s survivors there are $\xi_s = s!(s-1)!/2^{s-1}$ equally likely *labeled histories*, which are leaf labeled topologies whose internal nodes are order by their time. Thus, for a topology T with s leaves and $H(T)$ labeled histories, its probability is

$$P(T|t,\lambda,\mu) = \frac{H(T)}{\xi_s} p(s,t), \text{ where} \tag{5.8}$$

$$H(T) = \prod_{v \in I(T)} \binom{|I(T_{right(v)})| + |I(T_{left(v)})|}{|I(T_{right(v)})|}. \tag{5.9}$$

5.6.2 Doomed lineages

In addition to factoring the tree, there are two caveats to consider. The first to consider is the possibility of lineages in the gene tree that are hidden from observation because they have gone extinct, i.e. they leave no descendants in the leaves of the species tree. We call such lineages *doomed*, and this extinction process must be accounted for in our topology prior.

Let $d(u)$ be the probability that a lineage starting at node u in the species tree will be doomed, that is losses occur such that no descendants exist at the leaves of the species tree. This probability $d(u)$ is the product of the probability of extinction occurring in both the left and right subtrees beneath node u. For a

child branch $b(c)$ where $parent(c) = u$, we must consider two possibilities. Either the gene lineage goes extinct in $b(c)$ with probability $p(0,t(c))$ (Equation 5.7), or it survives and leaves i survivors, each of which themselves are doomed with probability $d(c)$. Thus, this probability can be expressed recursively as

$$d(u) = \begin{cases} \prod_{c \in child(u)} \sum_{i=0}^{\infty} p(i,t(c)) d(c)^i & \text{if } u \in I(S) \\ 0 & \text{if } u \in L(S) \end{cases} \tag{5.10}$$

The value $d(u)$ can be computed efficiently for each node u in the species tree S by dynamic programing following a post-order traversal of S.

5.6.3 Labeled and unlabeled nodes

The second caveat of the topology prior computation is distinguishing between labeled and unlabeled nodes within the gene tree. In Equation 5.8, we give the probability of a BD process generating a labeled topology T. Each duplication subtree $sub(T,v,u)$ is generated by one BD process, however, only duplication subtrees with extant leaves (i.e. $L(sub(T,v,u)) \subseteq L(T)$) are labeled topologies. All other duplication subtrees have leaves that are speciation nodes, and thus are unlabeled topologies.

To properly account for labeled and unlabeled nodes, we envision the duplication-loss (DL) model as a three step process. First, a gene tree T' is generated by repeated use of the BD process, after which as all extant and speciation nodes are labeled. The probability of this tree is $P(T',R|\theta)$ and it can be computed by factoring T' into duplication subtrees, each of which has a known probability (Equation 5.8).

Second, a mapping U is applied to T' that removes all labels to produce an unlabeled gene tree T''. The probability $P(T'',R|\theta)$ is thus the sum of the probability of each T' that becomes T'' after removing labels,

$$P(T'',R|\theta) = \sum_{\{T':T''=U(T')\}} P(T',R|\theta). \tag{5.11}$$

We call two trees T'_i and T'_j equivalently labeled if $U(T'_i) = U(T'_j)$. Since equivalently labeled trees T'_i all have equal probability, the probability $P(T'',R|\theta)$ is simply the probability of T' times the number of equivalent labelings. The number of equivalent labelings is computed as a product of correction terms, one for each duplication subtree. Specifically, for each internal subtree T_2 (i.e. leaves are speciations nodes) we multiply by the term $N_2(T,T_2,R)$ and for each external subtree T_2 (i.e. leaves are extant genes) we multiply by $N_1(T_2,R)$. We derive these terms in the following sections.

In the third and final step, labels are added back to the leaves of T'' to create our desired leaf labeled

gene tree topology T. Since each labeling is equally likely to be generated by this process, the probability $P(T,R|\theta)$ is $P(T'',R|\theta)$ divided by the number of ways to relabel T''. This final correction factor is $1/N_1(T,R)$ and we derive this next.

Step 3: the number of ways gene names can be added to a reconciled gene tree

The easiest step to describe first is step 3. In step 3, we add back labels to the leaves of T'' to create a labeled topology T. Let $N_1(T,R)$ be the number of ways to relabel T'' into T. Also assume that we can compute $P(T'',R|\theta)$, which is the probability of the unlabeled reconciled gene tree (T'',R) being generated by steps 1 and 2. Since each possible relabeling is equally likely in the DL model, the topology prior can be computed as

$$P(T,R|\theta) = \frac{1}{N_1(T,R)} P(T'',R|\theta). \tag{5.12}$$

Defining $N_1(T,R)$ can be done in the following way. Notice that after step 2, the reconciled gene tree (T'',R) has leaves that are only distinguished only by their species, but have no gene names. An example of such a gene tree in Newick notation would be "((Scer,Scer),Spar)", where "Scer" and "Spar" represent the species *S. cerevisiae* and *S. paradoxus*, respectively.

We can think of species as "coloring" the leaves of the gene tree T''. Define a *colored topology* as an unlabeled topology whose leaves are colored. Imagine we have a set of gene names (e.g. "Scer1", "Scer2", "Spar1","Spar2", etc.) which must be used to label the leaves of the gene tree. Notice, that each gene name ("Scer1") must be assigned to a leaf that already belongs to the corresponding species ("Scer"). Our question is, how many ways $N_1(T,R)$ can we assign these names?

From the set of names, we can calculate how many names $c(T,u)$ each species u has. Expressed in our notation we have

$$c(T,u) = |\{v : v \in L(T), R(v) = u\}|. \tag{5.13}$$

Naively, we might then conclude that

$$N_1(T,R) = \prod_{u \in L(S)} c(T,u)!. \tag{5.14}$$

This is naive, because many of the "attempted" labelings above actually represent equivalent labeled topologies. For example, if our colored topology was "((Scer,Scer),Spar)", the naive strategy above would

attempt the labelings "((Scer1,Scer2),Spar1)" and "((Scer2,Scer1),Spar1)", which are actually equivalent. For example, say we have the colored topology "((Scer,Spar),(Scer,Spar))", then these two attempted labelings would also be equivalent:

"((Scer1,Spar1),(Scer2,Spar2))" and "((Scer2,Spar2),(Scer1,Spar1))".

These are equivalent because we can swap the left and right children of the root to turn one tree into the other. In fact, any node v with children that have the same colored topology beneath them (e.g "(Scer,Spar)" is the common colored topology in the above example) can have its children swapped to create another attempted labeling. We call such nodes *mirrors*. This means the number of true labelings is reduced by half for each mirror node in the tree. Let $M(T,R)$ represent the number of mirrors in tree T, which can be calculated as follows,

$$M(T,R) = |\{v : v \in I(T), C(T,R,left_T(v)) = C(T,R,right_T(v))\}|, \tag{5.15}$$

where

$$C(T,R,v) = \begin{cases} R(v) & \text{if } v \in L(T) \\ \{C(T,R,w) : parent_T(w) = v\} & \text{if } v \in I(T). \end{cases} \tag{5.16}$$

The function $C(T,R,v)$ represents a colored topology using nested sets to represent the topology of the tree. Putting this together, we can define the number of ways of labeling a colored topology as,

$$N_1(T,R) = 2^{-M(T,R)} \prod_{u \in L(S)} c(T,u)!. \tag{5.17}$$

Steps 1 and 2: calculating the probability of a unlabeled gene tree T''.

We have shown how the term $N_1(T,R)$ can be used to compute the topology prior given the probability $P(T_2'',R|\theta)$ of an unlabeled topology T_2'' being generated by steps 1 and 2 of the DL model. Here, we define how to compute $P(T_2'',R|\theta)$.

In step 2, we use the mapping U to remove labels from the leaves and speciation nodes of tree T' to produce an unlabeled topology T''. When we remove these labels, gene trees that were once distinct suddenly become equivalent. The probability of an unlabeled gene tree T'' is thus the sum of the probability of all the labeled gene trees T_i' where $T'' = U(T')$. Since each T_i' is equally likely in the model, knowing

how many T_i' there are for a T'' is sufficient for computing the sum. We can imagine going through the gene tree T' and removing labels from each duplication subtree T_2' one at a time. When we consider a duplication subtree T_2' we have two cases (i) and (ii).

In case (i) the leaves of the subtree T_2' are extant genes that are labeled by their gene names (i.e. $L(T_2') \subset L(T')$). We need to ask, how many labeled subtrees would produce the same unlabeled subtree? It is also equivalent to ask the question in reverse, how many ways can you label an unlabeled topology? Notice, that we answered this question in the previous section for the entire tree T. When considering the question for a subtree $T_2' = sub(T', v, u)$, we have the difference that all of the leaves of T_2' are from the same species u (i.e. there is only one color). Therefore, we have a correction factor of

$$2^{-M(T_2,R)} L(T_2)!, \tag{5.18}$$

which is just a special case of $N_1(T,R)$ given in Equation 5.17.

In case (ii) the leaves of the subtree T_2' are speciations (i.e. $L(T_2') \subseteq spec(T)$) and are labeled, since the BD process generates labeled trees. The purpose of step 2 is to remove these labels, since ancestral nodes are supposed to be unlabeled. When we remove these labels, speciation nodes will become indistinguishable from one another if and only if they contain the same colored topology beneath them. Thus, the leaves of a duplication subtree T_2 are not completely interchangeable. If they were, the correction factor would be $|L(T_2)|!$, the number of ways of relabeling the subtree leaves $L(T_2)$. However, many of these labelings imply equivalent colored labeled topologies. Thus, we must again account for color mirrors. We define $M(T,T_2,R)$ to count the number of color mirrors within subtree T_2. Thus, the correction factor for each T_2 of case (ii) is

$$N_2(T,T_2,R) = 2^{-M(T,T_2,R)} |L(T_2)|! \tag{5.19}$$

$$M(T,T_2,R) = |\{v : v \in I(T_2), C(T,R,left_T(v)) = C(T,R,right_T(v))\}|. \tag{5.20}$$

5.6.4 The full topology prior

Combining the factoring as well as the equations of Chapter 5.6.3, we can compute the probability of a gene tree topology from the DL model. The full probability of a gene tree topology from the DL model is the probability of a gene tree being generated from step 1, times the correction factors $N_1(T_2,R)$ for each case (i) and $N_2(T,T_2,R)$ for each case (ii), and divided by the final correction factor $N_1(T,R)$ for step (3). Thus,

we have

$$P(T,R|S,\boldsymbol{t},\lambda,\mu) = \frac{1}{N_1(T,R)} \prod_{v \in spec(T)} \prod_{u \in child(R(v))} g(v,u,sub(T,v,u)) \tag{5.21a}$$

$$g(v,u,T_2) = f(T,T_2,R) \sum_{i=0}^{\infty} \binom{|L(T_2)|+i}{i} p(|L(T_2)|+i,t(u)) d(u)^i \tag{5.21b}$$

$$f(T,T_2,R) = \begin{cases} N_2(T,T_2,R) H(T_2)/\xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq I(S) \\ N_1(T_2,R) H(T_2)/\xi_{|L(T_2)|} & \text{if } L(T_2) \subseteq L(S) \end{cases} \tag{5.21c}$$

The sum in Equation 5.21 is a sum over how many doomed lineages i might have been present at node u. Within the sum, we find the probability that a BD process generates the survivors $L(T_2)$ that are present plus i hidden doomed lineages. The term $d(u)^i$ is the probability that those i lineages go extinct. The permutation term describes the number of ways to choose i doomed lineages from the total number of survivors $i + |L(T_2)|$.

Although this calculation involves an infinite sum, it can be computed analytically and the total computation of the topology prior takes at most $O(|V(T)||V(S)|)$ run time [6]. Currently, we only consider reconciliations R that are maximally parsimonious for duplications and losses. This approximation is likely reasonable, as we find that the true reconciliation is the most parsimonious one in 98% of gene trees simulated using our species tree (Figure 9.1) and independently estimated duplication and loss rates [64], agreeing with results from similar studies [32].

5.7 Computing the branch length prior

The final term in our model is the branch length prior $P(l|T,R,\theta)$, which is the prior probability of the branch lengths l given the topology T, reconciliation R and model parameters θ. This term helps SPIMAP choose gene trees that have branch lengths that are more reasonable given the time span implied by the reconciliation and our prior knowledge of the substitution rates.

We will explain the calculation of this term in a top-down fashion, breaking it into smaller parts until each part is defined. We begin by viewing the branch prior as a marginal over the gene rate g of the family in consideration,

$$P(l|T,R,\theta) = \int P(l|g,T,R,\theta)P(g|\alpha_G,\beta_G)dg. \tag{5.22}$$

Once conditioned on the gene rate g, many of the branch lengths of T become independent since we know their common scale factor g. However, those branches that surround a duplication node are still non-independent because their lengths depend on the time of the duplication, which is unknown. However, if we partition T into a set of subtrees $\mathbb{T}$ by segmenting at each speciation node $v \in spec(T)$ (without adding implied speciation nodes), each subtree $\tau \in \mathbb{T}$ will contain branch lengths that are independent of the other subtrees. In particular, each subtree τ is rooted by a speciation node, its leaves are either extant or are speciation nodes, and all other internal nodes are duplication nodes. We refer to branch lengths for each subtree τ as l^τ, its divergence times as t^τ, and its substitution rates as r^τ. Thus, $l^\tau = (l(w_1), l(w_2), ..., l(w_k))$ and $t^\tau = (t(w_1), ..., t(w_k))$ where $w_1, w_2, ..., w_k$ are the non-root nodes of subtree τ. Using this notation, we can continue to factor,

$$P(l|g,T,R,\theta) = \prod_{\tau \in \mathbb{T}} P(l^\tau|g,T,R,\theta). \tag{5.23}$$

The branch lengths within l^τ are non-independent because they depend on the duplication times. However, if we condition on the branch times t^τ, each branch length l_i^τ becomes a simple function of the branch rate r_i^τ, since $l_i^\tau = t_i^\tau r_i^\tau$. Since we model all branch rates as being independent of one another, we can then finally factor the branch prior as a product of the probability of each branch length l_i^τ,

$$P(l^\tau|g,T,R,\theta) = \int P(l^\tau|t^\tau,g,T,R,\theta)P(t^\tau|g,T,R,\theta)dt^\tau \tag{5.24}$$

$$\text{where } P(l^\tau|t^\tau,g,T,R,\theta) = \prod_i P(l_i^\tau|t_i^\tau,g,T,R,\theta), \tag{5.25}$$

and where $P(t^\tau|g,T,R,\theta)$ describes the distribution of branch times in subtree τ which is defined by the birth-death process. We have integrated over the branch times t^τ, since they are unknown.

The last term to define is the distribution of a single branch length $l(v_i)$. In the simplest case (see the next section for a caveat), the distribution can be derived as follows

$$l(v_i) = g \times t(v_i) \times s(v_i) \sim g \times t(v_i) \times Gamma\left(\alpha_{R(v_i)}, \beta_{R(v_i)}\right) = Gamma\left(\alpha_{R(v_i)}, \frac{\beta_{R(v_i)}}{g \times t(v_i)}\right). \quad (5.26)$$

where, $s(v_k)$ is the species-specific rate for branch $b(v_k)$. In our implementation of computing the branch prior, we integrate over gene rates g (Equation 5.22) by approximating with a summation with equally probable gene rates. Also, the integral over times t^τ (Equation 5.25) is performed with Monte Carlo by sampling from $P(t^\tau|g,T,R,\theta)$.

5.7.1 Handling implied speciation nodes

One complexity not considered in equation 5.26 is the effect of implied speciation nodes. In such a case, we can have a branch length $l(v_i)$ that spans multiple species branches. For example, the branch b_2 in Figure 5.3a spans the species B and u_1. Also note, that the length of branch b_2 is the sum of two smaller branches: one within species branch B and one within species branch u_1. Thus, to complete our description of the branch prior, we must define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$ for branches that span multiple species.

To handle these cases, we introduce a topology T' that is defined as the topology T with implied speciation nodes added. Also let l' and t' be the length and time vectors of T', and R' be a reconciliation of T' to S. For each branch $b(v_i) = (v_i, w_i)$ in T where w_i is the parent of v_i in T, there is a path $p = (v_i, ..., w_i)$ in T'. Let $p(v_i)$ be the set of all vertices in p excluding the top node w_i. Thus, the branch lengths and times in tree T can be expressed as sums of branch lengths and times in tree T',

$$l(v_i) = \sum_{v'_k \in p(v_i)} l(v'_k) \text{ and } t(v_i) = \sum_{v'_k \in p(v_i)} t(v'_k). \quad (5.27)$$

The distribution of each $l(v'_k)$ is the same as the distribution given in Equation 5.26 using R' as the reconciliation. To define the probability $P(l(v_i)|t(v_i),g,T,R,\theta)$, we note that $l(v_i)$ is simply the sum of independent gamma random variables, and methods exist to compute this probability efficiently [106].

5.7.2 Branches near the root

If a gene branch contains the root, then it is still distributed by a sum of gamma distributions and thus can use the same methods developed here. For nodes that reconcile before the species tree root, we still treat them as being generated by a BD process in the basal branch of the species tree. We model the length T_0 of the basal branch as expontentially distributed with mean λ_0 and model the species-specific subsitution rate as a gamma distributioned random variable with mean and variarance that is the average of the other species-specific rate distributions.

5.7.3 Distribution of a sum of branch lengths

I will need to be able to calculate the probability of seeing a particular branch length which is the linear combination of independent gammas. First, we know that independent Gammas with the same shape parameter β add together into another gamma and we know a gamma scaled is another gamma. But what about adding together several gammas with differing βs? I have found a citation Moschopoulos 1985[106], that gives a method for computing the PDF of the sum.

First, M. uses the following definition for the PDF of a gamma distributed variable X_i

$$f_i(x_i) = x_i^{\alpha_i - 1} \exp(-x_i/\beta_i)/[\beta_i^{a_i}\Gamma(\alpha_i)], \;\; x_i > 0$$

and $f_i(x_i) = 0$ elsewhere. Let $Y = X_1 + ... + X_n$. The PDF $g(y)$ of Y is then

$$\begin{aligned}
\beta_1 &= \min_i \beta_i \\
C &= \prod_{i=1}^{n} (\beta_1/\beta_i)_i^{\alpha} \\
\rho &= \sum_{i=1}^{n} \alpha_i \\
\gamma_k &= \sum_{i=1}^{n} \alpha_i (1 - \beta_1/\beta_i)^k/k, \;\; k = 1,2,... \\
\delta_0 &= 1 \\
\delta_{k+1} &= \frac{1}{k+1} \sum_{i=1}^{k+1} i\gamma_i \delta_{k+1-i}, \;\; k = 0,1,2,... \\
g(y) &= C \sum_{K=0}^{\infty} \delta_k y^{\rho+k-1} \exp(-y/\beta_1)/[\Gamma(\rho+k)\beta_1^{\rho+k}]
\end{aligned}$$

The CDF of the gamma sum is

$$G(w) = C \sum_{k=0}^{\infty} \delta_k \int_0^w (y^{p+k-1} e^{-y/\beta_1} / [\Gamma(p+k)\beta_1^{p+k}]) dy$$

5.7.4 Rapid tree search

To compute the argmax in Equation 5.4c, we search over the space of possible gene tree topologies T, branch lengths l, and reconciliations R using a hill climbing approach to find the maximum *a posteriori* reconciled gene tree $(\hat{T}, \hat{l}, \hat{R})$. We begin our search with an initial tree constructed using the Neighbor-Joining algorithm [127]. We use subtree pruning and regrafting (SPR) to propose additional topologies T. For each T, branch lengths l are proposed using numerical optimization (Newton-Raphson) of the likelihood term $P(\mathbf{D}|\mathbf{l}, T)$.

One unique feature of our search, is that we use the gene tree topology prior $P(T, R|\theta)$, a relatively fast computation compared to computing $P(\mathbf{D}|\mathbf{l}, T)$ by 2-3 orders of magnitude, to prescreen topology proposals for those that are likely to have high posterior probability. Given the best topology T thus far, we make $N \in [100, 1000]$ unique rearrangements T_i and compute their topology prior $k_i = P(T_i, R_i|\theta)$, where R_i is the maximum parsimonious reconciliation. As our next proposal we then choose a topology T_i from $T_1, ..., T_N$ with probability $p_i = \frac{c}{N} + \frac{(1-c)k_i}{\sum_i k_i}$, where parameter $c \in (0,1)$ defines a mixing between the weights k_i and the uniform distribution. In practice, we use $c = 0.2$.

We have found that this simple adjustment to our search strategy greatly increases the speed of finding the MAP gene tree (Chapter 9 and Table 9.3).

Chapter 6

A learning strategy for gene tree reconstruction

6.1 An Empirical Bayes strategy for gene tree reconstruction

Our model has several parameters, such as gene duplication-loss rates and substitution rates, that are shared across all the gene trees in a clade of species. Attempting to learn these parameters while reconstructing all gene trees simultaneously would be computationally prohibitive. Thus, we have devised two algorithms that can learn these model parameters from simpler but large genome-wide data prior to gene tree reconstruction. Using these parameters, especially the substitution rates, greatly increases reconstruction accuracy, more so than would be possible if gene trees were considered independently (Chapter 9). These two algorithms constitute a "training" step and represent an Empirical Bayes strategy. In this section, we present the two training algorithms for estimating (1) the duplication and loss rates $\theta_t = (\lambda, \mu)$, and (2) the substitution rate parameters $\theta_b = (\alpha_G, \beta_G, \alpha, \beta)$.

6.2 Estimating duplication and loss rate parameters

Our reconstruction method requires parameters for the gene duplication and loss rates $\theta_t = (\lambda, \mu)$. We estimate these parameters before reconstruction using an Empirical Bayes approach. To do this, we have implemented a variant of the method introduced by Hahn *et al.* [64] to perform a maximum likelihood estimation (MLE) of these parameters. However, unlike Hahn *et al.* we do not require λ and μ to be equal. Figure 5.1b illustrates how this estimation fits within the larger phylogenomic pipeline. In this section, we briefly review the duplication and loss rate estimation method.

The method takes as input a species tree S and set of gene family clusters. For each cluster, we only need the gene copy number count for each species. Thus, a gene tree is not needed to perform this estimation. The method assumes that gene counts vary along the branches of the species tree according to a birth-death (BD) process with constant birth rate λ and death rate μ. By assuming each gene family started with one gene copy in the common ancestral species, we can use the model to inferred maximum likelihood estimates of λ and μ.

To perform this inference, we build upon results from the BD process. First, consider a single gene family and a single branch in the species tree. We can compute the probability of a gene copies at the top of species branch of length t becoming b gene copies at its end by using the following formula [80, 8],

$$P(B=b|A=a,t,\lambda,\mu) = \sum_{j=0}^{min(a,b)} \binom{a}{j} \binom{a+b-j-1}{a-1} \alpha^{a-j}\beta^{b-j}(1-\alpha-\beta)^j \tag{6.1a}$$

$$\alpha = \mu\frac{e^{(\lambda-\mu)t}-1}{\lambda e^{(\lambda-\mu)t}-\mu} \tag{6.1b}$$

$$\beta = \lambda\frac{e^{(\lambda-\mu)t}-1}{\lambda e^{(\lambda-\mu)t}-\mu}, \tag{6.1c}$$

where $\lambda \neq \mu$. For the case where $\lambda = \mu$, we have

$$P(B=b|A=a,t,\lambda,\mu) = \sum_{j=0}^{min(a,b)} \binom{a}{j} \binom{a+b-j-1}{a-1} \alpha^{a+b-2j}(1-2\alpha)^j \tag{6.2a}$$

$$\alpha = \frac{\lambda t}{1+\lambda t}. \tag{6.2b}$$

Given these equations, we can now express the probability of observing gene counts as they vary across the entire species S. Assume for the moment that we knew the gene counts C present at every node of the species tree. Let C represent a mapping from vertices $v \in V(S)$ to integers $[0,\infty)$, where $C(v)$ is the number of gene copies that belong to a particular species v. Assuming a BD model for each branch, and given the number of genes starting at the root of species tree $C(root(S))$, we can compute the probability of seeing a family with gene counts C as

$$P(C|C(root(S)),\lambda,\mu) = \prod_{v\in V(S),v\neq root(S)} P(B=C(v)|A=C(parent(v)),t(v),\lambda,\mu). \tag{6.3}$$

In practice, we will only have the gene counts from the extant species $L(S)$. Let us call this mapping $C^L : L(S) \rightarrow [0,\infty)$. Therefore, the probability $P(C^L|C(root(S)),\lambda,\mu)$ is a marginal of $P(C|C(root(S)),\lambda,\mu)$

over all possible gene counts at the internal nodes $I(S)$. Let $\mathbb{L}(v,x,\lambda,\mu)$ be the likelihood of the tree at node v and below, given $C(v) = x$. We can then define the likelihood function recursively as

$$P(C^L|C(root(S)) = x,\lambda,\mu) = \mathbb{L}(root(S),x,\lambda,\mu) \tag{6.4}$$

$$\mathbb{L}(v,x,\lambda,\mu) = \begin{cases} \prod_{c\in child(v)} \sum_{i=0}^{\infty} P(B = i|A = x,\lambda,\mu)\mathbb{L}(c,i,\lambda,\mu) & \text{if } v \in I(S) \\ 1 & \text{if } v \in L(S), C(v) = x \\ 0 & \text{if } v \in L(S), C(v) \neq x. \end{cases} \tag{6.5}$$

For most cases, a decent approximation can be achieved by only computing the first 10 to 20 terms of the infinite sum above.

Lastly, let us consider the case where we have N gene families. Let $F = \{C^L 1, C_2^L, ..., C_N^L\}$ represent the extant gene counts for all the gene families, where C_i^L is the extant gene count for family i. Since each of these families evolve independently given λ and μ, we can compute the total probability of observing these gene families as the product of the individual family probabilities, namely

$$\prod_{C_i^L \in F} P(C_i^L|C_i(root(S)),\lambda,\mu). \tag{6.6}$$

To compute Maximum Likelihood Estimates (MLEs) of our duplication and loss rate parameters λ and μ, we use a gradient descent method on Equation 6.6. We assume $C(root(S))$ is 1 for each gene family. Gene counts F are determined by our gene clustering method which we developed in previously [10].

6.2.1 Estimating duplication and loss rates

We have used this estimation procedure extensively in our evaluation of the SPIMAP method (Chapter 9). We have implemented a simulation program based on our DL model, which we have used to create gene trees for the 12 *Drosophila* and 16 fungi clades (Figure 9.1). The gene trees were simulated with duplication and loss that were 1X, 2X, and 4X the rate of these events in real species. In Table 6.1, we show a comparison of true and estimated duplication and loss rates for data sets of simulated 12 flies and 16 fungi gene families.

species	dup/loss rate-setting	actual rates dup	actual rates loss	estimated rates dup	estimated rates loss
flies	1X,1X	0.0012	0.0012	0.001460	0.001170
flies	1X,4X	0.0012	0.0048	0.001203	0.004607
flies	2X,2X	0.0024	0.0024	0.002467	0.002268
flies	4X,1X	0.0048	0.0012	0.004852	0.001290
flies	4X,4X	0.0048	0.0048	0.004928	0.004608
fungi	1X,1X	0.000732	0.000859	0.000726	0.000805
fungi	2X,2X	0.001464	0.001718	0.001546	0.001705
fungi	4X,4X	0.002928	0.003426	0.002873	0.003102
fungi	4X,1X	0.002928	0.000859	0.002940	0.000879
fungi	1X,4X	0.000732	0.003426	0.000783	0.003237

Table 6.1: **Recovery of duplication and loss rates estimated for simulated datasets.** For each simulated dataset we choose duplication and loss rates at 1X, 2X, and 4X the rate estimated from real datasets. Above are the *actual* rates (events/gene/million years) specified in our simulation program and the rates *estimated* from gene counts of the resulting 500 simulated gene trees using the duplication loss estimation procedure. Estimated rates closely follow true rates for each dataset.

6.3 Estimating substitution rate parameters

The second training procedure that we have developed estimates the parameters of our substitution rate model from genome-wide data. As discussed previously (Chapter 5.4), our substitution rate model is able to describe rate variation that occurs in both gene- and species-specific ways. In order to achieve this it requires the estimation several parameters $\theta_b = (\alpha_G, \beta_G, \alpha, \beta)$. One unique approach in our method is that we estimate these parameters prior to reconstruction by analyzing substitution rates from multiple loci with known phylogenetic trees. Figure 5.1b illustrates how this estimation fits within the larger phylogenomic pipeline.

Currently for our training dataset, we use trees of one-to-one orthologous gene alignments (e.g. syntenic orthologs or unambiguous best reciprocal BLAST hits) where we can be reasonably confident that the gene tree topology is congruent to the species tree. Fixing the gene tree topology, we estimate the ML branch lengths for N trees with $M = |E(S)|$ branches each, in order to construct a matrix $\boldsymbol{L}$ of branch lengths, such that l_{ij} is the length of the j^{th} branch in the i^{th} tree. We then use the $\boldsymbol{L}$ matrix along with a species tree S and its branch lengths $\boldsymbol{t}$ to estimate the parameters θ_b. Since the gene rates $\boldsymbol{g}$ of these trees are not known, we treat them as hidden data and use an Expectation Maximization (EM) algorithm to estimate our parameters.

6.3.1 Variables of the model

The variables of the substitution rate training model are as follows. A gene tree will have a *gene rate* g, a vector of *species rates* $\boldsymbol{s}$ (measured in substitutions/site/unit time), and a vector of *branch lengths* $\boldsymbol{l}$ (measured in substitutions/site). Thus, for a single gene tree, we have the following variables

$$g,\ \boldsymbol{l} = [l_1,...,l_M]^T,\ \boldsymbol{s} = [s_1,...,s_M]^T,\ \boldsymbol{t} = [t_1,...,t_M]^T,\ \text{with } l_i = g s_i t_i. \tag{6.7}$$

For a set of N gene trees indexed by j, we can describe them using the variables

$$\boldsymbol{g} = [g_1,...,g_N]^T,\ \boldsymbol{L} = [\boldsymbol{l}_1,...,\boldsymbol{l}_N],\ \boldsymbol{S} = [\boldsymbol{s}_1,...,\boldsymbol{s}_N],\ \text{with } l_{ij} = g_j s_{ij} t_i. \tag{6.8}$$

We have designed this method to assume that $\boldsymbol{L}$ is directly observed and is given as input along with the divergence times $\boldsymbol{t}$. In contrast, the gene rates $\boldsymbol{g}$ and species rates $\boldsymbol{S}$ are not directly observed and have to be inferred from the model.

As for the distribution of these variables, recall that g_j are i.i.d. by the inverse gamma $InvGamma(\alpha_G, \beta_G)$ and that s_{ij} are independently distributed by $Gamma(\alpha_i, \beta_i)$. Thus, the distribution of the branch length matrix $\boldsymbol{L}$ is

$$P(\boldsymbol{L}|\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_G,\beta_G) = \prod_j P(\boldsymbol{l}_j|\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\beta},\alpha_G,\beta_G) \tag{6.9a}$$

$$= \prod_j \int_0^\infty P(g_j|\alpha_G,\beta_G) P(\boldsymbol{l}_j|g_j,\boldsymbol{t},\boldsymbol{\alpha},\boldsymbol{\beta}) dg_j \tag{6.9b}$$

$$= \prod_j \int_0^\infty InvGamma(g_j|\alpha_G,\beta_G) \prod_i Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) dg_j. \tag{6.9c}$$

6.3.2 EM method for estimation model parameters

In our EM algorithm, the branch length matrix $\boldsymbol{L}$ is the observed data and the gene rate vector $\boldsymbol{g}$ is the hidden data. The EM method guarantees that if we use the following iterative method, that we will converge on a locally maximum likelihood estimate of our parameters.

$$\theta_b{}^{h+1} = \underset{\theta_b}{\text{argmax}} \int P(\boldsymbol{g}|\boldsymbol{L},\theta_b{}^h) \log P(\boldsymbol{L},\boldsymbol{g}|\theta_b) d\boldsymbol{g} \tag{6.10a}$$

$$= \underset{\theta_b}{\text{argmax}} \int ... \int P(\boldsymbol{g}|\boldsymbol{L},\theta_b{}^h) \log P(\boldsymbol{L},\boldsymbol{g}|\theta_b) dg_1 ... dg_N. \tag{6.10b}$$

We will show in the rest of this section how to compute this expression efficiently.

First, we take advantage of several independence assumptions from our model. Note, that the variables g_j and l_j for a tree j are defined to be independent from variables $g_{j'}$ and $l_{j'}$ from any other tree j'. Therefore, we can factor the expression as

$$\theta_b{}^{h+1} = \underset{\theta_b}{\text{argmax}} \int \ldots \int \left[\prod_k P(g_k|l_k, \theta_b{}^h) \right] \left[\log \prod_j P(l_j, g_j|\theta_b) \right] dg_1 \ldots dg_N. \tag{6.11}$$

Rearranging the product and logarithm gives us

$$\theta_b{}^{h+1} = \underset{\theta_b}{\text{argmax}} \int \ldots \int \left[\prod_k P(g_k|l_k, \theta_b{}^h) \right] \left[\sum_j \log P(l_j, g_j|\theta_b) \right] dg_1 \ldots dg_N \tag{6.12a}$$

$$= \underset{\theta_b}{\text{argmax}} \int \ldots \int \left[\sum_j \left[\prod_k P(g_k|l_k, \theta_b{}^h) \right] \log P(l_j, g_j|\theta_b) \right] dg_1 \ldots dg_N. \tag{6.12b}$$

Now, if we pull out the term $P(g_j|l_j, \theta_b{}^h)$ from the product, we can move the integrals over $g_1, \ldots, g_N$ within the product

$$\theta_b{}^{k+1} = \underset{\theta_b}{\text{argmax}} \sum_j \int \ldots \int \left[\prod_{k \neq j} P(g_k|l_k, \theta_b{}^h) \right] P(g_j|l_j, \theta_b{}^h) \log P(l_j, g_j|\theta_b) dg_1 \ldots dg_N \tag{6.13a}$$

$$= \underset{\theta_b}{\text{argmax}} \sum_j \left[\prod_{k \neq j} \int P(g_k|l_k, \theta_b{}^h) dg_k \right] \int P(g_j|l_j, \theta_b{}^h) \log P(l_j, g_j|\theta_b) dg_j. \tag{6.13b}$$

Lastly, we take advantage of the fact that

$$\int P(g_k|l_k, \theta_b{}^h) dg_k = 1, \tag{6.14}$$

which gives us the simplified expression

$$\theta_b{}^{h+1} = \underset{\theta_b}{\text{argmax}} \sum_j \int P(g_j|l_j, \theta_b{}^h) \log P(l_j, g_j|\theta_b) dg_j. \tag{6.15}$$

In terms of the overall, EM algorithm, computing the term $P(g_j|l_j, \theta_b{}^h)$ (i.e. the probability of hidden data) constitutes the E-step, which we will outline in the next section. However, before we continue can

simplify this expression further. For example, the term $logP(\boldsymbol{l}_j, g_j|\theta_b)$ can be written as

$$\log P(\boldsymbol{l}_j, g_j|\theta_b) = \log P(\boldsymbol{l}_j|g_j, \theta_b) + \log P(g_j|\alpha_G, \beta_G) \tag{6.16a}$$

$$= \sum_i \log Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) + \log InvGamma(g_j|\alpha_G, \beta_G). \tag{6.16b}$$

When we plug this term back in, we can move the integrals inward and move the summations outward, giving us

$$\theta_b{}^{h+1} = \underset{\theta_b}{\operatorname{argmax}} \sum_j \int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \left[\sum_i \log Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) + \log InvGamma(g_j|\alpha_G, \beta_G)\right] dg_j \tag{6.17a}$$

$$= \underset{\theta_b}{\operatorname{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \sum_i \log Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) dg_j\right] + \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log InvGamma(g_j|\alpha_G, \beta_G) dg_j\right] \tag{6.17b}$$

$$= \underset{\theta_b}{\operatorname{argmax}} \sum_i \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) dg_j\right] + \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log InvGamma(g_j|\alpha_G, \beta_G) dg_j\right]. \tag{6.17c}$$

Since each parameter appears within its own term within this summation, we can compute the argmax by optimizing each term separately. Therefore, we can write the EM optimization as $N+1$ argmax equations

$$\alpha_G^{h+1}, \beta_G^{h+1} = \underset{\alpha_G, \beta_G}{\operatorname{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log InvGamma(g_j|\alpha_G, \beta_G) dg_j\right] \tag{6.18}$$

$$\alpha_i^{h+1}, \beta_i^{h+1} = \underset{\alpha_i, \beta_i}{\operatorname{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) dg_j\right]. \tag{6.19}$$

One last restriction we add to the model is to require that the gene rate distribution has a mean of one. Without this restriction the model is over-parameterized, since a gene rate distribution with a higher mean could be offset by species-specific rates with lower means. When this restriction is added, the gene rate argmax can be rewritten as follows

$$\beta_G^{h+1} = \underset{\beta_G}{\operatorname{argmax}} \sum_j \left[\int P(g_j|\boldsymbol{l}_j, \theta_b{}^h) \log InvGamma(g_j|\beta_G) dg_j\right]. \tag{6.20}$$

M-step: the gradient

To maximize the equations above, we use the BFGS method which requires the gradient with respect to θ_b. In this section, we give the derivatives of Equations 6.19 and 6.20 with respect to each parameter.

The integrals in Equations 6.19 and 6.20 are approximated by discretizing the gene rate g into K classes and computing the following lookup tables. The table igtab[j,k] represents the gene rate for the j^{th} gene family and the k^{th} gene rate class, where the table pigtab[j,k] gives the probability of that gene rate, namely $P(g_j|l_j,{\theta_b}^h)$. Populating these tables constitutes the E-step (see next section).

In the following equations, we will use the function f to represent the right-hand side of the argmax, namely

$$ {\theta_b}^{h+1} = \underset{\theta_b}{\operatorname{argmax}} f({\theta_b}^h, \theta_b). \tag{6.21} $$

Thus, for equation 6.20, the derivative is

$$ \frac{\partial}{\partial \beta_G} f = \sum_j \left[\int P(g_j|l_j,{\theta_b}^h) \frac{InvGamma'_\beta(g_j|\beta_G)}{InvGamma(g_j|\beta_G)} dg_j \right] \tag{6.22a} $$

$$ \approx \sum_j \sum_k \text{pigtab[j,k]} \frac{InvGamma'_\beta(\text{igtab}[j,k]|\beta_G)}{InvGamma(\text{igtab}[j,k]|\beta_G)}. \tag{6.22b} $$

Since this is one dimensional, we will use root finding on its derivative to optimize it. Such a method may need the second derivative, which is

$$ \frac{\partial^2}{\partial^2 \beta_G} f = \frac{\partial}{\partial \beta_G} \sum_j \left[\int P(g_j|l_j,{\theta_b}^h) \frac{Gamma'_\beta(g_j|\beta_G)}{Gamma(g_j|\beta_G)} dg_j \right] \tag{6.23a} $$

$$ = \sum_j \left[\int P(g_j|l_j,{\theta_b}^h) \frac{Gamma(g_j|\beta_G) Gamma''_\beta(g_j|\beta_G) - Gamma'_\beta(g_j|\beta_G)^2}{Gamma(g_j|\beta_G)^2} dg_j \right] \tag{6.23b} $$

The derivative with respect to the lineage rate parameters α_i and β_i are

$$\frac{\partial}{\partial \alpha_i} f = \sum_j \left[\int P(g_j | l_j, \theta_b{}^h) \frac{Gamma'_\alpha\left(l_{ij} | \alpha_i, \frac{\beta_i}{g_j t_i}\right)}{Gamma\left(l_{ij} | \alpha_i, \frac{\beta_i}{g_j t_i}\right)} dg_j \right] \tag{6.24a}$$

$$\approx \sum_j \sum_k \text{pigtab}[j,k] \frac{Gamma'_\alpha\left(l_{ij} | \alpha_i, \frac{\beta_i}{\text{igtab}[j,k] t_i}\right)}{Gamma\left(l_{ij} | \alpha_i, \frac{\beta_i}{\text{igtab}[j,k] t_i}\right)} \tag{6.24b}$$

$$\frac{\partial}{\partial \beta_i} f = \sum_j \left[\int P(g_j | l_j, \theta_b{}^h) \frac{Gamma'_\beta\left(l_{ij} | \alpha_i, \frac{\beta_i}{g_j t_i}\right)}{Gamma\left(l_{ij} | \alpha_i, \frac{\beta_i}{g_j t_i}\right) g_j t_j} dg_j \right] \tag{6.24c}$$

$$\approx \sum_j \sum_k \text{pigtab}[j,k] \frac{Gamma'_\beta\left(l_{ij} | \alpha_i, \frac{\beta_i}{\text{igtab}[j,k] t_i}\right)}{Gamma\left(l_{ij} | \alpha_i, \frac{\beta_i}{\text{igtab}[j,k] t_i}\right) \text{igtab}[j,k] t_i}. \tag{6.24d}$$

Derivatives of the gamma and inverse gamma distributions.

For our gradient, we need several derivatives of the gamma and inverse gamma probability distribution functions, which we give here

$$\frac{\partial}{\partial \alpha} Gamma(x|\alpha,\beta) = Gamma'_\alpha(x|\alpha,\beta) \tag{6.25}$$

$$= e^{-\beta x} \Gamma(\alpha)^{-1} \beta^\alpha x^{\alpha-1} \left[\log(x) + \log(\beta) - \psi^{(0)}(\alpha) \right] \tag{6.26}$$

$$\frac{\partial}{\partial \beta} Gamma(x|\alpha,\beta) = Gamma'_\beta(x|\alpha,\beta) \tag{6.27}$$

$$= x^{\alpha-1} \Gamma(\alpha)^{-1} (\alpha \beta^{\alpha-1} e^{-\beta x} - x \beta^\alpha e^{-\beta x}) \tag{6.28}$$

$$\frac{\partial}{\partial x} Gamma(x|\alpha,\beta) = Gamma'_x(x|\alpha,\beta) \tag{6.29}$$

$$= \beta^\alpha \Gamma(\alpha)^{-1} \left[(\alpha-1) x^{\alpha-2} e^{-\beta x} - \beta x^{\alpha-1} e^{-\beta x} \right], \tag{6.30}$$

where $\Gamma(x)$ is the gamma function and $\psi^{(0)}(x) = \Gamma'(x)/\Gamma(x)$ is the PolyGamma function.

For the inverse gamma distribution, where $\alpha = \beta + 1$, we have the following first and second derivatives

$$\begin{aligned}
\frac{\partial}{\partial\beta} InvGamma(x|\beta) =& (1/\Gamma(1+\beta)\beta^{\beta}\exp(-\beta/x)(1/x)^{(\beta+3)}) \\
& \left[\beta(x-1)+x+\beta x(\log(\beta)+\log(1/x)) - \beta x\psi^{(0)}(\beta+1)\right] \\
\frac{\partial^2}{\partial^2\beta} InvGamma(x|\beta) =& (1/\Gamma(1+\beta))\beta^{\beta}\exp(-\beta/x)(1/x)^{(\beta+4)} \\
& [\beta - 2(1+\beta)x + (3+\beta)x^2 + \\
& x(\log(\beta)+\log(1/x))(2(\beta(x-1)+x)+\beta x(\log(\beta)+\log(1/x)))+ \\
& -2x(\beta(x-1)+x+\beta x(\log(\beta)+\log(1/x)))\psi^{(0)}(1+\beta)+ \\
& \beta x^2\psi^{(0)}(1+\beta)^2 - \beta x^2\psi^{(1)}(1+\beta)\Big]
\end{aligned}$$

E-step

The goal of the E-step is to compute the probability of hidden data g_j given the previous iteration's parameters ${\theta_b}^h$, which is

$$P(g_j|\boldsymbol{l}_j, {\theta_b}^h) = \frac{P(\boldsymbol{l}_j|g_j, {\theta_b}^h)P(g_j|{\theta_b}^h)}{\int P(\boldsymbol{l}_j|g_j, {\theta_b}^h)P(g_j|{\theta_b}^h)dg_j}. \tag{6.31}$$

Since $P(g_j|{\theta_b}^h)$ is the inverse gamma distribution and $P(\boldsymbol{l}_j|g_j, {\theta_b}^h)$ is the gamma distribution, we can use conjugate priors to rewrite the probability of the hidden data as

$$P(g_j|\boldsymbol{l}_j, {\theta_b}^h) \propto InvGamma(g_j|\alpha_G, \beta_G)\prod_i Gamma\left(l_{ij}|\alpha_i, \frac{\beta_i}{g_j t_i}\right) \tag{6.32a}$$

$$= InvGamma(g_j|a', b'). \tag{6.32b}$$

We can derive the parameters (a', b') of the posterior distribution as follows. First we look at the

probability of the data $\boldsymbol{l}_j$.

$$\begin{aligned}
P(\boldsymbol{l}_j|g_j,t_i,\theta_b) &= \prod_i Gamma(l_{ij}|\alpha_i,\frac{\beta_i}{t_i g_j}) \\
&= \prod_i l_{ij}^{\alpha_i-1}\left(\frac{\beta_i}{t_i g_j}\right)^{\alpha_i}\exp\left(-\frac{\beta_i}{t_i g_j}l_{ij}\right)\Gamma(\alpha_i)^{-1} \\
&= \left[\prod_i l_{ij}^{\alpha_i-1}\Gamma(\alpha_i)^{-1}\left(\frac{\beta_i}{t_i}\right)^{\alpha_i}\right]\left[\exp\left(-\sum_i\frac{\beta_i}{t_i g_j}l_{ij}\right)\prod_i\left(\frac{1}{g_j}\right)^{\alpha_i}\right] \\
&\propto \exp\left(-\frac{1}{g_j}\sum_i\frac{\beta_i l_{ij}}{t_i}\right)\left(\frac{1}{g_j}\right)^{A} \\
&= \exp\left(-\frac{S}{g_j}\right)\left(\frac{1}{g_j}\right)^{A},
\end{aligned}$$

where

$$S=\sum_i\frac{\beta_i l_{ij}}{t_i} \qquad\qquad A=\sum_i\alpha_i.$$

Notice the likelihood can be factored such that we have a single term that depends on the parameter of interest g_j and only on sufficient statistics S and n. Since the conjugate prior is proportional to this term, the conjugate prior is the inverse gamma distribution

$$InvGamma(x|a,b)=\frac{b^a}{\Gamma(a)}(1/x)^{a+1}\exp(-b/x).$$

Also, we know the posterior distribution is proportional to the product of the isolated term above and the prior.

$$\begin{aligned}
P(g_j|\boldsymbol{l}_j,\boldsymbol{t},\theta_b) &\propto \frac{b^a}{\Gamma(a)}(1/g_j)^{a+1}\exp(-b/g_j)\exp\left(-\frac{S}{g_j}\right)\left(\frac{1}{g_j}\right)^{A} \\
&\propto \left(\frac{1}{g_j}\right)^{A+a+1}\exp\left(-\frac{S+b}{g_j}\right) \\
&\propto InvGamma(g_j|a',b'),
\end{aligned}$$

where

$$a'=A+a \qquad\qquad b'=S+b.$$

Rewriting these variables in the our notation gives

$$a' = \alpha_G + \sum_i \alpha_i \qquad\qquad b' = \beta_G + \sum_i \frac{\beta_i l_{ij}}{t_i}$$

Thus, we have

$$P(g_j|\boldsymbol{l}_j, \boldsymbol{t}, \theta) = InvGamma\left(g_j|\alpha_G + \sum_i \alpha_i, \beta_G + \sum_i \frac{\beta_i l_{ij}}{t_i}\right). \tag{6.33}$$

I have formulated my EM algorithm to use a discretized posterior distribution. First, the probability distribution function (PDF) is partitioned into W parts indexed by k. For a tree j, the gene rate g associated with the kth partition will be stored in the table entry `igtab[j,k]` and its probability will be stored in `pigtab[j,k]`. To cover the range of the distribution, we define divisions around the mode of the distribution, which for inverse gamma is $b'/(a'+1)$.

6.3.3 Extension: multiple gene rates

One possible extension to this model that may be interesting to explore in the future, is to model multiple gene rates that may exist in distinct parts of the species. This can be done with only a few modifications. The main idea is to break up the species tree into K subtrees indexed by m, each with their own gene rate g_{jm}. The tree partitioning can be described by a mapping $\mathcal{G}(i) = m$, which gives the gene rate partition m for each branch $i \in E(S)$. Let us also define the reverse mapping

$$\mathcal{G}'(m) = \{i|\mathcal{G}(m) = i\}. \tag{6.34}$$

When we incorporate this extension, the variables for each gene tree becomes

$$\boldsymbol{g} = [g_1, ..., g_K]^T,\ \boldsymbol{l} = [l_1, ..., l_M]^T,\ \boldsymbol{s} = [s_1, ..., s_M]^T,\ \boldsymbol{t} = [t_1, ..., t_M]^T,\ \text{with } l_i = g_{\mathcal{G}(i)} s_i t_i. \tag{6.35}$$

For a set of N gene trees indexed by j, we can describe them using the variables

$$\boldsymbol{G} = [\boldsymbol{g}_1, ..., \boldsymbol{g}_N]^T,\ \boldsymbol{L} = [\boldsymbol{l}_1, ..., \boldsymbol{l}_N],\ \boldsymbol{S} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_N],\ \text{with } l_{ij} = g_{j\mathcal{G}(i)} s_{ij} t_i, \tag{6.36}$$

with parameters

$$\boldsymbol{\alpha}_G = [\alpha_{G1}, ..., \alpha_{GK}]; \qquad \boldsymbol{\beta}_G = [\beta_{G1}, ..., \alpha_{GK}]; \qquad \boldsymbol{\alpha} = [\alpha_1, ..., \alpha_N]; \qquad \boldsymbol{\beta} = [\beta_1, ..., \beta_N]. \tag{6.37}$$

For notational convenience, let us define the super-script m to denote the subset of branches in within gene class m,

$$\boldsymbol{l}_j^m = [l_{ij} | \mathcal{G}(i) = m]. \tag{6.38}$$

The distribution for $\boldsymbol{G}$ becomes

$$g_{jm} \sim InvGamma(\alpha_{Gm}, \beta_{Gm}), \tag{6.39}$$

where each g_{jm} is sampled independently.

Since the gene rate classes are independent, we can factor our usual likelihood function along these classes, which gives us

$$P(\boldsymbol{L}|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}_G, \boldsymbol{\beta}_G) = \prod_j P(\boldsymbol{l}_j | \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}_G, \boldsymbol{\beta}_G) \tag{6.40a}$$

$$= \prod_j \prod_{m=1}^{K} \int_0^\infty P(\boldsymbol{l}_j^m | \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_{Gm}, \beta_{Gm}). \tag{6.40b}$$

We can now treat the terms inside of the product as marginals over the gene rates

$$P(\boldsymbol{l}_j^m | \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}_G, \beta_G) = \int_0^\infty P(\boldsymbol{l}_j^m, g_{jm} | \boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_{Gm}, \beta_{Gm}) dg_{jm}. \tag{6.41}$$

Next, we can factor out the gene rate probability,

$$P(\boldsymbol{l}_j^m, g_{jm} | \boldsymbol{t}^m, \boldsymbol{\alpha}, \boldsymbol{\beta}, \alpha_{Gm}, \beta_{Gm}) = P(\boldsymbol{l}_j^m | g_{jm}, \boldsymbol{t}^m, \boldsymbol{\alpha}, \boldsymbol{\beta}) P(g_{jm} | \alpha_{Gm}, \beta_{Gm}) \tag{6.42a}$$

$$= InvGamma(g_{jm} | \alpha_{Gm}, \beta_{Gm}) \prod_{i \in \mathcal{G}'(m)} P(l_{ij} | g_{jm}, t_i, \alpha_i, \beta_i). \tag{6.42b}$$

In conclusion, our likelihood function is

$$P(\boldsymbol{L}|\boldsymbol{t}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}_G, \beta_G) = \prod_j \prod_m \int_0^\infty InvGamma(g_{jm} | \alpha_{Gm}, \beta_{Gm}) \prod_{i \in \mathcal{G}'(m)} Gamma\left(l_{ij} | \alpha_i, \frac{\beta_i}{g_{jm} t_i}\right) dg_{jm}. \tag{6.43}$$

EM method for multiple gene rates

An EM method for the model with multiple gene rates can be easily constructed from multiple instances of the EM method for a single gene rate. Notice that in likelihood function of Equation 6.43, we could pull the product over gene classes $m \in 1,...,K$ out as the outer most product. Once this is done all of the model parameters split into separate groups, one for each gene rate class. Thus optimizing the likelihood for the parameters in each gene class separately, optimizes the over all likelihood.

6.3.4 Fitting the model

Using gene trees of one-to-one syntenic gene alignment from the 12 *Drosophila* and 16 fungi clades (Figure 9.1), we estimated parameters for our substitution rates model. In Figure 6.1, we compare the estimated distributions of the substitution rates (black lines) to the empirical rates (red lines) seen in the 16 fungi clade. In Figure 6.2, we show the comparison between the true rate distributions (black lines) of simulated 12 fly trees with the estimated rates (red lines).

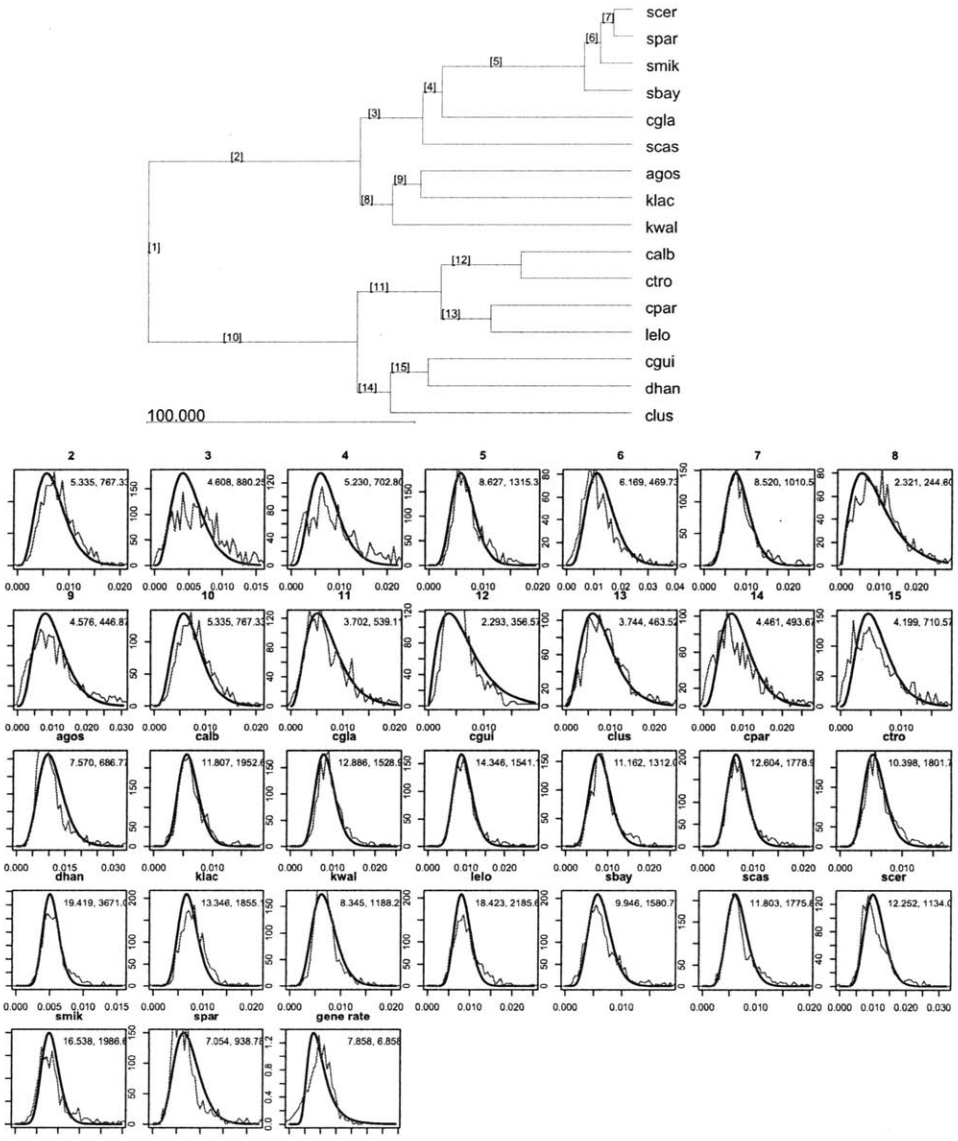


Figure 6.1: Agreement between observed substitution rate distributions and fitted model for each branch of the 16 fungi species tree. Each branch i has a gamma distributed rate with parameters α_i, β_i. Distributions estimated from real data are drawn in black and their parameters α_i, β_i are given in the top row within each plot. Red lines show distribution of relative branch lengths from real gene trees. Gene trees with fewer than 30 substitutions were filtered out of training. Branches are named as drawn in tree above.

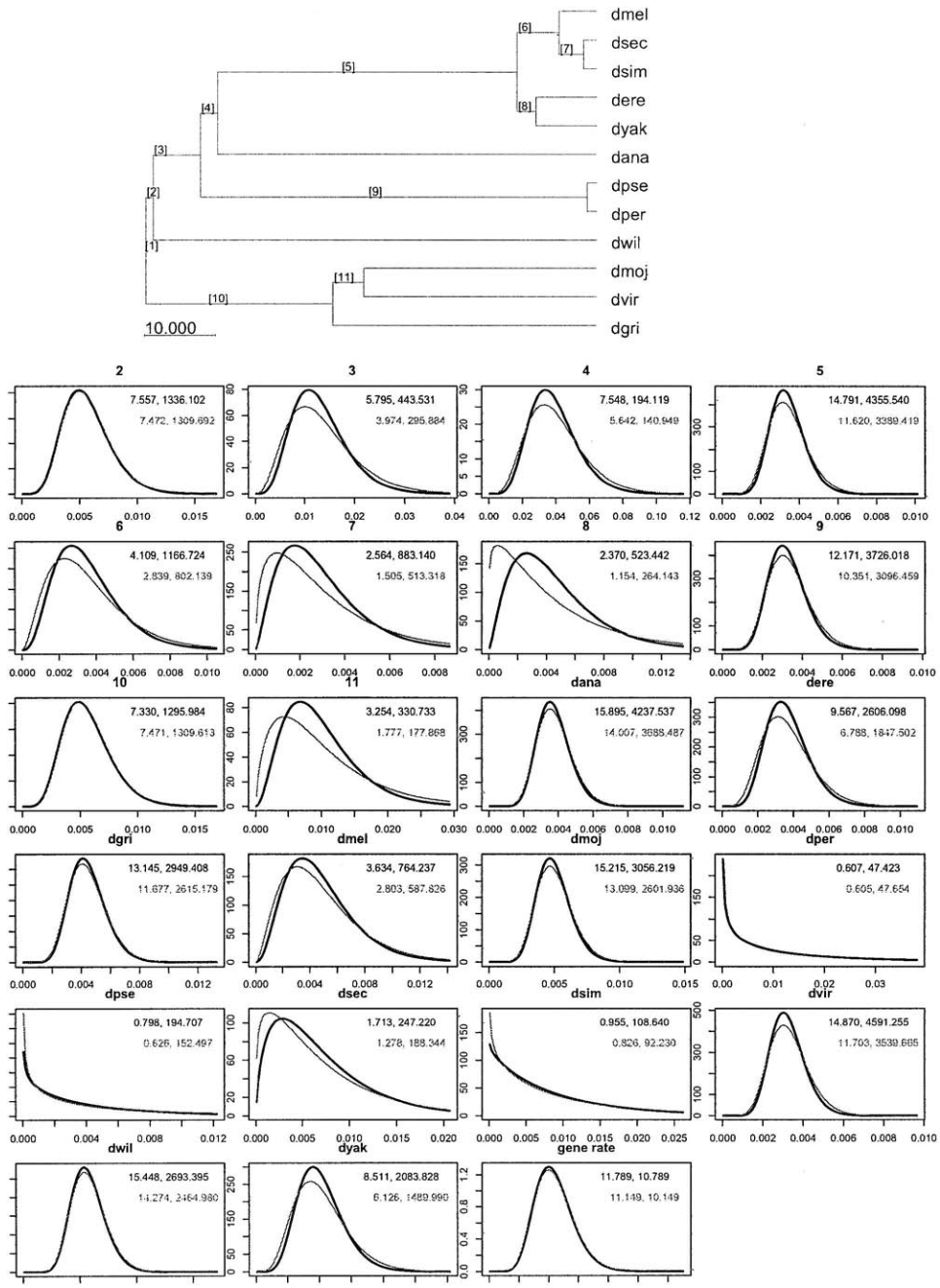


Figure 6.2: Agreement between true and estimated substitution rate parameters in simulated data for each branch of the 12 *Drosophila* species tree. Each branch i has a gamma distributed rate with parameters α_i, β_i. Distributions estimated from real data are drawn in black and their parameters α_i, β_i are given in the top row within each plot. Red lines and bottom row parameters represent distributions estimated from simulated data which were generated using the same parameters estimated from real data. Branches are named as drawn in tree above.

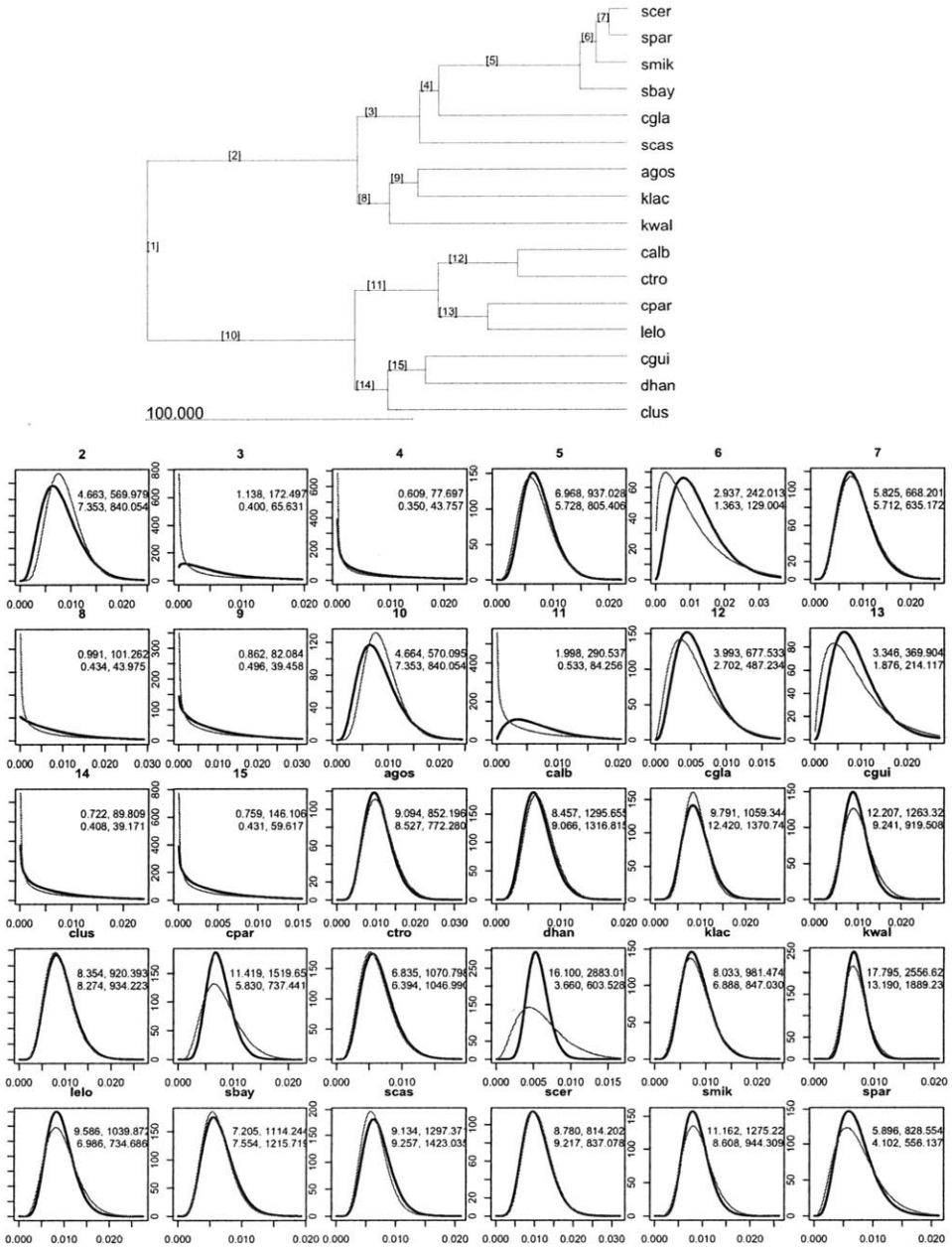


Figure 6.3: Agreement between true and estimated substitution rate parameters in simulated data for each branch of the the 16 fungi species tree. Each branch i has a gamma distributed rate with parameters α_i, β_i. Distributions estimated from real data are drawn in black and their parameters α_i, β_i are given in the top row within each plot. Red lines and bottom row parameters represent distributions estimated from simulated data which were generated using the same parameters estimated from real data. Branches are named as drawn in tree above.

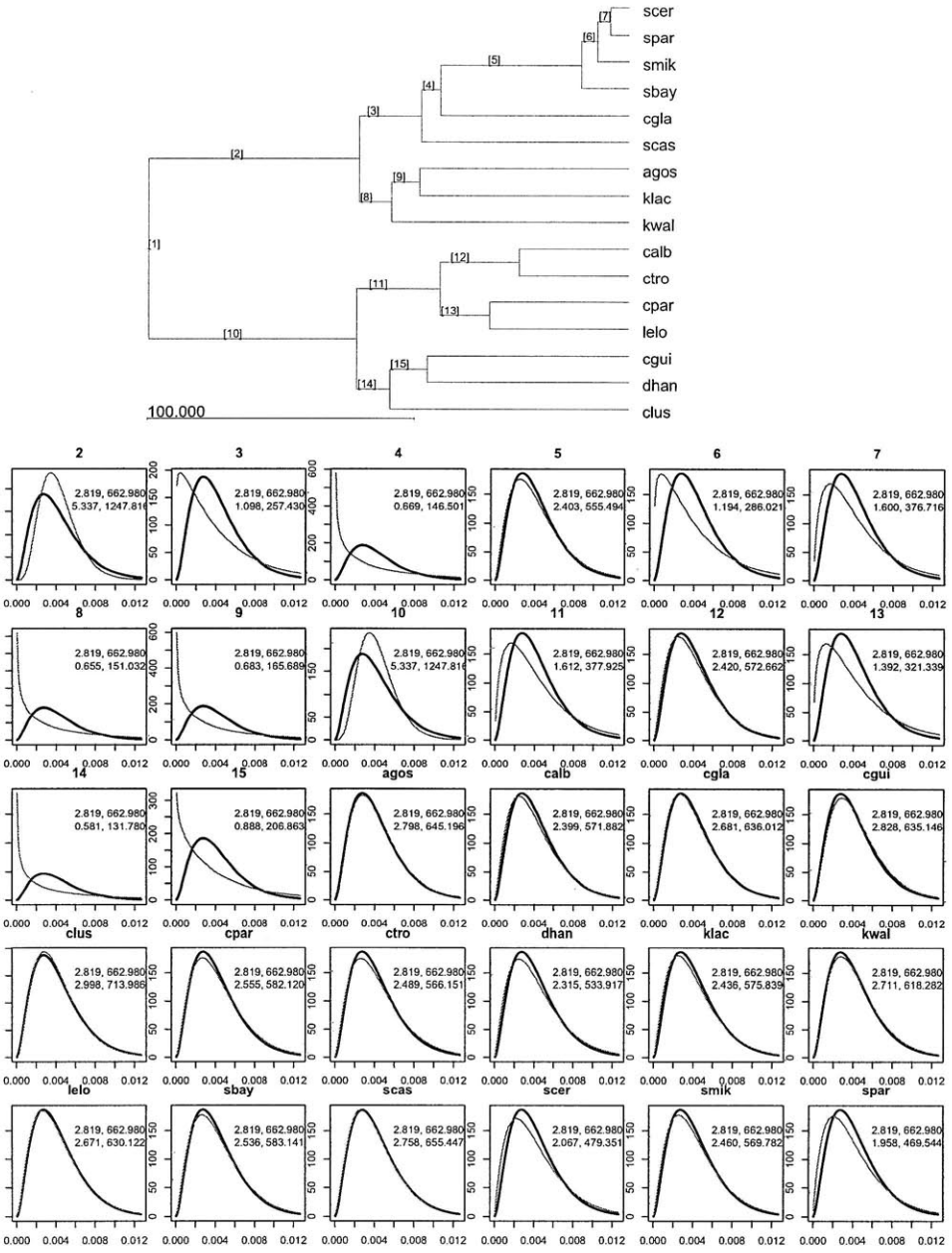


Figure 6.4: Substitution rate parameters used in the search speed evaluation. A 16 fungi datatset was simulated assuming i.i.d. species-specific gamma distributed rates (black curves). From a simulated dataset, SPIMAP then estimated its own rate parameters (red curves) for reconstructing gene trees. Most parameter disagreement happen on short branches where substitution rate estimation suffers most from errors due to small counts. Branches are named as drawn in tree above.

Chapter 7

Drosophila case study

7.1 New model of sequence substitution rates

Our sequence substitution model was motivated by rate distributions observed in the *Drosophila* and fungal genomes [121]. This model provided a key advantage in producing a more informative prior and thus better reconstruction accuracy.

To understand how branch lengths could be modeled, we revisited our 5154 one-to-one syntenic ortholog *Drosophila* gene alignments [121], only this time we built maximum likelihood gene trees with PHYML [62] while requiring a fixed topology congruent with the *Drosophila* species tree [136, 66]. Although each of the gene trees are the same in topology, they vary greatly in branch lengths (Figure 7.1; top row). However, when we normalized the gene trees by their total branch length to produce relative branch lengths (Figure 7.1; bottom row), we found the branch lengths to be much similar, although with some variation remaining.

From this, we concluded that we could model gene families as evolving according to a *gene-specific*

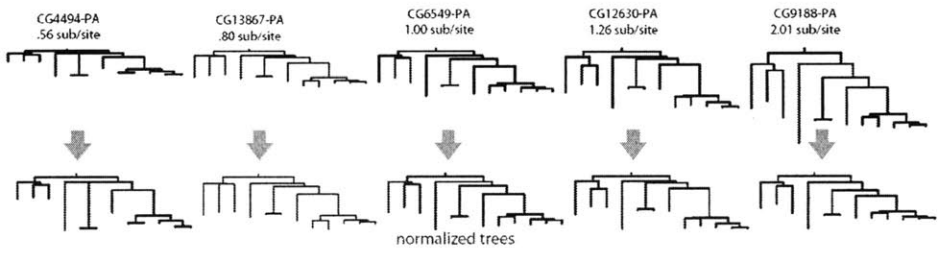


Figure 7.1: **Examples of various one-to-one orthologous gene trees.** Trees are displayed with absolute (top) and relative (bottom) branch lengths. Although gene trees vary greatly in total tree length, relative branch lengths are fairly consistent.

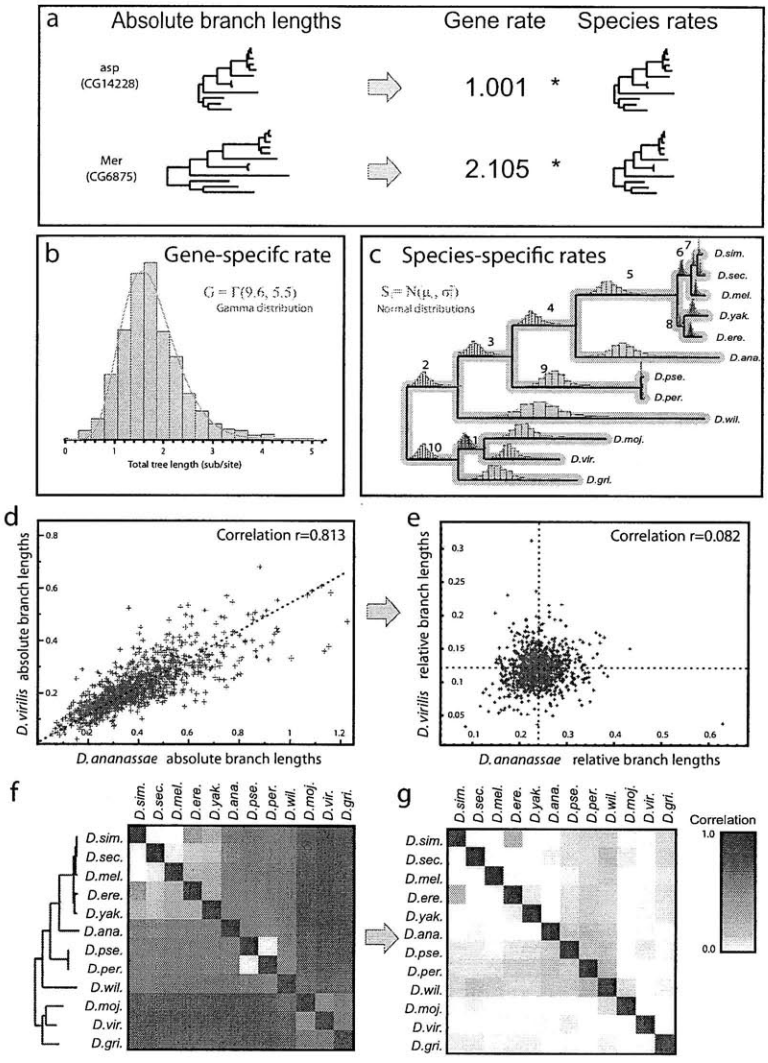


Figure 7.2: **Evolutionary rates decoupled into gene-specific and species-specific components.** (a) Syntenic ortholog trees appear as scaled versions of a common species tree, and can be expressed as the product of a gene-specific rate, and species-specific rates. (b) Gene-specific rates of 5,154 fly orthologs follow a gamma distribution. (c) Species-specific rates for each lineage follow normal distributions. (d) Unnormalized (absolute) branch lengths are highly correlated. Lengths for *D. virilis* and *D. ananassae* since their last common ancestor across the 5154 orthologs show correlation r=0.813. (e) Relative branch lengths become independent after normalization by the gene-specific rate (r=0.082). (f) Correlations are high for all species pairs before normalization, except for very closely related species. (g) Relative lengths are uncorrelated for all species pairs.

rate, that is a rate present throughout all branches in a gene tree that effectively scales the branch lengths together. Thus, a gene evolving fast in one species, is likely to be fast evolving in all species. In Figure 7.2a, we show the gene rate (expressed as the total tree length) of two such gene trees. In Figure 7.2b, we show the distribution of the gene rate across all 5154 gene families. We find that the distribution can be well approximated by either a gamma distribution (shown in Figure 7.2b) or the inverse gamma distribution (see Figure 6.3).

Modeling gene-specific rates has been done previously in several contexts [152, 48, 133, 81]. The 12 *Drosophila* provided a good dataset for testing this approximation for a large number of gene families.

One by-product of the scaling effect of gene rates can be observed in correlations between different species lineages. In Figure 7.2d, we see the absolute lengths of the *D. grimshawi* and *D. ananassae* branches show strong correlation ($r = 0.813$) across all 5154 gene trees. This high correlation is also seen for all pairs of branches in the fly tree (Figure 7.2f), where we find an average correlation of $r = 0.61$.

However, when we produce the same plots for the relative branch lengths, we see a significantly different picture. For the same two species, we find that the correlation of relative branch lengths between *D. grimshawi* and *D. ananassae* is dramatically lower ($r = 0.082$). In addition, this reduction in correlation is present throughout all pairs of species (Figure 7.2; $r = 0.09$). This indicates that relative branch lengths can be well approximated as being independent of each other.

When we plot the distribution of the relative branch lengths, we find that we can approximate them with either normal distributions (Figure 7.2c) or gamma distributions (Figure 6.3). Each of the distributions have a distinct mean and variance specific to each species. Therefore, we think of these rates as being *species-specific*. The mean of each normal distribution incorporates the time duration of that species branch as well as a genome-wide rate acceleration that is in effect for that species. In our later work, we have separated these two effects, such that we can isolate true rate changes from time span differences (Chapter 6.3).

We have found these same effects for other species as well. In Figure 7.3, we show the pair-wise correlations for absolute branches lengths (mean $r = 0.37$) and relative branch lengths (mean $r = 0.01$) of 739 gene trees from a clade of 16 fungi. In Figure 6.3, we show how the branch length distributions fit the gamma distribution.

One likely limitation of this model is that we may only be able to model a gene-specific rate for species that are some what closely related. For example, between the *Saccharomyces* and *Candida* clades there is less branch correlation than within the clades. Perhaps over long periods of evolutionary time, the function or architecture of a gene changes significantly enough that it alters the percentage of sites that are free to substitute without negative effect. Such a scenario might indicate that each major clade could be modeled

with its own independent gene-specific rate. In Chapter 6.3.3, I have presented a model and estimation procedure that could be appropriate for such data.

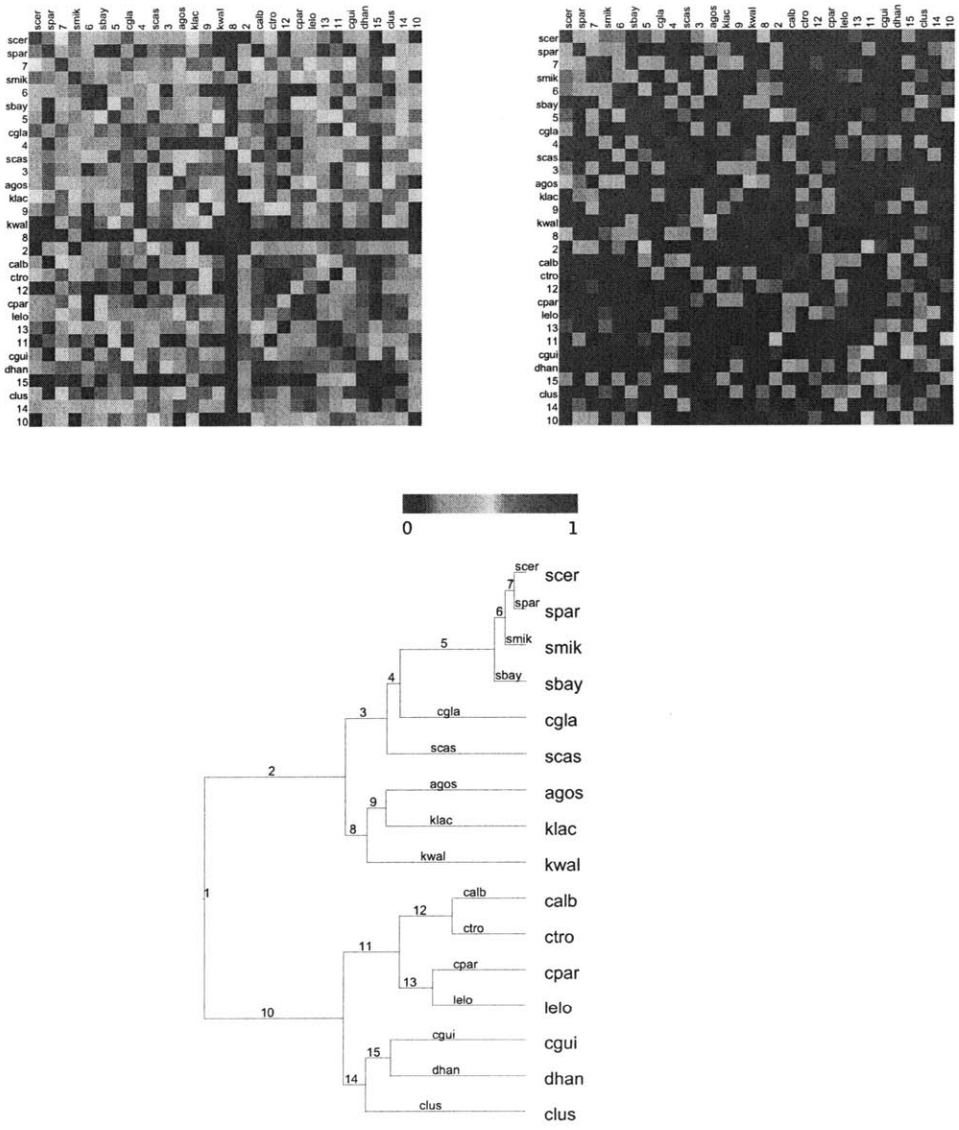


Figure 7.3: **Branch length correlations for 739 fungi gene families.** Pairwise correlations of branch lengths in the 16 fungi clade are correlated in absolute branch lengths (left; mean $r = 0.37$) and fairly uncorrelated for relative branch lengths (right; mean $r = 0.01$). The naming scheme for branches in the 16 fungi clade is given the phylogenetic tree (bottom).

Chapter 8

Candida gene family analysis

8.1 Candida genome sequencing

Invasive candidiasis is the leading cause of death resulting from fungal infections in the United States, and over 95% of these infections are caused by four particular species: *C. albicans*, *C. glabrata*, *C. tropicalis*, and *C. parapsilosis* [115]. Among these species *C. albicans* is the most prevalent in the human population, where it acts as an opportunistic pathogen, benignly existing in 80% of people, but occasionally becoming pathogenic, infecting oral cavities, genitals, and the blood stream. *C. albicans* can be life-threatening for immunocompromised patients and is an increasingly frequent cause of infections during hospital care.

The whole-genome sequence of *C. albicans* (strain SC534) was first published in 2004 [74]. Because of its pathogenicity, it has been the focus of many anti-fungal agent studies [92, 100, 115], where the use of gene expression profiles has been an especially useful tool.

The *Candida* species have also been studied for their unique use of a tRNA that performs a non-standard translation of the CUG codon to a serine residue instead of the usual leucine [101]. This alternative translation is common for all the *Candida* species and is hypothesized to have evolved at the base of the clade (Figure 8.1). In *C. albicans*, CUG codons are translated both with the standard and alternate residues in an apparently stochastic fashion, thus a single gene can express a great number of proteins [56]. It has been hypothesized stochastic translation may contribute to the species ability to adapt to new host environments or evade host detection.

In an effort to better understand this species, the Broad Institute along with a consortium of collaborators at the Sanger Center and the Candida Genome Database (CGD) have sequenced five related fungal species (*C. tropicalis*, *C. parapsilosis*, *L. elongisporus*, *C. guillermondii*, and *C. lusitaniae*) as well as a second strain of *C. albicans*, strain WO-1 [10]. These species, along with two additional species (*D. hansenii* and

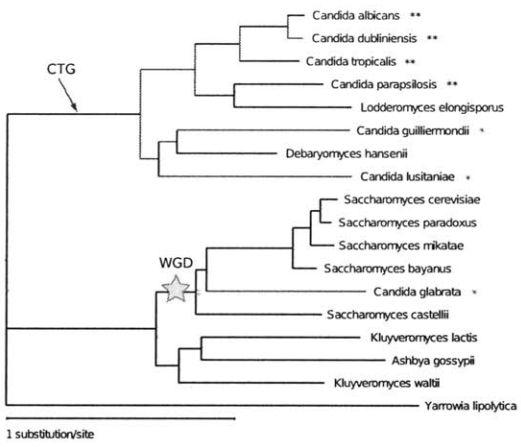


Figure 8.1: Species phylogeny of *Candida* and *Saccharomyces* clades. The estimated positions of the alternative CTG translation origin (arrow) and the *Saccharomyces* whole-genome duplication (star) are indicated. Species that are frequently found in human fungal infections are denoted as "strong pathogens" (**) and those that are only rarely seen as infectious are denoted "weak pathogens" (*). In our duplication enrichment analysis, lineages marked in red were assumed to be pathogenic, and lineages marked in dark red were assumed to be "strongly" pathogenic.

C. dublinensis) sequenced by other projects, represent a clade spanning approximately 100 million years of evolution.

As part of the genome sequencing project, we were interested in studying the evolving of these species. One question we focused on in particular was how the varying levels of pathogenicity may have evolved amongst these species. Figure 8.1 illustrates the distribution of these species on a phylogenetic tree and how they are related to the *Saccharomyces*. We have denoted four of the species as "strong pathogens", indicating that they are frequently found to be the cause of human infections. Another three species, which we denote "weak pathogens", are also pathogenic, but are rarely seen in human infections. Although most of the pathogenic species are near each other in the species tree, they are not monophyletic, that is they are interspersed with non-pathogenic species. Given these genome sequences, can we detect any changes in their gene content as the level of pathogenicity evolved across this clade?

8.1.1 Phylogenomics approach to Candida evolution

We sought to study the evolution of pathogenicity by using a phylogenomic approach. In particular, we asked whether the historical transition from non-pathogenic to pathogenic status had left any patterns in their gene

evolution. Moreover, could we identify possible pathogenic related gene families by finding those that show significantly different evolution within pathogenic lineages compared to the non-pathogenic?

Combining all annotations and computational predictions, the 16 species together contain about 111,745 genes. Since *C. dublinensis* was not part of the consortium sequencing, it was not included in this analysis. Using a sequence clustering procedure, we were able to assign 89,924 genes to 9209 families [10]. Families range from as small as 2 genes to the largest containing 126 glucose transporters which are present in all 16 species (Figure 8.2). There are 2865 families with 16 or more genes, 401 (83.8%) of which are *persistent* (black bars), that is they contain at least one gene from each species. 1521 families are specific to the *Candida* clade, and 3765 are specific to *Saccharomyces*.

For each gene family cluster, we aligned protein sequences with MUSCLE ([40]), and mapped coding sequences onto the peptide alignments (replacing each residue with a codon and every gap with a triplet of gaps). We then used a preliminary version of our SPIMAP algorithm to reconstruct gene trees for each of the 9209 families. By reconciling each gene tree to the species tree, we were able to infer all gene duplications and losses (Figure 8.3). We also inferred the appearance of a new gene families. These families exist within only a subset of the species and contain no easily identified orthologs in the other species. This may occur due to a significantly new gene architectures created by gene fusion or fission, or due to a high rate of sequence substitution.

8.2 Gene-specific rates in 16 fungi

Using our gene evolution model, we estimated the gene-specific rate for each family. Similar to our previous work with *Drosophila* gene families, we found that the distribution of fungal gene rates can be approximated by a gamma distribution (Figure 8.4). Of families with at least 10 genes, the fastest 10% of gene rates are significantly enriched (hypergeometric test) for several Gene Ontology (GO) terms related to adaption, including regulation ($P < 8.8 \times 10^{-15}$), transcription ($P < 8.7 \times 10^{-8}$), and pseudo-hyphal growth ($P < 5.8 \times 10^{-8}$). In contrast, the gene families with slowest 10% of gene rates are enriched for conserved core processes such as ribosome related processes ($P < 2.4 \times 10^{-46}$), and translation elongation ($P < 6.4 \times 10^{-8}$).

8.3 Pathogen-associated gene duplication

Within the *Candida* clade of species, only *L. elongisporus* and *D. hansenii* are rarely found in infections, where as *C. albicans*, *C. tropicalis*, and *C. parapsilosis* are the most aggressive pathogens. Among these

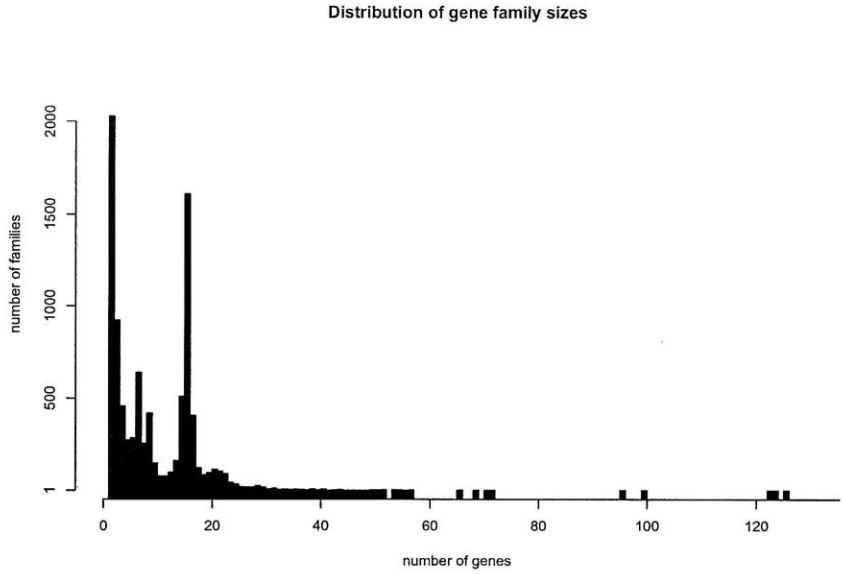


Figure 8.2: Distribution of family sizes across 16 species of fungi. Of families with 16 or more genes, 83.8% are persistent (at least one gene present in each species; black bars).

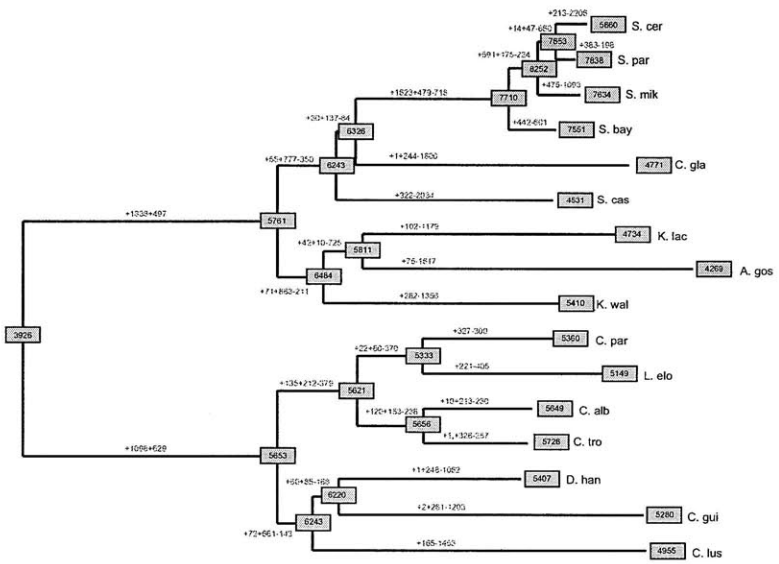


Figure 8.3: Inferred counts of all gene counts (boxes), gene duplications (green), gene losses (red), and gene appearances (blue) in fungal species phylogeny. Each branch is labeled with the following: "+gene appearances + duplications - losses".

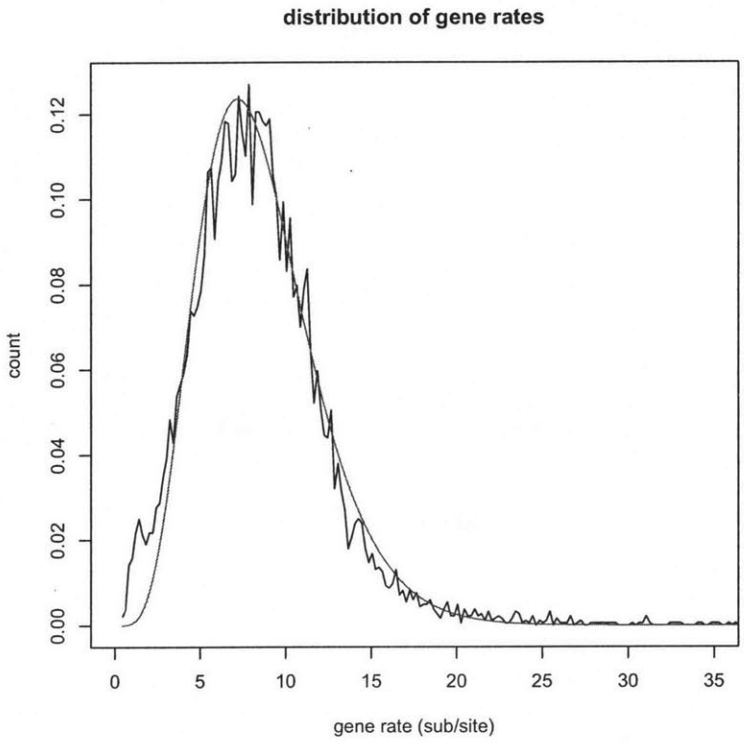


Figure 8.4: Distribution of gene-specific rates in 9209 gene families approximately follows a gamma distribution ($\alpha = 6.14, \beta = 0.71$).

family	annotation	genes	PD	D	L	P value	FDR	gene rate
1. CF7734	Leucine-rich repeat (IFA/FGR38-like)	34	32	32	5	1.1e-14	5.2e-11	18.3
2. CF8711	GPI family 18 (Hyr/Iff-like)	66	46	52	11	1.2e-14	5.2e-11	16.2
3. CF10581	Reductase family	54	30	31	30	1.6e-12	5e-9	2.3
4. CF10326	Ferric reductase family	55	28	30	25	1.1e-10	2.6e-7	2.4
5. CF1318	GPI family 17 (ALS-like adhesins)	36	27	29	4	2.8e-10	5.2e-7	20.5
6. CF10555	Oligopeptide transporters	44	21	23	11	7.5e-08	1.2e-4	6.7
7. CF9063	GPI family 13 (Pga30-like)	41	22	25	5	1.6e-07	2.0e-4	14.8
8. CF10133	Cell wall mannoprotein biosynthesis	56	18	19	34	1.8e-07	2.0e-4	2.1
9. CF1039	Short chain dehydrogenases	41	14	14	11	8.0e-07	7.3e-4	7.9
10. CF11051	Major facilitator transporters	32	14	14	17	7.0e-07	7.3e-4	2.0
11. CF739	Unclassified	20	13	13	3	2.2e-06	1.8e-3	15.8
12. CF10190	Amino acid permeases	38	11	11	18	1.6e-05	0.012	1.7
13. CF5137	Sphingomyelin phosphodiesterases	23	11	11	9	1.6e-05	0.012	7.4
14. CF2015	Short chain dehydrogenases	40	13	14	5	2.0e-05	0.012	8.6
15. CF11000	RNA binding proteins	26	13	14	43	2.0e-05	0.012	3.0
16. CF10270	GPI family 6 (Pga59/62-like)	29	14	16	32	4.1e-05	0.023	1.7
17. CF896	Phosphoglycerate mutase	41	10	10	8	4.4e-05	0.023	7.7
18. CF10316	MFS/sugar transporter	15	10	10	6	4.4e-05	0.023	1.6
19. CF1699	Unclassified	37	16	20	0	9.6e-05	0.046	10.9
20. CF4945	Unclassified	34	13	15	6	9.9e-05	0.046	11.1

Figure 8.5: Families with duplications enriched in pathogenic species. Duplications (D), losses (L), and pathogenic duplications (PD) were identified by reconciling each gene tree to the species tree.

family	annotation	genes	SPD	D	L	P value	FDR	gene rate
1. CF7734	Leucine-rich repeat (IFA/FGR38-like)	34	32	32	5	1.8e-17	3.0e-14	18.3
2. CF1318	GPI-family 17 (ALS-like Adhesins)	36	26	29	4	3.5e-11	2.9e-8	20.5
3. CF9063	GPI-family 13 (Pga30-like)	41	19	25	5	3.0e-6	1.7e-3	14.8
4. CF739	Unclassified	20	12	13	3	5.4e-6	2.2e-3	15.8
5. CF10333	Formate dehydrogenase	27	12	14	4	2.7e-5	9.0e-3	4.8
6. CF8711	GPI-family 18 (Hyr/Iff-like)	66	30	52	11	3.3e-5	9.2e-3	16.2
7. CF10555	Oligopeptide transporter	44	15	21	11	1.2e-4	0.02	6.7
8. CF7425	FGR6 family (filamentous growth)	13	7	7	1	2.3e-4	0.04	14.5

Figure 8.6: Families with duplications enriched in strong pathogenic species. Duplications (D), losses (L), and strong pathogenic duplications (SPD) were identified by reconciling each gene tree to the species tree.

species, we expect to find common properties of pathogenesis to be inherited from a common ancestor. Using our reconstructions of all duplications and losses, we asked if any families are significantly enriched with genes or duplications from the pathogenic species and lineages (see Figure 8.1 for lineage definitions). Using the hypergeometric test and requiring a false discovery rate of 5%, we found 20 families significantly enriched for duplications within pathogenic lineages (Figure 8.5). In addition, duplications were also significantly enriched in the "strong" pathogenic species for at least 8 gene families (Figure 8.6).

These families were associated with several pathogenic relevant functions such as the cell wall, transport, secretion, and filamentous growth. Cell wall and secretion proteins play important roles in adhesion to the host as well as avoiding detection by the host. Filamentous growth is a stage of growth that *C. albicans* often

adopts when it is invading host tissue during infection. Lastly, for several of the families we identified, very little is known about them (CF739, CF1699, CF4945). This analysis suggests that these families may be related to pathogenesis and may be good candidates for future functional analysis related to pathogenicity.

8.4 Pathogen-associated positive selection

In addition to studying duplication and loss within our families, we also searched for positive selection, specifically in association with pathogenicity. Using a branch-site model [149], as implemented in PAML, we looked for positive selection that is specific to only the strong pathogenic lineages (foreground branches: terminal branches leading to *C. albicans*, *C. parapsilosis*, *C. tropicalis*, and the branch immediately ancestral to *C. albicans* and *C. tropicalis*). We used the gene tree topologies found by SPIMAP and codon aligned nucleotide alignments for this analysis.

To test for significant positive selection, a Likelihood Ratio Test (LRT) was performed between model A and its null for each family (for details on PAML's models see [149]). Out of all 9209 families, 4927 families contained both strong pathogenic and as well as other lineages, and 64 showed significant positive selection (FDR < 0.001) within the strong pathogenic lineages (Figure 8.7). These families have faster substitution rates on average, showing a gene-specific rate of 8.13 substitutions/site compared to the overall average of 5.8.

Using the Candida-specific GO SLIM terms [20], we found 18 GO terms enriched within our positively selected set (Figure 8.8). Many of these terms are similar to the functions found in our pathogenic gene duplication analysis, although a different set of families were found. For example, 12 families are related to either hyphal, pseudohyphal, fileamentous growth, or biofilm formation and 6 families were previously associated with pathogenesis. One such example is the ERG3 family (CF3105), a C-5 sterol desaturase that is essential for yeast ergosterol biosynthesis[105]. ERG3 has been found to be up-regulated in azole-resistant strains of *C. albicans* [76].

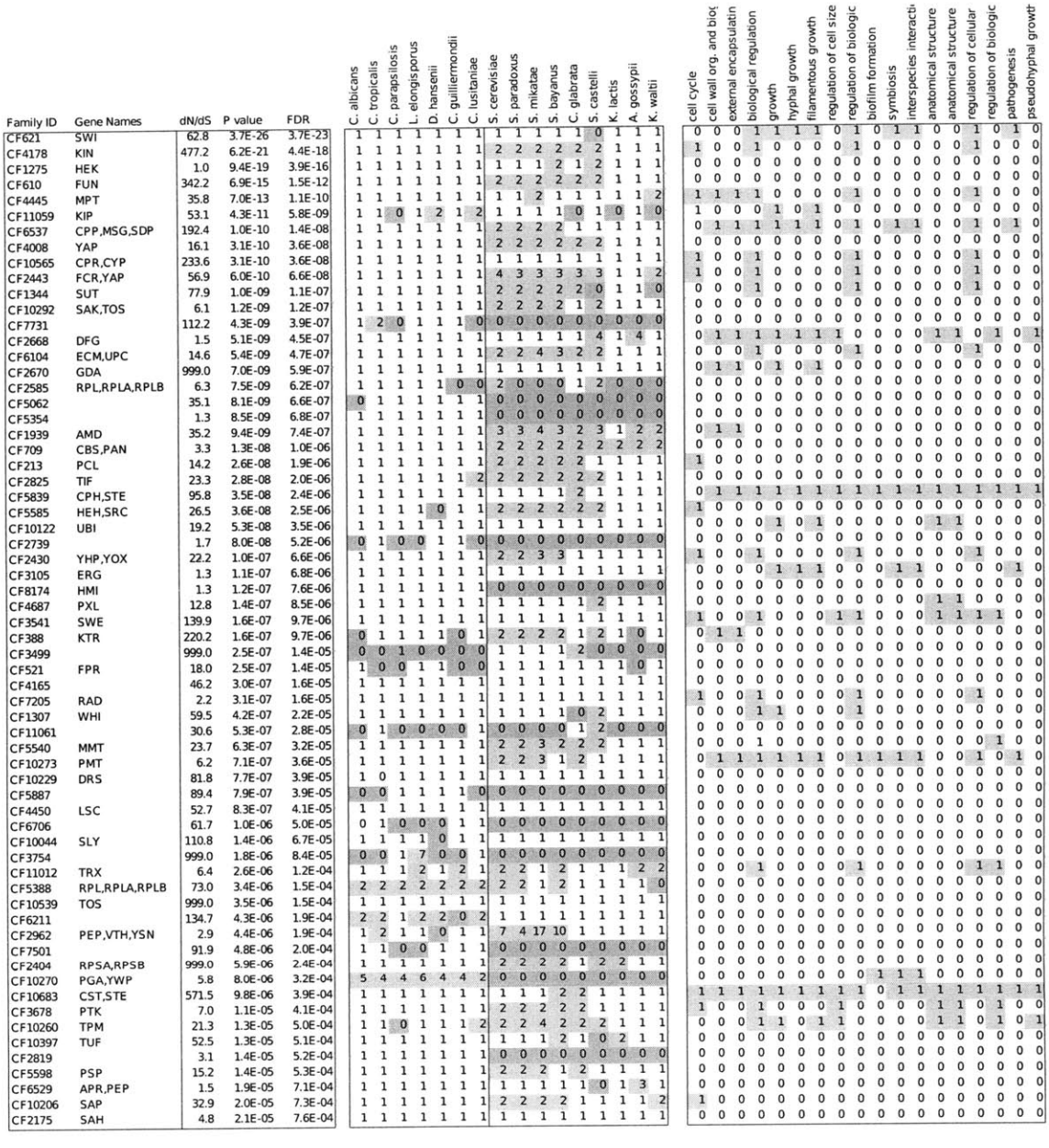


Figure 8.7: Gene trees with significant positive selection specifically within pathogenetic lineages. P values were computed using a likelihood ratio test. Gene counts for each species and family are indicated. Several GO terms associated with these families were enriched for positive selected set.

GO term	positively selected families	other families	P value	FDR	fold
cell cycle	13	311	2.1×10^{-4}	1.3×10^{-2}	3.0
cell wall organization and biogenesis	9	154	2.2×10^{-4}	1.3×10^{-2}	4.3
external encapsulating structure org. and biogen.	9	154	2.2×10^{-4}	1.3×10^{-2}	4.3
biological regulation	20	683	3.8×10^{-4}	1.7×10^{-2}	2.2
growth	12	319	9.5×10^{-4}	3.4×10^{-2}	2.8
hyphal growth	7	119	1.1×10^{-3}	3.4×10^{-2}	4.3
filamentous growth	11	288	1.4×10^{-3}	3.6×10^{-2}	2.8
regulation of cell size	6	94	1.7×10^{-3}	3.9×10^{-2}	4.6
regulation of biological process	16	555	2.0×10^{-3}	4.1×10^{-2}	2.2
biofilm formation	3	20	3.1×10^{-3}	4.4×10^{-2}	10.0
symbiosis	7	144	3.2×10^{-3}	4.4×10^{-2}	3.6
interspecies interaction between organisms	7	145	3.3×10^{-3}	4.4×10^{-2}	3.5
anatomical structure morphogenesis	8	191	3.9×10^{-3}	4.4×10^{-2}	3.1
anatomical structure development	8	191	3.9×10^{-3}	4.4×10^{-2}	3.1
regulation of cellular process	15	539	4.0×10^{-3}	4.4×10^{-2}	2.1
regulation of biological quality	8	194	4.3×10^{-3}	4.4×10^{-2}	3.0
pathogenesis	6	114	4.4×10^{-3}	4.4×10^{-2}	3.8
pseudohyphal growth	4	48	4.4×10^{-3}	4.4×10^{-2}	5.9

Figure 8.8: Gene Ontology (GO) terms enriched in the 64 families positively selected in the pathogenic lineages. P values were calculated using the hypergeometric test. Many GO terms describe functions or features related to pathogenicity (e.g. hyphal, pseudohyphal, fileamentous growth, biofilm formation, etc.).

Chapter 9

Extensive benchmarks for phylogenomics

9.1 Phylogenomic datasets

To evaluate our approach for gene tree reconstruction, we have reconstructed gene trees for both real and simulated datasets. For our real dataset, we have used 16 fungi species (Figure 9.1a) whose genomes have been sequenced to either draft or high coverage quality [55, 79, 18, 78, 74, 35, 31, 10]. For our simulated datasets, we simulated gene alignments that share many properties of real gene trees, by using a model with parameters estimated from real datasets. Thus, we have simulated gene trees that capture the properties of the 16 fungal genomes as well as 12 fully sequenced *Drosophila* genomes [1, 123, 17] (Figure 9.1b). By using both clades, we can evaluate the performance of phylogenetic methods across a variety of species tree topologies, divergence times, and gene duplication and loss rates.

For the species trees, we obtained the topologies and divergence times from several data sources. For the 16 fungi, we used the species phylogeny as constructed in Butler *et al.* [10] and estimated time divergence using the `r8s` program [128] with an estimate of 180 million years [101] for the clade depth (Figure 9.1a). For the 12 flies, we used the same topology and divergence times as used in several recent studies [136, 66] (Figure 9.1b).

9.2 Training SPIMAP's model parameters

To run SPIMAP in our evaluations, we applied our training algorithms to estimate the parameters of our gene family model. These parameters were also used to generate the simulated datasets. Here, we describe how we prepared the input data for our training procedure for both the 16 fungi and 12 *Drosophila* datasets. Our training procedure contains two methods: one to estimate our substitution rate parameters $\theta_b = (\beta_G, \alpha, \beta)$

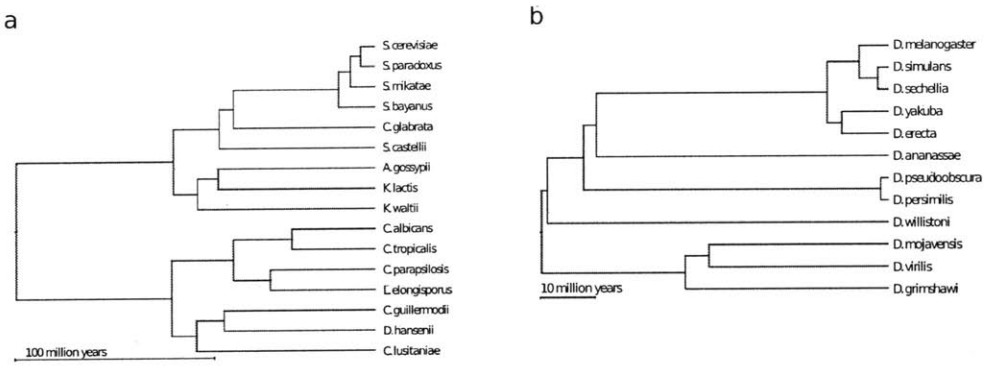


Figure 9.1: **Species and phylogenies used in evaluation.** (a) Phylogeny of 16 fungal species used for the reconstruction pipelines of real and simulated evaluation datasets. The phylogeny was estimated in [10] with divergence times estimated by the `r8s` program [128] assuming 180 million years [101] for the divergence depth. (b) The phylogeny of 12 *Drosophila* species used in our simulation evaluation. Phylogeny was estimated by Tamura *et al* [136].

and one to estimate our duplication and loss rates $\theta_t = (\lambda, \mu)$.

The first method (*Estimating substitution rate parameters*) estimates our substitution rate parameters from a dataset of one-to-one orthologous gene trees that are congruent to the species tree. To obtain such trees, we identified families that are highly likely to be one-to-one orthologous (i.e. one gene from each species in the clade). For the 16 fungi, we previously identified 739 confident one-to-one orthologous families [10]. This was done by identifying synteny blocks containing at least 3 consecutive genes and spanning across the *Saccharomyces* or *Candidae* clades. Pairs of syntenic clusters with best reciprocal BLAST hits spanning across the clades were merged, resulting in 739 families. For the 12 flies clade, we previously identified 5154 one-to-one families where genes belong to a synteny block spanning all 12 species and contains at least three consecutive genes along each chromosome [121]. Next, for each one-to-one family, we made peptide multiple alignments using MUSCLE [40]. Coding sequences were mapped onto the alignments to produce codon-aligned nucleotide alignments, substituting every amino acid with the corresponding codon and every gap with a triplet of gaps. PHYML v2.4.4 [62] was run on each nucleotide alignment using the HKY+Γ+I model and a fixed topology (congruent with the species tree), resulting in estimates for the branch lengths of each gene tree. Lastly, these branch lengths $\boldsymbol{L}$ were used in our EM method to estimate the model parameters (Figure 6.2 and Figure 6.3).

The second method (*Estimating duplication and loss rate parameters*) estimates our duplication and loss parameters from gene counts present within gene family clusters that contain duplications and losses. For the 16 fungi, we used gene counts from gene families previously clustered [10] to estimate the gene

duplication and loss rates $\lambda = 0.000732, \mu = 0.000859$ (events/gene/million years). For the 12 *Drosophila* clade, we used the duplication and loss rates $\lambda = 0.0012, \mu = 0.0012$ that were previously estimated [66].

9.3 Reconstructing gene families from 16 fungi

In our first evaluation, we analyzed the performance of SPIMAP versus several other popular phylogeny programs on a dataset of 16 fungi species. We have included three "traditional" methods: PHYML v2.4.4 [62] (Maximum Likelihood), BIONJ [53] (Neighbor-Joining), and MrBayes v3.1.1 [125] (Bayesian). We have also evaluated several other methods that use species-related information, which we call "species tree aware". These include our previous method [121] SPIDIR, SYNERGY [145], and PRIME-GSR [3].

For our 16 fungi real dataset, we downloaded coding sequences and peptides from the January, 2009 update of fungi dataset used by the SYNERGY method [145, 146]. By using this data as the input for all the other methods, we can compare against the trees constructed by SYNERGY (also downloaded from the January, 2009 update). We focused the analysis on the same 16 species as used in [10], which is a tree that also agrees with the one used by SYNERGY. We used the same gene clusters as defined by SYNERGY's trees, in effect using SYNERGY as the clustering step for the phylogenomic pipeline (Figure 5.1a). Peptide alignments were made using MUSCLE [40] and coding sequences were mapped onto them to produce nucleotide alignments. In addition, from the nucleotide alignments, we also produced RY-encoded alignments, which only indicate whether a base is purine (R) or pyrimidine (Y). No other information from SYNERGY trees was made available to the other methods.

We used the following parameters for each of the methods. For PHYML and BIONJ, we used a HKY+Γ+I model of nucleotide substitution. We configured MrBayes with four chains, an automatic stop rule, a 25% burn-in, sampled every 10 generations from a total of 10,000 generations, a 4by4 model for nucleotides, and enforced a binary tree. For methods that do not produce reconciled trees (i.e. PHYML, MrBayes, BIONJ), we have used maximum parsimonious reconciliation (MPR) to infer duplications and losses. For SPIDIR, we used duplication and loss penalties of .001 and an error cost of -600. For PRIME-GSR, we used 50,000 iterations, the JTT model, gamma distributed rates, and our own species tree (Figure 9.1). The tree search was initialized by an ML tree found by PHYML. We also ran PRIME-GSR with 1,000,000 iterations (as recommended by Åkerborg *et al.* [3]) but for only 500 trees randomly chosen from the dataset in order to limit the computational run time. SPIMAP was executed with two settings: "long" (2000 iterations with 1000 prescreening iterations) and "short" (100 iterations with 1000 prescreening iterations). For all other programs and options, defaults were used.

Table 9.1: **Evaluation of several phylogenetic programs on gene trees from 16 fungi**

Program	% Orthologs[a]	# Orthologs[b]	# Dup[b]	# Loss[b]	avg. run time
SPIMAP (quick)[c]	96.2%	550,800	5,541	10,884	1.0 m
SPIMAP (long)[c]	96.5%	557,981	5,407	10,384	21.9 m
SPIMAP (iid)[d]	93.9%	547,976	6,201	13,428	21.6 m
SPIDIR	83.3%	524,292	10,177	33,550	2.2 m
SYNERGY	99.2%	595,289	4,604	8,179	–[e]
PRIME-GSR (quick)[c]	88.9%	527,153	7,951	21,099	53.1 m
PRIME-GSR (long)[c]	90.7%	–	–	–	20.7 h
MrBayes	63.9%	460,510	21,307	65,238	43.2 s
PHYML	64.2%	464,479	21,264	64,391	45.3 s
BIONJ	60.4%	439,193	22,396	71,231	0.5 s

[a] Percentage of syntenic orthologs recovered.
[b] Number of pair-wise orthologs, duplications, and losses inferred from trees.
[c] Both SPIMAP and PRIME-GSR were run with a few iterations ("quick") of 100 and 50,000 and with many iterations ("long") 2000 and 1,000,000.
[d] SPIMAP was also run using a i.i.d. species-specific rate model.
[e] Since SYNERGYtrees were downloaded, no run time was estimated.

Although, a ground truth is not known for real datasets, we have used several informative metrics to assess the quality of gene trees, gene duplications, and losses inferred by these methods. Each of these metrics also illustrate different advantages and short-comings of each method.

9.3.1 Recovering syntenic orthologs

The first metric we investigated was the ability to infer syntenic orthologs – pairs of genes that are highly likely to be orthologous given their surrounding conserved gene order. Although not all orthologous pairs are syntenic, synteny information does allow us to identify a conservative set of orthologous genes using a method independent of phylogenetics, and thus provides a useful gold standard to test against. See Chapter A.2 for a description of our synteny determination method. When we construct trees on families that contain such genes, we expect a syntenic gene pair to appear within the reconstructed gene tree such that their most recent common ancestor is a speciation, and thus are inferred as orthologs.

SPIMAP recovered syntenic orthologs with 96.5% sensitivity followed by PRIME-GSR at 88.9% and PHYML at 64.1% (Table 9.1). Since SYNERGY uses synteny as one of its inputs, this test alone cannot assess its accuracy, and indeed 99.2% of syntenic genes are orthologs in SYNERGY's trees. When given more iterations, PRIME-GSR's accuracy increases to 90.7% but computational time increases dramatically, 24-fold from 53 minutes to 20 hours. In contrast, SPIMAP achieved its accuracy of 96.5% in 29.1 minutes

species	A	C	G	T	R (AG)	Y (CT)	W (AT)	S (GC)
S. cerevisiae	0.3291	0.1907	0.2050	0.2752	0.5341	0.4659	0.6044	0.3956
S. paradoxus	0.3276	0.1921	0.2065	0.2738	0.5341	0.4659	0.6014	0.3986
S. mikatae	0.3311	0.1873	0.2045	0.2772	0.5356	0.4644	0.6083	0.3917
S. bayanus	0.3225	0.2024	0.2109	0.2643	0.5334	0.4666	0.5868	0.4132
C. glabrata	0.3278	0.1901	0.2141	0.2680	0.5419	0.4581	0.5958	0.4042
S. castellii	0.3397	0.1775	0.1979	0.2849	0.5376	0.4624	0.6246	0.3754
K. waltii	0.2891	0.2267	0.2332	0.2511	0.5222	0.4778	0.5401	0.4599
K. lactis	0.3218	0.1909	0.2089	0.2784	0.5307	0.4693	0.6002	0.3998
A. gossypii	0.2551	0.2494	0.2739	0.2217	0.5290	0.4710	0.4767	0.5233
C. albicans	0.3492	0.1659	0.1863	0.2987	0.5354	0.4646	0.6479	0.3521
C. tropicalis	0.3476	0.1603	0.1847	0.3075	0.5323	0.4677	0.6551	0.3449
L. elongisporus	0.3286	0.1930	0.2113	0.2671	0.5400	0.4600	0.5957	0.4043
C. parapsilosis	0.3244	0.1869	0.2112	0.2775	0.5356	0.4644	0.6020	0.3980
D. hansenii	0.3368	0.1721	0.2023	0.2887	0.5391	0.4609	0.6256	0.3744
C. guillermondii	0.2909	0.2217	0.2255	0.2619	0.5164	0.4836	0.5528	0.4472
C. lusitaniae	0.2779	0.2290	0.2375	0.2556	0.5154	0.4846	0.5334	0.4666
standard deviation	0.0264	0.0244	0.0215	0.0199	0.0078	0.0078	0.0454	0.0454
CV	8.29%	12.46%	10.08%	7.33%	1.47%	1.67%	7.69%	11.10%

Table 9.2: **Stability of purine (R) and pyrimidine (Y) frequency across fungal species.** Although within coding sequence, GC content varies greatly across the 16 fungal species (standard deviation 0.0454), the frequency of purines (R) and pyrimidines (Y) is more consistent (standard deviation 0.0078, coefficient of variation (CV) 1.47%-1.67%).

on average per tree, and can achieve as much as 96.2% accuracy even when limited to an average run time of 1.0 minute ("quick" mode). Also, SPIMAP achieves 96.3% ortholog accuracy when assessing the same 500 tree subset as PRIME-GSR's "long" mode. Note that the species tree aware programs (SPIMAP, SYNERGY, and PRIME-GSR) predict as much as 20% more ortholog pairs than the leading competing traditional program (PHYML).

For SPIMAP, performance was greater on RY-encoded alignments (96.5%) versus the full nucleotide alignments (92%, data not shown). This is likely due to that fact that the nucleotide alignments contained a GC bias that varies across species (Table 9.2), thus violating the stationarity assumption made in our implemented sequence evolution model (HKY). Reconstruction accuracy of PHYML and MrBayes was slightly diminished on RY-encoded alignments (63.0% and 61.1%, respectively), mostly due to the their lower information content.

One important distinction between SPIMAP and PRIME-GSR is that SPIMAP models species-specific rate variation. To investigate the effect of this difference, we configured SPIMAP to learn an i.i.d. rates model, similar to PRIME-GSR. For each branch, our modified training step estimated ($\alpha_i = 2.819, \beta_i = 663.0$) as the parameters for the i.i.d. gamma distributions. Reconstructing gene trees using these parameters, we found fewer syntenic orthologs (93.9%) and greater numbers of duplications and losses.

9.3.2 Counting duplication and loss events

Second, we evaluated the total numbers of duplications and losses inferred across the clade (Figure 9.2). We found similar estimates between SPIMAP and SYNERGY(5,407 vs. 4,604 duplications and 10,407 vs. 8,179 losses). In contrast, traditional methods that do not use the species tree, infer many more events on nearly every branch, especially for short interior branches. The distribution of duplication and loss events that occur within each gene tree is illustrated in Figure 9.3. Interestingly, each of the other traditional methods inferred over four times as many gene duplication events and six times as many gene loss events as SPIMAP. For the traditional methods, duplications are more frequent near the root of the species tree and losses are more frequent near the leaves, a pattern suggesting that these events are erroneous [63].

9.3.3 Duplication consistency score

With our third metric, we sought to characterize the plausibility of the inferred duplications using the *duplication consistency score*, introduced by Ensembl for evaluating their phylogenomic pipeline [143]. The consistency of a duplication node with children l and r, is define as $|A \cap B|/|A \cup B|$, where A and B are the set of species represented in descendants of l and r, respectively (see example in Figure 9.4a). The consistency score is designed to detect duplications that are wrongly inferred due to phylogenetic reconstruction errors, since such false duplications are often followed by many compensating losses [63, 143] (i.e. low species overlap $|A \cap B|$). Figure 9.4 depicts the distribution for the duplication consistency score for each program. Both SPIMAP and SYNERGY showed similar consistency distributions that are heavily shifted towards 1 (47.8%-49.0% and 4.2%-17.2% of duplications with a score of 1 and 0, respectively; Figure 9.4). The traditional methods have many low scoring duplications ($<11\%$ and $>70\%$ with scores 1 and 0, respectively), an effect seen previously [143]. PRIME-GSR's distribution lies in between these extremes with 30.0% and 42.1% for scores 1 and 0, respectively. Lastly, the i.i.d. version of SPIMAP also scored lower than SPIMAP, inferring 10% more duplications with a consistency score of zero (Figure 9.4).

9.3.4 Recovering gene conversions

The four metric was specifically designed to test the case where species-level information is misleading, effectively testing the ability of species-aware methods to properly weigh species information against conflicting sequence information.

The fungal clade contains a whole-genome duplication (WGD) event, such that every gene duplicated simultaneously followed by many gene losses [148, 78]. Of the paralog pairs that are still present in the *S.*

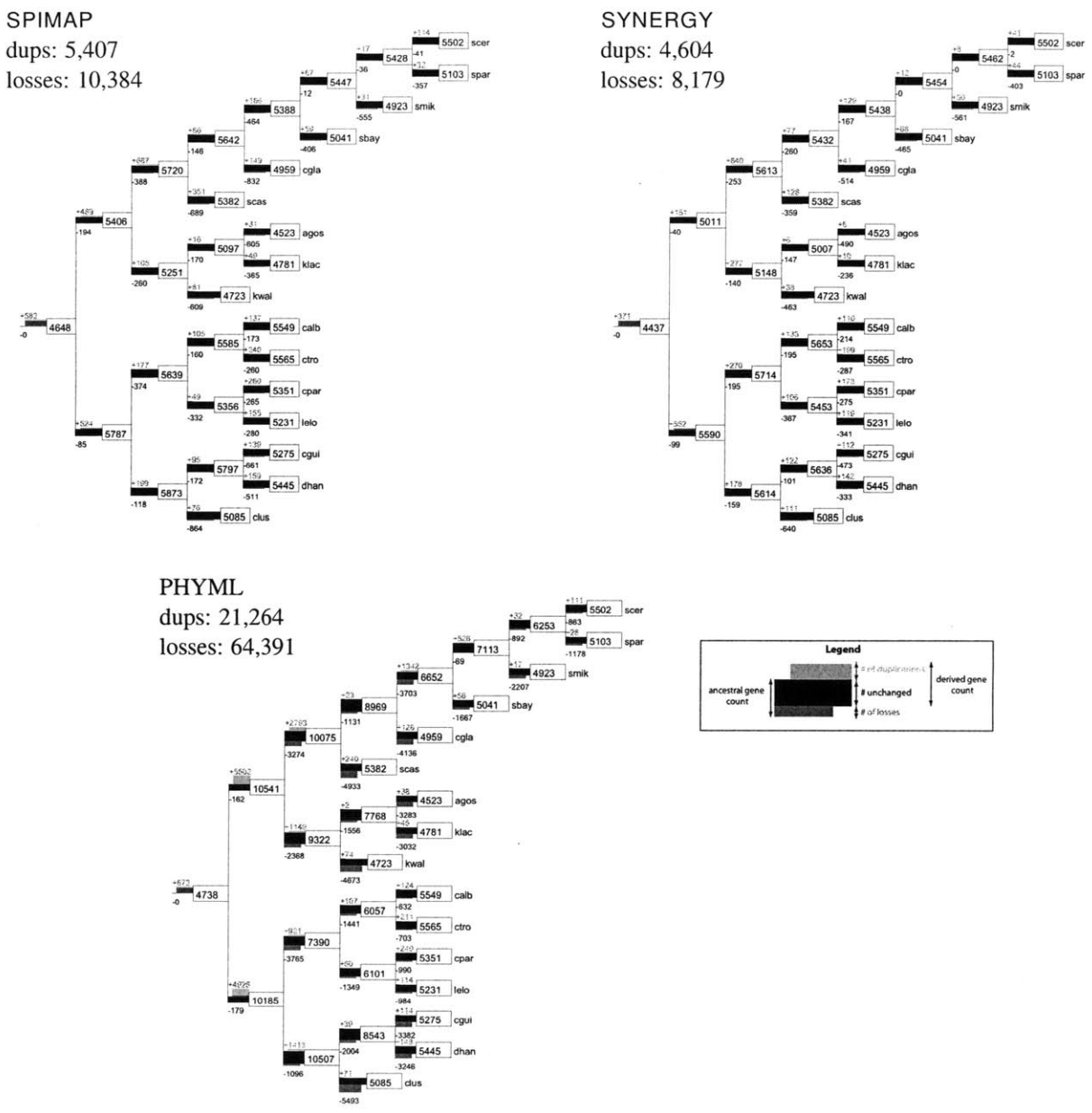


Figure 9.2: Total counts of duplications and losses inferred on the 16 fungi species phylogeny by SPIMAP (left), SYNERGY (right) and PHYML (bottom). Both SPIMAP and SYNERGY find similar numbers of events while PHYML infers many more ancient duplications followed by many compensating losses. Duplications for each branch are indicated by green text and bars, losses are indicated by red text and bars, and gene appearances are indicated by blue text and bars. Thickness of bars is proportional to the number of genes duplicated, lost, appearing, or inherited. The total thickness of the branch represents genome size along each branch.

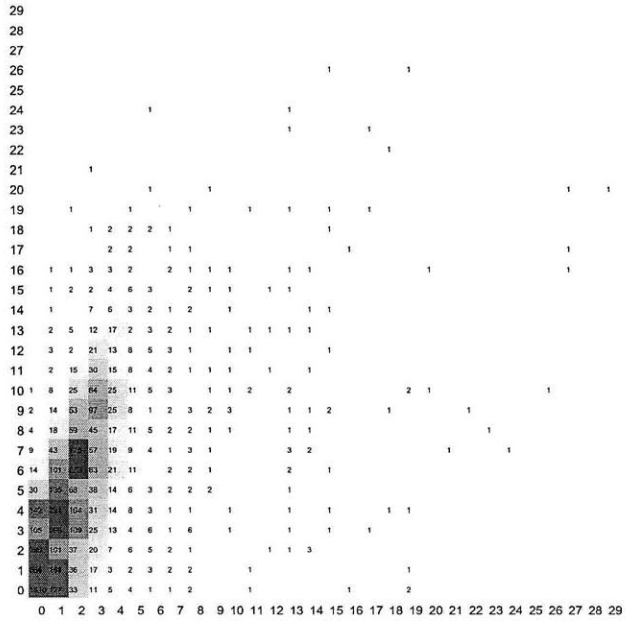


Figure 9.3: Distribution of gene duplication (horizontal) and loss (vertical) events per gene tree as inferred by SPIMAP for the 16 fungi dataset (5351 trees).

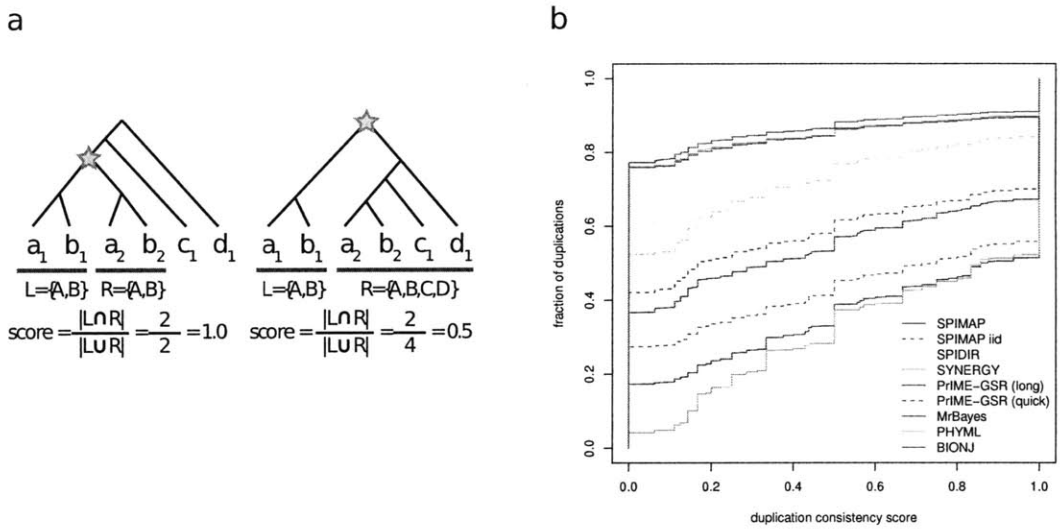
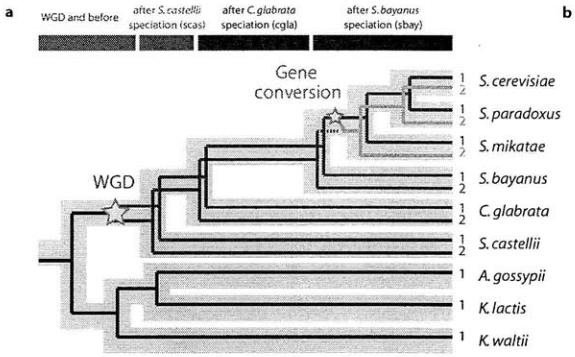


Figure 9.4: **The duplication consistency score for assessing phylogenetic methods.** (a) Duplication consistency score, computed on two example trees. For each duplication node (star), this score computes the number of species present in both the left and right subtrees divided by the total number of species descendant from the duplication node. Erroneous duplications show an increased rate of compensating losses, and thus lower scores. (b) Cumulative distribution of duplication consistency scores for all duplications inferred in the 16 fungi dataset by each method. SPIMAP (blue) and SYNERGY (green) perform best according to this metric, having the fewest duplications with low consistency scores. SPIMAP trained with an i.i.d. model similar to PRIME-GSR (dashed blue) infers duplications with overall lower consistency scores. These are followed by PRIME-GSR (dark green) and SPIDIR (dashed light blue) that show more moderate performance. Lastly, the three traditional methods implemented in the programs MrBayes, PHYML, and BIONJ, all have similar and significantly lower score distributions.



b

Program	WGD	scas	cgla	sbay	% conversion recovered
SPIMAP	5	1	16	15	83.8%
SPIDIR	9	4	18	6	64.9%
SYNERGY	34	2	1	0	2.7%
PrIME-GSR	4	0	24	9	89.2%
MrBayes	5	0	21	11	86.5%
PHYML	5	0	19	13	86.5%
BIONJ	5	1	19	12	83.8%

Figure 9.5: **Inferred duplication times for recent *S. cerevisiae* gene conversions.** (a) Typical gene-tree topology for 37 paralogous gene pairs originally arising from whole-genome duplication (WGD) and previously reported [52] to have undergone gene conversion events (small star) near or after the speciation of *S. cerevisiae* and *S. bayanus*, such that one gene copy (green) is replaced by the other (red), followed by subsequent nucleotide divergence (orange). The correct inferred duplication of the two *S. cerevisiae* paralogs (red and orange lines, denoted 1 and 2) should occur within the time span indicated by the top brown bars. However, we expect methods that are heavily biased to follow the known species tree to incorrectly infer these events further up the tree. (b) We evaluated both traditional and species-aware methods in their ability to recover the correct trees in these cases, and report the counts of where different gene conversion events are inferred for each method. We find that both SPIMAP and PRIME-GSR, as well as all traditional methods find the vast majority of these paralogs duplicates near or after *S. bayanus* speciation. However, SYNERGY incorrectly infers a WGD topology, most likely due to strong reliance in synteny information which is misleading in this case.

cerevisiae genome, 37 of them have a K_s less than the average K_s between the *S. cerevisiae* and *S. bayanus* genomes of 1.05, indicating that these paralogs have undergone recent gene conversions near or after the speciation of the *S. cerevisiae* and *S. bayanus* lineages [52] (see an example in Figure 9.5a). Also indicative of gene conversion [110], these genes have a significantly elevated GC frequency of 42.0% in the third codon position, compared to a frequency of 37.9% for all *S. cerevisiae* genes ($P < 1.7^{-09}$; Mann-Whitney U). Of these paralogs, SPIMAP infers 15 of them happening after the *S. bayanus* speciation and 31 after the *C. glabrata* speciation (Figure 9.5b). In comparison, SYNERGY infers none of the paralogs duplicating after the *S. bayanus* speciation and only 1 after the *C. glabrata* speciation. Instead 34 of the 37 paralogs are inferred as occurring on the branch containing the WGD, thus indicating that synteny information between *S. cerevisiae* and other post-duplication species overrides sequence information in the vast majority of cases. For 33 families, the SPIMAP-constructed tree has a higher likelihood than the SYNERGY tree and for 22 families the likelihood is significantly higher ($P<.01$; SH-test). In contrast, SYNERGY never has significantly higher likelihood.

Together these four metrics applied to real gene trees from 16 fungi suggest that SPIMAP often out-

performs both traditional and species-aware methods. From these trees, we observe what appears to be an over estimation of duplication and loss events by the other methods, an error which has been observed in previous empirical studies [63]. To better understand how phylogenetic errors influence the accuracy of event inference, we turn now to simulated data.

9.4 Reconstructing simulated gene trees

To test our method on a dataset where the correct phylogeny is unambiguously known, we implemented a simulation program based on our model for gene family evolution. Our intent was to make the simulations realistic by capturing the same gene and species-specific rate variation as well as gene duplication and loss rates as seen in real gene trees. Thus, the same model parameters and species phylogeny were used as those estimated for both the 12 flies and 16 fungi clades.

For each clade, we simulated 1000 gene trees and generated the corresponding nucleotide alignments (Figure 9.6 and Figure 9.7). Next, we reconstructed gene trees from these simulated alignments using SPIMAP and the other traditional phylogenetic methods. Since PRIME-GSR uses an i.i.d. model for species-specific rates, we choose to exclude it from this analysis (see *Search efficiency* for a comparison). SPIMAP's substitution rate parameters were estimated on a simulated dataset with no duplications and losses (Figure 6.2 and Figure 6.3). Its duplication and loss parameters were trained from the gene counts of each simulated dataset (Table 6.1).

First, we measured topology accuracy across all of the methods. SPIMAP outperforms the other programs by 7%-29% on the simulated 12 flies dataset and by 52%-81% for the 16 fungi dataset (Figure 9.8a). The accuracy improvement for SPIMAP is larger on the fungi dataset, which has a more complex and divergent phylogeny.

Second, we assessed partial topology correctness using the percent of branches accurately reconstructed. For the flies, SPIMAP consistently performs better but by only a few percent (Figure 9.8b). However, for the fungi, SPIMAP again shows a larger accuracy improvement at 20%-39% over other methods.

Third, we looked at the percentage of orthologs inferred correctly, where we noticed a surprising trend. Although topologies and branches had high error rates for many methods, there was also a high percentage of correctly inferred ortholog pairs (Figure 9.8c). Upon closer inspection, we found that often when a branch is misplaced it only disrupts a small fraction of the pair-wise orthologs. Thus, it appears that orthology discovery at the pair-wise level is quite robust to phylogenetic errors. In addition, we noticed that false positive orthologs calls are rarely made, although false negatives are more frequent, especially on the fungal

clade.

Fourth, we looked at the accuracy of inferring gene duplications and losses, which is very important for studies interested in study the rate of such events. As opposed to the ortholog pair-wise metric, we find that duplications and losses are very sensitive to phylogenetic errors. Notice, that although branch accuracy may be high for some programs and datasets, even a small number of errors can lead to dramatic overestimation duplications and losses (Figure 9.8c,d and Figure 9.10). In general, all programs are able to recover duplication and loss events for the flies and fungi datasets with similar sensitivity (<6% difference, with SPIDIR and BIONJ as outliers). However, in terms of precision SPIMAP has a dramatic improvement in event estimation: 21%-27% and 45%-53% for the flies duplication and loss, respectively, and 58%-69% and 75%-80% for fungi duplications and losses (with BIONJ as an outlier in each case). This 2 to 3 fold over prediction of events by the other phylogenetic methods (Figure 9.8c,d) is an effect similar to that seen in the real data.

Lastly, we find that these results also hold when simulations are performed with unusually high duplication and loss rates at twice (2X) and four-times (4X) the estimated true rates (1X). We performed simulations with five different settings 1X-1X, 2X-2X, 4X-4X, 4X-1X, 1X-4X for duplication and loss rates respectively. We find that SPIMAP has increased performance for topology, branch, and event accuracy for all of these rate settings (Figure 9.9).

9.5 Search efficiency

In addition, to evaluating reconstruction accuracy we also evaluated reconstruction speed. Our goal with SPIMAP, was to develop a method that is feasible enough to include in a phylogenomic pipeline containing thousands of trees and a variety of family sizes.

From the reconstruction of genes from our real dataset (Table 9.1), we found that SPIMAP has an average reconstruction time per tree (1.0 minutes) that is only slightly longer than that of PHYML (43.2 seconds). To investigate how our search strategy influences reconstruction run-time, we generated a simulated dataset of 500 gene families using 16 fungi species tree. For this simulation, we used i.i.d. species-specific rates ($\alpha_i = 2.819, \beta_i = 663.0$), no variation occurs in the gene rate, and the Jukes-Cantor model. We also used the same gene duplication and loss rates as estimated from real fungi gene families ($\lambda = 0.000732, \mu = 0.000859$). SPIMAP's substitution rate model was trained on a dataset with the same parameters but no duplications and losses. The parameters used by SPIMAP during reconstruction are given in Figure 6.4.

Although, we have not implemented many optimizations for SPIMAP, our prescreening search strategy

Table 9.3: **Evaluation of search time for several phylogenetic methods**

Program	iterations[a]	prescreens[b]	bootstraps	topology	branch	run time
PHYML	–	–	0	26.0%	83.9%	25.8 s
PHYML	–	–	100	26.0%	83.9%	13.9 m
SPIMAP	50	1	0	32.4%	81.4%	7.2 s
SPIMAP	100	1	0	50.8%	87.1%	12.7 s
SPIMAP	500	1	0	83.8%	96.0%	1.2 m
SPIMAP	1000	1	0	88.6%	97.5%	2.0 m
SPIMAP	50	100	0	84.8%	96.7%	8.5 s
SPIMAP	1000	100	0	90.8%	98.1%	2.3 m
SPIMAP	50	100	100	86.4%	97.1%	11.1 m

[a] Number of iterations used for each method.
[b] Number of prescreening iterations used for SPIMAP.

allows SPIMAP to compete with the highly optimized PHYML program (Table 9.3). We believe this is because the gene family model, through the use of the species tree in the prior, produces a posterior distribution that is far more concentrated than the likelihood. Thus, many seemingly equivalent trees from a likelihood perspective are significantly different based on their priors and posteriors. In addition, our prescreening search strategy (*Rapid tree search*) appears to greatly help in speeding up discovery of the MAP gene tree. For example, with no prescreening, SPIMAP achieves a topology accuracy of 32.4% with an average run time of 7.2 seconds. By using 100 prescreening iterations, accuracy increases to 84.8% while run time only increases to 8.5 seconds. For comparison, PHYML achieves 26.0% topology accuracy in about 25.8 seconds on average.

SPIMAP is currently implemented as a Maximum *a posteriori* (MAP) method, thus if branch support values are needed, bootstrapping will be required. Given the speed of our search, we can perform 100 bootstraps in about 11.1 minutes to achieve 86.4% accuracy. This run time is comparable to 100 bootstraps of PHYML at 13.9 minutes and 26.0% accuracy. Thus, bootstrap analysis is quite feasible for SPIMAP, and the method should be efficient and practical enough for any pipeline that uses phylogenetic programs with run times on the order of PHYML's.

Lastly, we evaluated the influence of run time and family size on reconstruction accuracy. Using the same parameters above, we simulated more gene trees from the 16 fungal species tree and divided them into six classes based on the their number of extant genes: 5-9, 10-19, 20-29, 30-39, 40-49, and 50-59. Each size class was populated with 100 simulated trees and alignments. SPIMAP was run in two modes, one without bootstrapping (1000 iterations and 100 prescreens) and one with 100 bootstraps (100 iterations and 100 prescreens). For the middle gene size class 20-29, SPIMAP achieved average run-times of 5.3 minutes

and 50.4 minutes, respectively. For each dataset, PRIME-GSR was also executed, using the same amount of time as SPIMAP, which required 7300 iterations (quick mode) and 77,000 iterations (long mode) . We find that for smaller trees with 5 to 29 extant genes, that both SPIMAP runs and PRIME-GSR's long mode achieve similar topology accuracy in the range of 80%-100% (Figure 9.11). However, for larger gene trees with 30-49 extant genes, as accuracy degrades for all methods, both modes of SPIMAP have a 20% increase in topology accuracy over PRIME-GSR. Improvements in inferring duplication and loss accuracy is also seen for the larger trees ($> 10\%$ increase in duplication precision and $> 30\%$ increase in loss precision).

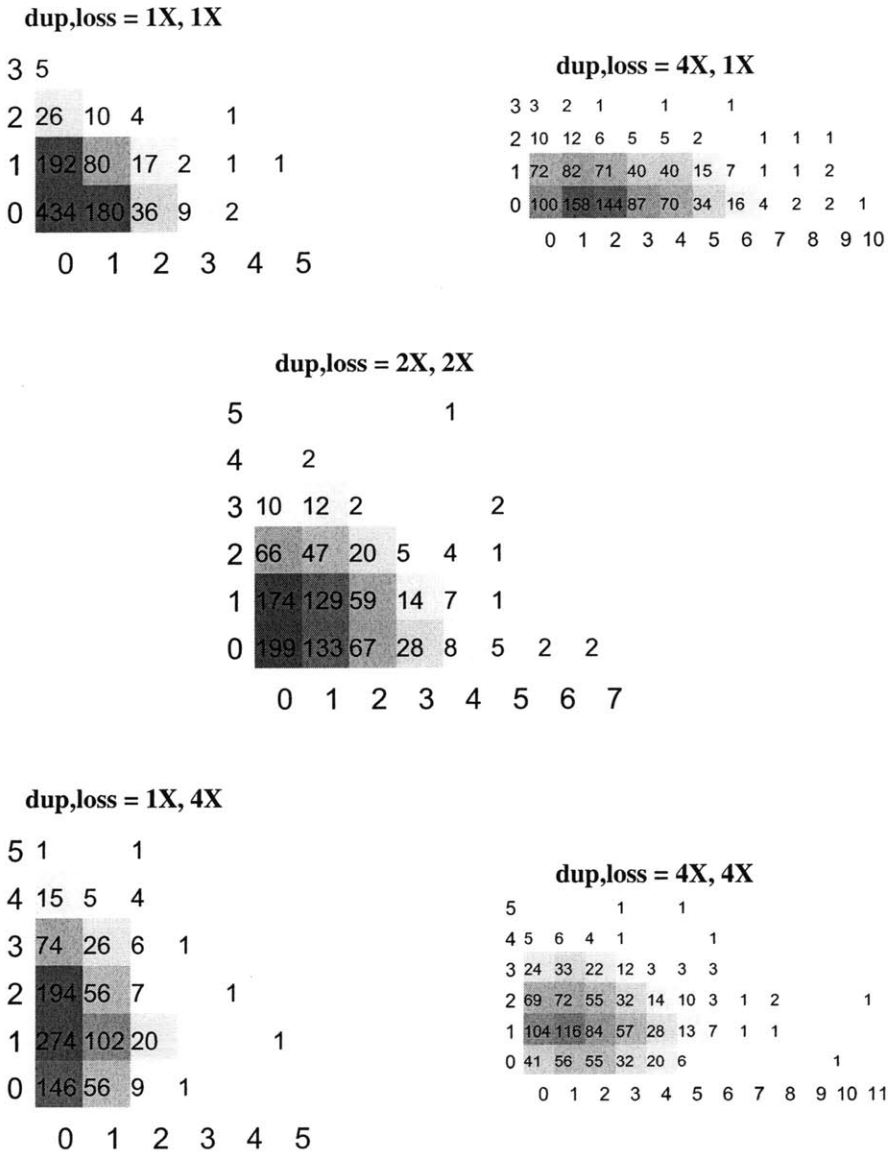


Figure 9.6: Distribution of observed gene duplication (horizontal) and loss (vertical) events per tree in the simulated 12 flies dataset (500 trees each). Event distributions are shown for each of the 5 duplication and loss rate settings. These rate settings provide a variety of gene tree sizes for evaluating the phylogenetic methods.

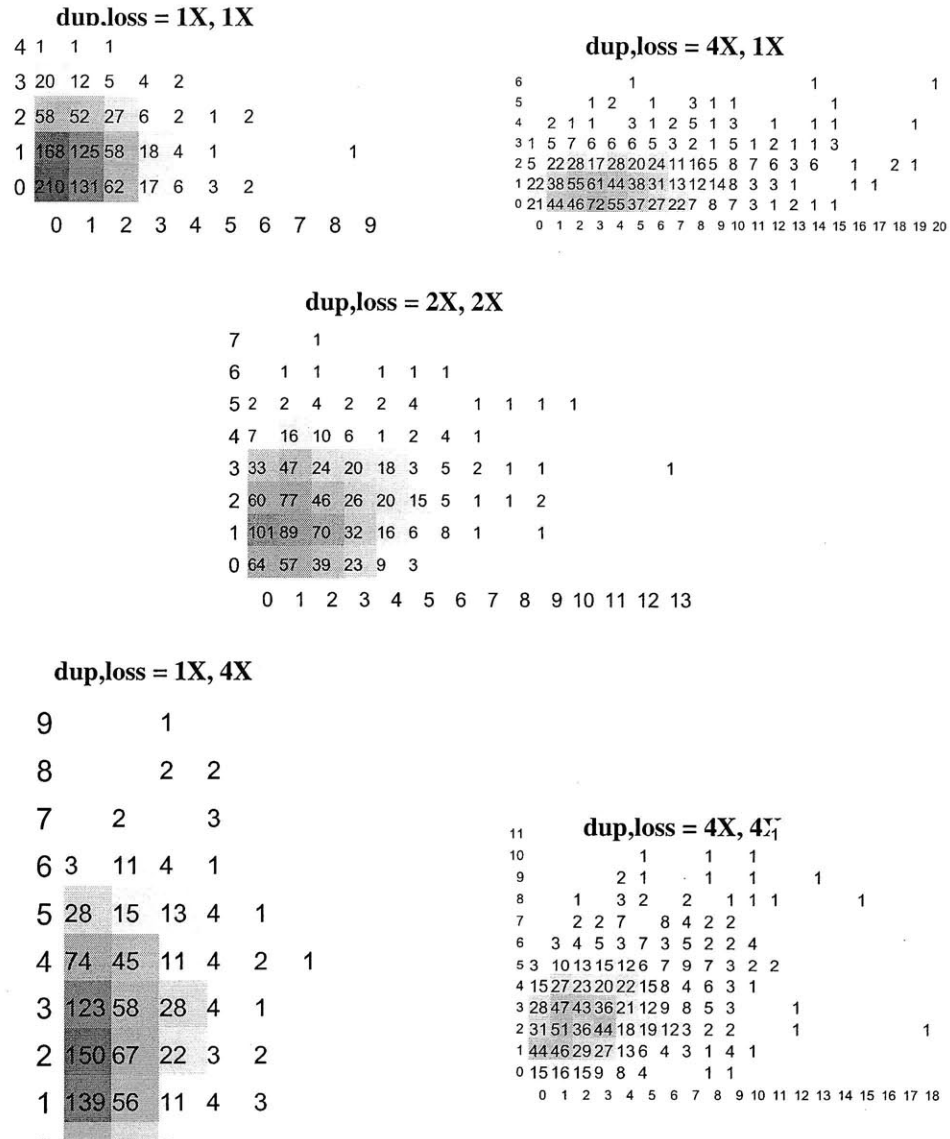


Figure 9.7: Distribution of observed gene duplication (horizontal) and loss (vertical) events per tree in the simulated 16 fungi dataset (500 trees each). Event distributions are shown for each of the 5 duplication and loss rate settings. These rate settings provide a variety of gene tree sizes for evaluating the phylogenetic methods.

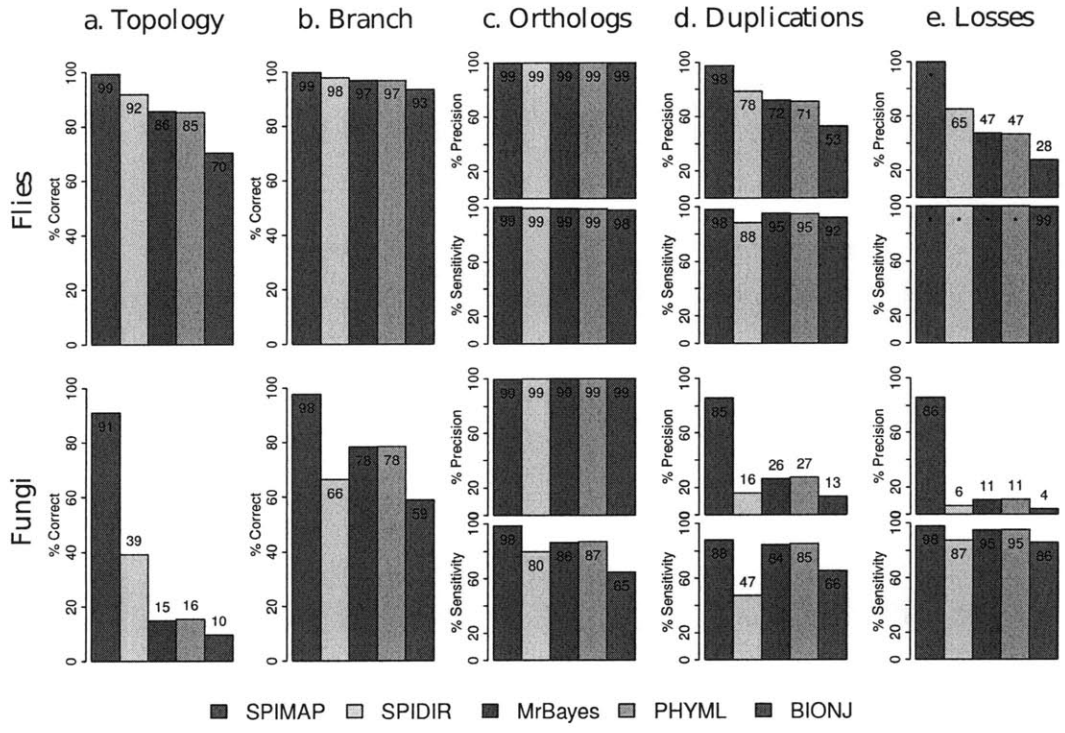


Figure 9.8: **Metrics of phylogenetic accuracy on simulated datasets for 12 *Drosophila* and 16 fungal species.** (a) SPIMAP has a higher reconstruction accuracy for correctly inferring the full gene-tree topology for both fly and fungal datasets. (b) The percent of accurately reconstructed branches is similar across methods for the 12 flies but a larger improvement is seen for SPIMAP on the larger and more diverse 16 fungi clade. (c) Despite topological and branch inaccuracies, pair-wise ortholog detection is robust in both precision and sensitivity. (d,e) In contrast, duplication and loss inference is very sensitive to phylogenetic errors, especially in terms of precision. Stars indicate 100%.

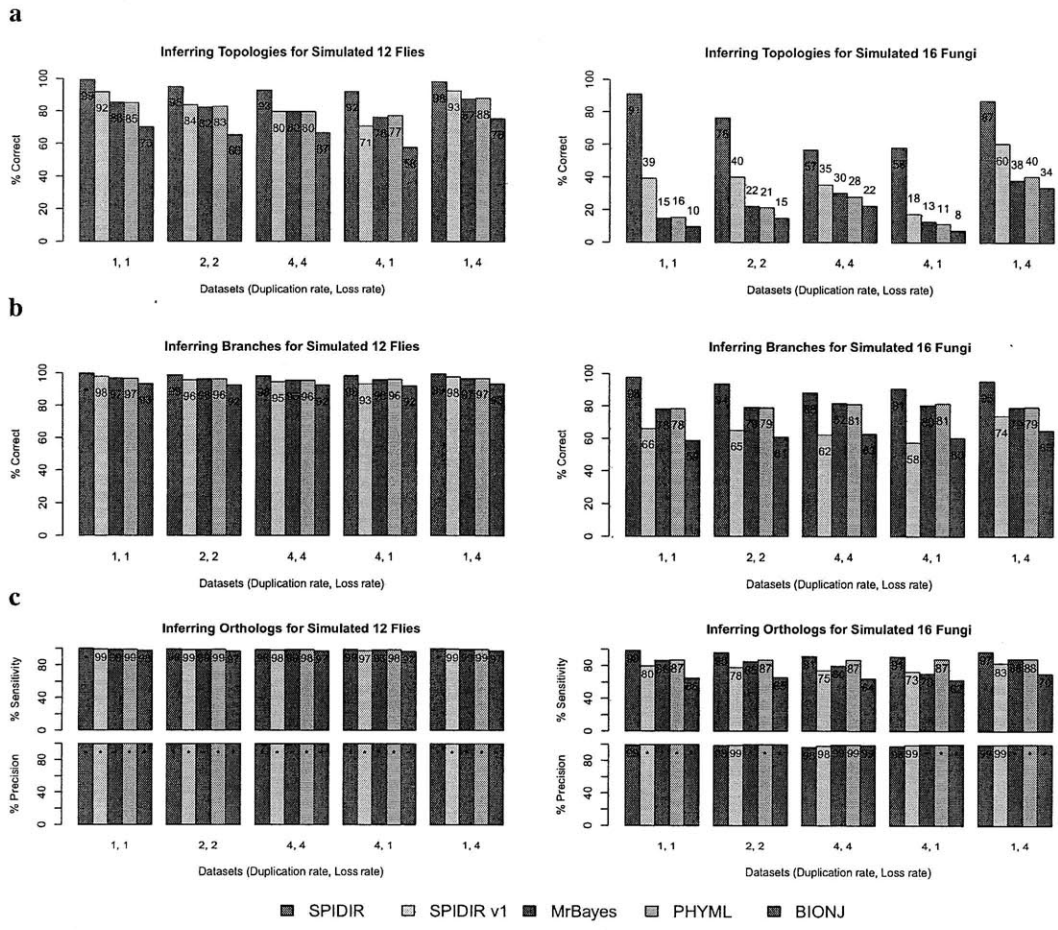


Figure 9.9: Reconstruction accuracy at increasing duplication and loss rates. Accuracy is measured for (a) topology, (b) branch, and (c) pair-wise orthology on both the 12 *Drosophila* and 16 fungi simulation datasets. 1000 alignments were simulated for each duplication and loss rate setting. Simulations were done with the same rates of duplication and loss as found in the real datasets (1,1), with twice the rate (2,2), and four times the rate (4,4). We also simulated more extreme cases such as 1,4 and 4,1. SPIMAP shows consistently higher accuracy in both clades, especially in the fungi due their larger trees. Orthology accuracy appears more robust to phylogenetic errors and is fairly high for many of the methods. PRIME-GSR was evaluated on this dataset since this data is not simulated with i.i.d. rates.

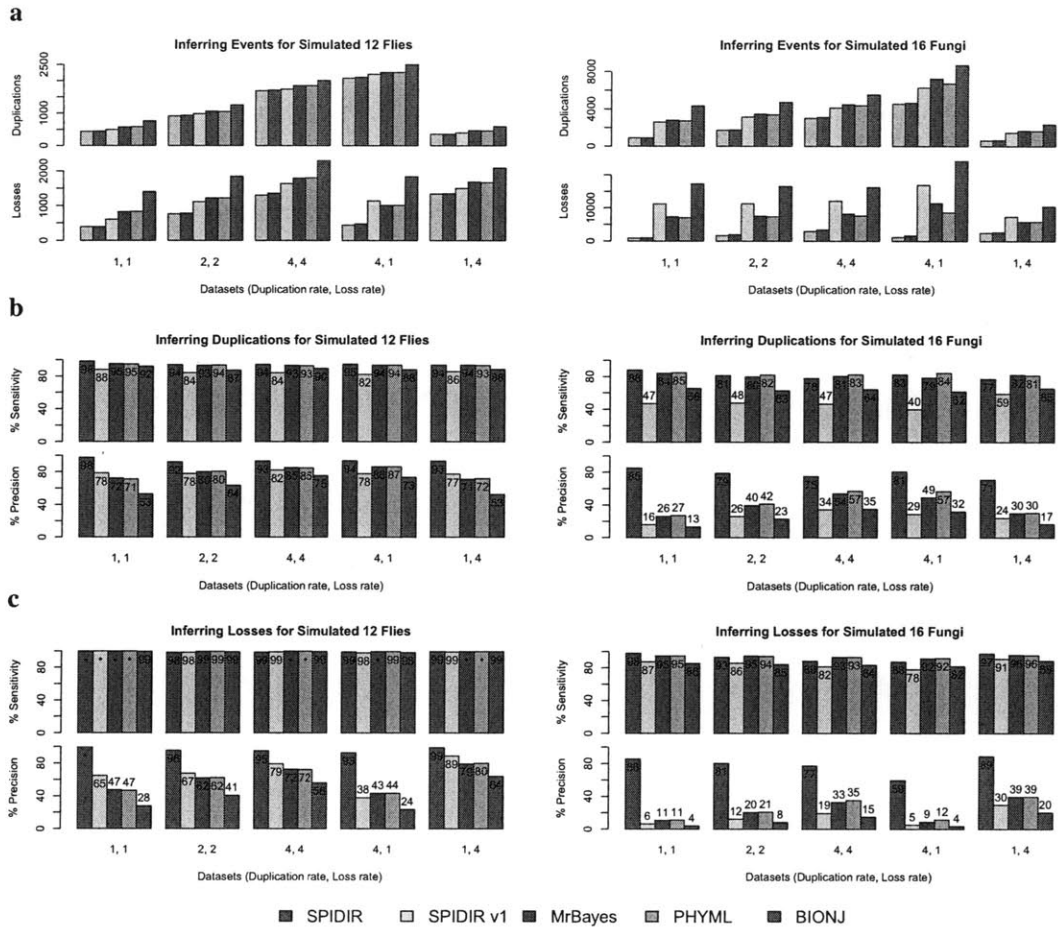


Figure 9.10: Event estimation by each program for increasing duplication and loss rates in the simulated datasets. (a) Number of inferred events by each program for each simulated dataset. The actual number of events are shown in grey bars. (b) Sensitivity and precision of estimating duplication events for the 12 *Drosophila* and 16 fungi simulation datasets. (c) Sensitivity and precision of estimating loss events for the 12 *Drosophila* and 16 fungi simulation datasets. Event estimation for SPIMAP remains high even for fast rate of duplication and loss.

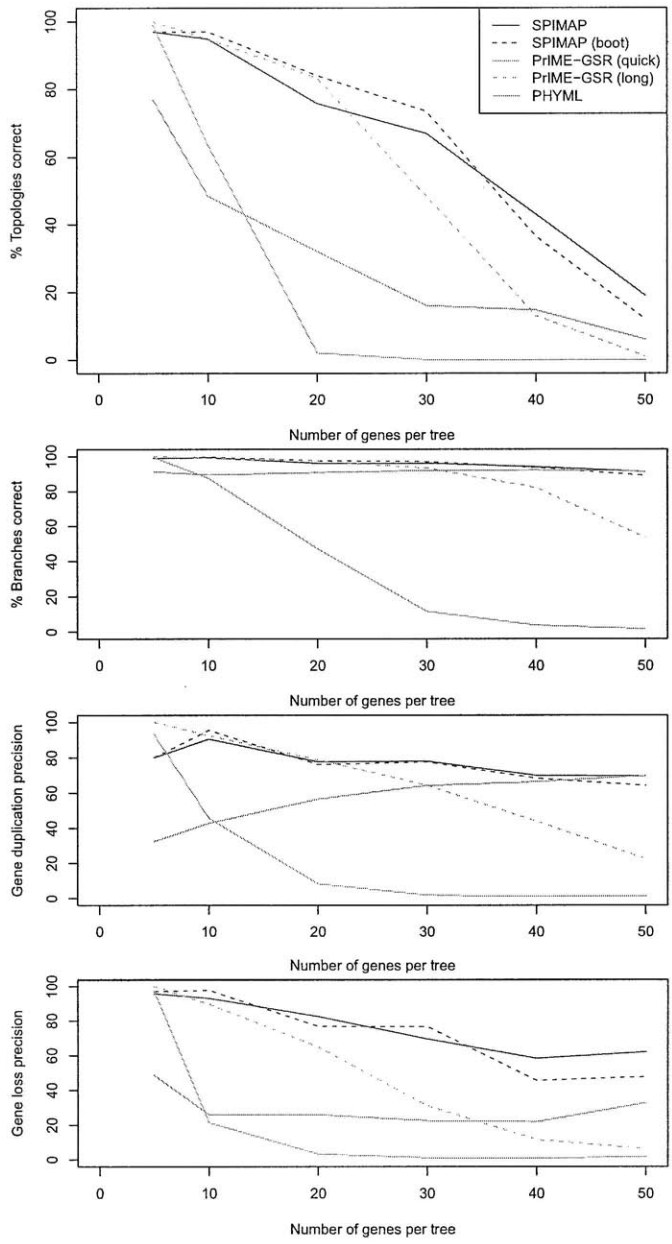


Figure 9.11: Reconstruction performance of several phylogenetic methods for gene trees of increasing size for 16 simulated fungi. Gene trees were simulated and divided into six classes based on the number of extant genes: 5-9, 10-19, 20-29, 30-39, 40-49, 50-59. Each size class was populated with 100 simulated trees and alignments. SPIMAP was run both with bootstrapping (100 iterations and 100 prescreens) and without bootstrapping (1000 iterations and 100 prescreens). For each dataset, PRIME-GSR was executed for the same amount of time taken by SPIMAPwith and without bootstrapping.

Chapter 10

Conclusions

10.1 Discussion

In this thesis, I have presented a novel probabilistic model and algorithm for gene tree reconstruction. The approach uses a Bayesian framework to model sequence evolution, gene duplication, loss, and substitution rate variation, thus incorporating many disparate types of information in a principled way. This unified framework presents many advantages.

In contrast to previous gene tree reconstruction methods [151, 34, 36, 143], where a gene tree is reconciled only after full reconstruction by a method such as Neighbor-Joining or ML, our method finds a reconciliation and gene tree simultaneously. In addition, the parameters of our model are interpretable (e.g. substitutions rates and duplication/loss rates), and we have provided training algorithms for each one. This provides an advantage over a method like SYNERGY [145] which optimizes a parsimony-based cost function for several different events such duplications, loss, and syntenic relationships. Without a probabilistic basis, the weights of these costs and the behavior of their combination are more difficult to determine and analyze. Our study of gene conversions demonstrates more work is needed to understand how synteny information should be weighed against conflicting sources of information.

Our method models rate variation that is correlated across all branches of the tree (gene-specific rate) as well as rates specific to each species lineage (species-specific rates). We have found that when both of these effects are modeled, the result is a more informative prior which leads to increased reconstruction accuracy (see the i.i.d. version of SPIMAP in Table 9.1 and Figure 9.4). In contrast, PRIME-GSR uses identical and independent gamma distributions for rate variation which do not model species-specific rate variation. Thus species with rate acceleration or decelerated across the genome will have branches that are consistently penalized by an i.i.d. rate prior. One complication for modeling species-specific rates is possibility of over-

parameterizing the model. We addressed this issue by learning the rate distributions prior to reconstruction from a dataset of multiple orthologous gene trees. By combining data across loci, the rate variation prior can be estimated more accurately than if the gene trees were considered in isolation.

The rate prior of our current work builds upon a previously developed method, SPIDIR [121]. We designed SPIDIR to be a distance-based likelihood method that exploits the rate variations we had observed in the 12 fly and 9 fungal genomes. Although the method proved effective, its reliance on pair-wise distances did not fully utilize the available character information and it lacked an explicit model for duplication and loss rates. Indeed, we find in our latest comparison that SPIMAP has more consistent accuracy improvements than SPIDIR even for large species trees (16 fungi) and fast rates of duplication and loss (Figure 9.8, Figure 9.9, and Figure 9.10).

We envision SPIMAP participating in a larger phylogenomic pipeline. We believe that within most clades of interest, there will be sufficient data for training our model. For example, in the 12 sequenced *Drosophila* species, about one third of all genes are syntenic across all 12 species [17, 121], and thus can serve as a training set for our substitution rates model. Once a model is learned from simple gene families, it can then be applied to reconstruct gene families with more complicated histories of gene duplication and loss. Given these advances and many others, phylogenetics will likely play an ever increasing role in understanding the evolution and function of genomes.

In Chapter 8, we implemented such a phylogenomic pipeline in order to study the gene families of 16 fungal species including *S. cerevisiae* and *C. albicans*. The trees allowed us to inferred gene families enriched in duplications specifically in pathogenic lineages. In addition, we were able to use the trees and their reconciliations to infer positive selection that specifically occurred in pathogenic species. Many families and protein functions currently known to be related to pathogenicity were found as well as several families where little is known thus far. Thus, evolutionary analysis such as these can provide clues about the functions of genes with currently unknown function.

10.2 Current directions

Going forward, there are many promising directions for how to further develop the work presented here. Several extensions to the substitution rate variation model are possible and one such possibility was presented in Chapter 6.3.3. However, it remains to be seen how much power they actually add in gene tree reconstruction in practice.

One particularly interesting line of work is to consider how to model duplications and losses in a popula-

tion setting. As more genomes are sequenced, our species trees will become more dense and the time spans we must consider will become as short as one or two million years. For example, in the primates clade there will soon be about 15 whole-genome sequences available [132]. Studying gene innovation through duplication and loss will be especially exciting in this clade, as it will shed light on the early stages of human evolution. However, such analyses will be quite challenging with our current models and algorithms.

Throughout this thesis and in the broader field, models for gene duplications and losses (both parsimony-based and probabilistic) currently assume that population-related effects are negligible, even though for many clades of interest this assumption may not be reasonable. In analyses of one-to-one orthologs, the coalescent model [83, 120, 25] has been used to study the distribution of the age the most recent common ancestor (MRCA) of two or more individuals. Although such a model will be import to use to study duplications in the primates, all coalescent models assume duplications and losses do not occur.

I believe it is possible to relax the assumptions of both the duplication-loss model as well as the multispecies coalescent, in order to form a new model that describes all of these event simultaneously. Such a model would lead to several useful algorithms for a variety of problems such as reconciliation, gene tree reconstruction, and species reconstruction. It may even have applications for problems within population genetics.

Appendix A

Supplementary material

A.1 M. Hasegawa and H. Kishino and T. Yano (HKY) model

This is a review of the HKY model [68].

The *rate matrix* for HKY is

$$Q = c \times \begin{pmatrix} - & \pi_c & \kappa\pi_g & \pi_t \\ \pi_a & - & \pi_g & \kappa\pi_t \\ \kappa\pi_a & \pi_c & - & \pi_t \\ \pi_a & \kappa\pi_c & \pi_g & - \end{pmatrix}.$$

with nucleotides ordered by A, C, G, T. The variables $\pi_a, \pi_c, \pi_g, \pi_t$ represent the equilibrium base frequency as thus sum to 1.

The constant c scales the overall rate of substitution. In most circumstances, the rate of substitution cannot be estimated separately from the duration of time. Thus, it is usually the convention to choose the scale factor c, such that the average rate of substitution at equilibrium is 1, namely

$$\text{avgerage rate} = \sum_i \pi_i \sum_{j \neq i} Q_{ij} = 1.$$

This convention is convenient, because estimating time t will also give the number of substitutions/site.

The coefficient κ is called the *transition/transversion ratio* and represents the ratio of the transition rate over the transversion rate. However, it should be noted that there is also another popular definition of the transition/transversion ratio. For example, Felsenstein will often use a variable R to represent this concept, however its definition takes into account the equilibrium base frequency. Given the rate of transitions t_s and

	A	C	G	T
A	-	v	s	v
C	v	-	v	s
G	s	v	-	v
T	v	s	v	-

Figure A.1: The location of transitions (s) and transversions (v) in a nucleotide substitution matrix. There are twice as many ways to preform a transversion (8) than a transition (4).

transversions t_v, we have

$$t_s = \pi_a Q_{ag} + \pi_g Q_{ga} + \pi_c Q_{ct} + \pi_t Q_{tc}$$
$$t_v = \pi_a(Q_{ac} + Q_{at}) + \pi_c(Q_{ca} + Q_{cg}) + \pi_g(Q_{gc} + Q_{gt}) + \pi_t(Q_{ta} + Q_{tg})$$
$$R = t_s/t_v.$$

Thus if there is no bias for transitions over transversion, such as in the Jukes-Cantor model, then $\kappa = 1$. However, for the other definition of this ratio we have $R = \frac{1}{2}$, because there are twice as many transversions (8) as transitions (4) (see Figure A.1).

For the HKY model, the two definitions of the transition/transversion ratio can be converted as follows:

$$R = t_s/t_v = \frac{2\kappa\pi_a\pi_g + 2\kappa\pi_c\pi_t}{2\pi_a\pi_c + 2\pi_a\pi_t + 2\pi_c\pi_g + 2\pi_g\pi_t}$$
$$= \frac{\pi_t\pi_c + \pi_a\pi_g}{\pi_y\pi_r}\kappa.$$

For the HKY model, the substitution matrix can be solved analytically. I have implemented the HKY model for my own work. For convenience, I include the relevant formulas for using this model.

Definitions:

j =destination base

i =source base

t =time

π_i =background/prior distribution of base$i$

R =Transition/Transversion ratio

κ =Transition/Transversion ratio (easier to specify)

Parameterization:

$$\pi_r = \pi_a + \pi_g$$

$$\pi_y = \pi_c + \pi_t$$

$$R = (\pi_t\pi_c + \pi_a\pi_g)\kappa/(\pi_y\pi_r)$$

$$\beta = 1/(2\pi_r\pi_y(1+R))$$

$$\rho = \pi_r/\pi_y$$

$$\alpha_y = \frac{\pi_r\pi_y R - \pi_a\pi_g - \pi_c\pi_t}{2(1+R)(\pi_y\pi_a\pi_g\rho + \pi_r\pi_c\pi_t)}$$

$$\alpha_r = \rho\alpha_y.$$

Convenience variables:

$$\alpha_i = \begin{cases} \alpha_r & \text{if } i \in \{A,G\} \\ \alpha_y & \text{otherwise} \end{cases}$$

$$\pi_{ry} = \begin{cases} \pi_r & \text{if } i \in \{A,G\} \\ \pi_y & \text{otherwise} \end{cases}$$

$$\delta_{ij} = [i = j]$$

$$e_{ij} = [(i \in \{A,G\}) = (j \in \{A,G\})],$$

where $[X]$ is 1 if X is true and 0 otherwise.

Formula: The probability of seeing a base j given a starting base i and a duration of time t is

$$\begin{aligned} P_{ij} = P(j|i,t,\pi,R) = &\exp(-(\alpha_i+\beta)t)\delta_{ij}+ \\ &\exp(-\beta t)(1-\exp(-\alpha_i t))(\pi_j e_{ij}/\pi_{ry})+ \\ &(1-\exp(-\beta t))\pi_j \end{aligned}$$

Derivatives: There are many cases where one may want to find the maximum likelihood estimate of t. In these cases the following derivative will be useful:

$$\begin{aligned} \frac{d}{dt}P(j|i,t,\pi,R) = &-\delta_{ij}(\alpha_i+\beta)\exp(-(\alpha_i+\beta)t)+ \\ &(\pi_j e_{ij}/\pi_{ry})[-\beta\exp(-\beta t)+(\alpha_i+\beta)\exp(-(\alpha_i+\beta)t)]+ \\ &\pi_j\beta\exp(-\beta t). \end{aligned}$$

The second derivative of the likelihood is

$$\begin{aligned} \frac{d^2}{dt^2}P(j|i,t,\pi,R) = &\delta_{ij}(\alpha_i+\beta)^2\exp(-(\alpha_i+\beta)t)+ \\ &(\pi_j e_{ij}/\pi_{ry})[\beta^2\exp(-\beta t)-(\alpha_i+\beta)^2\exp(-(\alpha_i+\beta)t)]- \\ &\pi_j\beta^2\exp(-\beta t)). \end{aligned}$$

A.1.1 Use of HKY in ML

Within the ML algorithm, we will have to compute the likelihood of a single branch. In addition, we will like to maximize the likelihood of a branch.

We will have a likelihood table with the following form: $f(n,j,k)$, where n is a node, j is a site, and k is a character.

$$\begin{aligned} P(D|root,t) &= \prod_j \sum_k \pi_k f(root,j,k) \\ &= \prod_j \sum_k \pi_k (\sum_x P(x|k,t_a)f(a,j,x))(\sum_y P(y|k,t_b)f(b,j,y)) \end{aligned}$$

The above equation is written specifically for the root node in the tree. Let it have child nodes a and b with branch lengths $t_a = t$ and $t_b = 0$, such that $t_a + t_b = t$. Lets now express the log likelihood

$$\begin{aligned}\log P(D|root,t) &= \sum_j \log(\sum_k \pi_k (\sum_x P(x|k,t) f(a,j,x))(\sum_y P(y|k,0) f(b,j,y))) \\ &= \sum_j \log(\sum_k \pi_k K(b,j,k)(\sum_x P(x|k,t) f(a,j,x))\end{aligned}$$

Where

$$K(b,j,k) = \sum_y P(y|k,0) f(b,j,y).$$

Let

$$g(t,j) = \sum_k \pi_k K(b,j,k)(\sum_x P(x|k,t_a) f(a,j,x)).$$

Then, the derivative of the function is

$$\begin{aligned}\frac{d}{dt}\log P(D|root,t) &= \sum_j \frac{d}{dt}\log(g(t,j)) \\ &= \sum_j \frac{g'(t,j)}{g(t,j)} \\ g'(t,j) &= \sum_k \pi_k K(b,j,k)(\sum_x P'(x|k,t_a) f(a,j,x)).\end{aligned}$$

The second derivative is

$$\begin{aligned}\frac{d^2}{dt^2}\log P(D|root,t) &= \sum_j \frac{d}{dt}\frac{g'(t,j)}{g(t,j)} \\ &= \sum_j -\frac{g'(t,j)^2}{g(t,j)^2} + \frac{g''(t,j)}{g(t,j)} \\ g'(t,j) &= \sum_k \pi_k K(b,j,k)(\sum_x P'(x|k,t_a) f(a,j,x)) \\ g''(t,j) &= \sum_k \pi_k K(b,j,k)(\sum_x P''(x|k,t_a) f(a,j,x)).\end{aligned}$$

A.2 Synteny

I have used synteny in many of my analyses. Here are a few of the algorithms that I use to identify synteny.

A.2.1 BLAST

First I perform BLASTs on peptides between all pairs of species. I usually threshold BLAST hits on an e-value (e.g. $1x10^{-5}$), a minimum alignment length (e.g. 30AA), and a percent identity (e.g. 60%). Next, I discard genes (and all their hits) that are promiscuous: have more than say 10 significant hits.

A.2.2 Fuzzy synteny blocks

Fuzzy synteny gets its name by allowing synteny blocks to skip over genes that appear to participate in another block. This allows one to find syntenic blocks in the presence of segmental or genome-wide duplication. The alternative form of synteny identification is called *strict synteny*.

Two hits are defined to be in a synteny block if their genes (a,b) and (c,d) are within a specified window of base pairs in both genomes. An additional requirement can be made that the hits have the same orientation and are properly ordered. If a gene's strand is specified as either +1 or -1, then the orientation of a hit (a, b) is $a.strand * b.strand$. To test for correct order the following test is performed (Figure A.2). In essence, it allows handles the case where genes overlap.

```
def samedir_hits(hit1, hit2):
    a, b = hit1
    c, d = hit2
    dir1 = a.strand * b.strand
    dir2 = c.strand * d.strand

    # check same orientation
    if dir1 != dir2:
        return False

    # check for proper order
    if dir1 > 0:
        return ((c.end >= a.start and d.end >= b.start) or
                (c.start <= a.end and d.start <= b.end))
    else:
        return ((c.start <= a.end and d.end >= b.start) or
                (c.end >= a.start and d.start <= b.end))
```

Figure A.2: Pseudo-code for determining whether two hits are the same orientation and are properly ordered. These two criterion are used for clustering hits into synteny blocks

After clustering hits into synteny blocks, we often need to filter blocks for those at a significant size. I defined the size of fuzzy synteny block as the number of its associated hits that best bi-direction (BBH). I

typically specify a threshold of (BBH size >= 3).

A.2.3 Orthologous synteny blocks

Orthologous synteny refers to genes in conserved order that are also orthologous. When identifying such blocks, we take additional steps to ensure paralogous synteny blocks and hits (resulting from segmental duplication) are removed.

For example in whole genome duplication, ohnologs (paralogs from WGD) can often be found in conserved gene order. In such cases, a region of genome A will have two (or more) regions in genome B that are syntenic or vice-versa. Two blocks *overlap* if their regions in either genome overlap. An *overlap set* is the single linkage cluster of blocks that overlap in one genome. We find that paralogous blocks tend to have lower block scores than their overlapping orthologous blocks. Thus, we filter out paralogous blocks by only keeping blocks that have the highest score in their overlap set.

Any hits that remain after these filters are called *syntenic orthologs* and they provide a confident gold standard set of orthologs for testing phylogenetic methods.

A.3 Relevant distributions

A.3.1 Exponential distribution

PDF

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & ,x >= 0 \\ 0 & ,x < 0 \end{cases}$$

$$\mu = 1/\lambda$$
$$\sigma^2 = 1/\lambda^2$$

Sampling

$$x \sim \frac{-\ln(Uniform(0,1))}{\lambda}$$

A.3.2 Gamma distribution

PDF

$$f(x;\alpha,\beta) = x^{\alpha-1}\beta^{\alpha}e^{-\beta x}\Gamma(\alpha)^{-1}$$

$$\mu = \alpha/\beta$$
$$\sigma^2 = \alpha/\beta^2$$

Another parameterization Using shape parameter k and scale parameter θ we have

$$f(x;k,\theta) = x^{k-1}\theta^{-k}e^{-x/\theta}\Gamma(k)^{-1}$$

for $x > 0$ and $k, \theta > 0$. To convert between the parameterizations use

$$k = \alpha$$
$$\theta = 1/\beta$$

Summation

$$X_i \sim Gamma(\alpha_i, \beta), \text{indepently distributed}$$
$$\sum_{i=1}^{N} X_i \sim Gamma(\sum_{i=1}^{N} \alpha_i, \beta)$$

Scaling

$$t > 0$$
$$X \sim Gamma(\alpha, \beta)$$
$$tX \sim Gamma(\alpha, \beta/t)$$

Gamma function The definition of the Gamma function is

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

It behaves similar to the factorial for positive integers n:

$$\Gamma(n) = (n-1)!$$

and obeys the following for real and complex numbers:

$$\Gamma(z+1) = z\Gamma(z)$$

Bibliography

[1] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidn-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, and J. C. Venter. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, Mar 2000.

[2] Louigi Addario-Berry, Benny Chor, Mike Hallett, Jens Lagergren, Alessandro Panconesi, and Todd Wareham. Ancestral maximum likelihood of evolutionary trees is hard. *J Bioinform Comput Biol*, 2(2):257–271, Jun 2004.

[3] Orjan Åkerborg, Bengt Sennblad, Lars Arvestad, and Jens Lagergren. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*, 106(14):5714–5719, Apr 2009.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[5] L Arvestad, AC Berglund, J Lagergren, and B Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, pages 326–335, 2004.

[6] L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its associated probabilities. *Journal of the ACM (JACM)*, 56(2):1–44, 2009.

[7] Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, 19 Suppl 1:i7–15, 2003.

[8] N Bailey. *The elements of stochastic processes*. John Wiley & Sons, Inc., New York, 1964.

[9] Stefanie De Bodt, Steven Maere, and Yves Van de Peer. Genome duplication and the origin of angiosperms. *Trends Ecol Evol*, 20(11):591–597, Nov 2005.

[10] Geraldine Butler, Matthew D Rasmussen, Michael F Lin, Manuel A S Santos, Sharadha Sakthikumar, Carol A Munro, Esther Rheinbay, Manfred Grabherr, Anja Forche, Jennifer L Reedy, Ino Agrafioti, Martha B Arnaud, Steven Bates, Alistair J P Brown, Sascha Brunke, Maria C Costanzo, David A Fitzpatrick, Piet W J de Groot, David Harris, Lois L Hoyer, Bernhard Hube, Frans M Klis, Chinnappa Kodira, Nicola Lennard, Mary E Logue, Ronny Martin, Aaron M Neiman, Elissavet Nikolaou, Michael A Quail, Janet Quinn, Maria C Santos, Florian F Schmitzberger, Gavin Sherlock, Prachi Shah, Kevin A T Silverstein, Marek S Skrzypek, David Soll, Rodney Staggs, Ian Stansfield, Michael P H Stumpf, Peter E Sudbery, Thyagarajan Srikantha, Qiandong Zeng, Judith Berman, Matthew Berriman, Joseph Heitman, Neil A R Gow, Michael C Lorenz, Bruce W Birren, Manolis Kellis, and Christina A Cuomo. Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature*, 459(7247):657–662, Jun 2009.

[11] LL Cavalli-Sforza and AWF Edwards. Analysis of human evolution. *Proc. 11th Internat. Congr. Genetics, The Hague*, page 923?933, 1963.

[12] W.C. Chang and O. Eulenstein. Reconciling gene trees with apparent polytomies. *Computing and Combinatorics*, pages 235–244, 2006.

[13] Cedric Chauve, Jean-Philippe Doyon, and Nadia El-Mabrouk. Gene family evolution by duplication, speciation, and loss. *J Comput Biol*, 15(8):1043–1062, Oct 2008.

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

[15] B. Chor and T. Tuller. Maximum likelihood of evolutionary trees is hard. pages 296–310, 2005.

[16] Francesca D Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287, Mar 2006.

[17] Andrew G Clark, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, Therese A Markow, Thomas C Kaufman, Manolis Kellis, William Gelbart, Venky N Iyer, Daniel A Pollard, Timothy B Sackton, Amanda M Larracuente, Nadia D Singh, Jose P Abad, Dawn N Abt, Boris Adryan, Montserrat Aguade, Hiroshi Akashi, Wyatt W Anderson, Charles F Aquadro, David H Ardell, Roman Arguello, Carlo G Artieri, Daniel A Barbash, Daniel Barker, Paolo Barsanti, Phil

Batterham, Serafim Batzoglou, Dave Begun, Arjun Bhutkar, Enrico Blanco, Stephanie A Bosak, Robert K Bradley, Adrianne D Brand, Michael R Brent, Angela N Brooks, Randall H Brown, Roger K Butlin, Corrado Caggese, Brian R Calvi, A. Bernardo de Carvalho, Anat Caspi, Sergio Castrezana, Susan E Celniker, Jean L Chang, Charles Chapple, Sourav Chatterji, Asif Chinwalla, Alberto Civetta, Sandra W Clifton, Josep M Comeron, James C Costello, Jerry A Coyne, Jennifer Daub, Robert G David, Arthur L Delcher, Kim Delehaunty, Chuong B Do, Heather Ebling, Kevin Edwards, Thomas Eickbush, Jay D Evans, Alan Filipski, Sven Findeiss, Eva Freyhult, Lucinda Fulton, Robert Fulton, Ana C L Garcia, Anastasia Gardiner, David A Garfield, Barry E Garvin, Greg Gibson, Don Gilbert, Sante Gnerre, Jennifer Godfrey, Robert Good, Valer Gotea, Brenton Gravely, Anthony J Greenberg, Sam Griffiths-Jones, Samuel Gross, Roderic Guigo, Erik A Gustafson, Wilfried Haerty, Matthew W Hahn, Daniel L Halligan, Aaron L Halpern, Gillian M Halter, Mira V Han, Andreas Heger, LaDeana Hillier, Angie S Hinrichs, Ian Holmes, Roger A Hoskins, Melissa J Hubisz, Dan Hultmark, Melanie A Huntley, David B Jaffe, Santosh Jagadeeshan, William R Jeck, Justin Johnson, Corbin D Jones, William C Jordan, Gary H Karpen, Eiko Kataoka, Peter D Keightley, Pouya Kheradpour, Ewen F Kirkness, Leonardo B Koerich, Karsten Kristiansen, Dave Kudrna, Rob J Kulathinal, Sudhir Kumar, Roberta Kwok, Eric Lander, Charles H Langley, Richard Lapoint, Brian P Lazzaro, So-Jeong Lee, Lisa Levesque, Ruiqiang Li, Chiao-Feng Lin, Michael F Lin, Kerstin Lindblad-Toh, Ana Llopart, Manyuan Long, Lloyd Low, Elena Lozovsky, Jian Lu, Meizhong Luo, Carlos A Machado, Wojciech Makalowski, Mar Marzo, Muneo Matsuda, Luciano Matzkin, Bryant McAllister, Carolyn S McBride, Brendan McKernan, Kevin McKernan, Maria Mendez-Lago, Patrick Minx, Michael U Mollenhauer, Kristi Montooth, Stephen M Mount, Xu Mu, Eugene Myers, Barbara Negre, Stuart Newfeld, Rasmus Nielsen, Mohamed A F Noor, Patrick O'Grady, Lior Pachter, Montserrat Papaceit, Matthew J Parisi, Michael Parisi, Leopold Parts, Jakob S Pedersen, Graziano Pesole, Adam M Phillippy, Chris P Ponting, Mihai Pop, Damiano Porcelli, Jeffrey R Powell, Sonja Prohaska, Kim Pruitt, Marta Puig, Hadi Quesneville, Kristipati Ravi Ram, David Rand, Matthew D Rasmussen, Laura K Reed, Robert Reenan, Amy Reily, Karin A Remington, Tania T Rieger, Michael G Ritchie, Charles Robin, Yu-Hui Rogers, Claudia Rohde, Julio Rozas, Marc J Rubenfield, Alfredo Ruiz, Susan Russo, Steven L Salzberg, Alejandro Sanchez-Gracia, David J Saranga, Hajime Sato, Stephen W Schaeffer, Michael C Schatz, Todd Schlenke, Russell Schwartz, Carmen Segarra, Rama S Singh, Laura Sirot, Marina Sirota, Nicholas B Sisneros, Chris D Smith, Temple F Smith, John Spieth, Deborah E Stage, Alexander Stark, Wolfgang Stephan, Robert L Strausberg, Sebastian Strempel, David Sturgill, Granger Sutton, Granger G Sutton, Wei Tao, Sarah Teichmann, Yoshiko N Tobari, Yoshihiko Tomimura, Jason M Tsolas, Vera L S Valente, Eli Venter, J. Craig Venter, Saverio Vicario, Filipe G Vieira, Albert J Vilella, Alfredo Villasante, Brian Walenz, Jun Wang, Marvin Wasserman, Thomas Watts, Derek Wilson, Richard K Wilson, Rod A Wing, Mariana F Wolfner, Alex Wong, Gane Ka-Shu Wong, Chung-I. Wu, Gabriel Wu, Daisuke Yamamoto, Hsiao-Pei Yang, Shiaw-Pyng Yang, James A Yorke, Kiyohito Yoshida, Evgeny Zdobnov, Peili Zhang, Yu Zhang, Aleksey V Zimin, Jennifer Baldwin, Amr Abdouelleil, Jamal Abdulkadir, Adal Abebe, Brikti Abera, Justin Abreu, St Christophe Acer, Lynne Aftuck, Allen Alexander, Peter An, Erica Anderson, Scott Anderson, Harindra Arachi, Marc Azer, Pasang Bachantsang, Andrew Barry, Tashi Bayul, Aaron Berlin, Daniel Bessette, Toby Bloom, Jason Blye, Leonid Boguslavskiy, Claude Bonnet, Boris Boukhgalter, Imane Bourzgui, Adam Brown, Patrick Cahill, Sheridon Channer, Yama Cheshatsang, Lisa Chuda, Mieke Citroen, Alville Collymore, Patrick Cooke, Maura Costello, Katie D'Aco, Riza Daza, Georgius De Haan, Stuart DeGray, Christina DeMaso, Norbu Dhargay, Kimberly Dooley, Erin Dooley, Missole Doricent, Passang Dorje, Kunsang Dorjee, Alan Dupes, Richard Elong, Jill Falk, Abderrahim Farina, Susan Faro, Diallo Ferguson, Sheila Fisher, Chelsea D Foley, Alicia Franke, Dennis Friedrich, Loryn Gadbois, Gary Gearin, Christina R Gearin, Georgia Giannoukos, Tina Goode, Joseph Graham,

Edward Grandbois, Sharleen Grewal, Kunsang Gyaltsen, Nabil Hafez, Birhane Hagos, Jennifer Hall, Charlotte Henson, Andrew Hollinger, Tracey Honan, . Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203–218, Nov 2007.

[18] Paul Cliften, Priya Sudarsanam, Ashwin Desikan, Lucinda Fulton, Bob Fulton, John Majors, Robert Waterston, Barak A Cohen, and Mark Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, Jul 2003.

[19] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, Apr 2003.

[20] Maria C Costanzo, Martha B Arnaud, Marek S Skrzypek, Gail Binkley, Christopher Lane, Stuart R Miyasato, and Gavin Sherlock. The candida genome database: facilitating research on candida albicans molecular biology. *FEMS Yeast Res*, 6(5):671–684, Aug 2006.

[21] C. J. Creevey and J. O. McInerney. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, 21(3):390–392, Feb 2005.

[22] Ruchira S Datta, Christopher Meacham, Bushra Samad, Christoph Neyer, and Kimmen Sjlander. Berkeley phog: Phylofacts orthology group prediction web server. *Nucleic Acids Res*, 37(Web Server issue):W84–W89, Jul 2009.

[23] W.H.E. Day, D.S. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical biosciences*, 81(33-42):299, 1986.

[24] W.H.E. Day and D. Sankoff. Computational complexity of inferring phylogenies by compatibility. *Systematic Zoology*, 35(2):224–229, 1986.

[25] James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*, 24(6):332–340, Jun 2009.

[26] Paramvir Dehal and Jeffrey L Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*, 3(10):e314, Oct 2005.

[27] Paramvir Dehal, Yutaka Satou, Robert K Campbell, Jarrod Chapman, Bernard Degnan, Anthony De Tomaso, Brad Davidson, Anna Di Gregorio, Maarten Gelpke, David M Goodstein, Naoe Harafuji, Kenneth E M Hastings, Isaac Ho, Kohji Hotta, Wayne Huang, Takeshi Kawashima, Patrick Lemaire, Diego Martinez, Ian A Meinertzhagen, Simona Necula, Masaru Nonaka, Nik Putnam, Sam Rash, Hidetoshi Saiga, Masanobu Satake, Astrid Terry, Lixy Yamada, Hong-Gang Wang, Satoko Awazu, Kaoru Azumi, Jeffrey Boore, Margherita Branno, Stephen Chin-Bow, Rosaria DeSantis, Sharon Doyle, Pilar Francino, David N Keys, Shinobu Haga, Hiroko Hayashi, Kyosuke Hino, Kaoru S Imai, Kazuo Inaba, Shungo Kano, Kenji Kobayashi, Mari Kobayashi, Byung-In Lee, Kazuhiro W Makabe, Chitra Manohar, Giorgio Matassi, Monica Medina, Yasuaki Mochizuki, Steve Mount, Tomomi Morishita, Sachiko Miura, Akie Nakayama, Satoko Nishizaka, Hisayo Nomoto, Fumiko Ohta, Kazuko Oishi, Isidore Rigoutsos, Masako Sano, Akane Sasaki, Yasunori Sasakura, Eiichi Shoguchi, Tadasu Shin-i, Antoinetta Spagnuolo, Didier Stainier, Miho M Suzuki, Olivier Tassy, Naohito Takatori, Miki Tokuoka, Kasumi Yagi, Fumiko Yoshizaki, Shuichi Wada, Cindy Zhang, P Douglas Hyatt, Frank Larimer, Chris Detter, Norman Doggett, Tijana Glavina, Trevor Hawkins, Paul Richardson, Susan Lucas, Yuji Kohara, Michael Levine, Nori Satoh, and Daniel S Rokhsar. The draft genome of ciona intestinalis: insights into chordate and vertebrate origins. *Science*, 298(5601):2157–2167, December 2002.

[28] Paramvir S Dehal and Jeffrey L Boore. A phylogenomic gene cluster resource: the phylogenetically inferred groups (phigs) database. *BMC Bioinformatics*, 7:201, 2006.

[29] Jeffery P Demuth and Matthew W Hahn. The life and death of gene families. *Bioessays*, 31(1):29–39, Jan 2009.

[30] JP Demuth, TD Bie, JE Stajich, N Cristianini, and MW Hahn. The evolution of mammalian gene families. *PLoS ONE*, 1(1):e85, 2006.

[31] Fred S Dietrich, Sylvia Voegeli, Sophie Brachat, Anita Lerch, Krista Gates, Sabine Steiner, Christine Mohr, Rainer Phlmann, Philippe Luedi, Sangdun Choi, Rod A Wing, Albert Flavier, Thomas D Gaffney, and Peter Philippsen. The ashbya gossypii genome as a tool for mapping the ancient saccharomyces cerevisiae genome. *Science*, 304(5668):304–307, Apr 2004.

[32] Jean-Philippe Doyon, Cedric Chauve, and Sylvie Hamel. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol*, 16(10):1399–1418, Oct 2009.

[33] Jean-Philippe Doyon, Sylvie Hamel, and Cedric Chauve. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. 2010.

[34] Jean-François Dufayard, Laurent Duret, Simon Penel, Manolo Gouy, François Rechenmann, and Guy Perrière. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603, Jun 2005.

[35] Bernard Dujon, David Sherman, Gilles Fischer, Pascal Durrens, Serge Casaregola, Ingrid Lafontaine, Jacky De Montigny, Christian Marck, Ccile Neuvglise, Emmanuel Talla, Nicolas Goffard, Lionel Frangeul, Michel Aigle, Vronique Anthouard, Anna Babour, Valrie Barbe, Stphanie Barnay, Sylvie Blanchin, Jean-Marie Beckerich, Emmanuelle Beyne, Claudine Bleykasten, Anita Boisram, Jeanne Boyer, Laurence Cattolico, Fabrice Confanioleri, Antoine De Daruvar, Laurence Despons, Emmanuelle Fabre, Ccile Fairhead, Hlne Ferry-Dumazet, Alexis Groppi, Florence Hantraye, Christophe Hennequin, Nicolas Jauniaux, Philippe Joyet, Rym Kachouri, Alix Kerrest, Romain Koszul, Marc Lemaire, Isabelle Lesur, Laurence Ma, Hlose Muller, Jean-Marc Nicaud, Macha Nikolski, Sophie Oztas, Odile Ozier-Kalogeropoulos, Stefan Pellenz, Serge Potier, Guy-Franck Richard, Marie-Laure Straub, Audrey Suleau, Dominique Swennen, Fredj Tekaia, Micheline Wsolowski-Louvel, Eric Westhof, Bndicte Wirth, Maria Zeniou-Meyer, Ivan Zivanovic, Monique Bolotin-Fukuhara, Agns Thierry, Christiane Bouchier, Bernard Caudron, Claude Scarpelli, Claude Gaillardin, Jean Weissenbach, Patrick Wincker, and Jean-Luc Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, Jul 2004.

[36] Dannie Durand, Bjarni V Halldorsson, and Benjamin Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol*, 13(2):320–335, Mar 2006.

[37] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[38] Julien Y Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K Uyenoyama, and Mikkel H Schierup. Ancestral population genomics: the coalescent hidden markov model approach. *Genetics*, 183(1):259–274, Sep 2009.

[39] S. R. Eddy. Hmmer: profile hidden markov models for biological sequence analysis., 2000.

[40] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797, 2004.

[41] J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, Mar 1998.

[42] Jonathan A Eisen and Claire M Fraser. Phylogenomics: intersection of evolution and genomics. *Science*, 300(5626):1706–1707, Jun 2003.

[43] Isaac Elias and Jens Lagergren. Fast neighbor joining. In *Proc. of the 32nd International Colloquium on Automata, Languages and Programming (ICALP'05)*, volume 3580 of *Lecture Notes in Computer Science*, pages 1263–1274. Springer-Verlag, July 2005.

[44] O. Eulenstein and G.M.D.F. Informationstechnik. A linear time algorithm for tree mapping. *Arbeitspapiere der GMD*, 1046, 1997.

[45] W. Feller. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1):11–40, 1939.

[46] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[47] J Felsenstein. Phylip (phylogeny inference package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington*, 2005.

[48] J. Felsenstein and G. A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*, 13(1):93–104, Jan 1996.

[49] W.M. Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19(2):99–113, 1970.

[50] W.M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool*, 20(4), 1971.

[51] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, Apr 1999.

[52] Li-zhi Gao and Hideki Innan. Very low gene duplication rate in the yeast genome. *Science*, 306(5700):1367–1370, 2004.

[53] Olivier Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*, 14(7):685–695, Jul 1997.

[54] Y. Gilad, V. Wiebe, M. Przeworski, D. Lancet, and S. P
"a
"abo. Correction: Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biology*, 5(6), 2007.

[55] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546, 563–546, 567, Oct 1996.

[56] Ana Gomes, Isabel Miranda, Raquel Silva, Gabriela Moura, Benjamin Thomas, Alexandre Akoulitchev, and Manuel Santos. A genetic code alteration generates a proteome of high diversity in the human pathogen candida albicans. *Genome Biol*, 8(10):R206, Oct 2007.

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

[58] P. Górecki and J. Tiuryn. Dls-trees: a model of evolutionary scenarios. *Theoretical Computer Science*, 359(1-3):378–399, 2006.

[59] Pawel Górecki. Reconciliation problems for duplication, loss and horizontal gene transfer. pages 316–325, 2004.

[60] Xun Gu and Hongmei Zhang. Genome phylogenetic analysis based on extended gene contents. *Mol Biol Evol*, 21(7):1401–1408, Jul 2004.

[61] Zhenglong Gu, Andre Cavalcanti, Feng-Chi Chen, Peter Bouman, and Wen-Hsiung Li. Extent of gene duplication in the genomes of drosophila, nematode, and yeast. *Mol Biol Evol*, 19(3):256–262, Mar 2002.

[62] Stphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, Oct 2003.

[63] Matthew Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol*, 8(7):R141, Jul 2007.

[64] Matthew W. Hahn, Tijl De Bie, Jason E. Stajich, Chi Nguyen, and Nello Cristianini. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.*, 15(8):1153–1160, 2005.

[65] Matthew W Hahn, Jeffery P Demuth, and Sang-Gook Han. Accelerated rate of gene gain and loss in primates. *Genetics*, 177(3):1941–1949, Nov 2007.

[66] Matthew W Hahn, Mira V Han, and Sang-Gook Han. Gene family evolution across 12 drosophila genomes. *PLoS Genet*, 3(11):e197, Nov 2007.

[67] Mike Hallett, Jens Lagergren, and Ali Tofigh. Simultaneous identification of duplications and lateral transfers. pages 347–356, 2004.

[68] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160–174, 1985.

[69] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, 3(2):e7, Feb 2007.

[70] Simone Hoegg and Axel Meyer. Hox clusters as models for vertebrate genome evolution. *Trends Genet*, 21(8):421–424, Aug 2005.

[71] Jaime Huerta-Cepas, Hernan Dopazo, Joaquin Dopazo, and Toni Gabaldon. The human phylome. *Genome Biol*, 8(6):R109, Jun 2007.

[72] G. H. Jacobs, M. Neitz, J. F. Deegan, and J. Neitz. Trichromatic colour vision in new world monkeys. *Nature*, 382(6587):156–158, Jul 1996.

[73] Olivier Jaillon, Jean-Marc Aury, Frdric Brunet, Jean-Louis Petit, Nicole Stange-Thomann, Evan Mauceli, Laurence Bouneau, Ccile Fischer, Catherine Ozouf-Costaz, Alain Bernot, Sophie Nicaud, David Jaffe, Sheila Fisher, Georges Lutfalla, Carole Dossat, Batrice Segurens, Corinne Dasilva, Marcel Salanoubat, Michael Levy, Nathalie Boudet, Sergi Castellano, Vronique Anthouard, Claire Jubin, Vanina Castelli, Michael Katinka, Benot Vacherie, Christian Bimont, Zineb Skalli, Laurence Cattolico, Julie Poulain, Vronique De Berardinis, Corinne Cruaud, Simone Duprat, Philippe Brottier, Jean-Pierre Coutanceau, Jrme Gouzy, Genis Parra, Guillaume Lardier, Charles Chapple, Kevin J McKernan, Paul McEwan, Stephanie Bosak, Manolis Kellis, Jean-Nicolas Volff, Roderic Guig, Michael C Zody, Jill Mesirov, Kerstin Lindblad-Toh, Bruce Birren, Chad Nusbaum, Daniel Kahn, Marc Robinson-Rechavi, Vincent Laudet, Vincent Schachter, Francis Qutier, William Saurin, Claude Scarpelli, Patrick Wincker, Eric S Lander, Jean Weissenbach, and Hugues Roest Crollius. Genome duplication in the teleost fish tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, Oct 2004.

[74] Ted Jones, Nancy A Federspiel, Hiroji Chibana, Jan Dungan, Sue Kalman, B. B. Magee, George Newport, Yvonne R Thorstenson, Nina Agabian, P. T. Magee, Ronald W Davis, and Stewart Scherer. The diploid genome sequence of candida albicans. *Proc Natl Acad Sci U S A*, 101(19):7329–7334, May 2004.

[75] TH Jukes and CR Cantor. Evolution of protein molecules: Pp. 21–132 in mammalian protein metabolism (hn munro, ed.). *Academic Press, New York*, 3:1–571, 1969.

[76] Mahir Karababa, Alix T Coste, Bndicte Rognon, Jacques Bille, and Dominique Sanglard. Comparison of gene expression profiles of candida albicans azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. *Antimicrob Agents Chemother*, 48(8):3064–3079, Aug 2004.

[77] Masanori Kasahara. The 2r hypothesis: an update. *Curr Opin Immunol*, 19(5):547–552, Oct 2007.

[78] Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, Apr 2004.

[79] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, May 2003.

[80] David G. Kendall. Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282, 1949.

[81] Su Yeon Kim and Jonathan K Pritchard. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet*, 3(9):e147, Sep 2007.

[82] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.

[83] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[84] E. V. Koonin, K. S. Makarova, and L. Aravind. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, 2001.

[85] L. B. Koski and G. B. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Evol*, 52(6):540–542, Jun 2001.

[86] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, 11(3):459–468, May 1994.

[87] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):86–93, 1984.

[88] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[89] Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim Heriche, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, and Richard Durbin. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34(Database issue):D572–D580, Jan 2006.

[90] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003.

[91] Liang Liu and Dennis K Pearl. Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol*, 56(3):504–514, Jun 2007.

[92] Teresa T Liu, Robin E B Lee, Katherine S Barker, Richard E Lee, Lai Wei, Ramin Homayouni, and P. David Rogers. Genome-wide expression profiling of the response to azole, polyene, echinocandin, and pyrimidine antifungal agents in candida albicans. *Antimicrob Agents Chemother*, 49(6):2226–2236, Jun 2005.

[93] B. Llorente, P. Durrens, A. Malpertuy, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola, B. Dujon, J. de Montigny, A. Lpingle, C. Neuvglise, O. Ozier-Kalogeropoulos, S. Potier, W. Saurin, F. Tekaia, C. Toffano-Nioche, M. Wsolowski-Louvel, P. Wincker, J. Weissenbach, J. Souciet, and C. Gaillardin. Genomic exploration of the hemiascomycetous yeasts: 20. evolution of gene redundancy compared to saccharomyces cerevisiae. *FEBS Lett*, 487(1):122–133, Dec 2000.

[94] B. Llorente, A. Malpertuy, C. Neuvglise, J. de Montigny, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola, P. Durrens, C. Gaillardin, A. Lpingle, O. Ozier-Kalogropoulos, S. Potier, W. Saurin, F. Tekaia, C. Toffano-Nioche, M. Wsolowski-Louvel, P. Wincker, J. Weissenbach, J. Souciet, and B. Dujon. Genomic exploration of the hemiascomycetous yeasts: 18. comparative analysis of chromosome maps and synteny with saccharomyces cerevisiae. *FEBS Lett*, 487(1):101–112, Dec 2000.

[95] L. G. Lundin. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. *Genomics*, 16(1):1–19, Apr 1993.

[96] Michael Lynch and John S Conery. The evolutionary demography of duplicate genes. *J Struct Funct Genomics*, 3(1-4):35–44, 2003.

[97] Michael Lynch, Martin O'Hely, Bruce Walsh, and Allan Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001.

[98] B Ma, HK Kowloon, M Li, O Waterloo, L Zhang, and BI Center. From gene trees to species trees. *SIAM J Comput*, 30(3):729–752, Jul 2000.

[99] Wayne P Maddison and L. Lacey Knowles. Inferring phylogeny despite incomplete lineage sorting. *Syst Biol*, 55(1):21–30, Feb 2006.

[100] Véronique Marchais, Marie Kempf, Patricia Licznar, Corinne Lefrançois, Jean-Philippe Bouchara, Raymond Robert, and Jane Cottin. Dna array analysis of candida albicans gene expression in response to adherence to polystyrene. *FEMS Microbiol Lett*, 245(1):25–32, Apr 2005.

[101] Steven E Massey, Gabriela Moura, Pedro Beltro, Ricardo Almeida, James R Garey, Mick F Tuite, and Manuel A S Santos. Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the ctg codon in candida spp. *Genome Res*, 13(4):544–557, Apr 2003.

[102] Atsushi Matsui, Yasuhiro Go, and Yoshihito Niimura. Degeneration of olfactory receptor gene repertories in primates: No direct link to full trichromatic vision. *Mol Biol Evol*, Jan 2010.

[103] V. W. Mayer and A. Aguilera. High levels of chromosome instability in polyploids of saccharomyces cerevisiae. *Mutat Res*, 231(2):177–186, Aug 1990.

[104] Michael R McGowen, Clay Clark, and John Gatesy. The vestigial olfactory receptor subgenome of odontocete whales: phylogenetic congruence between gene-tree reconciliation and supermatrix methods. *Syst Biol*, 57(4):574–590, Aug 2008.

[105] Y. Miyazaki, A. Geber, H. Miyazaki, D. Falconer, T. Parkinson, C. Hitchcock, B. Grimberg, K. Nyswaner, and J. E. Bennett. Cloning, sequencing, expression and allelic sequence diversity of erg3 (c-5 sterol desaturase gene) in candida albicans. *Gene*, 236(1):43–51, Aug 1999.

[106] P.G. Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985.

[107] H. J. Muller. Bar duplication. *Nature*, 83:528–530, 1936.

[108] J. Nathans, D. Thomas, and D. S. Hogness. Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science*, 232(4747):193–202, Apr 1986.

[109] Yoshihito Niimura and Masatoshi Nei. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One*, 2(1):e708, 2007.

[110] James P Noonan, Jane Grimwood, Jeremy Schmutz, Mark Dickson, and Richard M Myers. Gene conversion and the evolution of protocadherin gene cluster diversity. *Genome Res*, 14(3):354–366, Mar 2004.

[111] Genome 10K Community of Scientists. Genome 10k: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered*, 100(6):659–674, 2009.

[112] S Ohno. *Evolution by Gene Duplication.* Springer-Verlag New York, 1970.

[113] R. D. Page and M. A. Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, 7(2):231–240, Apr 1997.

[114] RDM Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43(1):58–77, 1994.

[115] M. A. Pfaller and D. J. Diekema. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev*, 20(1):133–163, Jan 2007.

[116] Herve Philippe, Yan Zhou, Henner Brinkmann, Nicolas Rodrigue, and Frederic Delsuc. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(1):50, 2005.

[117] 1000 Genomes Project. 1000 genomes project, 2010.

[118] Nicholas H Putnam, Thomas Butts, David E K Ferrier, Rebecca F Furlong, Uffe Hellsten, Takeshi Kawashima, Marc Robinson-Rechavi, Eiichi Shoguchi, Astrid Terry, Jr-Kai Yu, E. Lia Benito-Gutirrez, Inna Dubchak, Jordi Garcia-Fernndez, Jeremy J Gibson-Brown, Igor V Grigoriev, Amy C Horton, Pieter J de Jong, Jerzy Jurka, Vladimir V Kapitonov, Yuji Kohara, Yoko Kuroki, Erika Lindquist, Susan Lucas, Kazutoyo Osoegawa, Len A Pennacchio, Asaf A Salamov, Yutaka Satou, Tatjana Sauka-Spengler, Jeremy Schmutz, Tadasu Shin-I, Atsushi Toyoda, Marianne Bronner-Fraser, Asao Fujiyama, Linda Z Holland, Peter W H Holland, Nori Satoh, and Daniel S Rokhsar. The amphioxus genome and the evolution of the chordate karyotype. *Nature*, 453(7198):1064–1071, Jun 2008.

[119] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, 43(3):304–311, Sep 1996.

[120] Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656, Aug 2003.

[121] Matthew D. Rasmussen and Manolis Kellis. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.*, 17:1932–1942, 2007.

[122] M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 314(5):1041–1052, Dec 2001.

[123] Stephen Richards, Yue Liu, Brian R Bettencourt, Pavel Hradecky, Stan Letovsky, Rasmus Nielsen, Kevin Thornton, Melissa J Hubisz, Rui Chen, Richard P Meisel, Olivier Couronne, Sujun Hua, Mark A Smith, Peili Zhang, Jing Liu, Harmen J Bussemaker, Marinus F van Batenburg, Sally L Howells, Steven E Scherer, Erica Sodergren, Beverly B Matthews, Madeline A Crosby, Andrew J Schroeder, Daniel Ortiz-Barrientos, Catharine M Rives, Michael L Metzker, Donna M Muzny, Graham Scott, David Steffen, David A Wheeler, Kim C Worley, Paul Havlak, K. James Durbin, Amy Egan, Rachel Gill, Jennifer Hume, Margaret B Morgan, George Miner, Cerissa Hamilton, Yanmei Huang, Lene Waldron, Daniel Verduzco, Kerstin P Clerc-Blankenburg, Inna Dubchak, Mohamed A F Noor, Wyatt Anderson, Kevin P White, Andrew G Clark, Stephen W Schaeffer, William Gelbart, George M Weinstock, and Richard A Gibbs. Comparative genome sequencing of drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res*, 15(1):1–18, Jan 2005.

[124] Antonis Rokas, Barry L Williams, Nicole King, and Sean B Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003.

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

[126] N Saitou and T Imanishi. Relative efficiencies of the fitch-margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol Biol Evol*, 6(5):514–525, 1989.

[127] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, Jul 1987.

[128] Michael J Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, Jan 2003.

[129] Devin R Scannell, Kevin P Byrne, Jonathan L Gordon, Simon Wong, and Kenneth H Wolfe. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440(7082):341–345, Mar 2006.

[130] Devin R. Scannell and Kenneth H. Wolfe. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.*, page gr.6341207, 2007.

[131] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*, 16(8):1114–1116, 1999.

[132] Adam Siepel. Phylogenomics of primates and their ancestral populations. *Genome Res*, 19(11):1929–1941, Nov 2009.

[133] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, Aug 2005.

[134] M. Simonsen, T. Mailund, and C. Pedersen. Rapid neighbour-joining. *Algorithms in Bioinformatics*, pages 113–122, 2008.

[135] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of saitou and nei. *Mol Biol Evol*, 5(6):729–731, Nov 1988.

[136] Koichiro Tamura, Sankar Subramanian, and Sudhir Kumar. Temporal patterns of fruit fly (drosophila) evolution revealed by mutation clocks. *Mol Biol Evol*, 21(1):36–44, Jan 2004.

[137] Y. Tateno, N. Takezaki, and M. Nei. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol Biol Evol*, 11(2):261–277, Mar 1994.

[138] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B. Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren A Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sep 2003.

[139] Akihisa Terakita. The opsins. *Genome Biol*, 6(3):213, 2005.

[140] C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull Math Biol*, 59(3):581–607, May 1997.

[141] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam,

J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guig, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, Feb 2001.

[142] Benjamin Vernot, Maureen Stolzer, Aiton Goldman, and Dannie Durand. Reconciliation with non-binary species trees. *J Comput Biol*, 15(8):981–1006, Oct 2008.

[143] Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, 19(2):327–335, Feb 2009.

[144] J. Wakeley. *Coalescent theory: an introduction.* Roberts & Company Publishers, 2009.

[145] Ilan Wapinski, Avi Pfeffer, Nir Friedman, and Aviv Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449(7158):54–61, September 2007.

[146] Ilan Wapinski, Avi Pfeffer, Nir Friedman, and Aviv Regev. Synergy dataset january 2009 update, 2009.

[147] MS Waterman and TF Smith. On the similarity of dendrograms. *Journal of Theoretical Biology*, 73(789-800):732, 1978.

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

[149] Ziheng Yang and Rasmus Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917, Jun 2002.

[150] C. M. Zmasek and S. R. Eddy. A simple algorithm to infer gene duplication and speciation events on a gene tree. In *Bioinformatics* [114], pages 821–828.

[151] Christian M Zmasek and Sean R Eddy. Rio: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3:14, May 2002.

[152] E Zuckerkandl and L Pauling. *Horizons in Biochemistry.* Academic Press, New York, 1962.