

# 1

## An overview of phylogeny reconstruction

---

1.1	Introduction.....	1-1
1.2	What are phylogenies used for in biological research? .....	1-2
1.3	The steps of a phylogenetic analysis.....	1-4
	Designing the study • Collecting organisms in the field • In the lab • Multiple Sequence Alignment • Phylogenetic reconstruction • Support assessment	
1.4	Research problems in molecular phylogenetics ....	1-18
	Performance analysis of algorithms • Phylogenetic reconstruction on molecular sequences • Multiple Sequence Alignment (MSA) • Special challenges involved in large-scale phylogenetics	
1.5	Special topic: Supertree methods.....	1-28
	Introduction • Tree Compatibility • Matrix Representation Parsimony • Other supertree methods • Open problems	
1.6	Special topic: Genomic phylogeny reconstruction	1-30
1.7	Conclusions and suggestions for further reading ..	1-32
1.8	Acknowledgments .....	1-33

C. Randal Linder  
*The University of Texas at Austin*

Tandy Warnow  
*The University of Texas at Austin*

### 1.1 Introduction

---

The best evidence strongly supports that all life currently on earth is descended from a single common ancestor. Over a period of at least 3.8 billion years, that single original ancestor has split repeatedly into new and independent lineages (i.e., species), and, on occasion, some of these otherwise independent lineages have come back together to form yet other lineages or to exchange genetic information. The evolutionary relationships among these species is referred to as their “phylogeny”, and phylogenetic reconstruction is concerned with inferring the phylogeny of groups of organisms. The ultimate goal is to infer the phylogeny of life on earth.

Accurate phylogenetic reconstructions involve significant effort due to the difficulties of acquiring the primary biological data and the computational complexity of the underlying optimization problems. Not surprisingly, phylogenetic inference is providing interesting and hard problems to the computer science algorithms research community – as witnessed by the three chapters in this volume on novel algorithmic research for phylogeny reconstruction. The limitations of existing phylogenetic reconstruction methods have a direct impact on the ability of systematists (that is, biologists who study the evolutionary history of a group

of organisms) to analyze their data with adequate accuracy and efficiency, so that their subsequent scientific inferences are reliable.

The purpose of this chapter is to help computer scientists develop sufficient knowledge and taste in the area of computational phylogenetics, so that they will be able to develop new methods for phylogeny reconstruction that can help the practicing molecular systematist. Much of our discussion will therefore be from the viewpoint of a molecular systematist, and will (a) elucidate promising areas for additional research, (b) provide a context with which to understand the potential for impact of algorithmic innovations, and (c) give the algorithms research community a better understanding of the strengths and limitations of data collection and analysis in current practice. The final point is important because a better understanding of the type and amount of raw data that biologists can routinely obtain and analyze will help mathematicians and computer scientists in designing better methods. In particular, we will draw attention to the sources of potential error in the primary data, and to those aspects of the input data that have the potential to impact reconstruction methods in significant, and potentially different, ways.

Phylogenies and phylogenetic reconstruction are important to biology in many ways. So much so, that phylogenies have become an integral part of much biological research, including biomedical research, drug design, and other areas of bioinformatics, including protein structure prediction and multiple sequence alignment. Understanding the use of phylogenies for these various applications will help the algorithms designer appreciate where errors in phylogeny estimation can be tolerated, and where they will have a more serious impact.

We therefore begin our discussion in Section 1.2 with an overview of how phylogenies are used in biology, focusing on the questions that can be answered once the phylogeny is obtained. We continue in Section 1.3 with a description of the process a biologist goes through, from the inception of a project to the production of a publishable phylogenetic inference. In Section 1.4 we review each of the steps involved in phylogenetic inference, and discuss their major methodological and algorithmic issues. In Section 1.5 and Section 1.6, we describe advances on two specialized research problems – supertree methods (the subject of one of the chapters in this volume) and gene order phylogenetics (which is discussed in another chapter), respectively. We close in Section 1.7 with some comments about algorithmic research in phylogeny and recommendations for additional reading.

Finally, a caveat. Phylogenies are generally represented as rooted binary trees since speciation events are generally bifurcating, i.e., speciation usually occurs when an ancestral lineage splits into two new, independent lineages. The assumption is that reticulation, i.e., horizontal gene transfer and hybrid speciation, is rare. Nonetheless, reticulation events are known to occur and are more common in certain groups of organisms, e.g., hybrid speciation is relatively common in plants [40]. Therefore, the evolutionary history of all life is not properly represented as a tree. Instead the appropriate graphical model of evolution for all life is really, a directed acyclic graph (DAG) which we call a “phylogenetic network” [37, 36]. Despite the reality of reticulation, in order to keep the chapter to a reasonable size, we will confine most of our discussion to phylogenetic trees and their reconstruction and will only talk about reticulate evolution insofar as it creates difficulties for tree reconstruction. Readers who want to learn more about reticulate phylogenies may wish to read the tutorial on reticulate evolution at the DIMACS web site for reticulate evolution [31].

## 1.2 What are phylogenies used for in biological research?

---

Phylogenies are reconstructed on the basis of character data, where a “character” is any feature of an organism that can have different states. A typical biological example of a character is a nucleotide position in a DNA sequence, with the character state being the particular nucleotide (A,G,C,T) occupying that position. From a mathematical standpoint, a character is just a function that maps the set of taxa to its set of states. When the set of states is discrete, the character is said to be “qualitative” or “discrete”, and when the set of states is continuous, then the character is said to be “quantitative”. Molecular phylogenetics research is concerned not only with the evolutionary history of different organisms, but also with how the different characters evolve in the course of that history. Thus, characters are used to infer species trees, but can also be of interest in their own right.

The uses for phylogenies, beyond elucidating the evolutionary relationships of biological species, are many and growing; here we highlight the most common uses and some of the most intriguing.

The most common use of a phylogeny is for a comparative study [4, 25]. A comparative study is one where a particular question is addressed by comparing how certain biological characters have evolved in different lineages in the context of a phylogeny. This information is used to infer important aspects of the evolution of those characters. That statement is vague and general because of the many types of characters studied using a comparative approach. Some examples of areas in which the comparative method can be applied includes adaptation, development, physiology, gene function, vaccine design, and modes of speciation. In essence, a comparative approach can be taken any time a biologist wishes to examine a process or aspect of biological organisms that has evolved on a time scale that is greater than the time of an individual species or lineage. An evolutionary perspective, through the use of an accurately estimated phylogeny, makes it possible to ensure that the number of independent data points used in the comparative study is not over estimated, and to determine the order of the events.

Consider the following example, where a biologist is studying the occurrence of a particular character state (e.g., blue egg color), and finds that it occurs in 50 of the 100 species in a clade (all the descendants from a single ancestor or node). That particular character state might have evolved independently as many as 50 times, or as few as once, and knowing the phylogeny for the clade makes it feasible to estimate the evolution of that character, and hence answer these questions with greater accuracy. Similarly, the biologist might also like to know if the trait is *ancestral* (i.e., that it evolved before the clade of interest) or *derived* (i.e., that it evolved at some point within the clade under study). These questions fall more generally into the question of understanding the evolution of a particular character within a clade of the evolutionary tree for a particular group.

More generally, for quantitative characters the researcher might want to know about the rate at which the character has changed in the clade overall or in different parts of the clade, and for qualitative characters the researcher might like to know the number of times a trait was gained or lost and how those gains and losses are distributed. For both types of traits, the researcher might also like to know if there are correlations between changes in sets of traits or among traits and particular external (environmental) factors. In order to ensure independence between the inferences made in a comparative analysis and the phylogeny that is used to make the inference, biologists usually strive to use different sets of characters for inferring phylogenies and for comparative studies. Comparative studies have become extremely common in biology, and one can easily find examples of them in almost any issue of experimental biological journals, as well as in the general science journals, *Nature* and *Science*.

A second common use of phylogenies is to test biogeographic hypotheses. Biogeography is concerned with the geographical distribution of organisms, extant and extinct. For example, a researcher may be interested in whether a particular group of species has colonized a set of islands a single time or repeatedly. This can be assessed by determining whether all of the species on the island arose from a single most recent mainland common ancestor or whether there are multiple independent mainland ancestors.

Phylogenies can be used to look at the mode and tempo of speciation. Regular bifurcating speciation is hypothesized to occur by several mechanisms. One of these allopatric speciation, is caused by the geographical separation of a single ancestral species into two geographically isolated groups. Hence, one can look at the geographical distributions of species in a clade to determine whether there is support for allopatric speciation. For example, if the geographical ranges of a set of species is well known, it is sometimes possible to correlate speciation events with large scale geological events such as mountain building or plate tectonics. If there are adequate fossil records or if a molecular clock (the assumption that each character evolves at a rate that is proportional to elapsed time) can reasonably be assumed, one can also compare and contrast the rates of speciation in different clades and in different parts of a clade. However, the assessment of speciation rates is somewhat clouded by the fact that our knowledge of the number of extinct species in a clade is often poor. Since every extinct lineage must also have had a speciation event, it is possible to significantly underestimate the absolute number of speciation events, and their relative proportions between clades. For the same reasons, it is often difficult to use phylogenies to say much about rates of extinction in different clades.

One can also use a phylogeny to attempt to infer the amino acid sequence of extinct proteins. These putative extinct proteins can then be synthesized or an artificial gene coding for them can be produced, and the functional characteristics of the protein that are of interest to the researcher can be tested.

In a more practical vein, phylogenies can be used to track the evolution of diseases, which can, in turn, be used to help design drugs and vaccines that are more likely to be effective against the currently dominant strains. The most prominent example of this use is the flu vaccine, which is altered from year-to-year as medical experts work to keep track of the influenza types most likely to dominate in a given flu season [6].

Finally, phylogenies have even been used in criminal cases, most famously, in a case where a doctor in Louisiana was accused of having deliberately infected his girlfriend with HIV [34]. The phylogenetic evidence featured prominently in the trial and the doctor was ultimately convicted of attempted second degree murder.

The key here is that phylogenies are useful in any endeavor where the historical and hierarchical structure of the evolution of species can be used to infer the history of the point of interest.

### 1.3 The steps of a phylogenetic analysis

---

Biologists are reconstructing phylogenies for hundreds of sets of organisms representing all of the major divisions of the “Tree of Life,” e.g., bacteria, plants, animals, etc. While that might lead one to think that the practice of inferring phylogenies could be very different depending upon the taxonomic group that is being studied, there are many details and steps in the generation of the primary data that are similar whether the researcher is studying microbes, invertebrates, vertebrates, plants or fungi. And if DNA sequence data are used as the primary data for reconstruction, the steps after purified DNA has been extracted from the group of interest are very similar, if not identical. In this section, we outline the

process of inferring a phylogeny, from its conception to the assessment of support for an inferred phylogeny.

Our aim in providing this information is to give developers of phylogenetic methods a feeling for what a biologist must do to produce a phylogeny for a group of organisms. Some steps are relatively easy and not very time consuming; others are rate limiting and have significant impact on the type, quantity and quality of raw data that are available for analysis. A better understanding of how biologists work, what data they can easily augment and what they must struggle to provide, will help researchers to produce methods that will be most useful to biologists.

### 1.3.1 Designing the study

Before a biologist collects her first organism, the scope and nature of her study must be delineated, and a plan must be put in place to accomplish its goals. The first decision to be made is why the study is being conducted. This decision will determine the taxonomic scope of the project, e.g., a small recently evolved clade (such as a single genus of the great apes), or a clade that encompasses an older group (such as all insects), or even the relationships among the kingdoms of life<sup>1</sup>. However, the broader the taxonomic scope, the more likely a supertree analysis will be needed, for reasons which we will discuss below. In particular, the phylogeny of all life cannot be directly reconstructed. Instead, phylogenies of many subsets of the complete phylogeny are being independently inferred by hundreds of systematists around the world with the goal of ultimately combining the subsets into a single phylogeny for all life. This all-inclusive phylogeny is a major goal of phylogenetic reconstruction.

However, there are many reasons for wanting to reconstruct phylogenies at all levels, and the reasons for estimating the phylogeny of a particular group can be as varied as the purposes to which a phylogeny can be put (Section 1.2). However, once the taxonomic scope of the work is determined, many of the facets of the study (the sampling scheme, the markers to be used, etc.) will either be determined or at least constrained.

In this context a “marker” can refer to a particular region of DNA, a protein, a morphological character, or the order of genes on a chromosome. Each type of marker has its own special challenges and techniques. Rather than attempting to provide a comprehensive overview of how all of the different markers are collected and used, we will focus on the most common type in use today, DNA sequences, with occasional comments about the other types.

#### *Taxon selection and sampling*

Although a phylogenetic reconstruction can occasionally encompass over 10,000 species [49], the majority of studies where the researchers generate the primary data from scratch

---

<sup>1</sup>In traditional taxonomy, organisms are hierarchically grouped according to a standard terminology. From least to most inclusive the hierarchy goes species, genus, family, order, class, phylum, kingdom. Generically, any level can be referred to as a taxon. Ideally each hierarchical level represents a clade of organisms, but there are still many taxonomic groups where it is not yet known if they represent clades, and many organisms have to be reclassified as their phylogenetic relationships become clearer. Since there are at least several million species on earth, the number of taxonomic categories is obviously insufficient to give a name to every level of clades in the phylogeny of life. It is also important to note that the evolutionary depth of a given taxonomic level is not consistent from one group of organisms to the next. For example, *not* all genera represent clades that are approximately 4 million years old.

involve fewer than 100 species. Therefore, unless researchers are working with a small clade of organisms, a sampling scheme needs to be decided upon since it will not be possible to include all of the species in the clade. For example, if a researcher decides to look at floral evolution in a genus of plants with over two hundred described species distributed throughout the globe, several aspects of the study become immediately evident. First, several issues may prevent collection of specimens for every species in the genus: The number of species may be too great to collect and process; the species may be too geographically widespread to allow all collections to be made, even if sympathetic workers in other countries agree to help collect specimens locally; it may be politically infeasible to get collections from some countries; some countries may have highly restrictive laws regarding study of their flora by foreign scientists; and it might not be possible to obtain sufficient funds for all of the travel and permits required. These considerations will necessitate strategic decisions about which taxa are most critical to the study and what sort of sampling scheme will be most likely to yield scientifically valid results, given the goals of the study. It also means that methods designed for inferring phylogenies need to be robust to missing taxa. It is important to note that even if a researcher is able to sample all of the extant taxa for his or her group, there are likely to be some missing taxa due to extinction.

Because our example study is interested in the evolution of characters, the researcher will also have to decide how many individuals from each species will need to be examined to have a statistically valid sample of the range of character variation within and between individual species. Also, if the species are recently evolved and poorly circumscribed taxonomically, the researcher may need to sample several individuals from more than one population of each species to ensure variation that is used for reconstructing the phylogeny truly represents interspecific variation rather than ancestral variation that has been inherited by several species from a common ancestor. If the species for which a phylogeny will be reconstructed are more distantly related, a smaller number of samples will usually be required from each species. Finally, because in nearly all cases biologists want to reconstruct the rooted phylogeny, the researcher will need to select species to be used as *outgroups*. Outgroup species are ones that are not included in the clade of interest (the ingroup) and which, therefore, can be used to root the clade of interest. This method of rooting works because the common ancestor of the ingroup and the outgroup lies further back in time than the common ancestor of the ingroup. If the outgroup is correctly chosen, and a phylogeny is reconstructed that is correct (as an unrooted tree), then the rooted phylogeny obtained by rooting the reconstructed phylogeny on the branch between the ingroup and the outgroup will be the correct as a rooted phylogeny.

Selecting the outgroup is tricky. The first issue is the obvious one: a taxon may seem to be an outgroup, but it may not be. And if the taxon is not actually an outgroup, the resultant rooting of the reconstructed phylogeny can be incorrect, and hence the inferred order of speciation events will also at least in part be incorrect. Since many of the uses of phylogenies are strongly based upon accurate reconstruction of the order of speciation events, this can have negative consequences for scientific inferences later on. This is not just a theoretical danger: in many cases, work has shown that species that were considered outgroups were misclassified and actually belonged in the ingroup. For example, potatoes and tomatoes were originally thought to be in different genera, but are now known to be close relatives in the same genus.

Since more closely related groups of species are expected to be more similar to one another (because they will have inherited a larger number of unaltered ancestral traits from their most recent common ancestor), it would seem that adding a taxon (as an outgroup) which is most different from the ingroup taxa would be the “right thing” to do; and, indeed, such a choice would avoid the problem of inadvertently picking a member of the ingroup instead of

an outgroup. However, this too is potentially problematic. Suppose we take this approach seriously, and we pick taxon  $A$  as an outgroup to a set  $S$  of taxa. Let us also suppose that  $A$  is in fact a true outgroup, and that it is quite dissimilar to every taxon in  $S$ . The problem here is that the more dissimilar taxon  $A$  is to the taxa of  $S$ , the harder it is for any phylogeny reconstruction method to connect the taxon  $A$  to the phylogeny on  $S$ . That is, evolution is modeled as a random process that is operating on a phylogeny, producing sequences at the nodes of the tree. The more dissimilar  $A$  is to  $S$ , the closer to random the sequence for  $A$  is to the sequences for the taxa in  $S$ . The more random these sequences look relative to each other, the more random the placement of the edge that connects  $A$  to the rest of the tree. This has the consequence that, once again, the rooting of the resultant phylogeny can be incorrectly placed, and much of the subsequent scientific analysis can be based upon false premises.

Thus, a good outgroup taxon must be dissimilar enough to be definitely an outgroup taxon, but not so dissimilar that the phylogenetic reconstruction method cannot reconstruct the phylogeny correctly. If a molecular clock hypothesis is reasonable for the dataset (including the outgroup), these decisions are much easier to make, but in general this is quite a tricky and difficult problem.

### **Marker selection**

Another critical part of the planning is to choose markers that are appropriate for the study. When a phylogeny is reconstructed for a region of DNA, what is really being reconstructed is the phylogeny of that DNA region, which does not necessarily have to be the same as the phylogeny of the species from which the DNA was taken (see below); this is the classic gene tree/species tree problem [32, 33, 41, 43, 44, 42]. What can a researcher do to increase the probability that their DNA region will produce a gene tree whose evolutionary history is identical to that of the species tree? Achieving this ideal requires picking markers with the following characteristics:

1. *The marker is unrecombined.* When sex cells (eggs, sperm, ovules, pollen, etc.) are produced by normal diploid organisms (organisms that inherit one copy of each nuclear chromosome from their mother and one from their father<sup>2</sup>), they undergo a process called recombination in every generation, which causes individual strands of DNA to be a mixture of two or more different gene histories. Details about recombination (or more formally “meiotic recombination”) are available in any introductory genetics text. What is important here are not the details of how recombination produces DNA sequences having multiple evolutionary histories, but rather the effect this can have on phylogenetic reconstruction. If a recombination event has taken place within a marker used for phylogenetic analysis it will be composed of two different evolutionary histories. If the researcher can determine where the recombination event occurred, then the marker can be broken into these two regions, and each region can be analyzed separately – and each can produce a potentially different gene tree. But since only rarely does a researcher know that a marker has undergone recombination, it is more likely that she will analyze the full DNA sequence without considering the separate histories of the different parts of the sequence. Depending on the reconstruction method

---

<sup>2</sup>The one exception to this rule is the sex chromosomes. The sex that is heterogametic (XY, ZW, etc.) only has one copy of each type, e.g., human males have one copy of the X and one copy of the Y chromosomes.

used, analysis of such combined histories can produce a phylogeny that neither reflects the gene trees nor the species tree. To avoid the problem of recombined sequences, in many cases researchers use sequences that are predominantly uniparentally inherited such as mitochondrial and chloroplast sequences. Because these organelles are inherited from only one parent the sequences in these organelles are haploid and do not undergo recombination. This greatly reduces the probability that different genes will have different trees.

2. *The marker is single copy or is subject to rapid concerted evolution.* Another way that gene trees and species trees can differ is the presence of gene duplication, i.e., when a gene is copied at two or more locations (loci) on one or more chromosomes. If a gene is duplicated one or more times before a clade originates and then different copies of the duplications are randomly lost in different lineages (“random assortment”), the leaves of the tree will have different sets of copies of the gene. Depending upon how the systematist analyzes the data (a complicated situation which we will not discuss here - see [31] for more information), this can cause the reconstructed gene tree to differ from the species tree.

The gene tree is more likely to match the species tree if there is only one copy of a marker, and for this reason single copy markers are the most desirable in phylogenetics. However, some highly duplicated genes can be useful in a phylogenetic analysis, provided they undergo “concerted evolution”, a process that rapidly homogenizes all the copies of a gene to a single type with the same sequence. As long as concerted evolution is homogenizing the sequences of the copies at a rate significantly faster than the rate of speciation, then the probability that the phylogeny for the marker (the “gene tree”) will be identical to that of the species tree will be high. Thus, picking markers with multiple copies can also work, provided that the region has sufficiently rapid concerted evolution. The ribosomal DNA repeat is one such region, and it is broadly used by systematists.

3. *Finally, it is preferable to sequence the same allele of a gene for reconstruction.* Because diploid organisms have two copies of each autosome (non-sex chromosomes), they have two copies (alleles) of every gene at a position (called a “locus”) on a chromosome. Although each individual in a species can only have up to two different alleles at a locus, collectively, all the individuals in the species can, and often do, have three or more alleles at that locus. Hence, even single copy genes can experience the same assortment and sampling problems expected with duplicated genes. The only way for a researcher to make sure he/she is using the same allele for phylogenetic reconstruction is to sequence extensively and determine the phylogenetic relationships among the alleles of all the species in his/her clade. Because this is so much extra work, most systematists focus on organellar DNA sequences or nuclear sequences that undergo rapid concerted evolution, which homogenizes the repeats on both copies of the chromosome.

In addition to the issues surrounding gene trees and species trees, the researcher is also looking for the following characteristics in a DNA region.

1. *The marker evolves quickly enough to distinguish among the most recently evolved species in the ingroup, but does not evolve so quickly that it is either impossible or extremely difficult to infer a reliable multiple sequence alignment* (see Section 1.3.4). Markers that evolve very slowly, such as the ribosomal DNA genes, can be used to reconstruct relationships among organisms from different kingdoms, the broadest traditional taxonomic group, whereas, very rapidly evolving markers,

such as mitochondrial sequences in animals, may only be suitable for reconstructing genera, the least inclusive taxonomic group above species. The incomplete taxonomic coverage of all but a few slow evolving DNA regions is part of the reason supertree methods are needed for reconstructing large numbers of species from diverse taxonomic groups.

2. *The marker is readily amplifiable by polymerase chain reaction (PCR).* At present, almost all DNA sequencing is preceded by amplification of the region to be sequenced using PCR. PCR requires highly conserved sequences at either end of the marker so that “primers” (single stranded DNA sequences that have 18-30 nucleotides) will bind to them for all of the species that will be sequenced. If a priming sequence for a marker varies significantly within the set of species in the study, it can be very difficult to get DNA sequence data for all species in a study.
3. *The marker can be sequenced.* Some DNA regions can be amplified but are very difficult to sequence due to repetitive elements in the sequence. This problem can be caused by either very short repeats (one to four nucleotides in length) that cause the DNA polymerase enzyme to stutter and produce different numbers of repeats, or very large repeats that make it difficult to sequence through that region. Modern DNA sequencing methods use primers in a fashion similar to but not identical to PCR. For a given primer only 600-1000 nucleotides can be sequenced. If contiguous repeats are longer than twice this upper limit, it may be impossible to sequence through that repeat region. See Section 1.3.4 for a discussion of the problems with aligning repeats.

For example, if the ingroup is closely related, rapidly evolving markers that have been vetted by the systematics community for the characteristics enumerated above will be a systematist’s first choice, but if these prove to have too little variation to distinguish the species, the researcher may have to invest significant time and resources into developing a new marker or markers. In general, for each kingdom of organisms, there is a fairly small set of DNA regions that are considered acceptable for phylogenetic analysis. There are undoubtedly more regions that could be used, but to save time and expense, researchers use the ones that are already developed first.

Having made these critical decisions (and hopefully having secured funding), a systematist can now turn to the next step.

### 1.3.2 Collecting organisms in the field

Depending on the taxonomic group that is under study, collecting the organisms which will be studied can involve a small number of trips close to the home institution of the researcher or a number of far flung trips to locations that are difficult to access for both geographical and political reasons. Often the researcher will have to obtain permission from one or more governmental agencies to collect specimens. This is especially true in less developed countries where concern about bioprospecting is high. The rules for collecting in different countries are as varied as the countries themselves, and the researcher will have to negotiate the legal web of the countries in which he/she needs to collect. In some cases, it will be possible to get tissue for DNA for some species by arranging to borrow museum or herbarium collections recently made by other systematists. The older the specimens, the less likely they will have intact DNA.

When the researcher is in the field he/she must collect and preserve specimens of every species that will be studied. How the specimens are preserved and returned to the lab depends on the taxonomic group studied. For most vertebrates, specimens will either be

frozen in the field and then processed back at the lab or placed in a preserving solution. Insects can be easily captured, killed and preserved in the field with particular body parts, e.g., a leg, being harvested for DNA extraction back in the lab. Plants are usually placed in presses in the field to produce dried vouchers that will be kept permanently at a herbarium. At the time of collection, some material (usually leaves or flowers) from the voucher will be quickly dried in silica gel. With plants, the researcher also often has the option of collecting seeds which can later be grown in a greenhouse for vouchering and fresh material for DNA extraction.

For microorganisms or fungi, a researcher might collect from a particular locality and then culture the organisms back in the lab for identification and DNA extraction. However, the vast majority of species in these groups are not culturable. In these cases, the researcher will collect material, e.g., soil or water, from a locality and preserve it until it can be brought back to the lab where molecular techniques can be used to determine the species in the sample.

Collecting is often conducted over a period of several years, so the researcher is at pains to plan well before the trips are made. It may be prohibitively expensive or politically infeasible to return to an area a second time in rapid succession.

### 1.3.3 In the lab

Once a portion of the species in a study has been collected, work can begin on gathering the primary data for phylogenetic analysis. We focus here on DNA sequence data, researcher might also collect morphological, RNA, protein or gene order data.

For multicellular organisms, generally, a small piece of tissue from the organism is taken through a series of physical and chemical steps to release and purify the DNA from cells and organelles. For single celled organisms, either single species are cultured under sterile conditions and samples are taken from these cultures, or a mixture of many species is extracted simultaneously, e.g., from a soil sample. The steps for extracting and purifying DNA from different groups of organisms differ, but generally, it is easier to get DNA from animals than it is from plants, fungi, and some groups of microbes. Plants and fungi often have secondary compounds that either damage DNA when it is extracted or that co-extract with it, thereby complicating the purification process. Plants, fungi, and some microorganisms also often have either a secondary cell wall or other cell-wall structures that can interfere with DNA extraction. The researcher must often try several different extraction and purification procedures before sufficient quantities of high quality DNA, free of interfering compounds, are reliably obtained. In some cases, months can be spent just on determining an effective extraction and purification procedure.

Once high quality DNA has been obtained, the researcher will usually conduct a preliminary study on a subset of their taxa to determine which markers are likely to have enough informative variation for phylogenetic reconstruction. This study will consist of some taxa that are expected to be closely related and some that are expected to be distantly related, in an attempt to determine whether a given marker has sufficient variation to distinguish among closely related taxa but is not evolving so rapidly that it will be unalignable for the distantly related taxa and the outgroups.

Generally, PCRs will be set up to amplify the region of interest. If the amplifications are successful, they will be purified and sequenced. If they are unsuccessful, the researcher will attempt to determine if there was a problem with the PCR or with the template DNA and will correct the problems with the PCR or try other methods of getting purified DNA, respectively. Although modern DNA sequencing methods make it possible to sequence large quantities of DNA, and the sequence of nucleotides can be called with a fair degree of

accuracy by machine when the raw sequence data are of sufficient quality, the automated process is not infallible. Researchers proofread their sequences by eye and resequence regions that are ambiguous or of low quality.

If these preliminary runs indicate the marker is good for the group of interest, the researcher will amplify and sequence the marker for all of the taxa in the study. Because it is often the case that some taxa are more difficult to obtain or extract than others, a researcher often does not complete the sequencing for all of the taxa simultaneously. When this occurs he/she may perform preliminary phylogenetic analyses on the taxa that are more readily available for sequencing.

Usually, a researcher will sequence two or more markers and then check whether the phylogenetic analyses of each marker produces trees that are topologically identical. Topologically different trees produced by different markers can be caused by several things: gene tree/species tree problems, reticulation, lack of support for parts of the phylogeny, and use of different reconstruction methods for different markers. If a researcher finds that markers and the methods of analysis produce conflicting phylogenies, he/she will usually take additional steps to determine the source of the conflict. Conflicting tree topologies from multiple markers from the same organelle (or multiple markers from a region experiencing concerted evolution) cannot usually be caused by reticulation or gene tree/species tree problems since the markers usually do not contain multiple evolutionary histories. Therefore, in the absence of compelling evidence that two or more organellar markers have well supported conflicting phylogenies (see Section 1.3.6), it is usually assumed in these cases that conflict is due to particular model conditions (e.g., short edges) that make it hard to fully resolve the evolutionary tree. Such conditions usually result in aspects of the reconstructed trees that are not well supported by the data. To solve this problem, the researcher will often sequence more of the organellar genome or the region that is evolving concertedly. On the other hand, if organellar and nuclear or two or more nuclear markers produce well supported conflicting phylogenies, the researcher will have to try to decide whether the cause is gene tree/species tree problems or reticulation. At present, the methods for making this judgment solely on the basis of the sequences are lacking. However, the researcher may be able to make a judgment using other biological information that we do not discuss here.

### 1.3.4 Multiple Sequence Alignment

Once the sequence data are available, they need to be put in a multiple sequence alignment (MSA) before a phylogeny reconstruction method can be applied. A multiple sequence alignment of a set  $S$  of sequences is defined by a matrix where the rows are the sequences in  $S$ , and the entries within each column are “homologous”. For the purposes of a MSA, homology means “positional homology”, which asserts that all the nucleotides in the same column have evolved from a common ancestor. However, multiple sequence alignments (especially of protein sequences) can also be defined in terms of structural homology, so that columns identify residues that produce identical structural features in the three-dimensional folding of the protein. To a large extent, structural alignments and phylogenetically driven alignments are either the same or very similar, but there can be differences because non-homologous regions in proteins can sometimes evolve to have the same functional/structural form. “Convergent evolution” is the term used to describe this phenomenon, whereby similar characters in different species can evolve from nonhomologous genes or gene regions. Convergent evolution can take place at many levels in organisms, e.g., similar structural features in a protein or the spines on cacti and other desert plants.

The usual procedures for producing a multiple sequence alignment operate by inserting gaps (represented by dashes) into the sequences, so that the final resultant sequences are

all the same length. This limitation means that if the sequences submitted to the MSA method do not begin and end at homologous sequence positions, the leading and trailing bases for which at least some of the sequences lack homologous positions will often not be aligned correctly. In some cases this can confuse the alignment algorithms and produce very poor alignments as they attempt to make all of the sequences the same length. To avoid this problem, most researchers trim their sequences to begin and end at what they believe are the same homologous positions before submitting them for multiple alignment.

However, equalizing the length of the sequences is not the objective but rather a feature of the MSA process, as the following discussion should make clear. For example, if we begin with  $n$  DNA sequences, with maximum sequence length  $k$ , the result of an MSA will be a set of  $n$  sequences over the alphabet  $\{A, C, T, G, -\}$ , each of length  $k' \geq k$ . These sequences can then be placed in an  $n \times k'$  matrix, and hence the correspondence between matrices and multiple alignments.

Many systematists obtain multiple sequence alignments at least partly by hand, either aligning sequences themselves, or taking the output of some software (e.g. ClustalW [61]) and then modifying the alignment. While this may seem inefficient, the poor quality of the current set of MSA algorithms necessitates it. There are a number of features that are difficult for the current set of methods to handle. The two most prevalent problems are (1) sequences that have diverged so much that similarity is difficult to infer and (2) introduction of large gaps, especially when these are due to different numbers of repeat sequences. When sequences from different taxa have differing numbers of imperfect repeats, the current set of MSA algorithms usually cannot determine which repeats from one sequence should be aligned with which repeats from the other. Consider the simplest example with two sequences. If sequence A has 4 repeats and sequence B has 2 repeats, which of the repeats in A should be aligned with the repeats in B? In some cases, the researcher will simply delete the repeats from his/her analyses. Alternatively, if the researcher thinks the repeats have important phylogenetic information and the repeats are long enough and varied enough, the researcher can try to determine the phylogenetic relationships of the repeats, by producing a MSA consisting of each copy of each repeat from each taxon and then using that MSA as input to a phylogenetic analysis. In this way the researcher can produce a hypothesis of which repeats are homologous (they will appear in clades together) and which are not. The phylogeny can then be used to guide a hand alignment of the putatively homologous repeats within the larger alignment of the marker. An MSA algorithm that could at least partially automate this process would be very useful.

Finally, after the alignment is obtained, it is often further modified in order to eliminate unreliable sites (columns) or taxa (rows). For example, the systematist may eliminate those columns that contain too many gaps to have confidence in the positional homology of that region of the alignment, or that have a low “score” as computed by ClustalW; he/she may also elect to eliminate taxa (rows) that contain too many gaps. The objective of these modifications is to reduce the noise in the alignment that arises from poor quality data, since excessive noise (especially due to large numbers of unequal length gaps, which are often introduced in regions that are hard to align well) will result in poorly estimated phylogenies. Consequently, by eliminating the problematic components of the alignment, it may become possible to obtain an accurate reconstruction of the phylogeny on the remaining data.

### 1.3.5 Phylogenetic reconstruction

After the multiple alignment is obtained and before proceeding to reconstruct the phylogeny, several intermediate steps take place. The first involves deciding how to best analyze datasets which contain multiple markers for the same set of organisms. In this case, the

systematist must decide between doing a phylogenetic analysis on a “combined” dataset (obtained by concatenating the individual datasets), or doing phylogenetic analyses on the individual datasets and then comparing the resultant phylogenies. As discussed above, this determination requires ascertaining whether the datasets have the same evolutionary history, so that issues such as lineage sorting or reticulate evolution can be ruled out as causes for making the evolutionary histories different. Methods for determining when it is safe to combine datasets exist, but these methods are not necessarily sufficiently accurate [26]. Most do not even consider whether reticulate evolution is a reasonable explanation for not combining the sets. As pointed out above, sequences from the same uniparentally inherited organelle are generally considered safe to combine, but unless the assumption of uniparental inheritance is explicitly tested for each species—a time consuming and sometimes highly impractical task—combined analyses can be misleading even for these sequences.

The second issue has to do with the choice of phylogenetic reconstruction method. If the systematist wants to use one of the statistical estimation techniques (i.e., Maximum Likelihood or Bayesian MCMC), she needs to decide which stochastic model would be most appropriate for her data. If she is using multiple markers and wishes to perform a combined analysis (combining the datasets into one dataset), then the model selection will in general be different for the different markers, and her phylogenetic analysis will need to be able to maintain these partitions during the inference phase.

### ***Standard stochastic models of evolution***

There are many stochastic models of site evolution, most of which have been described in terms of DNA sequence evolution. The models used for DNA sequence evolution do not usually involve bringing in constraints on the evolutionary history that would arise due to structural issues (such as secondary structures for RNA and tertiary structures for regions that code for amino-acid sequences), and so are simpler than stochastic models for either RNA or amino-acid sequences. Stochastic models of DNA sequence evolution used in practice range from the simplest Jukes-Cantor (JC) Markov model, to the fairly complex General Time Reversible (GTR) Markov model. These models describe the evolution of a sequence beginning at the root and evolving down the tree as a sequence of *point mutations*. Thus, no insertions, deletions, or duplications occur, and instead the only changes possible on an edge of the tree are at individual nucleotide positions where the current state of a nucleotide changes to another state. The result of these assumptions is that if the root sequence has length  $k$ , then the result of this evolutionary process is that every leaf in the tree is assigned a sequence also of length  $k$ . Furthermore, standard models assume that all the positions (sites) within the sequence evolve under identical processes, and independently of each other.

### ***Rate variation across sites***

The reader may be aware that it is often assumed that the sites evolve at different rates, with some sites evolving more quickly than others; furthermore, some sites may be *invariable*, and so not be allowed to change at all.<sup>3</sup> However, these observations do not violate the assertion that standard models assume that the sites are evolving identically

---

<sup>3</sup>Note that a site that does not vary its state on a particular dataset is said to be “invariant”; however, that does not mean that the site is invariable. Its rate of evolution may be so slow as to not exhibit a change on a particular dataset. Thus, saying a site is invariable is a statement about the model, and not about its evolution on a particular dataset.

and independently (i.i.d.). That is, when rate variation is incorporated into these models, these rates are assumed to be taken from a distribution (typically a gamma distribution); thus, each site has a probability of being invariable (i.e., having rate 0), but if it is allowed to vary, it then selects its rate from a common distribution, typically the gamma distribution, and then maintains this relative rate on all edges of the tree. Thus, even when sites have variable rates or may be invariable, this way of defining the rates has the consequence that the sites are still evolving under identical and independent processes.

Thus, standard stochastic models of evolution are essentially defined by two considerations: how a random site evolves, since all sites will follow the same model, and then the distribution of rates across sites. Typically these rates are taken from a gamma distribution, and can be incorporated into any single site evolution model. Here we now focus on the single site evolution models.

Different stochastic models of evolution differ substantially with respect to their impact on the resultant phylogenetic analysis, and yet the mathematics involved in the theory underlying the models generally in use in phylogenetics (i.e., from the simplest Jukes-Cantor (JC) model to the complex Generalized Time Reversible (GTR) model) is the same. That is, all of these models are identifiable (the probability distribution a model tree defines on the “patterns” suffices to identify the model tree). To understand how a character defines patterns, suppose that a character has  $r$  states and there are  $n$  taxa. Then there are  $r^n$  possible ways that the character can assign states to the leaves, and these are the “patterns” for the character. Furthermore, the parameter values for the model (such as the substitution probabilities on each edge) determine the probabilities of each pattern occurring. Thus, the model tree itself defines a probability distribution on the  $r^n$  possible patterns at the leaves. Saying that a model is identifiable means that knowing the probability distribution of the patterns is sufficient to define the tree. Thus, all the models that are under general consideration, from the JC to the GTR model, are identifiable. However, not all stochastic models are identifiable – the “no-common-mechanism” model in [64] is one such model which is not identifiable. Also, if we use a standard site evolution model but go outside the usual models of how the rates of evolution can vary across sites, we can also get non-identifiability (see [9, 11, 57, 72]).

Thus, all the standard models used in phylogenetic reconstruction are identifiable. Furthermore, methods that have provably good performance under the simplest of these models (i.e., under JC) will also have provably good performance under GTR. In essence, therefore, there is little mathematical difference between any two models in current use in phylogenetic reconstruction.

### ***The Jukes-Cantor (JC) model***

The assumption of the JC model which characterizes it is that if a site changes its state, it changes with equal probability to the other states. Hence, in the JC model we can specify the evolutionary process on the tree  $T$  by the assigning of substitution probabilities  $p(e)$  to each edge  $e$ , where  $p(e)$  indicates the probability that a site changes on the edge  $e$ .

### ***The Generalized Time Reversible (GTR) model***

For the GTR model, we do not make the assumption of equiprobable nucleotide substitutions; instead we have the  $p(e)$  parameters for each edge, and also a  $4 \times 4$  substitution matrix  $M$  which governs the evolution across the tree. This matrix indicates the probability of each nucleotide being replaced by another nucleotide, given that there is a substitution. Thus, the GTR model contains the JC model, the Kimura 2-parameter (K2P) model, etc., as special cases, but allows greater complexity by having additional free parameters.

### **Phylogenetic analysis using stochastic models of evolution**

Various statistical techniques have been developed which make it possible to select the best fitting model within this spectrum – from JC to GTR – for a given DNA sequence dataset [47]. Thus, if a statistical estimation is desired, the first step in a phylogenetic analysis is generally to use one of these statistical tests to select the model under which the data will be analyzed. Once the model is selected, the researcher can then decide how to analyze his/her data – whether with a method that uses the model explicitly, or with one that does not. Phylogeny reconstruction methods come in essentially three flavors: (a) distance-based methods, such as Neighbor Joining [50], which tend to be polynomial time and are very fast in practice; (b) heuristics for either maximum likelihood or maximum parsimony, two hard optimization problems, and (c) Markov Chain Monte Carlo (MCMC) methods. Of these, only distance-based methods are polynomial time; despite this, most systematists prefer to use one of the other types of analyses, because these are assumed to provide better estimates of evolutionary history.

Distance-based methods operate by computing a matrix of pairwise “distances” between the sequences (these are typically not just edit distances, but distances which are supposed to approximate the evolutionary distance, and so are derived from statistically-based distance calculations), and then use only that distance matrix to estimate the tree. Maximum Parsimony is an optimization problem in which the tree with the minimum total number of changes is sought (i.e., it is the Hamming Distance Steiner Tree problem). This NP-hard optimization problem [16] is favored by some systematists. Maximum Likelihood (ML) is another NP-hard optimization problem [10], but this problem is defined in terms of an explicit parametric stochastic model of evolution. The optimization problem is then to find the tree and its associated parameters (typically substitution probabilities) that maximizes the probability of the data. The theoretical advantage of Maximum Likelihood over Maximum Parsimony is that it is “statistically consistent” under most models; this means that it is guaranteed to return the correct tree with high probability [8] if the sequences are sufficiently long – something which is not true of Maximum Parsimony [12]. However, ML is even harder in practice than MP, and heuristics for both problems require very substantial amounts of time (weeks or months) for acceptable levels of accuracy on even moderate sized datasets. MCMC methods also explicitly reference a parametric stochastic model of evolution, but rather than trying to solve the ML problem under the model, they perform a random walk through the space of model trees. After some burn-in period, statistics are gathered on the set of trees that are subsequently visited. Thus, the output of an MCMC method is not so much a single tree, but a probability distribution on either trees, or aspects of evolutionary history.

Thus, the major methods for phylogenetic inference, if run a sufficiently long time to obtain an acceptable level of accuracy, can take weeks, months or longer on large datasets, and even moderate datasets (with just a hundred or so taxa) can take several days. Since there are no well established techniques for determining whether a phylogenetic analysis has run for a sufficient time, most systematists use *ad hoc* methods to determine when to stop.

#### **1.3.6 Support assessment**

At the end of the process of reconstructing a phylogeny, a systematist may or may not have a single best reconstruction. This is particularly common for maximum parsimony analyses, where for some datasets there can be thousands of equally good trees. With maximum likelihood this is less likely to happen (because of the real-valued optimization,

the optimal solution is more likely to be unique), but it can happen with neighbor joining. However, all methods have the potential to produce a set of trees that are very close in score to the best score achievable on the dataset, and which are probably statistically no better. In these cases, the researcher would like to have an objective measure of the support for the best phylogeny (i.e., the one that optimizes the objective criterion, such as MP or ML). In the case where the best tree does not have significant support (such as will happen if the second best trees are not substantially worse than the best tree with respect to the objective criterion), the researcher will still want to know which aspects of the evolutionary history implied by the best tree are reliable, where “reliable” means a measure of how well supported the reconstruction is given the data and the method used. Usually reliability is assessed at the level of individual edges in the tree. Reliability can be addressed through statistical techniques, or through more purely combinatorial or “data-mining” techniques.

### *The combinatorial approach: consensus techniques*

The first step of the combinatorial approach to estimating reliability is to select the profile of trees to evaluate. This profile consists of those trees that are close enough to optimal (with respect to the objective criterion) to be considered equally reliable. From this set, a “consensus” tree will be inferred. Of the many ways of defining consensus trees, the most frequently used consensus methods in systematics are the “strict consensus” and “majority consensus” trees. These are defined in terms of edge-induced bipartitions, in a natural way which we now describe.

Let  $S$  be a set of species,  $T$  be a tree leaf-labelled by  $S$ , and  $e$  be an edge in  $T$ . The deletion of  $e$  from  $T$  splits the tree into two pieces, and hence creates a bipartition  $c_e$  on the set  $S$  of leaves. We can thus identify each edge  $e$  in  $T$  by the bipartition  $c_e$ . Furthermore, the tree  $T$  is uniquely identified by its set  $C(T) = \{c_e : e \in E(T)\}$ , and that set is called the “character encoding” of  $T$ . We can now define the strict and majority consensus trees.

Given a collection  $\mathcal{T}$  of trees, so  $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$ , the tree  $T_{sc}$  such that  $C(T_{sc}) = \bigcap_i C(T_i)$  is called the **strict consensus tree**. It is not hard to see that the strict consensus tree always exists, and that is the most refined common contraction of all the trees in  $\mathcal{T}$ . The tree  $T_{maj}$  defined by  $C(T_{maj}) = \{c_e : |\{i : c_e \in C(T_i)\}| > k/2\}$  is called the **majority consensus tree**. The majority consensus tree always exists, and it always refines or equals the Strict Consensus tree. Both the strict consensus and the majority trees can be computed in polynomial time. In practice, the Majority Consensus tree is more often reconstructed than the Strict Consensus Tree. If desired, other consensus trees can also be computed. In particular, instead of picking the tree whose bipartitions appear in more than half the trees, one can pick a different threshold, and every threshold  $p > 1/2$  will define a tree that will necessarily exist (if  $p < 1/2$  this consensus tree may not exist).

However, other consensus methods also exist (see for example [19, 29, 46, 71] for particular examples, and [5] for overviews of consensus methods (including a discussion of some of these methods)).

Other methods of assessing support are statistical in nature. For MP, ML, or distance-based phylogenetic reconstructions, the most common method of support is a bootstrap analysis, although a jackknife analysis is also sometimes applied. When a Bayesian MCMC approach is used to estimate the tree, posterior probabilities are often estimated.

### *The Bootstrap*

Bootstrap analyses take two basic forms: non-parametric and parametric. A non-parametric bootstrap resamples with replacement each of the positions in the original dataset, creating a new dataset drawn from the same distribution. The researcher then uses the same method

by which the phylogeny was reconstructed, and the resulting “bootstrap phylogeny” is compared with the reconstructed phylogeny. Many replicated bootstrap runs are performed and the proportion of times that each branch in the reconstructed phylogeny appears in the bootstrap phylogenies is recorded and interpreted as support for the branch.

The interpretation of the non-parametric bootstrap is an important issue, and it is often assumed that the bootstrap support for an edge somehow indicates the likelihood that the estimated edge is correct (that is, that it appears in the true tree). However, this is clearly overly simplified. It is possible for a phylogenetic reconstruction method to always return the same tree, thereby producing bootstrap proportions that are all 100%, and yet the tree (and hence all its edges) may be incorrect. Thus, strictly speaking, the best way to interpret the bootstrap support for an edge is that it indicates the probability that the edge would continue to be reconstructed if the same phylogenetic estimation procedure were applied to datasets having the same distribution as the original dataset. Thus, when the phylogenetic reconstruction method is statistically consistent for the model, the bootstrap proportions indicate the strength of support for that edge in the original dataset.

Parametric bootstrapping can only be performed when we assume an explicit model of sequence evolution, such as in ML or distance-based phylogenetic analyses. In this case, we use the original data to estimate the parameters of the stochastic model of evolution on the tree we constructed for the dataset. These parameters will generally include site-specific rates of evolution or the distribution from which the rates of evolution are drawn, substitution probabilities on each edge, and an overall substitution matrix governing the evolutionary process across the tree. Once this model tree is constructed, we can simulate evolution down the model tree, producing datasets of the same length as the original dataset. We then apply the same method used to construct the original tree, and estimate the tree on these new datasets. As with the nonparametric bootstrap, support is assessed by the proportion of times the branches in the phylogeny reconstructed using the original dataset appear in the parametrically bootstrapped reconstructions.

### ***The Jackknife***

Jackknifing involves repeatedly deleting some proportion of the original sites (or in some cases original taxa) at random and then using the same method as was used for the full dataset to reconstruct trees on the reduced dataset. Here again, the aim is to determine the strength of support for the branches the tree constructed using the full dataset by assessing the proportion of times that the branches in the analysis of the full dataset appear in the jackknife reconstructions.

### ***Bayesian MCMC methods***

When the phylogenetic reconstruction uses a Bayesian MCMC analysis, the output itself is in the form of an estimation of the statistical support for each of the hypotheses. That is, instead of a single best tree being returned, the output contains a frequency count for each of the trees that is visited after burn-in. These values are the “posterior probabilities” of the different trees. Just as with the bootstrap and jackknife, the interpretation of these values is complicated and subtle (and like the parametric bootstrap, this interpretation will depend upon issues having to do with the model of evolution used to analyze the data).

### ***Interpreting support:***

Interpretations of each of these three ways of estimating support need to be done carefully, and in each case real support for an edge can only be inferred if the phylogenetic reconstruction method is statistically consistent. Also, there are often severe computational limits to

the use of these methods on large datasets. If the original reconstruction method required a significant period of time to complete (days, weeks or months), then the time to complete a full bootstrap or jackknife analysis can be prohibitively long. Thus, one clear advantage of posterior probabilities over bootstrap methods is that they are a natural byproduct of the search for the best reconstruction, so no additional computational time is needed to produce the support measures after the reconstruction is completed.

## 1.4 Research problems in molecular phylogenetics

---

Molecular phylogenetics is concerned with the estimation of phylogenies from molecular (i.e., DNA, RNA, or amino-acid) data, which usually consists of sequences, but for DNA it can also be gene orders. There are many research issues in molecular phylogenetics, each of which could have potentially a large impact on practice in systematics. Some of these issues involve database research, others involve statistical inference (including developing better models of the evolutionary process), and some are algorithmic in nature. Rather than attempting to be comprehensive, in this chapter we will limit ourselves to discussing algorithmic research involved in phylogenetic inference, since the target audience is algorithms researchers in computer science. Also, in order to keep this chapter reasonably moderate in size, we will restrict our discussion to issues involved in reconstructing phylogenetic *trees*, rather than the more general problem of reconstructing phylogenetic networks.

### 1.4.1 Performance analysis of algorithms

Before discussing research problems in phylogenetics, it is important to discuss how algorithms are evaluated in this community. Algorithm developers are familiar with designing algorithms for numeric or combinatorial optimization problems, like vertex coloring, travelling salesperson, etc. Algorithms for polynomial time problems are generally exact - i.e., they find optimal solutions - and hence they are compared in terms of their running time (whether asymptotic running times or on benchmark datasets). By contrast, if a problem is NP-hard, then algorithms may not be exact, but may instead offer bounded-error guarantees (e.g., obtaining a solution no more than twice the cost of the optimal solution), or simply be designed to find hopefully good local (rather than global) optima. In these cases, algorithms (or heuristics, as they are often called) can be evaluated in terms of the scores they find on benchmark datasets and the time they take to find these scores. Benchmarks can be real datasets (typically from some application domain), or random datasets simulated from some distribution (such as random graphs of some sort). Comparing algorithms for NP-hard problems is thus a bit more complicated than comparing algorithms for polynomial time problems, since there is a trade-off between performance with respect to the optimization criterion and running time. Thus, algorithms for the usual numeric or combinatorial optimization problems can be evaluated both theoretically and empirically, with criteria that include running time and accuracy with respect to the objective criterion.

In this light we now consider phylogenetic estimation. Here, too, we have numeric optimization problems, and for the most part they are hard (either proven to be NP-hard, or conjectured to be so); maximum parsimony and maximum likelihood are two obvious examples. (The evaluation of Bayesian MCMC methods is more complicated, since the output is not a single tree with a score for some objective criterion, but rather a probability distribution on trees. We will discuss the issues in evaluating these methods later in this chapter.)

As with all algorithms for hard optimization problems, methods for MP and ML can

be evaluated using the same kind of criteria and methodology as described above. Thus, benchmark datasets can be obtained, some real and some randomly generated, and reconstruction methods can be compared in terms of their accuracy with respect to the referenced objective criterion on these benchmarks. When the method is a local search heuristic (i.e., it keeps on searching for improved solutions), the methods can also be evaluated at different points in time, and their performance can then be measured with respect to how long it takes each method to obtain an acceptable level of accuracy. These studies can help evaluate the performance of phylogenetic reconstruction methods to the extent that the user is interested in solving the particular optimization problem (whether it be MP, ML, or something else).

However, phylogeny reconstruction problems are different from the usual combinatorial optimization problems, in several significant ways. First, we are trying to estimate the “true” tree, not just solve some numeric problem, and so our real criteria for success must include how close our reconstructed trees are to the tree that actually generated the data. This statement itself makes it clear that phylogeny reconstruction can be considered a *statistical estimation* problem, whereby we are trying to infer something (the tree) from data generated by a stochastic process (defined by the model tree). Thus, issues such as *how much data does a method need to get an accurate reconstruction with high probability* are just as significant as the running time of the method (if not more so) (see the chapter in this volume on Disk Covering Methods, and [70], for some work on this topic). These issues can be evaluated theoretically, or in simulation.

#### **Simulation studies:**

Despite the interest in solving MP and ML, the purpose of a phylogenetic estimation is *not* to optimize some objective criterion, but rather to recover the evolutionary history as accurately as possible. For this reason, accuracy is often measured *topologically* with respect to known history. That is, a reconstructed evolutionary tree is compared against the “true tree”, and the degree of error quantified. These comparisons are made primarily in simulation studies, as follows. In a simulation study, a stochastic model of evolution (such as Jukes-Cantor or the Generalized Time Reversible model) is selected, and a model tree (that is, a rooted tree  $T$  along with the parameters necessary to define the evolutionary process) is specified. Then a sequence of some length is placed at the root of the tree  $T$  and evolved down the tree according to the specified model. At the end of this process there are sequences at each leaf of the tree, and these can be given as input to a phylogeny reconstruction method (such as neighbor joining, or a heuristic for MP, etc.), thus producing an estimated tree  $T'$ . The estimated tree  $T'$  is then compared to the model tree  $T$  with respect to topological accuracy.

Of the various ways of measuring topological accuracy, the Robinson-Foulds (RF) metric [48] is the standard technique.

The RF metric between trees is defined in terms of the character encoding of a tree (as described in Section 1.3.6 earlier). Let  $T$  be the true tree for a set  $S$  of  $n$  taxa, and let  $T'$  be a reconstructed tree on  $S$ . Then  $RF(T, T') = \frac{|C(T) - C(T')| + |C(T') - C(T)|}{2^{(n-3)}}$ . Note that  $0 \leq RF(T, T') \leq 1$ , and that  $T = T'$  if and only if  $RF(T, T') = 0$ . RF rates below 10% are generally required, unless the data themselves are so poor that a good estimation of the true tree is unlikely.

The major advantage of using simulations as compared to real data is that for almost all real datasets, it is not possible to know precisely the correct evolutionary history, and those aspects of the evolutionary history that are reliable are also generally easy to infer using any method. Thus, it is not particularly helpful nor straightforward to try to evaluate

methods with respect to topological accuracy on real datasets. Since topological accuracy is so important, simulations have become a standard methodology for evaluating the accuracy of phylogenetic reconstruction methods.

There are, however, distinct issues (and disadvantages) in using simulation studies. The most compelling of these issues is that the mathematical models used to define the evolutionary processes are not nearly as complex as those that operate on real organisms and genomes; thus, inferences about accuracy of reconstruction methods must be taken with a certain amount of salt, metaphorically speaking. Additionally, to the extent that simulations are used to evaluate running times for heuristics for hard optimization problems, the landscape produced by these models are also smoother (easier to navigate and find optimal solutions) than the landscapes of real datasets. All in all, however, simulations are important (otherwise we can rarely, if ever, have a real benchmark), and have changed practice within molecular systematics dramatically.

### 1.4.2 Phylogenetic reconstruction on molecular sequences

Although the objective in a phylogenetic analysis is a topologically accurate phylogeny, this objective cannot be directly addressed. Instead, systematists either try to solve some explicit optimization problem (such as MP or ML), or use Bayesian MCMC methods (which can be seen as an attempt to solve a non-standard optimization problem), and hope that a very good solution to the optimization problem will yield a reasonably accurate phylogeny. Of the various optimization problems that have been posed for phylogeny reconstruction, maximum parsimony (MP) and maximum likelihood (ML) are the two most popular criteria. The community of practicing systematists is not uniformly in favor of one of these two criteria, but both criteria have strong support in the community. Unfortunately MP and ML are both NP-hard, and also very hard to solve in practice except for small datasets, and so exact solutions are not generally feasible.

#### *Heuristic searches for MP and ML*

Heuristics for MP and ML (largely based upon hill-climbing) are in very broad use in the systematics user community, and seem to provide quite accurate solutions on small to moderate sized datasets. These heuristics differ from each other in various ways, but most use the same set of transformations for moving from one tree to another. First, a fast method (neighbor joining, or a greedy insertion of taxa into a tree to optimize MP or ML) is used to obtain an initial tree. Then, a neighborhood of the tree is examined, and each tree scored with respect to the objective criterion (MP or ML). If a better tree is found, the search continues from the new tree; otherwise, the current tree is a local optimum and the search may terminate. Some heuristics include additional techniques in order to get out of local optima. The Ratchet [39] is one of the most successful of these techniques for getting out of local optima; it randomly perturbs the sequence data, and then hill-climbs from the current tree (but using the perturbed data to score each visited tree) until a local optimum is found. Then the data are returned to their original values, and the hill-climbing resumes.

It should be clear from the description that these methods may not terminate in any acceptable time period, especially if randomness is included; thus, the systematist must decide when the search has gone on long enough.

The best of these heuristics, as implemented in the popular software packages PAUP\* [59], TNT [18], and others, are quite effective at producing good MP analyses on even fairly big datasets (containing a few hundred sequences), provided enough time is allotted. The limits for maximum parsimony analyses using currently available software is probably

1000 taxa, and maximum likelihood analyses are probably limited to 100 (or fewer) taxa. (Bayesian MCMC methods are reputed to be able to do well on large datasets, but as it is not clear how to evaluate the performance of these methods, this needs additional study.)

However, with the success of DNA sequencing projects, systematists are attempting to reconstruct phylogenies on larger datasets, where the number of sequences can be easily above a thousand. These datasets are much harder to analyze well in a “reasonable” time (of perhaps a few days or even a few weeks).

Thus, in general, large-scale phylogenetic analysis is quite difficult to do in a reasonable amount of time, and much of the focus of algorithm development (and of our discussion) is on developing better heuristics – ones that can provide sufficient analyses on large datasets in a matter of days rather than months or years. However, since MP and ML searches can take a long time to find optimal solutions, it is also important to be able to assess when it is safe to stop searching for better trees. The current technique is to run the analysis until it seems to have converged. However, it is very difficult to establish convergence, and there is clearly a real possibility that the method has not converged so much as slowed down dramatically with respect to finding improved scores.

Thus, a natural combinatorial optimization problem that is also relevant to practice is to obtain better bounds for MP and ML. If good lower bounds could be obtained, these could be used to evaluate how close to optimal a current best tree is, and that in turn could be used to evaluate whether it was reasonably safe to stop running the heuristic search.

### **Maximum Parsimony**

We begin with Maximum Parsimony, which is the simplest of these optimization problems. Heuristics for this problem have been used to construct perhaps the majority of published phylogenies, and so MP is a major approach to phylogeny estimation. However, optimal solutions to MP may not be correct reconstructions of evolution. There is no guarantee that MP will yield a correct solution, even given infinitely long sequences because MP is not statistically consistent in general. Still, MP is an important problem, and improved algorithms for MP would represent an important advance for computational phylogenetics.

#### **The Maximum Parsimony Search Problem:**

- *Input:* Set  $S$  of sequences, each of length  $k$ , in a multiple alignment
- *Output:* Tree  $T$  leaf-labeled by  $S$  and with additional sequences, all of length  $k$ , labeling the internal nodes of  $T$ , so as to minimize

$$\sum_{(u,v) \in E} H(s_u, s_v),$$

where  $s_u$  and  $s_v$  denote the sequence labelling the nodes  $u$  and  $v$ , respectively,  $H(x, y)$  denotes the Hamming distance between  $x$  and  $y$ , and  $E$  denotes the edge set of  $T$ . (The Hamming distance between two sequences is the number of positions in which they differ.)

Maximum Parsimony (MP), is thus the *Hamming distance Steiner Tree problem*. Although MP is NP-hard [16], it, like Steiner Tree problems in general, can be approximated. Also, although finding the best tree is NP-hard, it is possible to score a given fixed tree in linear time (i.e., it is possible to compute sequences at the internal nodes of a given fixed tree so as to minimize the total number of steps in  $O(nk)$  time, where there are  $n$  sequences each of length  $k$ ), using the well known Fitch-Hartigan algorithm [15]. Therefore, finding the best MP trees can be solved exactly through techniques such as exhaustive search and

branch-and-bound, but such techniques are limited to about 25 or 30 taxa, since the number of trees on  $n$  leaves is  $(2n - 5)!!$ . For larger datasets, heuristic search techniques are used to analyze essentially all datasets of interest to systematists. On moderate sized datasets (up to a few hundred sequences) these heuristics probably work quite well (though there are no theoretical guarantees bounding the error of these heuristics), but much less is understood about their performance on larger datasets, especially datasets with more than a thousand sequences.

### *Research questions for MP*

In addition to the general challenges we discussed earlier in the context of both MP and ML, there are many research questions of particular relevance to MP. One question that is particularly intriguing is to explain, mathematically, why MP is as good as it is. That is, statistical theory has established that MP is not a statistically consistent method for even simple DNA sequence evolution models, and so cannot be guaranteed to reconstruct the true tree (with high probability) even on extremely long sequences. Yet MP's performance in simulation studies (when the model trees are sufficiently large and biological) is clearly not bad. In some cases, it can be better than statistically consistent methods like neighbor joining, and comparable to ML. Why? There must be some theory to explain this phenomenon.

### *Maximum Likelihood*

The usual maximum likelihood problem in phylogenetics is to find a tree and its associated parameters so as to maximize the probability of the observed data. Since stochastic models differ according to the parameters that must be specified (and those that are fixed for the model), the use of a maximum likelihood analysis requires that the stochastic model of evolution already be explicitly specified. While the model choice definitely affects the running time of the software for finding the best ML trees (the more parameters in the model that must be estimated, the more computationally intensive), in essence the mathematics of ML estimation does not change between the simplest model (JC) and the most complex of the standard models (the GTR model). Thus, we will discuss Maximum Likelihood estimation under the JC model.

Recall that, like all of the standard models, the JC model assumes that the sites evolve identically and independently down a tree  $T$ , and that the state of each site at the root of  $T$  is drawn from a given distribution (usually the uniform distribution or a distribution estimated from the dataset). The model is then simply defined in terms of the evolution of a single site. The main feature of the JC model is that it asserts that if a site changes on an edge  $e$ , it changes with equal probability to the remaining states. Thus, the entire evolutionary model can be described by the pair  $(T, p)$ , where  $T$  is a rooted tree, and  $p$  is a function  $p : E(T) \rightarrow (0, 3/4)$ , so that  $p(e)$  is the substitution probability on edge  $e$ .

Once a model tree (that is, the tree  $T$  and its associated parameters as defined by the function  $p$ ) is specified, it is possible to define the probability  $Pr[S|T, p]$  of a given set  $S$  of sequences placed at the leaves being generated by the model tree  $(T, p)$ . Furthermore, this quantity can be calculated in polynomial time, using dynamic programming [13].

### *The ML score of a fixed tree*

Let  $T$  be a fixed tree. The score of the tree under ML (for a given model) is defined to be  $score_{ML}(T) = \sup_p \{Pr[S|T, p]\}$ . Note that we have used the supremum, indicated by  $\sup$ , instead of the maximum. This is because the maximum may not exist, but the supremum will (because the set  $\{Pr[S|T, p]\}$  is bounded from above by 1).

**Maximum Likelihood for the Jukes-Cantor model:**

We now define the ML search problem, again in the context of the Jukes-Cantor model (though the definitions and discussion extend in the obvious way to models with more parameters).

The objective in an ML search is to find the tree with the highest ML score; however, this optimal solution may not exist, because the set is not closed (in the same sense in which we can say there is no largest number in the open interval  $(0, 3/4)$ ). Therefore we will state the ML problem as a decision problem (i.e., does there exist a solution of at least score  $B$ ?).

- *Input:* A set  $S$  of sequences, each of the same length, and a value  $B$
- *Output:* A model tree  $(T, p)$  (where  $p : E(T) \rightarrow (0, 3/4)$  defines the substitution probabilities on the edges of  $T$ ) such that  $Pr[S|T, p] \geq B$  (if such a model tree exists); otherwise, *FAIL*.

From a computational viewpoint, ML is very difficult. Solving ML tree under the Jukes-Cantor model is NP-hard [10] (and conjectured NP-hard under the other models), and harder in practice than MP. Worse, even the problem of finding the optimal parameters for a fixed tree is potentially NP-hard, even for trees with only four leaves! Existing approaches to find optimal parameters on a fixed tree, which utilize hill-climbing on the finite dimensional real-parameter space, are not known to solve the problem exactly [54]. Note that the usual exhaustive search strategy isn't feasible, since this optimization is over a continuous space rather than a discrete space.

**Research questions for ML**

While ML and MP are clearly different, both ML and MP share the same search-space issues, and so heuristics that improve techniques for searching through “treespace” will help speed up both ML and MP analyses. Thus, many of the questions that we discussed in the context of MP apply here as well. However, ML has some additional challenges that are not shared by MP, and which make ML analyses additionally challenging.

The main challenge is computing the ML score of a given tree  $T$ ; in practice, this amounts to finding parameter settings that will produce (up to some tolerated error) the maximum probability of the data. The computational complexity of this problem is open, and current techniques use hill-climbing strategies which may not find global optima [54]. By contrast, the corresponding problem for MP (computing the “length” of a fixed tree) can be solved in linear time using dynamic programming. Thus, parameter estimation on a given tree is the real bottleneck for ML searches. In fact, this is such a time-consuming step in ML searches, that the popular heuristics for ML do not actually try to find optimal parameters on every tree, but only on some. The cost of computing these optimal parameters is just too high.

Consider then the possibility of *not* computing the optimal parameters. Instead, suppose we could quickly compute an upper bound on the ML score of the tree (that is, the probability of the data under the best possible settings of the parameter values). If we could do this, efficiently, we might be able to speed up solutions to ML. That is, during the heuristic search through treespace, instead of performing the computationally intensive task of computing optimal parameters on a tree we visit, we would simply check that the upper bound we have on its score is at least as big as our current best score. If it is not, we can eliminate this tree from consideration. In this way, we can (rigorously) select those trees which are worth actually spending the time to score them exactly, and thus potentially speed up the search.

Closely related to this is the question of simply comparing two trees for their possible scores, rather than scoring either one. Consider the following question:

For fixed phylogenetic trees  $T_1$  and  $T_2$  on set  $S$ , is

$$\text{score}_{ML}(T_1) > \text{score}_{ML}(T_2)?$$

Suppose we could answer this question in a fraction of the time it takes to find the optimal parameters on a tree (i.e. faster than it takes to actually compute  $\text{score}_{ML}(T)$ ). In this case, we could also traverse tree space more quickly, and thus get improved solutions to ML.

The challenge in these approaches is to be able to make these comparisons rigorously and efficiently, rather than in an *ad hoc* fashion. In addition, the objective is to obtain an empirical advantage, and not just a theoretical one.

### **MrBayes, and other Bayesian MCMC methods**

It should be clear that Bayesian MCMC methods are not trying to solve maximum likelihood in the sense we have defined, whereby the model tree (that is, the tree with parameter values) is returned that maximizes the probability of the data. However, there is a kind of maximum likelihood problem which Bayesian MCMC methods *can* be used to solve. This problem is the “maximum integrated likelihood” problem, described in [56], and which we now define. Recall that if we are given a set  $S$  of sequences and a model tree  $T$  with an assignment  $p$  of the parameter values to the tree, we can compute  $Pr[S|T, p]$  in polynomial time. Thus, for a fixed tree  $T$ , we can define the “integrated likelihood” of  $T$  to be the *integral* of this quantity, over all the possible parameter settings. In other words, the integrated likelihood of a tree  $T$ , which we denote  $IL(T)$ , is defined by

$$IL(T) = \int Pr[S|T, p]dF(p|T),$$

where  $F(p|T)$  is the distribution function of the parameters  $p$  on  $T$ . In general,  $(F(p|T))$  has a probability density function  $f(p|T)$ , so that we can write this as

$$IL(T) = \int Pr[S|T, p]f(p|T)dp.$$

The tree  $T$  which has the maximum possible integrated likelihood value is called the “maximum integrated likelihood tree”. The maximum integrated likelihood tree has many desirable properties, some of which are quite surprising [54]. In particular, as Penny and Steel point out in [54], the integrated likelihood of a tree  $T$  is proportional to its posterior probability.

Despite the popularity of MrBayes [27] and other Bayesian MCMC methods, not much is known about how to run the methods so as to obtain good analyses, nor about how to evaluate the performance of a Bayesian MCMC method. By contrast, much more is known about MP and ML. Consider, for example, the question of how to evaluate the performance of a heuristic for MP or ML. We can assemble benchmark datasets, and we can analyze each dataset using various heuristics, and record the best score found by each heuristic under various conditions. This is a legitimate way to compare methods, provided that the conditions are identical.

Some systematists use Bayesian MCMC methods as heuristics for ML; rather than using the normal output (i.e., posterior probabilities), they simply return either the most frequently visited tree, or the model tree which had the best likelihood score. If Bayesian MCMC methods are used in this way, then it is reasonable to use the same methodology for evaluating ML methods to evaluate Bayesian MCMC methods. But this is not what Bayesian MCMC methods are really designed for – they produce posterior distributions on

trees, not single trees with scores. Therefore, how should we evaluate a Bayesian MCMC method? To do this, we need to know what the “correct” output should be, and failing that, whether one posterior distribution is better than another.

### **Research questions for Bayesian MCMC methods**

The most fundamental problem for Bayesian MCMC is to be able to say what the “correct” output is, and to be able to analytically compute that (even if it would take a long time to obtain an answer). Recall the observation made earlier that the integrated likelihood of a tree  $T$  should be proportional to its posterior distribution (i.e., the frequency with which it is visited once the MCMC method has reached the stationary distribution). Hence, if we can calculate the integrated likelihood of each tree exactly, we can actually evaluate the accuracy of a Bayesian MCMC method. The question then becomes: how can we calculate this integral exactly? Once again, this seems difficult, and the problem is that we are trying to calculate something that is in a multi-dimensional continuous space, not a discrete space.

The difficulty in how to evaluate the output of a MCMC method is part of both the appeal and the problem with using MCMC methods – if there is no explicit way to evaluate the quality of the output, then stopping early (and hence finishing quickly) is potentially acceptable. On the other hand, if one wishes to be more conservative about the use of this technology, it becomes necessary to have tools for evaluating how long one should take for the burn-in period, before sampling from what is hoped to be the stationary distribution. Thus, two algorithmic research problems that present themselves in the context of MCMC methods are (1) developing analytical techniques for obtaining bounded-error estimations of the integrated likelihood of fixed trees, and (2) determining when a sufficient amount of time has elapsed so that the burn-in period can be considered over, and the sampling of trees can begin. It is easy to approach these problems using *ad hoc* techniques; the challenge here is (if possible) to develop techniques with a firm theoretical foundation. More generally, of course, designing new MCMC methods with better convergence rates is always beneficial.

### **1.4.3 Multiple Sequence Alignment (MSA)**

MSA remains one of the most significant open problems related to phylogeny estimation, with no really satisfactory software. One of the major challenges for the algorithms designer in developing better MSA methods is that no objective criterion for MSA has been met with general acceptance in the phylogeny research community. Instead, MSA methods (especially MSA methods for protein sequences) are evaluated with respect to accuracy on specific real datasets for which correct structural alignments are known. This makes the development of improved methods for MSA difficult to achieve since the structures of most molecular sequences are not known in advance, and the alignment must be obtained without that knowledge. However, as noted above, a correct structural alignment may not produce an alignment that maximizes positional homology (i.e., structural alignments need not produce a set of columns where each column has character states that are the result of a common evolutionary history for those character states). While not necessarily frequent, datasets where these criteria can lead to different alignments do come up in practice. In summary, for the purposes of a phylogenetic reconstruction, then, the current approaches for MSA are inadequate.

The rest of this section will describe two numeric optimization problems that have been suggested for MSA. Each of these optimization problems defines the cost of a given multiple alignment on the basis of a set of pairwise alignments. Therefore, we begin by describing

how pairwise alignments are scored.

### **Pairwise alignments**

Typically, the cost of a pairwise alignment depends upon the number of each type of substitution, and the number and length of the gaps in the pairwise alignment. The cost of each type of substitution is given by a substitution cost matrix which for computational purposes can be quite arbitrary, although there are standard matrices used in the community. The cost of a gap is more complicated. In general, affine gap penalties are the most frequently used. These penalties are of the form  $C_0 + C_1(l - 1)$ , where  $l$  is the length of the gap, and  $C_0$  and  $C_1$  are two positive real numbers. (Note that if we allowed  $C_0 = \infty$ , then no gaps would be permitted.) Given any such function for the cost of a pairwise alignment, we can then define the obvious optimization problem – given two sequences, find the pairwise alignment of minimum cost. As long as the gap penalty function is affine, the *pairwise alignment* problem can be solved in polynomial time (standard dynamic programming techniques give an  $O(mn)$  time algorithm, where  $m$  and  $n$  are the lengths of the two sequences).

Consider now a multiple alignment  $A$  on the set  $S = \{s_1, s_2, \dots, s_n\}$ , and consider two sequences  $s_i$  and  $s_j$  in  $S$ . The pairwise alignment induced by  $A$  on sequences  $s_i$  and  $s_j$  is obtained by examining the alignment  $A$ , and restricting the attention to just the  $i^{\text{th}}$  and  $j^{\text{th}}$  rows. We denote this induced pairwise alignment by  $A(s_i, s_j)$ . Then if we are given a cost function  $f(\cdot, \cdot)$  on pairwise alignments, we can extend it to any multiple alignment in the obvious way: simply score every induced pairwise alignment, and add up the scores. We formalize this discussion as follows:

### **Sum-of-Pairs (SOP) alignment**

- *Input:* A set  $S = \{s_1, s_2, \dots, s_n\}$ , of sequences and a function  $f(\cdot, \cdot)$  for computing the cost of a pairwise alignment between two sequences given as input.
- *Output:* A multiple alignment  $A$  on  $S$  such that  $\sum_{i,j} f(A(s_i, s_j))$  is minimized.

This natural optimization problem is NP-hard to solve exactly [28], but can also be approximated. Despite its natural appeal, it has no demonstrated connection to evolution. Therefore, consider the following alternative way of looking at multiple sequence alignments. This problem (described in [52] in the context of a linear gap penalty function) can be seen as an extension of the maximum parsimony optimization problem in which we allow for insertions and deletions of gaps in the evolutionary process.

### **Phylogenetic MSA**

- *Input:* A set  $S$  of sequences and a function  $f(\cdot, \cdot)$  for computing the cost of a pairwise alignment between two sequences, given as input.
- *Output:* A tree  $T$  which is leaf-labelled by the set  $S$  and with additional sequences labelling the internal nodes of  $T$  so as to minimize

$$\sum_{(v,w) \in E} f(A_{opt}(s_v, s_w)),$$

where  $s_v$  and  $s_w$  are the sequences assigned to nodes  $v$  and  $w$  respectively,  $A_{opt}(s_v, s_w)$  denotes an optimal pairwise alignment on  $s_v$  and  $s_w$ , and  $E$  is the edge set of  $T$ .

It is easy to see that the Phylogenetic MSA problem is NP-hard, since the special case where gaps have infinite cost (and hence are not permitted) is the maximum parsimony

(MP) problem, which is NP-hard. However, the Phylogenetic MSA problem is even harder than the MP problem, since it is still NP-hard even when the tree  $T$  is specified even for simple affine gap penalties. In other words, finding an optimal assignment of sequences to internal nodes of a fixed tree  $T$  is itself an NP-hard problem [65, 69] (current exact methods for this problem have running times on the order of  $O(2^n L^n)$ , where  $L$  is the maximum sequence length and  $n$  is the number of leaves in the tree) [52]. Thus, exact solutions to Phylogenetic MSA are achievable only for relatively small datasets. The fixed tree Phylogenetic MSA problem does have a PTAS (polynomial time approximation scheme) [66], but the running time needed for a good approximation is too large. Heuristics for both the fixed-tree problem and the search problem have been implemented but are not in general use (and it is not clear how well these heuristics perform).

Maximum likelihood approaches can also be used for phylogenetic multiple sequence alignment, but these require an explicit model of evolution which incorporates insertions and deletions (and perhaps also duplications) as well as site substitutions. Some such models exist, but they are in their infancy (see [62, 63]), and maximum likelihood methods under such models are likely to be extremely computationally intensive.

### **Research questions for MSA**

The main challenge here, from the point of view of phylogenetic estimation, is developing MSA techniques that are appropriate for phylogenetic reconstruction, so that accurate trees can be obtained when the input data are not yet aligned. To establish such a method, however, better models of sequence evolution (ones that include events that make a multiple alignment necessary) need to be developed, so that methods can be tested on simulated data. Such events include duplications of genes, insertions and deletions of DNA regions, and large-scale events such as inversions and transpositions. Realistic simulators should incorporate all these events, while still keeping the flexibility of the standard DNA sequence models which do not enforce molecular clocks or constant rates across sites. No simulator available today has all the flexibility needed to be of real use in testing alignment algorithms. Until we can test methods in simulation, we will not know if the Phylogenetic MSA problem is really the “right” problem to solve. We may find out that it is a good problem to solve, but that better problems may be needed.

Thus, there are really two main challenges: first, to develop good stochastic models that reflect the properties of real datasets, and then to develop methods for alignment (and perhaps simultaneous alignment and phylogeny reconstruction) that enable accurate phylogenies to be inferred. The general challenge of developing better, more biologically realistic, models of evolution applies to *all* aspects of phylogenetic inference.

#### **1.4.4 Special challenges involved in large-scale phylogenetics**

Since most approaches for estimating phylogenies involve solving hard optimization problems, phylogeny reconstruction is generally difficult, and the larger the dataset the harder the problem. This is the obvious challenge in analyzing large datasets. But certain other problems become particularly difficult when large datasets are analyzed. In particular, as mentioned before, assessing confidence in estimated phylogenies using bootstrapping becomes infeasible, unless fast reconstruction methods are used instead of computationally intensive ones.

Another challenge that comes up is storing and analyzing the set of best trees found during an analysis. Even for moderate sized datasets this can be a large number (running in the thousands), and the number of best trees for larger datasets may conceivably run into

the millions. How to store these datasets in a space-efficient manner, and so that consensus methods and other data mining techniques can be applied to the set, is still largely an open problem.

Finally, with datasets containing many taxa, the incidence of missing data and difficult multiple sequence alignment increases, thus making the usual approaches to phylogeny estimation difficult. In these cases, the systematist may wish to consider approaches for phylogeny reconstruction which first divide the full data matrix into smaller (probably overlapping) subsets (which may have less missing data, or be easier to align); such subsets should be easier to analyze phylogenetically. These smaller trees on subsets of the taxa may then be used in a *supertree analysis* (the subject of the next section) in order to obtain a phylogeny on the full dataset.

## 1.5 Special topic: Supertree methods

---

### 1.5.1 Introduction

Supertree methods attempt to estimate the evolutionary history of a set  $S$  of sequences given estimates of evolutionary history for subsets of  $S$ . Thus, supertree methods take as input a collection of trees (which may be rooted or unrooted, depending upon the method used to estimate evolution), and they produce a tree on the union of all the input leaf sets.

There are at least three situations in which supertree methods are used. Two of these situations are obvious: supertree methods may be used explicitly in a meta-analysis (merging previously computed trees on subsets of a group, and merging them into a tree on the full group), or as part of a divide-and-conquer strategy (see the chapter on Disk-Covering Methods in this volume); the use of supertree methods for meta-analyses is the more well known. Less obvious, however, is how the data may themselves indicate the potential need for a supertree analysis. For example, missing data are a serious problem which can lead to sites or taxa being eliminated from an analysis. Recall that systematists routinely use multiple markers for phylogenetic reconstruction and that often each marker will be analyzed in separate analyses. If missing data cause overlapping sets of taxa to be available for the different markers, the only way to get a phylogeny on all the taxa is to do a supertree method. Thus, datasets with a lot of missing data (i.e., gaps in the multiple alignment) may be analyzed more accurately if done in the context of a supertree method.

Each of these situations provides somewhat different challenges. For example, in most cases, a supertree method will need to be able to accept arbitrary inputs (and perform well on them). However, supertree methods used in the context of a divide-and-conquer strategy need not be designed to handle (or perform well on) arbitrary inputs, since the input trees can be assumed to have certain overlap patterns (since the divide-and-conquer strategy can produce subset decompositions that are favorable to the supertree method).

Finally, in some cases, supertree methods are not the only option for a phylogenetic analysis – *supermatrix* analyses are the other. The basic idea is to combine all the data matrices into one data matrix. If the submatrices do not all share the same taxa, then some characters will be missing states for some taxa. In this case, when the supermatrix is created, the missing entries are simply coded as “missing”. Phylogeny reconstruction methods are generally adapted for such data, since missing data are fairly common, and so the newly created supermatrix can then be given as input to a phylogenetic reconstruction (In this case, however, a reconstruction method may keep track of the way the new supermatrix is composed of submatrices, so that different stochastic models can be used on the different parts of the supermatrix during the estimation of the phylogeny.)

The utility of a supertree method must always be considered in comparison to the obvious

supermatrix analysis; both approaches have theoretical advantages and disadvantages, and the relative quality of these approaches is not yet known. Note, however, that a supermatrix approach is not possible for meta-analyses since in these cases, the trees will be available, but the original character data may not be.

### 1.5.2 Tree Compatibility

The most obvious computational problem related to supertree construction is to determine if a collection of trees is compatible, and if so, to construct a supertree consistent with all the input trees. We now make these concepts precise.

#### Terminology

We begin by defining tree compatibility when all the trees are on the same set.

Let  $T$  and  $T'$  be two trees on  $S$ . Then  $T$  is said to *refine*  $T'$  if  $T'$  can be obtained from  $T$  by a sequence of edge-contractions. Thus, every tree refines the star-tree on the same set of leaves, and if  $T$  refines  $T'$  and both are binary trees (or more simply have the same number of edges), then  $T = T'$ . Finally, trees  $T$  and  $T'$  on the same leaf set are *compatible* if there is a tree  $T''$  such that  $T''$  refines both  $T$  and  $T'$ . Note that if  $T''$  exists, it may be equal to  $T$  or  $T'$ .

These definitions can be extended to the case where the input trees are not on the same leaf set, by considering trees restricted to subsets of their leaf sets. We can restrict a tree  $T$  to a subset  $A$  of its leaves in the obvious way: including only the subtree of  $T$  connecting the leaves in  $A$  and suppressing nodes of degree two. The resultant tree is denoted by  $T_A$ . We now define tree compatibility.

#### Tree compatibility

- **Input:** Set  $\mathcal{T} = \{T_1, T_2, \dots, T_k\}$  of trees on sets  $S_1, S_2, \dots, S_k$ , respectively.
- **Output:** Tree  $T$ , if it exists, such that for each  $i$ ,  $T|_{S_i}$  refines  $T_i$ .

This tree compatibility problem is NP-hard [55], and so is difficult to solve. But algorithms for this problem are likely to fail to be useful because most inputs will simply not be compatible. Error in phylogenetic estimations are to be expected, and so supertree methods need to be able to handle inputs with errors.

Now consider the case where the input trees are rooted. In this case, we are looking for a rooted supertree which is consistent with all the inputs – thus, the rooted trees must be correct with respect to the location of their roots, as well as topologically. This is even less likely to be the case, since locating roots is itself difficult. Hence, again this approach is not likely to be useful. On the other hand, when the trees are rooted the compatibility problem is polynomial [1].

### 1.5.3 Matrix Representation Parsimony

Since estimates of evolutionary trees may not be completely correct, supertree methods need to be able to handle incompatibility in their inputs. Fortunately, there are approaches for constructing supertrees from incompatible trees. The most popular method is *Matrix Representation Parsimony* (MRP). This method can be applied to rooted or unrooted trees, and uses maximum parsimony to analyze the data matrix that it creates. The technique used to create the data matrix depends upon whether the trees are rooted or unrooted.

If the trees are unrooted, then each input tree is replaced by a datamatrix of partial binary characters on the entire set of taxa, where by “partial binary character” we mean

characters whose states are 0, 1, and ?, where the ? indicates that the state is missing for that taxon. An example will make this clear. Let  $T$  be one of the trees, and let  $T$  have leaf set  $A \subset S$ . Let  $e \in E(T)$  be an edge in  $T$ , and let  $A_0|A_1$  be the bipartition on  $A$  created by deleting  $e$  (but not its endpoints) from  $T$ . Then  $e$  is represented by the partial binary character  $c_e$  defined by  $c_e(a) = 0$  if  $a \in A_i$ ,  $i = 0, 1$ , and  $c_e(s) = ?$  if  $s \in S - A$ . Note that  $c_e$  can be defined in two ways, depending upon the definition of  $A_0$  and  $A_1$ ; however, this is irrelevant to the computation that ensues. Thus  $T$  is replaced by the data matrix on  $S$  with character set  $\{c_e : e \in E(T)\}$ . We represent each tree in the profile by its data matrix, and concatenate all the data matrices.

The same representation is used for rooted trees, except that the characters  $c_e$  are defined in such a way that the part of the tree containing the root is assigned the zero state. Also, an additional row consisting of all zeros is added to the datamatrix that is created to represent the root, and the trees that are obtained are rooted at this added node. In this way, MRP can be used to analyze either unrooted or rooted trees.

This concatenated data matrix is then analyzed using maximum parsimony, and the output of the maximum parsimony method is returned. Note that if all the trees are compatible (meaning that they can be combined into one supertree  $T$  without loss of accuracy), then this technique will construct the correct supertree (along with any other supertree which is consistent with all the input trees). However, because MRP involves solving an NP-hard problem, its running time is not generally efficient. Furthermore, most inputs will have some error, and it is not clear how errors in the input trees affect the quality of the MRP tree. Thus, despite the nice theoretical property of MRP, its performance in practice is less clear.

#### 1.5.4 Other supertree methods

Finally, other supertree methods also exist which can handle unrooted or rooted trees which have errors. For example, Gordon's strict consensus supertree [19] method is an interesting method, but not in general use. Another method quite similar to Gordon's strict consensus supertree method is used in the "Disk-Covering Methods" (DCMs) described in this volume. These DCMs are divide-and-conquer methods used to speed up maximum parsimony analyses, as well as improve other phylogenetic reconstruction methods. Both of these supertree methods are guaranteed to solve tree compatibility *if* the input trees are "big enough" (i.e. have enough overlap) and are correct.

#### 1.5.5 Open problems

Despite the potential for supertree methods to be useful, and the interest in them (see the book chapter in this volume on supertrees, and also [3]), little is really known about how well they work on real data analyses. Thus, the main research that needs to be done is to evaluate how well they work in comparison to each other, and also in comparison to supermatrix approaches. In the likely event that current approaches are not able to produce high quality supertrees, new methods should be developed. As a first step, it is likely that new optimization problems should be developed. As these are likely to be NP-hard (as almost everything in phylogeny estimation is), heuristics for these problems will need to be developed.

### 1.6 Special topic: Genomic phylogeny reconstruction

---

In the previous section we described phylogenetic inference when the input is a set of aligned molecular (DNA, RNA, or amino-acid) sequences. In this section, we will discuss phylogenetic inference for a different kind of data – whole genomes, in which the information is the order and strandedness of the genes within the genomes.

Just as evolution changes the individual nucleotides within gene sequences, other events also take place which affect the “chromosomal architecture” of whole genomes. Some of these events, such as inversions (which pick up a region and replace it in the same location but in the reverse order, and on the opposite strand) and transpositions (which move a segment from one location to another within the same chromosome) change the order and strandedness of genes within individual chromosomes; others, such as translocations (which move genomic segments from one chromosome to another) duplications, insertions, and deletions can change both the order and also the number of copies of a given gene within a chromosome. Finally, fissions (which split a chromosome into two) and fusions (which merge two chromosomes) change the number of chromosomes within a genome.

All these events are less frequent than individual point mutations, and so have inspired biologists to consider using gene orders as a source of phylogenetic signal in the hope that they might allow evolutionary histories to be reconstructed at a deeper level than is typically possible using molecular phylogenetics. On the other hand, much less is really understood about how genomes evolve, and so the statistical models describing the evolution of whole genomes are not as well developed. Equally problematic is the fact that from a computational standpoint, whole genome phylogenetic analysis is much more difficult than comparable approaches in molecular phylogenetics.

Research in gene order phylogeny has largely focused on two basic approaches: parsimony-style methods that seek to find trees with a minimum total “length”, and distance-based methods. Both types of approaches require the ability to compute either edit distances between two genomes (that is, determining the minimum number of events needed to transform one gene order into another) or estimating true evolutionary distances (i.e., estimating the actual number of events that occurred in the evolutionary history between the two gene orders). Algorithms to compute edit distances tend to involve graph-theoretic algorithms (see [22, 23, 24] for initial work in the area) whereas algorithms to estimate true evolutionary distances involve probability and statistics (see [35, 67, 68]).

However, despite the interest in these problems, not much has really been established. First, we can only solve problems for the special case in which all genomes have exactly one chromosome and exactly one copy of each gene. In other words, the only events that are permitted are inversions and transpositions, since all other events change the gene content. Even for this special case, however, many computational problems are known or conjectured to be NP-hard. For example:

- Computing the inversion distance is solvable in polynomial time [2, 23].
- True evolutionary distances (under standard Markov models) can be estimated in polynomial time using statistical techniques [35, 67, 68],
- Computing the transposition distance is of unknown computational complexity.
- Computing the inversion median of three genomes is NP-hard [7].
- Computing the breakpoint median of three genomes is NP-hard (though it reduces to the well studied travelling salesperson problem, and hence can often be solved quickly in practice) [51]

In other words, under even fairly idealized conditions (where the only events are inversions and transpositions), most optimization problems are hard to solve. Heuristics without

proven performance guarantees have been developed for these idealized conditions [20], but even these only have good performance under certain conditions.

### ***Research questions in whole genome phylogeny***

There are essentially two main obstacles for whole genome phylogeny. The first is that almost all the methods that have been developed (and their stochastic models) assume that all the genomes have exactly one copy of each gene (an assumption that is widely violated), and the second is that despite a fair amount of effort, we still do not have methods that can reliably analyze even moderately large datasets, even under these idealized conditions. Thus, work in both directions needs to be done.

The problem with the assumption that all genomes have one copy of each gene is that events that change gene content (such as insertions, deletions, and duplications) occur in enough datasets to make the current approaches inapplicable without some preprocessing. Thus, in fact, essentially all the datasets that have been analyzed using existing methods that assume equal gene content are first processed to remove duplicate genes. While in some cases this processing is acceptable (and should not change the phylogenetic reconstruction), it is not always clear how to do this rigorously (and in any event, throwing out data is almost always not desirable). Thus, making progress on developing new stochastic models that incorporate these events that change gene content is important, and will allow us to then test the performance of methods that we develop to infer evolutionary history from the full range of data.

Also, the usual stochastic models of gene order evolution make assumptions that all events of the same type are equiprobable, so that (for example) any two inversions have the same probability. However, research now suggests that “short inversions” may have a higher probability than “long inversions”. Incorporating these changed assumptions into phylogenetic inference changes the computational problems in interesting ways. For example, instead of edit distances we would have weighted edit distances (so the cost of an event would reflect its probability), and estimations of true evolutionary distances would also need to be changed to reflect the additional complexity of the model. While some progress has been made towards estimating these distances [58], much still needs to be done.

Finally, methods for whole genome phylogeny reconstruction are quite computationally intensive – more so than the corresponding problems for DNA sequence phylogenetics by far. For example, computing the inversion length of a fixed tree on just three leaves can take a long time on some instances! While some progress has been made to provide speed-ups for whole genome phylogeny so that they can analyze large datasets, so far these speed-ups are limited to datasets with certain properties (no long edges, in particular). Therefore, a natural research area is to develop techniques for handling large datasets which do not have as significant limitations as the current set of methods.

## **1.7 Conclusions and suggestions for further reading**

---

Research into methods for phylogeny reconstruction offers surprisingly deep and interesting challenges to algorithms developers. Yet understanding the data, the methods, and how biologists use phylogenies is necessary in order for the development to be productive. We hope this chapter will help the reader appreciate the difference between pure algorithmic research, and that which could make a tremendous difference to practice.

There is a wealth of books and papers on phylogenetics, from all the different fields (biology, statistics, and computer science). The following list is just a sample of some of these books and papers, that will provide additional grounding in the field of computational

and mathematical phylogenetics.

For more information from the perspective of a systematist, see [14, 60]. For books with a greater emphasis on mathematical and/or computational aspects, see [17, 53]. Expositions with a greater emphasis on stochastic models can be found in [21, 38]. Texts that are intermediate between these include [45]. For an on-line tutorial on phylogenetics, see [30].

## 1.8 Acknowledgments

---

This work was supported in part by the National Science Foundation, the David and Lucile Packard Foundation, and the Institute for Cellular and Molecular Biology at the University of Texas at Austin. Special thanks are due to Martin Nowak, and to the Program for Evolutionary Dynamics at Harvard University which supported both the authors for the 2004-2005 academic year. The authors also wish to thank Erick Matsen, Mike Steel, and Michelle Swenson, for their comments on earlier drafts of this chapter.

## References

---

### References

- [1] A. Aho, Y. Sagiv, T. Szymanski, and J. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. In *Proc. 16th Ann. Allerton Conf. on Commun., Contr. and Comptg.*, pages 54–63, 1978.
- [2] D. A. Bader, B.M.E. Moret, and M. Yan. A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. *J. Comput. Biol.*, 8(5):483–491, 2001.
- [3] O.R.P. Bininda-Emonds, editor. *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, volume 3 of *Computational Biology*. Kluwer Academics, 2004.
- [4] D.R. Brooks and D.A. McLennan. *Phylogeny, Ecology, and Behavior*. University of Chicago Press, Chicago, 1991.
- [5] D. Bryant. A classification of consensus methods for phylogenetics. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 1991.
- [6] R.M. Bush, C.A. Bender, K. Subbarao, N.J. Cox, and W.M. Fitch. Predicting the evolution of human influenza A. *Science*, 286:1921–1925, 1999.
- [7] A. Caprara. Formulations and complexity of multiple sorting by reversals. In S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB-99)*, pages 84–93, 1999.
- [8] J. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.
- [9] J. T. Chang. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, 134:189–215, 1996.
- [10] B. Chor and T. Tuller. Maximum likelihood of evolutionary trees is hard. In *RECOMB 2005*, 2005.
- [11] S.N. Evans and T. Warnow. Unidentifiable divergence times in rates-across-sites models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:130–134, 2005.

- [12] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.
- [13] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.
- [14] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.
- [15] W. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Biology*, 20:406–416, 1971.
- [16] L. R. Foulds and R. L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3:43–49, 1982.
- [17] O. Gascuel, editor. *Mathematics of Evolution and Phylogeny*. Oxford Univ. Press, 2005.
- [18] P.A. Goloboff. Analyzing large data sets in reasonable times: solution for composite optima. *Cladistics*, 15:415–428, 1999.
- [19] A. D. Gordon. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *J. Classif.*, 3:335–348, 1986.
- [20] GRAPPA (genome rearrangements analysis under parsimony and other phylogenetic algorithms). <http://www.cs.unm.edu/~moret/GRAPPA/>.
- [21] D. Grauer and W.-H. Li. *Fundamentals of Molecular Evolution*. Sinauer Publishers, 2000.
- [22] S. Hannenhalli and P.A. Pevzner. Towards computational theory of genome rearrangements. *Computer Science Today: Recent Trends and Developments. Lecture Notes in Computer Science*, 1000:184–202, 1995. Invited paper in a jubilee volume.
- [23] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. of the 27th Annual Symposium on the Theory of Computing (STOC 95)*, pages 178–189, 1995. Las Vegas, Nevada.
- [24] S. Hannenhalli and P.A. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problem). In *Proc. of the 36 Annual Symposium on Foundations of Computer Science (FOCS 95)*, pages 581–592, 1995. Milwaukee, Wisconsin.
- [25] P.H. Harvey and M.D. Pagel. *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, 1991.
- [26] J.P. Huelsenbeck, J.J. Bull, and C.W. Cunningham. Combining data in phylogenetic analysis. *Trends in Ecology and Evolution*, 11(4):152–158, 1996.
- [27] J.P. Huelsenbeck and R. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.
- [28] W. Just. Computational complexity of multiple sequence alignment with SP-score. *Journal of Computational Biology*, 8(6):615–623, 2001.
- [29] S. Kannan, T. Warnow, and S. Yooseph. Computing the local consensus of trees. *SIAM J. Computing*, 27(6):1695–1724, 1995.
- [30] J. Kim and T. Warnow. Tutorial on phylogenetic tree estimation, 1999. Presented at the ISMB 1999 conference, available on-line at <http://kim.bio.upenn.edu/jkim/media/ISMBtutorial.pdf>.
- [31] C. Randal Linder and B.M.E. Moret. Presented at the dimacs workshop on reticulate evolution. the slides can be obtained at [http://dimacs.rutgers.edu/workshops/reticulated\\_wg/slides/slides.html](http://dimacs.rutgers.edu/workshops/reticulated_wg/slides/slides.html).
- [32] B. Ma, M. Li, and L. Zhang. On reconstructing species trees from gene trees in terms of duplications and losses. In *Proc. 2nd Ann. Int'l Conf. Comput. Mol. Biol. RECOMB98*, 1998.
- [33] W. Maddison. Gene trees in species trees. *Syst. Bio.*, 46(3):523–536, 1997.

- [34] M.L. Metzker, D.P. Mindell, X.M. Liu, R.G. Ptak, R.A. Gibbs, and D.M. Hillis. Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14292–14297, 2002. OCT 29.
- [35] B. Moret, L.-S. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies based on gene order. In *Proceedings of 9th Int'l Conf. on Intelligent Systems for Molecular Biology (ISMB'01)*, pages 165–173, 2001.
- [36] B.M.E. Moret, L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Transactions on Computational Biology and Biocomputing*, 1(1), 2004.
- [37] L. Nakhleh, J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, and A. Tholse. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. In *Proc. 8th Pacific Symp. on Biocomputing (PSB 2003)*, 2003.
- [38] M. Nei, Sudhir Kumar, and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, 2003.
- [39] K.C. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407–414, 1999.
- [40] S.P. Otto and J. Whitton. Polyploid incidence and evolution. *Annual Review of Genetics*, 34:401–437, 2000. ANNU REV GENET.
- [41] R. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, 43:58–77, 1994.
- [42] R. Page. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinf.*, 14(9):819–820, 1998.
- [43] R. Page and M. Charleston. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phyl. Evol.*, 7:231–240, 1997.
- [44] R. Page and M. Charleston. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzehtsky, editors, *Mathematical hierarchies in biology*, volume 37. American Math. Soc., 1997.
- [45] R. Page and E. Holmes. *Molecular Evolution: A phylogenetic approach*. Blackwell Publishers, 1998.
- [46] C.A. Phillips and T. Warnow. The asymmetric median tree: a new model for building consensus trees. *Discrete Applied Mathematics*, 71:311–335, 1996.
- [47] D. Posada and K.A. Crandall. Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14(9):817–818, 1998.
- [48] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
- [49] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. IEEE Computer Society Bioinformatics Conference CSB 2004*, 2004. Stanford University.
- [50] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
- [51] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.
- [52] D. Sankoff and R.J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. In *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*, pages 253–264. Addison-Wesley, 1983. Chapter 9.
- [53] C. Semple and M. Steel. *Phylogenetics*. Oxford Series in Mathematics and its Applications, 2004.

- [54] M. A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology*, 43(4):560–564, 1994.
- [55] M.A. Steel. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9:91–116, 1992.
- [56] M.A. Steel and D. Penny. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.*, 17(6):839–850, 2000.
- [57] M.A. Steel, L.A. Székely, and M.D. Hendy. Reconstructing trees when sequence sites evolve at variable rates. *J. Comp. Biol.*, 1:153–163, 1994.
- [58] K.M. Swenson, M. Marron, J.V. Earnest-DeYoung, and B.M.E. Moret. Approximating the true evolutionary distance between two genomes. In *Proc. 7th Workshop on Algorithm Engineering and Experiments (ALENEX'05), Vancouver (Canada)*. SIAM Press, 2005.
- [59] D. L. Swofford. PAUP\*: Phylogenetic analysis using parsimony (and other methods), 1996. Sinauer Associates, Underland, Massachusetts, Version 4.0.
- [60] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B.K. Mable, editors, *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts, 1996.
- [61] J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
- [62] J.L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124, 1991.
- [63] J.L. Thorne, H. Kishino, and J. Felsenstein. Inching towards reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16, 1992.
- [64] C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of Mathematical Biology*, 59:581–607, 1997.
- [65] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.
- [66] L. Wang, T. Jiang, and D. Gusfield. A more efficient approximation scheme for tree alignment. *SIAM J. Computing*, 30(1):283–299, 2000.
- [67] L.-S. Wang. Exact-IEBP: A new technique for estimating evolutionary distances between whole genomes. In *Lecture Notes for Computer Sciences No. 2149: Proceedings for the First Workshop on Algorithms in BioInformatics (WABI'01)*, pages 175–188, 2001.
- [68] L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proceedings of the Thirty-Third Annual ACM Symposium on the Theory of Computing (STOC'01)*, pages 637–646, 2001.
- [69] H. T. Wareham. A simplified proof of the NP- and MAX SNP-hardness of multiple sequence tree alignment. *Journal of Computational Biology*, 2(4):509–514, 1995.
- [70] T. Warnow, B. M. Moret, and K. St. John. Absolute convergence: true trees from short sequences. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA 01)*, pages 186–195, 2001.
- [71] M. Wilkinson. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Syst. Biol.*, 43(3):343–368, 1994.
- [72] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1996.

# Index

---

- allele, 1-8
- assessing support for a phylogeny, 1-15–1-18
  - Bayesian MCMC posterior probabilities, 1-17
  - bootstrap support, 1-16–1-17
  - consensus methods, 1-16
  - jackknife support, 1-17
- character, 1-3
- Clustal, 1-12
- combined dataset, 1-8, 1-12, 1-13, 1-28
- concerted evolution, 1-8
- consensus tree, 1-16
- convergent evolution, 1-11
- designing a phylogenetic study, 1-5–1-9
  - marker selection, 1-7–1-9
  - taxon selection and sampling, 1-5–1-7
- gene order phylogeny reconstruction, 1-31–1-32
- gene tree/species tree problem, 1-7, 1-11
- hybrid speciation, 1-2
- identifiability, 1-14
- ingroup, 1-6
- methods of phylogenetic reconstruction
  - distance-based, 1-15
- methods of phylogenetic reconstruction
  - distance-based, 1-15
  - Markov Chain Monte Carlo, 1-15
  - maximum likelihood, 1-15
  - maximum parsimony, 1-15
  - neighbor joining, 1-15
- multiple sequence alignment, 1-11–1-12
- outgroup, 1-6
- performance analysis of algorithms, 1-18–1-20
  - benchmark datasets, 1-18
  - Robinson-Foulds metric, 1-19
  - simulation studies, 1-19
- phylogenetic markers, 1-5
- phylogenetic reconstruction on molecular sequences, 1-20–1-25
  - heuristic searches, 1-20
- point mutations, 1-13
- positional homology, 1-11
- rates across sites, 1-13
- recombination, 1-7
- repetitive elements, 1-9, 1-12
- reticulate evolution, 1-2, 1-11
- statistical consistency, 1-15
- stochastic models of evolution, 1-13–1-15
  - Generalized Time Reversible, 1-13, 1-14
  - Jukes-Cantor, 1-13, 1-14, 1-22
- supermatrix methods, 1-28
- supertree methods, 1-9, 1-28–1-30
- testing models of evolution, 1-15
- uses of phylogenies, 1-3–1-4
  - biogeography, 1-3
  - comparative method, 1-3
  - disease evolution, 1-4
  - modes of speciation, 1-4
  - protein reconstruction, 1-4
  - tempo of speciation, 1-4