Computational Challenges from the Tree of Life

B.M.E Moret*

Abstract

A phylogeny is a reconstruction of the evolutionary history of a group of organisms. Phylogenies are used throughout the life sciences, as they offer a structure around which to organize the knowledge and data accumulated by researchers. Computational phylogenetics has been a rich area for algorithm design over the last 15 years.

The biology community has embarked on an enormously ambitious project, the assembly of the Tree of Life—the phylogeny of all organisms on this planet. This project presents a true computational grand challenge: current phylogenetic methods can barely handle a few hundred organisms, yet the Tree of Life has an estimated 10–100 million organisms. Thus, while data collection and cataloguing is one of the major tasks, algorithm design and engineering is a crucial enabling component of the project.

In this paper, I briefly introduce the principles of phylogenetic analysis, then discuss the computational challenges posed by the Tree of Life project, from the design and validation of computationally useful models of evolution to the actual computation and assessment of the Tree of Life itself.

1 Introduction

The life sciences, and especially the study of evolution, have been almost completely redefined by modern information technology, both in terms of data acquisition (e.g., new genomic data accumulate at a rate exceeding Moore's law) and in terms of analysis (e.g., the literature shows over $10,000$ citations to the top three phylogenetic software packages, with exponentially growing rates).

The study of evolution is the foundation of the life sciences: biological knowledge, much of which consists of large amounts of data, is organized through an understanding of evolutionary relationships between organisms at every level, from DNA data to epidemiology and population ecology. The broad-scale history of genetic descent during organismal evolution takes the form of a single, enormous "Tree of Life"—see Figure 1 for a high-level draft of such a tree. This phylogeny stands as one of science's great discoveries. Its implication—that all living things on Earth today (from bacteria, to mushrooms, to humans) are related—has forever changed our perception of the world around us. Over the last 30 years, biologists have come to embrace reconstruction of this phylogeny as a major research goal [23, 34, 60, 63]. The use of phylogenetic principles is almost as ubiquitous today as the idea of Darwinian evolution. Phylogeneticists have formulated specific models and questions that can now be addressed using recent advances in database technology and optimization algorithms.

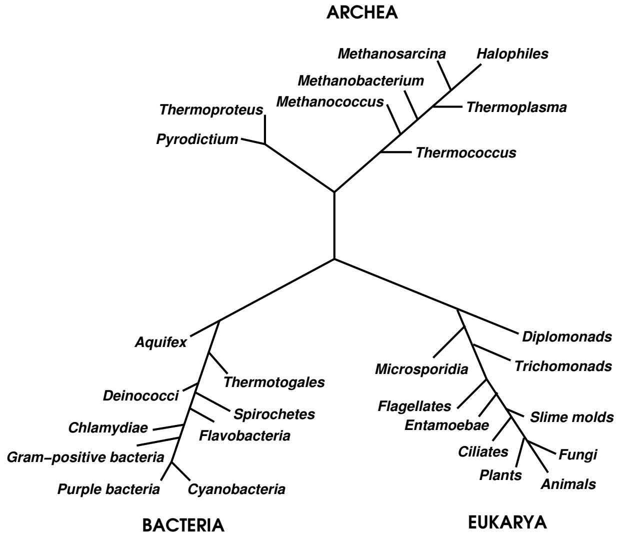

Figure 1: A rough sketch of the Tree of Life, showing only some of the main branches (after [23]).

2 Phylogenies: What and Why?

A phylogeny is a reconstruction of the evolutionary history of a group of organisms or other entities subject to evolution—these entities are usually referred to as *taxa*. Because evolution traces the descent of contemporary characteristics from common ancestors, a phylogeny usually takes the form of a tree, although certain evolutionary events, such as hybridization, may cause it to assume the form of a directed acyclic graph. Figures 2 and 3 illustrate phylogenies from the domains of epidemiology (the evolution and spread of the West Nile encephalitis virus) and virology (the relationships among herpes viruses that affect humans). Note that the first of these phylogenies is rooted (it has an implied evolutionary flow from left to right), whereas the second (like the sketch of the Tree of Life in Figure 1) is not. As we shall see, most phylogenetic reconstruction methods produce unrooted trees. Phylogenies are reconstructed

[*] Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA, moret@cs.unm.edu

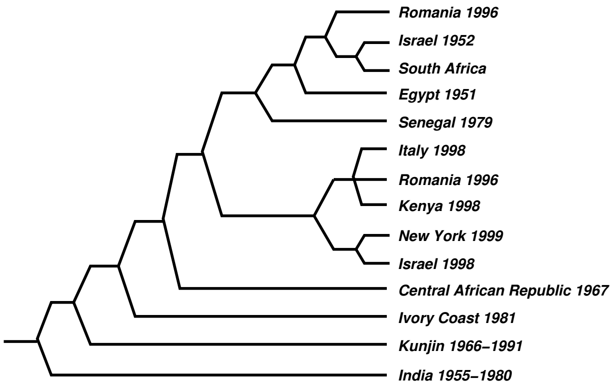

Figure 2: Epidemiology of the West Nile encephalitis virus (after [72]).

using data of all kinds, from molecular data (DNA sequences) through whole-genome data, metabolic data, morphological data, to geographical and geological data.

Because it reflects the history of transmission of life's genetic information, phylogeny has unique power to organize our knowledge of diverse organisms, genomes, and molecules. A reconstructed phylogeny guides our interpretation of the evolution of organismal characteristics, indicating in what lineages traits arose and under what circumstances, thus playing a vital role in studies of adaptation and evolutionary constraints [27, 32, 58, 75, 78, 82, 83, 137]. Patterns of divergence of species lineages indicated by the phylogeny inform us of the dynamics of speciation and extinction, the forces that generate and reduce biodiversity [20, 37], including the assembly and maintenance of species in ecological communities [136]. Phylogeny informs far more than evolutionary biology, however. The evolutionary histories of genes bear the marks of the functional demands to which they have been subjected, thus allowing phylogenetic analysis to elucidate functional relationships within living cells [38, 45, 139]. Thus, for instance, pharmaceutical companies are increasingly using phylogenetic analyses to make functional predictions from sequence data banks of gene families [4], for ligand prediction [21, 126, 140], and in the development of vaccines [50] and antimicrobials and herbicides [12, 98, 109, 138]. Molecular biologists use phylogenies to determine the relevance of model organisms [16, 30, 36, 62, 92]. Phylogenetic analysis is also implicitly used in the inference of secondary structure of RNAs [17, 40, 48, 49], as well as in predicting the structure of proteins or making proteins in the lab [22, 40]. Finally, phylogenetic reconstruction is used well beyond the boundaries of biology and biomedicine: it is a crucial tool in forensic studies (see, for instance, the dentist case [61, 97], the HIV murder trial [84], and recent work on the anthrax terrorist attacks), in security applications for networks and computers, and in a variety of disciplines such as historical linguistics [135].

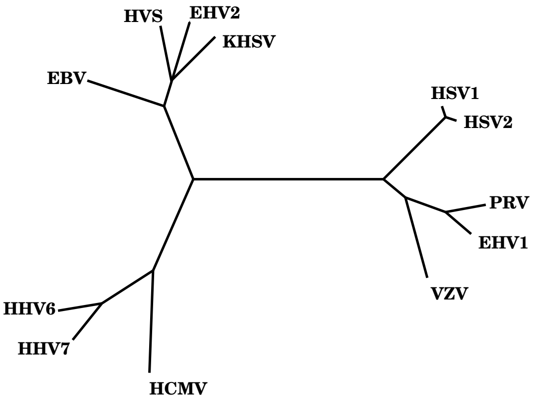

Figure 3: Species of herpes viruses that affect humans (after [85]).

3 Phylogenetic Reconstruction

Reconstructing a phylogeny for a group of taxa requires data about these taxa, a model of evolution for the data and for this group of taxa, and an algorithm. The model of is necessity simplified, as the reconstruction algorithm must, implicitly or explicitly, invert the model; and the algorithm is either *ad hoc* or an approximation algorithm for a difficult optimization problem.

3.1 Data While many types of data have been, and continue to be, used, the dominant choice today is molecular data [125]—typically, the DNA sequences of a few genes. Molecular data have the significant advantage of being exact and reproducible, at least within experimental error, not to mention fairly easy to obtain. Each nucleotide in a DNA or RNA sequence (or each codon) is, by itself, a well defined *character*, whereas morphological data, for instance, must first be encoded into characters, with all the attending problems of interpretation, discretization, etc. While genomic sequences (nucleotides or codons) remain the main source of molecular data, promising new types of genomic data are appearing, most notably gene rearrangement data [89].

In sequence data, characters are individual positions in the string and so can assume one of a few states: 4 states for nucleotides or 20 states for amino-acids. Such data evolve through *point mutations*, i.e., changes in the state of a character, plus *insertions* and *deletions* (or just *indels* for short)—the three main string editing operations in conventional string algorithms [47].

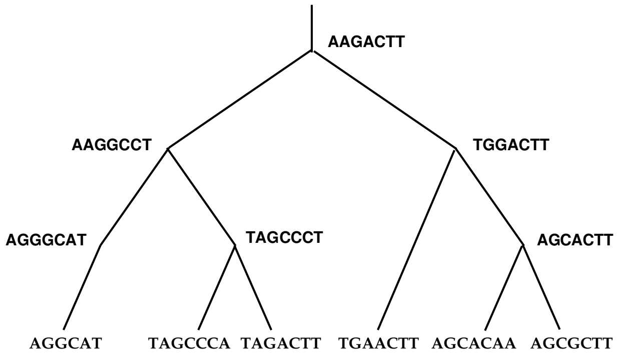

Figure 4: Sequence evolution: the sequences at the leaves are modern data evolved from the common ancestral sequence *AAGACTT*.

Figure 4 illustrates the concept. Sequence data are by far the most common form of molecular data used in phylogenetic analyses. Large amounts of sequence data are easily available from databases such as GenBank, along with search tools and annotations; moreover, the volume of such data grows at an exponential pace. Many analysis tools have been developed for such data: packages such as `PAUP*` [124], MacClade [77], Mesquite [80], Phylip [33], MEGA [70], MrBayes [64], and TNT [41], all available either freely or for a modest fee, are in widespread use.

Sequence data suffer from some problems. The relatively fast pace of mutation in many regions of the genome, combined with the fact that each character can assume one of only a few states, results in *silent* changes—changes that are subsequently reversed in the course of evolution, leaving no trace in modern organisms. Sequence data must therefore be selected to fit the problem at hand: stable regions to reconstruct old events, and variable regions to reconstruct recent history. Most importantly, the evolution of any given gene (or region of the sequence) need not be identical to that of the organism: this is the *gene tree vs. species tree* problem [79, 100]. Because of that problem, an analysis that uses all available genes risks running into internal contradictions, while one based on individual genes will typically yield different trees for the different genes, trees that must then be reconciled through a process known as *reconciliation* or *lineage sorting* [76, 99, 100, 102]. Sequence data also suffer from computational problems: most prominently, the problem of multiple sequence alignment is currently only poorly solved—indeed, most systematists will align sequence data by hand, or at least edit by hand the alignments proposed by the software.

The newer gene rearrangement data consist of lists of genes in the order in which they are placed along one or more chromosomes. Each gene along a chromosome is identified by some number, a number shared with its *homologs* (i.e., similar genes that are expected to have evolved from a common ancestral gene) on other chromosomes or, for that matter, on the same chromosome in case of gene duplications. In a signed gene order, each number is signed to indicate the strandedness of the gene. The entire gene order thus forms a *single* character that can assume a huge number of states. In addition to insertions (including gene *duplications*) and deletions, which modify the gene content, a gene order evolves through rearrangements, which leave the gene content unchanged. A rearrangement acts on a contiguous fragment of the chromosome (a piece of the gene order); known rearrangements are *inversions*, sometimes also called reversals (well documented in chloroplast organelles [67, 101]), and *transpositions* (strongly suspected in mitochondria [10, 11]). For instance, an inversion from the third to the fifth position causes the following change:

$$(1\,2\,3\,4\,5\,6\,7) \longrightarrow (1\,2\,\text{-}5\,\text{-}4\,\text{-}3\,6\,7)$$

and a transposition of the second and third elements to the sixth position in turn causes the following change:

$$(1\,2\,\text{-}5\,\text{-}4\,\text{-}3\,6\,7) \longrightarrow (1\,\text{-}4\,\text{-}3\,6\,2\,\text{-}5\,7)$$

Finally, in the case of genomes with multiple chromosomes, additional operations include *fusion* and *fission* of chromosomes, as well as *translocations*, which move a piece of a chromosome to another chromosome (in effect, they are transpositions where the target locations reside in a different chromosome).

The use of gene-content and gene-order data in phylogenetic reconstruction is attractive for several reasons: (i) because the entire genome is studied at once, there is no gene tree vs. species tree problem; (ii) there is no need for alignment; and (iii) gene rearrangements and duplications are "rare genomic events" in the sense of Rokas and Holland [110] and thus enable us to trace evolution farther back than sequence data. On the other hand, we have so far very little whole-genome data, are gathering such data at a relatively slow pace, and lack good models for the evolution of gene content and gene order. Moreover, the mathematics of gene orders is far more complex than that of DNA sequences—we have far more open problems than results and most results are proofs of NP-hardness for relatively simple tasks, such as computing the edit distance between two gene orders.

3.2 Models Most algorithms for phylogenetic reconstruction attempt to reverse a given *model of evolution*, which embodies certain knowledge and assumptions about the process of evolution, such as characteristics of speciation and details about evolutionary changes that affect the content of molecular sequences. Models of evolution vary in their complexity. For instance, the Jukes-Cantor model [68], which assumes that all characters evolve identically and independently and that all substitutions are equally likely, requires just one parameter per edge of the tree, viz., the expected number of changes of a random site on that edge; thus a rooted Jukes-Cantor tree with n leaves requires $2n-2$ parameters. Under more complex models of evolution, the process operating on a single edge can require up to 12 parameters for nucleotide data, far more for codon data. These parameters describe how a single site evolves down the tree and so require additional assumptions in order to describe how different sites evolve. Usually the sites are assumed to evolve independently; sometimes they are also assumed to evolve identically. Moreover, the different sites are assumed either to evolve under the same process or to have rates of evolution that vary depending upon the site. For more on stochastic models of (sequence) evolution, see [31, 69, 74, 125]. Tree generation models typically have parameters regulating speciation rates, but also inheritance characteristics, etc. For an interesting discussion of models of tree generation, see [56, 86].

By studying the performance of methods under explicit stochastic models of evolution, we can assess the relative strengths of different methods, as well as understand how the methods can fail. Such studies can be theoretical, for instance proving *statistical consistency*: given long enough sequences, the method will return the true tree with arbitrarily high probability. Others can use simulations to study the performance of the methods under conditions closely approximating practice. In a simulation, sequences are evolved down different model trees and then given to different methods for reconstruction; the reconstructions can then be compared against the model trees that generated the data. Such studies provide important quantifications of the relative merits of phylogenetic reconstruction methods. Thus, from an algorithm engineering point of view, models of evolution play two crucial roles: (i) inverting them forms the basis of most reconstruction algorithms; and (ii) they provide the datasets needed to assess and refine the performance of the algorithms.

3.3 Algorithms Algorithms for phylogenetic reconstructions can be roughly partitioned into two categories: *distance-based* methods, which operate from a pairwise distance matrix and typically only produce a tree, and *criterion-based* methods, which attempt to optimize a selected criterion and typically infer additional data (such as character states at internal nodes or parameter values for the model). In addition, *meta-methods* have been proposed that decompose the dataset into smaller subsets, construct trees on these subsets using a base method from the previous two categories, and then combine the resulting trees into a phylogeny for the entire dataset.

3.3.1 Distance-based methods These methods first estimate pairwise distances between every pair of taxa, then rely solely on the matrix of pairwise distances to compute an edge-weighted tree. The statistical consistency (if any) of these methods requires that a statistically consistent distance estimator and an appropriate distance-based algorithm be used. The distance estimator should return a value that approaches the expected number of times a random site changes on the path between the two taxa: thus, the estimation of pairwise distances must be done with respect to some assumed stochastic model of evolution. Naïvely defined distances, such as the Hamming distance, typically underestimate the number of changes that took place in the evolutionary history; thus the first step of a distance-based method is to *correct* the naïvely defined distance into one that accurately accounts for the expected number of unseen back-and-forth changes in a site. Such corrections are not without problems: as the measured distance grows larger, the variance in the estimator increases, causing increasing errors in reconstruction. The simplest and most commonly used distance-based method is *neighbor-joining (NJ)* [113]; improved variants include BioNJ [39] and *Weighbor* [13]. NJ is known to be statistically consistent under most models of evolution.

3.3.2 Criterion-based methods These methods attempt to optimize a criterion in order to approach the "truth," i.e., the actual evolutionary history. One widely used criterion is *parsimony* (also appearing under a slightly different guise as minimum evolution). Parsimony-based methods seek to solve the *maximum parsimony (MP)* problem: find the tree, along with character sequences labelling its internal nodes, that together minimize the total number of evolutionary changes (viewed as distances summed along all edges of the tree). This problem is NP-hard [25]; its point estimation version (given a fixed tree, find the labelling for its internal nodes that optimizes the criterion), however, is solvable in linear time [35]. Current approaches to solving MP are heuristics based on iterative improvement techniques; they appear to return very good solutions for up to a few hundred taxa. Many software packages implement such heuristics, among them MEGA [70], PAUP* [124], Phylip [33], and TNT [41].

If one postulates a model of evolution, it becomes possible to ask for a phylogeny that is the best fit for the data under the model; this approach forms the basis for the likelihood-based criteria. The *maximum likelihood (ML)* problem asks for the tree and associated model parameter values that maximizes the probability of producing the given set of character sequences. ML is not known to be NP-hard, although a version that also asks for labelling internal nodes with character sequences consistent with the model choices, called ancestral maximum likelihood, is known to be NP-hard [1]. However, it appears considerably more difficult than MP: its point estimation problem (given a fixed tree, estimate the model parameter to obtain the maximum likelihood value for that tree) does not currently have any exact solution (except for some trivial instances of four taxa) [121]. Current approaches to ML are heuristic searches through tree space with heuristics for point estimation and are typically limited to fewer than 100 taxa; they are implemented in various software packages, including PAUP* [124], Phylip [33], FastDNAml [96], and PhyML [46].

The computational problems associated with ML approaches, along with some statistical considerations, have motivated a Bayesian approach to the problem, in which one attempts to estimate, under a given model, the posterior probability of trees given the data. This approach is invariably implemented with a Markov Chain Monte Carlo (MCMC) approach, most notably in the software package MrBayes [64]. The MCMC approach scales better than the direct ML approach, but remains limited to a few hundred taxa at best.

3.4 Meta-methods The most successful meta-method is the *Disk-covering method (DCM)* developed by Warnow and her colleagues in a series of papers [65, 66, 112, 128]. All versions of the method work in three phases: (i) they decompose the set of taxa into a number of overlapping subsets; (ii) they apply a so-called *base method* (one of those discussed above) to each subset; and (iii) they combine the resulting trees to produce a phylogeny for the original dataset. They are thus a type of divide-and-conquer methodology; the overlap between the subproblems, which is uncharacteristic of divide-and-conquer, is required here in order to combine the smaller trees into a final large tree. Combining smaller trees into a large tree is a well established problem in phylogenetic reconstruction, known as the *supertree* problem [8]; the DCM approaches, by controlling the decomposition into subproblems, also enable better recombination of the pieces [111].

DCM approaches have two main advantages. By running the expensive base methods on smaller datasets only (the latest uses of DCM are recursive [112, 128] in order to ensure a maximum size for each subset), they avoid the exponential growth in running time of these methods and can be applied to much larger datasets—tens of thousands of sequences instead of a few hundred [112] and thousands of gene orders instead of a dozen [128]. By carefully decomposing the dataset, they also produce subsets that are better conditioned for reconstruction: in particular they minimize the problems due to large ratios between the largest and the smallest pairwise distances that plague all base methods.

4 Assessment of Phylogenetic Reconstruction Methods

In phylogenetic reconstruction, an assessment must take into account the accuracy of the reconstruction (in terms of the chosen optimization criterion but also, and more importantly, in terms of the biological significance of the results) as well as the scaling up of resource consumption (time and space). In turn, conducting such an assessment requires the use of a carefully designed set of benchmark datasets [90].

4.1 Choosing benchmark sets *Biological datasets* test performance where it really matters, but they can typically be used only for ranking (because we do not know the “true” answer) and are too few to permit quantitative evaluations. (When it comes to the Tree of Life itself, it is a single dataset!) Moreover, the analysis of any large biological dataset is hard to evaluate: a 10,000-taxon tree is a very complex object and not directly amenable to human evaluation. Thus biological datasets are good for “reality checks,” a capacity in which they are indispensable, as no simulation can be accurate enough to replace real data. *Simulated datasets* permit absolute evaluations of solution quality (because the model, and thus the “true” answer, is known) and can be generated in large numbers to ensure statistical significance, as well as tailored to answer specific performance questions. Thus a combination of large-scale simulations and reasonable numbers of biological datasets is the only way to obtain valid characterizations of algorithms for phylogenetic reconstruction.

4.2 Phylogenetic considerations A typical simulation study runs as follows:

1. Generate a rooted binary tree (according to a chosen model of speciation and extinction) with the appropriate number of leaves—this tree is known as the *model tree.*

2. Assign a “length” (i.e., a number of evolutionary events) to each edge of the tree according to a chosen model of divergence.
3. Place a label of suitable size and composition (e.g., a DNA sequence or a gene order) at the root.
4. Evolve the labels down the tree, i.e., transform the parent label along each edge to its children according to the number of evolutionary events on that edge and to the chosen model of evolution.
5. Collect the labels thus generated at the leaves and use them as input to the reconstruction algorithm under test.
6. Compare the topology (and, if desired, the internal labels) of the reconstructed tree with that of the model tree.

This sequence of operations is run many times for the same parameter values (number of taxa, size of labels, parameters of the model of evolution, distribution of edge lengths, etc.) to ensure statistical significance. In five years of experimentation, we have found a few useful guidelines (see [89, 90] for details):

- Tree shape, as determined by the model of speciation and extinction, plays a surprisingly large role.
- The evolutionary models for divergence and label evolution are important. In particular, most reconstruction methods exhibit poor accuracy when the *diameter* of the dataset (the ratio of the largest to the smallest pairwise distance in the dataset) is large.
- Testing a large range of parameters and using many runs for each setting to estimate variance are essential parts of any testing strategy. In the huge parameter space induced by even the simplest of models, it is easy to fall within an uncharacteristic region and draw wrong conclusions about the behavior of the algorithm.

5 Algorithmic Challenges with Sequence Data

Very few methods offer any performance guarantees, except in purely theoretical terms. *Statistical consistency* has been viewed as an important attribute: a statistically consistent method will, given sufficient data generated under a process obeying the conditions of the chosen model, produce the true tree with high probability. ML is known to be statistically consistent under most models; but the same cannot be said of its heuristic implementation, in which many corners are cut. Even neighbor-joining, which is also statistically consistent under most models and is implemented exactly, may return very poor trees: statistical consistency only implies good performance in the limit, as sequence lengths become sufficiently large. Unfortunately, the sequence length for a single gene appears to be bounded in nature, to a few thousand base pairs; and attempts to obtain longer sequences by concatenating the sequences of several genes tend to exacerbate difficulties with alignments and of course give rise to the gene tree vs. species tree problem. One way to get around this problem is to devise methods for which the implied convergence is rapid: whereas all that is known for most statistically consistent methods is that they converge for sequences of length exponential in the diameter of the dataset, *fast-converging* methods have been devised [24, 134] that only require sequences of length polynomial in the diameter. The approach proposed by Warnow *et al.* [134] can convert any distance-based method into a fast-converging method, yet little remains known of the convergence properties of the more powerful criterion-based approaches. One of our findings, however, has been that very high accuracy in optimization is required to produce good trees [112]: with trees of 10,000 taxa, we need MP scores that are better than 99.99% of optimal in order to obtain tree with less than 5% of edges in error! Such accuracy is unheard of in the world of (practical) approximation, yet it can (at least sometimes) be achieved with MP optimization; similar constraints no doubt hold for ML and Bayesian approaches and need to be studied.

Because biologists and biochemists have been studying DNA sequences for many decades, we have a fairly good understanding of the process by which a DNA character evolves—mutations and indels. Many statistical models have been proposed for nucleotide or codon mutation, using 4×4 or 20×20 transition matrices [42, 125]. However, the problem of sequence evolution has not been well addressed to date: most models assume that each character evolves independently of all others, which is clearly false in nature, and, moreover, that each character evolves according to the same model, which is another serious oversimplification. Of course, an ML approach can easily postulate a model in which the evolution of each character obeys its own model and depends on the evolution of all other characters, but such a model would have far too many parameters to be useful. A challenge, then, is to design a model of sequence evolution that takes into account dependencies among characters (perhaps within some distance) and allows flexibility in the choice of model parameters for different characters, yet does not require much larger quantities of data nor much larger computational resources.

We mentioned earlier that multiple sequence alignment (MSA) is not well solved. It is particularly poorly solved for phylogenetic uses, as the main evaluation criterion has long been the sum of all pairwise alignment scores—whereas, in reality, the alignment scores between distantly related taxa are not very important while those between closely related taxa are crucial. As long ago as 1975, Sankoff [115] introduced the *phylogenetic tree alignment* problem: given a collection of sequences, find the tree leaf-labelled by these sequences and an assignment of (new) sequences to its internal nodes such that the sum, over all edges of the tree, of the pairwise alignments of the sequences labeling the endpoints of each edge, is minimized. This formulation leads to an NP-hard problem [130], for which a PTAS exists [131]. Using an iterative refinement approach described by Sankoff *et al.* [119], several alignment programs, such as GESTALT [71], produce such tree alignments. However, most multiple sequence alignments are still created by software designed to optimize the sum of all pairwise alignment scores, such as ClustalW [59] and TCoffee [95]. The challenge here is to develop an algorithm that runs reasonably quickly on a collection of unaligned sequences and returns accurate phylogenies. If the phylogeny is correct, the tree alignment problem reduces to assigning the best possible sequences to internal nodes for a fixed leaf-labelled tree, a problem that can be solved by dynamic programming.

Finally, in order to approach the reconstruction of the Tree of Life, we need to scale up existing reconstruction methods or develop entirely new ones. The largest sequence-based reconstructions to date have reached nearly 15,000 taxa [112]; the methodology we used in that work appears capable of scaling to significantly larger datasets (most likely to 100,000 taxa), yet we do not expect it to be applicable to datasets with millions of taxa. Further research on DCM and related methods and on supertree construction is clearly necessary. Enabling these codes to run in parallel on large machines is a plus: in 2000, it enabled our group to solve in 24 hours on a 512-CPU machine a problem that would have required over a year of computation on a workstation [3].

6 Algorithmic Challenges with Gene-Order Data

As mentioned earlier, gene-order data give rise to some very complex combinatorial problems. We consider two such problems here: how to compute pairwise distances and how to compute the median of three gene orders.

6.1 Distance computations The simplest version of the problem is to computer the edit distance between two gene orders with identical gene content and no duplication (i.e., two permutations of the set $\{1, \ldots, n\}$) under a single rearrangement operation. When the allowed operation is inversion, the measure is called the *inversion distance*; when it is transposition, the measure is called the *transposition distance.* Computing the inversion distance is NP-hard when the genes are not signed (i.e., when we do not know on what strand they reside) [18], but solvable in linear time [5] when the genes are signed, using the elaborate theory of Hannenhalli and Pevzner [52]. The best result known to date for the transposition distance is a 1.5-approximation [53, 54]—the complexity of the problem remains unknown. Combining inversions and transpositions makes for a much more difficult problem, with only a 2-approximation known [44]; if the transpositions are also inverted, then a 1.5-approximation exists [55].

All of these methods consider all inversions and transpositions to have equal weight—i.e., to be equally likely. Yet there is evidence that transpositions are more common than inversions in some genomes (such as mitochondria) and less common in others (such as chloroplasts). Moreover, at least in bacterial genomes, short inversions appear more common than long ones [73], confirming a result from Sankoff [117] in which he showed that short inversions tend to preserve gene clusters (unordered groups of genes, similar to the operon groups commonly found in bacteria). Since gene clusters are of independent interest, much research has been conducted on identifying such clusters [57] and using them in computing pairwise distances [7]. Almost no work to date has been done on the identification of so-called "hot spots"—that is, locations in the genome where the a DNA strand is easier to break and thus where rearrangements and insertions are more likely to occur, although one study seems to indicate that such hot spots are common in some mammalian genomes [104].

Every method discussed so far assumes that the genomes have equal gene content and no duplicate genes—that every genome is a (signed) permutation of the same set. In biology, however, such is never the case: the gene content varies, sometimes drastically even among closely related species, and duplications are frequent and often produce very large gene families—it is not uncommon to find families of size 50–100 in bacteria, while some gene families in eukaryotic genomes may reach sizes in the thousands. El-Mabrouk [29] extended the theory of Hannenhalli and Pevzner to account for deletions, but only in the absence of duplications. Sankoff [116] proposed to get rid of all duplicates (for computational purposes) by selecting one *exemplar* from each gene family, namely that homolog whose selection minimizes the edit distance; however, selecting an exemplar is itself an NP-hard problem [14]. Marron *et al.* [81] gave a bounded approximation for the edit distance between a permutation and an arbitrary gene order, while Swenson *et al.* [123] reported good results with a heuristic to approximate directly the true evolutionary distance between two arbitrary genomes and Tang *et al.* [129] used simple enumeration to handle datasets with small difference in gene content. All of these methods, however, are limited to just inversions and assume that all inversions have equal weight.

Finally, we have already noted that edit distances underestimate the true evolutionary distance. Alone among the methods described so far, the heuristic of Swenson *et al.* approximates the true evolutionary distance directly. Earlier efforts resulted in distance corrections for special cases: estimating the true evolutionary distance (in terms of inversions) from the breakpoint distance[1] through a formal derivation [132] and from the inversion distance through an empirical formula [88], both under the assumption of equal gene content and no duplication (see also [133]).

Thus the challenges in distance computation for gene orders include developing a theoretical framework and suitable algorithms for the following problems: (i) the transposition distance; (ii) an edit distance that combine inversions and transpositions; (iii) weighted versions of these first two distances according to both locations and lengths of the operations; (iv) combining these distances with deletions and insertions in the absence of duplications; and (v) adding duplications (including specific models of duplications, such as tandem duplications) to the existing frameworks. Since many of these problems are likely to be computationally intractable, a crucial aspect of research will be the development and validation of fast heuristics.

6.2 Computing the median of three gene orders Once a suitable distance measure has been defined and an algorithm designed and implemented for its computation, one may proceed to use gene-content and gene-order data in phylogenetic reconstruction. The most successful approach to date has been one based loosely on maximum parsimony, first proposed by Sankoff [118] and then extended by us to produce the software suite GRAPPA, which has been refined over several years [88, 91, 128, 129]. The algorithm evaluates each tree topology in turn; a tree is scored by reconstructing gene orders at internal nodes and summing the pairwise distances along the edges of the tree.

The internal gene orders are computed so as to minimize that sum of edge lengths through an iterative improvement process: first the algorithm assigns initial gene orders to the internal nodes in some manner, then it refines these gene orders by computing, at each internal node, the *median* of the gene orders of its three neighbors (i.e., a gene order that minimizes the sum of the distances to its three neighboring gene orders), repeating the process throughout the tree until convergence. While this approach nests two exponential processes (the number of trees grows exponentially with the number of taxa and computing the median of three gene orders is NP-hard [103]) and does not guarantee an optimal solution (the iterative improvement process can stop at local optima), it has done well in practice [89].

The key step is the computation of the median of three gene orders under a given distance measure; this is relatively simple for breakpoint distances (in which case the problem reduces to a highly structured version of the Travelling Salesperson Problem), but quite difficult for other distances, even the simple inversion distance. Indeed, current methods [19, 120] simply conduct an exhaustive search of the space of gene orders around the three neighbors, a search that becomes impractical as soon as the distance to the median is at all large.

When the gene contents of the three neighbors differ, the problem gets much more complex. Tang and Moret proposed to solve it in two steps, by first computing the gene content alone, then only computing the gene order [127, 129], an approach that worked well for small chloroplast datasets, but proved more problematic when used with much larger bacterial genomes [28].

Thus the main algorithmic challenge is how to compute the median exactly (or with very strong guarantees) in the presence of potentially large pairwise distances and for a variety of distance measures (as described in the previous section).

7 Other Algorithmic Challenges

Reconstructing the Tree of Life will require the collection, curation, storage, and analysis of large quantities of heterogeneous data. Thus numerous informatics problems arise, most centered around the proper design and deployment of data models. One obvious computational challenge will be the integration of various types of data—today, integrating even closely related data such as DNA sequences and gene orders remains beyond our reach. Most of these challenges, however, do not yet qualify as algorithmic, although the advent of the database as the central repository of knowledge about a biological problem introduces to computational biology the model of preprocessing and querying long used in, e.g., computational biology [26], a model that phylogenetic databases will have to support [93].

Closer to algorithm design is the issue of tree generation for simulations. Computer scientists have used trees selected uniformly at random from all leaf-labelled

[1] The breakpoint distance, proposed by Sankoff [9, 118] for use in phylogenetic reconstruction, simply counts the number of gene adjacencies present in one genome. but not in the other.

trees on n leaves, while biologists typically prefer to use so-called birth-death trees, generated by a coalescent process (as implemented in `r8s` [114], for instance). Neither type of tree appears to match the shapes observed in published phylogenies [86]: random trees are too imbalanced and birth-death trees too balanced. A better model was studied by Heard [56] and includes a notion of inheritance for speciation—a well known biological property. Our experiments in the last few years have shown that tree shape plays a larger role than one might expect in the performance of methods, so that it is crucial to use good tree shapes in evaluating reconstruction algorithms and thus to use a model for tree generation that is biologically plausible and matches observed data. The challenge here is to design and validate a model that provides biological mechanisms for tree generation and uses few parameters—the latter of particular importance for algorithm assessment, since every additional parameter exponentially increases the space of configurations that must be tested.

For a final challenge problem, let us return to an early remark: the Tree of Life is not really a tree, but a DAG, because of evolutionary events such as hybridization and lateral gene transfer, both of which result in one organism having parents from two different lineages. Biologists call such non-tree evolutionary events *reticulations*. Reticulations are common among plants (hybridization) and prokaryotes (hybridization and lateral gene transfer) and present in other areas of life. Very little is known about reticulate evolution; most of our knowledge comes from populations genetics, where recombinations present much the same picture as reticulations [108], but where the time scale and the distance between taxa are orders of magnitude smaller than in phylogenetic analysis. Posada and Crandall [107] showed that ignoring reticulations could lead to the reconstruction of inaccurate trees; in other work [105, 106], they investigated the reconstruction of DAGs rather than trees, a step that had been advocated by several authors [43, 122]. Working from sequence data, the main tool for detecting reticulation has been the presence of incongruent gene trees [79, 94]; with gene order data, reconciliation amounts to matching duplicates and this analysis can be combined with detection of horizontal gene transfer [2, 51]. Detecting conflicts in the data (but not resolving these conflicts) can be done by the method of splits [6], implemented by Bryant and Moulton as `NeighborNet` [15]. None of these methods, however, can reconstruct a phylogenetic network (the reticulated equivalent of a phylogenetic tree). We have now provided an error metric [87] to compare network reconstructions; this metric will allow algorithm designers to assess the accuracy of their algorithms in much the same way as is done for phylogenetic trees. Challenges here abound: how do we reliably detect reticulations? how can we estimate their number and location with high reliability? how can sequence and gene order data be combined to overcome some of these challenges? what is a good model for reticulation events (e.g., are there sequence-level or genome-level characteristics that can predict the likely occurrence of hybridization or gene transfer)? and finally, of course, how do we go about reconstructing evolutionary networks at the level of accuracy to which we are accustomed with trees?

8 Conclusion

We have briefly surveyed the field of computational phylogenetics, with a particular emphasis on the ultimate challenge, the reconstruction of the Tree of Life. The algorithmic challenges we have listed are but a small fraction of the challenges in just this one small area of computational biology; and as these challenges are met and overcome, their solutions will lead to even more complex problems, such as relating in a computational manner phylogeny to population genetics, evolution to development, etc. There are enough problems here, already formulated or yet to be developed, to keep teams of algorithm designers busy for many years and just the right combination of real data, credible simulation, and scaling issues to make it the ideal testing ground for algorithm engineering.

9 Acknowledgments

Our work in this area is supported in part by the National Science Foundation under grants DEB 01-20709 (on a subcontract to U. Texas at Austin), IIS 01-13095, IIS 01-21377, ANI 02-03584. and EF 03-31654, by the National Institutes of Health under grant 2R01GM056120-05A1 (on a subcontract to U. Arizona), and by the IBM Corporation under a contract from the Defense Advanced Research Projects Agency (the PERCS project).

References

[1] L. Addario-Berry, B. Chor, M. Hallett, J. Lagergren, A. Panconesi, and T. Wareham. Ancestral maximum likelihood of phylogenetic trees is hard. In *Proc. 3rd Int'l Workshop Algs. in Bioinformatics (WABI'03)*, volume 2812 of *Lecture Notes in Computer Science*, pages 202–215. Springer Verlag, 2003.

[2] L. Arvestad, A.-C. Berglund, J. Lagergren, and B. Sennblad. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i7–i15, 2003.

[3] D.A. Bader and B.M.E. Moret. GRAPPA runs in record time. *HPC Wire*, 9(47), 2000.

[4] D.A. Bader, B.M.E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In H. J. Siegel, editor, *Proc. SPIE Commercial Applications for High-Performance Computing*, volume 4528, pages 159–168, Denver, CO, 2001. SPIE.

[5] D.A. Bader, B.M.E. Moret, and M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, 8(5):483–491, 2001.

[6] H. J. Bandelt and A. W. M. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phyl. Evol.*, 1:242–252, 1992.

[7] A. Bergeron, S. Heber, and J. Stoye. Common intervals and sorting by reversals: a marriage of necessity. In *Proc. 2nd European Conf. Comput. Biol. ECCB'02*, pages 54–63, 2002.

[8] O.R.P. Bininda-Edmonds, editor. *Phylogenetic Supertrees: Combining information to reveal the Tree of Life*. Kluwer Academic Publishers, 2004.

[9] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, Tokyo, 1997.

[10] J.L. Boore and W.M. Brown. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Curr. Opinion Genet. Dev.*, 8(6):668–674, 1998.

[11] J.L. Boore, T. Collins, D. Stanton, L. Daehler, and W.M. Brown. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, 376:163–165, 1995.

[12] J.R. Brown and P.V. Warren. Antibiotic discovery: Is it in the genes? *Drug Discovery Today*, 3:564–566, 1998.

[13] W.J. Bruno, N.D. Socci, and A.L. Halpern. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.*, 17(1):189–197, 2000.

[14] D. Bryant. The complexity of calculating exemplar distances. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pages 207–212. Kluwer Academic Pubs., Dordrecht, Netherlands, 2000.

[15] D. Bryant and V. Moulton. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. In *Proc. 2nd Int'l Workshop Algs. in Bioinformatics (WABI'02)*, volume 2452 of *Lecture Notes in Computer Science*, pages 375–391. Springer Verlag, 2002.

[16] R.M. Bush, C.B. Smith, N.J. Cox, and W.M. Fitch. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc. Nat. Acad. Sci., USA*, 97:6974–6980, 2000.

[17] J.J. Cannone and S. Subramanian *et al.* The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BioMed Central Bioinformatics*, 3(2), 2002.

[18] A. Caprara. Formulations and hardness of multiple sorting by reversals. In *Proc. 3rd Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'99)*, pages 84–93. ACM Press, New York, 1999.

[19] A. Caprara. On the practical solution of the reversal median problem. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 238–251. Springer Verlag, 2001.

[20] S.B. Carroll, J.K. Grenier, and S.D. Weatherbee. *From DNA to Diversity*. Blackwell Science, 2001.

[21] J.K. Chambers and L.E. Macdonald *et al.* A G protein-coupled receptor for UDP-glucose. *J. Biol. Chem.*, 275(15):10767–10771, 2000.

[22] B.S. Chang and M.J. Donoghue. Recreating ancestral proteins. *Trends Ecol. Evol.*, 15:109–114, 2000.

[23] J. Cracraft and M.J. Donoghue, editors. *Assembling the Tree of Life*. Oxford University Press, 2004.

[24] M. Csürös and M. Y. Kao. Recovering evolutionary trees through harmonic greedy triplets. *Proc. 10th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'99)*, pages 261–270, 1999.

[25] W.H.E. Day. Computationally difficult parsimony problems in phylogenetic systematics. *J. Theoretical Biology*, 103:429–438, 1983.

[26] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer Verlag, 2000.

[27] M.J. Donoghue, R.H. Ree, and D.A. Baum. Phylogeny and the evolution of flower symmetry in the asteridae. *Trends in Plant Science*, 3:311–317, 1998.

[28] J. Earnest-DeYoung, E. Lerat, and B.M.E. Moret. Reversing gene erosion: reconstructing ancestral bacterial genomes from gene-content and gene-order data. In *Proc. 4th Int'l Workshop Algs. in Bioinformatics (WABI'04)*, volume 3240 of *Lecture Notes in Computer Science*, pages 1–13. Springer Verlag, 2004.

[29] N. El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *Proc. 11th Ann. Symp. Combin. Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Computer Science*, pages 222–234. Springer Verlag, 2000.

[30] E.M. Akam Abouheif *et al.* Homology and developmental genes. *Trends Genet.*, 13(11):432–433, 1997.

[31] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[32] J. Felsenstein. Phylogenies and the comparative method. *Amer Nat.*, 125:1–15, 1985.

[33] J. Felsenstein. *Phylogenetic Inference Package (PHYLIP), Version 3.5*. University of Washington, Seattle, 1993.

[34] J. Felsenstein. The troubled growth of statistical phylogenetics. *Syst. Biol.*, 50(4):465–467, 2001.

[35] W.M. Fitch. On the problem of discovering the most

parsimonious tree. *American Naturalist*, 111:223–257, 1977.

[36] A.N. Fox, R.J. Pitts, H.M. Robertson, J.R. Carlson, and L.J. Zwiebel. Candidate odorant receptors form the malaria vector mosquito Anopheles gambiae and evidence of down-regulation in response to blood feeding. *Proc. Nat. Acad. Sci., USA*, 98:14693–14697, 2001.

[37] D.J. Futuyma. *Evolutionary Biology 3rd ed.* Sinauer Assoc., Sunderland, MA, 1998.

[38] M.Y. Galperin and E.V. Koonin. Comparative genome analysis. *Methods Biochem. Anal.*, 43:359–392, 2001.

[39] O. Gascuel. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, 14(7):685–695, 1997.

[40] N. Goldman, J.L. Thorne, and D.T. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, 263:196–208, 1996.

[41] P. Goloboff. Analyzing large datasets in reasonable times: solutions for composite optima. *Cladistics*, 15:415–428, 1999.

[42] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution.* Sinauer Assoc., Sunderland, MA, 1999.

[43] R.C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.

[44] Q.-P. Gu, S. Peng, and H. Sudborough. A 2-approximation algorithm for genome rearrangements by reversals and transpositions. *Theor. Computer Science*, 210(2):327–339, 1999.

[45] X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.*, 18(4):453–464, 2001.

[46] S. Guindon and O. Gascuel. PHYML—a simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696–704, 2003.

[47] D. Gusfield. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.* Cambridge University Press, 1997.

[48] R.R. Gutell. Comparative studies of RNA: Inferring higher-order structure from patterns of sequence variation. *Current Opinions in Structural Biology*, 3:313–322, 1993.

[49] R.R. Gutell, J.J. Cannonne, D. Konings, and D. Gautheret. Predicting U-turns in ribosomal RNA with comparative sequence analysis. *J. Mol. Biol.*, 300:791–803, 2000.

[50] P. Halbur, M.A. Lum, X. Meng, I. Morozov, and P.S. Paul. New porcine reproductive and respiratory syndrome virus DNA and proteins encoded by open reading frames of an Iowa strain of the virus are used in vaccines against PRRSV in pigs. Patent filing WO9606619-A1, 1994. (priority date).

[51] M.T. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. In *Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'04)*, pages 347–356, New York, 2004. ACM Press.

[52] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Ann. ACM Symp. Theory of Comput. (STOC'95)*, pages 178–189. ACM Press, New York, 1995.

[53] S. Hannenhalli and P.A. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problems). In *Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95)*, pages 581–592. IEEE Press, Piscataway, NJ, 1995.

[54] T. Hartman. A simpler 1.5-approximation algorithm for sorting by transpositions. In *Proc. 14th Ann. Symp. Combin. Pattern Matching (CPM'03)*, volume 2676 of *Lecture Notes in Computer Science*, pages 156–169. Springer Verlag, 2003.

[55] T. Hartman and R. Sharan. A 1.5-approximation algorithm for sorting by transpositions and transreversals. In *Proc. 4th Int'l Workshop Algs. in Bioinformatics (WABI'04)*, volume 3240 of *Lecture Notes in Computer Science*, pages 50–61. Springer Verlag, 2004.

[56] S.B. Heard. Patterns in phylogenetic tree balance with variable and evolving speciation rates. *Evol.*, 50:2141–2148, 1996.

[57] S. Heber and J. Stoye. Algorithms for finding gene clusters. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 252–263. Springer Verlag, 2001.

[58] D.S. Hibbett, L.B. Gilbert, and M.J. Donoghue. Evolutionary instability of ectomycorrhizal symbioses in basidiomycetes. *Nature*, 407:506–508, 2000.

[59] D. Higgins, J. Thompson, T. Gibson, J.D. Thompson, D.G. Higginss, 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 Res.*, 22:4673–4680, 1994.

[60] D.M. Hillis. Primer: Phylogenetic analyis. *Current Biology*, 7:R129–R131, 1997.

[61] D.M. Hillis and J.P. Huelsenbeck. Support for dental HIV transmission. *Nature*, 369:24–25, 1994.

[62] E.C. Homles and M. Worobey *et al.* Phylogenetic evidence for recombination in dengue virus. *Mol. Biol. Evol.*, 16(3):405–409, 1999.

[63] J.P. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: testing hypothesis in an evolutionary context. *Science*, 276:227–232, 1997.

[64] J.P. Huelsenbeck and F. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754b, 2001. Available at `morphbank.ebc.uu.se/mrbayes/`.

[65] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *J. Comput. Biol.*, 6(3):369–386, 1999.

[66] D. Huson, L. Vawter, and T. Warnow. Solving large scale phylogenetic problems using DCM-2. In *Proc. 7th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'99)*, 1999.

[67] R.K. Jansen and J.D. Palmer. A chloroplast DNA

inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). *Proc. Nat'l Acad. Sci., USA*, 84:5818–5822, 1987.

[68] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[69] J. Kim and T. Warnow. Phylogenetic tree estimation, 1999. Tutorial available at `ismb99.gmd.de/TUTORIALS/Kim/4KimTutorial.ps`.

[70] S. Kumar, K. Tamura, I.B. Jakobsen, and M. Nei. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics*, 17(12):1244–1245, 2001.

[71] G. Lancia and R. Ravi. GESTALT: GEnomic STeiner ALigmenTs. In *Proc. 10th Ann. Symp. Combin. Pattern Matching (CPM'99)*, volume 1645 of *Lecture Notes in Computer Science*, pages 101–114. Springer Verlag, 1999.

[72] R.S. Lanciotti, J.T. Roehrig, and V. Deubel *et al.* Origin of the West Nile virus responsible for an outbreak of encephalities in the northeastern United States. *Science*, 286:2333–2337, 1999.

[73] J.-F. Lefebvre, N. El-Mabrouk, E.R.M. Tillier, and D. Sankoff. Detection and validation of single gene inversions. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i190–i196. Oxford U. Press, 2003.

[74] W.-H. Li. *Molecular Evolution.* Sinauer Assoc., Sunderland, MA, 1997.

[75] D.A. Liberles, D.R. Schreiber, S. Govindarajan, S.G. Chamberlin, and S.A. Benner. The adaptive evolution database (TAED). *Genome Biol.*, 2(8):RESEARCH0028, 2001.

[76] 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. (RECOMB'98)*, pages 182–191. ACM Press, New York, 1998.

[77] D.R. Maddison and W.P. Maddison. *MacClade version 4: Analysis of phylogeny and character evolution*, 2000.

[78] W.P. Maddison. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evol.*, 44:539–557, 1990.

[79] W.P. Maddison. Gene trees in species trees. *Syst. Biol.*, 46(3):523–536, 1997.

[80] W.P. Maddison and D.R. Maddison. *Mesquite: a modular system for evolutionary analyses, version 0.98*, 2001. `mesquiteproject.org`.

[81] M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. *Theor. Computer Science*, 325(3):347–360, 2004.

[82] E.P. Martins. Phylogenies and comparative data, a microevolutionary perspective. *Phil. Trans. R. Soc. Lond. B*, 349:85–91, 1995.

[83] T.J. Merritt and J.M. Quattro. Evidence for a period of directional selection following gene duplication in a neurally expressed locs of triosephosphate isomerase. *Genetics*, 159:689–697, 2001.

[84] 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. *Proc. Nat'l Acad. Sci., USA*, 99(22):14292–14297, 2002.

[85] M.G. Montague and C.A. Hutchinson III. Gene content and phylogeny of herpesviruses. *Proc. Nat'l Acad. Sci., USA*, 97:5334–5339, 2000.

[86] A.O. Mooers and S.B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Rev. Biol.*, 72:31–54, 1997.

[87] B.M.E. Moret and L. Nakhleh *et al.* Phylogenetic networks: modeling, reconstructibility, and accuracy. *ACM/IEEE Trans. on Comput. Biology and Bioinformatics*, 1(1):13–23, 2004.

[88] B.M.E. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525, 2002.

[89] B.M.E. Moret, J. Tang, and T. Warnow. Reconstructing phylogenies from gene-content and gene-order data. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 321–352. Oxford University Press, 2005.

[90] B.M.E. Moret and T. Warnow. Reconstructing optimal phylogenetic trees: A challenge in experimental algorithmics. In R. Fleischer, B.M.E. Moret, and E.M. Schmidt, editors, *Experimental Algorithmics*, volume 2547 of *Lecture Notes in Computer Science*, pages 163–180. Springer Verlag, 2002.

[91] B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, pages 583–594. World Scientific Pub., 2001.

[92] J.H. Nadeau and P.L. Grant *et al.* A rosetta stone of mammalian genetics. *Nature*, 373:363–365, 1995.

[93] L. Nakhleh, D. Miranker, F. Barbancon, W. Piel, and M. Donoghue. Requirements of phylogenetic databases. In *Proc. 3rd IEEE Symp. on Bioinformatics and Bioengineering BIBE'03*, pages 141–148, 2003.

[94] L. Nakhleh, T. Warnow, and C.R. Linder. Reconstructing reticulate evolution in species—theory and practice. In *Proc. 8th Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'04)*, pages 337–346, New York, 2004. ACM Press.

[95] C. Notredame, D. Higgins, and J. Heringa. T-Coffee: A novel method for multiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[96] G.J. Olsen, H. Matsuda, R. Hagstrom, and R. Overbeek. FastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computations in Applied Biosciences*, 10(1):41–48, 1994.

[97] C.Y. Ou and C.A. Ciesielski *et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science*, 22(256):1165–1171, 1992.

[98] R. Overbeek and N. Larsen *et al.* WIT: integrated

system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, 28:123–125, 2000.

[99] 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, pages 57–70. American Math. Soc., 1997.

[100] R.D.M. Page and M.A. Charleston. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Mol. Phyl. Evol.*, 7:231–240, 1997.

[101] J.D. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Springer Verlag, 1992.

[102] P. Pamilo and M. Nei. Relationship between gene trees and species trees. *Mol. Biol. Evol.*, 5:568–583, 1998.

[103] I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Elec. Colloq. on Comput. Complexity*, 71, 1998.

[104] P. Pevzner and G. Tesler. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Nat'l Acad. Sci., USA*, 100(13):7672–7677, 2003.

[105] D. Posada and K.A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proc. Nat'l Acad. Sci., USA*, 98:13757–13762, 2001.

[106] D. Posada and K.A. Crandall. Intraspecific gene geneaologies: trees grafting into networks. *Trends in Ecol. and Evol.*, 16(1):37–45, 2001.

[107] D. Posada and K.A. Crandall. The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.*, 54(3):396–402, 2002.

[108] D. Posada, K.A. Crandall, and E.C. Holmes. Recombination in evolutionary genomics. *Annu. Rev. Genet.*, 36:75–97, 2002.

[109] F. Roberts and C.W. Roberts *et al.* Evidence for the shikimate pathway in apicomplexan parasites. *Nature*, pages 801–805, 1998.

[110] A. Rokas and P.W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, 15:454–459, 2000.

[111] U. Roshan, B.M.E. Moret, T. Warnow, and T.L. Williams. Performance of supertree methods on various dataset decompositions. In O.R.P. Bininda-Edmonds, editor, *Phylogenetic Supertrees: Combining information to reveal the Tree of Life*, pages 301–328. Kluwer Academic Publishers, 2004.

[112] 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. 3rd IEEE Computational Systems Bioinformatics Conf. CSB'04*, pages 98–109. IEEE Press, Piscataway, NJ, 2004.

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

[114] M.J. Sanderson. r8s: inferring absolute rates of evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19:301–302, 2003.

[115] D. Sankoff. Minimal mutation trees of sequences. *SIAM J. Applied Math.*, 28(1):35–42, 1975.

[116] D. Sankoff. Genome rearrangement with gene families. *Bioinformatics*, 15(11):990–917, 1999.

[117] D. Sankoff. Short inversions and conserved gene cluster. *Bioinformatics*, 18(10):1305, 2002.

[118] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.

[119] D. Sankoff, R.J. Cedergren, and G. Lapalme. Frequency of insertion-deletion, transversion, and transition in the evolution of 5s ribosomal rna. *J. Mol. Evol.*, 7:133–149, 1976.

[120] A.C. Siepel and B.M.E. Moret. Finding an optimal inversion median: experimental results. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01)*, volume 2149 of *Lecture Notes in Computer Science*, pages 189–203. Springer Verlag, 2001.

[121] M.A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biol.*, 43(4):560–564, 1994.

[122] K. Strimmer, C. Wiuf, and V. Moulton. Recombination analysis using directed graphical models. *Mol. Biol. Evol.*, 18(1):97–99, 2001.

[123] 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 SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*. SIAM Press, Philadelphia, 2005.

[124] D.L. Swofford. *PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b8*, 2001.

[125] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, B.K. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, MA, 1996.

[126] P.G. Szekeres and A.I. Muir *et al.* Neuromedin U is a potent agonist at the orphan G protein-coupled receptor FM3. *J. Biol. Chem.*, 275(27):20247–20250, 2000.

[127] J. Tang and B.M.E. Moret. Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In *Proc. 8th Int'l. Workshop on Algs. and Data Structures (WADS'03)*, volume 2748 of *Lecture Notes in Computer Science*, pages 37–46. Springer Verlag, 2003.

[128] J. Tang and B.M.E. Moret. Scaling up accurate phylogenetic reconstruction from gene-order data. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i305–i312. Oxford U. Press, 2003.

[129] J. Tang, B.M.E. Moret, L. Cui, and C.W. dePamphilis. Phylogenetic reconstruction from arbitrary gene-order data. In *Proc. 4th IEEE Symp. on Bioinformatics and Bioengineering BIBE'04*, pages 592–599. IEEE Press, Piscataway, NJ, 2004.

[130] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J. Comput. Biol.*, 1:337–348, 1994.

[131] L. Wang, T. Jiang, and D. Gusfield. A more efficient approximation scheme for tree alignment. *SIAM J. Comput.*, 30(1):283–299, 2000.

[132] L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01)*, pages 637–646. ACM Press, New York, 2001.

[133] L.-S. Wang and T. Warnow. Distance-based genome rearrangement phylogeny. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 353–383. Oxford University Press, 2005.

[134] T. Warnow, B.M.E. Moret, and K. St. John. Absolute convergence: true trees from short sequences. In *Proc. 12th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'01)*, pages 186–195. SIAM Press, 2001.

[135] T. Warnow, D. Ringe, and A. Taylor. Reconstructing the evolutionary history of natural languages. In *Proc. 7th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'96)*, pages 314–322, 1996.

[136] C.O. Webb, D.D. Ackerly, M. McPeek, and M.J. Donoghue. Phylogenies and community ecology. *Annual Rev. Ecol. Syst.*, 33, 2002. in press.

[137] G. Weiblen, R. Oyama, and M.J. Donoghue. Phylogenetic analysis of breeding system evolution in monocotyledons. *Am. Nat.*, 155:46–58, 2000.

[138] E.I. Wilding and J.R. Brown *et al.* Identification, essentiality and evolution of the mevalonate pathway for isopentenyl diphosphate biosynthesis in Gram-positive cocci. *J. Bacteriology*, 182:4319–4327, 2000.

[139] H. Zhu and J.F. Klemic *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genetics*, 26(3):283–289, 2000.

[140] Y. Zhu and D. Michalovich *et al.* Cloning, expression, and pharmacological characterization of a novel human histamine receptor. *Mol. Pharmacol.*, 59(3):434–441, 2001.