An Introduction to Phylogenetic Analysis

Olaf R.P. Bininda-Emonds
Faculty V, Institute for Biology and Environmental Sciences (IBU)
AG Systematics and Evolutionary Biology
Carl von Ossietzky Universität Oldenburg
D-26111 Oldenburg
Germany

Abstract. Biological species do not represent independent units of analysis because of the pattern of descent with modification through evolution. Although this fact has been appreciated since at least the time of Darwin, it is only in the past 15 years that biologists have increasingly appreciated the need to account for evolutionary relatedness in their analyses. However, phylogeny transcends pure biology to play an important role in related fields. For instance, alignment programs like Clustal rely on phylogenetic information in constructing their guide trees. Unfortunately, finding the optimal phylogeny for a given data set, like most other interesting problems in computational biology, is an NP-hard problem because the number of potential solutions (tree space) grows super-exponentially with the number of taxa in the analysis. With the rapid growth of phylogenomic data and thus the scope of the analyses based on them, phylogenetic inference has also garnered increasing attention from computer scientists eager to find fast solutions to this difficult problem. In this talk, I introduce the problem of phylogenetic inference in general and discuss several optimization criteria (parsimony, likelihood, and distance methods) and the strengths and weaknesses of each for building evolutionary trees.

Contents

1 Introduction

Evolutionary biology has been undergoing a renaissance recently, attracting renewed interest from the scientific community over the past two decades and especially in this, the bicentennial of the birth of Charles Darwin. Although the role of evolution as the framework underlying all biology was finally cemented in the early 20th century through the fusion of Darwinian evolution and Mendelian genetics (the *Modern Synthesis*), the realization of the need to account explicitly for the evolutionary history of the organisms under investigation is comparatively recent [14, 20]. Because all organisms have a shared evolutionary history, similarities between any two species can arise either through this shared history (phylogenetic inertia) and/or because of a common selection pressure. As such, deciphering the relative roles of these two respective causes of similarity requires knowledge of the phylogenetic relationships of the organisms in question. This realization in the mid-1980s increased the profile for phylogenetic systematics as a research field, with further fuel being added around the same time by the molecular revolution. Thanks to tremendous advances in sequencing technology, it was not only the case that phylogenetic trees were needed to perform good biology, but also that large-scale data sets became available to help meet this goal. Whereas typical morphology-based data sets continue to be only on the order of a few tens to hundreds of species and characters, multigene data sets now commonly include thousands of species and/or many tens of thousands of base pairs of DNA sequence data (e.g., [4, 30, 36, 46]).

But how do we make the jump from the data to the trees? What are the problems involved in this inference and what is the best method to obtain a phylogenetic tree? As I will show, the fundamentally difficult problem that is finding the optimal phylogenetic tree for a data set typically requires solutions that themselves are less than optimal. In this paper, I outline the basic hurdles that need to be overcome in phylogenetic inference and outline several different schools of thought on how to find the optimal tree. Despite their differences, these schools are all united in their desire to describe natural groups, clusters pertaining to real evolutionary entities.

2 The basic problem – navigating tree space

The inherent limitation to phylogenetic inference, quite simply, is that there are too many trees. Whereas there are only exactly three unique rooted, binary topologies

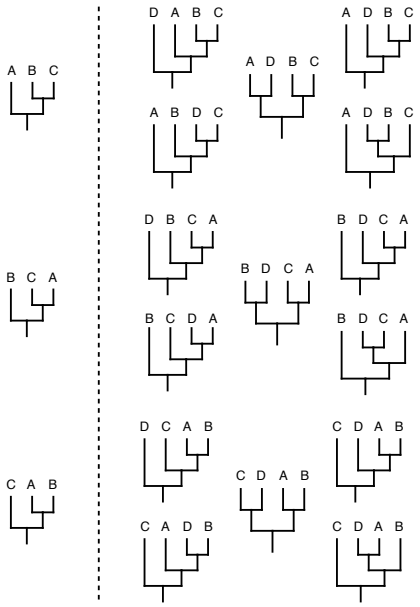

Figure 1: All possible binary rooted trees for three (left) and four (right) taxa.

for three taxa, there are 15 for four taxa, and 105 for five taxa (Fig. 1). The numbers become even larger when we do not require that each internal node in the tree have only two descendants (i.e., are binary). In fact, the increase in the number of possible topologies with the number of taxa is super-exponential. When restricted to binary, rooted topologies (the kind of trees that biologists prefer to work with), the number of trees for a given set of taxa (n) is [15]:

$$(2n-3)!! \quad = \quad \frac{(2n-3)!}{2^{n-1}(n-1)!} \quad = \quad 3 \times 5 \times 7 \times \ldots (2n-3). \tag{1}$$

As Table 1 shows, the number of possible tree topologies rises extremely quickly, such that with only 67 taxa, the number of topologies is on the order of 10^{110}, or roughly the volume of the universe in cubic Ångstroms ($= 10^{-10}$m). Given that these values are a lower bound (trees need not be fully resolved and unrooted solutions also exist), how can it ever be possible to find the best solution(s) for a given data set? And, yet, evidence exists that even if we are not finding the absolute optimal solution(s), we are usually finding trees that are probably only slightly suboptimal (e.g., [24]).

Number of taxa	Number of rooted binary trees
3	3
4	15
5	105
6	945
7	10395
...	...
10	34459425
...	...
50	2.75×10^{76}

Table 1: Number of rooted, binary tree topologies for a given number of taxa based on the formula from Felsenstein [15].

The philosophical answer to this problem is rooted in the principle of parsimony, also known as Ockham's Razor after the English Franciscan friar who formally described it, William of Ockham (1285 – 1349). Ockham's Razor states quite simply that the preferred solution is the one that minimizes the number of ad hoc assumptions. In a phylogenetic sense, the simplest solution is the one that posits the fewest number of evolutionary changes. It is far more parsimonious to assume that a shared feature in two species arose once in the common ancestor of those species (= homology) than independently in either lineage (= homoplasy). This assumption is both in accord with the idea underlying evolution (namely descent with modification) and also gives us a mechanism to infer clusters of closely related organisms on a tree. Thus, through the principle of parsimony, we assume that hair evolved once in the common ancestor of all $\sim$5000 species of mammal and not several thousand times. The presence of hair also serves as a character supporting the recognition of mammals as a natural group. Conflicts do occur, however, and similar features can and have arisen independently on multiple occasions. For instance, it is clear that wings have evolved independently among several animal groups: insects, birds, pterosaurs, and bats to name but a few. Revealing those features that are homologous from those that are homoplasious requires the application of global parsimony over all characters to find the tree topology that minimizes the number of instances of homoplasy (= ad hoc assumptions). Using this approach, we could determine on the basis of a larger data set, for instance, that hair is a shared homology of mammals, whereas the presence of wings is homoplasious between mammals (bats) and the other groups listed. (But, at the same time, this procedure will also enable us to determine that the presence of wings within mammals represents a shared homology for the bats!)

3 Finding the optimal tree

Ockham's Razor, however, really just provides us with a criterion for assessing the best tree. It does not provide a mechanism for actually finding that tree out of the super-exponentially many that are possible for a given data set. Here, instead, we need to rely on search strategies to find the path through tree space that leads to the optimal solution.

For small data sets, exact solutions that guarantee finding the optimal tree for a data set and any and all equally optimal trees are possible. The more direct solution here is an exhaustive, brute-force method that explicitly searches every possible tree topology. However, even with the recent increases in computer speed, the super-exponential increase in the number of trees means that exhaustive searches can only be performed on data sets with about 10 taxa at most (at least on a single desktop PC).

A more sophisticated exact method is branch-and-bound [22], which is able to preemptively discard unproductive paths through tree space. Branch-and-bound algorithms essentially operate by establishing an initial path through tree space to a particular tree. The optimality score of the tree is noted (the bound) and then other paths to other trees are examined. Once the tree length along the new path exceeds the stored bound, the path is aborted, regardless of whether or not a full tree has been reached or not. Continuing along the path can only lead to an increasingly less optimal solution. If, however, the new path does lead to a shorter tree in the end, then the length of the latter is stored as the new bound. In any case, the procedure repeats until all possible paths have been at least implicitly examined.

More precisely, the branch-and-bound method proceeds as follows. The different paths are built up through the process of taxon addition such that the trees are grown in a stepwise fashion. For the initial step, three taxa are chosen and joined together, with the optimality score of this subtree noted. A fourth taxon is then added to one of the three possible positions (effectively demonstrated in Fig. 1) and the score of this subtree is stored. The procedure then repeats until all taxa have been added in turn to yield a complete tree. The optimality score of this tree represents the initial value of the bound. Thereafter, a new round of taxon addition begins (the new path in tree space). If the length of any subtree in this process exceeds the current bound, the process of taxon addition is aborted and the remainder of that path is not examined: adding more taxa can only result in an increasingly less optimal solution. Thus, depending on how well the initial bound is chosen, a branch-and-bound method can quickly abort many unproductive paths up a tree and so be used on larger data sets than can an exhaustive search. Even so, the upper bound here for such searches is on the order of only about 20 taxa.

Beyond this point, there is no other option but to apply heuristic search strategies that, although usually very effective, cannot guarantee that any or all optimal solutions will be found. Most heuristic search strategies navigate tree space by perturbing an initial starting tree (typically obtained via taxon addition as for branch-and-bound searches except that each new taxon is added to its optimal position in the subtree)

through some form of random branch swapping: a branch on the tree is pruned, regrafted on to some other part of the tree, and the optimality scores of the two respective trees are compared. If the new tree is more optimal than the old one, then it is retained for additional rounds of branch swapping. Should it be less optimal, then it is discarded and the initial tree is used again instead. The procedure continues until some preset limit has been obtained (e.g., time taken, number of equally optimal solutions found, or number of branch swaps performed) or no more swapping is possible.

As effective as this basic procedure can be, it also runs the real danger of becoming stuck in regions that are only locally, and not globally, optimal. Most heuristic algorithms apply a search strategy known as hill-climbing: new solutions can be retained only if they are more optimal. Less optimal solutions are always discarded. Thus, the solution obtained depends not only on the shape of tree space (e.g., is there one clear peak or many?), but also on the starting point and the robustness of the branch-swapping algorithm applied. Consider the following example. Say we want to determine the highest point in Germany (the Zugspitze at 2962 m) using a simple hill-climbing solution: we can only go up, not down. If we start near Garmisch-Partenkirchen (elevation 708 m and already in the Alps close to the Zugspitze), then we stand a good chance of obtaining a reasonably optimal solution, if not reaching the top of the Zugspitze itself. If, however, we start in Oldenburg (elevation 4 m), we are likely to end up somewhere in the nearby Harz mountains (highest elevation 1141 m). Although these mountains are the highest point in the nearby countryside, they are merely a local optimum and definitely not the highest point in Germany.

One way to improve our chances of escaping a local optimum is to increase the degree of the tree perturbations that are performed, essentially enabling us to bridge wider valleys between the hills. For instance, we could employ more robust forms of branch swapping (see [52]). Whereas nearest-neighbour interchanges (NNI) can only swap neighbouring branches (small steps), subtree pruning and regrafting (SPR) or tree bisection and reconnection (TBR) can move entire clusters and to far removed parts of the tree (large steps). This increased thoroughness (NNI moves being a subset of SPR moves, which, in turn, are a subset of TBR moves) comes at the price of (much) longer analysis times. Other solutions include algorithms like the (parsimony) ratchet [38, 55], which uses character reweighting to jump between peaks in tree space, or any other of a host of new algorithms and shortcuts (e.g. [18, 48]), many of which represents contributions from the computer science community. In all cases, it is helpful to begin many independent searches from different starting points, thereby increasing the chances of starting the search close to Garmisch-Partenkirchen and not Oldenburg.

4 Schools of phylogenetic thought

The history of phylogenetic research has been loud and controversial. Whereas early disagreements dealt entirely with what the true set of relationships for a given group were, later discussion has revolved around what the right methodology is to find this

tree. The latter discussion gained steam in the 1960s with the English translation of the book *Phylogenetic Systematics* [23] in which the German entomologist Willi Hennig (1913 – 1976) set forth his ideas on how to conduct systematic research. These ideas were readily adopted (albeit with some delay) by the phylogenetic school now known as cladistics, who fought a long-running battle with the established school at the time (phenetics) before essentially winning by the early 1980s. (For a fascinating account of this entire period, see [27].) The victory, however, was short-lived because of the vast increase in sequence-based molecular data fueling the rise of statistical phylogenetics, which is now arguably the most commonly used method, at least for the analysis of molecular sequence data. In this last section, I briefly examine each of the three main schools, largely on the basis of the optimization criteria underlying them to bring some background information to the debate of the best method, a question with no clear answer. My focus will be largely on the application of these methods to the analysis of molecular sequence data, by far the most commonly analyzed data type today, although many of the same arguments still apply for other forms of data (e.g., morphological data). More detailed information about the different methods, including their implementations and mathematical foundations, can be found elsewhere. Two excellent references here are the book *Inferring Phylogenies* by Joe Felsenstein [13] or the German language book *Gene und Stammbäume* by Volker Knoop and Kai Müller [32].

4.1 Neighbour joining (phenetics)

The school of phenetics essentially holds that the phylogenetic relationships of organisms can best be revealed on the basis of their overall similarity: fish look like other fish, birds like other birds, and mammals like other mammals. Again, this tends to agree with our understanding of evolution and the process of descent with modification causing more closely related organisms to be similar to one another phenotypically (and genotypically). The pheneticists of the 1950s and 1960s attempted to apply this principle through an objective analysis of the entire phenotype (or at least of as many characters as possible), thereby attempting to remove the subjective selection and analysis of the data that had occurred before this point. Two of the ultimate death blows to this form of phenetics, however, was that not all similarity is in fact informative (see below) and that no single, objective optimization criterion existed, with different criteria often giving different results.

Phenetics survives today in the form of distance methods, and neighbour joining (NJ) in particular. These methods characterize the overall evolutionary distance between two taxa, something that is easy to calculate with DNA sequence data for a given model of evolution. Differences in the distances, however, cannot be traced back easily to differences in the raw character data. Thus, we would only know that the two taxon pairs A and B, and B and C are each separated by an evolutionary distance of, say, 2, and not whether the differences characterizing A and B are the same as those for B and C (in whole or in part).

This loss of information, however, is counteracted by the speed of the distance analyses, with NJ easily being the fastest of the major methods and able to analyze huge data sets of several thousand taxa within hours, if not minutes, on an ordinary desktop PC [26, 34]. The exact details of the NJ algorithm can be found elsewhere (e.g., in either of the Felsenstein or Knoop and Müller books mentioned above) and, for small trees, the NJ score is easily worked out by hand. Essentially, NJ is a heuristic for finding the tree with the shortest total branch length (the minimum evolution tree). Its speed (on the order of $O(n_3)$) is therefore slightly deceptive compared to the other methods. Unlike the other methods, NJ only engages in a single round of taxon addition with no subsequent branch swapping. Thus, the method is both fast and yields a single solution because it completely forgoes the arduous task of combing tree space for increasingly optimal solutions. (Indeed, it is the intensive nature of branch swapping that contributes in large part to making both maximum parsimony and maximum likelihood NP-hard [17] problems [6, 42]. However, tree building from incomplete distance matrices is also NP-hard [9] and the use of distance methods does not preclude branch swapping. If branch swapping is tacked on to the end of a NJ analysis, it is referred to instead as a minimum evolution (ME) analysis, which does attempt to find the tree with the shortest branch length.) For many, this fast, decisive analysis has much to recommend it; however, it has been shown repeatedly that NJ typically shows less accuracy than the remaining methods in recovering known model trees in simulation [5, 25, 53]. The solutions are not necessarily bad in the absolute sense, just not as good as those delivered by the other methods. The speed of NJ, however, cannot be denied, and NJ is often used to generate fast starting trees upon which to perform branch swapping in the other methods (e.g., the maximum likelihood program PHYML [19]).

4.2 Maximum parsimony (cladistics)

One of the key insights by Willi Hennig was that not all similarity is accurate when reconstructing evolutionary relationships. Lungfish look more similar to other fish on the whole than they do to us, but they are still more closely related to us! (Put another way, lungfish and humans exclusively share a more recent common ancestor than either do with other fish.) The key is to distinguish between primitive similarities inherited from a distant common ancestor (the general fish-like form of lungfish and other fish) and derived similarities from the most recent common ancestor (lungs and the skeletal morphology of the limbs in lungfish and humans). These two types of similarity are referred to as symplesiomorphies and synapomorphies, respectively, and only the latter are informative for defining clusters (= clades) on evolutionary trees. As mentioned above, it was this realization that helped cladistics triumph over phenetics, together with a lot of appeals to the philosophical soundness of cladistics and Popperian hypothetico-deductive falsification in particular (see [1]).

Cladistic analyses employ maximum parsimony (MP) as the optimization criterion, and the two terms have become so closely linked so as to have become virtually inseparable today. MP, however, is nothing more than a direct application of Ock-

ham's Razor and by no means exclusive to cladistics. The objective function in this case is to find the tree with the minimum number of evolutionary changes, thereby minimizing the number of homoplasies.

Although MP has apparently performed well with morphological data (although there has, until recently, been no alternative), its objective function has shown some problems with molecular sequence data and its limited character state space. For instance, DNA sequence data present only four character states, the nucleotide bases A, C, G, and T. As such, even two species that are infinitely distantly related will have DNA sequences that are about 25% similar on average, and these similarities will generally be entirely homoplasious. MP, however, will often interpret such large numbers of convergent similarities as homologies (the more parsimonious solution) and so will tend to cluster taxa together that sit isolated on long branches, where there is the strong likelihood for such convergent similarities to arise. This phenomenon, first identified by Felsenstein [11] has been come to known as the long-branch attraction (LBA) problem (for a review, see [3]). Worse yet, under conditions where MP is susceptible to this problem, it is statistically inconsistent, meaning that increasing the amount of sequence data in the analysis will only exacerbate the problem to increase the chances of going astray [11]. Thus, where MP is susceptible to LBA, using infinite amounts of data only guarantees getting the wrong answer!

Many of the problems exhibited by MP in this case also derive from it not being able to fully incorporate a model of evolution for molecular sequence data (including parameters such as base frequencies, transition probabilities between bases, or site-to-site rate heterogeneity), which would help it to distinguish homoplasies as such. Problems with LBA in MP analyses can be ameliorated to some degree by including specific taxa designed to subdivide suspected long branches. However, suitable taxa are not always available such that some taxa are always fated to sit problematically on a long branch (e.g., many of the basal angiosperm taxa and *Amborella* in particular; see [50]). Another potential solution is to delete fast-evolving sites beforehand (which will contribute the most to LBA) and it is often argued that third-codon positions for coding DNA sequences should be removed for this very reason (e.g., [56], but see [29, 45]).

The fact that MP does not include an explicit model of evolution has been hailed as a philosophical advantage of the method by some (e.g., [44]) because it minimizes the number of strong auxiliary assumptions being made. It also enables MP to be used as an optimization criterion for most data types, be they morphological or molecular. However, it has been shown through simulation that incorporating differential transition probabilities for sequence data in a MP analysis via weighting schemes does result in greater accuracy (e.g., [24]). In any case, the relative simplicity of the objective function makes it comparatively rapid and efficient solutions also exist for it (e.g., Fitch [16] and Farris [10] optimization for unordered and ordered characters, respectively). Thus, as mentioned previously, the speed bottleneck for MP analyses derives largely from the intensive nature of branch swapping needed for tree surfing and not from determining the parsimony score for any given tree, which is only on the

order of $O(nm)$, where n = the number of taxa and m = the number of characters.

Practically, MP searches often bog down because many trees possess the same parsimony score and so are equally optimal. Searching through and saving all these alternatives, and there can often be hundreds of thousands, strains the memory capacity of most desktop PCs (not to mention taking a great deal of time), typically causing the analysis to be aborted prematurely. Depending on the shape of tree space, this phenomenon will often prevent the MP analysis from proceeding to find trees that are increasingly optimal. This problem can be circumvented by reducing the number of equally optimal solutions that are saved at any point in the analysis. However, this is also not without its problems given that this strategy can prevent those few equally optimal paths that point further up the hill from being examined. A more sophisticated implementation of this idea is the parsimony ratchet [38], which through its combination of random character reweighting and restrictions on the numbers of equally optimal trees sampled causes the analysis to quickly visit more disperse points in tree space and thus escape local optima.

4.3 Maximum likelihood and Bayesian inference (statistical phylogenetics)

Likelihood has its roots in Bayes' Theorem, developed by the British mathematician and Presbyterian minister Thomas Bayes (1701/1702 (?) – 1756) and published posthumously in 1764:

$$P(H|D) = \frac{P(D|H) \times P(H)}{P(D)}, \tag{2}$$

where D = the data (e.g., the multiple sequence alignment), and H = the hypothesis (a given phylogenetic tree).

Pure maximum likelihood (ML) analyses largely concern themselves with the conditional probability $P(D|H)$ – how likely it is to see the data if the hypothesis is true? – and attempt to find the tree that maximizes the likelihood of observing the data. Both $P(D)$ and $P(H)$ are effectively ignored, the former being unknown in most cases and the latter being assumed to be equal across all trees. As such, the fundamental problem of phylogenetic analysis again pertains, namely finding those solutions among the super-exponentially large number that exist that are optimal for the given data set.

ML analyses add an extra computational wrinkle to this problem, however, because of their explicit model-based nature. Thus, the likelihood for a given tree depends not only on the tree topology (as for MP), but also on the branch lengths for a given topology (∼amount of evolution that has occurred), and the parameters for the model of evolution. To make this problem somewhat tractable, several simplifying assumptions are made. First, the existing models of evolution are stochastic. Second, they are time reversible, meaning that it does not matter if sequence one evolves into sequence two or vice versa. This assumption in particular enables the use of unrooted trees, eliminating another parameter (the position of the root) from the calculations. Third, most models assume that the evolution of different sites and different lineages

within the tree are independent. Certain models, however, can explicitly account for potential non-independence among sites. One such model is the doublet model of Schoniger and von Haeseler [43], which can account for nucleotide pairing (e.g., as occurs in the stem regions of ribosomal sequences).

However, even with these simplifying assumptions, ML remains a daunting computational problem because of the many parameters that need to be optimized in addition to simply the tree topology. Even calculating the likelihood of a tree with a fixed topology, fixed branch lengths, and fixed parameters for the chosen model of evolution is more intensive than calculating the analogous parsimony score because one needs to integrate over all possible combinations of nucleotides for all internal nodes (only these data for the tips of the tree are normally known, these being our observed sequence alignment) for each position in the alignment. Thus, both the number of taxa and the number of characters in a ML analysis will impact on its speed (whereas the latter was comparatively negligible for MP).

Interestingly, despite the much greater computational intensity of ML analyses compared to other methods, recent implementations of the method such as RAxML [47] or PHYML are faster than many MP implementations. This is especially true when compared to the program PAUP* [51], which, despite being the workhorse for MP analyses, uses search algorithms that are at least seven years old. The speed of these ML implementations is due in part to the use of dirtier search heuristics (e.g., the first versions of PHYML only used NNI branch swapping) plus various tricks from the computer science community to speed up the likelihood calculations or to reduce the memory footprint of the program. Thus, ML analyses of thousands of taxa on an ordinary desktop PC are now feasible, if one is prepared to wait a little bit. Even more remarkably, support values for trees of the same dimensions can also be calculated relatively efficiently [2, 48].

Related to ML is Bayesian inference (BI), which, unsurprisingly, is also based on Bayes' Theorem. Key differences with ML include BI focusing instead on the left-hand side of Bayes' Theorem (i.e., the posterior probability $P(H|D)$) and also its implementation via Markov chain Monte Carlo (MCMC) methods (see the contribution by Katzgraber in this volume [31]). Thus, instead of attempting to find the tree providing the highest likelihood, BI seeks to find those clades with the highest posterior probabilities. However, doing so now requires some knowledge of $P(H)$, the so-called prior probabilities that are ignored in ML analyses. Priors can be defined for almost any variable relevant to the analysis: tree shapes allowed, model parameters such as transition probabilities or base frequencies, or the probability of any single tree, to name but a few. These priors are usually defined based on previous knowledge and experience and are generally not fixed such that they can change during the analysis. They also have a markedly reduced influence in larger data sets, where they tend to be overruled in favour of the actual observations (i.e., the sequence alignment). Practically, however, the defining of priors tends to be a strong assumption that most biologists shy away from, such that flat (also known as Dirichlet) priors are usually used, essentially mimicking ML analyses where $P(H)$ is assumed to be equal among all alternatives.

Another important difference in most BI implementations is that they do not search for the tree with the highest posterior probability, but instead sample from the stable posterior distribution of nearly optimal trees to determine individual clades with the highest posteriors (i.e., those that appear in the distribution the most often). This procedure then yields a tree that automatically provides support values for the individual nodes, namely their posterior probabilities. This sampling strategy, however, cannot operate using a classic hill-climbing strategy, which is geared to find the peak only. BI instead wants to wander around the peak and so requires a mechanism enabling it to occasionally go downhill as well. It doesn't want to wander too far down the hill and away from the posterior distribution, however, and so only accepts less optimal trees with a probability proportional to the ratio of the difference in likelihood between the two trees being tested. A common implementation of this strategy in an MCMC framework is the well-known Metropolis-Hastings algorithm [21, 37].

BI has grown increasingly popular for phylogenetic inference because of the power and flexibility of the approach, combined with an excellent, full-featured implementation in MrBayes [40]. BI also appeals to biologists in that it automatically provides many hundreds and thousands of trees (those sampled from the posterior distribution), thereby providing an estimate of the phylogenetic uncertainty surrounding the main hypothesis that can be incorporated into subsequent analyses using the tree as a framework. (One must add, however, that similar estimates could easily be obtained for the other methods as well through methods such as the non-parametric bootstrap [12].) A fundamental problem in BI remains that there is no objective method to determine when the analysis has reached the stable posterior distribution, a point known as stationarity or convergence. The Metropolis-Hastings algorithm will reach this point after an infinite number of generations, but few people are willing to wait that long for the results. Fortunately, MrBayes and a number of other programs incorporate several robust diagnostics to help determine when convergence has likely been achieved. In the end, however, the best strategies still remain to run the analyses for longer (e.g., many millions of generations depending on the size and complexity of the data set) and then in parallel (so-called multiple chain MCMC or MC3). The latter suggestion is based on the reasoning that it is unlikely that two fully independent runs will end up in the same region of tree space unless this region is the posterior distribution. Whether or not this logic is absolutely true remains to be shown (and is probably data set dependent), but it is similar in spirit to performing multiple runs in a pure hill-climbing strategy to increase the likelihood of reaching the global optimum.

5 Conclusions

Phylogenetic analysis remains a field in a state of flux and one that has seen great methodological advances in the past decade. As mentioned in the Introduction, many

of these advances derive from the increasing input of computer scientists, who have built upon the earlier work of several pioneering, computer literate biologists and/or mathematical biologists. The rise of DNA sequence data as the premier data source for phylogenetic analysis has also played a role, with it presenting a much more interesting and tractable problem for the computer scientists. Likelihood models also exist for non-sequence data (e.g., [33, 54]) and, most recently, for combining tree topologies in a supertree framework [49], but have failed to generate the same kind of interest and widespread usage.

These advances, however, have only made biologists hungry to infer even larger phylogenetic trees. Whereas trees with hundreds of taxa were remarkable at the turn of the century, trees with thousands of taxa are becoming increasingly common, and the data potentially available in databases like GenBank (`www.ncbi.nlm.nih.gov`) offer the promise of even larger trees. Solving problems of this scale will require, in the first instance, ever more efficient implementations of existing methods. For instance, by relying on the construction of nodal profiles instead of building a traditional distance matrix, the ME program FastTree [39] was able to infer a tree for 186743 aligned sequences from the GreenGenes database of 16S rDNA sequences [7] in only 29 hours (albeit relying on NNI moves only).

However, in solving such problems, it will also be important to develop new search strategies in addition to using faster implementations on faster computers or larger computer clusters. A potential solution here is a divide-and-conquer strategy, which splits the large global problem into several smaller problems that are computationally more tractable, both because they are smaller and because they are of a more restricted phylogenetic diameter (i.e., comprise sets of more closely related species). The latter consideration potentially combats an emergent property of extremely large data sets, namely that it becomes increasingly difficult to align sequence data across increasingly distantly related organisms. Although early results using divide-and-conquer approaches have been promising (e.g., [5, 41]), the speed and accuracy increases that have been observed to date have not been as impressive as was hoped. As such, the approach has yet to find favour among phylogeneticists, who have instead turned more to parallel computing as a solution for the analysis of large phylogenetic data sets.

A final consideration is that the *Tree of Life* most phylogeneticists have long been striving to infer might actually not be very tree-like at all for many groups. Instead, reticulate processes such as horizontal gene transfer or hybridization cause the tree to resemble a network. This is particularly true for groups such as Prokaryota (which is likely not a natural group), which display rampant horizontal gene transfer, even between distantly related species [8, 35]. Fortunately, a strict tree-like structure (whether fully resolved or not) merely represents a special case of a network and the past decade has seen increasing interest into developing methods to infer phylogenetic networks (see [28]).

All told, the present day represents an exciting time for phylogenetic systematics, with the promise of new data, new methods, and especially new insights into the phylogenetic history of life. If the tremendous developments in phylogenetic methodology

of the past decade carry on into the near future, then we might very well soon be able to derive a robust vision of the *"Tree" of Life*.

References

[1] V. A. Albert (ed.). *Parsimony, phylogeny, and genomics*. Oxford University Press, Oxford, 2005.

[2] M. Anisimova and O. Gascuel. Approximate likelihood ratio test for branchs: a fast, accurate and powerful alternative. *Syst. Biol.*, 55:539-552, 2006.

[3] J. Bergsten. A review of long-branch attraction. *Cladistics*, 21:163-193, 2005.

[4] O. R. P. Bininda-Emonds, M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman and A. Purvis. The delayed rise of present-day mammals. *Nature*, 446:507-512, 2007.

[5] O. R. P. Bininda-Emonds and A. Stamatakis. Taxon sampling versus computational complexity and their impact on obtaining the Tree of Life. In T. R. Hodkinson and J. A. N. Parnell (eds), *Reconstructing the Tree of Life: taxonomy and systematics of species rich taxa*, pages 77-95, CRC Press, 2006.

[6] B. Chor and T. Tuller. Maximum likelihood of evolutionary trees: hardness and approximation. *Bioinformatics*, 21:97-106, 2005.

[7] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72:5069-5072, 2006.

[8] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124-2129, 1999.

[9] M. Farach, S. Kannan and T. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155-179, 1995.

[10] J. S. Farris. Methods of computing Wagner trees. *Syst. Zool.*, 19:83-92, 1970.

[11] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 27:401-410, 1978.

[12] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783-791, 1985.

[13] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.

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

[15] J. Felsenstein. The number of evolutionary trees. *Syst. Zool.*, 27:27-33, 1978.

[16] W. M. Fitch. Toward defining the course of evolution: minimum change for a specic tree topology. *Syst. Zool.*, 20:406-416, 1971.

[17] M. R. Garey and D. S. Johnson. *Computers and intractability: a guide to the theory of NP-completenes*. W. H. Freeman, San Francisco, 1979.

[18] P. A. Goloboff. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics*, 15:415-428, 1999.

[19] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52:696-704, 2003.

[20] P. H. Harvey and M. D. Pagel. *The comparative method in evolutionary biology*. Oxford University Press, Oxford, 1991.

[21] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97-109, 1970.

[22] M. D. Hendy and D. Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.*, 59:277-290, 1982.

[23] W. Hennig. *Phylogenetic systematics*. University of Illinois Press, Urbana, Illinois, 1966.

[24] D. M. Hillis. Inferring complex phylogenies. *Nature*, 383:130-131, 1996.

[25] D. M. Hillis, J. P. Huelsenbeck and C. W. Cunningham. Application and accuracy of molecular phylogenies. *Science*, 264:671-677, 1994.

[26] K. Howe, A. Bateman and R. Durbin. QuickTree: building huge neighbour-joining trees of protein sequences. *Bioinformatics*, 18:1546-1547, 2002.

[27] D. L. Hull. *Science as a process*. University of Chicago Press, Chicago, 1980.

[28] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23:254-267, 2006.

[29] M. Källersjö, V. A. Albert and J. S. Farris. Homoplasy increases phylogenetic structure. *Cladistics*, 15:91-93, 1999.

[30] M. Källersjö, J. S. Farris, M. W. Chase, B. Bremer, M. F. Fay, C. J. Humphries, G. Petersen, O. Seberg and K. Bremer. Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Pl. Syst. Evol.*, 213:259-287, 1998.

[31] H. G. Katzgraber. Introduction to Monte Carlo Methods. To appear in R. Leidl and A.K. Hartmann (eds.), *Modern Computational Science 09*, BIS-Verlag, Oldenburg, 2009.

[32] V. Knoop and K. Müller. Gene und Stammbäume. Ein Handbuch zur molekularen Phylogenetik. Spektrum Akademischer Verlag, Heidelberg, 2009.

[33] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.*, 50:913-925, 2001.

[34] T. Mailund and C. N. Pedersen. QuickJoin - fast neighbour-joining tree reconstruction, *Bioinformatics*, 20:3261-3262, 2004.

[35] J. O. McInerney, J. A. Cotton and D. Pisani. The prokaryotic tree of life: past, present ... and future? *Trends Ecol. Evol.*, 23:276-281, 2008.

[36] M. M. McMahon and M. J. Sanderson. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Syst. Biol.*, 55:818-836, 2006.

[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087-1092, 1953.

[38] K. C. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407-414, 1999.

[39] M. N. Price, P. S. Dehal and A. P. Arkin. FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, In press.

[40] F. Ronquist and J. P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572-1574, 2003.

[41] U. Roshan. *Fast algorithmic techniques for large-scale phylogenetic reconstruction.* PhD dissertation, University of Texas at Austin, 2004.

[42] M. J. Sanderson and J. Kim. Parametric phylogenetics? *Syst. Biol.*, 49:817-829, 2000.

[43] M. Schoniger and A. von Haeseler. A stochastic model and the evolution of autocorrelated DNA sequences. *Mol. Phylogen. Evol.*, 3:204-247, 1994.

[44] M. E. Siddall. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Cladistics*, 17:395-399, 2001.

[45] M. P. Simmons, L. B. Zhang, C. T. Webb and A. Reeves. How can third codon positions outperform first and second codon positions in phylogenetic inference? An empirical example from the seed plants. *Syst. Biol.*, 55:245-258, 2006.

[46] S. A. Smith, J. M. Beaulieu and M. J. Donoghue. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evol. Biol.*, 9:37, 2009.

[47] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22:2688-2690, 2006.

[48] A. Stamatakis, P. Hoover and J. Rougemont. A rapid bootstrap algorithm for the RAxML web-servers. *Syst. Biol.*, 75:758-771, 2008.

[49] M. Steel and A. Rodrigo. Maximum likelihood supertrees. *Syst. Biol.*, 57:243-250, 2008.

[50] S. Stefanovic, D. W. Rice and J. D. Palmer. Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evol. Biol.*, 4:35, 2004.

[51] D. L. Swofford. *PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts, 2002.

[52] D. L. Swofford. When are phylogeny estimates from molecular and morphological data incongruent? In M. M. Miyamoto and J. Cracraft (eds), *Phylogenetic analysis of DNA sequences* , pages 295-333, Oxford University Press, Oxford, 1991.

[53] D. L. Swofford, P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis and J. S. Rogers. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, 50:525-539, 2001.

[54] C. Tuffley and M. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.*, 59:581-607, 1997.

[55] R. A. Vos. Accelerated likelihood surface exploration: the likelihood ratchet. *Syst. Biol.*, 52:368-373, 2003.

[56] M. Zvelebil and J. Baum. *Understanding bioinformatics.* Garland Science, New York, 2007.