8 The Analysis of Molecular Sequences in Large Data Sets: Where Should We Put Our Effort?

W. C. Wheeler
Division of Invertebrate Zoology, American Museum of Natural History, New York City, USA

CONTENTS

ABSTRACT

The problems of nucleotide homology determination and tree search are intertwined and complex issues for phylogenetic reconstruction. Both present NP-hard optimisations. One step and two step heuristic procedures are reviewed and compared through the analysis of example data sets using multiple sequence alignment plus tree search and direct optimisation techniques. The examples here show that extraordinary effort on the tree search side cannot overcome the shortcomings of poor sequence homology heuristics. Direct optimisation using the most simple heuristics can offer solutions with 30% better optimality scores in larger data sets.

8.1 THE PROBLEM PRESENTED BY UNALIGNED SEQUENCE DATA

DNA sequences are rich sources of data for systematic analysis. Three problems rise above all others today in the use of such data: cladogram search, homology determination and optimality choice. All three of these are particularly relevant to the analysis of large collections of sequences from species rich taxa as well as smaller data sets. Cladogram search, given a set of homologies and an optimality criterion, has long been understood to be a computational challenge; homology determination less so. These are the topics of this discussion. The notion of whether to employ parsimony or likelihood as a criterion to identify 'best' hypotheses is beyond the scope of this discussion. Systematists expend a great deal of computational effort in optimal topology search, that is searching tree space for the best solution, supported by algorithmic research, especially in large (currently defined as >500 taxa) data sets. The result of this is clear; we are able to generate better (that is, more optimal) solutions than ever before. What of homology determination of unaligned DNA sequences? Phylogenetic trees of DNA sequences are based on assumptions of character homology (usually alignment of nucleotides), and although this process of homology determination is critical to the analysis, less attention has been paid to this component of the problem than cladogram search or choice of optimality criterion. We would like to know how homology determination compares to cladogram search in influencing the final result.

8.1.1 Two Step versus One Step Analysis

Phylogenetic analysis of a sequence data set often begins with a process of multiple alignment, where unaligned sequence data are arrayed in a series of columns via the insertion of gaps denoting insertion–deletion (indel) events. These aligned data now have an equal number of positions forming a putative or primary homology statement scheme sensu DePinna[1]. The aligned data are then subjected to a cladogram search (for example, using PAUP[2] or TNT[3]). This is referred to as 'two step' phylogenetics[4] in opposition to the 'one step' optimisation approach[5]. In this methodology, there is no separation of a homology determination phase from that of cladogram diagnosis and evaluation[6–8]. Such methods attempt to create an optimal set of homologies specific to each cladogram topology. As such, they do not create multiple alignments for processing through standard phylogeny reconstruction programs. The transformation series specific to each of these topologies can be arrayed and presented in the form of an implied alignment[9]. The use of an implied alignment from an optimisation analysis as a multiple alignment is problematic, since it was constructed with a particular phylogenetic hypothesis (cladogram) as a foundation; hence its use to evaluate competing topologies may seem 'unfair'. This does, however, provide a framework for comparison of the effectiveness of analysis, since optimality values can be directly compared.

8.1.2 NP-Completeness

The complexities of the problem of converting observations into cladograms come from the difficulty in optimising its two components. The joint problem is composed of two NP-complete (nondeterministic polynomial complete) problems. Both cladogram search and the assignment of

optimal ancestral sequences, such that any overall tree is optimal, are NP-hard optimisations[10]. Hence, heuristic solutions are required to find usable solutions to both of these problems. Their joint nature makes the challenge even greater. Two step multiple alignment analysis seeks to simplify the problem by separating homology and cladogram search into separate, tractable operations. One step optimisation methods attack the issue as a nested problem, dealing directly with the complexity of both operations.

8.1.3 What to Do?

Given a world of finite computational resources, how do we apportion our effort? Much has been achieved in the realm of cladogram search heuristics, increasing the manageable size of data sets from tens to thousands in the past few years[3,11]. Are these advances paying off in the homology problem of unaligned sequences?

8.2 CLADOGRAM SEARCH HEURISTICS

Ever larger data sets have required increasingly powerful heuristic search procedures to search tree space (the set of possible trees given the number of taxa). Briefly reviewed here are several major approaches.

8.2.1 Initial Build

Often called the Wagner procedure or Wagner tree[12], initial cladogram construction is usually a straightforward procedure where taxa are added in turn to each possible location on a cladogram (Figure 8.1). The computational complexity of this operation is proportional to the square of the number of taxa [$O(n^2)$]. This operation is repeated until all taxa are added and an initial, complete cladogram is produced.

8.2.2 Local Search Refinement

The initial Wagner tree may be complete, but it is quite unlikely to be very near a usefully optimal value. To improve on this solution, tree rearrangements can be performed where the basic Wagner tree is progressively rearranged and improved (Figure 8.2). These refinement operations yield a local search within the optimality neighbourhood defined by the initial tree (and include such methods as nearest neighbour interchange (NNI), subtree pruning and reconnection (SPR) and tree bisection and reconnection (TBR), depending on the level of rearrangement[13]).

8.2.3 Random Starting Points

Both the initial build and rearrangement are trajectory methods that follow a predefined and completely repeatable sequence to find a solution. An early method to attempt to break out of the local minima found in such searches was the randomisation of the taxon addition sequence used in the initial Wagner build. Such randomisations greatly improved search results and have been components in phylogenetic software for some time[14,15].

8.2.4 Simulated Annealing

As with many complex optimisation problems, simple heuristics can get stuck in local optima. Randomisation can help find more global solutions, but local search techniques such as NNI, SPR and TBR are prone to wallowing in local minima. This is due to the fact that the path to a more

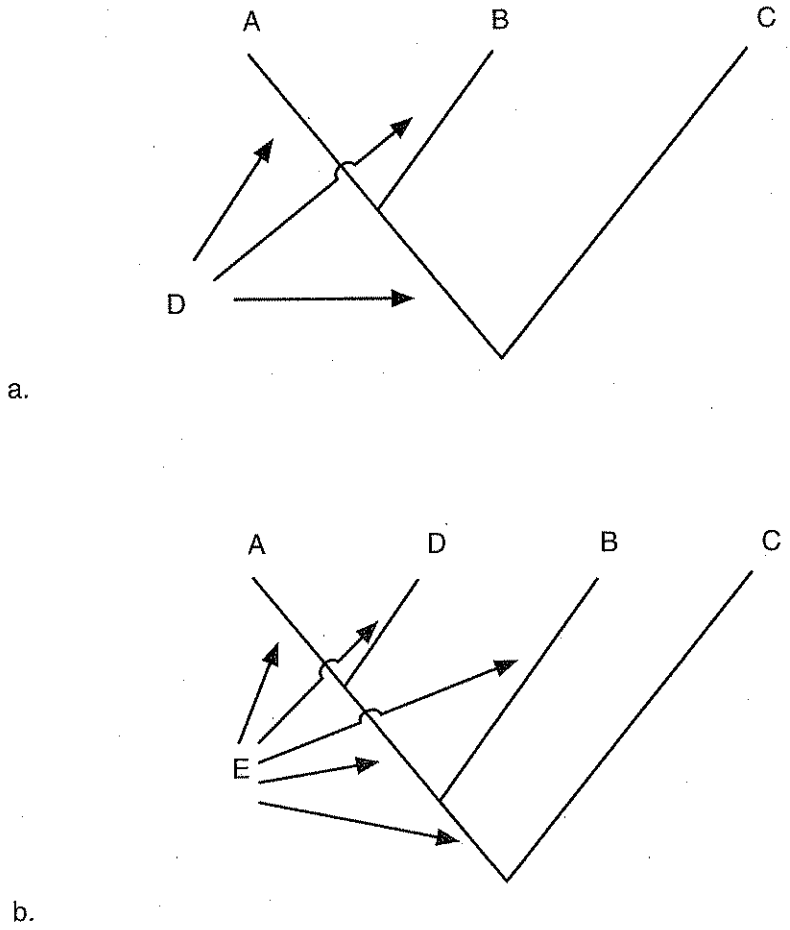

FIGURE 8.1 Basic Wagner build procedure of Farris[12] showing the addition of each taxon in turn to each possible edge (branch) on the tree. (a) The fourth taxon D is added to each of three places on the rooted tree. (b) The fifth taxon E is added at each of five positions.

globally satisfying solution may require traveling through suboptimal, intermediate states (Figure 8.3). This is the situation presented by the annealing of metals and applied to computational problems by Metropolis et al.[16].

Ratcheting. Nixon[17] brought simulated annealing to phylogenetic analysis. In Nixon's approach, called the 'ratchet', characters are randomly reweighted and searches performed on the newly weighted data. The weights are then set back to their initial values, and a search is performed with the reweighted tree as a starting point. The method has been extremely effective in finding lower-cost solutions in data sets thought to be refractory to further analysis, such as a large angiosperm matrix[18].

Drifting. A phylogenetic method much closer to the original description of simulated annealing was proposed by Goloboff[11] and termed 'drifting'. Unlike the ratchet, where suboptimal solutions were arrived at via weighting, drifting explicitly creates a probability of topology acceptance based on the extent of its suboptimality. This is implemented in TNT[3].

Monte Carlo Markov Chain. A probabilistic form of local search uses the relative probability of successive tree rearrangements as a criterion for the acceptance of a rearrangement. As with drifting and other simulated annealing techniques, suboptimal solutions can be accepted as intermediate solutions on the way to more globally optimal scenarios. As a search strategy, Monte Carlo Markov Chains have had their greatest impact on Bayesian estimates of clade probabilities[19].

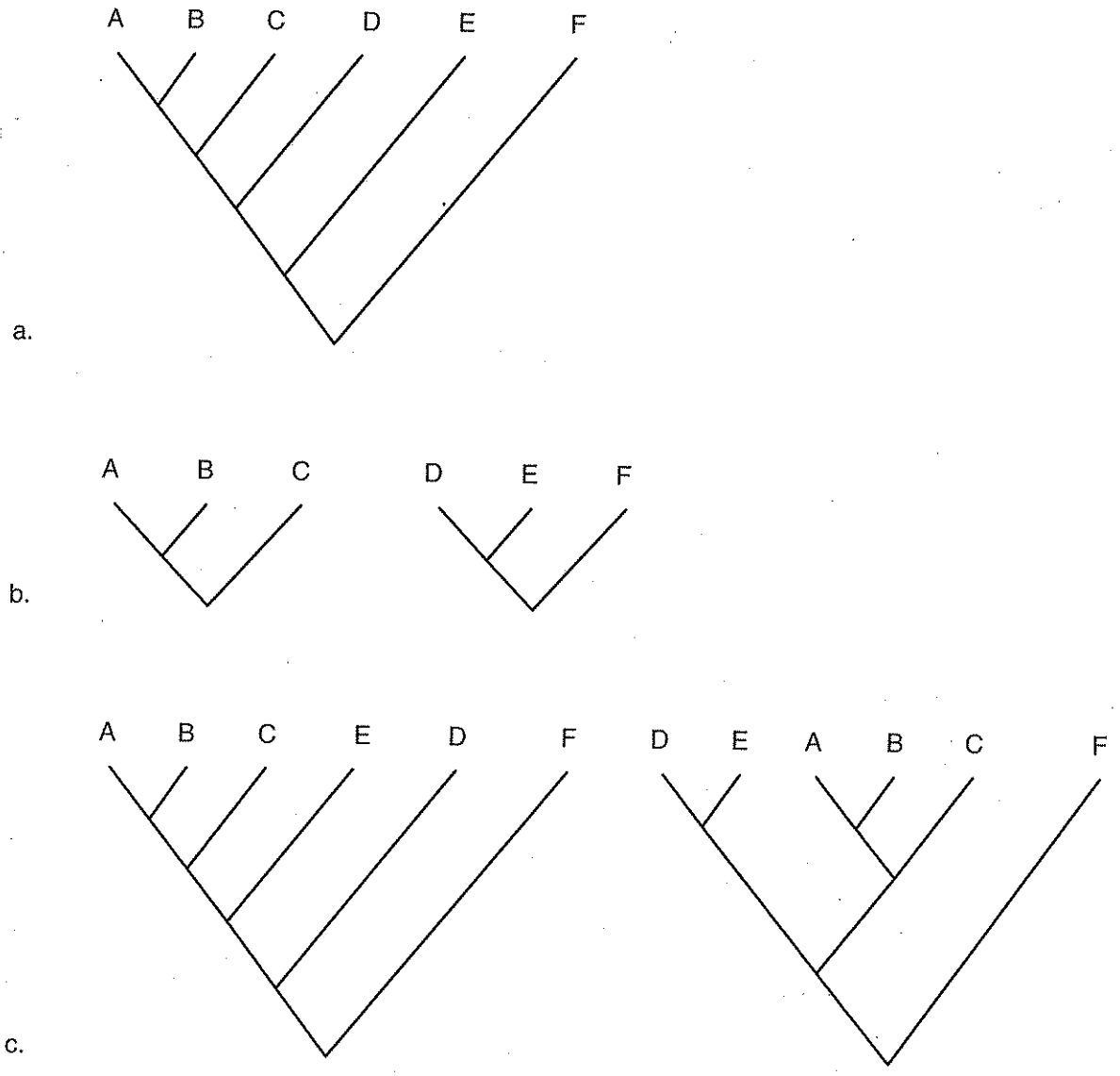

FIGURE 8.2 Simple tree rearrangement showing SPR branch swapping. Clade of tree (a) is pruned off leaving two subtrees (b). The subtree is then added back to each possible place on the subtree (c), avoiding its original position and yielding new trees closely related topologically to the first.

8.2.5 Genetical Algorithm

An extremely productive class of algorithms mimic the process of natural selection and go collectively under the name 'genetical algorithms' (GA). Unlike trajectory searches and simulated annealing perturbations, GA methods operate on collections or populations of trees. This was first introduced to systematics by Moilanen[20]. There are usually three steps to such a procedure, namely: generation of variation, recombination and selection. Beginning with a collection of locally optimal or near optimal trees, variation can be generated variously, drifting and ratcheting being example strategies. After this, a recombination stage occurs where, in this case, trees exchange compatible branches (Figure 8.4; this has been called 'tree fusing' by Goloboff[11,13]). Selection then takes place based on the optimality values at hand, retaining optimal and perhaps suboptimal trees at an intermediate stage to preserve variation. Combined search strategies implemented in TNT[3] have reduced search times for some iconic data sets (for example, 'Zilla'[21]) from months to minutes on commodity hardware.

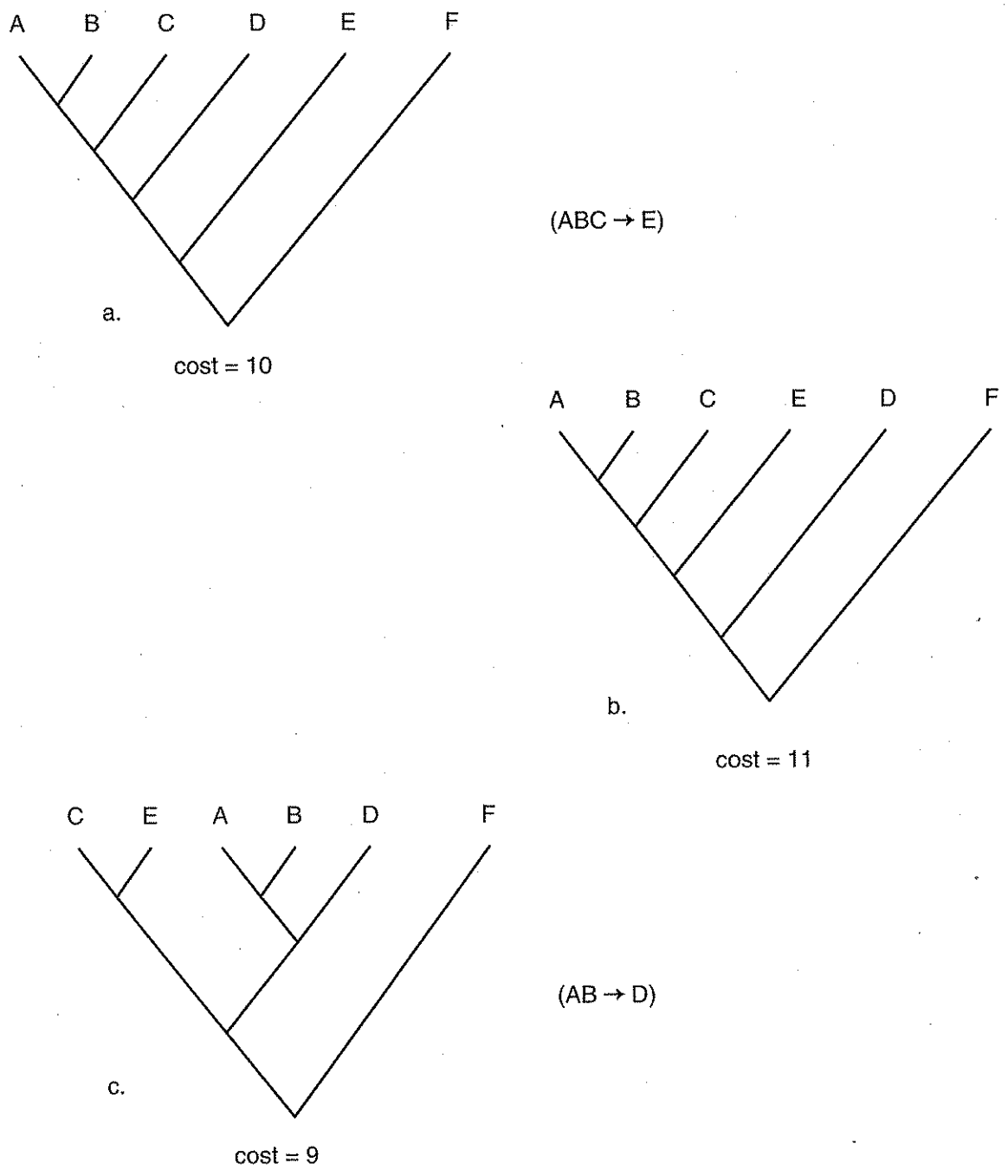

FIGURE 8.3 A simulated annealing trajectory. Globally optimal topology (c) can only be reached from locally optimal (a) by passing through the suboptimal topology (b).

8.2.6 Sectorial Searches

Goloboff[11] introduced the notion of effective taxon reduction as a means of focusing the search on the relationships among segments or sectors of a large tree. The central idea is that there may be subsets of taxa in optimal or near optimal arrangements that are in suboptimal arrangements with respect to each other. Goloboff[11] showed that random addition of taxa and simple TBR branch swapping can be very effective at creating the well ordered sectors of a cladogram, but when the number of taxa approaches hundreds or thousands, this strategy is much less effective at determining the overall structure of the tree. Basically, the topology of a candidate tree is divided into a series of

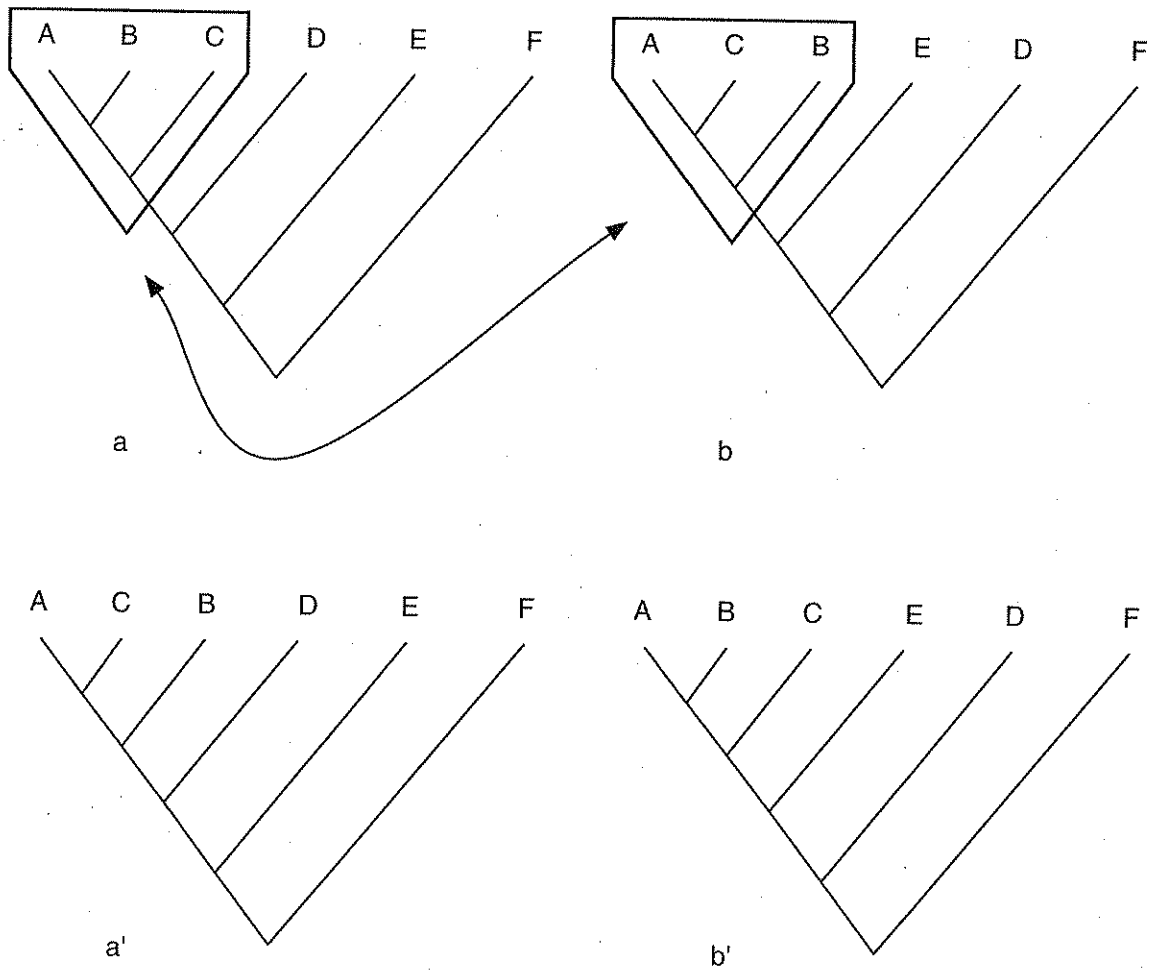

FIGURE 8.4 Tree fusing (Goloboff[11]) component of genetical algorithm (Moilanen[20]). Cladograms a and b exchange the (ABC) groups yielding two new arrangements a′ and b′.

sectors of 35–50 taxa that are then treated as a single terminal (Figure 8.5). Branch swapping is then performed on this reduced data set. Sectors and searches are dynamically defined and alternated as the topology evolves until a stable solution is found. This approach has been further explored as 'disk covering methods'[22,23] (see also Wilkinson and Cotton, *Chapter 5;* Bininda-Emonds and Stamatakis, *Chapter 6*), yielding improvements in many areas of phylogenetic tree searching.

8.2.7 Complex Search Strategies

The search methods described above are most profitably used in concert. This approach was advocated by Goloboff[11] when he first described tree fusing, drifting and sectorial searches. Goloboff described a variety of combinatorial strategies for different size data sets with different analytical properties. This sort of flexible approach is implemented in TNT[3], allowing the user to explore these procedures and their combinations to solve very large (thousands of taxa) data sets.

8.3 HOMOLOGY DETERMINATION HEURISTICS

As pointed out in Wheeler et al.[5], multiple sequence alignment and optimisation techniques can be viewed as alternate modes of homology heuristic. Both strive to create optimal (that is, lowest cost) cladograms (but see Simmons[24] for a different view), yet the approaches differ significantly.

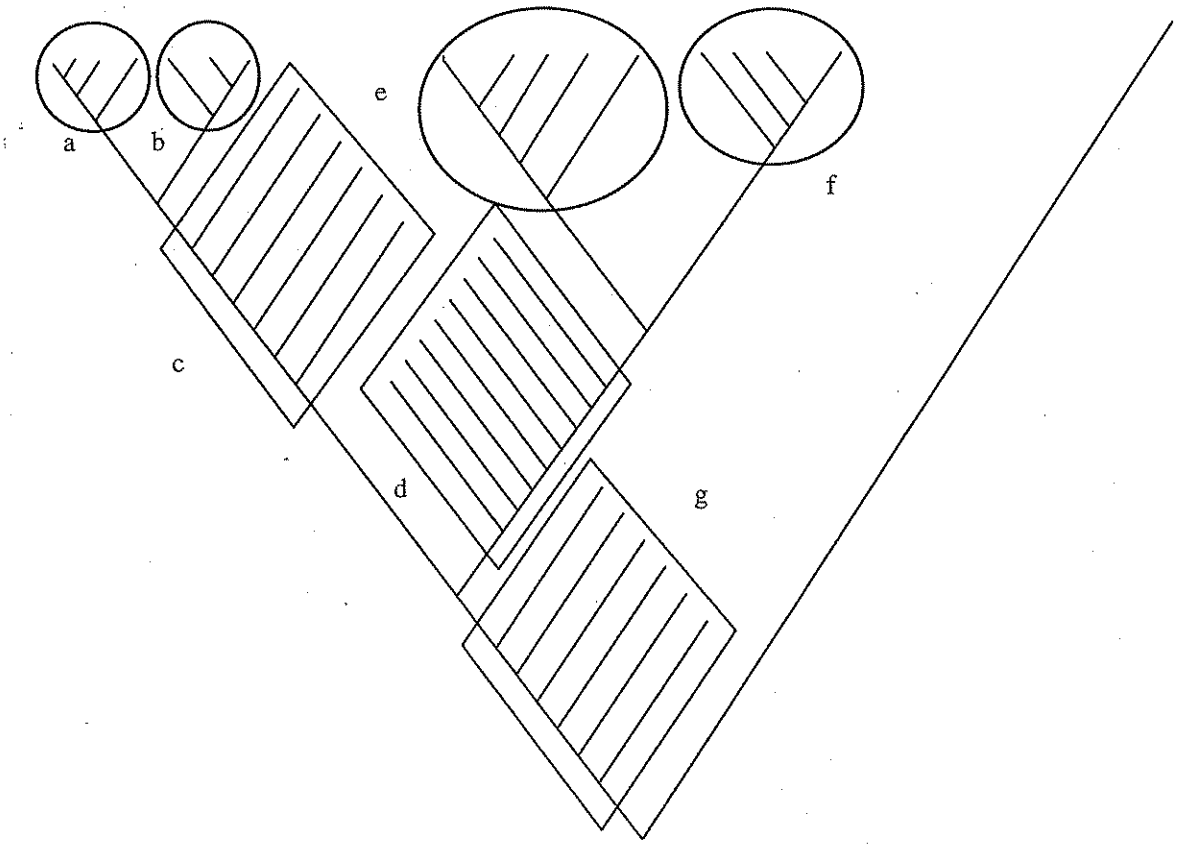

FIGURE 8.5 Definition of cladogram sectors for use in a sectorial search (Goloboff[11]); (a–g) usually 35–50 taxa.

Multiple alignment methods seek, in general, a single set of homologies, upon which all cladograms are evaluated, whereas optimisation methods create potentially unique schemes for each cladogram. The fundamental methods are briefly reviewed below.

8.3.1 Multiple Alignment

As described by Sankoff and Cedergren[25], an exact multiple alignment can be constructed via a recursive method, if the tree of relationships is known. This will have a time complexity of $O(n^k 2^k)$ for k sequences of length n; an impossibly large number for real data sets. This represents the enumeration of all possible multiple alignments for a given tree, described by Wang and Jiang[10] as a 'tree alignment'. Two problems are presented by this approach: time complexity and the lack of a 'known' tree. Sankoff and Cedergren[25] suggest an approximation in $O(n^3)$ time, but even this can be extremely daunting. In general, multiple alignment methods reduce the problem to a series of k–1 pairwise $[O(n^2)]$ alignments using a guide tree. Since the alignments are performed two at a time, no phylogenetic tree is required to determine alignment costs (Figure 8.6). This type of approach is commonly used, for example in CLUSTALW[26] reviewed by Phillips et al.[27].

One of the key issues is the generation of the guide tree. CLUSTALW generates a neighbour joining distance tree[28] based on pairwise alignment distances. Other methods used an interactive tree alignment building procedure (TREEALIGN[29,30]) or an explicit search on multiple guide trees attempting to find the alignment of lowest cost (MALIGN[31,32]). In general, pairwise, guide tree–based multiple alignment procedures are fast, but tend to become coarse as the number of sequences increases[27].

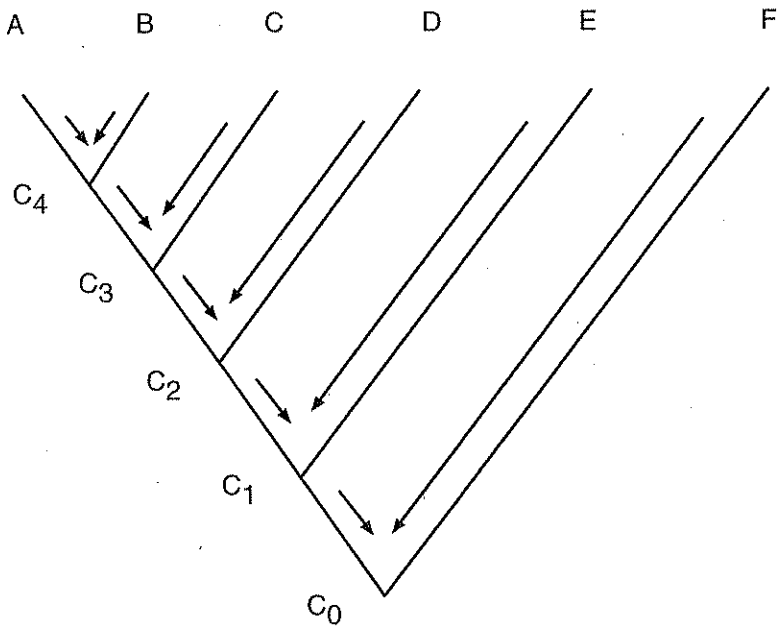

FIGURE 8.6 Guide tree-based multiple alignment. Sequences are accreted in turn as the procedure moves from the tips of the tree (A–E) to the root. Intermediate vertices (C_i) may be consensus sequences as in CLUSTALW (Thompson et al.[26]) or partial alignments as in MALIGN[32].

8.3.2 Optimisation Approaches

As opposed to multiple alignment methods, optimisation approaches seek to deal with the cladogram directly by determining the sequence states of internal vertices without the use of a pre-existing alignment. Each topology, in essence, would be granted its own set of ancestral sequences and transformations, hence homology relationships. Sankoff[33] pioneered this approach with an $O(n^k)$ for k sequences of length n (as above). As with multiple alignment, this time complexity placed the procedure well beyond the reach of real data sets. A series of heuristic solutions to this problem have been proposed for parsimony[6,34–36] and likelihood[7,8,37]. The simplifications can be divided into approaches where medians are calculated from two and three sequences to reduce the complexity to a manageable level, and those where no medians are calculated, but the set of possible vertex sequences specified a priori.

Median approaches. Median-based heuristics calculate vertex sequences from adjacent nodes (Figure 8.7). This is done with the same objective as the exact case[33], so that the total cladogram cost be minimised. An $O(n^2)$ method was described by Wheeler (1996) that used string matching to create a median sequence with the constraint that there are no sequence gaps in the median sequences. This was improved[35] in the sense of a better (that is, lower cost) median, but with additional complexity [$O(n^3)$] based on Sankoff and Blanchette[38] and Gladstein[39], again without sequence gaps in the medians. Likelihood medians have been derived for complex models[7,8] and simple ones[37]. Given the greater numerical complexity of likelihood calculations, these procedures are far more time consuming than parsimony based methods.

Search-based approaches. Unlike median based heuristics, search-based approaches[36,40] make no effort to calculate sequences de novo, but choose them from a prespecified set. This has the benefit of being relatively rapid when the set is not too large. After determining the edit cost between each pair of candidate sequences (a series of pairwise alignments), dynamic programming[41] is used to choose the optimal set of sequences for each vertex (Figure 8.8). For a single cladogram, with n sequences of length m and a set of k additional candidate sequences, such an optimisation would require

$$O(n \cdot (n+k)^2)$$

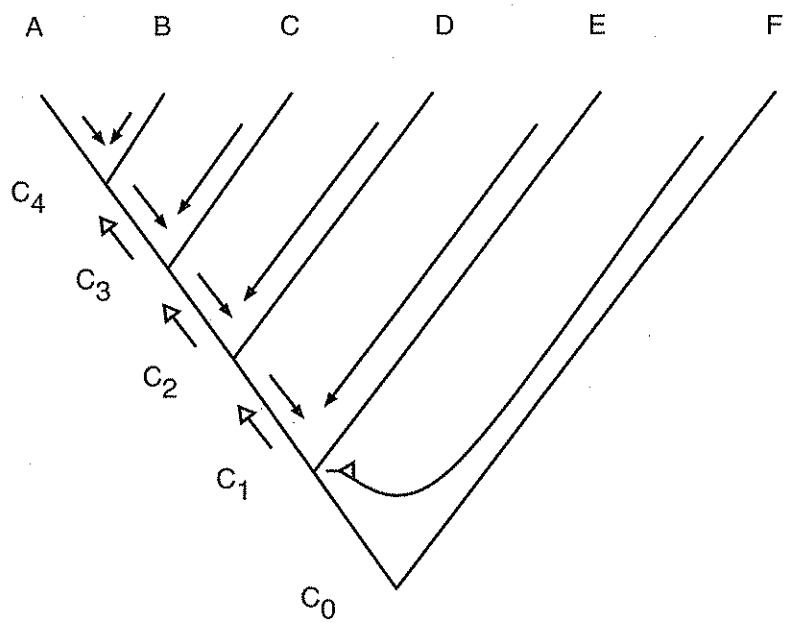

FIGURE 8.7 $O(n^2)$ and $O(n^3)$ sequence optimisation medians. $O(n^2)$ (closed arrows) and $O(n^3)$ (closed and open arrows). The median sequences (C_i) are calculated based on either their two descendants, or their descendants and immediate ancestor. C_5 would not be calculated in the $O(n^3)$ case. Multiple passes may be performed on the cladograms to update the vertex sequences and improve median quality.

whereas a simple median approach would require $O(n \cdot m^2)$. As long as $n + k < m$, search-based methods should win out as far as time is concerned. The set size k, however, will determine the quality (cost) of the result, with k varying from 0 in fixed state optimisation[40] to the set of all sequences resulting in an exact solution through explicit enumeration[36]. Little work has been done to examine how large k should be, or how best to choose the sequences to be included in the heuristic set.

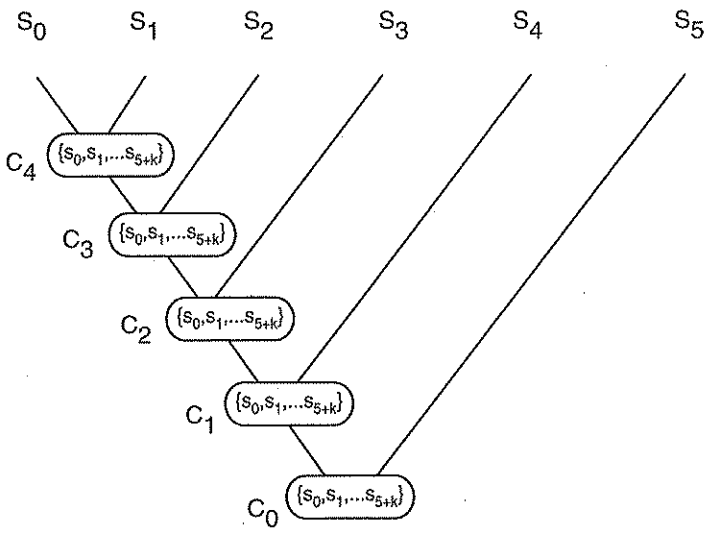

FIGURE 8.8 Search-based optimisation of a set of observed sequences (S_i) to determine vertex sequences (C_j) using a set of candidate sequences ($S_0,\ldots,S_{5+k}$).

8.4 EXAMPLE DATA

8.4.1 Data Sets

As a demonstration of the effects of these procedures on real data, four collections of unaligned sequences were analysed. The data sets were the 62-taxon set of mantid (Mantodea) 18S rDNA from Svenson and Whiting[42], a 208-sequence collection (18S rRNA) of Metozoa from G. Giribet (personal communication), a 585-taxon archaeal small subunit sequence data set from the European Ribosomal RNA Database (http://www.psb.ugent.be), and a 1,040 mitochondrial small subunit data set from the same source.

8.4.2 CLUSTALW Alignment

CLUSTALW[26] was used to align the four data sets under two sets of conditions. The first was with the default pairwise and multiple alignment parameters (initial indel cost of 10 and extension indel cost of 0.2). A second run was performed where all events were set to 1. Alignments were generated and then analysed using TNT[3] to assay the optimality (tree length) of the homology statements in terms of equal weighted parsimony. This was done under basic simple (10 random addition sequences and TBR swapping) and more aggressive searches (options 'mxram 1,000 hold 10,000 xmult = replications 10 ratchet 50 drift 20 fuse 5') (Table 8.1).

8.4.3 POY Optimisation

POY[43] was used to perform optimisation-based analyses. Searches were done under cost parameters matching CLUSTALW default and equal weighting scenarios. Searches were performed using direct optimization [$O(n^2)$ medians] and simple single addition sequence Wagner build

TABLE 8.1
Summary of CLUSTALW Multiple Alignment and POY Optimisation Analyses

Data Set	Parameters	Taxa	Aligned Length	Variable Positions	POY Cost	TNT Simple	TNT Aggressive
Mantodea	CLUS Default	62	1,873	510	—	1,052	1,052
	Equal		1,828	479	—	1,037	1,037
Metazoa	CLUS Default	208	2,868	2,521	—	30,980	30,980
	Equal		4,263	3,888	—	29,089	29,089
Archaea	CLUS Default	585	2,722	2,516	—	39,084	39,062
	Equal		1,768	1,679	—	39,533	39,499
Mitochondria	CLUS Default	1,040	3,054	3,002	—	90,640	90,619
	Equal		7,662	7,162	—	115,686	115,649
Mantodea	CLUS Default	62	1,900	847	4,582	1,292	1,292
	Equal		1,990	1,900	941	981	981
Metazoa	CLUS Default	208	21,430	21,238	171,949	48,675	48,675
	Equal		6,190	5,838	27,146	26,783	26,783
Archaea	CLUS Default	585	20,831	20,765	208,060	54,733	54,717
	Equal		6,725	6,682	37,931	37,751	37,750
Mitochondria	CLUS Default	1,040	30,648	30,638	486,950	132,586	132,529
	Equal		13,978	13,978	79,769	79,594	79,593

Note: Results from CLUSTALW are shown above the line (upper part of table) and those of POY shown below the line (lower part of table). 'CLUS Default' denotes CLUSTALW default parameters, 'Equal' all events = 1. The POY costs for CLUS Default runs are high due to the parameter setting ('-gap 50 -extensiongap 1 -change 5) in those runs. 'TNT Simple' denotes the TNT command 'mult', whilst 'TNT Aggressive' signifies 'xmult=replications 10 ratchet 50 drift 20 fuse 5'. The rightmost three column values are cladogram costs.

Source: Data from CLUSTALW Thompson et al.[26] and POY Wheeler et al.[43]

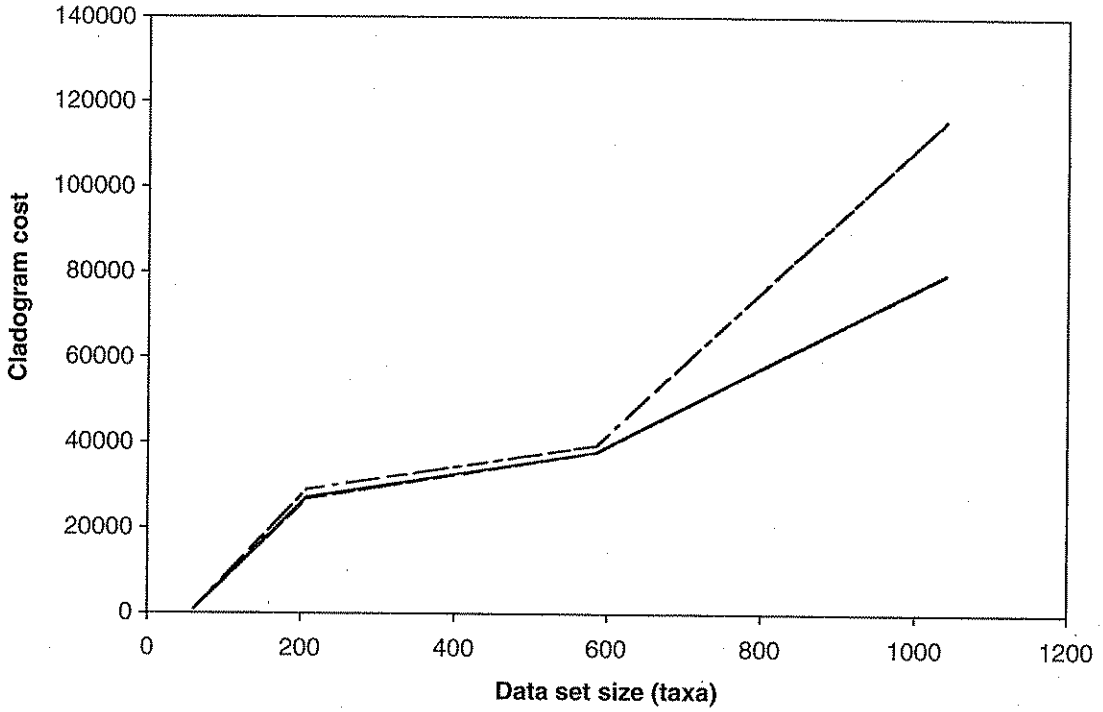

FIGURE 8.9 Cladogram optimality as a function of data set size based on the 1:1:1 values of Table 8.1. Dashed line represents POY, POY-TNT 'Simple' and POY-TNT 'Aggressive'. Solid line represents CLUSTALW-TNT 'Simple' and CLUSTALW-TNT 'Aggressive' (full values can be seen in Table 8.1).

without further refinement. Implied alignments[9] were created to allow for direct comparison with CLUSTALW results. These were subjected to the same TNT analysis conditions as the CLUSTALW alignments to find the tree lengths.

8.4.4 Results

The results of the CLUSTALW multiple alignment and POY optimisation analyses are shown in Table 8.1 and Figure 8.9.

8.5 COMPARISONS

Several patterns are immediately apparent. The CLUSTALW alignments do not differ much in optimality value (tree length) from the default to equal weighting scenarios. Given that the relative indel costs differ by a factor of 50, this is striking. POY analyses contrast sharply with this. The Mantodea data show a difference in equally weighted TNT cost (tree length) of 1.4%, whereas the POY runs are 32% different; Metazoa data show 6.5% (CLUSTALW) and 81% (POY), and for Archaea data the difference was 1.1% (CLUSTALW) and 44% (POY). The mitochondrial data set has a 22% difference for CLUSTALW, the highest of the alignment-based runs, but still lower than all the optimisation comparisons. This difference is so great that the CLUSTALW alignments were superior to the POY optimisations in every case where the homology and cladogram cost parameters differed (CLUSTALW default settings). Whilst the POY optimisation analyses are very responsive to cost parameters, the CLUSTALW runs are not. Whilst each case where equal weighted optimisation was used yielded superior (that is, lower cost) optimality values for POY optimisation, only half of the alignment cases showed this pattern.

The two methods also differed in their response to increased severity of cladogram search heuristics. Neither CLUSTALW nor POY implied alignments displayed any better solutions under

more exhaustive cladogram searching for the smaller Mantodea or Metazoa data sets. The 585-taxon archaeal and 1,040-taxon mitochondrial data sets did, with the CLUSTALW showing an average improvement factor of 4.93×10^{-4} and the POY a factor of 1.09×10^{-4}. The results based on these simple POY homology searches were from 4.7% (Metazoa) to 45% (mitochondrial) less costly than those based on the CLUSTALW alignments. This is especially pointed in concert with the search severity improvement being 20% as great for the POY runs. Cladogram search effort is much more productive for the CLUSTALW alignments. Even with the very aggressive phylogenetic search options of TNT, in no case where the homology and search parameters were the same did the CLUSTALW alignment match the cost of the rudimentary Wagner build procedure used in the POY optimisation.

8.6 WHAT IS HAPPENING IN LARGE DATA SETS?

As data sets grow (more taxa are added), however, the relative importance of homology and cladogram search heuristics change. For small data sets, homology heuristics seem to be relatively capable, and cladogram searching is more important. As the data sets grow, however, the importance of homology determination increases, eventually dominating the result. Multiple sequence alignment, at least as implemented in CLUSTALW, is not effective enough. The comparisons here are based on only the most simple homology heuristics. Methods such as more aggressive median calculation[35] and search[36] coupled with better cladogram search, will likely yield results 10% lower in cost for optimisation techniques, making the comparison even more stark.

Whilst tree search space is relatively well studied and understood, we are only at the beginning stages of understanding the space of homology optimisation. Homology problems are at least as computationally complex and certainly less well known. Increased understanding in this area will no doubt yield much greater improvements in the optimality of our results in the future. Cladogram searching has undergone a huge change in the last ten years; homology assessment needs the same amount of attention. A good cladogram search can never make up for a poor homology search. Systematics requires that more attention be paid, in both methodological innovation and computer time, to homology determination.

ACKNOWLEDGEMENTS

The US National Science Foundation and NASA for research support. Louise Crowley, Gonzalo Giribet, Megan Harrison, Camilo Mattoni, Kurt Pickett and Andrés Varón for data sets, discussion and commentary on this manuscript. The temporal forbearance of Trevor Hodkinson and John Parnell while reviewing this paper.

REFERENCES

1. DePinna, M.C.C., Concepts and tests of homology in the cladistic paradigm, *Cladistics,* 7, 367, 1991.
2. Swofford, D.L., *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods),* version 4.0b 10, Sinauer Associates, Sunderland, MA, 2002.
3. Goloboff, P.A., Farris, J.S., and Nixon, K., TNT (Tree analysis using New Technology) version 1.0 ver. beta test v. 0.2., 2003, Tucumán, Argentina (http://www.zmuc.dk/public/phylogeny/tnt).
4. Giribet, G., Generating implied alignments under direct optimization using POY, *Cladistics,* 21, 396, 2005.
5. Wheeler, W.C. et al., *Dynamic Homology and Phylogenetic Systematics: A Unified Approach Using POY,* American Museum of Natural History, 2005.
6. Wheeler, W.C., Optimization alignment: the end of multiple sequence alignment in phylogenetics? *Cladistics,* 12, 1, 1996.
7. Hein, J. et al., Statistical alignment: computational properties, homology testing, and goodness-of-fit, *J. Mol. Biol.,* 302, 265, 2000.

8. Hein, J., Jensen, C.J.L., and Pedersen C.N.S., Recursions for statistical multiple alignment, *Proc. Natl. Acad. Sci. USA,* 100, 14960, 2003.
9. Wheeler, W.C., Implied alignment, *Cladistics,* 19, 261, 2003.
10. Wang, L. and Jiang, T., On the complexity of multiple sequence alignment, *J. Comput. Biol.,* 1, 337, 1994.
11. Goloboff, P.A., Analyzing large data sets in reasonable times: solutions for composite optima, *Cladistics,* 15, 415, 1999.
12. Farris, J.S., A method for computing Wagner trees. *Syst. Zool.,* 19, 83, 1970.
13. Goloboff, P.A., Techniques for analysing large data sets, in *Techniques in Molecular Systematics and Evolution,* DeSalle, R., Giribet, G. and Wheeler, W., Eds., Birkhäuser Verlag, Basel, 2002, 7.
14. Felsenstein, J., *PHYLIP,* 1980 (http://evolution.genetics.washington.edu/phylip.html)
15. Mickevich, M.F. and Farris, J.S., *PHYSYS: Phylogenetic Analysis System,* 1980.
16. Metropolis, N.A. et al., Equation of state calculations by fast computing machine, *J. Chem. Phys,* 21, 1087, 1953.
17. Nixon, K.C., The parsimony ratchet, a new method for rapid parsimony analysis, *Cladistics,* 15, 407, 1999.
18. Rice, K.A., Donoghue, M.J., and Olmstead. R.G., Analyzing large data sets: rbcl 500 revisited, *Syst. Biol.,* 46, 554, 1997.
19. Huelsenbeck, J.P. and Ronquist, F., *MrBayes: Bayesian inference of phylogeny,* 3.0 edition, 2003. (http://mrbayes.csit.fsu.edu).
20. Moilanen, A., Searching for most parsimonious trees with simulated evolutionary optimization, *Cladistics,* 15, 39, 1999.
21. Chase, M.W. et al., Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcl, *Ann. Mol. Bot. Gard.,* 80, 528, 1993.
22. Huson, D., Nettles, S., and Warnow, T., Disk-covering, a fast converging method for phylogenetic tree reconstruction, *J. Comput. Biol.,* 6, 368, 1999.
23. Roshan, U., et al., Rec-i-dcm3: A fast algorithmic technique for reconstructing large phylogenetic tree, in *Proc. IEEE Computer Society Bioinformatics Conference CSB 2004,* Stanford University, 2004.
24. Simmons, M.P., Independence of alignment and tree search, *Mol. Phylogenet. Evol.,* 31, 874, 2004.
25. Sankoff, D.M. and Cedergren, R.J., Simultaneous comparison of three or more sequences related by a tree, in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison,* Sankoff, D.M. and Kruskall, J.B., Eds., Addison Wesley, Reading, MA, 1983, chap. 9.
26. Thompson, J.D., Higgins, D.G., and Gibson, T.J., 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, 1994.
27. Phillips, A., Janies, D., and Wheeler. W., Multiple sequence alignment in phylogenetic analysis, *Mol. Phylogenet. Evol.,* 16, 317, 2000.
28. Saitou, N. and Nei. M., The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.,* 4, 406, 1987.
29. Hein, J., A new method that simultaneously aligns and reconstruct ancestral sequences for any number of homologous sequences, when the phylogeny is given, *Mol. Biol. Evol.,* 6, 649, 1989.
30. Hein, J., A tree reconstruction method that is economical in the number of pairwise comparisons used, *Mol. Biol. Evol.,* 6, 669, 1989.
31. Wheeler, W.C., Sources of ambiguity in nucleic acid sequence alignment, in *Molecular Ecology and Evolution: Approaches and Applications,* Schierwater, G.W.B., Streit, B., and DeSalle, R., Eds., Birkhäuser Verlag, Basel Switzerland, 1994, 323.
32. Wheeler, W.C. and Gladstein, D.S., (documentation by Janies, D. and Wheeler, W.C.), *MALIGN,* New York, NY, 1991–1998 (http://research.amnh.org/scicomp/projects/malign.php).
33. Sankoff, D.M., Minimal mutation trees of sequences, *SIAM J. Appl. Math.,* 28, 35, 1975.
34. Wheeler, W.C., Fixed character states and the optimization of molecular sequence data, *Cladistics,* 15, 379, 1999.
35. Wheeler, W.C., Iterative pass optimization, *Cladistics,* 19, 254, 2003.
36. Wheeler, W.C., Search-based character optimization, *Cladistics,* 19, 348, 2003.
37. Wheeler, W.C., Dynamic homology and the likelihood criterion, *Cladistics,* 2005.

38. Sankoff, D.M. and Blanchette, M., The median problem for breakpoints in comparative genomics, *Computing and Combinatorics 3rd Annual Int. Conf. COCOON 97,* 1276, 251, 1997.
39. Gladstein, D.S., Efficient incremental character optimization, *Cladistics,* 13, 21, 1997.
40. Wheeler, W.C., Measuring topological congruence by extending character techniques, *Cladistics,* 15, 131, 1999.
41. Sankoff, D.M. and Rousseau, P., Locating the vertices of a Steiner tree in arbitrary space, *Math. Program.,* 9, 240, 1975.
42. Svenson, G.J., and Whiting, M.F., Phylogeny of Mantodea based on molecular data: evolution of a charismatic predator, *Syst. Ent.,* 29, 359, 2004.
43. Wheeler, W.C., Gladstein, D.S., and De Laet, J.D., (documentation by Janies, D. and Wheeler, W.C.—commandline documentation by De Laet, J.D. and Wheeler W.C.), *POY* version 3.0.11, American Museum of Natural History, New York, 1996–2005 (http://research.amnh.org/scicomp/projects/poy.php).