



Scaling up accurate phylogenetic reconstruction from gene-order data

Jijun Tang and Bernard M.E. Moret*

Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Phylogenetic reconstruction from gene-order data has attracted increasing attention from both biologists and computer scientists over the last few years. Methods used in reconstruction include distance-based methods (such as neighbor-joining), parsimony methods using sequence-based encodings, Bayesian approaches, and direct optimization. The latter, pioneered by Sankoff and extended by us with the software suite GRAPPA, is the most accurate approach, but cannot handle more than about 15 genomes of limited size (e.g. organelles).

Results: We report here on our successful efforts to scale up direct optimization through a two-step approach: the first step decomposes the dataset into smaller pieces and runs the direct optimization (GRAPPA) on the smaller pieces, while the second step builds a tree from the results obtained on the smaller pieces. We used the sophisticated disk-covering method (DCM) pioneered by Warnow and her group, suitably modified to take into account the computational limitations of GRAPPA. We find that DCM-GRAPPA scales gracefully to at least 1000 genomes of a few hundred genes each and retains surprisingly high accuracy throughout the range: in our experiments, the topological error rate rarely exceeded a few percent. Thus, reconstruction based on gene-order data can now be accomplished with high accuracy on datasets of significant size.

Availability: All of our software is available in source form under GPL at <http://www.compbio.unm.edu>

Contact: moret@cs.unm.edu

INTRODUCTION

Biologists can infer the ordering and strandedness of genes on a chromosome, and thus represent each chromosome by an ordering of signed genes (where the sign indicates the strand). These gene orders can be rearranged by evolutionary events such as inversions and transpositions and, because they evolve slowly, give biologists an important new source of data for phylogeny

reconstruction—see, e.g. Downie and Palmer (1992), Olmstead and Palmer (1994), Palmer (1992), and Rauberson and Jansen (1992). Appropriate tools for analyzing such data may help resolve some difficult phylogenetic reconstruction problems. Developing such tools is thus an important area of research—indeed, the recent DCAF symposium organized by Sankoff and Nadeau (2000) was devoted to this topic.

A natural optimization problem for phylogeny reconstruction from gene-order data is to reconstruct an evolutionary scenario with a minimum number of the permitted evolutionary events on the tree. This problem is NP-hard for most criteria—even the very simple problem of computing the median of just *three* genomes (the median of k genomes is a genome that minimizes the sum of the pairwise distances between itself and each of the k given genomes) under such models was proved NP-hard by Pe'er and Shamir (1998) and Caprara (1999).

For some datasets (e.g. chloroplast genomes of land plants), biologists conjecture that rearrangement events are predominantly *inversions* (also called reversals). In other datasets, transpositions and inverted transpositions are viewed as possible, but their relative preponderance with respect to inversions is unknown. Sankoff proposed the *breakpoint* distance (the number of pairwise gene adjacencies present in one genome but absent in the other), a measure of distance between genomes that is independent of any particular mechanism of rearrangement; the *breakpoint phylogeny*, introduced by Blanchette *et al.* (1997), is the most parsimonious tree with respect to breakpoint distances.

The two software packages for reconstructing the breakpoint phylogeny, the original BPAanalysis of Sankoff and Blanchette (1998) and the more recent and much faster GRAPPA of Moret *et al.* (2001), both use as their basic optimization tool an algorithm for computing the breakpoint median of three genomes, although GRAPPA also supports inversion medians and inversion distance—see Moret *et al.* (2002b)—, the latter through the linear-time algorithm of Bader *et al.* (2001) and the former through the algorithms of Caprara (2001) and Siepel and Moret

*To whom correspondence should be addressed.

(2001). Extensive testing has shown that the trees returned by GRAPPA are superior to those returned by other methods used in phylogenetic reconstruction based on gene orders, such as distance-based methods and parsimony based on encodings—see Moret *et al.* (2002c) and Wang *et al.* (2002) for reviews of these other methods. The closely related software of Bourque and Pevzner (2002) is the only method that approaches its accuracy. (A recent Bayesian approach due to Larget *et al.* (2002) and an effort based on local perturbation of a minimum spanning tree from Wu and Gu (2003), while both showing promise, were tested on just one or two datasets and thus cannot as yet be properly evaluated.)

Although GRAPPA runs over one billion times faster than the initial BPAAnalysis implementation, it remains an exponential-time algorithm. On a modern workstation, it typically takes one hour to finish a 13-taxon analysis, but nearly one month to finish a 15-taxon one. Bayesian methods rarely scale well even for sequence-based data: it may take months to run an analysis of a 1000-taxon dataset through Markov Chain Monte Carlo (MCMC) methods. Distance-based methods run quickly even on large datasets, but their accuracy decreases rapidly with increasing number of taxa, as shown by Moret *et al.* (2002a) and Nakhleh *et al.* (2002). Thus our best hope for accurate reconstruction is to design a way to scale the current GRAPPA software suite so as to tackle much larger problems; any such approach must reduce the size of the problem(s) that GRAPPA will be required to solve.

APPROACHES TO SCALING

A standard approach to scaling is to compute the smallest possible nontrivial trees: *quartet trees*, defined on just four taxa. Quartet methods rely on finding the optimal 4-leaf tree for each quartet and using this information to build the overall tree. Several theoretical methods (quartet cleaning and others) as well as one practical method (quartet puzzling) have been proposed to use quartet trees—see St. John *et al.* (2001) for a recent review and experimental comparison of these methods. In the case of gene orders, however, computing the best tree for a quartet is itself NP-hard (it includes finding the median of three genomes as a special case). Moreover, having to consider all quartets means that, on large datasets, many quartets will have very large pairwise distances, in which case determining the best quartet tree becomes chancy—each of the three possible trees will have poor scores. Finally, quartet-based methods, while running in polynomial time, tend to be slow: all of them must take $\Omega(n^4)$ time by definition. In previous work—see Tang *et al.* (2002)—, we conducted preliminary experiments with both quartet optimization methods and tree-building from quartets, the latter with the quartet-puzzling method of Strimmer and

von Haeseler (1996) (the best quartet-based method in the experiments of St. John *et al.* (2001) and the one used by biologists); we found that quartet-puzzling, even with optimal quartet trees, lagged far behind our new DCM-GRAPPA (discussed below) in both speed and accuracy, although it did construct more accurate trees than pure neighbor-joining.

A more sophisticated approach to the decomposition problem should use a type of divide-and-conquer approach, in which the set of taxa is decomposed into a collection of subsets, each of which optimizes some criterion designed to make reconstruction on the subset as accurate and efficient as possible. The best such approach to date is the family of disk-covering methods (DCM), introduced by Warnow and her group—see Huson *et al.* (1999a), Huson *et al.* (1999b), and Huson *et al.* (1999c)—and since shown to produce better results on sequence-based data than any other distance- or parsimony-based method through experimental studies of Moret *et al.* (2002a), and Nakhleh *et al.* (2001a,b). We combined the DCM2 approach of Huson *et al.* (1999c) with GRAPPA, limiting the size of the subsets (disks in the DCM terminology) to at most 13 taxa through a combination of threshold choices and recursive calls to the DCM decomposition itself, yielding a DCM-GRAPPA software for tree reconstruction from gene-order data. We then tested the performance of DCM-GRAPPA through extensive simulations.

OUR EXPERIMENTAL APPROACH

We ran simulation studies of DCM-GRAPPA, using neighbor-joining (NJ)—see Saitou and Nei (1987)—and DCM-NJ—see Huson *et al.* (1999c)—as controls. We generated both uniformly distributed trees and random birth-death trees, the latter with the program *r8s* of Sanderson (2002). We generated trees with 20, 40, 80, 160, 320, 640, and 1280 taxa (the last to test scalability); on each tree, we evolved signed permutations of 50, 100, and 200 genes (a range that covers organellar genomes), using evolutionary rates (r , the expected number of evolutionary events along a tree edge) of 2, 4, and 8.

For each combination of parameter settings, we generated 10 datasets and examined the mean and variance of the outcomes. All our experiments were run on Athlon 1900XP machines with 2GB of main memory running Linux.

Given an inferred tree (reconstructed phylogeny), we can assess the topological accuracy by computing the *Robinson-Foulds* (RF) distance due to Robinson and Foulds (1981) with respect to the true tree. (Note that the true tree may not be the model tree itself, because the evolutionary process may cause no changes on some edges of the model tree—the true tree is defined to be the result

of contracting such edges in the model tree.) For every tree there is a natural association between every edge and the bipartition on the leaf set induced by deleting the edge from the tree. An edge is said to be *missing* in a tree if there is no edge defining the same bipartition in the tree. If an edge in the true tree is missing in the inferred tree, this edge is then called a *false negative* (FN). Similarly, a *false positive* edge (FP) is an edge of the inferred tree that is missing in the true tree. The RF distance is the total number of false negative and positive negative edges. (These measures can also be normalized by dividing them by the number of internal edges in the true tree.)

Overall, we found that the reconstructions produced by DCM-GRAPPA demonstrated excellent topological accuracy (within a few percent of optimal) throughout the range of parameters tested.

BACKGROUND

We briefly review the DCM approaches, focusing on the DCM2 method that we used, and the basic ideas behind GRAPPA.

The disk-covering methods

The disk-covering methods are a class of phylogenetic reconstruction ‘meta-methods’ that operate in conjunction with a given ‘base method,’ such as maximum parsimony or maximum likelihood. These methods operate by dividing a set of taxa into overlapping subsets (the ‘disks’), constructing trees on the subsets using the base method, and then merging the subtrees into a supertree.

Warnow and her group devised two types of DCM. The first method, DCM1, produces many decompositions for the dataset; for each such decomposition, it computes a (possibly different) supertree; finally, it chooses one of these supertrees according to some criterion such as maximum parsimony or maximum likelihood. DCM1 was designed to be used with fast base methods, such as neighbor-joining, because it involves up to $O(n^3)$ phylogenetic reconstructions of subsets of taxa. In contrast, DCM2 produces fewer decompositions (potentially only one, although not in the way we used it) and thus can use computationally expensive methods such as maximum parsimony and maximum likelihood. Since our base method is the very expensive GRAPPA, we use DCM2.

Both methods operate by creating a graph from the distance matrix: each taxon becomes a vertex and an edge is placed between two taxa whenever their pairwise distance falls below a given threshold. (DCM2 typically uses the smallest threshold that results in a connected graph.) Edges are then added (greedily, since a minimum addition is NP-hard) to the graph to make it chordal (i.e. the graph does not contain simple cycles with more than 3 vertices). A chordal graph has a linear number of maximal

cliques and these cliques can be found in polynomial time; moreover, minimal vertex separators in chordal graphs are maximal cliques.

DCM2 uses a vertex separator technique for its decomposition: if $G = (V, E)$ is the chordal graph it has obtained, it computes (in quadratic time) a separator $X \subseteq V$ such that X is a maximal clique and $G' = (V - X, E')$ has components A_1, A_2, \dots, A_r where $\max_i |X \cup A_i|$ is minimized. The overlapping subproblems are then $X \cup A_i$ for $i = 1, 2, \dots, r$. These subproblems overlap in a single ‘spine,’ the separator X , a property exploited in the supertree merging phase, which uses a strict consensus merger specialized for DCM2 subtrees.

GRAPPA

GRAPPA is our re-implementation and elaboration on the original BPAAnalysis of Sankoff and Blanchette (1998). In order to identify the best reconstructed tree, the program examines every possible tree topology on the given taxa, scoring each (using a sum of tree edge lengths) and retaining the tree(s) of lowest score. Scoring each tree is itself an NP-hard problem, since it requires reconstructing internal genomes. In the code, it is carried out heuristically through local iterative improvement: initial internal genomes are assigned in some way, then the tree is repeatedly traversed, replacing each internal genome by the median of its three neighbors if such a replacement reduces the sum of tree edge lengths, and continuing until no change takes place. Finally, computing the median is itself NP-hard, but fast solutions have been provided by Moret *et al.* (2001) for breakpoint medians and by Caprara (2001) and Siepel and Moret (2001) for inversion medians—although it should be noted that all of these methods will display exponential behavior for large pairwise distances. The study of Moret *et al.* (2002b) showed unequivocally that inversion medians are preferable to breakpoint medians (even though exact breakpoint medians can be found faster than exact inversion medians), so inversion medians are used throughout this study.

OUR NEW METHOD: DCM-GRAPPA

In combining the DCM approach with GRAPPA, we have to face two issues. DCM2 uses for its threshold the smallest value that will produce a connected graph, but the sizes of the resulting disks are unpredictable (although larger thresholds tend to produce larger disks). Since we cannot realistically run GRAPPA on more than 15 taxa, we must either limit the choice of thresholds to those that produce sufficiently small disks (but may fail to produce connected graphs) or use a recursive decomposition of the larger disks. However, the proofs of convergence and guarantees offered for DCM2 hold only when the graph produced is connected and only for a one-level decomposi-

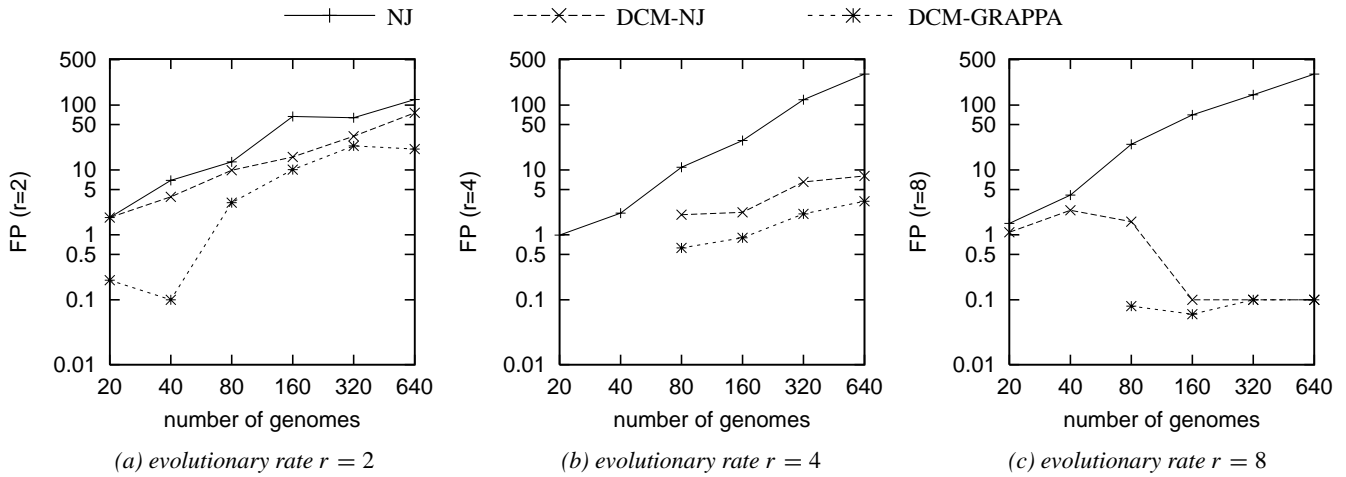


Fig. 2. Average numbers of false positives for the three algorithms as a function of the number of taxa, for 100 genes and three evolutionary rates. (Missing values equal 0.)

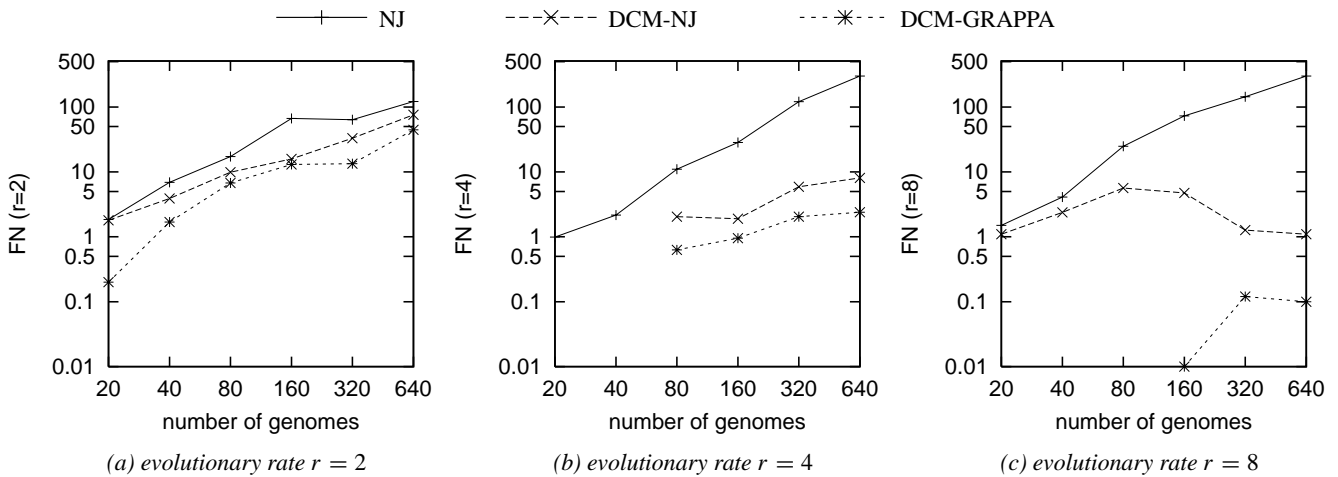


Fig. 3. Average numbers of false negatives for the three algorithms as a function of the number of taxa, for 100 genes and three evolutionary rates. (Missing values equal 0.)

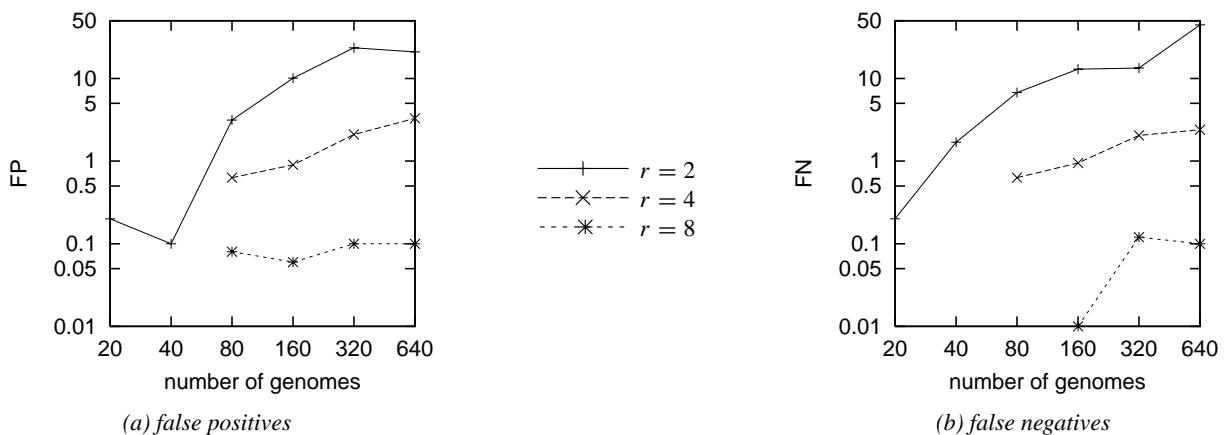


Fig. 4. Average numbers of false positives and false negatives for DCM-GRAPPA as a function of the number of taxa, for 100 genes and three evolutionary rates. (Missing values equal 0.)

problem can be remedied, at least in part, by better taxon sampling.

CONCLUSIONS

We have shown that the time-consuming, but accurate approach to phylogeny reconstruction from gene-order data first proposed by Sankoff, then refined by our group, can be placed within a divide-and-conquer framework to scale it up to larger problems. Our DCM-GRAPPA scales gracefully to a thousand genomes, returning remarkably accurate results in our simulations within a few hours to a day of computation. As larger datasets of gene-order data are produced by the many existing projects dealing with organellar evolution, we will have an accurate tool available for their phylogenetic analysis.

The principal remaining challenge is to handle unequal gene contents; using our GRAPPA framework, we have made significant strides in this direction in terms of handling gene duplication events—for which see Marron *et al.* (2003) and Tang and Moret (2003). The other major challenge is to extend our results from genomes of organellar size (no more than a few hundred genes) to the much larger nuclear genomes (thousands of genes) by devising fast new algorithms for computing inversion medians.

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation under grants ACI 00-81404, DEB 01-20709, EIA 01-13095, and EIA 01-21377.

REFERENCES

- Aldous, D. (2001) Stochastic models and descriptive statistics for phylogenetic trees, from Yule to Today. *Stat. Sci.*, **16**, 23–34.
- Bader, D., Moret, B. and Yan, M. (2001) A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, **8**, 483–491.
- Blanchette, M., Bourque, G. and Sankoff, D. (1997) Breakpoint phylogenies. In Miyano, S. and Takagi, T. (eds), *Genome Informatics*. Univ. Academy Press, pp. 25–34.
- Bourque, G. and Pevzner, P. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**, 26–36.
- Caprara, A. (1999) Formulations and hardness of multiple sorting by reversals. *Proc. 3rd Intl Conf. Comput. Mol. Biol. RECOMB99*. ACM Press, pp. 84–93.
- Caprara, A. (2001) On the practical solution of the reversal median problem. In Gascuel, O. and Moret, B. (eds), *Proc. 1st Intl Workshop on Algorithms in Bioinformatics (WABI'01)*, Lecture Notes in Computer Science 2149, Springer, pp. 238–251.
- Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T. and Wyman, S. (2000) An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In Sankoff, D. and Nadeau, J. (eds), *Comparative Genomics*. Kluwer Acad. Publ., pp. 99–122.
- Downie, S. and Palmer, J. (1992) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In Soltis, P., Soltis, D. and Doyle, J. (eds), *Plant Molecular Systematics*. Chapman and Hall, pp. 14–35.
- Huson, D., Nettles, S., Rice, K., Warnow, T. and Yooseph, S. (1999a) The hybrid tree reconstruction method. *ACM J. Exp. Algorithms*, **4**, <http://www.jea.acm.org/1999/HusonHybrid/>.
- Huson, D., Nettles, S. and Warnow, T. (1999b) Disk-covering, a fast converging method for phylogenetic tree reconstruction. *J. Comput. Biol.*, **6**, 369–386.
- Huson, D., Vawter, L. and Warnow, T. (1999c) Solving large scale phylogenetic problems using DCM-2. In *Proc. 7th Intl Conf. on Intelligent Systems for Molecular Biology (ISMB99)*, pp. 118–129.
- Larget, B., Kadane, J. and Simon, D. (2002) A Markov chain Monte Carlo approach to reconstructing ancestral genome rearrangements. *Technical Report*. Carnegie Mellon University, Pittsburgh, PA, available at www.stat.cmu.edu/tr/tr765/.
- Marron, M., Swenson, K.M. and Moret, B.M.E. (2003) Genomic distances under deletions and insertions. *Technical Report CS-2003-8*. Univ. New Mexico, Albuquerque, New Mexico.
- Moore, A.O. and Heard, S.B. (1997) Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.*, **72**, 31–54.
- Moret, B., Roshan, U. and Warnow, T. (2002a) Sequence length requirements for phylogenetic methods. In Guigo, R. and Gusfield, D. (eds), *Proc. 2nd Intl Workshop on Algorithms in Bioinformatics (WABI'02)*, Lecture Notes in Computer Science 2452, Springer, pp. 343–356.
- Moret, B., Siepel, A., Tang, J. and Liu, T. (2002b) Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In Guigo, R. and Gusfield, D. (eds), *Proc. 2nd Intl Workshop on Algorithms in Bioinformatics (WABI'02)*, Lecture Notes in Computer Science 2452, Springer, pp. 521–536.
- Moret, B., Tang, J., Wang, L.-S. and Warnow, T. (2002c) Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, **65**, 508–525.
- Moret, B., Wyman, S., Bader, D., Warnow, T. and Yan, M. (2001) A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing (PSB01)*. World Scientific Publ., pp. 583–594.
- Nakhleh, L., Roshan, U., St. John, K., Sun, J. and Warnow, T. (2001a) Designing fast converging phylogenetic methods. *Bioinformatics*, **17**(Suppl. 1), S190–S198.
- Nakhleh, L., Roshan, U., St. John, K., Sun, J. and Warnow, T. (2001b) The performance of phylogenetic methods on trees of bounded diameter. In Gascuel, O. and Moret, B. (eds), *Proc. 1st Intl Workshop on Algorithms in Bioinformatics (WABI'01)*, Lecture Notes in Computer Science 2149, Springer, pp. 214–226.
- Nakhleh, L., Moret, B., Roshan, U., John, K.S. and Warnow, T. (2002) The accuracy of fast phylogenetic methods for large datasets. In *Proc. 7th Pacific Symp. on Biocomputing (PSB02)*. World Scientific Publ., pp. 211–222.
- Olmstead, R. and Palmer, J. (1994) Chloroplast DNA systematics: a review of methods and data analysis. *Amer. J. Bot.*, **81**, 1205–1224.
- Palmer, J. (1992) Chloroplast and mitochondrial genome evolution in land plants. In Herrmann, R. (ed.), *Cell Organelles*. Springer, pp. 99–133.
- Pe'er, I. and Shamir, R. (1998) The median problems for breakpoints

- are NP-complete. *Elec. Colloq. Comput. Complexity*, **71**.
- Raubeson,L. and Jansen,R. (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. *Science*, **255**, 1697–1699.
- Raubeson,L., Moret,B., Tang,J., Wyman,S. and Warnow,T. (2001) Inferring phylogenetic relationships using whole genome data: A case study of photosynthetic organelles and chloroplast genomes. *Technical Report CS-2001-19*. University of New Mexico, Albuquerque, NM 18731.
- Rice,K., Donoghue,M. and Olmstead,R. (1997) Analyzing large datasets: rbcL500 revisited. *Syst. Biol.*, **46**, 554–563.
- Robinson,D. and Foulds,L. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
- Saitou,N. and Nei,M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
- Sanderson,M. (2002) r8s software package. Available from <http://ginger.ucdavis.edu/r8s/>.
- Sankoff,D. and Blanchette,M. (1998) Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, **5**, 555–570.
- Sankoff,D. and Nadeau,J. (eds) (2000) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*. Dordrecht. Kluwer Academic, Netherlands.
- Siepel,A. and Moret,B. (2001) Finding an optimal inversion median: experimental results. In Gascuel,O. and Moret,B. (eds), *Proc. 1st Intl Workshop on Algorithms in Bioinformatics (WABI'01)*, Lecture Notes in Computer Science 2149, Springer, pp. 189–203.
- St. John,K., Warnow,T., Moret,B. and Vawter,L. (2001) Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. In *Proc. 12th Ann. ACM/SIAM Symp. Discrete Algs. (SODA01)*. SIAM Press, pp. 196–205.
- Strimmer,K. and von Haeseler,A. (1996) Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, **13**, 964–969.
- Tang,J., Liu,T. and Moret,B. (2002) Scaling up accurate phylogenetic reconstruction from gene-order data. *Technical Report CS-2002-31*. University of New Mexico, Albuquerque, NM 87131.
- Tang,J. and Moret,B.M.E. (2003) Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. *Technical Report CS-2003-9*. University of New Mexico, Albuquerque, NM 87131.
- Wang,L., Jansen,R., Moret,B., Raubeson,L. and Warnow,T. (2002) Fast phylogenetic methods for genome rearrangement evolution: An empirical study. In *Proc. 7th Pacific Symp. on Biocomputing (PSB02)*. World Scientific Publ., pp. 524–535.
- Wu,S. and Gu,X. (2003) Algorithms for multiple genome rearrangement by signed reversals. In *Proc. 8th Pacific Symp. on Biocomputing (PSB03)*. World Scientific Publ., pp. 363–374.