Advances in Phylogeny Reconstruction from Gene Order and Content Data

Bernard M.E. Moret

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

Tandy Warnow

Department of Computer Sciences, University of Texas, Austin TX 78712

Abstract

Genomes can be viewed in terms of their gene content and the order in which the genes appear along each chromosome. Evolutionary events that affect the gene order or content are "rare genomic events" (rarer than events that affect the composition of the nucleotide sequences) and have been advocated by systematists for inferring deep evolutionary histories. This chapter surveys recent developments in the reconstruction of phylogenies from gene order and content, focusing on their performance under various stochastic models of evolution. Because such methods are currently quite restricted in the type of data they can analyze, we also present current research aimed at handling the full range of whole-genome data.

Key words:

1 Introduction: Molecular Sequence Phylogenetics

A phylogeny represents the evolutionary history of a collection of organisms, usually in the form of a tree. Sequence data are by far the most common form of molecular data used in phylogenetic analyses. We begin by briefly reviewing techniques for estimating phylogenies from molecular sequences, with emphasis on the computational and statistical issues involved.

Email addresses: moret@cs.unm.edu (Bernard M.E. Moret),
tandy@cs.utexas.edu (Tandy Warnow).
URLs: www.cs.unm.edu/∼moret/ (Bernard M.E. Moret),
www.cs.utexas.edu/users/tandy/ (Tandy Warnow).

1.1 Model trees and stochastic models of evolution

Most algorithms for phylogenetic reconstruction attempt to reverse a *model of evolution*. Such a model embodies certain knowledge and assumptions about the process of evolution, such as characteristics of speciation and details about evolutionary changes that affect the content of molecular sequences. Models of evolution vary in their complexity; in particular, they require different numbers of parameters. For instance, the Jukes-Cantor model, which assumes that all sites evolve identically and independently and that all substitutions are equally likely, requires just one parameter per edge of the tree, viz., the expected number of changes of a random site on that edge. Overall, then, a rooted Jukes-Cantor tree with n leaves requires $2n-2$ parameters. Under more complex models of evolution, the process operating on a single edge can require up to 12 parameters (for the General Markov model), although these models still requires $\Theta(n)$ parameters overall. If edge "lengths" are drawn from a distribution, however, the complexity can be reduced, since the evolutionary process operating on the model tree can then be described just by the parameters of the distribution.

These parameters describe how a single site evolves down the tree and so require additional assumptions in order to describe how different sites evolve. Usually the sites are assumed to evolve independently; sometimes they are also assumed to evolve identically. Moreover, the different sites are assumed either to evolve under the same process or to have rates of evolution that vary depending upon the site. In the latter case (in which each site has its own rate), an additional k parameters are needed, where k is the number of sites. However, if the rates are presumed to be drawn from a distribution (typically, the gamma distribution), then a single additional parameter suffices to describe the evolutionary process operating on the tree; furthermore, in this case, the sites still evolve under the *i.i.d.* assumption.

Tree generation models typically have parameters regulating speciation rates, but also inheritance characteristics, etc. For more on stochastic models of (sequence) evolution, see Felsenstein (1981), Kim and Warnow (1999), Li (1997), and Swofford et al. (1996); for an interesting discussion of models of tree generation, see Heard (1996) and Mooers and Heard (1997).

By studying the performance of methods under explicit stochastic models of evolution, it becomes possible to assess the relative strengths of different methods, as well as to understand how methods can fail. Such studies can be theoretical, for instance proving *statistical consistency*: given long enough sequences, the method will return the true tree with arbitrarily high probability. Others can use simulations to study the performance of the methods under conditions closely approximating practice. In a simulation, sequences are evolved down different model trees and then given to different methods for reconstruction; the reconstructions can then be compared against the model trees that generated the data. Such studies provide important quantifications of the relative merits of phylogenetic reconstruction methods.

1.2 Phylogeny reconstruction from molecular sequences

Three main types of methods are used to reconstruct phylogenies from molecular sequences: *distance-based* methods, *maximum parsimony* heuristics, and *maximum likelihood* heuristics.

1.2.1 Distance-based methods

Of the three types of methods, only distance-based methods include algorithms that run in polynomial-time. Distance-based methods operate in two phases:

(1) Pairwise distances between every pair of taxa are estimated.
(2) An algorithm is applied to the matrix of pairwise distances to compute an edge-weighted tree T.

The statistical consistency (if any) of such two-phase procedures rests on two assumptions: first, that a statistically consistent distance estimator is used in the first phase and, second, that an appropriate distance-based algorithm is used in the second phase. The requirements that the first phase be statistically consistent means that the distance estimator should return a value that approaches the expected number of times a random site changes on the path between the two taxa. Thus, the estimation of pairwise distances must be done with respect to some assumed stochastic model of evolution. As an example, in the Jukes-Cantor model of evolution, the estimated distance between sequences s_i and s_j is given by the formula

$$d_{ij} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{H_{ij}}{k}\right),$$

where k is the sequence length and H_{ij} denotes the Hamming distance (the number of positions in which s_i and s_j differ, which is the edit distance under mutation operations).

Algorithms that attempt to reconstruct trees from distance matrices are guaranteed to produce accurate reconstructions of the trees only when the distance matrix entries approach very closely the actual number of changes between the pair of sequences. (In the context of estimating model trees, this requirement means that the estimated distances need to be extremely close to the model distances, defined to be the expected number of times a random site changes on a leaf-to-leaf path. See Atteson (1999) and Kim and Warnow (1999) for more on this issue.) Naïvely defined distances, such as the Hamming distance, typically underestimate the number of changes that took place in the evolutionary history; thus the first step of a distance-based method is to *correct* the naïvely defined distance into one that accurately accounts for the expected number of unseen back-and-forth changes in a site. Such corrections are not without problems: as the measured distance grows larger, the variance in the estimator increases, causing increasing errors in reconstruction.

The most commonly used, and simplest, distance-based method is the *neighbor-joining (NJ)* algorithm of Saitou and Nei (1987); improved versions of this basic method include BioNJ (Gascuel, 1997) and a version known as Weighbor, which requires an estimate of the variance of the distance estimator (Bruno et al., 2000). NJ is known to be statistically consistent under most models of evolution.

1.2.2 Maximum Parsimony

Parsimony-based methods seek the tree, along with sequences labelling its internal nodes, that together minimize the total number of evolutionary changes (viewed as distances summed along all edges of the tree). Put formally, the problem is as follows: Given a set S of sequences in a multiple alignment, each of length k, find a tree T and a set of additional sequences S_0, all also of length k, so that, with the leaves of T are labelled by S and its internal nodes by S_0, the value $\sum_{e \in E(T)} Hamming(e)$ is minimized, where $Hamming(e)$ denotes the Hamming distance between the sequences labelling the endpoints of e. (Weighted or distance-corrected versions can also be defined.)

The maximum parsimony problem (MP) is thus an optimization problem—and a hard one: finding the best tree is provably NP-hard (Day, 1983). This property effectively rules out exact solutions for all but the smallest instances; indeed, in practice, exact solvers run within reasonable time on at most 30 taxa. Thus heuristics are the normal approach to the problem; most are based on iterative improvement techniques and appear to return very good solutions for up to a few hundred taxa. Many software packages implement such heuristics, among them MEGA (Kumar et al., 2001), PAUP* (Swofford, 2001), Phylip (Felsenstein, 1993), and TNT (Goloboff, 1999).

1.2.3 Maximum Likelihood

Like maximum parsimony, maximum likelihood (ML) is an optimization problem. ML seeks the tree and associated model parameter values that maximizes the probability of producing the given set of sequences. ML thus depends explicitly on the assumed model of evolution. For example, the ML problem under the Jukes-Cantor model needs to estimate one parameter (the substitution probability) for each edge of the tree, while under the General Markov model 12 parameters must be estimated on each edge. ML is much more computationally expensive than MP: even the problem of point estimation (scoring a tree), i.e., finding optimal edge parameters, for the simplest (Jukes-Cantor) model of evolution on a fixed tree is of unknown computational complexity, and computationally expensive without being provably accurate in practice (see Steel (1994) for a discussion), whereas it is easily accomplished in linear time for MP using Fitch's algorithm (Fitch, 1977). Provably correct solutions to ML are currently limited to some special cases of four-leaf model trees, exhaustive searches through tree space that use heuristics for scoring trees are limited to about ten taxa, and heuristic searches through tree space using similar heuristics for scoring trees are typically limited to fewer than 100 taxa. Various software packages provide heuristics for ML, including PAUP* (Swofford, 2001), Phylip (Felsenstein, 1993), FastDNAml (Olsen et al., 1994), PhyML (Guindon and Gascuel, 2003), and TrExML (Wolf et al., 2000).

1.3 Performance issues

Methods can be compared in terms of their performance guarantees, in terms of their resource requirements, and in terms of the quality of the trees they produce. Very few methods offer any performance guarantees, except in purely theoretical terms. For instance, while ML is known to be statistically consistent under most models, the same cannot be said of its heuristic implementation; and even neighbor-joining, which is statistically consistent and is implemented exactly, may return very poor trees—the guarantee of statistical consistency only implies good performance in the limit, as sequences lengths become sufficiently large. In terms of computational requirements, the comparison is easy: distance-based methods are efficient (running in polynomial-time with low coefficients); parsimony is much harder to solve (systematists are accustomed to running MP for weeks on a dataset of modest size); and maximum likelihood is much harder again than MP. These comparisons, however, all have limited value: as we saw, statistical consistency is a very weak guarantee, while a guarantee of fast running times is worthless if the returned solution is poor. Thus experimental studies are our best tool in the study of the relative performance of methods. Simulation studies, in particular, can establish the absolute accuracy of methods (whereas studies conducted with biological datasets can only assess relative performance in terms of the optimization criterion). Such studies have shown that MP methods can produce reasonably good trees under conditions where neighbor-joining can have high topological error (significantly worse than MP); this possibly surprising performance holds under many realistic model conditions—in particular, when the model tree has a high evolutionary diameter (Moret et al., 2002b; Nakhleh et al., 2001a,b, 2002; Roshan et al., 2004).

1.4 Limitations for molecular sequence phylogenetics

Although existing methods often yield good estimates of phylogenies on datasets of small to medium size, all methods based on molecular sequences suffer from similar limitations. Perhaps most seriously, deep evolutionary histories can be hard to reconstruct from molecular sequence data: the further back one goes in time, the harder the alignment of sequences becomes and the greater the impact of *homoplasy* (multiple point mutations at the same position). Under these conditions, we have established, through extensive simulation studies, that most major methods (heuristics for maximum parsimony as well as neighbor-joining) have poor topological accuracy (Moret et al., 2002b; Nakhleh et al., 2002, 2001b). The problem accrues from a combination of the small state space (only four nucleotides for DNA or RNA sequences and only 20 amino acids for protein sequences), the relatively high frequency of point mutations, and the limited amount of data. Concatenating—also called combining—gene sequences to obtain longer sequences may provide more data, but brings its own problems: different genes may follow different evolutionary paths (each potentially different from that of the organism as a whole)—a problem known as the *gene tree/species tree* problem (Ma et al., 1998; Maddison, 1997; Page and Charleston, 1997a; Pamilo and Nei, 1998)—, while reticulation events (such as hybridization, lateral gene transfer, gene conversion, etc., see Linder et al. (2004)) create convergent paths, with the result that tree-based analyses may run into contradictions and yield poor results. Current research on resolving the gene tree/species tree problem (a process known as *reconciliation* (Page, 1998; Page and Charleston, 1997a,b)) and on identifying and properly handling reticulation events has not yet produced reliably accurate and scalable methods. Thus phylogeny reconstruction based on site-evolution models will continue to suffer problems, for at least the near future, when attempting to infer deep evolutionary histories.

2 Whole-Genome Evolution

Systematists are interested in whole genomes because they have the potential to overcome two of the main problems afflicting sequence data. Since the entire genome is used, the data reflect organismal evolution, not the evolution of single genes, thereby avoiding the gene tree/species tree problem. (Naturally, however, if the phylogeny is based on organellar genomes, it need not coincide with the organismal phylogeny, nor will two phylogenies based on different organelles or plasmids necessarily agree.) Moreover, the events that affect

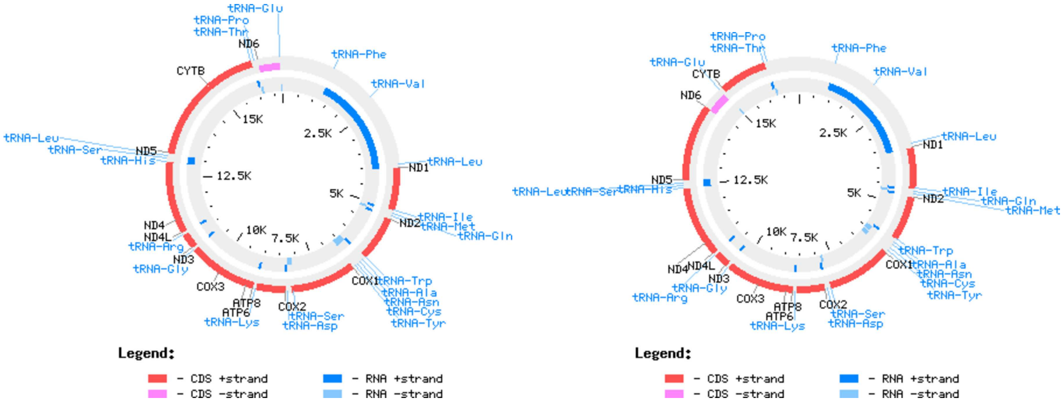


Fig. 1. The mitochondrial chromosomes of *Gallus gallus* and of *Felis catus* (from NCBI)

the whole genome by altering its gene content or rearranging its genes are so-called "rare genomic events" (Rokas and Holland, 2000): they occur rarely and come from a very large set of choices (for instance, there can be a quadratic number of distinct inversion events), so that they are unlikely to give rise to homoplasy, even in deep branches of the tree. Thus genome rearrangements, in particular, have been increasingly used in phylogenetic analysis (Boore et al., 1995; Cosner et al., 2000a,b; Downie and Palmer, 1992; Jansen and Palmer, 1987; Stein et al., 1992).

To date, the main approach to whole-genome phylogenetic analysis has used the ordering of the genes along the chromosomes as its primary data. That is, each chromosome is considered as a linear (or circular) ordering of genes, with each gene represented by an identifier that it shares with its homologs on other chromosomes (or, for that matter, on the same chromosome, in the case of gene duplications). The genome is thus simplified in the sense that point mutations are ignored and evolutionary history is inferred on the basis of the gene content and gene order within each chromosome. Typical single circular chromosomes for the mitochondrial organelles of the farm chicken (left) and the domestic cat (right) are shown in Figure 1. Animal mitochondrial genomes have 37 genes (DNA and RNA) and are quite similar—as is evident in the figure; plant mitochondrial genomes are much larger and quite variable, with up to several hundred genes; and plant choloroplast genomes have around 120 genes. In contrast, nuclear chromosomes in eukaryotes and free-living bacteria typically have many thousands of genes and, unlike the organellar genomes, often have many homologs of a gene (forming gene families).

2.1 Evolution of gene order and content

Events that change gene order (but not content) along a single chromosome include *inversions*, which are well documented (Jansen and Palmer, 1987;

Palmer, 1992), *transpositions*, which are strongly suspected in mitochondria (Boore and Brown, 1998; Boore et al., 1995), and *inverted transpositions*; these three operations are illustrated in Figure 2. In a multichromosomal genome, additional operations that do not affect gene content include *translocations*, which moves a piece of one chromosome into another chromosome (in effect, a transposition between chromosomes), and *fissions* and *fusions*, which split and merge chromosomes without affecting genes. Finally, a number of events can affect the gene content of genomes: *insertions* (of genes without existing homologs), *duplications* (of genes with existing homologs), and *deletions*. In multichromosomal organisms, colocation of genes on the same chromosome, or *synteny*, is an important evolutionary character and has been used in phylogenetic reconstruction (Nadeau and Taylor, 1984; Sankoff and Nadeau, 1996; Sankoff et al., 1997).

In this model, one whole chromosome forms a single character, whose state is affected by all of the operations just described. This one character can assume any of an enormous number of states—for a chromosome with n distinct, single-copy genes, the number of states is $2^{n-1}(n-1)!$, in sharp contrast to the 4 or 20 states possible for a sequence character. Even for a simple chloroplast genome, with a single small circular chromosome of 120 genes, the resulting number of states is very large—on the order of 10^{235}.

The use of gene-order and gene-content data in phylogenetic reconstruction is relatively recent and the subject of much current research. As mentioned earlier, such data present many advantages: (i) the identification of homologies can rest on a lot of information and thus tends to be quite accurate; (ii) because the data capture the entire genome, there is no gene tree vs. species tree problem; (iii) there is no need for multiple sequence alignment; and (iv) gene rearrangements and duplications are much rarer events than nucleotide mutations and thus enable us to trace evolution farther back in time—by two or more orders of magnitude.

However, there remain significant challenges. First is the lack of data: mapping

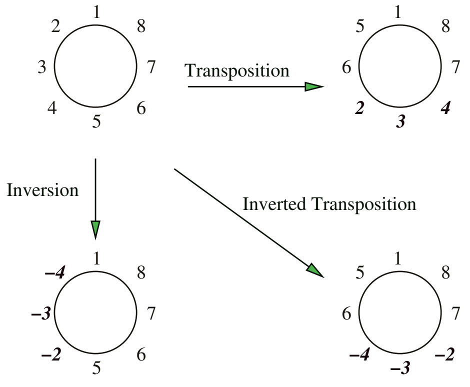


Fig. 2. The three rearrangements operating on a single chromosome

Table 1
Existing whole-genome data *ca.* 2003 (approximate values)

Type	Attributes	Numbers
Animal mitochondria	1 chromosome, 37 genes	500
Plant chloroplasts	1 chromosome, ∼120 genes	100
Bacteria	1–2 chromosomes, 500–5,000 genes	150
Eukaryotes	3–30 chromosomes, 2,000–40,000 genes	10

a full genome, while easier than sequencing it, remains far more demanding than sequencing a few genes. Table 1 gives a rough idea of the state of affairs around 2003. For obvious reasons, most of the bacteria sequenced to date are human pathogens (but nearly 500 additional bacterial genomes are expected by the end of 2005), while the few eukaryotes are the model species chosen in genome projects: human, mouse, fruit fly, round worm, mustard, yeast, etc. Although the number of sequenced eukaryotic genomes is growing quickly, coverage at this level of detail will not, for the foreseeable future, exceed a small fraction of the total number of described organisms.

This lack of data has so far prevented us from understanding very much about the relative probabilities of different events that modify gene order and content; consequently, the stochastic models proposed to date (see Section 2.2) remain fairly primitive.

Finally, the extreme (at least in comparison with sequence data) mathematical complexity of gene orders means that all reconstruction methods (even the distance-based ones) face major computational challenges even on small datasets (containing only ten or so genomes).

Table 2 summarizes the salient characteristics of sequence data and gene-order data.

Table 2
Main attributes of sequence and gene-order data

	Sequence	Gene-Order
evolution	fast	slow
data type	a few genes	whole genome
data quantity	abundant	sparse
# char. states	tiny	huge
models	good (sites) primitive (sequences)	primitive
computation	easy	hard

2.2 Stochastic models of evolution

Models of genome evolution have been largely limited to simple combinations of the main rearrangement operations: inversion, transposition, and inverted transposition. The original model was proposed by Nadeau and Taylor; it uses only inversions and assumes that all inversions are equally likely (Nadeau and Taylor, 1984). We extended this model to produce the Generalized Nadeau-Taylor (GNT) model (Wang and Warnow, 2001), which includes transpositions and inverted transpositions. Within each of the three types of events, any two events are equiprobable, but the relative probabilities of each type of event are specified by the two parameters of the model: α, the probability that a random event is a transposition and β, the probability that a random event is an inverted transposition. (The probability of an inversion is thus given by $1 - \alpha - \beta$.) The GNT model contains the Nadeau-Taylor model as a special case: just set $\alpha = \beta = 0$. Extensions of this simple model in which the probability of an event depends on the length of the segment affected by the event have been proposed (Bender et al., 2004), but no solid data exist to support one model over another; all that appears certain (see Lefebvre et al. (2003)) is that short inversions are more likely than long ones in prokaryotes. Similarly, multichromosomal rearrangements such as translocations and events that affect the gene content of chromosomes such as duplications, insertions, and deletions, have all been considered, but once again we have insufficient biological data to define any model with confidence.

3 Genomic Distances

The distance between two genomes (each represented by the order of its genes) can be defined in several ways. First, we have the *true evolutionary distance*, that is, the actual number of evolutionary events (mutations, deletions, etc.) that separate one genome from the other. This is the distance measure we would really want to have, but of course it cannot be inferred—as our earlier discussion of homoplasy made clear, we cannot infer such a distance even when we know the correct phylogeny and have correctly inferred ancestral data (at internal nodes of the tree). What we can define precisely and, in some cases, compute, is the *edit distance*, the minimum number of permitted evolutionary events that can transform one genome into the other. Since the edit distance invariably underestimates the true evolutionary distance, we can attempt to *correct* the edit distance according to an assumed model of evolution in order to produce the *estimated true evolutionary distance*—which, when derived from an assumed model of evolution, is actually a maximum likelihood estimate. Finally, we can attempt to estimate the true evolutionary distance directly through various heuristic techniques.

We begin by reviewing distance measures between two chromosomes with equal gene contents and no duplications—the simplest possible case—, then discuss current research on distance measures between multichromosomal genomes and between genomes with unequal gene content.

3.1 Distances between two chromosomes with equal gene content and no duplications

Here we consider chromosomes that have identical gene content and exactly one copy of each gene, so that each chromosome can be viewed as a (signed) permutation of the underlying set of genes. Two distance metrics have been used in this context, one based on observed differences and one based on allowable evolutionary operations. Assume our two chromosomes each have seven genes, numbered 1 through 7, and are circular. Genome G_1 is given by $(1, 2, -4, -3, 5, 6, 7)$ and genome G_2 by $(1, 2, 3, 4, 5, 6, 7)$.

- The *breakpoint distance* simply counts the number of gene adjacencies (read on either strand) present in one chromosome, but not in the other. In our example, we have two breakpoints: the adjacencies $2, 3$ and 4.5 are present in G_2, but not in G_1. (Note that the adjacency $3, 4$ is present on the forward strand on G_2 and on the reverse complement strand, as $-4, -3$, on G_1. Note also that the definition is symmetric: the adjacencies $2, -4$ and $-3, 5$ are present in G_1, but not in G_2.) The breakpoint distance is thus not based on evolutionary events, only on the end result of such events.
- The *inversion distance* is the edit distance under the single allowed event of inversion. We need only one inversion to transform G_1 into G_2: we invert the two-gene segment $-4, -3$.

Every inversion clearly creates (or, equivalently, removes) at most two breakpoints, so that the inversion distance is at least half the breakpoint distance. The number of breakpoints is also clearly at most n in a circular chromosome of n genes (and $n+1$ in a linear chromosome); less obviously, the same bound holds for the inversion distance (Meidanis et al., 2000).

Computing the breakpoint distance is trivially achievable in linear time. Computing the inversion distance, however, is a very complex problem. Indeed, it is computationally intractable (technically, it is NP-hard) for unsigned permutations (when we cannot tell on which strand each gene lies). For signed permutations, we showed that it can be computed in linear time (Bader et al., 2001), but this result is the culmination of many years of research and rests on the very elaborate and elegant theory of Hannenhalli and Pevzner (1995a,b). Obtaining an actual sequence of inversions (as opposed to just the number of required inversions) is computationally more demanding: the classic algorithm of Kaplan et al. (1999) takes $O(dn)$ time, where d is the distance and n the number of genes; thus, for large distances, this algorithm takes quadratic time, but was recently improved to $O(n\sqrt{n \log n})$ time (Tannier and Sagot, 2004).

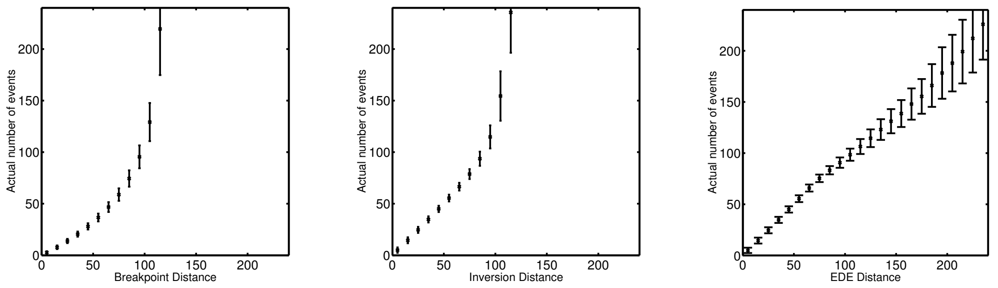
Transpositions are also of significant interest in biology. However, while some of the same theoretical framework can be used (Bafna and Pevzner, 1995), results here are disappointing: no efficient algorithm has yet been developed to compute the transposition distance. The best result to date remains an approximation algorithm that could suffer from up to a 50% error (Bafna and Pevzner, 1995; Hartman, 2003); this result was recently extended, with the same error bound, to the computation of edit distances under a combination of inversions and inverted transpositions, with equal weights assigned to each (Hartman and Sharan, 2004).

3.2 Distance corrections

None of these distances produces the true evolutionary distance; indeed, since all are bounded by n, all can produce arbitrary underestimates of the true evolutionary distance. In order to estimate the latter, we must use correction methods. Such methods are widely used in distance estimation between DNA sequences (Huson et al., 1999b; Swofford et al., 1996). We designed two techniques for "correcting" gene-order distances under the GNT model, one (IEBP) based on the breakpoint distance (Wang and Warnow, 2001) and the other (EDE) based on the inversion distance (Moret et al., 2001a, 2002d; Wang, 2003).

- The *IEBP* estimator takes as input the breakpoint distance and the values of the two GNT parameters, α and β, and returns a maximum likelihood estimate of the number of inversions, transpositions, and inverted transpositions under the specified relative probabilities of the different events. The method is analytical and mathematically exact and can be implemented to run in cubic time.
- The *EDE* estimator ("EDE" stands for "empirically derived estimator") takes as input the inversion distance and produce an approximation of the maximum likelihood estimate of the number of inversions, under a model in which all inversions are equally likely. The formula used was derived with numerical techniques from a large series of simulations and can be computed in quadratic time.

Figure 3 (Moret et al., 2002d) illustrates the EDE correction and compares it with the breakpoint and inversion distances, under an inversion-only scenario. The EDE estimator offers no theoretical guarantee and only takes inversions into account, but is easier to compute than the IEBP estimator, which also



breakpoint distance *inversion distance* *EDE*

Fig. 3. Edit distances vs. true evolutionary distances and the EDE correction

requires an accurate guess of the values of the two GNT parameters. Both estimators suffer from the simple fact that the variance of the true distance (as a function of the breakpoint or inversion distance) grows rapidly as the distance grows—as is clearly visible in Figure 3. The most important aspect of EDE's performance is that trees reconstructed by distance-based methods using EDE are generally more accurate than those reconstructed by the same methods using any of the other distances (including IEBP), even when the evolutionary model involves many (or even only) transpositions (Moret et al., 2001a, 2002d).

3.3 Distances between two chromosomes with unequal gene content

Unequal gene content implies the existence of evolutionary events that affect gene content: insertions (including duplications) and deletions. If we first assume that each gene exists in at most one copy (no duplication), then the problem is to handle insertions of new genes and deletions of existing ones. (Note that the same event can be viewed as a deletion or as a non-duplicating insertion, depending on the direction of time flow.) In a seminal paper, El-Mabrouk (2000) showed how to extend the theory of Hannenhalli and Pevzner for a single chromosome without duplicated genes to handle both inversions and deletions exactly; we showed that the corresponding distance measure can also be computed in linear time (Liu et al., 2003). In the same paper, El-Mabrouk also showed how to approximate the distance between two such genomes in the presence of both deletions and nonduplicating insertions along the same time flow.

Duplications are considerably harder from a computational standpoint, as they introduce a matching problem: which homolog in one genome corresponds to which in the other genome? Sankoff (1999) proposed to sidestep the entire issue by reducing the problem to one with no duplications using the *exemplar* approach. In that approach a single copy is chosen from each gene family, and all other homologs in each family are discarded, in such a way as to minimize the breakpoint or inversion distance between the two genomes. Unfortunately,

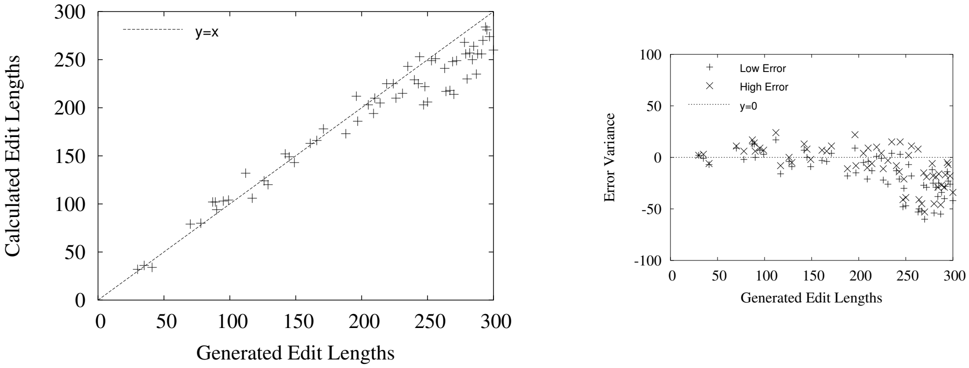


Fig. 4. Experimental results for 800 genes with expected edge length 40. Left: generated edit length vs. reconstructed length; right: the variance of computed distances per generated distance.

choosing the exemplars themselves is an NP-hard problem (Bryant, 2000); moreover, the loss of information in nuclear genomes (which can have tens or hundreds of genes in each gene family) is very large—large enough to give rise to problems in reconstruction (Tang and Moret, 2003a; Tang et al., 2004).

We recently proposed an approximation with provable guarantees (in terms of the edit distance) for the distance between two genomes under duplications, insertions, and deletions (Marron et al., 2003). We later refined the approach to estimate true evolutionary distances directly (Swenson et al., 2004) and used the new measure in a pilot study of a group of 13 γ-proteobacteria with widely differing gene contents, varying from 800 to over 5,000 genes in their single nuclear chromosome (Earnest-DeYoung et al., 2004). The latter study indicates that simple distance-based reconstructions can be very accurate even in the presence of enormous evolutionary distances: many of the edges in the final tree have several hundred evolutionary events along them. Our simulations show that the distance computation tracks the true evolutionary distance remarkably well up until saturation, which only occurs at extremely high levels of evolution (over 250 events on a genome of 800 genes, for instance). Figure 4 (Swenson et al., 2004) shows typical results from these simulations: genomes of 800 genes were generated in a simulation on a balanced tree of 16 taxa and all 120 pairwise distances computed and compared to the true evolutionary distances (the sum of the lengths of the edges in the true tree on the path connecting each pair). The figure shows the calculated edit lengths as a function of the generated edit lengths (the true evolutionary distances) on the left and an error plot on the right. The data used for this figure were generated using an expected edge length of 40, so that the pairwise distance between leaves whose common ancestor is the root has an expected value of 320 events. Events were a mix of 70% inversions, 16% deletions, 7% insertions, and 7% duplications; the inversions had a mean length of 20 and a standard deviation of 10, while the deletions, insertions, and duplications all had a mean length of 10 with a standard deviation of 5. The figure shows excellent tracking of the true evolutionary distance for up to about 250 events, after which the computation consistently returns values less than the true distance; of course, distance corrections could be devised to remedy the latter situation.

3.4 Distances between multichromosomal genomes

With multiple chromosomes, we can still make the same distinction as for single chromosomes, by first addressing genomes with equal gene content and no duplicate genes. The second of the two papers by Hannenhalli and Pevzner showed how to handle a combination of inversions and translocations for such genomes (Hannenhalli and Pevzner, 1995b), a result later improved by Tesler (2002) to handle any combination of inversions, translocations, fissions, and fusions. Using the same approach pioneered by Bader et al. (2001), one can devise linear-time algorithms to compute other distances covered by the Hannenhalli-Pevzner theory, such as the translocation distance (Li et al., 2004). However, multichromosomal genomes typically have large gene families, so that the handling of duplications is in fact crucial; moreover, few such genomes will have identical gene content—even such closely related genomes as two species of the nematode worm Caenorhabditis, *C. elegans* and *C. briggsae*, have different gene content. Thus no distance method currently exists that could be used in the reconstruction of phylogenies for multichromosomal organisms.

4 Phylogenetic Reconstruction From Whole Genomes

The same three general types of approaches to phylogenetic reconstruction that we just reviewed for sequence data can be used for whole-genome data.

4.1 Distance-based methods

We have run extensive simulations under a variety of GNT model conditions (Moret et al., 2001a, 2002d,e; Wang et al., 2002), using breakpoint distance, inversion distance, and both EDE and IEBP corrections, all using the standard neighbor-joining (Saitou and Nei, 1987) and Weighbor (Bruno et al., 2000) distance-based reconstruction algorithms, as well as within our own "DCM-boosted" extensions of the same (Huson et al., 1999a). Figure 5 (Moret et al., 2002d) shows typical results from these simulation studies: NJ is run with four distances (BP stands for breakpoint and INV for inversion) on datasets of various sizes (here, 10, 20, 40, 80, and 160 taxa), each taxon being

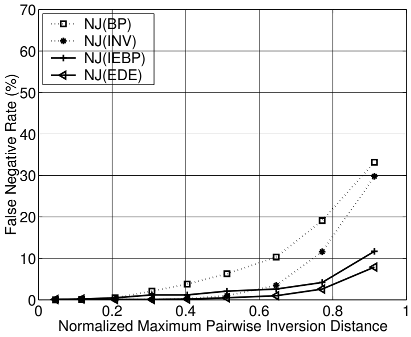


Fig. 5. False negative rates for NJ run with four genomic distance measures. The datasets have 10, 20, 40, 80, and 160 taxa; each taxon is a signed permutation of the same set of 120 genes, generated from the identity permutation at the root through an equal-weight mix of inversions, transpositions, and inverted transpositions.

given by a signed permutation of the same set of 120 genes, then the reconstructed trees are compared with the model trees. The figure shows the false negative rate (the percentage of edges present in the model tree, but missing in the reconstructed tree) of the trees reconstructed with each distance measure, as a function of the diameter of the dataset. At low rates of evolution, the results are much the same for all four distances measures, but the corrected measures perform much better than the uncorrected ones when the diameter is large.

Our simulations have established the following:

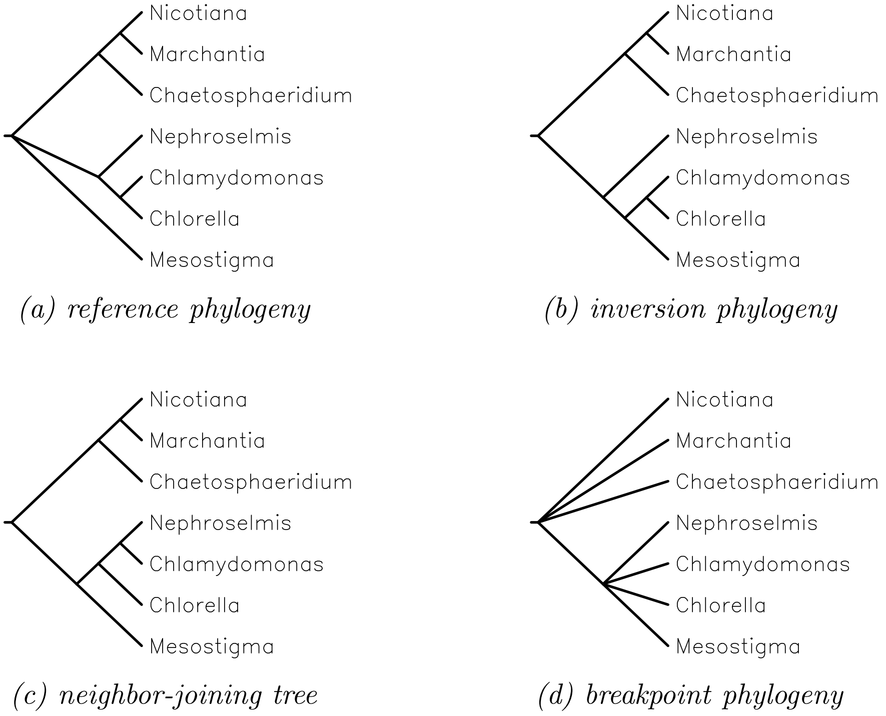
- IEBP distances are generally the most accurate, even when given incorrect estimates of the relative probabilities of the three rearrangement events. Although EDE is based on an inversion-only scenario, it produces highly accurate estimates of the actual number of events even under model conditions in which inversions play a minor role. Both EDE and IEBP are thus very robust to model violations.
- Phylogenies estimated using EDE are more accurate than phylogenies estimated using any other distance estimator (including IEBP), under all conditions, but especially when the model tree has a high evolutionary diameter. Phylogenies based upon either EDE or IEBP are better than phylogenies based upon inversion distances, which in turn are far better than phylogenies based upon breakpoint distances.

4.2 Parsimony-based methods

Given a way of defining a distance between two genomes, we can define the “length” of a tree in which every node is labelled by a genome to be the sum of the lengths of the edges. This measure of length is thus similar to that used in sequence-based phylogenetic analysis, enabling us to define parsimony problems for gene-order data. Using the breakpoint distance, for instance, we would seek a tree and a collection of new genomes (for internal nodes) such that this tree, leaf-labelled by the given set of genomes and with internal nodes labelled by the new genomes, minimizes the breakpoint length of the tree—this problem is called the *breakpoint phylogeny* (Sankoff and Blanchette, 1998). Similarly we can define the *inversion phylogeny* by replacing breakpoint distances with inversion distances. Both problems are NP-hard—indeed, they are harder than maximum parsimony for DNA sequences, as they remain NP-hard even for just three genomes (Caprara, 1999; Pe'er and Shamir, 1998). This last version is known as the *median problem*: given three genomes, find a fourth genome (to label the internal node connecting the three leaves) that minimizes the sum of its distances to the three given genomes. Exact solutions to the problem of finding a median of three genomes can be obtained for both the inversion and breakpoint distances (Caprara, 2001; Moret et al., 2002d; Siepel and Moret, 2001) and are implemented in our GRAPPA software suite. However, they take time exponential in the overall length and are thus applicable only to instances with modest pairwise distances. It should also be noted that all of these approaches are limited to unichromosomal genomes with equal gene content and no duplication.

Solving the parsimony problem for gene-order data requires the inference of "ancestral" genomes at the internal nodes of the candidate trees. Sankoff proposed to use the median problem in an iterative manner to refine rough initial guesses for these genomes (Sankoff and Blanchette, 1998), an approach that we implemented in GRAPPA with good results to date on small genomes of a few hundred genes (see (Moret et al., 2001b, 2002a; Tang, 2004; Tang and Moret, 2003a,b; Tang et al., 2004)) and that was also implemented, with less accurate heuristics, to handle a few larger genomes by Pevzner's group (Bourque and Pevzner, 2002). Identifying good candidate trees, however, is a much more expensive proposition; in that same paper, Sankoff proposed generating and scoring each possible tree in turn; the resulting `BPAnalysis` software is limited to breakpoint medians and to trees of 8 or fewer leaves. We reimplemented Sankoff's algorithm and optimized its components to improve its speed, gradually, by 7 orders of magnitude (Moret et al., 2001b, 2002a,d; Siepel and Moret, 2001), enabling the analysis of up to 16 taxa; more recently, we successfully coupled it with the Disk-Covering Method (the "DCM1" technique of Huson et al. (1999a)) to make it applicable to large datasets of several thousand taxa (Tang, 2004; Tang and Moret, 2003b).

Evidence to date from simulation studies as well as from the analysis of biological datasets indicates that even when the mechanism of evolution is based entirely on transpositions, solving the inversion phylogeny yields more accurate reconstructions than solving the breakpoint phylogeny—see, e.g., Moret et al. (2002c) and Tang et al. (2004). Solving the inversion phylogeny was also found to yield better results than using distance-based methods. Fig-



(a) reference phylogeny

(b) inversion phylogeny

(c) neighbor-joining tree

(d) breakpoint phylogeny

Fig. 6. Phylogenies on a 7-taxa cpDNA dataset

ure 6 (Tang et al., 2004) shows one example of such findings, on a small dataset of 7 chloroplast genomes from green plants. Note that the phylogeny produced by NJ (using inversion distances after equalizing the gene contents) has false positives (edges in the inferred tree that are not present in the reference tree), while the breakpoint phylogeny (computed on the same input) resolves only one single edge. In contrast, the inversion phylogeny matches the reference phylogeny, modulo the placement of *Mesostigma* (which remains in doubt).

A recent study (Lefebvre et al., 2003) indicates that, at least in prokaryotic genomes, short inversions are much more likely than long ones; work on developing new models of evolution that take such results into account is in progress. Applying these methods to large eukaryotic genomes may yield some surprises, however, since there is evidence that certain breakpoints in chromosomes are "hot spots" for rearrangements (Pevzner and Tesler, 2003); if an inversion is thus "anchored" at a fixed position in the chromosome, its net effect over several events is to produce one breakpoint per event, which might be better modelled with breakpoints than with inversions.

To date, our GRAPPA code remains the only parsimony-based reconstruction tool that can handle datasets with unequal gene content, albeit with only modest changes in gene content. In its current version (Tang et al., 2004), it has been used to reconstruct chloroplast phylogenies of green plants in which gene content differs by a few genes at most. The algorithm proceeds in two phases: first it computes gene contents for all internal nodes, then it proceeds (much as in the equal gene-content case) to establish an ordering for these contents based on median computations. Gene content is determined from the leaves inward, under standard biological assumptions such as independence of branches; under such assumptions, the likelihood of simultaneous identical changes on two sibling edges is vanishingly small compared to the reverse change on the third edge (Maddison, 1990; McLysaght et al., 2003).

4.3 Likelihood-based methods

Likelihood methods are based on a specific model of evolution and come in two main flavors: maximum likelihood (ML) and Bayesian methods. For a given tree T, ML methods estimate the parameter values that maximize the probability that T would produce the observed data; over all trees, they return that tree (and its associated parameter values) with the largest probability of producing the observed data. Because current statistical models for whole-genome evolution are so primitive and because any such models are bound to involve enormous complications due to the global nature of changes caused by a single event, no method yet exists to compute ML trees for gene-order data. Bayesian methods turn the tables: instead of seeking to maximize the probability of producing the data given the tree, they compute the probability of the tree given the data. In their standard implementation using Markov Chain Monte-Carlo (MCMC), Bayesian methods do not explicitly attempt to estimate parameters, but use a biased random walk through the space of trees and compute the equilibrium frequency with which each tree is visited during this walk. The relative simplicity of the MCMC approach makes it possible to apply it to gene-order data; a preliminary implementation of such a method for equal gene content has yielded some promising results (Larget et al., 2002).

5 Open Problems and Future Research

Nearly everything remains to be done! The work to date has convincingly demonstrated that gene-content and gene-order data can form the basis for highly accurate phylogenetic analyses; the demonstration is all the more impressive given the primitive state of knowledge in the area. We cannot give a detailed list of interesting open problems, as it would be far too long, but content ourselves with a short list of what, from today's perspective, seem to be the most promising or important avenues of exploration, from both computational and modelling perspectives.

- Solving the transposition distance problem.
- Handling transpositions along with inversions, preferably in a weighted framework.
- Adding length and location dependencies to the rearrangement framework.
- Formulating and providing reasonable approaches to solving the median problem in the above contexts.
- Developing a formal statistical model of evolution that includes all rearrangements discussed here, takes into account location within the chromosomes and length of affected segments, and obeys basic biological constraints (such as the need for telomeres and the presence of a single centromere).
- Designing a Bayesian approach to reconstruction within the framework just sketched.
- Combining DNA sequence data and rearrangement data—the sequence data may be used to rule out or favor certain rearrangements.
- Using rearrangement data in the context of network reconstruction (i.e., in the presence of past hybridizations, gene conversions, or lateral transfers).

The reader interested in more detail in the topics presented here and in some of the research problems just suggested should consult the survey articles of Wang and Warnow (2005) on distance corrections and distance-based methods and of Moret et al. (2005) on parsimony-based methods.

6 Acknowledgments

Research on this topic at the University of New Mexico is supported by the National Science Foundation under grants ANI 02-03584, EF 03-31654, IIS 01-13095, IIS 01-21377, and DEB 01-20709 (through a subcontract to the U. of Texas), by the NIH under grant 2R01GM056120-05A1 (through a subcontract to the U. of Arizona), and by IBM Corporation, under contract NBCH30390004 from the US Defense Advanced Research Projects Agency (the HPCS initiative). Research on this topic at the University of Texas is supported by the National Science Foundation under grants EF 03-31453, IIS 01-13654, IIS 01-21680, and DEB 01-20709, by the Program for Evolutionary Dynamics at Harvard, by the Institute for Cellular and Molecular Biology at the University of Texas at Austin, and by the David and Lucile Packard Foundation.

References

Atteson, K., 1999. The performance of the neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25 (2/3), 251–278.

Bader, D., Moret, B., Yan, M., 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. J. Comput. Biol. 8 (5), 483–491, a preliminary version appeared in WADS'01, pp. 365–376.

Bafna, V., Pevzner, P., 1995. Sorting permutations by transpositions. In: Proc. 6th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'95). SIAM Press, Philadelphia, pp. 614–623.

Bender, M., Ge, D., He, S., Hu, H., Pinter, R., Skiena, S., Swidan, F., 2004. Improved bounds on sorting with length-weighted reversals. In: Proc. 15th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'04). SIAM Press, Philadelphia, pp. 912–921.

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

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

Bourque, G., Pevzner, P., 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research 12, 26–36.

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

Bryant, D., 2000. The complexity of calculating exemplar distances. In: Sankoff, D., Nadeau, J. (Eds.), Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families. Kluwer Academic Pubs., Dordrecht, Netherlands, pp. 207–212.

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

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

Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T., Wyman, S., 2000a. An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In: Sankoff, D., Nadeau, J. (Eds.), Comparative Genomics. Kluwer Academic Publishers, pp. 99–122.

Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T., Wyman, S., 2000b. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In: Proc. 8th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'00). pp. 104–115.

Day, W., 1983. Computationally difficult parsimony problems in phylogenetic systematics. J. Theoretical Biology 103, 429–438.

Downie, S., Palmer, J., 1992. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis, P., Soltis, D., Doyle, J. (Eds.), Plant Molecular Systematics. Chapman and Hall, pp. 14–35.

Earnest-DeYoung, J., Lerat, E., Moret, B., 2004. Reversing gene erosion: reconstructing ancestral bacterial genomes from gene-content and gene-order data. In: Proc. 4th Int'l Workshop Algs. in Bioinformatics (WABI'04). Vol. 3240 of Lecture

Notes in Computer Science. Springer Verlag, pp. 1–13.

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

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

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

Fitch, W. M., 1977. On the problem of discovering the most parsimonious tree. American Naturalist 111, 223–257.

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

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

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

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

Hannenhalli, S., Pevzner, P., 1995b. Transforming mice into men (polynomial algorithm for genomic distance problems). In: Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95). IEEE Press, Piscataway, NJ, pp. 581–592.

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

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

Heard, S., 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evol. 50, 2141–2148.

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

Huson, D., Smith, K., Warnow, T., 1999b. Correcting large distances for phylogenetic reconstruction. In: Proc. 3rd Int'l Workshop Alg. Engineering (WAE'99). Vol. 1668 of Lecture Notes in Computer Science. Springer Verlag, pp. 273–286.

Jansen, R., Palmer, J., 1987. A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Nat'l Acad. Sci., USA 84, 5818–5822.

Kaplan, H., Shamir, R., Tarjan, R., 1999. Faster and simpler algorithm for sorting signed permutations by reversals. SIAM J. Computing 29 (3), 880–892.

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

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

Larget, B., Simon, D., Kadane, J., 2002. Bayesian phylogenetic inference from animal mitochondrial genome arrangements. J. Royal Stat. Soc. B 64 (4), 681–694.

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

Li, G., Qi, X., Wang, X., Zhu, B., 2004. A linear-time algorithm for computing translocation distance between signed genomes. In: Proc. 15th Ann. Symp. Combin. Pattern Matching (CPM'04). Vol. 3109 of Lecture Notes in Computer Science. Springer Verlag, pp. 323–332.

Li, W.-H., 1997. Molecular Evolution. Sinauer Assoc.

Linder, C., Moret, B., Nakhleh, L., Warnow, T., 2004. Network (reticulated) evolution: Biology, models, and algorithms. Tutorial available at `compbio.unm,edu/papers.html`.

Liu, T., Moret, B., Bader, D., 2003. An exact, linear-time algorithm for computing genomic distances under inversions and deletions. Tech. Rep. TR-CS-2003-31, Univ. of New Mexico.

Ma, B., Li, M., Zhang, L., 1998. On reconstructing species trees from gene trees in terms of duplications and losses. In: Proc. 2nd Ann. Int'l Conf. Comput. Mol. Biol. (RECOMB'98). ACM Press, New York, pp. 182–191.

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

Maddison, W., 1997. Gene trees in species trees. Syst. Biol. 46 (3), 523–536.

Marron, M., Swenson, K., Moret, B., 2003. Genomic distances under deletions and insertions. In: Proc. 9th Int'l Conf. Computing and Combinatorics (COCOON'03). Vol. 2697 of Lecture Notes in Computer Science. Springer Verlag, pp. 537–547.

McLysaght, A., Baldi, P., Gaut, B., 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc. Nat'l Acad. Sci., USA 100, 15655–15660.

Meidanis, J., Walter, M., Dias, Z., 2000. Reversal distance of signed circular chromosomes, unpublished.

Mooers, A., Heard, S., 1997. Inferring evolutionary process from phylogenetic tree shape. Quarterly Rev. Biol. 72, 31–54.

Moret, B., Bader, D., Warnow, T., 2002a. High-performance algorithm engineering for computational phylogenetics. J. Supercomputing 22, 99–111.

Moret, B., Roshan, U., Warnow, T., 2002b. Sequence length requirements for phylogenetic methods. In: Proc. 2nd Int'l Workshop Algs. in Bioinformatics (WABI'02). Vol. 2452 of Lecture Notes in Computer Science. Springer Verlag, pp. 343–356.

Moret, B., Siepel, A., Tang, J., Liu, T., 2002c. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Proc. 2nd Int'l Workshop Algs. in Bioinformatics (WABI'02). Vol. 2452 of Lecture Notes in Computer Science. Springer Verlag, pp. 521–536.

Moret, B., Tang, J., Wang, L.-S., Warnow, T., 2002d. Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci. 65 (3), 508–525.

Moret, B., Tang, J., Warnow, T., 2005. Reconstructing phylogenies from gene-content and gene-order data. In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press, pp. 321–352.

Moret, B., Wang, L.-S., Warnow, T., 2002e. New software for computational phylogenetics. IEEE Computer 35 (7), 55–64.

Moret, B., Wang, L.-S., Warnow, T., Wyman, S., 2001a. New approaches for reconstructing phylogenies from gene-order data. In: Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01). Vol. 17 of Bioinformatics. pp. S165–S173.

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

Nadeau, J., Taylor, B., 1984. Lengths of chromosome segments conserved since divergence of man and mouse. Proc. Nat'l Acad. Sci., USA 81, 814–818.

Nakhleh, L., Moret, B., Roshan, U., John, K. S., Warnow, T., 2002. The accuracy of fast phylogenetic methods for large datasets. In: Proc. 7th Pacific Symp. on Biocomputing (PSB'02). World Scientific Pub., pp. 211–222.

Nakhleh, L., Roshan, U., St. John, K., Sun, J., Warnow, T., 2001a. Designing fast converging phylogenetic methods. In: Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01). Vol. 17 of Bioinformatics. Oxford U. Press, pp. S190–S198.

Nakhleh, L., Roshan, U., St. John, K., Sun, J., Warnow, T., 2001b. The performance of phylogenetic methods on trees of bounded diameter. In: Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01). Vol. 2149 of Lecture Notes in Computer Science. Springer Verlag, pp. 214–226.

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

Page, R., 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinf. 14 (9), 819–820.

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

Page, R., Charleston, M., 1997b. Reconciled trees and incongruent gene and species trees. In: Mirkin, B., McMorris, F. R., Roberts, F. S., Rzehtsky, A. (Eds.), Mathematical Hierarchies in Biology. Vol. 37. American Math. Soc., pp. 57–70.

Palmer, J., 1992. Chloroplast and mitochondrial genome evolution in land plants. In: Herrmann, R. (Ed.), Cell Organelles. Springer Verlag, pp. 99–133.

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

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

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

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

Roshan, U., Moret, B., Williams, T., Warnow, T., 2004. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In: Proc. 3rd IEEE Computational Systems Bioinformatics Conf. CSB'04. IEEE Press, Piscataway, NJ, pp. 98–109.

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

Sankoff, D., 1999. Genome rearrangement with gene families. Bioinformatics 15 (11), 990–917.

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

Sankoff, D., Ferretti, V., Nadeau, J., 1997. Conserved segment identification. J. Comput. Biol. 4 (4), 559–565.

Sankoff, D., Nadeau, J., 1996. Conserved synteny as a measure of genomic distance. Disc. Appl. Math. 71 (1–3), 247–257.

Siepel, A., Moret, B., 2001. Finding an optimal inversion median: experimental results. In: Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01). Vol. 2149 of Lecture Notes in Computer Science. Springer Verlag, pp. 189–203.

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

Stein, D., Conant, D., Ahearn, M., Jordan, E., Kirch, S., Hasebe, M., Iwatsuki, K., Tan, M., Thomson, J., 1992. Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns. Proc. Nat'l Acad. Sci., USA 89, 1856–1860.

Swenson, K., Marron, M., Earnest-DeYoung, J., Moret, B., 2004. Approximating the true evolutionary distance between two genomes. Tech. Rep. TR-CS-2004-15, Univ. of New Mexico.

Swofford, D., 2001. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b8. Sunderland, MA.

Swofford, D., Olsen, G., Waddell, P., Hillis, D., 1996. Phylogenetic inference. In: Hillis, D. M., Mable, B. K., Moritz, C. (Eds.), Molecular Systematics. Sinauer Assoc., pp. 407–514.

Tang, J., 2004. Large-scale hylogeny reconstruction from arbitrary gene-order data. Ph.D. thesis, The University of New Mexico, Albuquerque, NM.

Tang, J., Moret, B., 2003a. Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In: Proc. 8th Int'l. Workshop on Algs. and Data Structures (WADS'03). Vol. 2748 of Lecture Notes in Computer Science. Springer Verlag, pp. 37–46.

Tang, J., Moret, B., 2003b. Scaling up accurate phylogenetic reconstruction from gene-order data. In: Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03). Vol. 19 of Bioinformatics. Oxford U. Press, pp. i305–i312.

Tang, J., Moret, B., Cui, L., dePamphilis, C., 2004. Phylogenetic reconstruction from arbitrary gene-order data. In: Proc. 4th IEEE Symp. on Bioinformatics and Bioengineering BIBE'04. IEEE Press, Piscataway, NJ, pp. 592–599.

Tannier, E., Sagot, M., 2004. Sorting by reversals in subquadratic time. In: Proc. 15th Ann. Symp. Combin. Pattern Matching (CPM'04). Vol. 3109 of Lecture Notes in Computer Science. Springer Verlag, pp. 1–13.

Tesler, G., 2002. Efficient algorithms for multichromosomal genome rearrangements. J. Comput. Syst. Sci. 65 (3), 587–609.

Wang, L.-S., 2003. Large-scale phylogenetic analysis. Ph.D. thesis, The University of Texas, Austin, TX.

Wang, L.-S., Jansen, R., Moret, B., Raubeson, L., Warnow, T., 2002. Fast phylogenetic methods for genome rearrangement evolution: An empirical study. In:

Proc. 7th Pacific Symp. on Biocomputing (PSB'02). World Scientific Pub., pp. 524–535.

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

Wang, L.-S., Warnow, T., 2005. Distance-based genome rearrangement phylogeny. In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press, pp. 353–383.

Wolf, M., Easteal, S., Kahn, M., McKay, B., Jermiin, L., 2000. TrExML: A maximum likelihood program for extensive tree-space exploration. Bioinformatics 16, 383–394.