Rearrangements in phylogenetic inference: Compare, model, or encode?

Bernard M.E. Moret[1], Yu Lin[1], and Jijun Tang[2]

[1] Laboratory for Computational Biology and Bioinformatics, EPFL,
EPFL-IC-LCBB INJ230, Station 14, CH-1015 Lausanne, Switzerland
{bernard.moret,yu.lin}@epfl.ch
[2] Department of Computer Science and Engineering,
University of South Carolina, Columbia, SC 29208, USA
jtang@cse.sc.edu

Abstract. Genomic rearrangements were first used for phylogenetic analysis in the late 1920s, but this work was largely ignored until the 1980s, when the sequencing of organellar genomes enabled the first genome-wide comparisons. G. Watterson *et al.* proposed to measure what later became the inversion distance between two genomes, J. Palmer *et al.* published papers on the evolution of mitochondrial and chloroplast genomes in plants, and D. Sankoff and W. Day published the first algorithmic paper on phylogenetic inference from rearrangement data, giving rise to a fertile field of mathematical, algorithmic, and biological research. By using rearrangement data, we bypass the pitfalls of multiple sequence alignment and benefit from rare genomic events more directly linked to function. As the cost of sequencing whole genomes has plummeted, we have witnessed a huge increase in work on duplications and rearrangements.

Distance measures for sequence data are simple to define, but those based on rearrangements proved to be complex mathematical objects. Whereas mutations events for sequence data are purely local and permutable at will, rearrangements are inherently nonlocal and subject to strong ordering constraints. The first approaches for phylogenetic inference from rearrangement data, due to D. Sankoff and various coauthors, used model-free distances, such as synteny (colocation on a chromosome) or breakpoints (disrupted adjacencies). The development of algorithms for distance and median computations led to modelling approaches based on documented biological mechanisms. However, the multiplicity of such mechanisms and the lack of knowledge about their relative preponderance pose serious challenges. A unifying framework, proposed by S. Yancopoulos *et al.* and popularized by D. Sankoff, has become the accepted model used in research on the algorithmics of rearrangements, leading to precise distance corrections and efficient algorithms for median estimation, and thereby enabling large-scale phylogenetic inference from high-resolution genomes using both distance and maximum-parsimony methods

In phylogenetic inference, likelihood-based methods outperform both distance and maximum-parsimony methods, but using such methods with rearrangement models has proved problematic. Thus we have returned to an idea we proposed twelve years ago: encoding the genome structure into sequences to use methods designed for 0/1 character data. By setting a simple bias in the ground probabilities, we have attained levels of performance comparable, in terms of both speed and accuracy, to the best sequence-based methods. Unsurprisingly, the idea of injecting such a bias was first proposed by D. Sankoff in 1998.

1 Introduction

Rearrangement data bypass multiple sequence alignment—an often troublesome step in sequence analysis; they represent the outcome of events much rarer than simple nucleotide mutations and thus hold the promise of high accuracy and of a reach extending back to the very distant past; and they are more closely tied to function (and through it to evolutionary selection) than point mutations. These attractive characteristics are mitigated by our limited understanding of the mechanisms causing large-scale structural changes in genomes.

The use of genomic rearrangements in phylogeny dates back to the early days of genetics, with a series of papers in the 1930's from A. Sturtevant and T. Dobzhansky on inversions in the genome of Drosophila pseudoobscura [21, 87]. However, this early foray had no successor until the 1980s, when G. Watterson proposed to build phylogenies from pairwise distances between circular genomes under inversions [100], J. Palmer and various coauthors published a series of papers on the structure and evolution of mitochondrial and chloroplast genomes in plants [37, 61], including studies of inversions in these genomes and their use in phylogenetic inference [22, 62], and D. Sankoff and W. Day published the first algorithmic paper on phylogenetic inference from rearrangement data [17]. Many hundreds of papers have been published since then on the use of rearrangement data in phylogenetic inference, by biologists, mathematicians, and computer scientists. A first major conference was organized by D. Sankoff and J. Nadeau in 2000 [79], followed by the yearly RECOMB Workshop on Comparative Genomics, started by J. Lagergren, B. Moret, and D. Sankoff.

Algorithms for phylogenetic inference (from any type of data) fall into three main categories. Simplest are the distance-based methods, which reduce the input data to a matrix of pairwise (evolutionary) distances. Next come the Maximum Parsimony (MP) methods, which attempt to find a tree that minimizes the total number of changes required to explain the data. Most complex are the probabilistic methods, either Maximum Likelihood (ML) or Bayesian, which attempt to find a tree (or a population of trees) that maximizes the conditional or posterior probability of observing the data.

In sequence-based phylogenetic inference, distance measures have a long history; all are simple and all have corresponding "distance corrections," that is, maximum-likelihood estimators that yield an estimate of the true (as opposed to observed) pairwise distance. Their simplicity derives in good part from the fact that they are based on observed results (rather than proposed mechanisms) and on local models of change. Distances based on rearrangements, however, turned out to be complex concepts: it remains possible to define a mechanism-free distance, as D. Sankoff did in 1992 [72, 78], but the model cannot be local, as a single rearrangement can alter the location of almost every gene in the genome. Other distance measures based on mechanisms, whether biological or mathematical (inversion distance, transposition distance, DCJ distance, etc.), have proved yet harder to characterize, with work still ongoing. The absence of localization of changes also means that the fundamental assumption of parsimony- and likelihood-based approaches, independence between different regions of the sequence, does not hold for rearrangement data, thus creating enormous algorithmic difficulties.

Thus the first approaches for phylogenetic inference from rearrangement data used simple distances [6, 74], such as the observed number of nonconserved adjacencies;

attempts were also made to encode observed adjacencies in order to use sequence-based inference methods [15, 96]. The development of better algorithms for distance and median computations led to a period during which most approaches were based on documented biological mechanisms for rearrangements, such as inversions, transpositions, translocations, etc. The ability to compute some of these distances efficiently also led researchers to devise suitable distance corrections, some of which resulted in substantial improvements in the quality of inference [50, 51]. The multiplicity of biological mechanisms posed a serious problem, however, since there were no data to quantify their relative preponderance during the course of evolution. Fortunately, a unifying framework was proposed by S. Yancopoulos *et al*. in 2005 [104] and popularized by D. Sankoff; since then nearly all research on the algorithmics of rearrangements have used this model, leading to very precise distance corrections [42] and efficient and accurate algorithms for median estimation [103] and tree scoring [102]. The precise distance estimates overcame, to a large extent, the weakness of distance-based methods, so that large-scale phylogenetic inference from high-resolution genomes became possible [68].

With sequence data, likelihood-based methods outperform distance-based and MP methods [97]—the two classes of methods used with rearrangement data to date. Direct use of likelihood-based methods was attempted once using a Bayesian approach [39], but the complex mechanisms of rearrangement created insurmountable convergence problems. In the latest step in the evolution of methods for phylogenetic inference from rearrangement data, we have returned to the idea of encoding the genome structure into binary sequences [41] that we first proposed a dozen years ago [15, 96]. This time, however, we use an ML method for inference and inject a bias in its ground probabilities to reflect our better understanding of the evolution of genomic structure; and we have attained levels of performance, in terms of both speed and accuracy, that compare favorably to the best sequence-based methods. Unsurprisingly, the idea of a bias in ground probabilities was first proposed by D. Sankoff in 1998 [75]. Today, then, after over 15 years of research by dozens of groups, phylogenetic inference from rearrangement data is best carried out using a mechanism-free approach and a simple statistical bias in the one operation allowed under the model (transitions between the 0 and the 1 state for each character), much as D. Sankoff advocated from the beginning.

What we present below is a survey of phylogenetic inference from rearrangement data, as viewed through the lens of the work of our group in this area, in tribute to David Sankoff, pioneer and mentor, who more than anyone is responsible for the blossoming of research, unfolding over the last 30 years, on models and algorithms for the evolution of genome structure through rearrangements, duplications, and losses.

2 Background

2.1 Genome representations

Each chromosome of the genome is represented by an ordered list of identifiers, each identifier referring to a syntenic block or, more commonly, to a member of a gene family. (In the following, we shall use the word “gene” in a broader sense to denote elements of such orderings and refer to such orderings as “gene orders.”) A gene is a stranded sequence of DNA that starts with a tail and ends with a head. The tail of a gene *a* is

denoted by a^t and its head by a^h. We are interested, not in the strand of one single gene, but in the connection of two consecutive genes in one chromosome. Due to different strandedness, two consecutive genes b and c can be connected by one *adjacency* of the following four types, $\{b^t, c^t\}$, $\{b^h, c^t\}$, $\{b^t, c^h\}$ and $\{b^h, c^h\}$. If gene d lies at one end of a linear chromosome, then we have a singleton set, $\{d^t\}$ or $\{d^h\}$, called *telomere*. (These definitions use the notation codified by Bergeron *et al.* [4].) Given a reference set of n genes $\{g_1, g_2, \cdots, g_n\}$, a genome can be represented by an *ordering* of some multi-subset of these genes. Each gene is given a sign to denote its orientation (strandedness). A genome can be *linear* or *circular*. A linear genome is simply a permutation on the multi-subset, while a circular genome can be represented in the same way under the implicit assumption that the permutation closes back on itself. A multiple-chromosome genome can be represented in the same manner, with telomeres indicating the start and end of a chromosome.

2.2 Evolutionary events

Let G be the genome with signed ordering of $g_1, g_2, \cdots, g_n$. An *inversion* between indices i and j $(i \leq j)$, produces the genome with linear ordering

$$g_1, g_2, \cdots, g_{i-1}, -g_j, -g_{j-1}, \cdots, -g_i, g_{j+1}, \cdots, g_n$$

A *transposition* on genome G acts on three indices i, j, k, with $i \leq j$ and $k \notin [i, j]$, picking up the interval $g_i, g_{i+1}, \cdots, g_j$ and inserting it immediately after g_k. Thus genome G is replaced by (assume $k > j$):

$$g_1, \cdots, g_{i-1}, g_{j+1}, \cdots, g_k, g_i, g_{i+1}, \cdots, g_j, g_{k+1}, \cdots, g_n$$

An *insertion* is the addition of one gene or a segment of genes, and a *deletion* is the loss of same. A section of the chromosome can be *duplicated*, through *tandem duplication*, in which the copied section is inserted immediately after the original, or through *transposed duplication*, in which the copied section is inserted somewhere else. With multichromosomal genomes, additional operations that involve two chromosomes come into play: *translocation* (one end segment in one chromosome is exchanged with one end segment in the other chromosome), *fission* (one chromosome splits and becomes two), and *fusion* (two chromosomes combine to become one).

2.3 Distance computation and pairwise genome comparison

Given two genomes G and G' on the same set of genes, a breakpoint in G is defined as an ordered pair of genes (g_i, g_j) that forms an adjacency in G but not in G'. The *breakpoint distance* [74] is simply the number of breakpoints in G relative to G'.

We define an *edit distance* as the minimum number of events required to transform one genome into the other. (The breakpoint distance is not an edit distance.) We can apply edit distances to any collection of evolutionary operations. Thus we can consider distances under inversions, insertions, and deletions; we can also consider transpositions (within the same chromosome) and translocations (across chromosome), as well

as duplications, whether of just one gene, a large segment of a chromosome, or, in the extreme case, the entire genome. The range of possible evolutionary events also include chromosome fusion (merging two chromosomes into one) and fission (the reverse operation), as well as chromosome linearization (turning a circular chromosome into a linear one) and circularization (the reverse event).

When we compute edit distances, we find the minimum number of events required to transform one genome into another, yet evolution need not have proceeded along the shortest path to the current genomes. What we would like to recover isthe *true evolutionary distance*, that is, the mean number of evolutionary events on the paths to the two genomes from their last common ancestor. By computing edit distances, we may significantly underestimate the true distance—a problem that also arises with pairwise distances between two sequences or, indeed, between any two objects evolving through some given operation. In sequence analysis, *distance corrections* were devised to provide maximum-likelihood estimators of the true distance given the edit distance. The same can be done, formally or empirically, for rearrangement distances.

2.4 Phylogenetic reconstruction and ancestral genome estimation

Working with genome rearrangement data is computationally much harder than with sequence data. For example, given a fixed phylogeny, the minimum number of evolutionary events can be found in linear time if the leaves of the phylogeny are labeled with DNA or protein sequences, whereas such a task for rearrangement data is NP-hard, even when the phylogeny has only three leaves [10, 63].

Methods to reconstruct phylogenies from genome rearrangement data include distance-based methods (such as neighbor-joining [70] and minimum-evolution [19]), maximum parsimony and likelihood methods based on encodings [15,35,41,96], and optimization methods, usually based on median computations.

Distance-based methods transform the input genomes into a matrix of pairwise distances, from which the phylogeny is then inferred. Any of the distances introduced above can be used, but inference using estimates of true distances generally outperforms inference based on edit distances [51]; similarly, any distance-based inference method can be used, but FastME [19, 20] appears to outperform most others.

Optimization methods to date have been based on maximum parsimony—they seek the tree and associated ancestral data that minimizes the total number of events. Because this minimization is hard even for a fixed tree of just three leaves, a heuristic first proposed by D. Sankoff in another context [71] and then reused by him for breakpoints (see, e.g., [73]), is widely used. The heuristic uses an iterative improvement approach, based on the recomputation of medians: given 3 genomes, find a single genome that minimizes the sum of the pairwise distances between itself and each of the 3 given genomes.

3 Comparing: Distance Computations

The first pairwise distance measure used to compare two genomes was a measure of conservation of synteny due to D. Sankoff and J. Nadeau [24, 80]. Inversion and DCJ distances are now the two most commonly used genomic distances. Neither inversions

nor DCJ events affect the gene content of a genome: they are pure rearrangements. Little by little, events that do affect gene content, such as deletions, insertions, and duplications, were included in distance computations—a necessity with real genomes. Although various methods have been proposed to combine rearrangements and duplications and losses [14, 84], the problem remains poorly solved.

3.1 Inversion distance

D. Sankoff [72] formulated the fundamental computational problem about inversions: given two signed permutations on the same index set, what is their edit distance under inversions? The breakthrough came in 1995, when S. Hannenhalli and P. Pevzner provided a polynomial-time algorithm to solve this problem [32]. Their algorithm is based on the *breakpoint graph* (see Fig. 1). Assume we are given two genomes, G_1 and G_2, with the same n genes and assume that G_2 is the identity. Add two extra "genes" (mathematical markers), gene 0 on the left of the genome and gene $n+1$ on the right. For each gene i in G_1, the breakpoint graph contains two vertices i^h and i^t, connected with two (colored) undirected edges, one for each genome. The *desire edges* (also called grey edges) connect i^h and $(i+1)^t$ for all $0 \leq i \leq n$ and represent the identity genome G_2; they are shown with dashed arcs in Fig. 1. The *reality edges* (also called black edges) represent the rearranged genome, G_1; for each adjacency (i, j) in G_1, a reality edge begins at vertex i^h if gene i is positive or at at vertex i^t if i is negative, and ends at vertex j^t if j is positive or at vertex j^h if j is negative. Reality edge are shown as solid lines in Fig.1. These edges form cycles of even length that alternate between reality and desire edges; denote by $c(G_1, G_2)$ the number of these cycles. Overlapping cycles in certain configurations create structures known as *hurdles*; denote by $h(G_1, G_2)$ the number of these hurdles. Finally, a very unlikely [88] configuration of hurdles can form a *fortress*. S. Hannenhalli and P. Pevzner [32] proved that the inversion distance between two signed permutations of n genes is given by: $n - c(G_1, G_2) + h(G_1, G_2) + (1 \text{ if a fortress is present}, 0 \text{ otherwise})$.

D. Bader, M. Yan, and B. Moret [1] later showed that this edit distance can be computed in linear time. Extending this distance to multichromosomal genomes can be done through a reduction to the unichromosomal case using "capping," a subtle process that required several iterations before it was done right [3, 33, 38, 93]. The various operations supported under this multichromosomal model (for which see the next section),

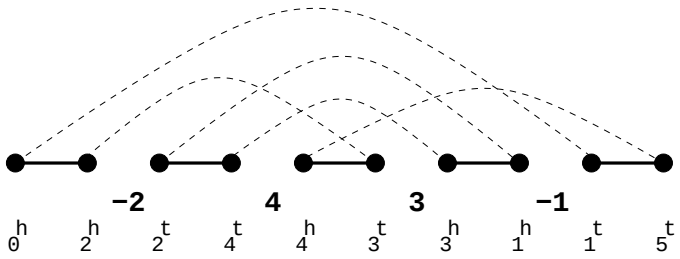


Fig. 1. Breakpoint graph between genome (-2 4 3 -1) and the identity genome (1 2 3 4).

all of which keep the gene content intact, give rise to what we shall call the HP-distance. The *transposition distance* is known to be NP-hard to compute [8]; attempts at defining distances combining transpositions and inversions have so far proved unsuccessful.

Moving to distances between genomes of unequal gene content has proved very challenging. N. El-Mabrouk [25] first extended the results of S. Hannenhalli and P. Pevzner to the computation of edit distances for inversions and deletions and gave a heuristic for inversions, deletions, and non-duplicating insertions. The distance computation is NP-hard when duplications and inversions are present [13]. M. Marron et al. [48] gave a guaranteed approximation for edit distances under arbitrary operations (including duplications and deletions).

3.2 DCJ distance

S. Yancopoulos, O. Attie, and R. Friedberg [104] proposed a double-cut-and-join (DCJ) operation that accounts for inversions, translocations, fissions, and fusions, yielding in a new genomic distance that can be computed in linear time. As its name indicates, a DCJ operation makes a pair of cuts in the chromosomes and reglues the four cut ends into two adjacencies or telomeres, giving rise to four cases:

- A pair of adjacencies $\{i^u, j^v\}$ and $\{p^x, q^y\}$ can be replaced by the pair $\{i^u, p^x\}$ and $\{j^v, q^y\}$ or by the pair $\{i^u, q^y\}$ and $\{j^v, p^x\}$.
- An adjacency $\{i^u, j^v\}$ and a telomere $\{p^x\}$ can be replaced by the adjacency $\{i^u, p^x\}$ and telomere $\{j^v\}$ or by the adjacency $\{j^v, p^x\}$ and telomere $\{i^u\}$.
- A pair of telomeres $\{i^u\}$ and $\{j^v\}$ can be replaced by the adjacency $\{i^u, j^v\}$.
- An adjacency $\{i^u, j^v\}$ can be replaced by the pair of telomeres $\{i^u\}$ and $\{j^v\}$.

Given two genomes G_1 and G_2, their DCJ distance can be computed using the adjacency graph $AG(G_1, G_2)$. The adjacency graph has a vertex for each adjacency and each telomere of G_1 and G_2 and, for each $u \in G_1$ and $v \in G_2$, has $|u \cap v|$ edges between u and v (see Fig. 2). S. Yancopoulos *et al.* [104] and, for the multichromosomal formulation, A. Bergeron *et al.* [4], showed that the DCJ distance between G_1 and G_2 is just $d_{DCJ}(G_1, G_2) = n - (C + I/2)$, where C is the number of cycles and I the number

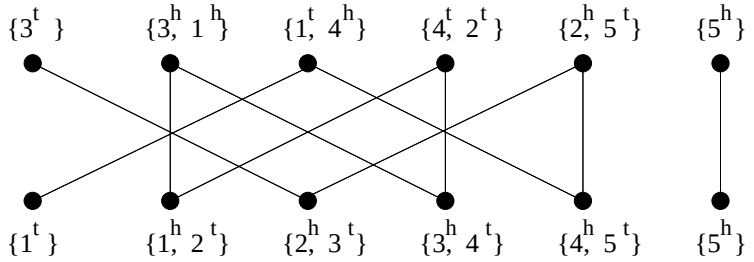


Fig. 2. Adjacency graph and DCJ distance of two genomes $G_1 = (3, -1, -4, 2, 5)$ and $G_2 = (1, 2, 3, 4, 5)$. The number of cycles C is 1, the number of odd paths I is 2, the DCJ distance is $N - (C + I/2) = 3$.

of odd paths in the adjacency graph. While there is no single biological mechanism to mirror the DCJ operation, researchers everywhere have adopted the DCJ model in their work because of its mathematical simplicity and because of its observed robustness in practice.

The DCJ operator, like the operator in the multichromosomal rearrangement model of S. Hannenhalli and P. Pevzner, preserves gene content. The DCJ model has been extended, with mixed results, to handle insertions and deletions [14, 84], duplications [2], and all three [82]. Inversion distances, HP distances, and DCJ distances are all implemented in the UniMoG software [34].

3.3 Estimating true distances

Edit distances underestimate the true number of events, particularly so when the two genomes are distant. Estimating the true distance through a maximum-likelihood approach requires an evolutionary model. Since models are the subject of the next section, we focus here on the estimators themselves. The IEBP estimator [98] (and its improved version [95]) uses a Markov chain formulation to provide an exact formula for the the expected number of events (inversions, transpositions, and inverted transpositions—the events in the Nadeau-Taylor model) as a function of the breakpoint (edit) distance. The EDE estimator [53] uses curve-fitting to approximate the most likely number of evolutionary events, derived by simulating known numbers of inversions and comparing that number against the computed inversion (edit) distance. The formula thus takes an inversion (edit) distance and returns an estimate of the true number of inversions. Although EDE was designed just for inversions, experience shows that it works well even with a significant number of transpositions and that its use significantly improves the accuracy of distance-based phylogenetic reconstruction [99].

K. Swenson *et al.* [89] proposed a heuristic to approximate the true evolutionary distance under inversions, duplications, insertions, and deletions for unichromosomal genomes. Y. Lin and B. Moret [42] developed a true distance estimator for the DCJ model based on the DCJ distance; later Y. Lin *et al.* [45] gave an estimator for the true number of events in the DCJ model based directly on the lists of gene adjacencies—a useful generalization as it can be used even for lists of adjacencies that do not define a "real" genome. Given genome G, for any genome G^* , we can divide the adjacencies and telomeres of G^* into four sets, $SA(G^*)$, $ST(G^*)$, $DA(G^*)$ and $DT(G^*)$, where $SA(G^*)$ is the set of adjacencies of G^* that also appear in G, $ST(G^*)$ is the set of telomeres of G^* that also appear in G, $DA(G^*)$ is the set of adjacencies of G^* that do not appear in G, and $DT(G^*)$ is the set of telomeres of G^* that do not appear in G. Then we can calculate a vector $V_G(G^*) = (SA^*, ST^*, DA^*, DT^*)$ to represent the genome G^* in terms of G, where SA^*, ST^*, DA^* and DT^* are the cardinalities of the sets $SA(G^*)$, $ST(G^*)$, $DA(G^*)$ and $DT(G^*)$, respectively. Obviously, we have $2n = 2SA^* + ST^* + 2DA^* + DT^*$. Let G^k be the genome obtained from $G = G^0$ by applying k randomly selected DCJ operations. The $(i+1)$st DCJ operation is selected from a uniform distribution of all possible DCJ operations on the current genome G^i. We can compute the vector $V_G(G^k) = (SA^k, ST^k, DA^k, DT^k)$ to represent the genome G^k with respect to G. For any integer $k > 0$, we can also produce the estimate $\widetilde{E}(V_G(G^k)) = (\widetilde{SA^k}, \widetilde{ST^k}, \widetilde{DA^k}, \widetilde{DT^k})$ for the ex-

pected vector $E(V_G(G^k))$. We then use $\widetilde{SA^k}$ to approximate the expected number of adjacencies present in both G and G^k and compute SA^F from G and G^F. The estimated true number of evolutionary events is then the integer k that minimizes the difference $|SA^F - \widetilde{SA^k}|$. This estimator is quite robust and achieves better performance than the EDE estimator; it was later extended [45] to include gene losses and duplications.

4 Modelling Genomic Evolution

Models for genomic rearrangements have been studied intensely over the last 30 years by biologists, computer scientists, and mathematicians—for an overview of the work of the latter two, see, e.g., [29, 52, 55]. We briefly review the main models used in phylogenetic inference.

4.1 Inversions

Of the various genomic rearrangements studied, perhaps the simplest and best documented is the *inversion* (also called reversal), through which a segment of a chromosome is inverted in place. In 1989, D. Sankoff and M. Goldstein formalized a probabilistic model of genomic inversions in which a chromosome is represented as a permutation of gene indices [77]; in this framework, an inversion acts on an interval of the permutation by reversing the order in which the indices appear within the interval. A year later, D. Sankoff *et al.* [76] also proposed to apply this model to the assessment of evolutionary relationships among a number of bacterial genomes using genetic maps. Two years later, D. Sankoff used gene-order data from 16 mitochondrial genomes from fungi and other eukaryotes to infer a species phylogeny; the study used nearly complete genomic sequences for the organelles [78].

4.2 The generalized Nadeau-Taylor model

The generalized Nadeau-Taylor (NT) model [57] deals with inversions, transpositions and inverted transpositions; it assumes that the number of each of those three events obeys a Poisson distribution on each edge, that the relative probabilities of each type of event are fixed across the tree, and that events of a given type have the same probability, regardless of the location and size of the affected region. The model thus has two main parameters—two of the three relative probabilities of occurrence of the three types of events. The generalized Nadeau-Taylor model can be extended to include insertions, deletions, or duplications in the triplet, by choosing the relative probabilities for all events. Nevertheless, the generalized NT model remains far from realistic, and many features of genome evolution are lacking in this model, such as hot spots (a much debated feature [18, 64, 65, 81]), operons, as well as fission and fusion events.

4.3 The HP model

The first model of multichromosomal rearrangements was given by S. Hannenhalli and P. Pevzner [33]. This HP model includes inversions, translocations, fusions, and fissions (see Fig. 3).

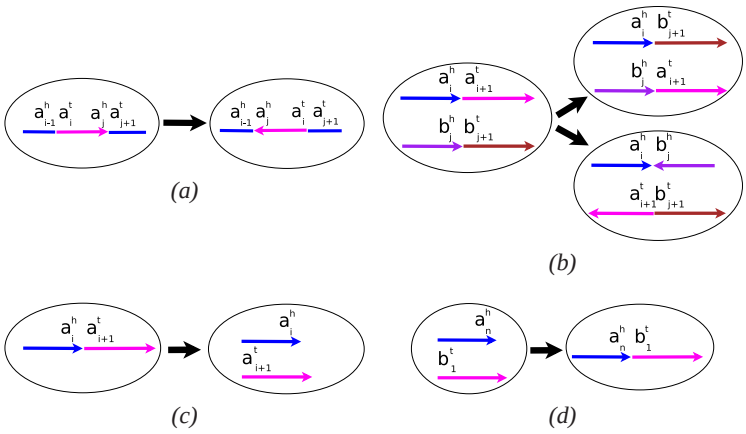


Fig. 3. Possible rearrangements in the HP model: (a) inversion, (b) translocation, (c) fission, and (d) fusion.

- *Inversion*: Given a linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$, reverse all genes between a_i and a_j yields $(a_1 \dots a_{i-1}\text{-}a_j \dots \text{-}a_i a_{j+1} \dots a_n)$. Two adjacencies, $\{a^h_{i-1}, a^t_i\}$ and $\{a^h_j, a^t_{j+1}\}$, are replaced by two others, $\{a^h_{i-1}, a^h_j\}$ and $\{a^t_i, a^t_{j+1}\}$.
- *Translocation*: Given two linear chromosomes $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $B = (b_1 \dots b_j b_{j+1} \dots b_m)$, exchange two segments between these two chromosomes. There are two possible outcomes, $(a_1 \dots a_i b_{j+1} \dots b_m)$ and $(b_1 \dots b_j a_{i+1} \dots a_n)$ or $(a_1 \dots a_i\text{-}b_j \dots \text{-}b_1)$ and $(\text{-}b_n \dots \text{-}b_{j+1} a_{i+1} \dots a_n)$. Two adjacencies, $\{a^h_i, a^t_{i+1}\}$ and $\{b^h_j, b^t_{j+1}\}$, are replaced by $\{a^h_i, b^h_{j+1}\}$ and $\{a^t_{i+1}, b^t_j\}$ or $\{a^h_i, b^h_j\}$ and $\{a^t_{i+1}, b^t_{j+1}\}$.
- *Fission*: Given a linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$, split A into two new linear chromosomes, $(a_1 \dots a_i)$ and $(a_{i+1} \dots a_n)$. The adjacency $\{a^h_i, a^t_{i+1}\}$ is replaced by two telomeres $\{a^h_i\}$ and $\{a^t_{i+1}\}$.
- *Fusion*: Given two linear chromosomes $A = (a_1 \dots a_n)$ and $B = (b_1 \dots b_m)$, concatenate these two linear chromosomes into a single new chromosome $(a_1 \dots a_n b_1 \dots b_m)$. Two telomeres, $\{a^h_n\}$ and $\{b^t_1\}$, are replaced by one adjacency $\{a^h_n, b^t_1\}$.

In 2009, A. Bergeron *et al.* [5] gave an optimal linear-time algorithm to compute the HP distance with a simple formula.

4.4 The DCJ model

The HP model cannot mix linear and circular chromosomes. In such a mix, additional operations come into play: linearization and circularization, which transform circular chromosomes into linear ones and vice versa. The DCJ operation, on the other hand, does model such a mix and supports every simple rearrangement: inversions, transpositions, block exchanges, circularizations, and linearizations, all of which act on a single chromosome, and translocations, fusions, and fissions, which act on a pair of chromosomes. The DCJ model is less well motivated than the HP model in terms of biology:

whereas inversions and translocations are well documented, the additional rearrangements possible in the DCJ model may not correspond to actual evolutionary events. For example, if the two cuts are in the same linear chromosome, one of the two nontrivial rejoinings causes a fission, excising a portion of the original chromosome and packaging that portion as a new circular chromosome—something usually called a "circular intermediate," the name itself denoting the opinion that such structures are at best ephemeral. (In the best tradition of biology, where "anything that can happen already has," the existence of circular intermediates has been inferred in vertebrates [23, 31].) A simple modification to the DCJ model (forbidding the least realistic operation) can lead genomes into the two stable structures (single circular chromosome or multiple linear chromosomes) found in the vast majority of prokaryotes and eukaryotes, respectively [43]. The main argument for the DCJ model is its mathematical simplicity: it is much easier to reason about and devise algorithms for a model with a single operation (however multifaceted) than for one with a "zoo" of operations.

4.5 Models for rearrangements, duplications, and losses

Gene (or segment) duplications and losses have long been studied by geneticists and molecular biologists. A particularly spectacular version of duplication is *whole genome doubling (WGD)*, the duplication of the entire genome—a very rare event, but one often viewed as responsible for much the diversity of life forms. WGD has been of particular interest to evolutionary biologists for many years. D. Sankoff *et al.* integrated WGD with rearrangements, and introduced the Genome Halving Problem [26–28]: from the present-day doubled and rearranged genome, recover the pre-doubling ancestral genome using a criterion of maximum parsimony. Since the ancestral solutions are often numerous, D. Sankoff *et al.* proposed to take into account external reference genomes as outgroups, to guide and narrow down the search [92, 107].

Segmental duplications and gene losses do not affect just gene content: they can mimic the effect of rearrangements; the most obvious example is transposition: instead of moving a segment from one location to another, one can envision deleting that segment from its original location and inserting it at its new location. N. El-Mabrouk [25] gave an exact algorithm to compute edit distances for inversions and losses and also a heuristic to approximate edit distances for inversions, losses, and nonduplicating insertions. Her work was then extended and refined by M. Marron *et al.* [48]. In 2008, S. Yancopoulos and R. Friedberg [105] gave an algorithm to compute edit distances under deletions, insertions, duplications, and DCJ operations, under the constraint that each deletion can only remove a single gene. These and other approaches targeted the edit distance, not the true evolutionary distance. K. Swenson *et al.* [89] gave a first heuristic to approximate the true evolutionary distance under inversions, duplications, and losses; more recently, Y. Lin *et al.* [45] gave an algorithm to estimate the true evolutionary distance under deletions, insertions, duplications, and inversions. Rearrangements, duplications and losses have also been addressed in the framework of ancestral reconstruction [47, 60], a topic of rapidly increasing interest.

4.6 Inferring phylogenies using models

In phylogenetic inference, the main use of models is to guide inference phrased as an optimization problem, using maximum parsimony or maximum likelihood criteria. The first program used for phylogenetic inference from rearrangement data, D. Sankoff's BPAnalysis, used the breakpoint model and median-based heuristics aimed at obtained a tree of maximum parsimony. The approach proposed by D. Sankoff is based on scoring each possible tree separately. Since scoring a single tree is NP-hard [63], the scoring procedure is a heuristic, using iterative improvement, that D. Sankoff himself had proposed much earlier for multiple sequence alignment on trees [71]. First, each internal node of the tree is initialized with some "ancestral genome." Then, and until no changes can be made to any internal node, each internal node is examined in turn (according to some chosen scheduling): the median of its immediate neighbors' "genomes" is computed and, if better (giving a lower parsimony score) than the current genome at the node in question, is used to replace that genome. Computing the median of three or more genomes is itself NP-hard [10, 63]—indeed, it is just the simplest case of parsimony scoring, for a tree of diameter 2. Thus the overall procedure nests some unknown number of instances of an NP-hard problem within an exponential loop: obviously such an approach cannot scale up very far.

D. Sankoff and M. Blanchette [74] provided the first software package for the problem, BPAnalysis, subsequently improved by our group with GRAPPA [16, 56]. BPAnalysis handled 8 genomes of 40 genes each; GRAPPA [56] scaled to 15 genomes and a few hundred genes and supported both breakpoint and inversion models (with both edit and estimated true distances). Reusing the GRAPPA code for inversion distance computation, P. Pevzner's group produced MGR [7], which could handle multichromosomal genomes and, rather than scoring every tree, used a construction heuristic to build a single tree in incremental fashion. In doing the latter, MGR came closer to sequence-based MP (or ML) algorithms, none of which, naturally, scores every tree; but the lack of tight bounding for subtrees (an indirect consequence of the complexity of rearrangement operations) meant that the construction heuristic gave poor results—in direct comparisons, the exhaustive approach of BPAnalysis and GRAPPA inferred better trees. Further developments of GRAPPA, such as very tight bounds on entire trees (but not on subtrees) [91], reduced the space to be explored by orders of magnitude, while high-performance methods [49] further increased its speed, yet scaling such an algorithm requires an entirely different algorithmic approach, such as the Disk-Covering Methods developed by T. Warnow *et al.* [36], used with some success in combination with GRAPPA [90].

D. Sankoff [73] had shown that seeking a median that minimizes the breakpoint distance can be transformed into a special instance of the well-studied Traveling Salesperson Problem and thus can be solved relatively efficiently (as shown in GRAPPA). In practice, however, computing breakpoint medians yields poor solutions; a better approach is to use the inversion median [50], although exact solvers for this problem scale poorly [11, 83, 106]. Thanks to the relative simplicity of the DCJ model, DCJ median solvers are easier to design and perform better than inversion solvers, so that parsimony methods using DCJ median solvers outperform other methods in terms of speed and accuracy [69]. Among existing DCJ median solvers, the best to date appears to be

ASMedian [101], due to W. Xu and D. Sankoff, and our extension GASTS [102], used to produce the parsimony score of a given tree.

As the algorithmic community expended considerable energy on improving the initialization, the computation of medians, and the exploration of the search space, it became clear that scoring through iterative improvement was going to cause serious problems, due to a combination of two factors. First, the number of local minima for parsimony scoring is huge, so that the number of times a median gets improved tends to one as the instance gets larger, with consequent poor results. Secondly, the error accumulates quickly as the distance from the leaves increases: what works reasonably well on very small trees of 10–20 leaves (in which most internal nodes are within 1–2 edges of a leaf) fails more and more (and worse and worse) as the trees became larger. Better median computation helps reduce the second problem, as better initialization helps reduce the first: both are deployed in the GASTS scoring software [102], but at a significant expense of computational time.

The difficulty of deriving bounds for the completion of partial trees means that the standard approaches used in sequence-based MP and ML inference cannot be used today with rearrangement data. There has been one attempt to use a probabilistic approach, using Bayesian inference (the BADGER tool) [39], but it could not achieve reasonable scaling and kept suffering from convergence problems since it was not clear what steps should be used in the construction of the Markov chains. Until new results in mathematics and algorithms are obtained to provide bounds on partial trees, model-based inference of phylogenies under MP or ML remains restricted to small instances.

5 Encodings

Distance-based methods suffer from the problem of *saturation*: the observed changes may be only a small piece of the history of changes and any attempt at estimating the true number of changes from a large observed number of changes will suffer from a very large variance. In good part, this problem stems from the pairwise approach of these methods: if two leaves in the tree have the root as their last common ancestor, then the pairwise distance is an unreliable predictor of the length of the tree path connecting these two leaves. Methods that score trees rather than pairs can take advantage of the smaller evolutionary steps represented by tree edges: they do not compute any pairwise distances between leaves, focusing instead on pairwise comparisons between the endpoints of a tree edge. Unfortunately the optimization problem for rearrangement data solved by such methods (maximum parsimony or maximum likelihood) is NP-hard even on a fixed tree. For sequence data, however, scoring a single tree under parsimony can be done in linear time, while fast and well tested packages exist to compute the MP or ML tree. Thus a natural approach is to produce sequences from the input permutations, use a sequence-based phylogenetic inference package, then analyze the resulting tree in terms of rearrangements.

5.1 Parsimonious methods

The idea of encoding the genome structure into sequences was introduced a dozen years ago [15, 96], based on an earlier characterization approach of D. Sankoff and M. Blanchette [74]. Two encoding methods were proposed:

- Maximum Parsimony on Binary Encodings (MPBE): each genome is translated into a binary sequence, where each site from the binary sequence corresponds to an adjacency present in at least one of the input genomes. Uninformative sites are discarded.
- Maximum Parsimony on Multistate Encodings (MPME): each genome of n genes is translated into a sequence of length $2n$, where site i takes the index of the gene immediately following gene i and site $n+i$ takes the index of the gene immediately following gene -i, for $1 \leq i \leq n$.

Results show that an encoding based on adjacencies preserves useful phylogenetic information that a parsimonious search can put to good use. However, both MPBE and MPME proved too computationally expensive compared to distance-based methods, while MPME was ill suited for use with existing MP and ML inference packages because of its large number of character states. Moreover, accuracy with MP inference was not significantly better than accuracy using distance-based methods.

5.2 Likelihood-based approaches

In the last few years, likelihood-based inference packages such as `RAxML` [86] and `FastTree` [67] have largely overcome computational limitations and allowed reconstructions of large trees (with tens of thousands of taxa) and the use of long sequences (to several hundred thousand characters). In 2011, F. Hu *et al.* [35] applied likelihood-based inference to an unusual encoding scheme, in which the same adjacency could be translated into multiple character positions. Results on bacterial genomes were promising, but difficult to explain, while the method appeared too time-consuming to handle eukaryotic genomes.

In 2013, we described MLWD (Maximum Likelihood on Whole-genome Data) [41], a new approach that encodes genomic structure into binary sequences using both gene adjacencies and gene content, estimates the transition parameters for the resulting binary sequence data, and finally uses sequence-based ML reconstruction to infer the tree. For each adjacency or telomere within the entire collection of genomes, there exists exactly one position in the sequence, with 1 indicating presence of this adjacency in a genome and 0 indicating absence. If the total number of distinct genes among the input genomes is n, then the total number of distinct adjacencies and telomeres cannot exceed $\binom{2n+2}{2}$, but the actual number is typically far smaller—it is usually linear in n rather than quadratic. For each gene family within the collection of genomes, there is exactly one position, with the same meaning attributed to the Boolean values. For the two toy genomes of Fig. 4, the resulting binary sequences and their derivation are shown in Table 1.

Since the encodings are binary sequences, the parameters of the model are simply the transition probability from presence (1) to absence (0) and that from absence (0)

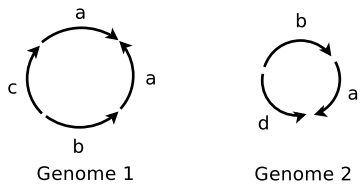


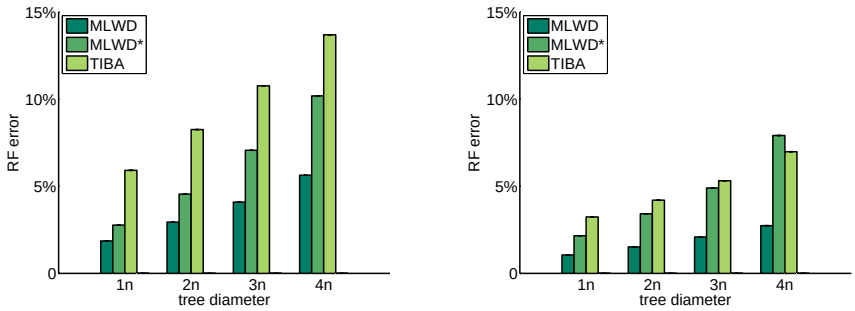
Fig. 4. Two toy genomes.

Table 1. The binary encodings for the two genomes of Fig. 4.

	adjacency information						content information			
	$\{a^h, a^h\}$	$\{a^t, b^h\}$	$\{a^t, c^h\}$	$\{b^t, c^t\}$	$\{a^h, d^h\}$	$\{b^t, d^t\}$	a	b	c	d
Genome 1	1	1	1	1	0	0	1	1	1	0
Genome 2	0	1	0	0	1	1	1	1	0	1

to presence (1). In rearrangements, every DCJ operation will select two adjacencies (or telomeres) uniformly at random, and, if adjacencies, break them to create two new adjacencies. Each genome has $n + O(1)$ adjacencies and telomeres ($O(1)$ is the number of linear chromosomes in the genome, viewed as a small constant). Thus the transition probability from 1 to 0 under one DCJ operation at some fixed index in the sequence is $\frac{2}{n+O(1)}$. Since there are up to $\binom{2n+2}{2}$ possible adjacencies and telomeres, the transition probability from 0 to 1 is $\frac{2}{2n^2+O(n)}$. Thus the transition from 0 to 1 is roughly $2n$ times less likely than that from 1 to 0. Despite the restrictive assumption that all DCJ operations are equally likely, this result is in line with general opinion about the probability of eventually breaking an ancestral adjacency (high) vs. that of creating a particular adjacency along several lineages (low)—a version of homoplasy for adjacencies. The probability of losing a gene independently along several lineages is high, whereas the probability of gaining the same gene independently along several lineages (the standard homoplasy) is low. Unsurprisingly, D. Sankoff first observed and studied such a bias in transitions of adjacencies in 1998 [75].

Fig. 5 shows the Robinson-Foulds (RF) error rates of three different approaches, MLWD, MLWD* and TIBA. (The RF rate measures the difference between two trees as the ratio of the number edges present in one tree but not in the other to the total number of edges and is the most commonly used measure of phylogenetic accuracy.) MLWD*, used as a control for bias, follows the same procedure as MLWD, but without setting the bias explicitly, while TIBA [44] is a fast distance-based tool to reconstruct phylogenies from rearrangement data, combining a pairwise distance estimator [42] and the FastME [19] distance-based reconstruction method. These simulations show that MLWD can reconstruct much more accurate phylogenies from rearrangement data than the distance-based approach TIBA, in line with experience in sequence-based reconstruction. MLWD also outperforms MLWD*, underlining the importance of estimating and setting the transition biases before applying the sequence-based maximum-likelihood method.



(a) 1'000 genes *(b) 5'000 genes*

Fig. 5. RF error rates for different approaches for trees with 100 species, with genomes of 1′000 and 5′000 genes and tree diameters from one to four times the number of genes, under the rearrangement model.

Fig. 6 (from [41]) shows the MLWD-inferred phylogeny for 68 eukaryotic genomes from the eGOB (Eukaryotic Gene Order Browser) database [46]. The database contains the order information of orthologous genes (identified by OrthoMCL [12]) of 74 different eukaryotic species; the total number of different gene markers in eGOB is around 100′000. We selected 68 genomes with from 3k to 42k gene markers—the remaining 6 genomes in the database have too few adjacencies (fewer than 3′000). We encoded the adjacency and gene content information of all 68 genomes into 68 binary sequences of length 652′000. Inferring this phylogeny (using RAxML and setting the bias ratio to 100) took under 3 hours on a desktop computer, showing that MLWD can easily handle high-resolution genomic data.

As shown in Fig. 6, all major groups in those 68 eukaryotic genomes are correctly identified, with the exception of Amoebozoa. Those incorrect branches with respect to Amoebozoa receive extremely low bootstrap values (0 and 2), indicating that they are very likely to be wrong. For the phylogeny of Metazoa, the tree is well supported from existing studies [66, 85]. For the phylogeny of model fish species (D. rerio, G. aculeatus, O. latipes, T. rubripes, and T. nigroviridis), two conflicting phylogenies have been published, using different choices of alignment tools and reconstruction methods for sequence data [58]. Our result supports the second phylogeny, which is considered as the correct one by the authors in their discussion [58]. For the phylogeny of Fungi, our results agree with most branches for common species in recent studies [30, 94]. It is worth mentioning that among three Chytridiomycota species C. cinereus, P. gramnis, and C. neoformans, our phylogeny shows that C. cinereus and P. gramnis are more closely related, which conflicts with the placement of C. cinereus and C. neoformans as sister taxa, but with very low support value (bootstrapping score 35) [94]. The phylogenetic placement of C. merolae, a primitive red algae, has been the topic of a long-running debate [59]. Our result suggests that C. merolae is closer to Alveolata than to Viridiplantae, in agreement with a recent finding obtained by sequencing and comparing expressed sequence tags from different genomes [9].

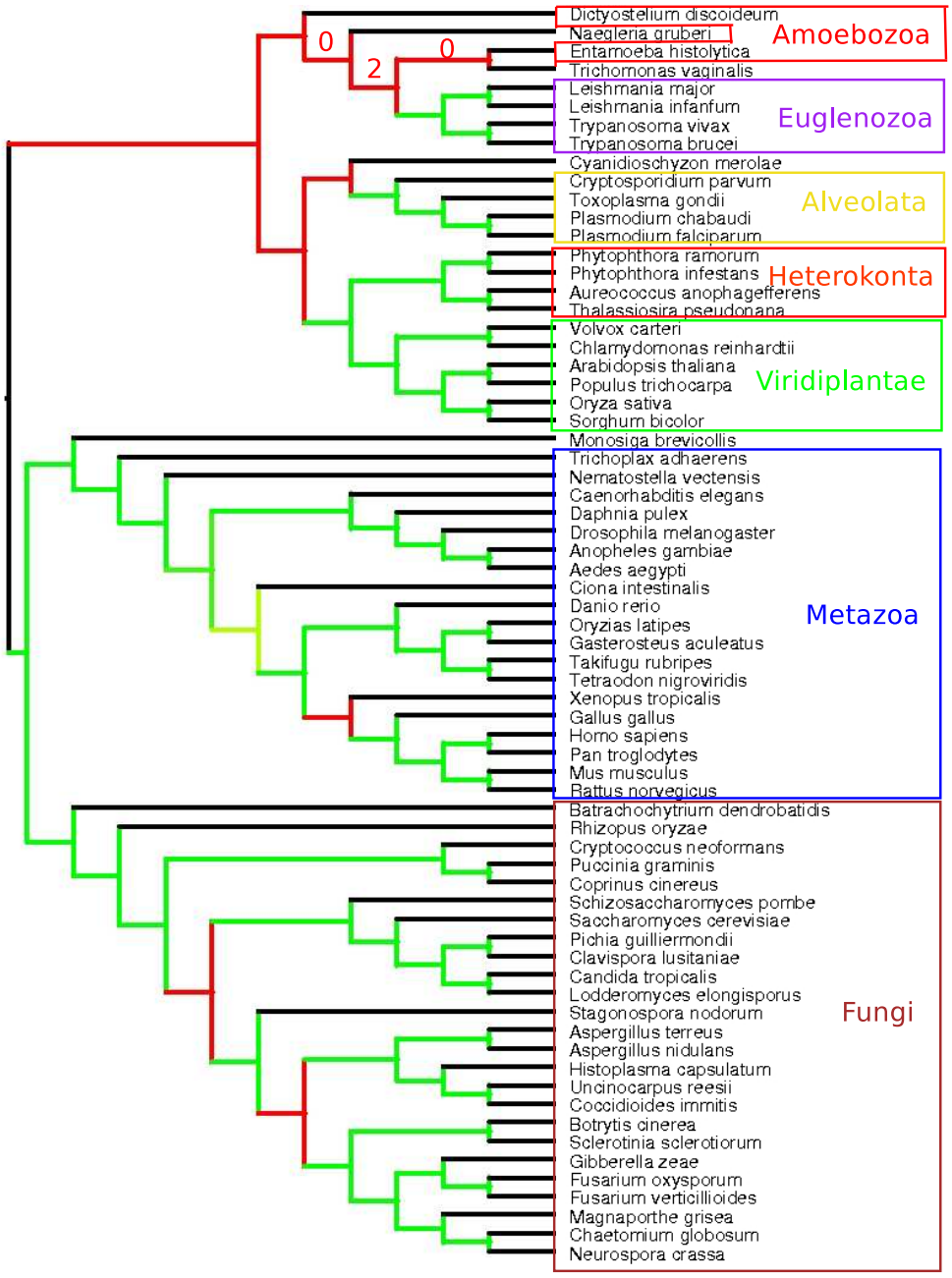


Fig. 6. The inferred phylogeny of 68 eukaryotic genomes, drawn with iTOL [40]. Internal branches are colored green, yellow, and red, to indicate, respectively, strong (bootstrap value > 90), medium (bootstrap value between 60 and 90), and weak support (bootstrap value < 60).

This approach opens the way to widespread use of whole-genome data in phylogenetic analysis, as it uses a fairly general model of genomic evolution (rearrangements plus duplications, insertions, and losses of genomic regions), is very accurate, scales as well as sequence-based approaches, and, importantly, supports standard bootstrapping methods. In addition, the nature of the encoding makes it robust against typical errors in genome assembly or in the identification of genes or syntenic blocks, as a few erroneous entries in a sequence of some hundreds of thousands of characters have little impact on the outcome. Moreover, the encoding can be modified to increase robustness by coding for proximity rather than just adjacency; such an encoding could use degrees of proximity to maintain discrimination among local rearrangements or deliberately treat all neighbors in the same manner to create invariance with respect to local rearrangements—a useful property when dealing with bacterial operons.

6 Conclusions

The shifting emphasis on simple comparisons, model-based distance computations, and encoding of features into sequences reflects both our increased understanding of rearrangements and duplication in genomes and the well known superiority (under most circumstances) of likelihood-based approaches in phylogenetic inference. Starting with very simple measures (the number of breakpoints) and with attempts at encoding the genomic structure into sequence data (in both cases because anything else remained unsolved), we have moved to computing model-based edit distances, then to estimate model-based true distances, then to use these as tools in median heuristics for a parsimonious approach. Most recently, we have returned to encodings, not for lack of alternatives, but because our deeper understanding of duplications and rearrangements in terms of adjacencies has led us to such a step. Yet the encoding is at least partly motivated by the impossibility, at this time, of using a direct approach to ML inference, of the style used for sequence data: a direct approach would require some parameterized model of genomic evolution under rearrangements and duplications and all models to date are both overly simplistic in terms of biology and far too complex for a Bayesian or ML inference strategy. (Even were such a model to emerge, a direct approach would remain a formidable algorithmic challenge, because of the lack of bounding methods for partial trees.) Thus the encoding of Y. Lin *et al.* is to evolutionary genomics much what binary character encoding is to morphological evolution: a way to take very rich and complex data produced through poorly understood events and to reduce them to a simple formulation that can be handled with today's phylogenetic inference tools.

From completing with some difficulty the inference of a phylogeny for fewer than 10 species with mitochondrial data featuring fewer than 50 common, single-copy genes using a rather inaccurate heuristic for maximum parsimony (the original `BPAnalysis`), we now have moved, thanks to this latest encoding approach, to easy handling of datasets of hundreds of species with tens of thousands of genes, many of them duplicated or missing in many of the species, using standard tools from sequence-based analysis. Yet it is clearly not the final word on phylogenetic reconstruction from rearrangement data: this area of research is little more than 15 years old and sufficient data to support it have been available for less than half that time. Challenges range from mod-

elling and algorithmic questions to implementation and assessment [54]. As new data accumulate at a frenetic pace and our understanding of the genome deepens with the daily additions to the research literature, we expect further insights, better models, refined methodology, and some breakthroughs—and, as has now been the case for nearly 30 years, these next developments are likely to be inspired by some past or forthcoming publication or remark of David Sankoff's.

References

1. Bader, D., Moret, B., Yan, M.: A fast linear-time algorithm for inversion distance with an experimental comparison. J. Comput. Biol. 8(5), 483–491 (2001)
2. Bader, M.: Genome rearrangements with duplications. BMC Bioinformatics 11(Supple 1), S27 (2010)
3. Bergeron, A., Mixtacki, J., Stoye, J.: On sorting by translocations. In: Proc. 9th Ann. Int'l Conf. on Research in Computational Molecular Biology (RECOMB'05). Lecture Notes in Comp. Sci., vol. 3500, pp. 615–629 (2005)
4. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Proc. 6th Workshop Algs. in Bioinf. (WABI'06). Lecture Notes in Comp. Sci., vol. 4175, pp. 163–173. Springer Verlag, Berlin (2006)
5. Bergeron, A., Mixtacki, J., Stoye, J.: A new linear time algorithm to compute the genomic distance via the double cut and join distance. Theor. Computer Science 410(51), 5300–5316 (2009)
6. Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. In: Miyano, S., Takagi, T. (eds.) Genome Informatics, pp. 25–34. Univ. Academy Press, Tokyo (1997)
7. Bourque, G., Pevzner, P.: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. 12, 26–36 (2002)
8. Bulteau, L., Fertin, G., Rusu, I.: Sorting by transpositions is difficult. In: Proc. 38th Int'l Colloq. on Automata, Languages, and Programming (ICALP 2011). Lecture Notes in Comp. Sci., vol. 6756. Springer Verlag, Berlin (2011)
9. Burki, F., et al.: Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE 2(8), e790 (2007)
10. Caprara, A.: Formulations and hardness of multiple sorting by reversals. In: Proc. 3rd Int'l Conf. Comput. Mol. Biol. (RECOMB'99). pp. 84–93. ACM Press, New York (1999)
11. Caprara, A.: On the practical solution of the reversal median problem. In: Proc. 1st Workshop Algs. in Bioinf. (WABI'01). Lecture Notes in Comp. Sci., vol. 2149, pp. 238–251. Springer Verlag, Berlin (2001)
12. Chen, F., Mackey, A., Vermunt, J., Roos, D.: Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2(4), e383 (2007)
13. Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S., Jiang, T.: Computing the assignment of orthologous genes via genome rearrangement. In: Proc. 3rd Asia Pacific Bioinf. Conf. (APBC'05). pp. 363–378. Imperial College Press, London (2005)
14. Compeau, P.: A simplified view of DCJ-Indel distance. In: Proc. 12th Workshop Algs. in Bioinf. (WABI'12). Lecture Notes in Comp. Sci., vol. 7534, pp. 365–377. Springer Verlag, Berlin (2012)
15. Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T., Wyman, S.: An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In: Sankoff, D., Nadeau, J. (eds.) Comparative Genomics, pp. 99–122. Kluwer Academic Publishers, Dordrecht, NL (2000)

16. Cosner, M., Jansen, R., Moret, B., Raubeson, L., Wang, L., Warnow, T., Wyman, S.: 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 (2000)
17. Day, W., Sankoff, D.: The computational complexity of inferring phylogenies from chromosome inversion data. J. Theor. Biol. 127, 213–218 (1987)
18. Demongeot, J., et al.: Hot spots in chromosomal breakage: From description to etiology. In: Sankoff, D., Nadeau, J. (eds.) Comparative Genomics, Computational Biology, vol. 1, pp. 71–83. Springer Verlag, Berlin (2000)
19. Desper, R., Gascuel, O.: Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9(5), 687–705 (2002)
20. Desper, R., Gascuel, O.: Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21(3), 587–598 (2003)
21. Dobzhansky, T., Sturtevant, A.: Inversions in the chromosomes of Drosophila pseudoobscura. Genetics 23(1), 28–64 (1938)
22. Downie, S.R., Palmer, J.D.: Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis, D., Soltis, P., Doyle, J. (eds.) Molecular Systematics of Plants, pp. 14–35. Chapman and Hall, New York (1992)
23. Durkin, K., et al.: Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature 482(7383), 81–84 (2012)
24. Ehrlich, J., Sankoff, D., Nadeau, J.: Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics 147, 289–296 (1997)
25. El-Mabrouk, N.: Genome rearrangement by reversals and insertions/deletions of contiguous segments. In: Proc. 11th Ann. Symp. Combin. Pattern Matching (CPM'00). Lecture Notes in Comp. Sci., vol. 1848, pp. 222–234. Springer Verlag, Berlin (2000)
26. El-Mabrouk, N., Bryant, D., Sankoff, D.: Reconstructing the pre-doubling genome. In: Proc. 3rd Int'l Conf. Comput. Mol. Biol. (RECOMB'99). pp. 154–163. ACM Press, New York (1999)
27. El-Mabrouk, N., Nadeau, J., Sankoff, D.: Genome halving. In: Proc. 9th Ann. Symp. Combin. Pattern Matching (CPM'98). pp. 235–250. Lecture Notes in Comp. Sci., Springer Verlag, Berlin (1998)
28. El-Mabrouk, N., Sankoff, D.: The reconstruction of doubled genomes. SIAM J. Computing 32(3), 754–792 (2003)
29. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: Combinatorics of Genome Rearrangements. MIT Press (2009)
30. Fitzpatrick, D., Logue, M., Stajich, J., Butler, G.: A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology 6(1), 99 (2006)
31. Fujimura, K., Conte, M., Kocher, T.: Circular DNA intermediate in the duplication of Nile Tilapia vasa genes. PLoS ONE 6(12 (e29477)) (2011)
32. Hannenhalli, S., Pevzner, P.: Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In: Proc. 27th Ann. ACM Symp. Theory of Comput. (STOC'95). pp. 178–189. ACM Press, New York (1995)
33. Hannenhalli, S., Pevzner, P.: Transforming mice into men (polynomial algorithm for genomic distance problems). In: Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95). pp. 581–592. IEEE Press, Piscataway, NJ (1995)
34. Hilker, R., Sickinger, C., Pedersen, C., Stoye, J.: UniMoG—a unifying framework for genomic distance calculation and sorting based on DCJ. Bioinformatics 28(19), 2509–2511 (2012)

35. Hu, F., Gao, N., Zhang, M., Tang, J.: Maximum likelihood phylogenetic reconstruction using gene order encodings. In: Proc. 2011 IEEE Symp. Comput. Intell. in Bioinf. & Comput. Biol. (CIBCB'11). pp. 117–122. IEEE Press, Piscataway, NJ (2011)
36. Huson, D., Nettles, S., Warnow, T.: Disk-covering, a fast converging method for phylogenetic tree reconstruction. J. Comput. Biol. 6(3), 369–386 (1999)
37. Jansen, R., Palmer, J.: A chloroplast DNA inversion marks an ancient evolutionary split in the sunflower family (Asteraceae). Proc. Nat'l Acad. Sci., USA 84, 5818–5822 (1987)
38. Jean, G., Nikolski, M.: Genome rearrangements: a correct algorithm for optimal capping. Inf. Process. Lett. 104(1), 14–20 (2007)
39. Larget, B., Simon, D., Kadane, J.: Bayesian phylogenetic inference from animal mitochondrial genome arrangements. J. Royal Stat. Soc. B 64(4), 681–694 (2002)
40. Letunic, I., Bork, P.: Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39(S2), W475–W478 (2011)
41. Lin, Y., Hu, F., Tang, J., Moret, B.: Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: Proc. 18th Pacific Symp. on Biocomputing (PSB'13). pp. 285–296. World Scientific Pub. (2013)
42. Lin, Y., Moret, B.: Estimating true evolutionary distances under the DCJ model. In: Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08). Bioinformatics, vol. 24(13), pp. i114–i122 (2008)
43. Lin, Y., Moret, B.: A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes. J. Comput. Biol. 18(9), 1055–1064 (2011)
44. Lin, Y., Rajan, V., Moret, B.: Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. J. Comput. Biol. 18(9), 1130–1139 (2011)
45. Lin, Y., Rajan, V., Swenson, K., Moret, B.: Estimating true evolutionary distances under rearrangements, duplications, and losses. In: Proc. 8th Asia Pacific Bioinf. Conf. (APBC'10). BMC Bioinformatics, vol. 11 (Suppl. 1):S54 (2010)
46. López, M., Samuelsson, T.: eGOB: Eukaryotic Gene Order Browser. Bioinformatics (2011)
47. Ma, J., Ratan, A., Raney, B., Suh, B., Miller, W., Haussler, D.: The infinite sites model of genome evolution. Proc. Nat'l Acad. Sci., USA 105(38), 14254–14261 (2008)
48. Marron, M., Swenson, K., Moret, B.: Genomic distances under deletions and insertions. Theor. Computer Science 325(3), 347–360 (2004)
49. Moret, B., Bader, D., Warnow, T.: High-performance algorithm engineering for computational phylogenetics. J. Supercomputing 22, 99–111 (2002)
50. Moret, B., Siepel, A., Tang, J., Liu, T.: Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Proc. 2nd Workshop Algs. in Bioinf. (WABI'02). Lecture Notes in Comp. Sci., vol. 2452, pp. 521–536. Springer Verlag, Berlin (2002)
51. Moret, B., Tang, J., Wang, L.S., Warnow, T.: Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci. 65(3), 508–525 (2002)
52. Moret, B., Tang, J., Warnow, T.: Reconstructing phylogenies from gene-content and gene-order data. In: Gascuel, O. (ed.) Mathematics of Evolution and Phylogeny, pp. 321–352. Oxford Univ. Press, UK (2005)
53. Moret, B., Wang, L.S., Warnow, T., Wyman, S.: New approaches for reconstructing phylogenies from gene-order data. In: Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01). Bioinformatics, vol. 17, pp. S165–S173 (2001)
54. Moret, B., Warnow, T.: Reconstructing optimal phylogenetic trees: A challenge in experimental algorithmics. In: Fleischer, R., Moret, B., Schmidt, E. (eds.) Experimental Algorithmics, Lecture Notes in Comp. Sci., vol. 2547, pp. 163–180. Springer Verlag, Berlin (2002)

55. Moret, B., Warnow, T.: Advances in phylogeny reconstruction from gene order and content data. In: Zimmer, E., Roalson, E. (eds.) Molecular Evolution: Producing the Biochemical Data, Part B, Methods in Enzymology, vol. 395, pp. 673–700. Elsevier (2005)
56. Moret, B., Wyman, S., Bader, D., Warnow, T., Yan, M.: A new implementation and detailed study of breakpoint analysis. In: Proc. 6th Pacific Symp. on Biocomputing (PSB'01). pp. 583–594. World Scientific Pub. (2001)
57. Nadeau, J., Taylor, B.: Lengths of chromosome segments conserved since divergence of man and mouse. Proc. Nat'l Acad. Sci., USA 81, 814–818 (1984)
58. Negrisolo, E., Kuhl, H., Forcato, C., Vitulo, N., Reinhardt, R., Patarnello, T., Bargelloni, L.: Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. Mol. Biol. Evol. 27(12), 2757–2774 (2010)
59. Nozaki, H., et al.: The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. J. Mol. Evol. 56(4), 485–497 (2003)
60. Ouangraoua, A., Boyer, F., McPherson, A., Tannier, E., Chauve, C.: Prediction of contiguous regions in the amniote ancestral genome. In: Proc. 5th Int'l Symp. Bioinformatics Research & Appls. (ISBRA'09). Lecture Notes in Comp. Sci., vol. 5542, pp. 173–185. Springer Verlag, Berlin (2009)
61. Palmer, J.D., Herbon, L.A.: Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. 27, 87–97 (1988)
62. Palmer, J.: Chloroplast and mitochondrial genome evolution in land plants. In: Herrmann, R. (ed.) Cell Organelles, pp. 99–133. Springer Verlag (1992)
63. Pe'er, I., Shamir, R.: The median problems for breakpoints are NP-complete. Elec. Colloq. on Comput. Complexity 71 (1998)
64. Peng, Q., Pevzner, P., Tesler, G.: The fragile breakage versus random breakage models of chromosome evolution. PLoS Comput. Biol. 2(2 (e14)) (2006)
65. Pevzner, P., Tesler, G.: Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc. Nat'l Acad. Sci., USA 100(13), 7672–7677 (2003)
66. Ponting, C.: The functional repertoires of metazoan genomes. Nat. Rev. Genet. 9(9), 689–698 (2008)
67. Price, M., Dehal, P., Arkin, A.: Fasttree: Computing large minimum-evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26, 1641–1650 (2009)
68. Rajan, V., Lin, Y., Moret, B.: TIBA: A tool for phylogeny inference from rearrangement data with bootstrap analysis. Bioinformatics 28(24), 3324–3325 (2012)
69. Rajan, V., Xu, A., Lin, Y., Swenson, K., Moret, B.: Heuristics for the inversion median problem. In: Proc. 8th Asia Pacific Bioinf. Conf. (APBC'10). BMC Bioinformatics, vol. 11 (Suppl. 1):S30 (2010)
70. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425 (1987)
71. Sankoff, D.: Minimal mutation trees of sequences. SIAM J. Applied Math. 28(1), 35–42 (1975)
72. Sankoff, D.: Edit distance for genome comparison based on non-local operations. In: Proc. 3rd Ann. Symp. Combin. Pattern Matching (CPM'92). Lecture Notes in Comp. Sci., vol. 644, pp. 121–135. Springer Verlag, Berlin (1992)
73. Sankoff, D., Blanchette, M.: The median problem for breakpoints in comparative genomics. In: Proc. 3rd Conf. Computing and Combinatorics (COCOON'97). Lecture Notes in Comp. Sci., vol. 1276, pp. 251–264. Springer Verlag, Berlin (1997)
74. Sankoff, D., Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. J. Comput. Biol. 5, 555–570 (1998)

75. Sankoff, D., Blanchette, M.: Phylogenetic invariants for metazoan mitochondrial genome evolution. In: Miyano, S., Takagi, T. (eds.) Genome Informatics, pp. 22–31. Univ. Academy Press, Tokyo (1998)
76. Sankoff, D., Cedergren, R., Abel, Y.: Genomic divergence through gene rearrangement. In: Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, Methods in Enzymology, vol. 183, pp. 428–438. Academic Press (1990)
77. Sankoff, D., Goldstein, M.: Probabilistic models for genome shuffling. Bull. Math. Biol. 51, 117–124 (1989)
78. Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B., Cedergren, R.: Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc. Nat'l Acad. Sci., USA 89(14), 6575–6579 (1992)
79. 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 Publishers, Dordrecht, NL (2000)
80. Sankoff, D., Nadeau, J.: Conserved synteny as a measure of genomic distance. Disc. Appl. Math. 71(1–3), 247–257 (1996)
81. Sankoff, D., Trinh, P.: Chromosomal breakpoint re-use in genome sequence rearrangement. In: Proc. 9th Int'l Conf. Comput. Mol. Biol. (RECOMB'05). Lecture Notes in Comp. Sci., vol. 3388, pp. 30–35. Springer Verlag, Berlin (2005)
82. Shao, M., Lin, Y.: Approximating the edit distance for genomes with duplicate genes under DCJ, insertion and deletion. BMC Bioinformatics 13(Suppl 19), S13 (2012)
83. Siepel, A., Moret, B.: Finding an optimal inversion median: Experimental results. In: Proc. 1st Workshop Algs. in Bioinf. (WABI'01). Lecture Notes in Comp. Sci., vol. 2149, pp. 189–203. Springer Verlag, Berlin (2001)
84. da Silva, P.H., Braga, M.D.V., Machado, R., Dantas, S.: DCJ-indel distance with distinct operation costs. In: Proc. 12th Workshop Algs. in Bioinf. (WABI'12). Lecture Notes in Comp. Sci., vol. 7534, pp. 378–390. Springer Verlag, Berlin (2012)
85. Srivastava, M., *et al.*: The functional repertoires of metazoan genomes. Nature 454(7207), 955–960 (2008)
86. Stamatakis, A.: RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21), 2688–2690 (2006)
87. Sturtevant, A., Dobzhansky, T.: Inversions in the third chromosome of wild races of Drosophila pseudoobscura and their use in the study of the history of the species. Proc. Nat'l Acad. Sci., USA 22, 448–450 (1936)
88. Swenson, K., Lin, Y., Rajan, V., Moret, B.: Hurdles and sorting by inversions: combinatorial, statistical, and experimental results. J. Comput. Biology 16(10), 1339–1351 (2009)
89. Swenson, K., Marron, M., Earnest-DeYoung, J., Moret, B.: Approximating the true evolutionary distance between two genomes. ACM J. Experimental Algorithmics 12 (2008)
90. Tang, J., Moret, B.: Scaling up accurate phylogenetic reconstruction from gene-order data. In: Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03). Bioinformatics, vol. 19, pp. i305–i312. Oxford U. Press (2003)
91. Tang, J., Moret, B.: Linear programming for phylogenetic reconstruction based on gene rearrangements. In: Proc. 16th Ann. Symp. Combin. Pattern Matching (CPM'05). Lecture Notes in Comp. Sci., vol. 3537, pp. 406–416. Springer Verlag, Berlin (2005)
92. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal genome median and halving problems. In: Proc. 8th Workshop Algs. in Bioinf. (WABI'08). Lecture Notes in Comp. Sci., vol. 5251, pp. 1–13. Springer Verlag, Berlin (2008)
93. Tesler, G.: Efficient algorithms for multichromosomal genome rearrangements. J. Comput. Syst. Sci. 63(5), 587–609 (2002)
94. Wang, H., Xu, Z., Gao, L., Hao, B.: A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evolutionary Biology 9(1), 195 (2009)

95. Wang, L.S.: Exact-IEBP: a new technique for estimating evolutionary distances between whole genomes. In: Proc. 1st Workshop Algs. in Bioinf. (WABI'01). Lecture Notes in Comp. Sci., vol. 2149, pp. 175–188. Springer Verlag, Berlin (2001)
96. Wang, L.S., Jansen, R., Moret, B., Raubeson, L., Warnow, T.: Fast phylogenetic methods for genome rearrangement evolution: An empirical study. In: Proc. 7th Pacific Symp. on Biocomputing (PSB'02). pp. 524–535. World Scientific Pub. (2002)
97. Wang, L.S., Leebens-Mack, J., Wall, P., Beckmann, K., dePamphilis, C., Warnow, T.: The impact of multiple protein sequence alignment on phylogenetic estimation. ACM/IEEE Trans. on Comput. Bio. & Bioinf. 8, 1108–1119 (2011)
98. Wang, L.S., Warnow, T.: Estimating true evolutionary distances between genomes. In: Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01). pp. 637–646. ACM Press, New York (2001)
99. Wang, L.S., Warnow, T.: Distance-based genome rearrangement phylogeny. In: Gascuel, O. (ed.) Mathematics of Evolution and Phylogeny, pp. 353–383. Oxford Univ. Press, UK (2005)
100. Watterson, G., Ewens, W., Hall, T., Morgan, A.: The chromosome inversion problem. J. Theor. Biol. 99(1), 1–7 (1982)
101. Xu, A.: A fast and exact algorithm for the median of three problem—a graph decomposition approach. J. Comput. Biol. 16(10), 1369–1381 (2009)
102. Xu, A., Moret, B.: GASTS: parsimony scoring under rearrangements. In: Proc. 11th Workshop Algs. in Bioinf. (WABI'11). Lecture Notes in Comp. Sci., vol. 6833, pp. 351–363. Springer Verlag, Berlin (2011)
103. Xu, A., Sankoff, D.: Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In: Proc. 8th Workshop Algs. in Bioinf. (WABI'08). Lecture Notes in Comp. Sci., vol. 5251, pp. 25–37. Springer Verlag, Berlin (2008)
104. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16), 3340–3346 (2005)
105. Yancopoulos, S., Friedberg, R.: Sorting genomes with insertions, deletions and duplications by DCJ. In: Proc. 6th RECOMB Workshop Comp. Genomics (RECOMB-CG'08). Lecture Notes in Comp. Sci., vol. 5267, pp. 170–183. Springer Verlag, Berlin (2008)
106. Zhang, M., Arndt, W., Tang, J.: A branch-and-bound method for the multichromosomal reversal median problem. In: Proc. 8th Workshop Algs. in Bioinf. (WABI'08). pp. 1–13 (2008)
107. Zheng, C., Zhu, Q., Adam, Z., Sankoff, D.: Guided genome halving: hardness, heuristics and the history of the hemiascomycetes. In: Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08). Bioinformatics, vol. 24(13), pp. i96–i104 (2008)