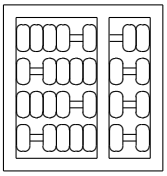

INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

A Survey on Genome Rearrangement Problems and Gene Order Based Phylogenies

Pedro C. Feijão *João Meidanis*

Technical Report - IC-08-033 - Relatório Técnico

December - 2008 - Dezembro

A Survey on Genome Rearrangement Problems and Gene Order Based Phylogenies

Pedro Cipriano Feijão João Meidanis*

Abstract

With the growing availability of complete genome sequence data, the genome rearrangement problem has received a lot of attention in the field of computational biology. In this problem, a genome is modeled as a sequence of regions that are conserved within a genome group. The goal is to find a plausible evolution scenario for this genome group, considering the position and orientation of the conserved regions, building a phylogenetic tree using pairwise distance estimates, optionally also rebuilding the ancestral genomes. This methodology has shown encouraging results when applied to the phylogenetic inference of several species groups. In this article, we review the state-of-the-art and applications in this field.

1 Introduction

Most of molecular phylogeny research is based upon the analysis of DNA and aminoacids sequences of individual genes or groups of genes. However, with the advent of completely sequenced genomes, these approaches are complemented by genome-wide comparisons.

The reconstruction of evolutionary scenarios using gene order information is a powerful tool and has been used to gain a better understanding of the evolution of several groups of species, such as mammals [12, 13], yeast [24], microbial genomes [20] and the entire metazoan kingdom [25].

In this survey, we will review recent advances on genome rearrangement problems and gene order based phylogenies. In Section 2 we describe recent genome rearrangement methods for the pairwise parsimonious distance calculation. Section 3 introduces the application of statistical models to find better estimates of such distances, and in Section 4 we talk about phylogenetic inference based on genome rearrangement events.

2 Pairwise Genome Rearrangement

The inference of evolutionary scenarios based on gene order started with the study of the *genome rearrangement problem* [43]: given two genomes, represented as sequences of conserved segments called *syntenic blocks*, find the most parsimonious sequence of rearrangement events that converts one genome into the other. In this sense, the distance between these two genomes is the sum of the weights of each these events.

*Institute of Computing, University of Campinas, 13081-970 Campinas, SP.

The initial approaches considered the case where the orientation of all genes is known and both genomes share the same gene set, without gene duplication. In this case, genomes can be represented as signed permutations over a set of integers, where each integer denotes a gene and its sign denotes the orientation.

We are interested in the most parsimonious rearrangement scenario, which uses the fewest rearrangement operations to convert one permutation into another. Finding this scenario can be a hard problem, depending on the allowed rearrangement operations. The most commonly used operations and proposed algorithms will be discussed throughout this section.

The first approaches considered a simpler case, where just one operation is allowed. In this type of problem, the rearrangement distance is defined as the minimum number of events needed to transform one permutation into another. Some algorithms find the distance only, while others also find the sequence of rearrangement events. Since transforming one permutation into another is usually equivalent to transforming one permutation into the identity, finding the minimum sequence of events is called *sorting*.

The most often used operations are reversals. *Unsigned reversal* is an operation that inverts the order of a block of adjacent genes, and *signed reversal* also inverts the sign of each gene in the block. The unsigned reversal problem has been shown to be NP-hard by Caprara [14]. Kececioglu and Sankoff [30] developed 2-approximation algorithms, which Bafna and Pevzner [3] improved to a factor of 7/4 approximation and Berman et al. [9] reduced the factor even further, to 1.375.

The signed reversal has drawn much more attention, since it seems to be more biologically relevant and fast algorithms are currently available. We will use the term "reversal" to refer to the signed case. Kececioglu and Sankoff [32] formulated the reversal sorting problem and derived lower and upper bounds for the reversal distance between two genomes. They developed an $O(n^2)$ approximation algorithm and a *branch-and-bound* exact algorithm. Later, Hannenhalli and Pevzner [27] solved the reversal distance problem in $O(n^4)$, using the *breakpoint graph*, a structure introduced by Bafna and Pevzner [2] and largely used on several sorting algorithms. It was the first time a polynomial algorithm was developed for a realistic model of genome rearrangement. Several faster implementations followed [8, 29] and the reversal distance can now be calculated in linear time [1], but the best *sorting* algorithm — which also finds the optimal sequence of reversals — runs in $O(n^{3/2}\sqrt{\log n})$ [60].

These algorithms ignored the lenght of the reversal, only counting their number. Pinter and Skiena [51] introduced a model where the lenght of the reversals plays a role. They derived a upper bound of $O(f(n)\lg^2 n)$ on the cost of sorting any n-element permutation, where $f(l)$ is a additive monotonic function representing the cost of an unsigned reversal of size l. They also developed a approximation algorithm with ratio $O(\lg^2 n)$ for sorting a given permutation.

Another largely studied rearrangement operator is the *transposition*, where two consecutive adjacent blocks are exchanged, without changing gene orientation. Since the work of Bafna and Pevzner [4], it was studied on several papers. The complexity of transposition sorting is still an open problem, and the current approaches strive to find the best approximation algorithms. The best algorithm so far is by Elias and Hartman [22], witch runs in $O(n^2)$ and has an approximation factor of 1.375. Dias and Meidanis [18] worked with a new

Table 1: Rearrangement Problem Complexity with Different Operations

Rearrangement Operation	Complexity	Algorithm Type	Authors
Unsigned Reversal	NP-hard	1.375-Approx.	Berman et al. [9]
Reversal	$O(n)$	Exact	Bader et al. [1]
Transposition	Open	1.375-Approx.	Elias and Hartman [22]
Prefix Transposition	Open	2-Approx.	Dias and Meidanis [18]
Block-Interchange	$O(n^2)$	Exact	Christie [16]
Block-Interchange and Reversal	$O(n^2)$	Exact	Mira and Meidanis [45] Lin et al. [38]
Translocation	$O(n)$	Exact	Bergeron et al. [6]
Fission, Fusion, Transposition	$O(n^2)$	Exact	Dias and Meidanis [17]
Fission, Fusion, Block-Interchange	$O(n^2)$	Exact	Lu et al. [40]
Double-Cut-and-Join	$O(n)$	Exact	Yancopoulos et al. [62]

related operator, the *prefix transposition*, where the first gene of the permutation is always moved. In their work, some bounds are given for the prefix transposition distance, as well as approximation algorithms with factors 2 and 3.

Another related operation is the *block-interchange*, which is a generalization of the transposition, where the exchanged blocks may not be consecutive. It was introduced by Christie [16], and the block-interchange sorting problem can be solved in $O(n^2)$. More precisely, the algorithm complexity is $O(\delta n)$, where δ is the block-interchange distance, which can be calculated in $O(n)$ in advance, as shown by Lin et al. [37]. On recent works, the block-interchange operation was used along with signed reversal [45, 38], and linear time algorithms are available when appropriate weights are used.

When multi-chromosomal genomes are considered, additional rearrangement operations are needed. The first studied multi-chromosomal operation was the *translocation*, defined by Kececioglu and Ravi [31] and Hannenhalli [26], where two chromosomes exchange blocks at their ends. Recently, Bergeron and colleagues [6] found an error in Hannenhalli's linear algorithm. In the same article, they developed a new theory and showed a linear algorithm for the translocation sorting problem.

Using a completely different approach, Ozery-Flato and Shamir [50] showed that the problem of sorting by transpositions has many similarities with sorting by reversals. They were able to transform both Bergeron's [5] score-based algorithm and Berman-Hannenhalli [8] algorithm into transposition sorting algorithms, mantaining the original complexity.

Other multi-chromosomal operations are *fissions* and *fusions*. In the fission operation, a chromosome is cut between two genes, forming two new chromosomes. The fusion is the inverse operation, joining two different chromosomes into one. Algorithms that use these operations usually include a third operation to move genes within the same chromosome. Dias and Meidanis [17] solved the problem using fissions, fusions and transpositions in $O(n^2)$. It was the first polynomial algorithm for a rearrangement problem involving transpositions. Using fissions, fusions and block-interchanges, Lu et al. [40] developed an $O(n^2)$ algorithm.

Instead of using specific rearrangement operations, some recent approaches focus on

finding a general rearrangement operation that can be used to model several different operations. Yancopoulos and colleagues [62] introduced a new concept of a unifying operation, called *double-cut-and-join* (DCJ), which is applied directly on the breakpoint graph [2]. It consists of cutting two edges (breakpoints) in the graph,and rejoining the resulting four unconnected vertices in two new pairs. Depending on the selected edges and the possible ways of rejoining the vertices, this same operation models reversals, translocations, fissions and fusions. They also developed an $O(n)$ sorting algorithm using the DCJ. Bergeron et al. [7] improved the DCJ theory introducing a structure called *adjacency graph*, a simple graph that is a union of paths and cycles and can be used to model genomes. Using this structure they developed an optimal greedy sorting algorithm that runs in $O(n)$.

A more theoretical approach is the one by Meidanis and Dias [42], where they use the theory of permutation groups to develop a rearrangement theory with a strong algebraic formalism. With this formalism many different operations can be easily modelled, without the need for graphs that are often used in rearrangement algorithms. Some of the papers already cited in this review applied this theory to develop their algorithms [17, 18, 45] and it seems that this is a promising research direction. The search for an unified operation can potentially simplify both rearrangement theory and algorithms.

An overview of the complexity of the rearrangement sorting problems considered so far is shown at Table 1, for each different set of operations, also citing the faster algorithms known to date.

To this point, we have reviewed genome rearrangement studies where the genomes have the same set of genes, without repetition. In practice, this is seldom the case. Gene duplication, deletion and horizontal transfers are abundant [36] and it is unlikely that genomes from different species satisfy this condition. One approach to this question is trying to find a common set of markers (genes, syntenic blocks), discarding duplicated genes and genes that are present in only one of the genomes [10, 12].

Another approach aims at understanding gene loss and gain mechanisms, trying to develop rearrangement models that allow these kind of operations, such as gene deletion and duplication. One of the first studies in this direction was done by El-Mabrouk [21], based on the reversal theory of Hannenhalli and Pevzner [27] with the addition of gene insertions/deletions of contiguous segments. The complexity of their algorithm is the same as Hannenhalli-Pevzner, which was improved to $O(n^2)$ in Kaplan, Shamir and Tarjan work [29].

Marron and Swenson [41] further developed El-Mabrouk's results allowing gene duplications. They also developed an approximation algorithm for the distance calculation under this model, and experimental results suggest that its performance is near optimal. On a follow-up work [57], some improvements were included, such as an estimate of the evolutionary distance between genomes, which is calculated applying an empirical correction factor on the rearrangement distance.

Another line of research focusing uniquely on gene gain and loss events is from Koonin and colleagues [33, 46], aiming to reconstruct a set of ancestor genes which is universal to all analyzed species. The hypothetical species comprising this set of genes was called *last universal common ancestor* (LUCA). Starting with a set of homolog genes, they showed that the phyletic patterns — presence-absence gene patterns on whole sequenced genomes — are not compatible with the hypothetical tree of the analyzed species. They developed

algorithms that search for the most parsimonious evolutionary scenario, minimizing gene loss and gain events on a given tree. These algorithms allow different penalties for gain and loss, as there is evidences showing that losses are more frequent. The greater the gene gain penalty, the more genes are present in the LUCA. After testing with different penalties the authors concluded that gene gain and loss may actually occur with the same frequency, and that horizontal gene transfer is an important factor in the evolution of prokaryotes.

3 Statistical estimates of evolutionary distance

A known issue on parsimonious distances, such as rearrangement distances, particularly in the context of phylogeny reconstructions, is the tendency of underestimating the evolutionary distance, notably with distant species. In the case of nucleotide mutations, for instance, since the Jukes-Cantor model of 1969, several statistical models where developed to better estimate the evolutionary distance, constructing more accurate phylogenies with the same data set [23, Chap. 11].

Unfortunately, as we saw in the previous section, the calculation of the parsimonious rearrangement distance is not as easy as the local mutation distance, giving rise to an NP-hard problem on several cases. Possibly because of this difficulty, statistical models of rearrangement evolution were not studied until recently. Such models would enable the inference of confidence intervals on distance estimations, as well as hypothesis testing about evolution mechanisms.

The first approaches focused on applying an empirical correction factor to the parsimonious rearrangement distance. One example is the EDE — *empirically derived estimator* — of Moret et al. [48], which is applied on the reversal distance. In a recent work, this estimator was compared to the uncorrected reversal distance when constructing phylogenies with distance-based methods, and generated trees with better quality.

Some recent approaches attempt to create statistical models for rearrangement evolution. Durret, York and colleagues worked on Bayesian models for rearrangement distance based on Markov chain Monte Carlo methods (MCMC). In a first attempt [63], the model calculated an estimate of the reversal distance. This model was improved in a subsequent article [19] to allow for multi-chromosomal genomes, adding the translocation operation. Another works in this line of research include the Bayesian model developed by Miklós [44] to estimate distance allowing the transposition operation, in addition to reversals, and the work of Larget and colleagues [35], where they made a bayesian analysis of metazoan mitochondrial genome arrangements.

Further enhancements were included in a recent article [64], where the size of the reversed block is considered to estimate the reversal probabilities. Also, pericentric – including the centromere – and paracentric – not including the centromere – reversals also have different probabilities, as it is believed that pericentric reversals are more frequent.

These contributions show that phylogeny reconstruction based on distance estimates perform better than parsimonious distances. Furthermore, the study of statistical models of rearrangement is of key importance to the understanding of large-scale evolutionary processes.

4 Rearrangement-based phylogeny

In Section 2 we reviewed several methods of calculating rearrangement distances between two genomes. In this section, we focus on the more general case: is it possible to calculate rearrangement distances between a group of genomes, inferring their evolutionary history, possibly reconstructing ancestral genomes?

Methods of phylogenetic tree reconstruction are divided in three classes [23]: i) distance-based, which use pairwise distance to establish evolutionary relations; ii) maximum parsimony, which looks for the tree with the minimum number of evolutionary events; iii) maximum likelihood, which reconstruct the most likely phylogenetic tree according to a well-defined probabilistic model of evolution.

Distance-based methods are computationally less demanding. After calculating a pairwise distance matrix of all genomes in the analyzed group, an algorithm such as Neighbor-Joining or Minimum Evolution is applied [23]. One of the first studies that built trees based on rearrangement distances used the *breakpoint* distance, defined as the number of genes that are adjacent on one genome but not on the other. Although very easy to calculate, breakpoint distance does not have a significant biological meaning. An alternative would be using more relevant distances, such as the reversal distance, but as we discussed, parsimonious distances may not be suitable for phylogeny reconstruction, given their tendency to underestimate the "true" distance. To support this claim, Wang et al. [61] tested breakpoint and reversal distances versus an empirical estimator, the EDE. The estimator recovered better phylogenies than both parsimonious distances.

Distance-based methods are fast and can build good trees, but cannot infer the ancestral genomes matching the internal nodes of the phylogenetic tree. The problem of reconstructing a tree with ancestral genomes with the minimum number of rearrangement events, creating a plausible evolution scenario is called *Multiple Genome Rearrangement Problem* (MGRP) [54].

The parsimonious approach has been used to solve the MGRP. Although there are instances where the pairwise rearrangement distance problem is easy, such as the reversal distance, this is not extended in a multiple genome context. Even the simplest case, with only three genomes, called *reversal median problem* (RMP) — finding the ancestral genome with minimal sum of reversal distances — is NP-hard [15].

Using a simpler distance measure for the median problem, the breakpoint distance, Sankoff and Blanchette proposed the first solution to the MGRP, with their software BP-Analysis [53]. After an initialization step, their algorithm iterates over a tree, repeatedly resetting the permutations of internal nodes to breakpoint medians of their three neighbors, until convergence is achieved. They discovered a reduction of the breakpoint median problem to the *Traveling Salesman Problem* (TSP) and were able to compute medians relatively efficiently.

Later, Moret and colleagues developed GRAPPA [48], based on BPAnalysis. Their algorithm runs much faster, with the use of better TSP bounds and parallelized algorithms. With the availability of a linear algorithm for reversal distance calculations [1], breakpoint distance was replaced by reversal distance [47], obtaining better results. Several improvements were integrated into GRAPPA in the last years: faster exact algorithms for the

RMP [55]; inclusion of the distance estimate EDE [49]; implementation of the *disk-covering method* [28, 52] for large phylogenetic tree reconstructions [58]; the use of Linear Programming to calculate better bounds and speed up the *branch-and-bound* [59]; quartet-based phylogeny reconstruction [39].

Another algorithm for the MGRP was developed by Bourque and Pevzner [11], the MGR-MEDIAN. In this algorithm, the RMP is not solved exactly. An heuristic searches for *good reversals* — reversals that bring one of the genomes closer to the other two — and applies them iteratively. When a good reversal is not found, a search to find the best reversal according to several tested criteria is performed. While there is no clear evidence of better results, MGR-MEDIAN is significantly faster than GRAPPA, because it does not solve the RMP exactly.

As we saw in Section 2, pairwise distance rearrangement algorithms allowing gene gain and loss events are quite recent, with the work of El-Mabrouk [21] being one of the first formal approaches. The same is true for phylogenetic reconstruction. Lake and Rivera [34] developed a method allowing horizontal gene transfer, using similar techniques as in nucleotide substitution models. Particularly, they used Markov chains with states A and P, representing absence or presence of a gene in a genome. They introduced the concept of *conditioning genome* (CG), a chosen set of genes. The CG is fundamental to the state attribution, for a gene is considered A (absent) in a certain genome only if it is not present in neither the genome nor the CG. In a recent work, Spencer et al. [56] evaluated this algorithm, finding good results. They were particularly interested in the choice of the conditioning genome and showed that, in some instances, different CGs lead to strong support for very different phylogenies. They proposed a supertree method to solve this inconsistency, combining information obtained with all the different CGs.

5 Conclusion

In this paper, we reviewed the state-of-the-art and recent applications in genome rearrangement problems. This methodology has shown encouraging results when applied to evolutionary distance estimation and the phylogenetic inference of several species groups, complementing other widely used approaches such as DNA sequence-based methods.

There are several promising research directions within this field. First, in pairwise parsimonious distance calculation, most of the developed theory and algorithms allow just one operation. We need better sorting algorithms that allow several rearrangement operations, possibly resulting in more significant distance measures. There are even fewer works that consider gene gain/loss operations, but we believe that algorithms that examine this type of operations would be extremely important.

All these ideas should also be extended to the MGRP, where there is the predominant use of signed reversals. Some authors already pointed out that MGRP algorithms that allow more operations can produce better evolutionary scenarios.

In parallel with these problems, advances in rearrangement theoretical background, such as the double-cut-and-join unified operation of Yancopoulos or the algebraic formalism of Meidanis and Dias can simplify both rearrangement theory and algorithms.

Finally, recent results in statistical models of rearrangement have shown that this is a promising area with vast room for exploration. Particularly, when distances are large, parsimonious distances are not well suited for phylogenetic reconstruction and better distance estimates are needed.

In summary, we expect to see more activity in the following areas:

- Modelss with multiple rearrangement operations
- Multiple genome rearrangement
- Unifying operations
- Statistical estimates of rearrangemente distance

References

[1] Bader, D. A., Moret, B. M., and 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.

[2] Bafna, V. and Pevzner, P. A. (1993). Genome rearrangements and sorting by reversals. In *FOCS*, pages 148–157. IEEE.

[3] Bafna, V. and Pevzner, P. A. (1996). Genome rearrangements and sorting by reversals. *SIAM J. Comput.*, 25(2):272–289.

[4] Bafna, V. and Pevzner, P. A. (1998). Sorting by transpositions. *SIAM J. Discrete Math.*, 11(2):224–240.

[5] Bergeron, A. (2005). A very elementary presentation of the hannenhalli-pevzner theory. *Discrete Applied Mathematics*, 146:134–145.

[6] Bergeron, A., Mixtacki, J., and Stoye, J. (2006a). On sorting by translocations. *J Comput Biol*, 13(2):567–578.

[7] Bergeron, A., Mixtacki, J., and Stoye, J. (2006b). A unified view of genome rearrangements. In *Algorithms in Bioinformatics*, volume 4175 of *Lecture Notes in Computer Science*, pages 163–173. Springer Berlin / Heidelberg.

[8] Berman, P. and Hannenhalli, S. (1996). Fast sorting by reversals. In *Combinatorial Pattern Matching, Proceedings of the 7th Annual Symposium (CPM'96)*, number 1075 in Lecture Notes in Computer Science, pages 168–185.

[9] Berman, P., Hannenhalli, S., and Karpinski, M. (2002). 1.375-approximation algorithm for sorting by reversals. 2461:200–210.

[10] Blair, J. E., Shah, P., and Hedges, S. B. (2005). Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics*, 6:53.

[11] Bourque, G. and Pevzner, P. A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res*, 12(1):26–36.

[12] Bourque, G., Pevzner, P. A., and Tesler, G. (2004). Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, 14(4):507–516.

[13] Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A., and Tesler, G. (2005). Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, 15(1):98–110.

[14] Caprara, A. (1997). Sorting by reversals is difficult. In *RECOMB '97: Proceedings of the first annual international conference on computational molecular biology*, pages 75–83, New York, NY, USA. ACM Press.

[15] Caprara, A. (1999). Formulations and hardness of multiple sorting by reversals. In *Proceedings of the third annual international conference on computational molecular biology*, pages 84–93. ACM Press.

[16] Christie, D. A. (1996). Sorting permutations by block-interchanges. *Information Processing Letters*, 60:165–169.

[17] Dias, Z. and Meidanis, J. (2001). Genome rearrangements distance by fusion, fission, and transposition is easy. In *Proceedings of the String Processing and Information Retrieval: 8th International Symposium, SPIRE 2001*, pages 250–253.

[18] Dias, Z. and Meidanis, J. (2002). Sorting by prefix transpositions. In *Proceedings of the String Processing and Information Retrieval: 9th International Symposium, SPIRE 2002*, volume 2476/2002 of *Lecture Notes in Computer Science*, pages 463–468. Springer Berlin / Heidelberg.

[19] Durrett, R., Nielsen, R., and York, T. L. (2004). Bayesian estimation of genomic distance. *Genetics*, 166(1):621–629.

[20] Earnest-DeYoung, J. V., Lerat, E., and Moret, B. M. (2004). Reversing gene erosion – reconstructing ancestral bacterial genomes from gene-content and order data. In *Algorithms in Bioinformatics, 4th International Workshop, WABI 2004*, volume 3240 of *Lecture Notes in Bioinformatics*, pages 26–37. Springer-Verlag.

[21] El-Mabrouk, N. (2001). Sorting signed permutations by reversals and insertions/deletions of contiguous segments. *Journal of Discrete Algorithms*, 1(1):105–122.

[22] Elias, I. and Hartman, T. (2006). A 1.375-approximation algorithm for sorting by transpositions. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(4):369–379.

[23] Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

[24] Fischer, G., Rocha, E. P. C., Brunet, F., Vergassola, M., and Dujon, B. (2006). Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. *PLoS Genet*, 2(3):e32.

[25] Fritzsch, G., Schlegel, M., and Stadler, P. F. (2006). Alignments of mitochondrial genome arrangements: applications to metazoan phylogeny. *J Theor Biol*, 240(4):511–520.

[26] Hannenhalli, S. (1996). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Appl. Math*, 71(1–3):137–151.

[27] Hannenhalli, S. and Pevzner, P. A. (1999). Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1):200–223.

[28] Huson, D. H., Nettles, S. M., and Warnow, T. J. (1999). Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J Comput Biol*, 6(3-4):369–386.

[29] Kaplan, H., Shamir, R., and Tarjan, R. E. (1999). Faster and simpler algorithm for sorting signed permutations by reversals. *Society for Industrial and Applied Mathematics, Jornal on Computing*, 29(3):880–892.

[30] Kececioglu, J. and Sankoff, D. (1993). Exact and approximation algorithms for the inversion distance between two chromosomes. In *Combinatorial Pattern Matching*, volume 684/1993, pages 87–105. Springer Berlin / Heidelberg.

[31] Kececioglu, J. D. and Ravi, R. (1995). Of mice and men: Algorithms for evolutionary distances between genomes with translocation. In *Proceedings of the sixth ACM-SIAM Symposium on Discrete Algorithm*, pages 604–613. Society of Industrial and Applied Mathematics.

[32] Kececioglu, J. D. and Sankoff, D. (1995). Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13(1/2):180–210.

[33] Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, 1(2):127–136.

[34] Lake, J. A. and Rivera, M. C. (2004). Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol Biol Evol*, 21(4):681–690.

[35] Larget, B., Simon, D. L., Kadane, J. B., and Sweet, D. (2005). A bayesian analysis of metazoan mitochondrial genome arrangements. *Mol Biol Evol*, 22(3):486–495.

[36] Lawrence, J. G. and Hendrickson, H. (2003). Lateral gene transfer: when will adolescence end? *Mol Microbiol*, 50(3):739–749.

[37] Lin, Y. C., Lu, C. L., Chang, H.-Y., and Tang, C. Y. (2005). An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *J Comput Biol*, 12(1):102–112.

[38] Lin, Y. C., Lu, C. L., Liu, Y.-C., and Tang, C. Y. (2006). Spring: a tool for the analysis of genome rearrangement using reversals and block-interchanges. *Nucleic Acids Res*, 34(Web Server issue):W696–W699.

[39] Liu, T., Tang, J., and Moret, B. M. E. (2005). Quartet-based phylogeny reconstruction from gene orders. In *Proceedings of the 11th Annual International Conference in Computing and Combinatorics (COCOON 05)*, pages 63–73.

[40] Lu, C. L., Huang, Y. L., Wang, T. C., and Chiu, H.-T. (2006). Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. *BMC Bioinformatics*, 7:295.

[41] Marron, M., Swenson, K. M., and Moret, B. M. E. (2004). Genomic distances under deletions and insertions. *Theor. Comput. Sci.*, 325(3):347–360.

[42] Meidanis, J. and Dias, Z. (2000). An alternative algebraic formalism for genome rearrangements. In Sankoff, D. and Nadeau, J., editors, *Comparative Genomics*, pages 213–223. Kluwer Academic Publishers.

[43] Meidanis, J. and Setubal, J. C. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing.

[44] Miklós, I. (2003). MCMC genome rearrangement. *Bioinformatics*, 19 Suppl 2:II130–II137.

[45] Mira, C. and Meidanis, J. (2007). Sorting by block-interchanges and signed reversals. In *Information Technology, 2007. ITNG '07. Fourth International Conference on*, pages 670–676.

[46] Mirkin, B. G., Fenner, T. I., Galperin, M. Y., and Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, 3:2.

[47] Moret, B. M., Siepel, A. C., Tang, J., and Liu1, T. (2002a). Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *Algorithms in Bioinformatics: Second International Workshop, WABI 2002, Rome, Italy, September 17-21, 2002. Proceedings*, volume 2452/2002 of *Lecture Notes in Computer Science*, pages 521–536. Springer Berlin / Heidelberg.

[48] Moret, B. M., Wang, L. S., Warnow, T., and Wyman, S. K. (2001). New approaches for reconstructing phylogenies from gene order data. *Bioinformatics*, 17 Suppl 1:S165–S173.

[49] Moret, B. M. E., Tang, J., Wang, L.-S., and Warnow, T. (2002b). Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525.

[50] Ozery-Flato, M. and Shamir, R. (2006). Sorting by translocations via reversals theory. In Bourque, G. and El-Mabrouk, N., editors, *Comparative Genomics*, volume 4205 of *Lecture Notes in Computer Science*, pages 87–98. Springer.

[51] Pinter, R. Y. and Skiena, S. (2002). Genomic sorting with length-weighted reversals. *Genome Inform*, 13:103–111.

[52] Roshan, U. W., Moret, B. M. E., Warnow, T., and Williams, T. L. (2004). Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proceedings of the IEEE Computational Systems Bioinformatics conference (CSB'04)*, pages 98–109, Los Alamitos, CA, USA. IEEE Computer Society.

[53] Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol*, 5(3):555–570.

[54] Sankoff, D., Sundaram, G., and Kececioglu, J. D. (1996). Steiner points in the space of genome rearrangements. *International Journal of Foundations of Computer Science*, 7(1):1–9.

[55] Siepel, A. (2001). Exact algorithms for the reversal median problem. Master's thesis, Cornell University.

[56] Spencer, M., Bryant, D., and Susko, E. (2007). Conditioned genome reconstruction: how to avoid choosing the conditioning genome. *Syst Biol*, 56(1):25–43.

[57] Swenson, K. M., Marron, M., Earnest-DeYoung, J. V., and Moret, B. M. (2005). Aproximating the true evolutionary distance between two genomes. In *7th Workshop on Algorithm Engineering and Experiments (ALENEX05)*, pages 37–46. SIAM Press.

[58] Tang, J. and Moret, B. M. E. (2003). Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, 19 Suppl 1:i305–i312.

[59] Tang, J. and Moret, B. M. E. (2005). Linear programming for phylogenetic reconstruction based on gene rearrangements. In *Proceedings of the 16th Annual Symposium on Combinatorial Pattern Matching (CPM 2005)*, pages 406–416.

[60] Tannier, E. and Sagot, M.-F. (2004). Sorting by reversals in subquadratic time. In *Combinatorial Pattern Matching*, volume 3109 of *Lecture Notes in Computer Science*, pages 1–13s. Springer Berlin / Heidelberg.

[61] Wang, L.-S., Warnow, T., Moret, B. M. E., Jansen, R. K., and Raubeson, L. A. (2006). Distance-based genome rearrangement phylogeny. *J Mol Evol*, 63(4):473–483.

[62] Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346.

[63] York, T. L., Durrett, R., and Nielsen, R. (2002). Bayesian estimation of the number of inversions in the history of two chromosomes. *J Comput Biol*, 9(6):805–818.

[64] York, T. L., Durrett, R., and Nielsen, R. (2007). Dependence of paracentric inversion rate on tract length. *BMC Bioinformatics*, 8:115.