Steiner points in the space of genome rearrangements*

David Sankoff† Gopalakrishnan Sundaram‡ John Kececioglu§

May 8, 1995; revised January 17, 1996

Abstract We present some experiences with the problem of multiple genome comparison, analogous to multiple sequence alignment in sequence comparison, under the inversion and transposition distance metrics, given a fixed phylogeny. We first describe a heuristic for the case in which phylogeny is a star on three vertices and then use this to approximate the multiple genome comparison problem via local search.

Keywords Permutation reversals, chromosome inversions, median problem, phylogeny

1 Introduction

Although the mathematical nature of the problems are very different, genome comparison has inherited much of the spirit of traditional research into sequence comparison. This paper explores the concept of multiple genome comparison, analogous to multiple sequence alignment in sequence comparison. And just as the original multiple alignment problem [17] was posed in terms of optimizing the internal nodes of a given phylogenetic tree by minimizing the sum of an edit distance over the branches of that tree, we will extend the notion of genome rearrangement distance to the optimal positioning of Steiner points in the appropriate space.[1]

The phylogenetic versions of the Steiner problem are generally decomposed into two subproblems, the first embedded in the second. The first, or inner problem, is usually amenable to a treatment specific to the metric space in which the problem is situated, and may not be computationally complex. It requires the optimization of the internal nodes of a given tree where the positions of the n labeled terminal nodes are known. The second, or outer problem involves optimizing over the set of all trees with n terminal nodes, a 'hard' problem which can be treated by worst-case exponential algorithms, or local optimization approaches, in a wide

*A version of this paper appeared in *International Journal of Foundations of Computer Science* 7:1, 1–9, 1996. An earlier version was presented at the *Symposium on Combinatorial Methods for Genome Rearrangements*, University of Southern California, March 18, 1994.

†Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada. Email: sankoff@ere.umontreal.ca.

‡Environmental Systems Research Institute, Redlands, CA 92373, USA. Email: gsundaram@esri.com

§Department of Computer Science, The University of Georgia, Athens, GA 30602, USA. Email: kece@cs.uga.edu

[1]We point out that some other versions of multiple alignment not involving a tree (such as minimizing the sum of pairwise distances) do not have obvious analogues in genomic rearrangement space.

variety of metric spaces. Here we will be discussing the inner problem only. Mathematically, the problem is defined as follows: Given a fixed phylogeny (tree) T, together with a set of k permutations (genomes), each of size n corresponding to the terminal (leaf) nodes, find a set of permutations corresponding to the internal nodes such that the total weight $w(T)$ is minimized, where $w(T)$ is defined as

$$w(T) = \sum_{(x,y)\in T} d(x,y).$$

Here $d(.,.)$ is the *genome rearrangement distance* metric defined on pairs of permutations. For example, in the tree given in Fig 1, we want to find the permutations corresponding to the internal nodes a, b, c and d.

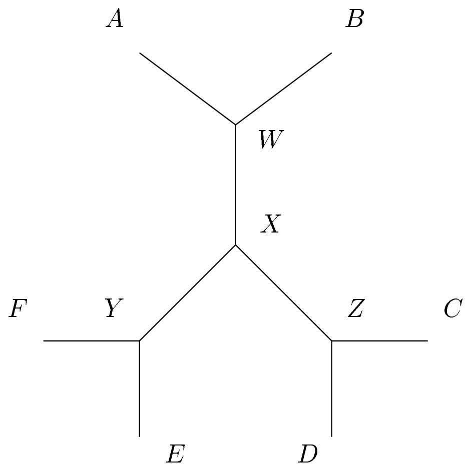

Figure 1 The problem of optimizing Steiner points. A, B, C, D, E and F are given permutations. X, Y, Z and W are permutations to be found so as to minimize the sum of the nine distances represented by the branches of the given tree.

In this paper we consider a heuristic for the problem of computing the internal nodes, given a fixed phylogeny and the permutations corresponding to the terminal nodes under the *inversion* and *transposition* distance metrics. First, we consider a more basic problem, the *median* problem, where T is a *star* on three vertices. We then divide the problem on an arbitrary binary tree into a number of overlapping median problems and apply the median algorithm iteratively to search for a heuristic solution to the original problem.

The paper is organized as follows. In Section 2, we review previous work on computing genome rearrangement distances between a given pair of permutations and its connection to our problem. In Section 3 we discuss some approximation algorithms for the median problem and a heuristic algorithm which gives a local optimum. In Section 4, we outline a heuristic algorithm for multiple genome comparison, based on the *iterative improvement method* of [15].

2 Previous work

The edit distances or genome rearrangement distances used in genomic comparisons can involve *inversions* in which the gene order along a chromosome is altered by the reversal of a segment of arbitrary length, *transpositions* where two adjacent segments interchange, *reciprocal translocation* of segments between two chromosomes, *duplication*, *deletion* or *insertion* of chromosome segments, the *fusion* of two chromosomes into one or the *fission* of one into two, and other processes. This terminology originates in microscopy-based cytogenetics (cf. Shulz-Schaeffer [18], Part VI, Swanson *et al.* [19], Ch. 6-7), but is readily interpretable on the molecular level. In formulations of the problem involving inversions, the elements of the permutations may be signed, indicating from which DNA strand the gene is read, so that when a segment is reversed, the signs of the elements in the segment change polarity. (Problems involving inversions of unsigned permutations are also important.) In addition, versions of these distances based only on intrachromosomal events, such as inversion or transposition, can be defined for circular as well as linear chromosomes. An effort at a formalization of these processes in a common framework can be found in earlier papers [13, 14].

Kececioglu and Sankoff [11] (extended version of [8, 9]) considered the problem of computing the minimum reversal distance between two given permutations in the unsigned case, including approximation algorithms and an exact algorithm feasible for moderately long permutations. Bafna and Pevzner [1] gave improved approximation algorithms for this problem. Kececioglu and Sankoff [10] also found tight lower and upper bounds for the signed case and implemented an exact algorithm which works rapidly for relatively long permutations. Recently Hannenhalli and Pevzner [5] have shown that the signed problem is only of polynomial complexity. Computation of the transposition distance between two permutations was considered by Bafna and Pevzner [2]. Sankoff *et al.* [16, 12] implemented and applied a heuristic to compute an edit distance which is a weighted combination of inversions, transpositions and deletions. Kececioglu and Ravi [7] began the investigation of translocation distances, and Hannenhalli [3] has shown that a formulation is of polynomial complexity.

3 Computing the median

Consider a special case of multiple genome comparison where the phylogeny is a star on three vertices. This is the *median* problem, formally defined as follows: Given three permutations A, B and C each of size n, we want to find a permutation S^* such that

$$S^* = \mathrm{argmin}_{S \in \Pi^n} \left\{ d(S,A) + d(S,B) + d(S,C) \right\},$$

where Π^n is the set of all permutations of length n and $d(A,B)$ is the rearrangement distance between permutations A and B.

3.1 Approximate solutions for the median

It is not known whether the problem of computing the median for genome rearrangement distances is NP-hard; indeed, except for translocations and signed inversions [3, 5], the complexity of computing the distance between two permutations remains open.

It is easy to find an approximation algorithm for the median problem, if there is an oracle to compute the distance between two permutations, as the weight of the minimum spanning tree on the three given points is at most $\frac{4}{3}$ the optimal tree T^*, for any arbitrary metric.

For comparing two permutations under the inversion metric, Hannenhalli and Pevzner [5] give a polynomial algorithm for the signed case. Hence in this case, we can construct a tree whose total weight cannot exceed $\frac{4}{3}$ times that of the optimal tree T^*, by solving the two-permutation problem for the pairs (A, B), (B, C) and (C, A) and outputing the minimum spanning tree. For the unsigned case, Bafna and Pevzner [1] give a $\frac{7}{4}$-approximation algorithm. This implies we can find a tree with associated reversals whose weight cannot exceed $\frac{7}{3}$ times the optimum. The problem of finding a tree with associated reversals whose weight does not exceed 2 times the optimum remains open. For the transposition metric, Bafna and Pevzner [2] give a $\frac{3}{2}$ approximation algorithm for the two permutation case; hence the minimum spanning tree for this metric is at most twice the length of the optimal median tree.

3.2 Metric lower bounds

Having an approximation algorithm is theoretically assuring, but in applications we must do better than a factor of 2 in the transposition case and a factor of $\frac{7}{3}$ in the unsigned reversals case. In the two-permutation case under the inversions metric, Kececioglu and Sankoff [10, 11] give algorithms that are effective in practice by providing relatively tight lower and upper bounds, though the guaranteed theoretical bound is only twice optimal. For the transpositions metric between two permutations, while Bafna and Pevzner's algorithm has a guarantee of 1.5, their lower bound based on the cycle decomposition is generally much better than this would suggest.

For the median problem, however, finding tight lower bounds appears to be more difficult. A simple lower bound, based on the triangle inequality, is

$$\frac{1}{2}\{d(A, B) + d(B, C) + d(C, A)\}.$$

This lower bound, which holds for any metric, implies that when the largest of the three pairwise distances is equal to the sum of the other two pairwise distances, the minimum spanning tree is an optimal median tree.

3.3 A heuristic for the median problem

Even though the metric lower bound is not sufficiently tight to make feasible branch-and-bound or other exhaustive searches for the median problem, we may still employ heuristics, sacrificing a guarantee of optimality in exchange for some computational experience that may by chance illuminate some of the features of a solution.

The following heuristic appears to work quite well on many instances from biological practice. First, we compute an *initial solution grid*, as follows for example for the inversions metric: For the pair of permutations A and B, we compute an optimal series of inversions using the branch and bound algorithm of Kececioglu and Sankoff [11]. Then, from each intermediate permutation X in this series, we again compute an optimal series of inversions between C and X. The entire procedure is repeated for the pairs B, C and C, A, as shown in Figure 2.

To construct the grid for the transposition metric, we can either use a greedy algorithm, or a branch-and-bound algorithm based on the lower bound of Bafna and Pevzner [2].

Then, from each permutation X in the initial solution grid, we initiate a search for a local optimum by finding the best search direction in a neighborhood of X, i.e. among all

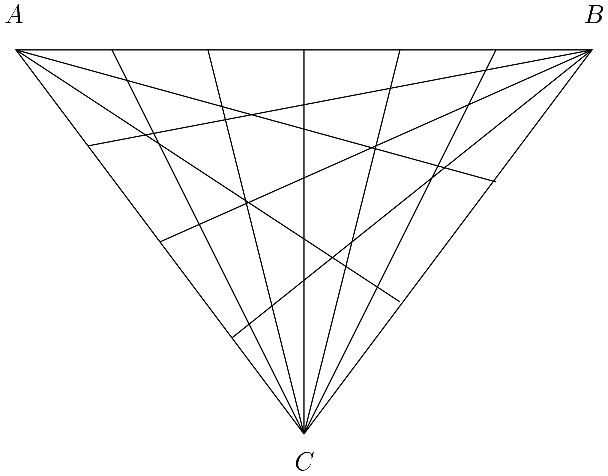

Figure 2 The initial solution grid.

permutations which are at a genome rearrangement distance of one from X. We stop each search when no further improvement can be made, and choose the best local optimum as our candidate for the global solution. We have implemented our heuristic using the code of Kececioglu and Sankoff [10, 11] for signed and unsigned inversions and a greedy algorithm for transpositions.

As an example, we ran our analysis on three signed permutations of length 33, representing mitochondrial gene orders in humans (A), sea urchin (B), and fruit fly (C):

A: (26 13 17 12 -24 15 18 32 -2 -16 -3 -33 4 -28 7 5 1 10 19 25 22 11 29 14 20 -21 - 8 6 30 -23 9 27 31)

B: (26 4 25 22 5 1 -28 19 11 29 20 -21 6 9 27 8 30 23 -24 16 14 -2 32 3 -31 15 -7 33 10 13 17 12 18)

C: (-26 -31 -27 12 -24 15 18 32 -3 -33 4 13 5 7 1 10 19 2 25 16 29 8 -9 -20 -11 -22 30 -23 21 6 28 -17 -14)

We found the following three solutions of length 39, yield an upper bound on the optimum value. (The three genomes, for example, are at distances 5, 19, and 15 from the first solution below.) The general metric lower bound of Section 3.2 for this data has value 37.5. Thus, the optimal solution to the median problem must be within one inversion of the metric lower bound. It is surprising that the general lower bound is so tight in this metric space.

(26 13 17 12 -24 15 18 32 25 22 11 29 14 -2 -19 -10 -1 -5 -3 -33 4 -28 7 16 20 -21 23 -30 -6 8 9 27 31)

(2 -14 -29 -11 -22 -25 -32 -18 -15 24 -12 -17 -13 -26 -31 -27 -9 23 -30 16 20 -21 -8 6 -7 28 -4 33 3 5 1 10 19)

(33 3 16 20 -21 6 -7 28 -32 -18 -15 24 -12 -17 -13 -26 -31 -27 -9 23 -30 -8 5 1 10 19 2 -14 -29 -11 -22 -25 -4)

A few observations that emerge from our simulations are illustrated by this example. First,

n	avg. no. sol.	avg. lower bound	avg. value
10	4	10	11
20	8	27	29
30	9	43	50
40	4	58	64
50	2	72	86

Table 1 Solutions of simulated problems using inversions; case of signed permutations.

n	avg. no. sol.	avg. lower bound	avg. value
10	3	8	9
20	7	17	18
30	8	24	27
40	4	35	37
50	8	43	47

Table 2 Solutions of simulated problems using transpositions on unsigned permutations.

the number of best solutions found by the heuristic is not very large. This contrasts with the two-permutation problem in which there are generally an extremely large number of optimal trajectories, and trajectory mid-points, between two permutations. Second, the solutions are relatively close together as can be readily seen by inspecting the three solutions in the example. A "triangulation" effect in the median problem seems to pin down the solution,[2] in contrast with the two-permutation case, where the average distance between the mid-points of optimal trajectories between two random permutations is roughly half the inversion distance.

Tables 1 and 2 illustrate experimental results for random permutations in the signed and unsigned cases.

4 A heuristic for multiple genome comparison

We return to the main problem, i.e. given a fixed tree T associated with known permutations at the terminal nodes, find a set of permutations for the internal nodes so that the total weight of the tree is minimized. Jiang, Lawler and Wang [6] show that given a fixed topology one can label the internal nodes with leaves such that the weight of the tree is at most twice optimal under any arbitrary metric.

Our heuristic for multiple genome comparison is analogous to the iterative improvement method of Sankoff, Cedergren and Lapalme [15]. We assume that the given phylogeny is an unrooted binary tree. An unrooted binary tree on k terminal nodes can be uniquely decomposed into $k-2$ 3-stars corresponding to the internal nodes. See Figure 3. We start

[2] Other examples of solutions to median problems are given by Hannenhalli et al.[4].

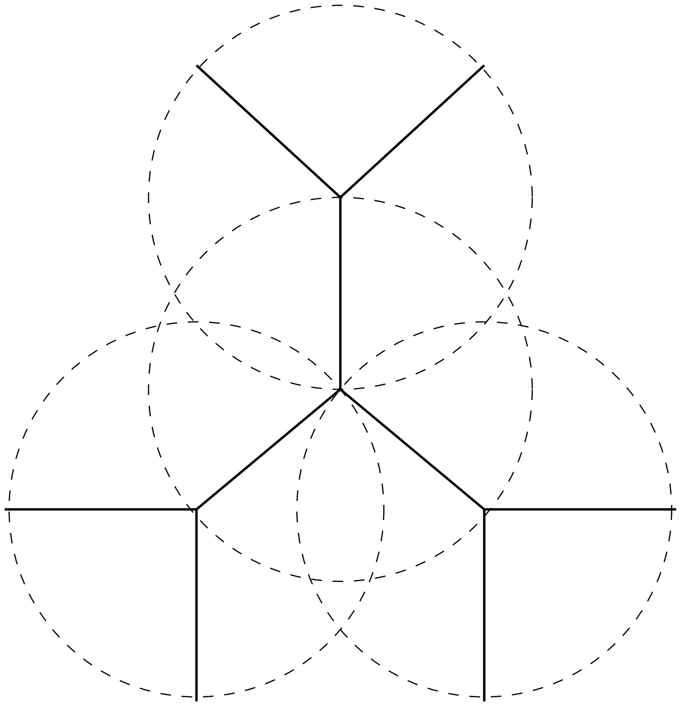

Figure 3 Decomposing the tree problem into a set of median problems.

with a good initial solution and iteratively improve the stars on three vertices, by the heuristic for median. This process will eventually converge to a local optimum.

We have implemented the algorithm and tested it in two kinds of experiments. First, we chose the innermost node of the tree in Figure 3 as the root and assigned the identity permutation to it. We then performed k random reversals along each branch of the tree to generate the permutations at the leaf nodes. For small values of k, the algorithm reconstructs the identity permutation in finding an optimal total tree length $(n-1)k$. For larger k the algorithm finds solutions more economical than the "true" configuration that generated the data. In the second kind of experiment, the terminal nodes were simply assigned random permutations. Evaluating the results of the analyses was not straightforward in this case. Not knowing better lower bounds than the one provided by Jiang *et al.* [6], we can not determine whether our solution is close to the optimum. Moreover, finding a good initial solution is crucial; assigning random permutations to the internal nodes for an initial solution converges in this context to a local optimum far from the global optimum.

Acknowledgements

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. DS is a fellow of the Canadian Institute for Advanced Research.

References

[1] V. Bafna and P.A. Pevzner. Genome rearrangements and sorting by reversals. In *34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993.

[2] V. Bafna and P.A. Pevzner. Sorting by transpositions. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 614-623, 1995.

[3] S. Hannenhalli. Polynomial algorithm for computing translocation distance between genomes. *Dept. of Computer Science and Engineering, Penn State University, Technical Report CSE-95-005*

[4] S. Hannenhalli, C. Chappey, E.V. Koonin and P.A. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: a test case. manuscript, Department of Computer Science, Penn State University. Earlier version presented at the Genome Rearrangement Workshop, University of Southern California, March 1994.

[5] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). *Dept. of Computer Science and Engineering, Penn State University, Technical Report CSE-95-004*

[6] T. Jiang, E.L. Lawler, and L. Wang. Aligning sequences via an evolutionary tree: complexity and approximation. *Proceedings of the 26th Symposium on Theory of Computing,* 760-769, 1994.

[7] J. Kececioglu and R. Ravi. Of mice and men. Evolutionary distances between genomes under translocation. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms,* pp. 604-613, 1995.

[8] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals. *Centre de recherches mathématiques Technical Report* 1824, July 1992.

[9] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two chromosomes. *Proceedings of the 4th Symposium on Combinatorial Pattern Matching*,Springer Verlag Lecture Notes in Computer Science 684:87-105, 1993.

[10] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. *Proceedings of the 5th Symposium on Combinatorial Pattern Matching*,Springer Verlag Lecture Notes in Computer Science 807:307-325. 1994.

[11] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica,* 13:180-210, 1995.

[12] D. Sankoff. Edit distance for genome comparison based on non-local operations. *Proceedings of the 3rd Symposium on Combinatorial Pattern Matching*,Springer Verlag Lecture Notes in Computer Science 644:121-135. 1992.

[13] D. Sankoff. Analytical approaches to genomic evolution. *Bilchimie*, 75:409–413, 1993.

[14] D. Sankoff. Models and analyses of genomic evolution. In *Second International Conference on Bioinformatics,Supercomputing and Complex Genome Analysis*, 1993.

[15] D. Sankoff, R.J. Cedergren, and G. Lapalme. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.*, 7:133–149, 1976.

[16] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B.F. Lang, and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA* 89, 6575-6579, 1992.

[17] D. Sankoff, C. Morel, and R.J. Cedergren. Evolution of 5S RNA and the non-randomness of base replacement. *Nature New Biology*, 245:232–234, 1973.

[18] J. Schulz-Schaeffer. *Cytogenetics*. Springer-Verlag, New York, 1980.

[19] C.P. Swanson, T. Merz, and W.J. Young. *Cytogenetics, 2nd ed.*. Prentice Hall, Englewood Cliffs, NJ 1981.