The Median Problem for Breakpoints in Comparative Genomics

David Sankoff[1] and Mathieu Blanchette[1]

Centre de recherches mathématiques, Université de Montréal
CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7
sankoff@ere.umontreal.ca blanchem@iro.umontreal.ca

Abstract. During evolution, chromosomal rearrangements, such as reciprocal translocation, transposition and inversion, disrupt gene content and gene order on chromosomes. We discuss algorithmic and statistical approaches to the analysis of comparative genomic data. In a phylogenetic context, a combined approach is suggested, leading to the *median problem for breakpoints*. We solve this problem first for the case where all genomes have the same gene content, and then for the general case.

1 Introduction

During biological evolution, inter- and intrachromosomal exchanges of chromosomal fragments disrupt the order of genes on a chromosome and, for multichromosomal genomes, the partition of genes among these chromosomes.

When comparing two evolutionarily diverging species, any (maximal) contiguous region of the genome in which gene content and order have been conserved in both species is called a *conserved segment*. Between any two adjacent conserved segments is a *breakpoint*. The number of conserved segments increases as they are disrupted by new events, so that they tend to become shorter over time. The number of chromosomal segments conserved during the divergence of two species, or equivalently, the number of breakpoints, can be used as a rough measure of their genomic distance.

Two approaches, the algorithmic and the statistical, have been taken to the reconstruction of genomic history based on the comparison of chromosomal gene content and order in two or more genomes. The first attempts to infer a most economical sequence of rearrangement events to account for the differences among the genomes, based only on the breakpoints, and neglects the contents of conserved segments. The second approach ignores the details of rearrangement history and assumes that a random model (the Nadeau-Taylor model) accounts for the differences in chromosomal gene content and order. In this paper, we discuss the strengths and weaknesses of the two approaches.

In the phylogenetic context, a compromise approach can be adopted, algorithmic, but not attempting to infer precise details of hypothesized evolutionary events. This leads to a new, tractable, problem, the *median problem for breakpoints*. We give a solution to the version of this problem where all genomes have the same gene content, and extend it to the case where the median and other genomes involved may have partially different gene sets.

2 The algorithmic approach

The algorithmic study of comparative genomics has focused on inferring the most economical explanation for observed differences in gene orders in two or more genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes, this has been formulated as the problem of calculating an edit distance between two linear orders on the same set of objects, representing the ordering of homologous genes in two genomes. In the most realistic version of the problem, a sign (plus or minus) is associated with each object in the linear order, representing the direction of transcription, or strandedness, of the corresponding gene. The elementary edit operations may include one or more of:
1) inversion, or reversal, of any number of consecutive terms in the ordered set, which, in the case of signed orders, also reverses the polarity of each term within the scope of the inversion. Kececioglu and Sankoff [13] 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. Recently, Caprara [4] showed this problem to be NP-complete. Kececioglu and Sankoff [12] also found tight lower and upper bounds for the signed case and implemented an exact algorithm which worked rapidly for long permutations. Indeed, Hannenhalli and Pevzner [8] showed in 1995 that the signed problem is only of polynomial complexity, and an improved polynomial algorithm was given by Kaplan, Shamir and Tarjan [10].
2) transposition of any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. This may or may not also involve an inversion. Computation of the transposition distance between two permutations was considered by Bafna and Pevzner [2]. Sankoff *et al.* [21, 18, 3] implemented and applied heuristics to compute an edit distance which is a weighted combination of inversions, transpositions and deletions.

In addition, for multi-chromosome genomes, a major role is played by:
3) reciprocal translocation. Kececioglu and Ravi [11] began the investigation of translocation distances, and Hannenhalli [7] has shown that a formulation is of polynomial complexity. A relaxed form of translocation distance was proposed by Ferretti *et al.*[6] and the complexity of its calculation was shown to be NP-complete by DasGupta *et al.* [5].

2.1 Shortcomings of the algorithmic approach

It would seem to be an advantage of the algorithmic approach that it actually constructs an optimizing series of events that accounts for the rearrangement of one genome with respect to another. There are two problems with this, however. One is the non-uniqueness of the solution, especially when all rearrangement events are weighted equally - one event equals one unit of the objective function being minimized - or even if the weights are integral multiples of some common factor. Though this problem is reduced with suitable event weights, a serious

measure of arbitrariness is thereby introduced. Some progress has recently been made in estimating appropriate weights empirically [18, 3].

The advantage of reconstructing a feasible history is thus diminished, since this history likely has no particular status with respect to many other equally parsimonious solutions. This problem is somewhat attenuated in the context of the median problem, to be discussed later. A more serious problem with reconstructed solutions is that when the number of steps approaches a certain proportion of the number of breakpoints, this number is almost certainly a serious underestimate [13]. Again, having a reconstructed history is a dubious advantage, since it inevitably contains some wrong steps and omits even more true events.

Finally, the algorithmic approach is very sensitive to errors and other small changes in the data. These are especially numerous when gene order has been determined by mapping techniques other than complete sequencing [20].

3 The statistical approach

Our formulation of the Nadeau-Taylor model of genomic divergence assumes that each reciprocal translocation breaks chromosomes at random points on two randomly chosen chromosomes. As a consequence when we compare two divergent genomes, the endpoints of the conserved segments making up each chromosome are uniformly and independently distributed along its length (spatial homogeneity of breakpoints). We also assume that which genes of a genome are discovered and mapped first does not depend on their position on the chromosome (spatial homogeneity of gene distribution), nor on their proximity to each other (independence of map positions).

In trying to count the number of conserved segments for the quantification of evolution, we must deal with underestimation due to conserved segments in which genes have not yet been identified in one or both species. This is particularly important if there are relatively few genes common to the data sets for a pair of species, so that many or most of the conserved segments are not represented in the comparison, and genomic distance may be severely underestimated. Nadeau and Taylor [16] in 1984 could only treat 13 segments out of the 100-200 now known to exist.

We model the genome as a single long unit broken at n random breakpoints into $n + 1$ segments, within each of which gene order has been conserved with reference to some other genome. (Little is lost in not distinguishing between breakpoints and concatenation boundaries separating two successive chromosomes[19].)

It is remarkable that to estimate n from m and the number of segments n_r observed to contain r genes, for $r = 1, 2 \ldots$, only the number of non-empty segments $a = \sum_{r>0} n_r$ is important [17].

Theorem 1. *The variable a is a sufficient statistic for the estimation of n.*

3.1 Estimating n from a

To estimate n, we study $P(a, m, n)$, the probability of observing a non-empty segments if there are m genes and n breakpoints. Combinatorial arguments give

$$P(a, m, n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}$$

After observing m and a it is an easy matter to find the value of n which maximizes P, i.e. the maximum likelihood estimate.

3.2 Weaknesses of the statistical approach

One weakness of the statistical approach is that it does not estimate a specific series of events, although we have discussed how this advantage of the algorithmic approach is dubious. The number of breakpoints (or conserved segments) cannot be deterministically converted into a number of events, since different types of rearrangement produce different numbers of breakpoints, and even a single type of event does not always produce the same number of breakpoints.

Perhaps the greatest potential weakness of this approach is that it depends on the applicability of a particular probabilistic model. This is a temporary problem, however, in that it gives rise to further research on better models [14].

4 The median problem

For phylogenetic purposes, it is useful to solve the following sort of problem: Given a distance or dissimilarity d, three genomes A, B and C, and a set of genes Σ, we want to find a genome S containing all the genes in Σ such that

$$d(S, A) + d(S, B) + d(S, C)$$

is minimized. How a solution to this problem can be the key to solving phylogenetic problems involving many genomes is discussed in [22, 6].

The set Σ may be determined by the phylogenetic problem under study, or may be defined by the analyst. For example, if A, B and C all contain the same set of genes, then it is natural to use this set for Σ. Another example is drawn from organelle evolution, where genes tend to be lost from the genome and not re-inserted. Then if A is ancestral to both B and C, the set Σ will be the union of the set of genes in B and the set of genes in C. Still another possibility, where the direction of evolution is less clear, is to include in Σ just those genes that are in at least two of A, B and C.

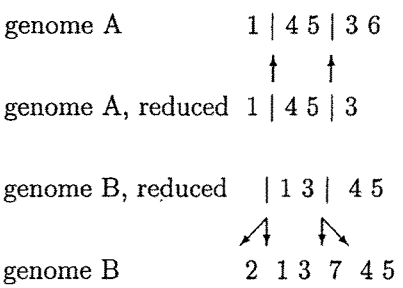

Fig. 1. Defining breakpoints for (circular) genomes with different gene contents. Position of breakpoints (vertical strokes) found first in reduced genomes with identical gene sets. This unambiguously determines breakpoints between 1 and 4 and between 5 and 3 in genome A. Breakpoint between 5 and 1 in genome B is "hidden" by gene 2; that between 3 and 4 is hidden by gene 7.

4.1 Breakpoints

Consider two genomes $A = a_1 \ldots a_n$ and $B = b_1 \ldots b_n$ on the same set of genes $\{g_1, \ldots, g_n\}$. We say a_i and a_{i+1} are adjacent in A (and a_n and a_1 are adjacent as well in circular genomes). If two genes g and h are adjacent in A but not in B, they determine a breakpoint in A. The number of breakpoints in A is clearly equal to the number of breakpoints in B.

For two genomes whose gene sets are not identical, to calculate the breakpoints, we first remove all genes that are present in only one of the genomes. We then find the breakpoints for the reduced genomes, now of identical composition. The positions of the breakpoints are well-defined in the reduced genomes. In the full genomes, there is a breakpoint between a_i and a_{i+1} only if this is a breakpoint for the reduced genome. If, as in Figure 1, there is a breakpoint between a_i and a_j in the reduced genome, where $j \neq i+1$, then there is a corresponding breakpoint in the full genome, but its position is ambiguous. We call it a *hidden* breakpoint; it is somewhere between a_i and a_j, which are not adjacent.

4.2 The median problem for a fixed gene set Σ

Now define d to be the number of breakpoints derived from the comparison of two genomes containing the same genes Σ. The median problem becomes one of finding the genome S on Σ that determines the fewest total breakpoints between itself and the three genomes $A.B$ and C.

In contrast to other genomic distances, this problem seems relatively tractable, although its computational complexity remains to be determined. We proceed by reduction to the Traveling Salesman Problem (TSP).

It will be convenient to describe genomes in graph-theoretical terms. The genes will be represented by the vertices of the graph and adjacency of two genes will be indicated by the existence of a corresponding edge in the graph.

Thus, only graphs consisting of a single complete cycle of the vertices represent (circular) genomes.

We first define G to be the complete graph whose vertices are the elements of Σ. For each edge gh in $E(G)$, let $u(gh)$ be the number of times g and h are adjacent in the three genomes. Set $w(gh) = 3 - u(gh)$. Then the solution to TSP on (G, w) traces out an optimal genome S on Σ, since if g and h are adjacent in S, but not in A, for example, then they form a breakpoint in S.

4.3 A lower bound

To solve this restricted form of TSP, we resort to a branch-and-bound algorithm based on the following lower bound:

Let the*edge-pool* $P \subseteq E(G)$, be disjoint from the *fragment* $F \subseteq E(G,)$ and let $score = \sum_{gh \in F} w(gh)$. Define $a(g)$, the *availability* of $g \in V(G)$, to be 2, 1 or 0, depending on whether g is incident to zero, one, or more than one edge in F, respectively. Let $\mu(g)$ be the sum of the $a(g)$ smallest weights of edges in P incident to g. ($\mu(g)$ is undefined if there are less than $a(g)$ such edges.)

If there is a TSP solution cycle S of weight W_S which includes all the edges in the fragment F and some additional edges drawn from the edge-pool P, let $\nu(g)$ be the sum of the weights of the exactly $a(g)$ edges of S in P incident to g. (In this case $\mu(g)$ is always defined.) Clearly $\mu(g) \leq \nu(g)$.

Now,

$$W_S = score + \sum_{gh \in E(S) \cap P} w(gh)$$

$$= score + \frac{1}{2} \sum_{g | gh \in E(S) \cap P} w(gh)$$

since each edge in $E(S) \cap P$ is counted twice in the sum. Thus

$$W_S = score + \frac{1}{2} \sum_{g | gh \in E(S) \cap P} \nu(g).$$

Defining

$$L(P) = \frac{1}{2} \sum_{g | gh \in E(S) \cap P} \mu(g),$$

$$score + L(P) \leq score + \frac{1}{2} \sum_{g | gh \in E(S) \cap P} \nu(g) = W_S.$$

We use $L(P)$ as a lower bound in the branch-and-bound algorithm in Section 4.4. When $P = E(G)$ and $F = \Phi$ this as a well-known bound on TSP (see, e.g., pp. 272-273 in [15]). There are a number of other bounds which can be used for the TSP, but this one is of particular interest in that it can be modified for use in the median problem with more general genomes as discussed in Section 5.2.

4.4 Algorithm BBF

input: weighted complete graph (G, w)
output: solution S to the TSP on (G, w)

initialization
{ $V(S) \leftarrow V(G)$
$F \leftarrow \Phi$
$P \leftarrow E(G)$
score$\leftarrow$ 0
best$\leftarrow$ ∞
}
procedure BBF(P, F, S,score,best)
{ **if** $|F| = |G|$ **and** score $<$ best **then**
{ store $S = F$ as current best solution
best$\leftarrow$score
}
if $|F| < |G|$ **then**
{ **if** $L(P)$ + score $<$ best **then**
{ choose $gh \in P$ to try to add to F
where $a(g) > 0, a(h) > 0$ and $w(gh)$ is as small as possible,
and $F \cup \{gh\}$ is not a cycle on less than $|G|$ vertices.
BBF($P - \{gh\}, F \cup \{gh\}, S$,score + $w(gh)$,best)
BBF($P - \{gh\}, F, S$,score,best)
}
}
}

The recursion functions as a "greedy" search until it first finds a cycle, which is necessarily an upper bound. If its cost $U = L(E(G))$, it is optimal.

4.5 Genomes with directionality

In the case of directed genomes, the notion of breakpoint must be modified to take into account the polarity of the two genes. If gh represents the order of two genes in one genome, then if another genome contains gh or $-h - g$ there is no breakpoint involved. However, between gh and hg there is a breakpoint, similarly between gh and $-g-h, g-h, -gh, h-g$ or $-hg$. Adjacency is no longer commutative. The reduction of the median problem to TSP must be somewhat different to take into account that the median genome contains g or $-g$ but not both. Let G be a complete graph with vertices $V = \{-g_n, \ldots, -g_1, g_1, \ldots, g_n\}$, where $\Sigma = \{g_1, \ldots, g_n\}$. For each edge gh in $E(G)$, let $u(gh)$ be the number of times $-g$ and h are adjacent in the three genomes A, B and C, and $w(gh) = 3 - u(gh)$, if $g \neq -h$. If $g = -h$, we simply set $w(gh) = -M$, where M is large enough to assure that a minimum weight cycle must contain the edge $-gg$.

Proposition: If $s = s_1, -s_1, s_2, -s_2, \ldots, s_n, -s_n$ is the solution of the TSP on (G, w), then the median is given by $S = s_1 s_2 \ldots s_n$.

Proof. $d(S,A) + d(S,B) + d(S,C) = \sum_{gh \in S, g \neq -h} w(gh)$

$$= nM + \sum_{gh \in s} w(gh).$$

Thus S minimizes $d(S,A) + d(S,B) + d(S,C)$ iff s is of minimal weight.

The same bound $L(G)$ may be constructed as before, though for directed genomes $\mu(g) = -M +$ smallest weight of any edge incident to g.

An implementation of the algorithm we have described finds the median of three directed genomes of size 50 in one minute, on average on an Origin 200 computer with a RISC 10000 processor. Random genomes are easily processed since $L(G)$ tends to be a fairly tight bound. Three similar genomes are also rapidly treated since the first $|G|$ "greedy" recursive steps are likely to produce an optimal solution. It is between these extremes that longer execution times are encountered.

4.6 Larger stars

The median problem can also be defined for $k > 3$ genomes. When all these genomes have identical gene sets, the BBF procedure is directly applicable to finding the median, the only difference being in the calculation of the weights where $w(gh)$ becomes $k - u(gh)$.

5 The case of a more general median gene set

5.1 Extension of the previous method

If the differences among the sets of genes in A, B, C and Σ consist of very few genes, the bound and algorithm in Section 4.2 can be adapted to function relatively efficiently. We redefine $w(gh)$ =(number of genomes containing both g and h) $- u(gh)$. In the algorithm in Section 4.4,in the call

BBF$(P - \{gh\}, E(S) \cup \{gh\}$,score $+ w(gh)$,best),

"score $+ w(gh)$" must be replaced "score $+ w(gh) + z(gh)$" where $z(gh)$ counts the "hidden" breakpoints (cf Section 5.2) caused by the addition of gh to the solution.

5.2 A better bound

In this section, we develop a bound designed for the situation where the gene content of Σ can differ considerably from that of A, B and/or C.

We assume all genes in A, B or C are also in Σ, and each gene in Σ is in at least one of A, B or C, since only these can contribute to the weight of a cycle.

There will, however, generally remain genes in Σ which are absent from some, but not all, of A, B and C, and as we shall see, this is the crux of the difficulty.

The bound in Section 4.3 was based on the fact that each vertex on a cycle is incident to two edges, and it was easy to bound the sum of their two weights. In the present context, when examining each vertex g on a cycle, we have to take into account that its incident edges may not be relevant to the breakpoint calculations with respect to one or more of the given genomes; we may have $gi \in S, ih \in S$, but i absent from A and g not adjacent to h in A. The breakpoint between g and h in S is hidden by gene i.

Suppose we wish to bound the contributions, to the cost of a cycle, of the edges"near" g in S, since the individual edges directly incident may not be relevant to all of A, B and C, as we have seen. We arbitrarily impose a directionality on S. If g is in genome X, let l_X and r_X be the closest vertices to the left and right of g in S that are also present in genome $X, X \in \{A, B, C\}$. If g is not in genome X, it cannot be involved in a breakpoint. The cost of the edges near g, summed over all $g \in \Sigma$, is then

$$W = \frac{1}{2} \sum_{X \in \{A,B,C\}} \sum_{g \in \Sigma \cap X} w(l_X g) + w(g r_X)$$

where $w(l_X g) = 0$ if l_X is adjacent to g in X, and $w(l_X g) = 1$ otherwise; similarly for $w(r g_X)$.

What is the configuration of the l_X and r_X around g in S? To the left of g we may have $l_{Y(1)} \ldots l_{Y(2)} \ldots l_{Y(3)}$, where $(Y(1), Y(2), Y(3))$ is a permutation of (A, B, C), and "rightward exclusion" prevails: $l_{Y(1)}$ is in genome $Y(1)$ but not in genomes $Y(2)$ or $Y(3)$; $l_{Y(2)}$ is in genome $Y(2)$ but not in genome $Y(3)$; $l_{Y(3)}$ is in genome $Y(3)$. If g is absent from one or two of genomes A, B or C, then there will be at most two or one l terms, respectively.

Other possibilities are that we may have only $l_{Y(1)} \ldots l_{Y(2)}$ left of g, and one of these genes is in two of the genomes A, B and C (rightward exclusion still obtains), or that there is only one l gene common to the three genomes.

A similar accounting of the possibilities can be made for the r genes, involving the notion of "leftward exclusion".

Then a lower bound on the cost of S is found by choosing, for each $g \in \Sigma$,
- up to three (depending on how many of A, B, C contain g) genes l_A, l_B, l_C, not necessarily distinct, each l_X in genome X, and some permutation $(Y(1), Y(2), Y(3))$ of (A, B, C) such that $l_{Y(1)} l_{Y(2)} l_{Y(3)}$ (or $l_{Y(1)} l_{Y(2)}$ if there are only two distinct genes) is rightward exclusive, and
- up to three (depending on how many of A, B, C contain g) genes r_A, r_B, r_C, not necessarily distinct, each r_X in genome X, and some permutation $(Z(1), Z(2), Z(3))$ of (A, B, C) such that $r_{Z(1)} r_{Z(2)} r_{Z(3)}$ (or $r_{Z(1)} r_{Z(2)}$ if there are only two distinct genes) is leftward exclusive, such that

$$t(g) = \sum_{X \in \{A,B,C\}, g \in X} w(l_X g) + w(g r_X)$$

is minimized by the cycle fragment $l_{Y(1)} l_{Y(2)} l_{Y(3)} g r_{Z(1)} r_{Z(2)} r_{Z(3)}$.

Then a lower bound on W is given by $\Lambda = \frac{1}{2}\sum_{g\in\Sigma} \min t(g)$, since each breakpoint is counted at most twice in the sum.

5.3 Adapting the bound for the stepwise construction of a cycle

Suppose, somewhat differently from Section 4.3, a candidate fragment of a path $F = s_1\, s_2 \ldots s_j$ of a solution cycle has already been constructed, and a pool Q of vertices remain to be tested for possible addition to the cycle. We define the availability $\alpha_X(g)$ to be 0 if g is not in X, and otherwise to be 2, 1 or 0, depending on whether g is not in F, g is the leftmost or rightmost in F of $V(X) \cap V(F)$, or is some other gene in F, respectively.

Then $\Lambda(Q) = \frac{1}{2}\sum_{g\in Q} \min t(g)$, is a lower bound on the weight of the remainder of the cycle, where the search for the minimizing cycle fragment for each g is constrained to respect the order of genes already in F and to use both an l_X and an r_X only if $\alpha_X(g) = 2$. Only one of l_X or r_X can be used if if $\alpha_X(g) = 1$.

5.4 Algorithm BBG

input: genomes A, B, C, median gene set Σ
output: solution S to the median problem

initialization
{ $F \leftarrow g$ (arbitrary choice)
$Q \leftarrow \Sigma - g$
score$\leftarrow$ 0
best$\leftarrow$ ∞
for $X = A, B, C$
{ $last(X) = g$ if g is in X, otherwise $last(X) = \Phi$
$\alpha_X(g) = 2$ if g is in X, otherwise $\alpha_X(g) = 0$
}
}
procedure BBG$(Q, F, S, \alpha, last$,score,best)
{ **if** there are $|\Sigma|$ edges in F **and** score $<$ best **then**
{ store $S = F$ as current best solution
best$\leftarrow$score
}
if there are less than $|\Sigma|$ edges in F **then**
{ **if** $\Lambda(Q)$ + score $<$ best **then**
{ choose $h \in Q$ to try to add to F, i.e.add edge s^*h, where $s^* = s_{|F|}$
(except if there are $|\Sigma| - 1$ edges in F: here we add s^*g)
for $X = A, B, C$ if $h \in X$
{ $\alpha_X(last(X)) = \alpha_X(last(X)) - 1$
$last'(X) = last(X)$
$last(X) = h$
}
BBG$(Q - \{h\}, F \cup s^*h, S, \alpha, last$,score$(F \cup s^*h)$,best)

```
            for X = A, B, C if h ∈ X
            {   α_X(last(X)) = α_X(last(X)) + 1
                last(X) = last'(X)
                    remove s*h from further consideration
            }
            BBG(Q, F, S, α, last,score,best)
        }
    }
}
```

Note that in the first recursive call of **BBG**,

$$\text{score}(F \cup s^*h) = \text{score} + \sum_{X|h \in X} w(last(X)h)$$

unless F contains $|\Sigma| - 1$ edges, in which case

$$\text{score}(F \cup s^*h) = \text{score} + \sum_{X|h \in X} w(last(X)h) + \sum_{X|h \in X} w(first(X)h).$$

6 Discussion

The problems of non-uniqueness and underestimation inherent in parsimonious analyses of genomic distance (or sequence distance) are attenuated when more than two genomes (or sequences) are compared. The median problem is the archetype of this effect: triangulation increases accuracy. With other methods of genomic distance, however, the median problem turns out to be much more difficult than a pairwise comparison [22].

The number of breakpoints between two genomes is not only the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution underlying the data, but it is also the easiest to calculate. We might then expect the median problem for breakpoints to be more tractable than for other measures, and the preliminary work reported here supports this hope. The relatively easy extension from 3 to k genomes is also a positive indication of its feasibility for phylogenetics.

7 Acknowledgements

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and a NSERC fellowship for graduate studies to MB. 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. *SIAM Journal of Computing* 25:272-289, 1996.
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. M. Blanchette, T .Kunisawa and D. Sankoff. Parametric genome rearrangement. *Gene* 172, GC 11-17, 1996.
4. A. Caprara. Sorting by Reversals is Difficult. *Proceeedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97),* 75-83, 1997.
5. B. DasGupta, T.Jiang, S.Kannan, M.Li, Z.Sweedyk. On the Complexity and Approximation of Syntenic Distance.*Proceeedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97),* 99-108,1997.
6. V. Ferretti, J.H. Nadeau and D. Sankoff. Original synteny. *Proceedings of the Seventh Annual Symposium on Combinatorial Pattern Matching,* D. Hirschberg and G. Myers ed., Springer Verlag Lecture Notes in Computer Science, 1075: 159-167, 1996.
7. S. Hannenhalli. Polynomial algorithm for computing translocation distance between genomes. *Proceedings of the 6th Symposium on Combinatorial Pattern Matching,* Springer Verlag Lecture Notes in Computer Science:162-176, 1995.
8. S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing,* pp. 178-189, 1995.
9. S. Hannenhalli and P.A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science,* 581-592, 1995.
10. H. Kaplan, R. Shamir, and R.E. Tarjan. Faster and Simpler Algorithm for Sorting Signed Permutations by Reversals. *Proceeedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 97),* 1997.
11. 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.
12. 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.
13. J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica,* 13:180-210, 1995.
14. I. Marchand. *Généralisations du modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* MSc thesis, Département de mathématiques et de statistique, Université de Montréal. 1997.
15. E. Minieka, *Optimization Algorithms for Networks and Graphs,* Industrial Engineering vol.1, New York: Marcel Dekker, chap. 7.3. pp. 272-273, 1978
16. J.H. Nadeau and B.A. Taylor Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA,* 81: 814-818, 1984.
17. M.-N. Parent. *Estimation du nombre de segments vides dans le modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* MSc thesis, Département de mathématiques et de statistique, Université de Montréal. 1997.

18. 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.
19. D. Sankoff and V. Ferretti. Karotype distributions in a stochastic model of reciprocal translocation. *Genome Research* 6, 1-9, 1996.
20. D. Sankoff, V. Ferretti and J.H. Nadeau. Conserved segment identification. *RECOMB 97. Proceedings of the First Annual International Conference on Computational Molecular Biology.* New York: ACM Press,1997, pp. 252-256. revised version to appear in *Journal of Computational Biology.*
21. 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.
22. D. Sankoff, G.Sundaram and J. Kececioglu. Steiner points in the space of genome rearrangements. *International Journal of the Foundations of Computer Science* 7, 1-9,1996,