

Genomic Sequence Comparison

Pavel Pevzner

Computer Science Department
The Pennsylvania State University
University Park, PA 16 802

June 26, 1995

[summary by Mireille Régnier]

1. Introduction

Sequence comparison is traditionally based on gene comparison based on local mutations (insertions, deletions or substitutions of nucleotides). Such comparisons do not yield evolutionary information. It appears that evolution is manifested as the divergence in gene order. For example, the large number of conserved segments in the maps of man and mouse suggests that multiple chromosome rearrangements have occurred since the divergence of lineages leading to human and mice. The number of such rearrangements has been recently estimated to be approximately 180. This leads to a major shift of sequence comparison toward the analysis of such rearrangements at the *genome level*. Nevertheless, there are almost no computer science results allowing a biologist to analyze gene rearrangements.

This talk addresses two problems: define and estimate the distance between two different species for a same gene and reconstruct the rearrangement scenario. A paper version can be found in [1].

2. State of the Art

Some genomes evolve so rapidly that the similarity between many genes is very low and is indistinguishable from the background noise. Nevertheless, according to Ohno's law, gene content of X chromosomes is assumed to have remained the same throughout mammalian development in the last 125 million years. However, the order of genes on X chromosomes has been disrupted several times, even though synteny has been almost completely conserved.

The order of genes in two organisms is represented by permutations $\pi = (\pi_1 \pi_2 \dots \pi_n)$ and $\sigma = (\sigma_1 \sigma_2 \dots \sigma_n)$.

DEFINITION 1. A *reversal* ρ of an interval $[i, j]$ is a permutation

$$\rho = [1, 2, \dots, i-1, j, j-1, \dots, i+1, i, j+1, \dots, n]$$

$\pi \cdot \rho$ has the effect of reversing genes $\pi_i, \pi_{i+1}, \dots, \pi_j$.

The *reversal distance problem* is to find a series of reversals $\rho_1, \rho_2, \dots, \rho_t$ such that $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_t = \sigma$ and t is minimum (Fig. 1a). The number t is called the *reversal distance*. *Sorting by reversals* is the problem of finding reversal distance $d(\pi)$ between π and identity i .

Reversals generate the symmetric group S_n . Given a set of generators of a permutation group, determining the shortest product of generators that equals π is NP-hard [2]. The problem is PSPACE-complete [4]. And in [5] it is conjectured that sorting by reversals is NP-complete even

when the generator set is fixed. Bounds for the related problem of sorting by prefix reversals can be found in [3]. Gollan conjectured that the reversal diameter of S_n , i.e. the maximal reversal distance between two permutations, is $d(n) = n - 1$, a bound achieved for only one permutation. Lower bound and verification for $n < 200$ are presented in [5].

3. Breakpoint graph

The key idea to sort by reversals is the definition of the *breakpoint graph* (see Fig. 1).

Let $\pi = (\pi_1 \dots \pi_n)$ be a permutation of the elements $\{1, \dots, n\}$. Denote $i \sim j$ if $|i - j| = 1$. Extend a permutation $\pi = (\pi_1 \dots \pi_n)$ by adding $\pi_0 = 0$ and $\pi_{n+1} = n + 1$. We call a pair of consecutive elements π_i and π_{i+1} , $0 \leq i \leq n$, of π a *breakpoint* if $\pi_i \not\sim \pi_{i+1}$. The *breakpoint graph* of π is an edge-coloured graph $G(\pi)$ with $n + 2$ vertices $\{\pi_0, \pi_1, \dots, \pi_n, \pi_{n+1}\}$. We join vertices π_i and π_j by a *black* edge if $i \sim j$ and by a *gray* edge if $\pi_i \sim \pi_j$. (See Fig. 1b). Later we also use the notion of breakpoint graph $G(\pi, \gamma)$ for *two* permutations π and γ which is defined as $G(\pi, \gamma) \equiv G(\pi\gamma^{-1})$ described earlier. A *cycle* in an edge-coloured graph G is *alternating* if the colours of every two consecutive edges of this cycle are distinct. In the following, by cycles we mean alternating cycles.

Let $\vec{\pi}$ be a *signed* permutation of $\{1, \dots, n\}$, i.e. a permutation with “+” or “-” sign associated with each element (Fig. 1c). In the signed case, every reversal of fragment $[i, j]$ changes *both* the order and the signs of the elements within that fragment. We are interested in the minimum number of reversals $d(\vec{\pi})$ required to transform a signed permutation $\vec{\pi}$ into the identity signed permutation $(+1 + 2 \dots + n)$. Define a transformation from a signed permutation $\vec{\pi}$ of order n to an (unsigned) permutation π of $\{1, \dots, 2n\}$ as follows. To model the signs of elements in $\vec{\pi}$ replace the positive elements $+x$ by $2x - 1, 2x$ and negative elements $-x$ by $2x, 2x - 1$ (Fig. 1c). We call the unsigned permutation π , the *image* of the signed permutation $\vec{\pi}$. In the breakpoint graph $G(\pi)$, elements $2x - 1$ and $2x$ are joined by both black and gray edges for $1 \leq x \leq n$. We define the breakpoint graph $G(\vec{\pi})$ of a signed permutation $\vec{\pi}$ as the breakpoint graph $G(\pi)$ with these $2n$ edges excluded. Observe that in $G(\vec{\pi})$ every vertex has degree 2 (Fig. 1c) and therefore the breakpoint graph of a signed permutation is a collection of disjoint cycles. Denote the number of such cycles as $c(\vec{\pi})$. We observe that the identity signed permutation of order n maps to the identity (unsigned) permutation of order $2n$, and the effect of a reversal on $\vec{\pi}$ can be mimicked by a reversal on π thus implying $d(\vec{\pi}) \geq d(\pi)$. In the following, by sorting the image $\pi = \pi_1\pi_2 \dots \pi_{2n}$ of a signed permutation $\vec{\pi} = \vec{\pi}_1\vec{\pi}_2 \dots \vec{\pi}_n$, we mean sorting of π by reversals $\rho(2i + 1, 2j)$ which “cut” *only after even positions* in π . In the rest of this section, π is an image of a signed permutation.

Cycle decompositions play an important role in estimating the reversal distance. Applying a reversal to a permutation may change the number of breakpoints, $b(\pi)$, as well as the number of cycles in a maximum decomposition, $c(\pi)$. The key idea in the algorithm of [1] is to take advantage of this strong correlation. One proves:

THEOREM 1. *For every permutation π and reversal ρ , one has:*

$$\Delta b(\pi, \rho) + \Delta c(\pi, \rho) \leq 1.$$

PROOF. (sketch): every reversal removes/adds at most two breakpoints. One considers all 5 potential values of Δb in a case-by-case fashion. \square

This immediately gives a new lower bound for the reversal distance:

THEOREM 2. *For every permutation π , $d(\pi) \geq b(\pi) - c(\pi)$.*

For all biological examples, one has $d(\pi) = b(\pi) - c(\pi)$. Hence, the use of the breakpoint graph reduces the reversal distance problem to maximal cycle decomposition problem. One shows:

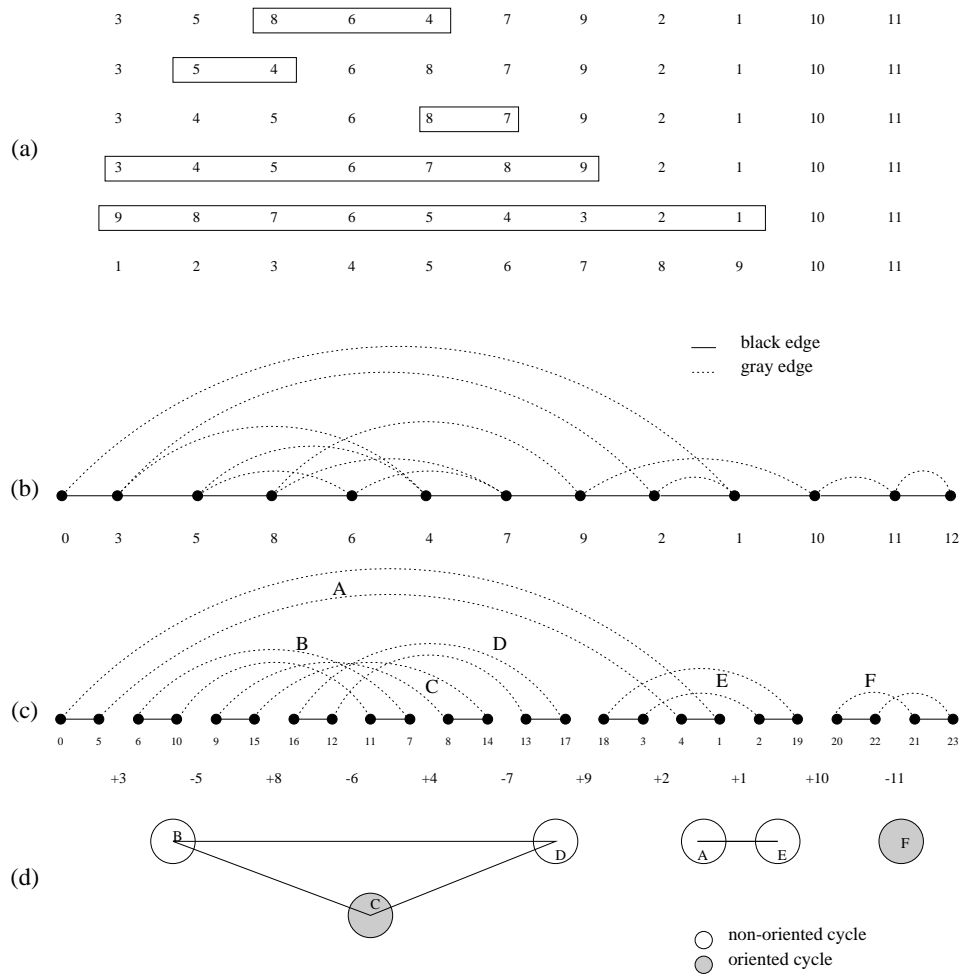


FIGURE 1. (a) Optimal sorting of a permutation $\sigma = (3\ 5\ 8\ 6\ 4\ 7\ 9\ 2\ 1\ 10\ 11)$ by 5 reversals. (b) Breakpoint graph $G(\sigma)$. (c) Transformation of a signed permutation into an unsigned permutation π and the breakpoint graph $G(\pi)$. Gray edges $(8, 9)$ and $(22, 23)$ are oriented while gray edges $(4, 5)$ and $(18, 19)$ are unoriented. Cycles C and F are oriented while cycles A, B, D and E are unoriented. Gray edges $(6, 7)$ and $(12, 13)$ are interleaving while gray edges $(6, 7)$ and $(4, 5)$ are non-interleaving. (d) Interleaving graph H_π with two oriented and one unoriented component.

THEOREM 3 (STRONG GOLLAN CONJECTURE). *For every n , the only permutations that require $n - 1$ reversals to be sorted are γ_n and its inverse γ_n^{-1} :*

$$\gamma_n = \begin{cases} (1, 3, 5, 7, \dots, n-1, n, \dots, 8, 6, 4, 2), & n \text{ even;} \\ (1, 3, 5, 7, n, n-1, \dots, 8, 6, 4, 2), & n \text{ odd.} \end{cases}$$

PROOF. (sketch). Let P_n be the set of n -permutations that satisfy $d(\pi) = n$. It contains γ_n and γ_n^{-1} . One inductively proves that there are no other elements. \square

Finally, one proves that the expected reversal distance is very close to the reversal diameter. The key idea is that a typical cycle is long, hence the number of cycles is small. More precisely: $E(d) \geq (1 - \frac{4}{\log n})n$

4. Algorithms

4.1. A greedy algorithm. Define a *strip* of π as an interval $[i, j]$ such that $(i-1, i)$ and $(j, j+1)$ are breakpoints, and no breakpoint lies between them. A strip is *increasing* if π_i, π_j , otherwise it is *decreasing*. A reversal can remove at most two breakpoints; therefore $d(\pi) \geq \frac{b(\pi)}{2}$. In [5] a greedy procedure is given, where one chooses a reversal that removes the most breakpoints of π , resolving ties in favour of reversals that leave a decreasing strip. An upper bound on the number on $d(\pi)$ that provides a performance guarantee of 2, follows from the lemma:

LEMMA 1. *If π is a decreasing permutation with a decreasing strip, then π allows a 1 or 2-reversal. Additionally, If every reversal that removes a breakpoint of π leaves a permutation with no decreasing strips, then π has a 2-reversal.*

4.2. An approximation algorithm for signed permutations. While the problem of sorting signed permutations is easier to handle, it is also more relevant to a biological point of view: genes are directed fragments of DNA sequences. Fortunately, the concept of breakpoint graph as well as strips extends naturally to signed permutations (see above). The algorithm SignedSort sorts a signed permutation π in at most $b(\pi) - \frac{1}{2}c_4(\pi)$ reversals, where c_4 is the number of 4-cycles. It provides an approximation ratio of $\frac{3}{2}$.

4.3. An approximation algorithm for sorting by reversals. As 2-reversals correspond to elimination of 4-cycles, one concentrates on finding a cycle decomposition with a large number of 4-cycles. The algorithm *ReversalSort* achieves an approximation ratio of $\frac{9}{5}$.

To conclude, let us cite among the remaining open problems the analysis of genome rearrangements in *multiple* genomes.

Bibliography

- [1] Bafna (V.) and Pevzner (P.). – Sorting by reversals: rearrangements in plant organelles and evolutionary history of mammalian chromosome. *Molecular Biology and Evolution*, vol. 12, 1994, pp. 239–246.
- [2] Even (S.) and Goldreich (O.). – The minimum-length generator sequence problem is NP-hard. *Journal of Algorithms*, vol. 2, 1981, pp. 311–313.
- [3] Gates (K. W.) and Papadimitriou (Ch.). – Bounds for sorting by prefix reversals. *Discrete Mathematics Algorithms*, vol. 27, 1979, pp. 47–57.
- [4] Jerrum (M.). – The complexity of finding minimum-length generator sequences. *Theoretical Computer Science*, vol. 36, 1985, pp. 265–289.
- [5] Kececioğlu (J.) and Sankoff (D.). – Exact and approximation algorithms for the reversal distance between two permutations. *Algorithmica*, 1995. – To appear.