Transposition distance between a permutation and its reverse

João Meidanis[1,5] Maria Emilia M. T. Walter[2,1,3] Zanoni Dias[1,4]

[1] University of Campinas, Institute of Computing, Brazil.
[2] University of Brasilia, Computer Science Department, Brazil.
[3] Partially supported by Brazilian agency CAPES.
[4] Supported by Brazilian agency CNPq.
[5] Partially supported by Brazilian agencies FAPESP and CNPq.

Abstract. In this note we solve an open question posed by Bafna and Pevzner [1], regarding chromosome distance with respect to transpositions: we show that the distance between a permutation and its reverse (without complementation) is $\lfloor n/2 \rfloor + 1$, where n is the size of the permutations. We also present an algorithm to compute an optimal series of transpositions.

1 Introduction

The huge amount of data resulting from genome sequencing in Molecular Biology is giving rise to an increasing interest in the development of algorithms for comparing genomes of related species. Particularly these data prompted research on mutational events acting on large portions of the chromosomes. Such events can be used to compare genomes for which the traditional methods of comparing DNA sequences are not conclusive. The field originated by the study of large mutations on chromosomes is known as *genome rearrangements*.

There are several mutational events affecting large fragments of genomes of organisms, including duplication, reversal, transposition (acting on a single chromosome), translocation, fusion, and fission (involving more than one chromosome). Each such event or combination of events gives rise to a theoretical problem of finding, given two genomes, the shortest series of events that transforms one genome into the other. We seek the shortest series because it has the largest likelihood of occurrence under a general principle of parsimony. Notice that in general more than one shortest series exists. The length of the shortest series is called the *distance* between the two genomes.

Chromosomes are usually represented as *permutations* of integers in a given range, each integer representing a gene. Sometimes the integers are signed to indicate the orientation of the gene. However, when the gene orientations are unknown or not relevant (as in the case of transpositions), the integers are unsigned.

In the last few years we have witnessed formidable advances in our understanding of genome rearrangements. A partial list of known results follows. With respect to the *reversal* event, Hannenhalli and Pevzner [6] presented the first polynomial time algorithm to find the distance, later improved on its running time by Berman and Hannenhalli [2], and Kaplan, Shamir, and Tarjan [7]. These results concern signed permutations. For the unsigned case, also involving reversals, Caprara [3] showed that finding the distance is NP-hard. Hannenhalli and Pevzner [5] studied a multichromosomal distance problem for signed genomes involving reversals, fusion, fission, and a specific form of translocation, producing a polynomial time algorithm in this case as well. Bafna and Pevzner [1] analyzed the problem with respect to transpositions, presenting several approximation algorithms, and leaving a number of open questions, among them the complexity of the problem and the diameter (largest possible distance between two permutations of size n). Gu, Peng, and Sudborough [4] gave approximation algorithms for the combination of events of reversal and transposition.

In this note we solve an open question posed by Bafna and Pevzner [1], regarding chromosome distance with respect to transpositions: we show that the distance between a permutation and its reverse (without complementation) is $\lfloor n/2 \rfloor + 1$, where n is the size of the permutations. Besides, we present an algorithm to compute an optimal series of transpositions.

2 Definitions

Chromosomes are represented by *permutations* of integers in the range $1..n$, where n is the number of genes of interest in the chromosome. For instance, (3 4 2 6 1 5) represents a chromosome with six genes. A *transposition* is an operation that transforms a permutation into another one, "cutting" a certain portion of the permutation and "pasting" it elsewhere in the same permutation. A transposition $\rho(i,j,k)$ is defined by three integers i, j, and k such that $1 \leq i < j \leq n+1$, $1 \leq k \leq n+1$, and $k \notin [i,j]$, in the following way. It "cuts" the portion between positions i and $j-1$, including the extremes, and "pastes" it just before position k. Thus, we can write

$$\rho(i,j,k) \cdot (\pi_1 \pi_2 \dots \pi_i \dots \pi_j \dots \pi_k \dots \pi_n) = \\ (\pi_1 \pi_2 \dots \pi_{i-1} \pi_j \dots \pi_{k-1} \pi_i \dots \pi_{j-1} \pi_k \dots \pi_n),$$

if $i < j < k$, and

$$\rho(i,j,k) \cdot (\pi_1 \pi_2 \dots \pi_k \dots \pi_i \dots \pi_j \dots \pi_n) = \\ (\pi_1 \pi_2 \dots \pi_{k-1} \pi_i \dots \pi_{j-1} \pi_k \dots \pi_{i-1} \pi_j \dots \pi_n),$$

if $k < i < j$. Notice that $\rho(i,j,k) = \rho(j,k,i)$ when $i < j < k$.

Given two permutations π and σ, the *transposition distance* or just *distance* between them is the minimum number t of transpositions $\varrho_1 \dots \varrho_t$ such that

$$\varrho_t \varrho_{t-1} \dots \varrho_1 \cdot \pi = \sigma.$$

We denote such distance by $d(\pi, \sigma)$. Because the inverse of a transposition is also a transposition, we have that $d(\pi, \sigma) = d(\sigma, \pi)$.

A powerful tool for studying the transposition distance is the *reality and desire diagram* of two permutations. Suppose we want to compute $d(\pi, \sigma)$. We construct this diagram writing the origin permutation π in the following way. Replace each integer i by a pair of points $-i$ and $+i$, in this order, and add two extra points, one called $+0$ at the beginning of the sequence, and one called $-(n+1)$ at the end of the sequence. Now draw oriented *reality* edges from $-\pi_1$ to $+0$, from $-\pi_{i+1}$ to $+\pi_i$, and from $-(n+1)$ to $+\pi_n$. Finally, draw oriented *desire* edges from $+0$ to $-\sigma_1$, from $+\sigma_i$ to $-\sigma_{i+1}$, and from $+\sigma_n$ to $-(n+1)$.

The diagram has exactly $n+1$ reality edges and the same number of desire edges. The idea is that reality edges indicate the situation as it is now, while desire edges indicate the situation sought. When reality equals desire in all edges, we have $\pi = \sigma$ and $d = 0$. Therefore, in a way, our goal is to apply transpositions so that reality becomes closer to desire. Figure 1 shows the diagram corresponding to a pair of permutations.

$$\pi = (8\ 5\ 1\ 4\ 3\ 2\ 7\ 6)$$

$$\sigma = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$$

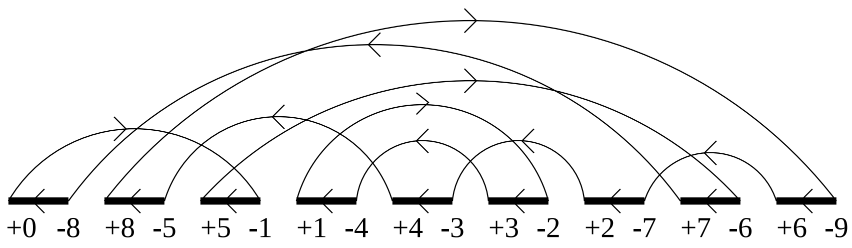

Fig. 1. Reality and desire diagram for two permutations, π and σ, as showed. In this figure, reality edges are represented by thick lines and desire edges by thin lines.

Bafna and Pevzner [1] made several useful results regarding the reality and desire diagram. One of them is that the diagram is composed of a number of cycles, with each cycle alternating between reality and desire edges. The *length* of a cycle is the number of reality edges in it (which is the same as the number of desire edges in it). One important remark follows.

Lemma 2.1 *The sum of the lengths of all cycles in any reality and desire diagram is always equal to* $n+1$.

Moreover, a transposition can affect the number of cycles in a very specific way, as the following lemma shows [1]. Denote $c(\pi, \sigma)$ the number of cycles in the diagram of π and σ.

Lemma 2.2 *For any permutations π and σ and any transposition ϱ we have*

$$c(\varrho \cdot \pi, \sigma) = c(\pi, \sigma) + x,$$

where $x = -2$, 0, or 2.

A transposition ϱ is called a -2-*move*, a *0-move*, or a *2-move* according to x being -2, 0, or 2 in the previous lemma. Since $c(\sigma, \sigma) = n + 1$, the maximum possible, we would like to perform 2-moves as much as possible.

If fact, a stronger statement can be made regarding the effect of a transposition on a diagram. Denote by $c_{odd}(\pi, \sigma)$ the number of cycles of odd length in the diagram of π and σ.

Lemma 2.3 *For any permutations π and σ and any transposition ϱ we have*

$$c_{odd}(\varrho \cdot \pi, \sigma) = c_{odd}(\pi, \sigma) + x,$$

where $x = -2$, 0, or 2.

From this lemma we have the following lower bound on the distance:

Theorem 2.1 *For any permutations π and σ we have*

$$d(\pi, \sigma) \geq \frac{(n+1) - c_{odd}}{2}$$

Now we make some definitions used in the following sections.

First we show a way to represent a cycle by its reality edges. We number the reality edges of the diagram assigning label i to a reality edge from π_{i+1} to π_i, with $0 \leq i \leq n$, so we label them from 1 to $n+1$. Let us consider a cycle C of size k, taking the reality edges in the order they appear in the cycle, $(i_1, \ldots, i_k)$. A cycle C can be represented in k possible ways, depending on the choice of the first reality edge. We will consider a *canonical representative of a cycle* C, taking the initial reality edge i_1 as the rightmost edge of C in π, that is, $i_1 = \max_{1 \leq t \leq k} i_t$. In the diagram of Figure 1 we have three cycles, with canonical representatives $C_1 = (9, 7, 5, 2)$, $C_2 = (8, 1, 3)$ and $C_3 = (6, 4)$.

Let us consider now three reality edges x, y, z belonging to the same cycle C in the diagram. C forces an order on x, y, z, and we have three possible representations of this order. We will choose as the *canonical representative of a triple* (x, y, z) the one starting from the rightmost reality edge $\max(x, y, z)$. A triple in the canonical order is *non-oriented* if $x > y > z$ and *oriented* if $y < z < x$. In the diagram of Figure 1 we have the following non-oriented triples: $(9, 7, 5)$, $(9, 7, 2)$, and $(7, 5, 2)$; and the oriented triple $(8, 1, 3)$.

Finally, we say that a cycle is *oriented* if it admits a 2-move, and *non-oriented* if there is no possible 2-moves acting on it. In the diagram of Figure 1 we have C_1 and C_3 non-oriented and C_2 oriented.

3 Computing the transposition distance

Given the permutations $\pi = (n\ n-1\ n-2\ \ldots\ 2\ 1)$ and $\tau = (1\ 2\ \ldots\ n-2\ n-1\ n)$ we want to compute the transposition distance $d(\pi, \tau)$. Of course, this distance will be the same for any pair consisting of a permutation and its reverse. Theorem 3.1 below establishes that $d(\pi, \tau) = 1$ if $n = 2$ and $d(\pi, \tau) = \lfloor \frac{n}{2} \rfloor + 1$ if $n > 2$.

However, in the proof of this theorem we need the following lemma, which can be easily proved. Bafna and Pevzner [1] mention part of this result in their work.

Lemma 3.1 *Let C be a cycle and (x, y, z) a triple of C in the canonical representation. Then we have*

$$(x, y, z) \text{ is oriented if and only if } \varrho(y, z, x) \text{ is a 2-move}$$

and

$$(x, y, z) \text{ is non-oriented if and only if } \varrho(y, z, x) \text{ is a 0-move}$$

Now we will state and prove the main theorem.

Theorem 3.1 *Given the permutations $\pi = (n\ n-1\ n-2\ \ldots\ 2\ 1)$ and $\tau = (1\ 2\ \ldots\ n-2\ n-1\ n)$, we have for $n \geq 2$*

$$d(\pi, \tau) = \begin{cases} 1 & \text{if } n = 2 \\ \lfloor \frac{n}{2} \rfloor + 1 & \text{if } n > 2 \end{cases}$$

Proof: From the work of Bafna and Pevzner [1], given $\pi = (n\ n-1\ n-2\ \ldots\ 2\ 1)$ we have

$$c_{odd}(\pi, \tau) = \begin{cases} 1 & \text{if } n \text{ is even} \\ 0 \text{ or } 2 & \text{if } n \text{ is odd} \end{cases}$$

Applying the lower bound given by Theorem 2.1 for $d(\pi, \tau)$ we have

$$\frac{(n+1) - c_{odd}}{2} = \begin{cases} \frac{(n+1)-1}{2} = \frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n+1}{2} \text{ or } \frac{n-1}{2} & \text{if } n \text{ is odd} \end{cases}$$

Notice that in order to attain the lower bound every transposition used must increase the number of odd cycles.

For π and τ as defined earlier in this section Bafna and Pevzner [1] proved the following upper bound

$$d(\pi, \tau) \leq \left\lfloor \frac{n}{2} \right\rfloor + 1$$

for all $n \geq 1$.

We have two cases:

1. When n is odd:

   – for the case of 0 odd cycles:

   $$\frac{n+1}{2} = \left\lfloor \frac{n}{2} \right\rfloor + 1$$

   In that case, the lower bound is exactly equal to the upper bound, and then $d(\pi, \tau) = \lfloor \frac{n}{2} \rfloor + 1$.

   – For the case of 2 odd cycles:

   $$\frac{n-1}{2} < \left\lfloor \frac{n}{2} \right\rfloor + 1$$

   and the gap is exactly 1. But here we have two non-oriented odd cycles. This implies that the next move cannot increase c_{odd}, and therefore we cannot reach the lower bound.

   So, when n is odd then we have $d(n\ n-1\ n-2\ \ldots\ 2\ 1) = \lfloor \frac{n}{2} \rfloor + 1$.

2. When n is even:

   $$\frac{n}{2} < \left\lfloor \frac{n}{2} \right\rfloor + 1$$

   and the gap is exactly 1. In this case, we have to prove that there is necessarily a transposition that will not increase c_{odd} during any transposition sequence that transforms π into τ.

   The first transposition is either a 0-move or a 2-move. We cannot apply a -2-move because this first diagram is formed by just one cycle.

   If we apply a 0-move, the unique odd cycle is transformed into another odd cycle, not increasing c_{odd}.

   So, we have to verify what happens if we apply a 2-move. We will show that any 2-move gives rise to a diagram with all cycles non-oriented. This will imply the result as follows. We have two possibilities. If the resulting diagram has one odd cycle and two even cycles, the first transposition did not increase c_{odd}. On the other hand, if we end up with three odd cycles, the second transposition of the series cannot increase c_{odd}, because all three cycles are non-oriented.

   Let us now study what happens when the first transposition is a 2-move. From Lemma 3.1 we know that every 2-move corresponds to an oriented triple in the unique cycle of the diagram. The order of reality edges in this cycle is such that all even labels appear together, in decreasing order, and all odd labels appear together, also in decreasing order. It follows that the possible 2-moves are of the form $\varrho(i, j, k)$ with i and j of opposite parity, k of same parity as i, and $1 \leq i < j < k \leq n+1$.

We now have to apply one such transposition and analyze the resulting diagram. Because $i < j < k$, we have

$$\begin{aligned}&\varrho(i,j,k) \cdot (n\ n-1\ n-2\ \ldots\ 2\ 1) = \\ &\quad (\ n-1\ \ldots\ n-i+2\quad n-j+1\ \ldots\ n-k+2 \\ &\quad n-i+1\ \ldots\ n-j+2\quad n-k+1\ \ldots\ 2\ 1).\end{aligned}$$

Figure 2 shows examples of such 2-moves.

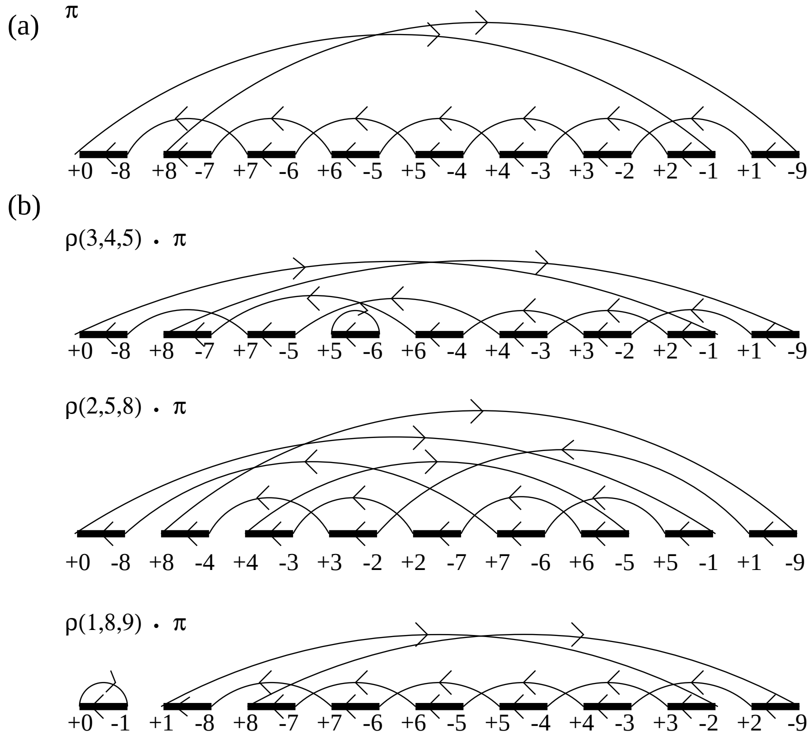

Fig. 2. This figure shows the diagram created by a strictly decreasing sequence with respect to the identity, and the diagrams created by some possible 2-moves applied to the first diagram. (a) The diagram created by the decreasing cycle $\pi = (8\ 7\ 6\ 5\ 4\ 3\ 2\ 1)$ with respect to $\tau = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$. (b) The diagrams created by some transpositions applied to π as indicated.

The important fact here is that, because of the opposite parity of j and k, and of i and j, the cycle involving reality edge $(n-i+1, n-k+2)$ is non-oriented. Likewise, the cycle involving reality edge $(n-k+1, n-j+2)$ is non-oriented because of the opposite parity of i and j (regardless of the parity of k), and the cycle involving reality edge $(n-j+1, n-i+2)$ is also non-oriented because j and k have opposite parity (regardless of the parity of i). Thus, in any case we end up with three non-oriented cycles, which proves the theorem.

$\square$

4 An algorithm to compute $d(\pi, \tau)$

We show now an algorithm to compute the transposition distance between a strictly decreasing sequence with respect to the identity. Note that the algorithm runs without using the reality and desire diagram. Instead, it uses an explicit series of transpositions that work in the case treated in this article. As the series has length $\lfloor n/2 \rfloor + 1$, the results in the previous section guarantee that it is a shortest series.

Algorithm

Input: $n > 2$, $\pi = (n\ n-1\ \ldots\ 2\ 1)$

Output: $t = d(\pi, \tau)$ and $\varrho_1, \varrho_2, \ldots, \varrho_t$

begin

1. $\pi_1 \leftarrow \varrho_1(1, \lceil \frac{n}{2} \rceil, n) \cdot \pi$
2. $t \leftarrow 1$
3. if n is even then

   $t \leftarrow t + 1$
   $\pi_2 \leftarrow \varrho_t(\frac{n}{2}, \frac{n}{2} + 1, n + 1) \cdot \pi_1$
   $k \leftarrow 1$
   $p \leftarrow 1$
4. if n is odd then

   $k \leftarrow 0$
   $p \leftarrow 0$
5. while $k < \lfloor \frac{n}{2} \rfloor$ do

   $t \leftarrow t + 1$
   $\pi_t \leftarrow \varrho_t(\lfloor \frac{n}{2} \rfloor - k, \lfloor \frac{n}{2} \rfloor - k + 2, n + 1 - k + p) \cdot \pi_{t-1}$
   $k \leftarrow k + 1$
6. return t, $\varrho_1, \varrho_2, \ldots, \varrho_t$

end

The four initial steps create, from the initial permutation, a new permutation with two decreasing subsequences on its left extremity, and an increasing sequence, on its right end. If n is even then we have

$$\pi_2 = (\ \frac{n}{2} + 1\ \ \frac{n}{2}\ \ldots\ 3\ )\ (\ n\ n-1\ \ldots\ \frac{n}{2} + 2\ )\ (\ 1\ 2\ )$$

Note that the decreasing subsequences have $\lceil \frac{n}{2} \rceil - 1$ elements each. We marked the subsequences with parenthesis.

If n is odd then we have

$$\pi_1 = (\ \frac{n+1}{2}\ \ldots\ 3\ 2\ )\ (\ n\ n-1\ \ldots\ \frac{n+1}{2}+1\ )\ (\ 1\ )$$

Analogously, in this case the first two subsequences also have $\lceil \frac{n}{2} \rceil - 1$ elements each.

The loop in step 5 moves both the last element of first subsequence and the first element of second subsequence to the right end of the permutation, where two other subsequences are being increased as the algorithm runs. Generically, if n is even then we have, after $k-1$ iterations of the loop,

$$\pi_{k+1} = (\frac{n}{2}+1\ \frac{n}{2}\ \ldots\ k+2)\ (n-k+1\ \ldots\ \frac{n}{2}+2)\ (1\ 2\ \ldots\ k+1)\ (n-k+2\ \ldots\ n).$$

If n is odd then we have, after k iterations,

$$\pi_{k+1} = (\frac{n+1}{2}\ \ldots\ k+2)\ (n-k\ \ldots\ \frac{n+1}{2}+1)\ (1\ 2\ \ldots\ k+1)\ (n-k+1\ \ldots\ n).$$

So this algorithm correctly transforms the permutation in its inverse, using transpositions. Also, the algorithm runs in $O(\lfloor \frac{n}{2} \rfloor + 1)$ steps, for $n > 2$.

Figure 3 shows examples of this algorithm executions for the decreasing sequences for $n = 6$ and $n = 7$, with respect to the identity.

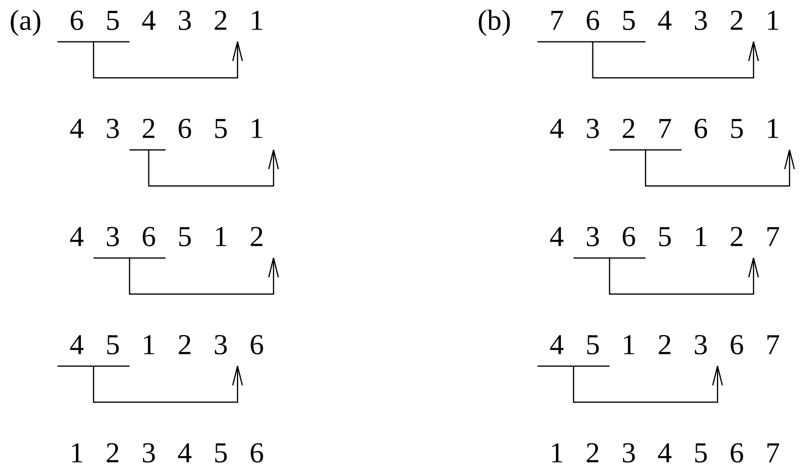

Fig. 3. This figure shows two executions of the algorithm. (a) Example with n even. (b) Example with n odd.

5 Conclusions

We demonstrated that the transposition distance between a permutation and its reverse (without complementation) is $\lfloor \frac{n}{2} \rfloor + 1$ for all $n > 2$, where n is the size of the permutation. We conjecture that this is in fact the value of the transposition diameter.

We also presented an algorithm to find an optimal series of sorting transpositions for the case studied.

References

1. V. Bafna and P. Pevzner. Sorting by transpositions. In *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms - SODA'95*, pages 614–623, 1995.
2. P. Berman and S. Hannenhalli. Fast sorting by reversals. In *Proc. 7th Annual Symposium on Combinatorial Pattern Matching - CPM'96*, pages 168–185, 1996.
3. A. Caprara. Sorting by reversals is difficult. In *Proc. 1st Annual International Conference on Computational Molecular Biology - RECOMB'97*, pages 75–83, 1997.
4. Qian-Ping Gu, Shietung Peng, and Hal Sudborough. Approximating algorithms for genome rearrangements. In *Proc. 7th Worshop on Genome Informatics - GIW'96*, 1996.
5. S. Hannenhalli and P. Pevzner. Tranforming men into mice (polynomial algorithmic for genomic distance problem). In *Proc. 36th Annual IEEE Symposium on Foundations of Computer Science - FOCS'95*, pages 581–592, 1995.
6. S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 20th Annual ACM Symposium on Theory of Computing - STOC'95*, pages 178–189, 1995.
7. H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. In *Proc. 8th Annual ACM-SIAM Symposium on Discrete Algorithms - SODA'97*, 1997.

This article was processed using the LaTeX 2_ε macro package with CUP_CS class