INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Genome Rearrangements Distance by Fusion, Fission, and Transposition is Easy

J. Meidanis and Z. Dias

Technical Report - IC-01-07 - Relatório Técnico

July - 2001 - Julho

Genome Rearrangements Distance by Fusion, Fission, and Transposition is Easy

Joao Meidanis[*] Zanoni Dias [†]

Abstract

Given two genomes represented as circularly ordered sequences of genes, we show a polynomial time algorithm for the minimum weight series of fusion, fissions, and transpositions (with transpositions weighing twice as much as fusions and fissions) that transforms one genome into the other. The algorithm is based on classical results of permutation group theory and is the first polynomial result for a genome rearrangement problem involving transpositions. It has been observed in real biological instances that transpositions occur with about half the frequency of reversals. Although we are not using reversals in this study, this observation motivated the double weight assigned to transpositions.

1 Introduction

With the advent of fast sequencing techniques, we are witnessing today a spectacular increase in the quantity of molecular data (DNA and protein sequences). More than 40 complete microbial genomes are known now, and about 170 others are in progress [6]. The great challenge we face now is how to process this huge amount of data and extract from it relevant biological information that could help design drugs, understand life and disease, improve crops, and so on. One way to structure this information is by comparative genomics, where we analyze data coming from distinct species and learn from the similarities and differences in related genomes. Among the several proposed ways of comparing genomes, the area of **genome rearrangements** has received a lot of attention recently. In this area, very large DNA molecules (usually entire chromosomes or large pieces of chromosomes) are investigated with respect to the relative order of genes in them. The goal is to determine a rearrangement distance, which is the minimum number of rearrangement events that could explain the differences between two such DNA molecules.

Many different events have been considered. Reversals, transpositions and translocations are the best studied ones from a theoretical point of view, although in practice events such as duplications and deletions are at least as important. As far as reversals are concerned, Hannenhalli and Pevzner presented the first polynomial time algorithm [8], subsequently improved by Kaplan, Shamir, and Tarjan [9]. Caprara showed that the reversal problem is NP-hard if we disregard the orientation

[*]University of Campinas, Institute of Computing, P.O.Box 6167, 13084-971, Campinas, Brazil. Research supported in part by CNPq and FAPESP

[†]University of Campinas, Institute of Computing, P.O.Box 6167, 13084-971, Campinas, Brazil. Research supported in part by FAPESP

of genes [3]. Hannenhalli and Pevzner also solved in polynomial time a multi-chromosomal problem involving translocation, fusion and fission [7]. Bafna and Pevzner [1] studied the transposition distance between two linear unsigned chromosomes, presenting several approximation algorithms. The best one has an approximation factor of 1.5 and runs in $O(n^2)$ time, but it is very complex. Christie [5] devised an alternative 1.5-approximation algorithm that runs in $O(n^4)$ time. Christie [4] also proposed and solved the problem of block-interchange distance. A block-interchange can be viewed as a generalization of a transposition. (In a block-interchange two non-intersecting substrings of any length are swapped in the permutation. In a transposition the substrings must be adjacent.) Transposition distance seems to be a harder problem, that has eluded researchers for many years now. Its computational complexity is still unknown.

We show a polynomial time algorithm for the minimum weight series of fusion, fissions, and transpositions (with transpositions weighing twice as much as fusions and fissions) that transforms one genome into the other. The algorithm is based on classic results of permutation group theory and it is the first polynomial result for a genome rearrangement problem involving transpositions. It has been observed [2] in real biological instances that transpositions occur with about half the frequency of reversals.

In the following sections we present definitions, the main result of this work, proof sketches and conclusions and plans for future work.

2 Definitions, Modeling, and Results

A **permutation** in group theory is a one-to-one mapping from a set E into itself. We will use permutations to represent genomes with circular chromosomes. The standard notation [10] for permutations is to represent in parenthesis an element followed by its successive images. For instance, if $E = \{1, 2, 3, 4, 5\}$ the permutation α such that

$$\alpha(1) = 3, \alpha(2) = 5, \alpha(3) = 2, \alpha(4) = 1, \alpha(5) = 4$$

is represented as

$$(1\ 3\ 2\ 5\ 4).$$

The representation is not unique since we could have started at an element other than 1: $(2\ 5\ 4\ 1\ 3)$, $(5\ 4\ 1\ 3\ 2)$, etc. are all equivalent.

In our model, the set E is the set of genes of the genome and the permutation indicates how genes follow each other in the chromosomes. Only circular chromosomes can be represented in this way, but an easy translation of some results from circular to linear chromosomes exists [11].

Permutations can represent multi-chromosomal genomes. For instance, if $E = \{1, 2, 3, 4, 5, 6, 7\}$ the permutation $(1\ 5\ 4\ 3)(2\ 7\ 6)$ represents the genome depicted in Figure 1.

An element x if fixed under a permutation α when $\alpha(x) = x$. Fixed elements can be omitted in the parenthesized notation for permutations. For instance, if α is such that

$$\alpha(1) = 1, \alpha(2) = 3, \alpha(3) = 4, \alpha(4) = 2, \alpha(5) = 5,$$

then we can write $\alpha = (2\ 3\ 4)$, or $\alpha = (2\ 3\ 4)(1)$ or $\alpha = (1)(2\ 3\ 4)(5)$. The missing elements are implicitly understood as fixed. The **support** of a permutation α is the set of elements not fixed

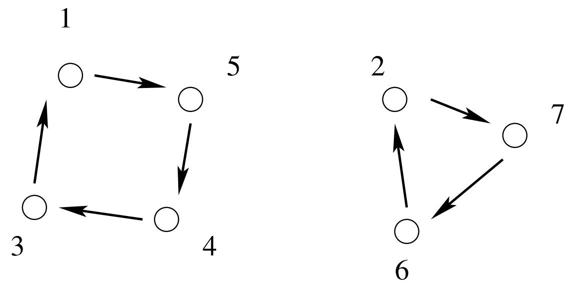

Figure 1: A multi-chromosomal genome.

by α. In the preceding example, we have $Supp(\alpha) = \{2, 3, 4\}$. The identity permutation, denoted simply by 1 (without parenthesis), fixes all elements and has empty support.

Permutations can be composed as mappings. This defines a product of permutations. For instance, if $E = \{1, 2, 3, 4, 5, 6\}$, $\alpha = (2\ 3\ 4)$ and $\beta = (3\ 1\ 5\ 2\ 6\ 4)$ we have $\alpha\beta$ defined as $\alpha\beta(x) = \alpha(\beta(x))$ for all $x \in E$, and therefore $\alpha\beta = (1\ 5\ 3)(2\ 6)$. Any two permutations over the same set can composed in this way. The operation is associative, the identity permutation is the identity element, and every permutation α has an inverse α^{-1} [10].

Composition of permutations is important in the context of genome rearrangements for at least two reasons:

- Some permutations ρ of small support can be viewed as re-arrangement events: $\rho\pi$ is then the result of event ρ acting on genome π.
- Given two genomes σ and π, the product $\sigma\pi^{-1}$describes in some sense the "differences" between the two genomes.

For instance, fusions and fissions can be seen as permutations of support size 2, and transpositions can be seen as permutations of support size 3, as we will see shortly. In addition, our main result establishes that the distance between two genomes can be computed as n minus the number of cycles in $\sigma\pi^{-1}$ (see Section 2.1 for a formal definition of cycles).

2.1 Orbits and Cycles

Any permutation can be written in a unique way as the product of cycles disjoint support (disjoint cycles). To understand cycles we need first the definition of orbit. An orbit can be defined intuitively as a set of the form $x, \alpha(x), \alpha^2(x), \ldots$ for some element x. Since we are dealing with finite sets, orbits are always finite, that is, there is a positive integer k such that $\alpha^k(x) = x$. A more formal definition is given below.

Definition 2.1 *An* ***orbit*** *of a permutation α is a minimal set of elements A such that for any two elements $x, y \in A$ there is an integer k such that $\alpha^k(x) = y$.*

For instance, if $E = \{1, 2, 3, 4, 5, 6\}$ and $\alpha = (2\ 3\ 5)(1\ 4)$ then the orbits are $\{2, 3, 5\}$, $\{1, 4\}$, and $\{6\}$. Restricting α to one of its orbits we obtain what is called a cycle of α. Formally, we have the following definition.

Definition 2.2 *A **cycle** of a permutation α is a permutation β such that there is an orbit A of α with*

$$\beta(x) = \begin{cases} \alpha(x) & \textit{if } x \in A \\ x & \textit{if } x \notin A \end{cases}$$

For instance, if $E = \{1, 2, 3, 4, 5, 6\}$ and $\alpha = (2\ 3\ 5)(1\ 4)$ then the cycles of α are $(2\ 3\ 5)$, $(1\ 4)$, and (6). A permutation that is a cycle of some permutation is called simply a **cycle**. The **size** of the cycle is the number of elements of its non-singleton orbit if there is one, or 1 if all orbits are singletons. A cycle of size k is also called a k**-cycle**.

It is important to note here a potentially confusion terminology used in the literature. The term "transposition" is used in permutation group theory meaning 2-cycle. The same term "transposition" is used in biology meaning a block move in a genome. Unfortunately, in both areas the term "transposition" is well-established and very unlikely to change. We had to make a choice in this paper, and decided to keep the biological meaning. Whenever we need to refer to the group theoretical meaning of "transposition" we will use the term "2-cycle" instead.

2.2 Rearrangement Events

We will define now the rearrangement events that are the main subject of this paper: fusion, fission, and transposition. Intuitively, a fusion joins two chromosomes into one; a fission breaks a chromosomes into two; and a transposition moves a block of consecutive genes from our place into another in the same chromosome.

Formally, fusions and fissions correspond to 2-cycles. Given a 2-cycle $\rho = (x\ y)$ and a permutation π, we have the following classical results:

- if x and y are in the same cycle of π, then in $\rho\pi$ this cycle is broken into two cycles, one containing x and the other y (among others elements). The remaining cycles of π are left unchanged.

- if x and y are in the distinct cycle of π, then in $\rho\pi$ these two cycles are joined into one. The remaining cycles of π are left unchanged.

These classical results show that 2-cycles correctly model fusions and fissions, because the cycles of a permutation correspond to circular chromosomes of a genome. However, notice that $\rho = (x\ y)$ can act as a fusion for some genomes and as a fission for others. Therefore, being a fusion (or a fission) is not an intrinsic property of ρ but rather depends also on the genome π on which ρ is being applied. Nevertheless, all fusions and fissions are captured by 2-cycles.

Formally, transpositions correspond to 3-cycles. Again, a 3-cycle is not intrinsically a transposition, but rather its transposition status depends on the particular genome on which it is being applied. More specifically, a 3-cycle $\rho = (x\ y\ z)$ is a transposition when it acts on a genome π where the elements x,y,z are all in the same cycle and appear in this order in the cycle. For instance, if $\rho = (7\ 3\ 2)$ and $\pi = (7\ 1\ 5\ 3\ 2\ 6\ 4)$ then $\rho\pi = (7\ 1\ 5\ 2\ 6\ 4\ 3)$ and ρ models a transposition.

In the next section we define our problem formally. Intuitively, given two genomes represented by permutations π and σ, we want to find a minimum weight series of events leading from π to σ. The main result of this paper is that this problem is solvable in polynomial time. This follows from a classical result in permutation group theory as we will see shortly.

3 Proof Sketches

In this section we will sketch some of the proofs need in our main result. We begin with some additional definitions, continue with a formal statement of the main result, and finish with the proof sketches.

Definition 3.1 *A permutation ρ is a* ***valid event*** *for another permutation (genome) π when either:*

1. *ρ is a 2-cycle, or*
2. *ρ is a transposition when applied to π.*

In case (1) the ***weight*** *of ρ, denoted by $w(\rho)$, is equal to 1; in case (2), $w(\rho) = 2$.*

Notice that if ρ is a valid event for π then ρ^{-1} is a valid event for $\rho\pi$. In other words, valid events can be "undone" by other valid events of the same weight.

Definition 3.2 *An ordered sequence of permutations $(\rho_1, \rho_2, \ldots, \rho_k)$ is a* ***series of valid events leading from*** *π* ***to*** *σ when:*

- *each ρ_i is a valid event for $\rho_{i-1}\rho_{i-2}\ldots\rho_2\rho_1\pi$, and*
- *$\rho_k\rho_{k-1}\ldots\rho_2\rho_1\pi = \sigma$*

We are interested in such a series with minimum total weight $\Sigma_{i=1}^k w(\rho_i)$. The minimum value of $\Sigma_{i=1}^k w(\rho_i)$ is called the distance between π and σ.

Definition 3.3 *For a permutation α, let $c(\alpha)$ denote the number of orbits of α. For two permutations (genomes) π and $sigma$, let $c(\pi, \sigma)$ denote the number of orbits of $\sigma\pi^{-1}$.*

For instance, $c(\pi, \pi) = n$ for any genome π, where $n = |E|$ is the number of genes.

Definition 3.4 *Given two permutations (genomes) π and σ and a valid event ρ for π, denote by $\Delta c(\rho, \pi, \sigma)$ the value:*

$$\Delta c(\rho, \pi, \sigma) = c(\rho\pi, \sigma) - c(\pi, \sigma)$$

The quantity $\Delta c(\rho, \pi, \sigma)$ is the increase in the number of orbits of $\sigma\pi^{-1}$ when π is replaced by $\rho\pi$. If this number is positive, $\rho\pi$ is "closer" to σ than π was.

Definition 3.5 *A valid event ρ for π is* ***good with respect to*** *σ when:*

$$\Delta c(\rho, \pi, \sigma) = w(\rho)$$

Our main result can be stated as follows.

Theorem 3.1 *Given two permutations (genomes) π and σ, the distance between them is $n - c(\pi, \sigma)$.*

The proof relies on the following two lemmas.

Lemma 3.1 *For any series of valid events (ρ_1, ρ_2, ..., ρ_k) leading from π to σ we have:*

$$\Sigma_{i=1}^{k} w(\rho_i) \geq n - c(\pi, \sigma)$$

with equality if and only if each ρ_i is good for $\rho_{i-1}\rho_{i-2} \ldots \rho_2\rho_1\pi$ with respect to σ.

Proof Sketch: Suppose that

$$\rho_k\rho_{k-1} \ldots \rho_2\rho_1\pi = \sigma \tag{1}$$

Each ρ_i is a 2-cycle or a 3-cycle, but any 3-cycle can be written as a product of two 2-cycles. Replacing every 3-cycle of equation (1) by a product of 2-cycles, and multiplying by π^{-1} on the right, we have:

$$c_{k'}c_{k'-1} \ldots c_2c_1 = \sigma\pi^{-1}$$

where each c_i is a 2-cycle. The number os 2-cycles involved is just:

$$k' = \Sigma_{i=1}^{k} w(\rho_i)$$

since 2-cycles have weight 1 and 3-cycles weigh 2. But there is a classical result that says that if a permutation α can be written as a product of 2-cycles, the number of 2-cycles is at least $n - c(\alpha)$ [10].

Therefore,

$$k' \geq n - c(\alpha)$$

or

$$\Sigma_{i=1}^{k} w(\rho_i) \geq n - c(\pi, \sigma). \tag{2}$$

To analyse the cases where there is equality in this formula, we need another classical result that says that, with the weights we are uding, we have always the following:

$$c(\sigma\pi^{-1}\rho_1^{-1} \ldots \rho_i^{-1}) - c(\sigma\pi^{-1}\rho_1^{-1} \ldots \rho_{i-1}^{-1}) \leq w(\rho_i).$$

Adding up these inequalities in the case where (ρ_1, ρ_2, ..., ρ_k) is a series of valid events leading from π to σ we have that equality holds in (2) only if every ρ_i is good for $\rho_{i-1}\rho_{i-2} \ldots \rho_2\rho_1\pi$ with respect to σ. ■

Lemma 3.2 *Given two distinct permutations (genomes) π and σ, there is always a good event for π with respect to σ*

Proof Sketch: Since $\pi \neq \sigma$ we have $\sigma\pi^{-1} \neq 1$ and there is a k-cycle in $\sigma\pi^{-1}$ with $k \geq 2$. Choose x and y as two distinct elements in this k-cycle. We claim that the event $\rho = (x\ y)$ is valid for π and is a good event with respect to σ.

The event ρ is valid for π since it is a 2-cycle, and therefore is either a fusion or a fission in π.

It is a good event with respect to σ because of the following argument. By choice we know that ρ splits a cycle of $\sigma\pi^{-1}$ into two. Therefore,

$$c(\sigma\pi^{-1}\rho) = c(\sigma\pi^{-1}) + 1$$

In addition, ρ is a 2-cycle so $\rho = \rho^{-1}$. But then:

$$\begin{aligned}
\Delta(\rho, \pi, \sigma) &= c(\rho\pi, \sigma) - c(\pi, \sigma) \\
&= c(\sigma\pi^{-1}\rho^{-1}) - c(\sigma\pi^{-1}) \\
&= c(\sigma\pi^{-1}) + 1 - c(\sigma\pi^{-1}) = 1 = w(\rho)
\end{aligned}$$

■

This suggests the following algorithm for finding the distance and an optimal series of events leading from π to σ.

FUSION, FISSION, AND TRANSPOSITION DISTANCE()

1 Input π, σ
2 $d \leftarrow 0$
3 **while** $\pi \neq \sigma$
4 **do** $\rho \leftarrow$ any valid event for π which is good with respect to σ
5   output ρ
6     $\pi \leftarrow \rho\pi$
7     $d \leftarrow d + 1$
8 output d

The complexity of this algorithm is $O(n^2)$, because the main loop is executed at most n times and consumes at most n steps per iteration. Lemma 3.2 guarantees that step 4 is well defined.

A word about the choice of ρ in step 4: Lemma 3.2 tells us that it suffices to take $\rho = (x\ y)$ with x, y in the same orbit of $\sigma\pi^{-1}$. It is instructive to have a similar criterium for transpositions. As stated in the algorithm, $\rho = (x\ y\ z)$ must be valid for π and good for π with respect to σ, for the particular values of variables π and σ at the moment. In Section 2.2 we saw that ρ is valid for π when x, y, z are in the same orbit of π and appear in this order in the corresponding cycle of π. To be good for π with respect to σ, x, y, z should be in the same orbit of $\sigma\pi^{-1}$ and appear in this order in the corresponding cycle of $\sigma\pi^{-1}$.

4 Conclusions

We have shown how a classical result on permutation groups leads to a polynomial time algorithm for weighted genome rearrangement distance involving fusions (with weight 1), fissions (with weight 1), and transpositions (with weight 2). We observe that the algorithm remains valid if transpositions have any weight greater than 2. This is the first complexity result for a rearrangement problem involving transpositions. We hope this result can be extended to more general problems, involving other events, arbitrary weights, and signed genomes.

References

[1] V. Bafna and P. A. Pevzner. Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224–240, May 1998.

[2] M. Blanchette, T. Kunisawa, and D. Sankoff. Parametric genome rearrangement. *Journal of Computational Biology*, 172:11–17, 1996.

[3] A. Caprara. Sorting permutations by reversals and eulerian cycle decompositions. *SIAM Journal on Discrete Mathematics*, 12(1):91–110, February 1999.

[4] D. A. Christie. Sorting permutations by block-interchanges. *Information Processing Letters*, 60(4):165–169, November 1996.

[5] D. A. Christie. *Genome Rearrangement Problems*. PhD thesis, Glasgow University, 1998.

[6] TIGR The Institute for Genomic Research, May 2001.

[7] S. Hannenhalli and P. A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science (FOCS'95)*, pages 581–592, Los Alamitos, USA, October 1995. IEEE Computer Society Press.

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

[9] H. Kaplan, R. Shamir, and R. E. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM Journal on Computing*, 29(3):880–892, June 2000.

[10] S. MacLane and G. Birkhoff. *Algebra*. The Macmillan Company, London, sixth printing edition, 1971.

[11] J. Meidanis and Z. Dias. An alternative algebraic formalism for genome rearrangements. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics: Empirical and Analyitical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*, pages 213–223. Kluwer Academic Publishers, Le Chantecler, Canada, September 2000. DCAF'2000.