Transforming Men into Mice
(polynomial algorithm for genomic distance problem)

Sridhar Hannenhalli[1] **and Pavel A. Pevzner**[1,*]

Department of Computer Science and Engineering
Institute of Molecular Evolutionary Genetics*
The Pennsylvania State University
University Park, PA 16802

> Then Puss said, "I understand that you have magical powers, that you can change yourself into any kind of animal... But, it must be easy to turn yourself into something huge. However, it must be impossible to turn into something very, very small - like a mouse".
>
> Brothers Grimm, *Puss N Boots*

Abstract

Many people (including ourselves) believe that transformations of humans into mice happen only in fairy tales. However, despite some differences in appearance and habits, men and mice are genetically very similar. In the pioneering paper, Nadeau and Taylor, 1984 estimated that surprisingly few genomic rearrangements (178 ± 39) happened since the divergence of human and mouse 80 million years ago. However, their analysis is non-constructive and no rearrangement scenario for human-mouse evolution has been suggested yet. The problem is complicated by the fact that rearrangements in multi-chromosomal genomes include inversions, translocations, fusions and fissions of chromosomes, a rather complex set of operations. As a result, at the first glance, a polynomial algorithm for the genomic distance problem with all these operations looks almost as improbable as the transformation of a (real) man into a (real) mouse. We prove a duality theorem which expresses the genomic distance in terms of easily computable parameters reflecting different combinatorial properties of sets of strings. This theorem leads to a polynomial-time algorithm for computing most parsimonious rearrangement scenarios. Based on this result and the latest comparative physical mapping data we have constructed a scenario of human-mouse evolution with 131 reversals/translocations/fusions/fissions. A combination of the genome rearrangement algorithm with the recently proposed experimental technique called ZOO-FISH suggests a new constructive approach to the 100-year old problem of reconstructing mammalian evolution.

[1]This work is supported by NSF Young Investigator award CCR-9457784 and NIH grant 1R01 HG00987. Authors' E-mail addresses: *hannenha@cse.psu.edu*, *pevzner@cse.psu.edu*

1 Introduction

When the Brothers Grimm described a transformation of a man into a mouse they could hardly anticipate that two centuries later human and mouse will be the most genetically studied mammals. Man-mouse comparative physical mapping started 20 years ago and currently more than 1300 pairs of homologous genes are mapped in these species. As a result, biologists found that the related genes in human and mouse are not chaotically distributed over the genomes but instead form "conserved blocks" (*synteny groups*).

Current comparative mapping data indicate that both human and mouse genomes are combined from approximately 150 blocks which are "shuffled" in human as compared to mouse (Copeland et al., 1993). Shuffling of blocks happens quite rarely (roughly once in a million years) thus making it possible to reconstruct a rearrangement scenario of human-mouse evolution. Below we present the combinatorial formulation of the problem.

In the model we consider, every *gene* is represented by an *identification* number (positive integer) and an associated *sign* ("+" or "−") reflecting the *direction* of the gene. A *chromosome* is defined as a *sequence* of genes, while a *genome* is defined as a *set* of chromosomes. Given two genomes Π and Γ with the same set of genes, we are interested in a most parsimonious scenario of *evolution* of Π into Γ, i.e. the shortest sequence of rearrangements (defined below) to transform Π into Γ. Fig. 1 illustrates 4 rearrangement events transforming one genome into another.

Let $\Pi = \{\pi(1), \ldots, \pi(N)\}$ be a genome consisting of N chromosomes and let $\pi(i) = (\pi(i)_1 \ldots \pi(i)_{n_i})$, n_i being the number of genes in the i^{th} chromosome. Every chromosome π can be viewed either from "left to right" (i.e. as $\pi = (\pi_1 \ldots \pi_n)$) or from "right to left" (i.e. as $-\pi = (-\pi_n \ldots -\pi_1)$) leading to two equivalent representations of the same chromosome. From this perspective a 3-chromosomal genome $\{\pi(1), \pi(2), \pi(3)\}$ is equivalent to $\{\pi(1), -\pi(2), \pi(3)\}$ or $\{-\pi(1), -\pi(2), -\pi(3)\}$, i.e. the *directions* of chromosomes are irrelevant. The four most common elementary rearrangement events in multi-chromosomal genomes are *reversals*, *translocations*, *fusions* and *fissions* as defined below.

Let $\pi = (\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_j\pi_{j+1} \ldots \pi_n)$ be a chromosome and $1 \leq i \leq j \leq n$. A *reversal* $\rho(\pi, i, j)$ rearranges the genes *inside* the chromosome and transforms π into a chromosome $(\pi_1 \ldots \pi_{i-1} - \pi_j \ldots - \pi_i\pi_{j+1} \ldots \pi_n)$. Let $\pi = (\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_n)$ and $\sigma = (\sigma_1 \ldots \sigma_{j-1}\sigma_j \ldots \sigma_m)$ be two chromosomes and $1 \leq i \leq n+1$, $1 \leq j \leq m+1$. A *translocation* $\rho(\pi, \sigma, i, j)$ exchanges genes *between* two chromosomes π and σ and transforms them into chromosomes $(\pi_1 \ldots \pi_{i-1}\sigma_j \ldots \sigma_m)$ and $(\sigma_1 \ldots \sigma_{j-1}\pi_i \ldots \pi_n)$ with $(i-1)+(m-j+1)$ and $(j-1)+(n-i+1)$ genes respectively. We denote $\Pi \cdot \rho$ as the genome obtained from Π as a result of a rearrangement ρ. Given genomes Π and Γ, the *genomic sorting* problem is to find a series of rearrangements (reversals and translocations) $\rho_1, \ldots, \rho_t$ such that $\Pi \cdot \rho_1 \cdots \rho_t = \Gamma$ and t is minimum. We call t the *genomic distance* between Π and Γ and denote it as $d(\Pi, \Gamma)$.

We distinguish between *internal* reversals which do not involve ends of the chromosomes (i.e. the reversals $\rho(\pi, i, j)$ of a n-gene chromosome with $1 < i \leq j < n$) and *prefix* reversals involving ends of the chromosomes (i.e. either $i = 1$ or $j = n$). Also note that a translocation $\rho(\pi, \sigma, n+1, 1)$ concatenates the chromosomes π and σ resulting in a chromosome $\pi_1 \ldots \pi_n\sigma_1 \ldots \sigma_m$ and an *empty* chromosome $\emptyset$. This special translocation leading to a reduction in the number of (non-empty) chromosomes is known in molecular biology as a *fusion*. The translocation $\rho(\pi, \emptyset, i, 1)$ for $1 < i \leq n$ "breaks" a chromosome π into two chromosomes $(\pi_1 \ldots \pi_{i-1})$ and $(\pi_i \ldots \pi_n)$. This translocation leading to an increase in the number of (non-empty) chromosomes is known as a *fission*. A translocation is *internal* if it is neither a fusion, nor a fission. Fusions and fissions are rather common in mammalian evolution. For example, the only difference in the overall genome organization of human and chimpanzee is the fusion of chimpanzee chromosomes 12 and 13 into human chromosome 2.

Genome rearrangements provide a multitude of challenges for computer scientists; see Pevzner and Waterman, 1995 for a review of combinatorial problems motivated by genome rearrangements. Kececioglu and Sankoff, 1993 suggested the first approximation algorithm to analyze genome rearrangements in uni-chromosomal genomes (reversals only). The problem was further studied by Bafna and Pevzner, 1993, Kececioglu and Sankoff, 1994, Kececioglu and Gusfield, 1994, Bafna and Pevzner, 1995a and Hannenhalli and Pevzner, 1995. See also Sankoff et al.,1992, Bafna and Pevzner, 1995b and Hannenhalli et al., 1995 for biological applications, as well as Gates and Papadimitriou, 1979, Even and Goldreich, 1981, Jerrum, 1985, Aigner and West, 1987, Cohen and Blum, 1993, Heydari and Sudborough, 1993 for studies of related combinatorial problems.

Kececioglu and Ravi, 1995 made the first attempt to analyze rearrangements of multi-chromosomal genomes by devising an approximation algorithm for genomes evolving by internal reversals and internal translocations. Recently Hannenhalli, 1995 devised a polynomial algorithm for the case when only internal translocations are allowed. All these algorithms address the case when both genomes contain the same number of chromosomes. This is a serious limitation since different organisms (in particular human and mouse) have different numbers of chromosomes. From this perspective, every realistic model of genome rearrangements should include fusions and fissions. Moreover, despite some shortcomings (see the last section for a discussion on centromeres) the reversals/translocations/fusions/fissions model adequately reflects the existing biological challenges (Joe Nadeau, personal communication). It turns out that fusions and fissions present a major difficulty in analyzing genome rearrangements: the problem of devising an *approximation* algorithm for genome rearrangements with reversals/translocations/fusions/fissions was raised by Kececioglu and Ravi, 1995. This paper presents an *exact* polynomial algorithm for this problem.

Every analysis of rearrangements involves revealing "hidden obstacles" which prevent a "fast" transformation of one genome into another (like dual variables in linear programming). Sixty years ago Dobzhansky and Sturtevant, 1938 already used a notion of breakpoint, which is the most obvious example of such an obstacle. Based on this notion, Kececioglu and Sankoff, 1994, devised a 2-approximation algorithm for uni-chromosomal genomes. However, an estimate of genomic distance in terms of breakpoints is very inaccurate. Bafna and Pevzner, 1993 revealed another obstacle (cycle decomposition) which significantly improves the bounds for genomic distance. Finally, Hannenhalli and Pevzner, 1995 found a duality theorem for uni-chromosomal genomes (reversals only) which expresses the genomic distance in terms of four parameters describing the combinatorial structure of permutations. In the case of multi-chromosomal genomes combinatorics of rearrangements becomes rather complicated. This paper presents the duality theorem for multi-chromosomal genomes which computes the genomic distance in terms of seven (!) parameters capturing different combinatorial properties of sets of strings.

Our analysis of multi-chromosomal genomes extensively uses the duality theorem for uni-chromosomal genomes (Hannenhalli and Pevzner, 1995), which is stated in section 2. In section 3 we introduce an idea called *flipping* of chromosomes to analyze a relatively simple case of so-called *co-tailed* genomes (approximation algorithm for this case was proposed by Kececioglu and Ravi, 1995). In section 4 we introduce another idea called *capping* of chromosomes. In section 5, using cappings, we take the first step towards a polynomial-time algorithm for genomic distance by proving a bound which is at most one rearrangement away from the genomic distance. This bound provides the intuition for a rather complicated potential function that leads to a polynomial-time algorithm and duality theorem for genomic sorting (section 6). Finally, in section 7 we present biological applications, formulate open problems, and discuss our result in relation to the recent experimental breakthrough in ZOO-FISH chromosome painting.

2 Cycles, hurdles and fortresses

In *reversal distance* problem, the order of genes in two (uni-chromosomal) genomes is represented by (unsigned) permutations $\pi = (\pi_1\pi_2\ldots\pi_n)$ and $\sigma = (\sigma_1\sigma_2\ldots\sigma_n)$ and the only considered rearrangements are reversals (i.e. both π and σ consist of a single chromosome and the genes do not have associated signs). Given (unsigned) permutations π and σ, the *reversal distance problem* is to find a series of reversals $\rho_1, \rho_2, \ldots, \rho_t$ such that $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_t = \sigma$ and t is minimum (Fig. 2a). We call t the *reversal distance* between π and σ. *Sorting π by reversals* is the problem of finding the reversal distance, $d(\pi)$, between π and the *identity* permutation $(12\ldots n)$.

Let $\pi = (\pi_1\ldots\pi_n)$ be a permutation of the elements $\{1,\ldots,n\}$. Denote $i \sim j$ if $|i-j| = 1$. Extend a permutation $\pi = (\pi_1\ldots\pi_n)$ by adding $\pi_0 = 0$ and $\pi_{n+1} = n+1$. We call a pair of consecutive elements π_i and π_{i+1}, $0 \le i \le n$, of π a *breakpoint* if $\pi_i \not\sim \pi_{i+1}$. The *breakpoint graph* of π is an edge-colored graph $G(\pi)$ with $n+2$ vertices $\{\pi_0, \pi_1, \ldots, \pi_n, \pi_{n+1}\}$. 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. 2b). 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-colored graph G is *alternating* if the colors 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 $\{1,\ldots,n\}$, i.e. a permutation with "+" or "−" sign associated with each element (Fig. 2c). 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 $(+1+2\ldots+n)$. Define a transformation from a signed permutation $\vec{\pi}$ of order n to an (unsigned) permutation π of $\{1,\ldots,2n\}$ 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. 2c). 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. 2c) 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 $\pi = \pi_1\pi_2\ldots\pi_{2n}$ of a signed permutation $\vec{\pi} = \vec{\pi}_1\vec{\pi}_2\ldots\vec{\pi}_n$, 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.

We say that reversal $\rho(i,j)$ *acts* on black edges (π_{i-1},π_i) and (π_j,π_{j+1}) in $G(\pi)$. We call $\rho(i,j)$ a *reversal on a cycle* if the black edges (π_{i-1},π_i) and (π_j,π_{j+1}) belong to the *same* cycle in $G(\pi)$. Every reversal increases $c(\pi)$ by at most 1, i.e., $c(\pi\rho)-c(\pi) \leq 1$ (Bafna and Pevzner, 1993). A gray edge g is *oriented* if for a reversal ρ acting on two black edges incident to g, $c(\pi\rho)-c(\pi)=1$ and *unoriented* otherwise. A cycle in $G(\pi)$ is *oriented* if it has an oriented gray edge and unoriented otherwise. Gray edges (π_i,π_j) and (π_k,π_t) in $G(\pi)$ are *interleaving* if the intervals $[i,j]$ and $[k,t]$ overlap but neither of them contains the other. Cycles C_1 and C_2 are *interleaving* if there exist interleaving gray edges $g_1 \in C_1$ and $g_2 \in C_2$. See Fig. 2c for examples.

Let $\mathcal{C}_\pi$ be the set of cycles in the breakpoint graph of a permutation π. Define an *interleaving* graph $H_\pi(\mathcal{C}_\pi,\mathcal{I}_\pi)$ of π with the edge set $\mathcal{I}_\pi = \{(C_1,C_2): C_1 \text{ and } C_2 \text{ are interleaving cycles in } \pi\}$ (Fig. 2d). The vertex set of H_π is partitioned into *oriented* and *unoriented* vertices (cycles in $\mathcal{C}_\pi$). A connected component of H_π is *oriented* if it has at least one oriented vertex and *unoriented* otherwise. In the following we use the terms edge of π, cycle in π and component of π instead of (more accurate) terms edge of $G(\pi)$, cycle in $G(\pi)$ and component of $H(\pi)$. A connected component U corresponds to the set of integers $\bar{U} = \{i: \pi_i \in C \in U\}$ representing the set of positions of the permutation belonging to cycles of U. For a set of integers U define $U_{min} = \min_{u\in U} u$ and $U_{max} = \max_{u\in U} u$.

Let $\prec$ be a partial order on a set P. An element $x \in P$ is called a *minimal* element in $\prec$ if there is no element $y \in P$ with $y \prec x$. An element $x \in P$ is the *greatest* in $\prec$ if $y \prec x$ for every $y \in P$ and $|P| > 1$. Let $\mathcal{U}$ be a collection of sets of integers. Define a partial order on $\mathcal{U}$ by the rule $U \prec W$ iff $[U_{min},U_{max}] \subset [W_{min},W_{max}]$ for $U,W \in \mathcal{U}$. We say that a set $U \in \mathcal{U}$ *separates* sets U' and U'' if there exists $u \in U$ such that $U'_{max} < u < U''_{min}$. A *hurdle for the set* $\mathcal{U}$ is defined as an unoriented component U in $\mathcal{U}$ which is either a minimal hurdle or the greatest hurdle where a *minimal hurdle* is a minimal element in $\prec$ and the *greatest hurdle* satisfies the following two conditions (i) U is the greatest element in $\prec$ *and* (ii) U does not separate any two sets in $\mathcal{U}$. A hurdle $K \in \mathcal{U}$ *protects* a non-hurdle $U \in \mathcal{U}$ if deleting K from $\mathcal{U}$ transforms U from a non-hurdle into a hurdle (i.e. U is a hurdle in $\mathcal{U} \setminus K$). A hurdle in π is a *superhurdle* if it protects a non-hurdle $U \in \mathcal{U}$.

Define a collection of sets of integers $\mathcal{U}_\pi = \{\bar{U}: U \text{ is an unoriented component of permutation } \pi\}$ and let $h(\pi)$ be the overall number of hurdles for the collection $\mathcal{U}_\pi$. Permutation π is called a *fortress* if it has an odd number of hurdles and all these hurdles are superhurdles. Define

$$f(\pi) = \begin{cases} 1, & \text{if } \pi \text{ is a fortress} \\ 0, & \text{otherwise} \end{cases}$$

For a signed permutation $\vec{\pi}$ with the image π we define $b(\vec{\pi}) = b(\pi)$, $c(\vec{\pi}) = c(\pi)$, $h(\vec{\pi}) = h(\pi)$ and $f(\vec{\pi}) = f(\pi)$.

Theorem 1 *(Hannenhalli and Pevzner, 1995) For a signed permutation $\vec{\pi}$ of order n, $d(\vec{\pi}) = b(\vec{\pi}) - c(\vec{\pi}) + h(\vec{\pi}) + f(\vec{\pi})$.*

3 Flipping the chromosomes

For a chromosome $\pi = (\pi_1 \dots \pi_n)$, the numbers $+\pi_1$ and $-\pi_n$ are called *tails* of π. Note that changing the direction of a chromosome does not change the set of its tails. Tails in a N-chromosomal genome Π comprise the set $\mathcal{T}(\Pi)$ of $2N$ tails. In this section we consider *co-tailed* genomes Π and Γ with $\mathcal{T}(\Pi) = \mathcal{T}(\Gamma)$. For co-tailed genomes internal reversals and translocations are sufficient for genomic sorting, i.e. prefix reversals, fusions and fissions can be ignored (the validity of this assumption will become clear later). For chromosomes $\pi = (\pi_1 \dots \pi_n)$ and $\sigma = (\sigma_1 \dots \sigma_m)$ denote the fusion $(\pi_1 \dots \pi_n \sigma_1 \dots \sigma_m)$ by $\pi + \sigma$ and the fusion $(\pi_1 \dots \pi_n - \sigma_m \dots - \sigma_1)$ by $\pi - \sigma$. Given an ordering of chromosomes $(\pi(1), \dots, \pi(N))$ in a genome Π and a *flip* vector $s = (s(1), \dots, s(N))$ with $s(i) \in \{-1, +1\}$ one can form a *concatenate* of Π as a permutation $\Pi(s) = s(1)\pi(1) + \dots + s(N)\pi(N)$ on $\sum_{i=1}^{N} n_i$ elements. Depending on the choice of a flip vector there exists 2^N concatenates of Π for each of $N!$ orderings of chromosomes in Π. If an order of chromosomes in a genome Π is fixed we call Π an *ordered* genome.

In this section we assume w.l.o.g. that $\Gamma = (\gamma_1, \dots, \gamma_N)$ is an (ordered) genome and $\gamma = \gamma_1 + \dots + \gamma_N$ is the identity permutation. We denote $d(\Pi) \equiv d(\Pi, \Gamma)$ and call a problem of genomic sorting of Π into Γ simply a *sorting of a genome* Π.

We use the following idea to analyze co-tailed genomes. Given a concatenate π of a genome Π one can optimally sort π by reversals (Hannenhalli and Pevzner, 1995). Every reversal in this sorting corresponds to a reversal or a translocation in a (not necessarily optimal) sorting of the genome Π. For example, a translocation $\rho(\pi, \sigma, i, j)$ acting on chromosomes $\pi = (\pi_1 \dots \pi_n)$ and $\sigma = (\sigma_1 \dots \sigma_m)$ can be alternatively viewed as a reversal $\rho(\pi - \sigma, i, n + (m - j + 1))$ acting on $\pi - \sigma$ (and vice versa). Define an *optimal concatenate* of Π as a concatenate π with minimum reversal distance $d(\pi)$ among all concatenates of Π. Below we prove that sorting of an optimal concatenate of Π mimics an optimal sorting of a genome Π. This approach reduces the problem of sorting Π to a problem of finding an optimal concatenate of Π.

Let π be a concatenate of $\Pi = (\pi(1), \dots, \pi(N))$. Every tail of $\pi(i)$ corresponds to two vertices of of the breakpoint graph $G(\pi)$, exactly one of which is a boundary (either leftmost or rightmost) vertex among the vertices of the chromosome $\pi(i)$ in the concatenate π. We extend the term *tail* to denote such vertices in $G(\pi)$. An edge in a breakpoint graph $G(\pi)$ of a concatenate π is *interchromosomal* if it connects vertices in different chromosomes of Π and *intrachromosomal* otherwise. A component of π is *interchromosomal* if it contains an interchromosomal edge and *intrachromosomal* otherwise.

Every interchromosomal black edge in $G(\pi)$ connects two tails. Let $b_{tail}(\Pi)$ (notice that $b_{tail}(\Pi) = N - 1$) be the number of interchromosomal black edges in $G(\pi)$. Note that for co-tailed genomes tails in $G(\Pi)$ are adjacent to tails only and hence a cycle containing a tail contains only tails. Let $c_{tail}(\Pi)$ be the number of cycles of $G(\pi)$ containing tails. Define $b(\Pi) = b(\pi) - b_{tail}(\pi)$ (notice that $b(\Pi) = n - N$ and $b(\pi) = n - 1$) and $c(\Pi) = c(\pi) - c_{tail}(\pi)$.

Consider the set of intrachromosomal unoriented components $\mathcal{IU}_\pi$ in π. Hurdles, superhurdles and fortresses for the set $\mathcal{IU}_\pi$ are called *knots*, *superknots* and *fortresses-of-knots* respectively. Let $k(\Pi)$ be the number of knots in a concatenate π of Π. Define $f(\Pi) = 1$ if π is a fortress-of-knots and $f(\Pi) = 0$ otherwise. Clearly, $b(\Pi)$, $c(\Pi)$, $k(\Pi)$ and $f(\Pi)$ do not depend on the choice of a concatenate π.

Lemma 1 *For co-tailed genomes Π and Γ, $d(\Pi) \geq b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi)$.*

Proof A more involved version of the proof of theorem 1 from Hannenhalli and Pevzner, 1995. ∎

Concatenates $\Pi(s)$ and $\Pi(s')$ of an (ordered) genome Π are *i-twins* if the directions of all chromosomes but i^{th} one in $\Pi(s)$ and $\Pi(s')$ coincide, i.e. $s(i) = -s'(i)$ and $s(j) = s'(j)$ for $j \neq i$. A chromosome $\pi(i)$ is *properly flipped* in $\Pi(s)$ if all interchromosomal edges originating in this chromosome belong to oriented components in $\Pi(s)$. A concatenate π is *properly flipped* if every chromosome in π is properly flipped. The following lemma proves the existence of a properly flipped concatenate.

Lemma 2 *If a chromosome $\pi(i)$ is not properly flipped in $\pi = \Pi(s)$ then it is properly flipped in the i-twin π' of π. Moreover, every properly flipped chromosome in π remains properly flipped in π'.*

Proof: Let g be an interchromosomal gray edge in π originating in the chromosome $\pi(i)$ and belonging to an unoriented component in π. Note that the orientation of any interchromosomal gray edge originating at $\pi(i)$ is different in π as compared to π' (i.e. a non-oriented edge in π becomes oriented in π' and vice versa). Since all edges interleaving with g in π are unoriented, every interchromosomal edge originating at $\pi(i)$ and interleaving with g in π is oriented in π'.

All interchromosomal edges originating in $\pi(i)$ which are not interleaving with g in π, interleave with g in π'. Since g is oriented in π' all such edges belong to an oriented component containing g in π'. Therefore, $\pi(i)$ is properly flipped in π'.

Let $\pi(j)$ be a properly flipped chromosome in π. If $\pi(j)$ is not properly flipped in π' then there exists an interchromosomal unoriented component U having an interchromosomal gray edge originating at $\pi(j)$ in π'. If U does not have an edge originating at $\pi(i)$ in π' then U is an unoriented component in π, implying that $\pi(j)$ was not properly flipped in π, a contradiction. If U has an (unoriented) gray edge h originating at $\pi(i)$ then, clearly, this edge does not interleave with g in π'. Therefore h interleaves with g in π and h is oriented in π thus implying that g belonged to an oriented component in π, a contradiction. ∎

Lemma 2 implies existence of a properly flipped concatenate $\pi = \Pi(s)$ with $h(\pi) = k(\Pi)$ and $f(\pi) = f(\Pi)$. Below we show that there exists a sorting of π by $b(\pi) - c(\pi) + h(\pi) + f(\pi)$ reversals which mimics a sorting of Π by $b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi)$ internal reversals and translocations.

Theorem 2 *For co-tailed genomes* Π *and* Γ, $d(\Pi, \Gamma) \equiv d(\Pi) = b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi)$.

Proof Assume the contrary and let Π be a genome with the minimum value of $b(\Pi)-c(\Pi)+h(\Pi)+f(\Pi)$ among the genomes for which the theorem fails. Let π be a properly flipped concatenate of Π with minimal value of $b_{tail}(\pi) - c_{tail}(\pi)$ among all properly flipped concatenates of Π.

If $b_{tail}(\pi) = c_{tail}(\pi)$ (*i.e.* every interchromosomal black edge is involved in a cycle of length two) then there exists an optimal sorting of π by $b(\pi)-c(\pi)+k(\pi)+f(\pi)$ reversals which act on intrachromosomal black edges (Hannenhalli and Pevzner, 1995). Every such reversal ρ can be mimicked as an internal reversal or an internal translocation on Π thus leading to sorting Π by $b(\pi) - c(\pi) + k(\pi) + f(\pi)$ internal reversals/translocations. Since π is a properly flipped concatenate, $b(\pi) = b(\Pi) + b_{tail}(\pi)$, $c(\pi) = c(\Pi) + c_{tail}(\pi)$, $h(\pi) = k(\Pi)$, $f(\pi) = f(\Pi)$. Therefore, optimal sorting of π mimics an optimal sorting of Π by $b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi)$ internal reversals/translocations.

If $b_{tail}(\pi) > c_{tail}(\pi)$ then there exists an interchromosomal black edge involved in a cycle of length greater than two and this edge belongs to an oriented component in π (since every interchromosomal black edge belongs to an oriented component in a properly flipped concatenate). Hannenhalli and Pevzner, 1995 proved that if there exists an oriented component in π then there exists a reversal ρ in π acting on the black edges of an oriented cycle in this component such that $c(\pi\rho) = c(\pi)+1$. Moreover, this reversal does not create new unoriented components in $\pi\rho$ and $h(\pi\rho) = h(\pi)$, $f(\pi\rho) = f(\pi)$. Note that every cycle containing tails of chromosomes belongs to an oriented component in π and consists entirely from edges between tails. Therefore ρ acts either on two intrachromosomal black edges or on two interchromosomal black edges belonging to some oriented cycle of this component.

A reversal ρ acting on two interchromosomal black edges can be interpreted as a transformation of a concatenate π of $\Pi = (\pi(1), \ldots, \pi(i-1), \pi(i), \ldots, \pi(j), \pi(j+1), \ldots, \pi(N))$ into a concatenate $\pi\rho' = \Pi'(s')$ where Π' is a new ordering $(\pi(1), \ldots, \pi(i-1), \pi(j), \ldots, \pi(i), \pi(j+1), \ldots, \pi(N))$ of the chromosomes and $s' = (s(1), \ldots, s(i-1), -s(j), \ldots, -s(i), s(j+1), \ldots, s(N))$. Therefore, $b_{tail}(\pi\rho) - c_{tail}(\pi\rho) = b_{tail}(\pi) - (c_{tail}(\pi) + 1)$ and $\pi\rho$ is a properly flipped concatenate of Π, a contradiction to minimality of $b_{tail}(\pi) - c_{tail}(\pi)$.

If reversal ρ acts on two intrachromosomal black edges then $\pi\rho$ is a properly flipped concatenate of $\Pi\rho$ implying that

$$b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi) = (b(\pi) - b_{tail}(\pi)) - (c(\pi) - c_{tail}(\pi)) + h(\pi) + f(\pi) =$$
$$(b(\pi\rho) - b_{tail}(\pi\rho)) - (c(\pi\rho) - 1 - c_{tail}(\pi\rho)) + h(\pi\rho) + f(\pi\rho) = b(\Pi\rho) - c(\Pi\rho) + h(\Pi\rho) + f(\Pi\rho) + 1$$

Since $b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi) > b(\Pi\rho) - c(\Pi\rho) + h(\Pi\rho) + f(\Pi\rho)$, the theorem holds for the genome $\Pi\rho$. Therefore $d(\Pi) \leq d(\Pi\rho) + 1 = b(\Pi) - c(\Pi) + k(\Pi) + f(\Pi)$ ∎

4 Capping the chromosomes

We now turn to the general case when genomes Π and Γ might have different sets of tails and different number of chromosomes. Below we describe an algorithm for computing $d(\Pi,\Gamma)$ which is polynomial in the number of genes but exponential in the number of chromosomes. This algorithm provides an intuition for the (truly) polynomial-time algorithm which is described in the following sections.

Let Π and Γ be two genomes with M and N chromosomes respectively. W.l.o.g. assume that $M \le N$ and extend Π by $N-M$ empty chromosomes. As a result $\Pi = \{\pi(1),\ldots,\pi(N)\}$ and $\Gamma = \{\gamma(1),\ldots,\gamma(N)\}$ contain the same number of chromosomes. Let $\{cap_0,\ldots,cap_{2N-1}\}$ be a set of $2N$ distinct positive integers (called *caps*) which are different from genes of Π (or equivalently, Γ). Let $\hat{\Pi} = \{\hat{\pi}(1),\ldots,\hat{\pi}(N)\}$ be a genome obtained from Π by adding caps to the ends of each chromosome, i.e. $\hat{\pi}(i) = cap_{2(i-1)}, \pi(i)_1,\ldots,\pi(i)_{n_i}, cap_{2(i-1)+1}$. Note that every reversal/translocation in Π corresponds to an *internal* reversal/translocation in $\hat{\Pi}$. If this translocation is a fission we assume that there are enough empty chromosomes in Π (the validity of this assumption will become clear later).

Every sorting of Π into Γ induces a sorting of $\hat{\Pi}$ into a genome $\hat{\Gamma} = \{\hat{\gamma}(1),\ldots,\hat{\gamma}(N)\}$ (called *capping* of Γ), where $\hat{\gamma}(i) = ((-1)^j cap_j, \hat{\gamma}(i)_1,\ldots,\hat{\gamma}(i)_{m_i}, (-1)^{k+1} cap_k)$ for $0 \le j,k \le 2N-1$. Genomes $\hat{\Pi}$ and $\hat{\Gamma}$ are co-tailed since $\mathcal{T}(\hat{\Pi}) = \mathcal{T}(\hat{\Gamma}) = \bigcup_{i=0}^{2N-1}(-1)^i cap_i$. There exist $(2N)!$ different cappings of Γ, each capping defined by the distribution of $2N$ caps of $\hat{\Pi}$ in $\hat{\Gamma}$. Denote the set of $(2N)!$ cappings of Γ as $\mathbf{\Gamma}$. The following lemma leads to an algorithm for computing genomic distance which is polynomial in the number of genes but exponential in the number of chromosomes N.

Lemma 3 $d(\Pi,\Gamma) = \min_{\hat{\Gamma}\in\mathbf{\Gamma}} b(\hat{\Pi},\hat{\Gamma}) - c(\hat{\Pi},\hat{\Gamma}) + k(\hat{\Pi},\hat{\Gamma}) + f(\hat{\Pi},\hat{\Gamma})$

Proof Follows from theorem 2 and an observation that every sorting of $\hat{\Pi}$ into a genome $\hat{\Gamma} \in \mathbf{\Gamma}$ by internal reversals/translocations induces a sorting of Π into Γ. ∎

Let $\hat{\pi}$ and $\hat{\gamma}$ be arbitrary concatenates of (ordered) cappings $\hat{\Pi}$ and $\hat{\Gamma}$. Let $G(\hat{\Pi},\hat{\Gamma})$ be a graph obtained from $G(\hat{\pi},\hat{\gamma})$ by deleting all tails (vertices of $G(\hat{\pi},\hat{\gamma})$) of genome $\hat{\Pi}$ (or equivalently, $\hat{\Gamma}$) from $G(\hat{\pi},\hat{\gamma})$. Different cappings $\hat{\Gamma}$ correspond to different graphs $G(\hat{\Pi},\hat{\Gamma})$. Graph $G(\hat{\Pi},\hat{\Gamma})$ has $2N$ vertices corresponding to caps, gray edges incident on these vertices completely define the capping $\hat{\Gamma}$. Therefore, deleting these $2N$ gray edges transforms $G(\hat{\Pi},\hat{\Gamma})$ into a graph $G(\Pi,\Gamma)$ which does not depend on capping $\hat{\Gamma}$ (Fig. 3a,b,c,d).

Graph $G(\Pi,\Gamma)$ contains $2N$ vertices of degree 1 corresponding to $2N$ caps of Π (called Π-*caps*) and $2N$ vertices of degree 1 corresponding to $2N$ tails of Γ (called Γ-*tails*). Therefore, $G(\Pi,\Gamma)$ is a collection of cycles and $2N$ paths, each path starting and ending with a black edge. A path is a $\Pi\Pi$-*path* ($\Gamma\Gamma$-*path*) if it starts and ends with Π-caps (Γ-tails) and a $\Pi\Gamma$-*path* if it starts with a Π-cap and ends with a Γ-tail. A vertex in $G(\Pi,\Gamma)$ is a $\Pi\Gamma$-*vertex* if it is a Π-cap in a $\Pi\Gamma$-path and a $\Pi\Pi$-*vertex* if it is a Π-cap in a $\Pi\Pi$-path. $\Gamma\Pi$- and $\Gamma\Gamma$-vertices are defined similarly (see Fig. 3d).

Every capping $\hat{\Gamma}$ corresponds to adding $2N$ gray edges to the graph $G(\Pi,\Gamma)$, each edge joining a Π-cap with a Γ-tail. These edges transform $G(\Pi,\Gamma)$ into the graph $G(\hat{\Pi},\hat{\Gamma})$ corresponding to a capping $\hat{\Gamma}$ (Fig. 3e).

Define $b(\Pi,\Gamma)$ as the number of black edges in $G(\Pi,\Gamma)$ and $c(\Pi,\Gamma)$ as the overall number of cycles and paths in $G(\Pi,\Gamma)$. The parameter $b(\Pi,\Gamma) = b(\hat{\Pi},\hat{\Gamma})$ does not depend on capping $\hat{\Gamma}$. Clearly $c(\hat{\Pi},\hat{\Gamma}) \le c(\Pi,\Gamma)$ with $c(\hat{\Pi},\hat{\Gamma}) = c(\Pi,\Gamma)$ if every path in $G(\Pi,\Gamma)$ is "closed" by a gray edge in $G(\hat{\Pi},\hat{\Gamma})$. An observation that every cycle in $G(\hat{\Pi},\hat{\Gamma})$ containing a $\Pi\Pi$-path contains at least one more path leads to an inequality $c(\hat{\Pi},\hat{\Gamma}) \le c(\Pi,\Gamma) - p(\Pi,\Gamma)$, where $p(\Pi,\Gamma)$ is the number of $\Pi\Pi$-paths (or equivalently, $\Gamma\Gamma$-paths) in $G(\Pi,\Gamma)$.

We define the notions of interleaving cycles/paths, oriented and unoriented components, etc. in the graph $G(\Pi,\Gamma)$ in a usual way (see Appendix) by making no distinction between cycles and paths in $G(\Pi,\Gamma)$. We say that a vertex π_j is *inside* a component U of π if $j \in [\bar{U}_{min},\bar{U}_{max}]$. An intrachromosomal component for genomes Π and Γ is called a *real* component if it has neither a Π-cap nor a Γ-tail inside.

For genomes Π and Γ define $\mathcal{RU}(\Pi,\Gamma)$ as the set of real components and $\mathcal{IU}(\Pi,\Gamma)$ as the set of intrachromosomal components (as defined by the graph $G(\Pi,\Gamma)$). Clearly $\mathcal{RU}(\Pi,\Gamma) \subseteq \mathcal{IU}(\Pi,\Gamma)$. Hurdles, superhurdles and fortresses for the set $\mathcal{RU}(\Pi,\Gamma)$ are called *real-knots*, *super-real-knots* and *fortresses-of-real-knots*. Let RK be the set of real-knots (i.e. hurdles for the set $\mathcal{RU}(\Pi,\Gamma)$) and K be the set of knots (i.e. hurdles for the set $\mathcal{IU}(\Pi,\Gamma)$). A knot from the set $K \setminus RK$ is a *semi-knot* if it does not contain a $\Pi\Pi$- or $\Gamma\Gamma$-vertex inside. Clearly, every semi-knot contains a $\Pi\Gamma$-path (otherwise, it would be a real-knot). Denote the number of real-knots and semi-knots for genomes Π and Γ as $r(\Pi,\Gamma)$ and $s(\Pi,\Gamma)$, respectively. Clearly $k(\hat{\Pi},\hat{\Gamma}) \geq r(\Pi,\Gamma)$ implying that

$$b(\hat{\Pi},\hat{\Gamma}) - c(\hat{\Pi},\hat{\Gamma}) + k(\hat{\Pi},\hat{\Gamma}) \leq b(\Pi,\Gamma) - c(\Pi,\Gamma) + p(\Pi,\Gamma) + r(\Pi,\Gamma)$$

However, this bound is not tight since it assumes that there exists a capping $\hat{\Gamma}$ which simultaneously maximizes $c(\hat{\Pi},\hat{\Gamma})$ and minimizes $k(\hat{\Pi},\hat{\Gamma})$. Taking $s(\Pi,\Gamma)$ into account leads to a better bound for genomic distance which is at most 1 rearrangement apart from the genomic distance (next section).

5 Caps and tails

Genomes Π and Γ are *correlated* if all the real-knots in $G(\Pi,\Gamma)$ are located on the same chromosome and *non-correlated* otherwise. In this section we restrict our analysis to non-correlated genomes (it turns out that the analysis of correlated genomes involves some additional technical difficulties) and prove a tight bound for $d(\Pi,\Gamma)$ (this bound provides an intuition for a rather complicated potential function used in the proof of the duality theorem):

$$b(\Pi,\Gamma) - c(\Pi,\Gamma) + p(\Pi,\Gamma) + r(\Pi,\Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil \leq d(\Pi,\Gamma) \leq$$

$$b(\Pi,\Gamma) - c(\Pi,\Gamma) + p(\Pi,\Gamma) + r(\Pi,\Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil + 1$$

The following lemmas suggest a way to connect some paths in $G(\Pi,\Gamma)$ by oriented edges.

Lemma 4 *For every $\Pi\Pi$-path and $\Gamma\Gamma$-path in $G(\Pi,\Gamma)$ there exists either an interchromosomal or an oriented gray edge which joins these paths into a $\Pi\Gamma$-path.*

Lemma 5 *For every two unoriented $\Pi\Gamma$-paths located on the same chromosome there exists an oriented gray edge which joins these paths into a $\Pi\Gamma$-path.*

In a search for an optimal capping we first ignore the term $f(\hat{\Pi},\hat{\Gamma})$ in lemma 3 and find a capping whose genomic distance $d(\hat{\Pi},\hat{\Gamma})$ is within 1 from the optimal. The following theorem suggests a way to find such an "almost optimal" capping $\hat{\Gamma}$.

Theorem 3 $\min_{\hat{\Gamma}\in\mathbf{\Gamma}} b(\hat{\Pi},\hat{\Gamma}) - c(\hat{\Pi},\hat{\Gamma}) + k(\hat{\Pi},\hat{\Gamma}) = b(\Pi,\Gamma) - c(\Pi,\Gamma) + p(\Pi,\Gamma) + r(\Pi,\Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$.

Proof Every capping $\hat{\Gamma}$ defines a transformation of $G(\Pi,\Gamma)$ into $G(\hat{\Pi},\hat{\Gamma})$ by consecutively adding $2N$ gray edges to $G(\Pi,\Gamma)$: $G(\Pi,\Gamma) = G_0 \stackrel{g_1}{\to} G_1 \stackrel{g_2}{\to} \ldots \stackrel{g_{2N}}{\to} G_{2N} = G(\hat{\Pi},\hat{\Gamma})$. For a graph G_i the parameters $b_i = b(G_i)$, $c_i = c(G_i)$, $p_i = p(G_i)$, $r_i = r(G_i)$ and $s_i = s(G_i)$ are defined in the same way as for the graph $G_0 = G(\Pi,\Gamma)$. For a parameter ϕ define $\Delta\phi_i$ as $\phi_i - \phi_{i-1}$, i.e. $\Delta c_i = c_i - c_{i-1}$, etc. Denote $\Delta_i = (c_i - p_i - r_i - \lceil \frac{s_i}{2} \rceil) - (c_{i-1} - p_{i-1} - r_{i-1} - \lceil \frac{s_{i-1}}{2} \rceil)$. Below we prove that $\Delta_i \leq 0$ for $1 \leq i \leq 2N$, i.e. adding a gray edge does not increase the parameter $c(\Pi,\Gamma) - p(\Pi,\Gamma) - r(\Pi,\Gamma) - \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$. For a fixed i we ignore index i, i.e. denote $\Delta = \Delta_i$, etc.

Depending on the edge g_i the following cases are possible (the analysis below assumes that Π and Γ are non-correlated):

Case 1: edge g_i "closes" a $\Pi\Gamma$-path (i.e. g_i connects a $\Pi\Gamma$-vertex with a $\Gamma\Pi$-vertex within the same $\Pi\Gamma$-path). If this vertex is the only $\Pi\Gamma$-vertex in a semi-knot, then $\Delta c = 0, \Delta p = 0, \Delta r = 1, \Delta s = -1$ (note that this might not be true for correlated genomes). Otherwise $\Delta c = 0, \Delta p = 0, \Delta r = 0, \Delta s = 0$. In both cases $\Delta \leq 0$.

Case 2: edge g_i connects a $\Pi\Gamma$-vertex with a $\Gamma\Pi$-vertex in a different $\Pi\Gamma$-path. This edge "destroys" at most two semi-knots and $\Delta c = -1, \Delta p = 0, \Delta r = 0, \Delta s \geq -2$. Therefore $\Delta \leq 0$.

Case 3: edge g_i connects a $\Pi\Gamma$-vertex with a $\Gamma\Gamma$-vertex (or a $\Gamma\Pi$-vertex with a $\Pi\Pi$-vertex). This edge "destroys" at most one semi-knot and $\Delta c = -1, \Delta p = 0, \Delta r = 0, \Delta s > -2$. It implies $\Delta \leq 0$.

Case 4: edge g_i connects a $\Pi\Pi$-vertex with a $\Gamma\Gamma$-vertex. This edge can not destroy any semi-knots and $\Delta c = -1, \Delta p = -1, \Delta r = 0, \Delta s \geq 0$. It implies $\Delta \leq 0$.

Note that $b_{2N} = b(\hat{\Pi}, \hat{\Gamma}) = b(\Pi, \Gamma) = b_0$, $c_{2N} = c(\hat{\Pi}, \hat{\Gamma})$, $p_{2N} = 0$, $s_{2N} = 0$ and $r_{2N} = k(\hat{\Pi}, \hat{\Gamma})$. Therefore $b(\hat{\Pi}, \hat{\Gamma}) - c(\hat{\Pi}, \hat{\Gamma}) + k(\hat{\Pi}, \hat{\Gamma}) = b_{2N} - c_{2N} + p_{2N} + r_{2N} + \lceil \frac{s_{2N}}{2} \rceil \geq b_0 - c_0 + p_0 + r_0 + \lceil \frac{s_0}{2} \rceil = b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$.

We now prove that there exists a capping $\hat{\Gamma}$ such that $b(\hat{\Pi}, \hat{\Gamma}) - c(\hat{\Pi}, \hat{\Gamma}) + k(\hat{\Pi}, \hat{\Gamma}) = b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$ by constructing a sequence of $2N$ gray edges $g_1, \ldots, g_{2N}$ connecting Π-caps with Γ-tails in $G(\Pi, \Gamma)$ such that $\Delta_i = 0$ for all $1 \leq i \leq 2N$.

Assume that the first $i-1$ such edges are already found and let G_{i-1} be the result of adding these $i-1$ edges to $G(\Pi, \Gamma)$. If G_{i-1} has a $\Pi\Pi$-path then it has a $\Gamma\Gamma$-path as well and, by lemma 4 there exists an interchromosomal or oriented gray edge joining these paths into an oriented $\Pi\Gamma$-path. Clearly $\Delta c = -1$, $\Delta p = -1$, $\Delta r = 0$, $\Delta s = 0$ for this edge, implying $\Delta = 0$.

If G_{i-1} has at least two semi-knots (i.e. $s_{i-1} > 1$) let v_1 and v_2 be a $\Pi\Gamma$- and a $\Gamma\Pi$-vertex in different semi-knots. If v_1 and v_2 are in different chromosomes of Π then the gray edge $g_i = (v_1, v_2)$ "destroys" both semi-knots. Therefore $\Delta c = -1, \Delta p = 0, \Delta r = 0, \Delta s = -2$ and $\Delta = 0$. If v_1 and v_2 belong to the same chromosome then by lemma 5 there exists an oriented gray edge joining these paths into an oriented $\Pi\Gamma$-path. This gray edge destroys two semi-knots. Therefore $\Delta = 0$ in this case also.

If G_{i-1} has the only semi-knot, let P_1 be a $\Pi\Gamma$-path in this semi-knot. If it is the only $\Pi\Gamma$-path in the semi-knot then for an edge g_i "closing" this path, $\Delta c = 0, \Delta p = 0, \Delta r = 1, \Delta s = -1$ implying that $\Delta = 0$. Otherwise, $\Delta c = 0, \Delta p = 0, \Delta r = 0, \Delta s = 0$ implying that $\Delta = 0$.

If G_{i-1} has neither a $\Pi\Pi$-path, nor a semi-knot then let g_i be an edge closing an arbitrary $\Pi\Gamma$-path in G_{i-1}. Since g_i does not belong to a semi-knot, $\Delta c = 0, \Delta p = 0, \Delta r = 0, \Delta s = 0$ and $\Delta = 0$. Therefore, the constructed sequence of edges $g_1, \ldots, g_{2N}$ transforms $G(\Pi, \Gamma)$ into $G(\hat{\Pi}, \hat{\Gamma})$ such that $b(\hat{\Pi}, \hat{\Gamma}) - c(\hat{\Pi}, \hat{\Gamma}) + k(\hat{\Pi}, \hat{\Gamma}) = b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$. ∎

Since $0 \leq f(\Pi, \Gamma) \leq 1$, lemma 3 and theorem 3 imply that $b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$ is within one rearrangement from the genomic distance $d(\Pi, \Gamma)$ for non-correlated genomes. In the following section we close the gap between $b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$ and $d(\Pi, \Gamma)$ for arbitrary genomes.

6 Duality theorem for genomic distance

The major difficulty in closing the gap between $b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + r(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma)}{2} \rceil$ and $d(\Pi, \Gamma)$ is to "uncover" remaining "obstacles" in the duality theorem. It turns out that the duality theorem involves 7 (!) parameters, thus making it very hard to explain an intuition behind it. Theorem 3 provides such an intuition for the first five parameters. Two more parameters are defined below.

A component in $G(\Pi, \Gamma)$ containing a $\Pi\Gamma$-path is *simple* if it is not a semi-knot.

Lemma 6 *There exists an optimal capping $\hat{\Gamma}$ which closes all $\Pi\Gamma$-paths in simple components.*

Let $\overline{G}$ be a graph obtained from $G(\Pi, \Gamma)$ by closing all $\Pi\Gamma$-paths in simple components. Without a confusion we can use the terms *real-knots, super-real-knots and fortress-of-real-knots in $\overline{G}$* and define $rr(\Pi, \Gamma)$ as the number of real-knots in $\overline{G}$. Note that $rr(\Pi, \Gamma)$ does not necessarily coincide with $r(\Pi, \Gamma)$.

Correlated genomes Π and Γ form a *weak-fortress-of-real-knots* if (i) they have an odd number of real-knots in $\overline{G}$, (ii) one of the real-knots is the greatest real-knot in $\overline{G}$, (iii) every real-knot but the greatest one is a super-real-knot in $\overline{G}$ and (iv) $s(\Pi,\Gamma)>0$. Notice that a weak-fortress-of-real-knots can be transformed into fortress-of-real-knots by closing $\Pi\Gamma$-paths contained in one of the semi-knots.

Define two more parameters as follows:

$$fr(\Pi,\Gamma)=\begin{cases}1, & \text{if } \Pi \text{ and } \Gamma \text{ form a fortress-of-real-knots or a weak-fortress-of-real-knots in } \overline{G}\\ 0, & \text{otherwise}\end{cases}$$

$$gr(\Pi,\Gamma)=\begin{cases}1, & \text{if there exists the greatest real-knot in } \overline{G} \text{ and } s(\Pi,\Gamma)>0\\ 0, & \text{otherwise}\end{cases}$$

The following duality theorem proves that the algorithm *Genomic_Sort* (Fig. 4) solves the genomic sorting problem. The running time of *Genomic_Sort* (dominated by the running time of sorting signed permutations by reversals[3]) is $O(n^4)$, where n is the overall number of genes (Hannenhalli and Pevzner, 1995).

Theorem 4 $d(\Pi,\Gamma)=b(\Pi,\Gamma)-c(\Pi,\Gamma)+p(\Pi,\Gamma)+rr(\Pi,\Gamma)+\lceil\frac{s(\Pi,\Gamma)-gr(\Pi,\Gamma)+fr(\Pi,\Gamma)}{2}\rceil$

Proof Let t be the number of $\Pi\Gamma$-paths in simple components of $G(\Pi,\Gamma)$ and let $\hat{\Gamma}$ be an optimal capping which closes all these $\Pi\Gamma$-paths in simple components (lemma 6). Similar to the proof of the theorem 3 we consider a transformation of $G(\Pi,\Gamma)$ into $G(\hat{\Pi},\hat{\Gamma})$ defined by $2N$ gray edges: $G(\Pi,\Gamma)=G_0\overset{g_1}{\to}G_1\overset{g_2}{\to}\ldots\overset{g_t}{\to}G_t=\overline{G}\overset{g_t+1}{\to}\ldots\overset{g_{2N}}{\to}G_{2N}=G(\hat{\Pi},\hat{\Gamma})$ and assume that the first t edges in this transformation close $\Pi\Gamma$-paths in simple components.

The parameters b_i, c_i, p_i, r_i, gr_i and fr_i are defined in the same way as in the theorem 3. Denote $\Delta_i=(c_i-p_i-r_i-\lceil\frac{s_i-gr_i+fr_i}{2}\rceil)-(c_{i-1}-p_{i-1}-r_{i-1}-\lceil\frac{s_{i-1}-gr_{i-1}+fr_{i-1}}{2}\rceil)$. Below we prove that $\Delta_i\le 0$ for $t+1\le i\le 2N$. For a fixed i we ignore index i, i.e. denote $\Delta=\Delta_i$, etc.

Depending on the edge g_i the following cases are possible:

Case 1: edge g_i closes a $\Pi\Gamma$-path P. If this path is the only $\Pi\Gamma$-path in a semi-knot S then we consider two sub-cases: $gr_i=1$ and $gr_i=0$. If $gr_i=1$ then there exists the greatest real-knot in G_i and $s_i>0$. Therefore edge g_i transforms S into the greatest real-knot in G_i. It implies that all real-knots in G_{i-1} (if any) are located on the same chromosome as S. Moreover, $r_{i-1}>0$ since otherwise S is the only real-knot in G_i, a contradiction to S being the greatest real-knot in G_i (see the definition of the greatest hurdle). Since $s_i>0$, S is not a semi-knot (since S is not a hurdle in $\mathcal{IU}$), a contradiction. If $gr_i=0$ then either $gr_{i-1}=0$ or $gr_{i-1}=1$. If $gr_{i-1}=0$ then $\Delta c=0,\Delta p=0,\Delta r=1,\Delta s=-1$, $\Delta gr=0$, $\Delta fr\ge -1$ implying that $\Delta\le 0$. If $gr_{i-1}=1$ then $\Delta c=0,\Delta p=0,\Delta r=0,\Delta s=-1$, $\Delta gr=-1$. If $fr_{i-1}=1$, G_{i-1} forms a fortress-of-real-knots or a weak-fortress-of-real-knots. One can see that in this case G_i is a fortress-of-real-knots and $fr_i=1$. Therefore $\Delta fr=0$ implying $\Delta\le 0$. If $fr_{i-1}=0$ then $\Delta fr\ge 0$ and $\Delta\le 0$.

If the path P is the only $\Pi\Gamma$-path in a simple component then after closing the path P, either $\Delta gr=0$ (in this case $\Delta c=0,\Delta p=0,\Delta r=0,\Delta s=0$, $\Delta fr\ge 0$) or $\Delta gr=1$ (in this case $\Delta c=0,\Delta p=0,\Delta r=1,\Delta s=0$, $\Delta fr\ge -1$). In both cases $\Delta\le 0$. If the path P is not the only $\Pi\Gamma$-path in its component then closing P does not destroy any semi-knots. Therefore, $\Delta c=0,\Delta p=0,\Delta r=0,\Delta s=0$, $\Delta gr=0$ and $\Delta fr=0$, implying $\Delta=0$.

Case 2: edge g_i connects a $\Pi\Gamma$-vertex with a $\Gamma\Pi$-vertex in a different $\Pi\Gamma$-path. This edge "destroys" at most two semi-knots. If it destroys less than two semi-knots then $\Delta c=-1,\Delta p=0,\Delta r=0,\Delta s\ge -1$. Since $\Delta gr\le 0$ and $\Delta fr\ge -1$, $\Delta\le 0$. If it does destroy two semi-knots, $\Delta c=-1,\Delta p=0,\Delta r=0,\Delta s=-2$. Clearly $\Delta gr\le 0$ in this case. If $\Delta gr=-1$ then $\Delta s-\Delta gr+\Delta fr\ge -2$ and $\Delta\le 0$. If $\Delta gr=0$ and $\Delta fr\ge 0$ then $\Delta\le 0$. If $\Delta gr=0$ and $\Delta fr=-1$ then $fr_{i-1}=1$ and $fr_i=0$. It implies that G_{i-1} is a weak-fortress-of-real-knots and $s_{i-1}=2, s_i=0$. It implies $gr_{i-1}=1$, $gr_i=0$ and $\Delta gr=-1$, a contradiction.

Case 3: edge g_i connects a $\Pi\Gamma$-vertex with a $\Gamma\Gamma$-vertex (or a $\Gamma\Pi$-vertex with a $\Pi\Pi$-vertex). This edge "destroys" at most one semi-knots and $\Delta c=-1,\Delta p=0,\Delta r=0,\Delta s\ge -1$. Since $\Delta gr\le 0$ and $\Delta fr\ge -1$, $\Delta\le 0$.

[3]Recently Berman and Hannenhalli further improved the running time for sorting signed permutations by reversals

Case 4: edge g_i connects a $\Pi\Pi$-vertex with a $\Gamma\Gamma$-vertex. This edge can not destroy any semi-knots and $\Delta c = -1, \Delta p = -1, \Delta r = 0, \Delta s \geq 0$, $\Delta gr \geq 0$, $\Delta fr \geq 0$. Note that if $\Delta gr = 1$ then both G_{i-1} and G_i have the greatest real-knots, $s_{i-1} = 0$ and $s_i = 1$. It implies $\Delta s = 1$ and $\Delta \leq 0$.

Note that $b_{2N} = b(\hat{\Pi}, \hat{\Gamma}) = b(\Pi, \Gamma)$, $c_{2N} = c(\hat{\Pi}, \hat{\Gamma})$, $p_{2N} = 0$, $r_{2N} = k(\hat{\Pi}, \hat{\Gamma})$, $s_{2N} = 0$, $gr_{2N} = 0$ and $fr_{2N} = f(\hat{\Pi}, \hat{\Gamma})$. Also $b_t = b(\Pi, \Gamma)$, $c_t = c(\Pi, \Gamma)$, $p_t = p(\Pi, \Gamma)$, $r_t = rr(\Pi, \Gamma)$, $s_t = s(\Pi, \Gamma)$, $gr_t = gr(\Pi, \Gamma)$ and $fr_t = fr(\Pi, \Gamma)$. Therefore for an optimal capping $\hat{\Gamma}$:

$$d(\hat{\Pi}, \hat{\Gamma}) = b(\hat{\Pi}, \hat{\Gamma}) - c(\hat{\Pi}, \hat{\Gamma}) + k(\hat{\Pi}, \hat{\Gamma}) + f(\hat{\Pi}, \hat{\Gamma}) = b(\hat{\Pi}, \hat{\Gamma}) - c_{2N} + p_{2N} + r_{2N} + \lceil \frac{s_{2N} - gr_{2N} + fr_{2N}}{2} \rceil \geq$$
$$b_t - c_t + p_t + r_t + \lceil \frac{s_t - gr_t + fr_t}{2} \rceil = b(\Pi, \Gamma) - c(\Pi, \Gamma) + p(\Pi, \Gamma) + rr(\Pi, \Gamma) + \lceil \frac{s(\Pi,\Gamma) - gr(\Pi,\Gamma) + fr(\Pi,\Gamma)}{2} \rceil$$

We now prove that there exists a capping $\hat{\Gamma}$ such that $c_{2N} - p_{2N} - r_{2N} - \lceil \frac{s_{2N} - gr_{2N} + fr_{2N}}{2} \rceil = c_t - p_t - r_t - \lceil \frac{s_t - gr_t + fr_t}{2} \rceil$ by building a sequence of $2N - t$ gray edges $g_{t+1}, \ldots, g_{2N}$ connecting Π-caps with Γ-tails in $\overline{G}$ such that $\Delta_i = 0$ for all $t + 1 \leq i \leq 2N$. The algorithm *Genomic_Sort* building this sequence of edges is shown in Fig. 4.

Closing a $\Pi\Gamma$-path inside a component having more than one $\Pi\Gamma$-path inside it (line 3) does not affect any of the parameters and hence $\Delta = 0$ for the gray edge closing such a path.

Connecting a $\Pi\Pi$-path with a $\Gamma\Gamma$-path via an interchromosomal or oriented edge (line 6) affects only two parameters ($\Delta c = -1$, $\Delta p = -1$) and $\Delta = 0$.

When number of semi-knots is greater than 2, for an edge "destroying" 2 semi-knots (line 8), $\Delta c = -1$ and $\Delta s = -2$. Other parameters do not change and hence, $\Delta = 0$.

When number of semi-knots is 2 and $gr_{i-1} = 1$ then for the edge closing the $\Pi\Gamma$-path in one of the semi-knots (line 11), $\Delta c = \Delta r = \Delta p = 0$ and $\Delta s = -1$. Clearly, $gr_i = 0$ hence $\Delta gr = -1$. Moreover $fr_{i-1} = fr_i$, hence $\Delta fr = 0$. Thus $\Delta = 0$. If the number of semi-knots is 2 and $gr_{i-1} = 0$ then for an edge "destroying" the two semi-knots (line 13), $\Delta c = -1$ and $\Delta s = -2$. Other parameters do not change, hence, $\Delta = 0$.

For the edge closing the $\Pi\Gamma$-path in the only semi-knot (line 15), if $gr_{i-1} = 1$ then $\Delta c = \Delta p = \Delta r = 0$, $\Delta s = -1$, $\Delta gr = -1$, $\Delta fr = 0$, hence $\Delta = 0$. Else if $gr_{i-1} = 0$, $\Delta c = \Delta p$, $\Delta r = 1$, $\Delta s = -1$, $\Delta gr = 0$, $fr_i = 0$. It can be verified that $\Delta = 0$ in this case.

Closing any other $\Pi\Gamma$-path (line 17) doesn't affect any parameters and hence $\Delta = 0$.

■

7 Applications and Open Problems

To derive human and mouse gene orders we used comparative mapping data from the Mouse Genome Database (Jackson Laboratory). Deriving gene orders is a non-trivial task since the map accuracy in human is significantly lower than in mouse (mice are much easier to breed!) and for some closely located genes in human the relative ordering is still unknown. Moreover, despite the fact that the average number of genes in a human-mouse "conserved block" is about 10, some of the blocks consist of a single gene thus making it impossible to infer a sign of these blocks. These problems forced us to make a number of arbitrary decisions while deriving the order of synteny groups in human and mouse. The rapid progress in human-mouse comparative mapping leaves no doubt that in a few years the complete and unambiguous information about human-mouse synteny groups will be obtained.

Centromeres represent another difficulty in analyzing chromosomal rearrangements. We have chosen to ignore the positions of centromeres since the molecular structure and evolution of centromeres are very poorly understood. In particular it is unclear whether a transformation of an *inactive* centromere into an *active* one and vice versa is a frequent evolutionary event. From this perspective, ignoring centromeres might be the most reasonable approach at the moment. We also ignore *transpositions* since they are extremely rare in chromosome evolution.

Under all these limitations we derived a human-mouse gene order consisting of 138 conserved gene blocks. For this (tentative) gene order, a most parsimonious scenario of human-mouse evolution involves 131 reversals/translocations/fusions/fissions, thus "improving" the Nadeau and Taylor, 1984 and more recent Copeland et al., 1993 estimates. Note that our estimate is constructive unlike all the previous estimates. At the same time this estimate should be taken with caution until a more reliable gene order is produced by experts in human-mouse comparative mapping.

Of course, gene orders for just two genomes are hardly sufficient to delineate a correct rearrangement scenario. Comparative gene mapping has made possible the generation of comparative maps for 28 species representing different mammalian orders (O'Brien and Graves, 1991). However, the resolution of these maps is significantly lower than the resolution of the human-mouse map. Since conventional comparative physical mapping is very laborious and time consuming, one can hardly expect that the tremendous efforts involved in obtaining the human-mouse map will be repeated for other mammalian genomes. However, a newly developed experimental technique called *chromosome painting* allows one to derive gene order without actually building an accurate "gene-based" map! In past, applications of chromosome painting were limited to primates (Jauch et al., 1992), and attempts to extend this approach to other mammals were not successful because of DNA sequence diversity between distantly related species. Very recently, an improved version of chromosome painting, called *ZOO-FISH* that is capable of detecting homologous chromosome fragments in distant mammalian species was developed (Scherthan et al., 1994). In April, 1995 Rettenberger et al., reported successful completion of the human-pig chromosome painting project. In a relatively inexpensive experiment Rettenberger et al., 1995 identified 47 conserved blocks common to human and pigs and used these data for analyzing human-pig evolution. The success of the human-pig chromosome painting project indicates that gene orders of many mammalian species can be generated with ZOO-FISH inexpensively. This provides an invaluable new source of data to attack a 100-years old problem of mammalian evolution with a new *constructive* approach versus previous ones based on the statistics of point mutations. This paper makes the first step in this direction but the problem of analyzing genome rearrangements in *multiple* genomes remains open.

8 Acknowledgments

We are indebted to Joe Nadeau for many helpful insights on biology of genome rearrangements and comparative human-mouse physical mapping. We are also grateful to Vineet Bafna and Webb Miller for many helpful comments and Jannan Eppig for her help with the Mouse Genome Database.

References

[1] M. Aigner and D. B. West. Sorting by insertion of leading element. *Journal of Combinatorial Theory*, 45:306–309, 1987.

[2] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. In *34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993. (to appear in SIAM J. Computing).

[3] V. Bafna and P. Pevzner. Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12:239–246, 1995a.

[4] V. Bafna and P. Pevzner. Sorting by transpositions. In *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 614–623, 1995b.

[5] D. Cohen and M. Blum. Improved bounds for sorting pancakes under a conjecture. 1993 (manuscript).

[6] N. G. Copeland, N. A. Jenkins, D. J. Gilbert, J. T. Eppig, L. J. Maltals, J. C. Miller, W. F. Dietrich, A. Weaver, S. E. Lincoln, R. G. Steen, L. D. Steen, J. H. Nadeau, and E. S. Lander. A genetic linkage map of the mouse: Current applications and future prospects. *Science*, 262:57–65, 1993.

[7] T. Dobzhansky and A.H.Sturtevant. Inversions in the chromosomes of *drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.

[8] S. Even and O. Goldreich. The minimum-length generator sequence problem is NP-hard. *Journal of Algorithms*, 2:311–313, 1981.

[9] W. H. Gates and C. H. Papadimitriou. Bounds for sorting by prefix reversals. *Discrete Mathematics*, 27:47–57, 1979.

[10] S. Hannenhalli. Polynomial algorithm for computing translocation distance between genomes. In *Combinatorial Pattern Matching, Proc. 6th Annual Symposium (CPM'95)*, Lecture Notes in Computer Science, pages 162–176. Springer-Verlag, Berlin, 1995.

[11] S. Hannenhalli, C. Chappey, E. Koonin, and P. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: A test case. In *Genomics*, 1995. (to appear).

[12] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Annual ACM Symposium on the Theory of Computing*, pages 178–189, 1995a.

[13] M. Heydari and I. H. Sudborough. On sorting by prefix reversals and the diameter of pancake networks. 1993 (manuscript).

[14] A. Jauch, J. Wienberg, Stanyon, N. Arnold, S. Tofanelli, T. Ishida, and T. Cremer. Reconstruction of genomic rearrangements in great apes gibbons by chromosome painting. *Proc. Natl. Acad. Sci.*, 89:8611–8615, 1992.

[15] M. Jerrum. The complexity of finding minimum-length generator sequences. *Theoretical Computer Science*, 36:265–289, 1985.

[16] J. Kececioglu and D. Gusfield. Reconstructing a history of recombinations from a set of sequences. In *5th Annual ACM-SIAM Symp. on Discrete Algorithms*, pages 471–480, 1994.

[17] J. Kececioglu and R. Ravi. Of mice and men: Evolutionary distances between genomes under translocation. In *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 604–613, 1995.

[18] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Combinatorial Pattern Matching, Proc. 4th Annual Symposium (CPM'93)*, volume 684 of *Lecture Notes in Computer Science*, pages 87–105. Springer-Verlag, Berlin, 1993. (Extended version has appeared in Algorithmica, 13: 180-210, 1995.).

[19] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. In *Combinatorial Pattern Matching, Proc. 5th Annual Symposium (CPM'94)*, volume 807 of *Lecture Notes in Computer Science 807*, pages 307–325. Springer-Verlag, Berlin, 1994.

[20] J. H. Nadeau and B. A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA*, 81:814–818, 1984.

[21] S. O'Brien and J. Graves. Report of the committee on comparative gene mapping in mammals. *Cytogenet. Cell Genet.*, 58:1124–1151, 1991.

[22] P.A. Pevzner and M.S. Waterman. Open combinatorial problems in computational molecular biology. In *3rd Israel Symposium on Theory of Computing and Systems*, pages 158–163. IEEE Computer Society Press, 1995.

[23] G. Rettenberger, C. Klett, U. Zechner, J. Kunz, W. Vogel, and H. Hameister. Visualization of the conservation of synteny between humans and pigs by hetereologous chromosomal painting. *Genomics*, 26:372–378, 1995.

[24] 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. *Proc. Natl. Acad. Sci. USA*, 89:6575–6579, 1992.

[25] H. Scherthan, T. Cremer, U. Arnason, H. Weier, A. Lima de Faria, and L. Fronicke. Comparative chromosomal painting discloses homologous segments in distantly related mammals. *Nature Genetics*, 6:342–347, April 1994.

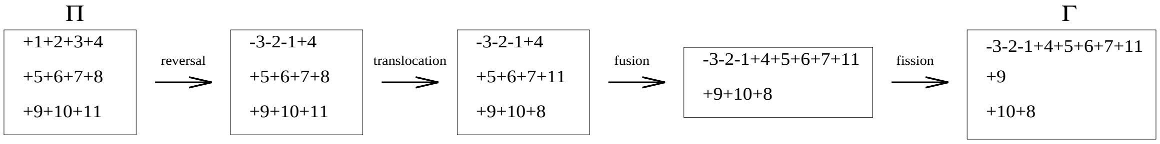

Figure 1: Evolution of genome Π into genome Γ

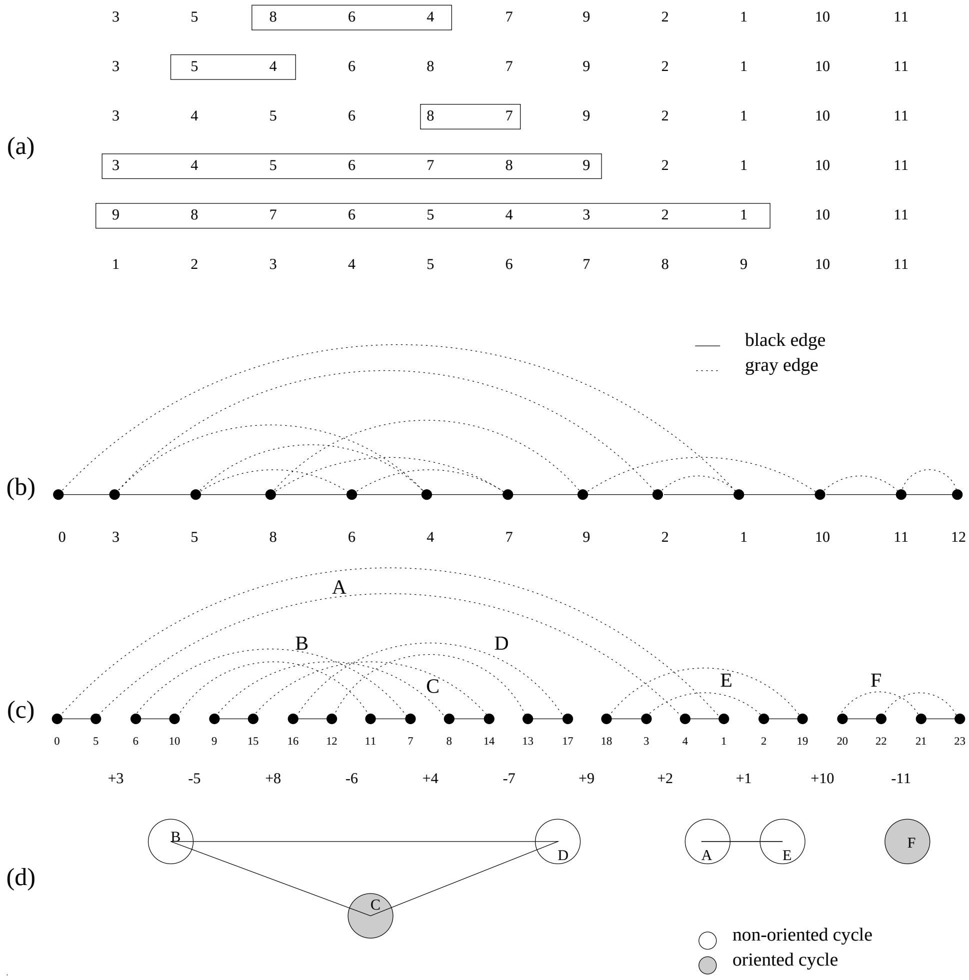

Figure 2: (a) Optimal sorting of a permutation $\sigma = (3\ 5\ 8\ 6\ 4\ 7\ 9\ 2\ 1\ 10\ 11)$ by 5 reversals. (b) Breakpoint graph $G(\sigma)$. (c) Transformation of a signed permutation into an unsigned permutation π and the breakpoint graph $G(\pi)$. Gray edges $(8,9)$ and $(22,23)$ are oriented while 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.

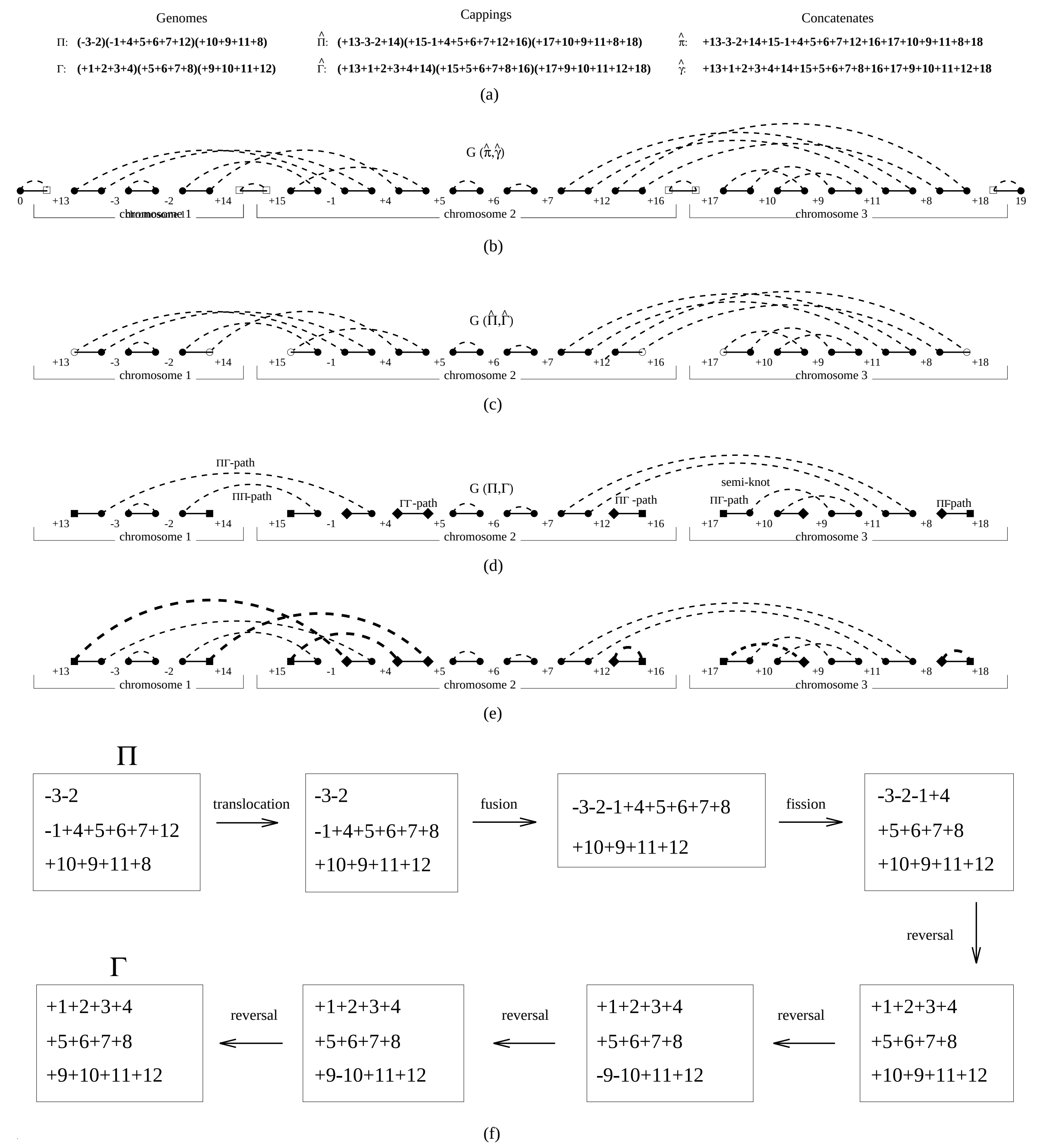

Figure 3: (a) Genomes Π and Γ, cappings $\hat{\Pi}$ and $\hat{\Gamma}$ and concatenates $\hat{\pi}$ and $\hat{\gamma}$. (b) Graph $G(\hat{\pi}, \hat{\gamma})$. Tails are shown as white boxes. (c) Graph $G(\hat{\Pi}, \hat{\Gamma})$ is obtained from $G(\hat{\pi}, \hat{\gamma})$ by deleting the tails. Caps are shown as white circles. (d) Graph $G(\Pi, \Gamma)$ with 4 cycles and 6 paths ($c(\Pi, \Gamma) = 10$). Π-caps are shown as boxes while Γ-tails are shown by diamonds. For genomes Π and Γ, $b(\Pi, \Gamma) = 15$, $r(\Pi, \Gamma) = 0$, $p(\Pi, \Gamma) = 1$, $s(\Pi, \Gamma) = 1$, $gr(\Pi, \Gamma) = fr(\Pi, \Gamma) = 0$. Therefore $d(\Pi, \Gamma) = 15 - 10 + 1 + 0 + \lceil \frac{1-0+0}{2} \rceil = 7$. (e) Graph $G(\hat{\Pi}, \hat{\Gamma})$ corresponding to an optimal capping of $\hat{\Gamma} = (+13 + 1 + 2 + 3 + 4 - 15)(-14 + 5 + 6 + 7 + 8 + 18)(+17 + 9 + 10 + 11 + 12 + 16)$. Added gray edges are shown by thick dashed lines. (f) optimal sorting of Π into Γ with seven rearrangements.

Algorithm *Genomic_Sort*(Π, Γ)

Construct the graph $G = G(\Pi, \Gamma)$
Close all $\Pi\Gamma$-paths in simple components of $G(\Pi, \Gamma)$ (lemma 6)
Close all but one $\Pi\Gamma$-path in components having more than one $\Pi\Gamma$-path inside them
while G contains a path
    if there exists a $\Pi\Pi$-path in G
        find an interchromosomal or an oriented edge g joining this $\Pi\Pi$-path with a $\Gamma\Gamma$-path (lemma 4)
    elseif G has more than 2 semi-knots
        find an interchromosomal or an oriented edge g joining $\Pi\Gamma$-paths in any two semi-knots (lemma 5)
    elseif G has 2 semi-knots
        if G has the greatest real-knot
            find an edge g closing the $\Pi\Gamma$-path in one of these semi-knots
        else
            find an interchromosomal or an oriented edge g joining $\Pi\Gamma$-paths in these semi-knots (lemma 5)
    elseif G has 1 semi-knot
        find edge g closing the $\Pi\Gamma$-path in this semi-knot
    else
        find edge g closing arbitrary $\Pi\Gamma$-path
    add edge g to the graph G, i.e $G \leftarrow G + \{g\}$
find a capping $\hat{\Gamma}$ defined by the graph $G = G(\hat{\Pi}, \hat{\Gamma})$
sort genome $\hat{\Pi}$ into $\hat{\Gamma}$ (theorem 3, Hannenhalli and Pevzner, 1995)
sorting of $\hat{\Pi}$ into $\hat{\Gamma}$ mimics sorting of Π into Γ

Figure 4: Algorithm *Genomic_Sort*