Chapter 10

Genome Rearrangements

10.1 Introduction

Genome Comparison versus Gene Comparison In the late 1980s, Jeffrey Palmer and his colleagues discovered a remarkable and novel pattern of evolutionary change in plant organelles. They compared the mitochondrial genomes of *Brassica oleracea* (cabbage) and *Brassica campestris* (turnip), which are very closely related (many genes are 99% identical). To their surprise, these molecules, which are almost identical in gene *sequences*, differ dramatically in gene *order* (Figure 10.1). This discovery and many other studies in the last decade convincingly proved that genome rearrangements represent a common mode of molecular evolution.

Every study of genome rearrangements involves solving a combinatorial "puzzle" to find a series of *rearrangements* that transform one genome into another. Three such rearrangements "transforming" cabbage into turnip are shown in Figure 10.1. Figure 1.5 presents a more complicated *rearrangement scenario* in which mouse X chromosome is transformed into human X chromosome. Extreme conservation of genes on X chromosomes across mammalian species (Ohno, 1967 [255]) provides an opportunity to study the evolutionary history of X chromosome independently of the rest of the genomes. According to Ohno's law, the gene content of X chromosomes has barely changed throughout mammalian development in the last 125 million years. However, the order of genes on X chromosomes has been disrupted several times.

It is not so easy to verify that the six evolutionary events in Figure 1.5 represent a *shortest* series of *reversals* transforming the mouse gene order into the human gene order on the X chromosome. Finding a shortest series of reversals between the gene order of the mitochondrial DNAs of worm *Ascaris suum* and humans presents an even more difficult computational challenge (Figure 10.2).

In cases of genomes consisting of a small number of "conserved blocks," Palmer and his co-workers were able to find the most parsimonious rearrangement

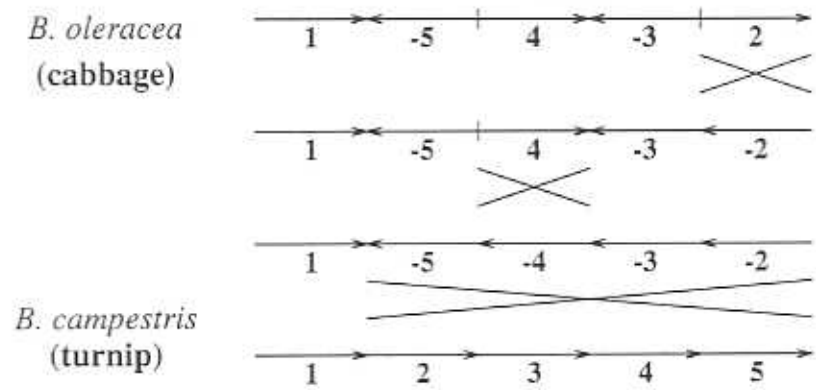

Figure 10.1: "Transformation" of cabbage into turnip.

scenarios. However, for genomes consisting of more than 10 blocks, exhaustive search over all potential solutions is far beyond the capabilities of "pen-and-pencil" methods. As a result, Palmer and Herbon, 1988 [259] and Makaroff and Palmer, 1988 [229] overlooked the most parsimonious rearrangement scenarios in more complicated cases such as turnip versus black mustard or turnip versus radish.

The traditional molecular evolutionary technique is a *gene* comparison, in which phylogenetic trees are being reconstructed based on point mutations of a single gene (or a small number of genes). In the "cabbage and turnip" case, the gene comparison approach is hardly suitable, since the rate of point mutations in cabbage and turnip mitochondrial genes is so low that their genes are almost identical. *Genome comparison* (i.e., comparison of gene orders) is the method of choice in the case of very slowly evolving genomes. Another example of an evolutionary problem for which genome comparison may be more conclusive than gene comparison is the evolution of rapidly evolving viruses.

Studies of the molecular evolution of herpes viruses have raised many more questions than they've answered. Genomes of herpes viruses evolve so rapidly that the extremes of present-day phenotypes may appear quite unrelated; the similarity between many genes in herpes viruses is so low that it is frequently indistinguishable from background noise. Therefore, classical methods of sequence comparison are not very useful for such highly diverged genomes; ventures into the quagmire of the molecular phylogeny of herpes viruses may lead to contradictions, since different genes give rise to different evolutionary trees. Herpes viruses have from 70 to about 200 genes; they all share seven conserved blocks that are rearranged in the genomes of different herpes viruses. Figure 10.3 presents different arrangements of these blocks in Cytomegalovirus (CMV) and Epstein-Barr virus (EBV) and a shortest series of reversals transforming CMV gene order into EBV gene

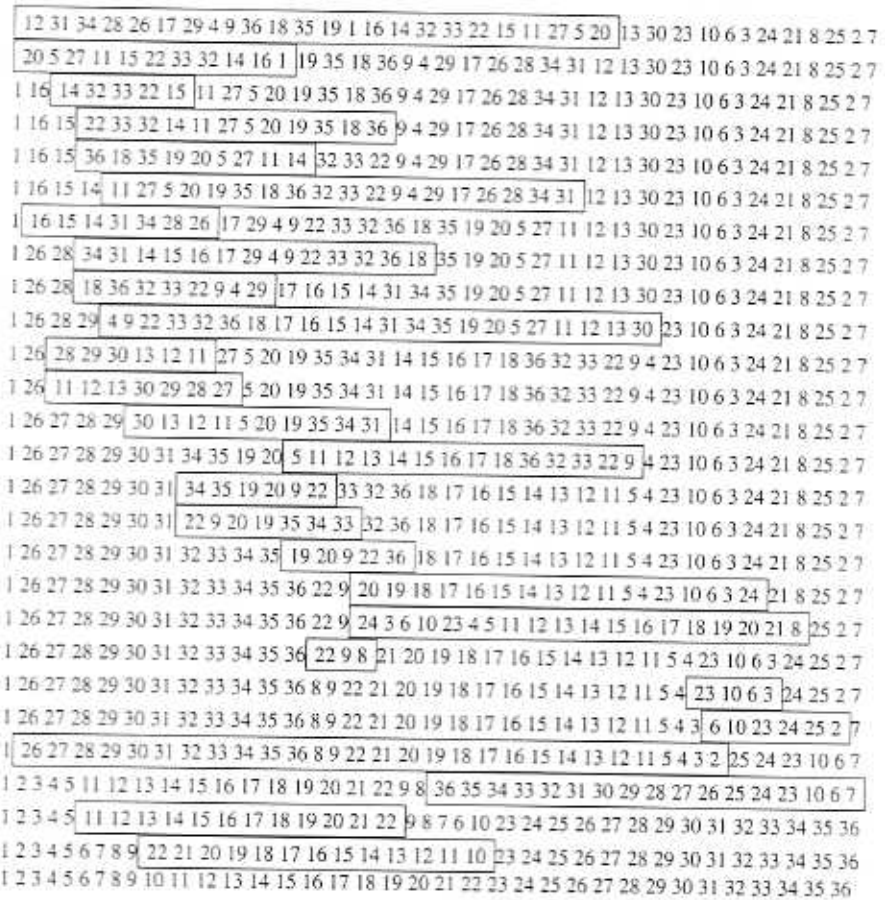

Figure 10.2: A most parsimonious rearrangement scenario for transformation of worm *Ascaris Suum* mitochondrial DNA into human mitochondrial DNA (26 reversals).

order (Hannenhalli et al., 1995 [152]). The number of such large-scale rearrangements (five reversals) is much smaller than the number of point mutations between CMV and EBV (hundred(s) of thousands). Therefore, the analysis of such rearrangements at the *genome* level may complement the analysis at the *gene* level traditionally used in molecular evolution. Genome comparison has certain merits and demerits as compared to classical gene comparison: genome comparison ignores actual DNA sequences of genes, while gene comparison ignores gene order. The ultimate goal would be to combine the merits of both genome and gene comparison in a single algorithm.

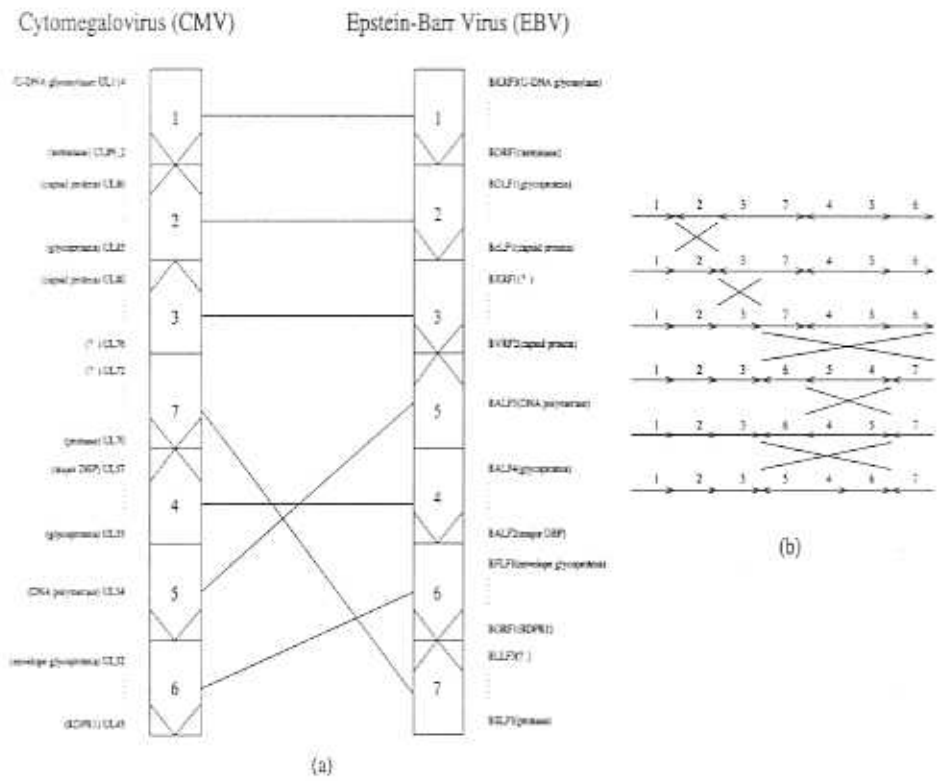

Figure 10.3: Comparative genome organization (a) and the shortest series of rearrangements transforming CMV gene order into EBV gene order (b).

The analysis of genome rearrangements in molecular biology was pioneered in the late 1930s by Dobzhansky and Sturtevant, who published a milestone paper presenting a rearrangement scenario with 17 inversions for the species of *Drosophila* fruit fly (Dobzhansky and Sturtevant, 1938 [87]). With the advent of large-scale mapping and sequencing, the number of *genome comparison* problems is rapidly growing in different areas, including viral, bacterial, yeast, plant, and animal evolution.

Sorting by Reversals A computational approach based on comparison of gene orders was pioneered by David Sankoff (Sankoff et al.; 1990, 1992 [302, 304] and Sankoff, 1992 [300]). Genome rearrangements can be modeled by a combinatorial problem of sorting by reversals, as described below. The order of genes in two

organisms is represented by permutations $\pi = \pi_1\pi_2\ldots\pi_n$ and $\sigma = \sigma_1\sigma_2\ldots\sigma_n$. A *reversal* $\rho(i,j)$ of an interval $[i,j]$ is the permutation

$$\begin{pmatrix} 1\ 2\ \ldots\ i-1\ \mathbf{i}\ \mathbf{i+1}\ \ldots\ \mathbf{j-1}\ \mathbf{j}\ j+1\ \ldots\ n \\ 1\ 2\ \ldots\ i-1\ \mathbf{j}\ \mathbf{j-1}\ \ldots\ \mathbf{i+1}\ \mathbf{i}\ j+1\ \ldots\ n \end{pmatrix}$$

Clearly $\rho(i,j)$ has the effect of reversing the order of $\pi_i\pi_{i+1}\ldots\pi_j$ and transforming $\pi_1\ldots\pi_{i-1}\pi_i\ldots\pi_j\pi_{j+1}\ldots\pi_n$ into $\pi\cdot\rho(i,j) = \pi_1\ldots\pi_{i-1}\pi_j\ldots\pi_i\pi_{j+1}\ldots\pi_n$.

Given 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 minimal. 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 $(12\ldots n)$.

Computer scientists have studied a related *sorting by prefix reversals* problem (also known as the *pancake flipping problem*): given an arbitrary permutation π, find $d_{pref}(\pi)$, which is the minimum number of reversals of the form $\rho(1,i)$ sorting π. The pancake flipping problem was inspired by the following "real-life" situation described by Harry Dweigter:

The chef in our place is sloppy, and when he prepares a stack of pancakes they come out all different sizes. Therefore, when I deliver them to a customer, on the way to a table I rearrange them (so that the smallest winds up on top, and so on, down to the largest at the bottom) by grabbing several from the top and flipping them over, repeating this (varying the number I flip) as many times as necessary. If there are n pancakes, what is the maximum number of flips that I will ever have to use to rearrange them?

Bill Gates (an undergraduate student at Harvard in late 1970s, now at Microsoft) and Cristos Papadimitriou made the first attempt to solve this problem (Gates and Papadimitriou, 1979 [120]). They proved that the *prefix reversal diameter* of the symmetric group, $d_{pref}(n) = \max_{\pi\in S_n} d_{pref}(\pi)$, is less than or equal to $\frac{5}{3}n + \frac{5}{3}$, and that for infinitely many n, $d_{pref}(n) \geq \frac{17}{16}n$. The pancake flipping problem still remains unsolved.

The Breakpoint Graph What makes it hard to sort a permutation? In the very first computational studies of genome rearrangements, Watterson et al., 1982 [366] and Nadeau and Taylor, 1984 [248] introduced the notion of a *breakpoint* and noticed some correlations between the reversal distance and the number of breakpoints. (In fact, Sturtevant and Dobzhansky, 1936 [331] implicitly discussed these correlations 60 years ago!) Below we define the notion of a breakpoint.

Let $i \sim j$ if $|i-j| = 1$. Extend a permutation $\pi = \pi_1\pi_2\ldots\pi_n$ by adding $\pi_0 = 0$ and $\pi_{n+1} = n+1$. We call a pair of elements (π_i, π_{i+1}), $0 \leq i \leq n$, of π an *adjacency* if $\pi_i \sim \pi_{i+1}$, and a *breakpoint* if $\pi_i \not\sim \pi_{i+1}$ (Figure 10.4). As the identity permutation has no breakpoints, sorting by reversals corresponds to

eliminating breakpoints. An observation that every reversal can eliminate *at most* 2 breakpoints immediately implies that $d(\pi) \geq \frac{b(\pi)}{2}$, where $b(\pi)$ is the number of breakpoints in π. Based on the notion of a breakpoint, Kececioglu and Sankoff, 1995 [194] found an approximation algorithm for sorting by reversals with performance guarantee 2. They also devised efficient bounds, solving the reversal distance problem almost optimally for n ranging from 30 to 50. This range covers the biologically important case of animal mitochondrial genomes.

However, the estimate of reversal distance in terms of breakpoints is very inaccurate. Bafna and Pevzner, 1996 [19] showed that another parameter (size of a maximum cycle decomposition of the breakpoint graph) estimates reversal distance with much greater accuracy.

The *breakpoint graph* of a permutation π is an edge-colored graph $G(\pi)$ with $n+2$ vertices $\{\pi_0, \pi_1, \ldots, \pi_n, \pi_{n-1}\} \equiv \{0, 1, \ldots, n, n+1\}$. We join vertices π_i and π_{i+1} by a *black* edge for $0 \leq i \leq n$. We join vertices π_i and π_j by a *gray* edge if $\pi_i \sim \pi_j$. Figure 10.4 demonstrates that a breakpoint graph is obtained by a superposition of a black path traversing the vertices $0, 1, \ldots, n, n+1$ in the order given by permutation π and a gray path traversing the vertices in the order given by the identity permutation.

A *cycle* in an edge-colored graph G is called *alternating* if the colors of every two consecutive edges of this cycle are distinct. In the following, by cycles we mean alternating cycles. A vertex v in a graph G is called *balanced* if the number of black edges incident to v equals the number of gray edges incident to v. A *balanced graph* is a graph in which every vertex is balanced. Clearly $G(\pi)$ is a balanced graph: therefore, it contains an alternating Eulerian cycle. Therefore, there exists a *cycle decomposition* of $G(\pi)$ into edge-disjoint alternating cycles (every edge in the graph belongs to exactly one cycle in the decomposition). Cycles in an edge decomposition may be self-intersecting. The breakpoint graph in Figure 10.4 can be decomposed into four cycles, one of which is self-intersecting. We are interested in the decomposition of the breakpoint graph into a *maximum* number $c(\pi)$ of edge-disjoint alternating cycles. For the permutation in Figure 10.4, $c(\pi) = 4$.

Cycle decompositions play an important role in estimating reversal distance. When we apply a reversal to a permutation, the number of cycles in a maximum decomposition can change by at most one (while the number of breakpoints can change by two). Bafna and Pevzner, 1996 [19] proved the bound $d(\pi) \geq n + 1 - c(\pi)$, which is much tighter than the bound in terms of breakpoints $d(\pi) \geq b(\pi)/2$. For most biological examples, $d(\pi) = n + 1 - c(\pi)$, thus reducing the reversal distance problem to the maximal cycle decomposition problem.

Duality Theorem for Signed Permutations Finding a maximal cycle decomposition is a difficult problem. Fortunately, in the more biologically relevant case of *signed permutations*, this problem is trivial. Genes are *directed* fragments of DNA,

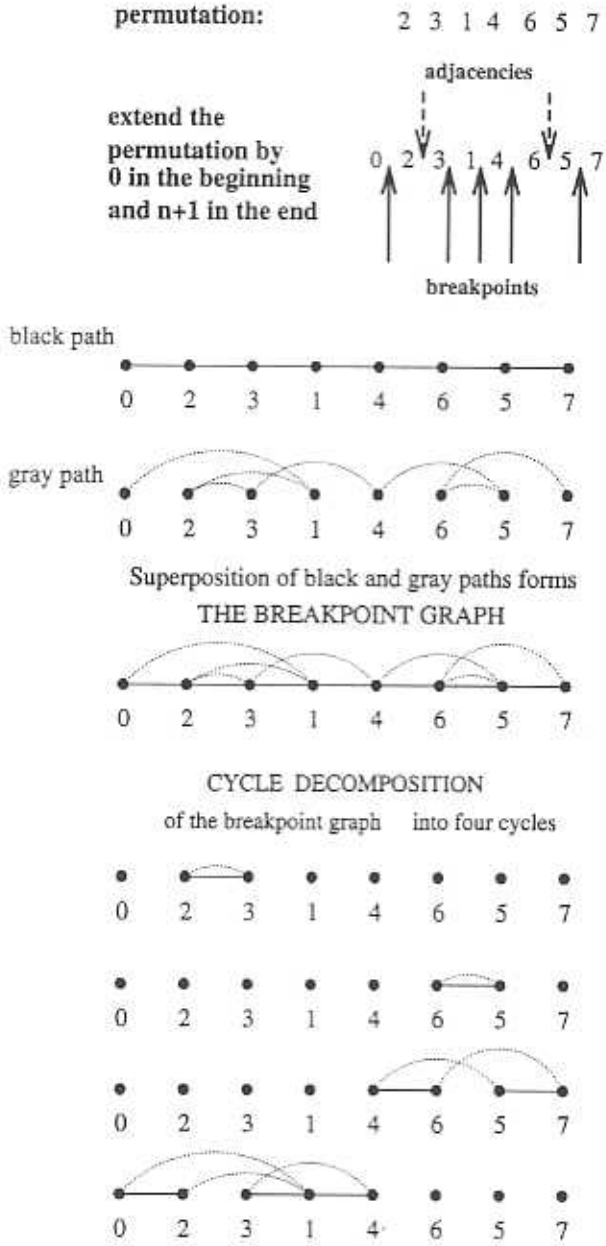

Figure 10.4: Breakpoints, breakpoint graph, and maximum cycle decomposition.

and a sequence of n genes in a genome is represented by a *signed* permutation on $\{1, \ldots n\}$ with a $+$ or $-$ sign associated with every element of π. For example, the gene order for *B. oleracea* presented in Figure 10.1 is modeled by the signed permutation $(+1 - 5 + 4 - 3 + 2)$. In the signed case, every reversal of fragment $[i, j]$ changes both the order and the signs of the elements within that fragment (Figure 10.1). We are interested in the minimum number of reversals $d(\pi)$ required to transform a signed permutation π into the identity signed permutation $(+1 + 2 \ldots + n)$.

Bafna and Pevzner, 1996 [19] noted that the concept of a breakpoint graph extends naturally to signed permutations by mimicking every directed element i by two undirected elements i_a and i_b, which substitute for the tail and the head of the directed element i (Figure 10.5).

For signed permutations, the bound $d(\pi) \geq n + 1 - c(\pi)$ approximates the reversal distance extremely well for both simulated and biological data. This intriguing performance raises the question of whether the bound $d(\pi) \geq n+1-c(\pi)$ overlooks another parameter (in addition to the size of a maximum cycle decomposition) that would allow closing the gap between $d(\pi)$ and $n+1-c(\pi)$. Hannenhalli and Pevzner, 1995 [154] revealed another "hidden" parameter (number of *hurdles* in π) making it harder to sort a signed permutation and showed that

$$n + 1 - c(\pi) + h(\pi) \leq d(\pi) \leq n + 2 - c(\pi) + h(\pi) \qquad (10.1)$$

where $h(\pi)$ is the number of hurdles in π. They also proved the duality theorem for signed permutations and developed a polynomial algorithm for computing $d(\pi)$.

Unsigned Permutations and Comparative Physical Mapping Since sorting (unsigned) permutations by reversals is NP-hard (Caprara, 1997 [57]), many researchers have tried to devise a practical approximation algorithm for sorting (unsigned permutations) by reversals.

A *block* of π is an interval $\pi_i \ldots \pi_j$ containing no breakpoints, i.e., (π_k, π_{k+1}) is an adjacency for $0 \leq i \leq k < j \leq n+1$. Define a *strip* of π as a maximal block, i.e., a block $\pi_i \ldots \pi_j$ such that (π_{i-1}, π_i) and (π_j, π_{j+1}) are breakpoints. A strip of one element is called a *singleton*, a strip of two elements is called a *2-strip*, and a strip with more than two elements is called a *long strip*. It turns out that singletons cause a major challenge in sorting unsigned permutations by reversals.

A reversal $\rho(i, j)$ *cuts* a strip $\pi_k \ldots \pi_l$ if either $k < i \leq l$ or $k \leq j < l$. A reversal cutting a strip separates elements that are consecutive in the identity permutation. Therefore, it is natural to expect that for every permutation π there exists an (optimal) sorting of π by reversals that does not cut strips. This, however, is false. Permutation 3412 requires three reversals if we do not cut strips, and yet it can be sorted with two: $3412 \rightarrow 1432 \rightarrow 1234$. Kececioglu and Sankoff, 1993 [192] conjectured that every permutation has an optimal sorting by reversals that does not cut long strips and does not increase the number of breakpoints.

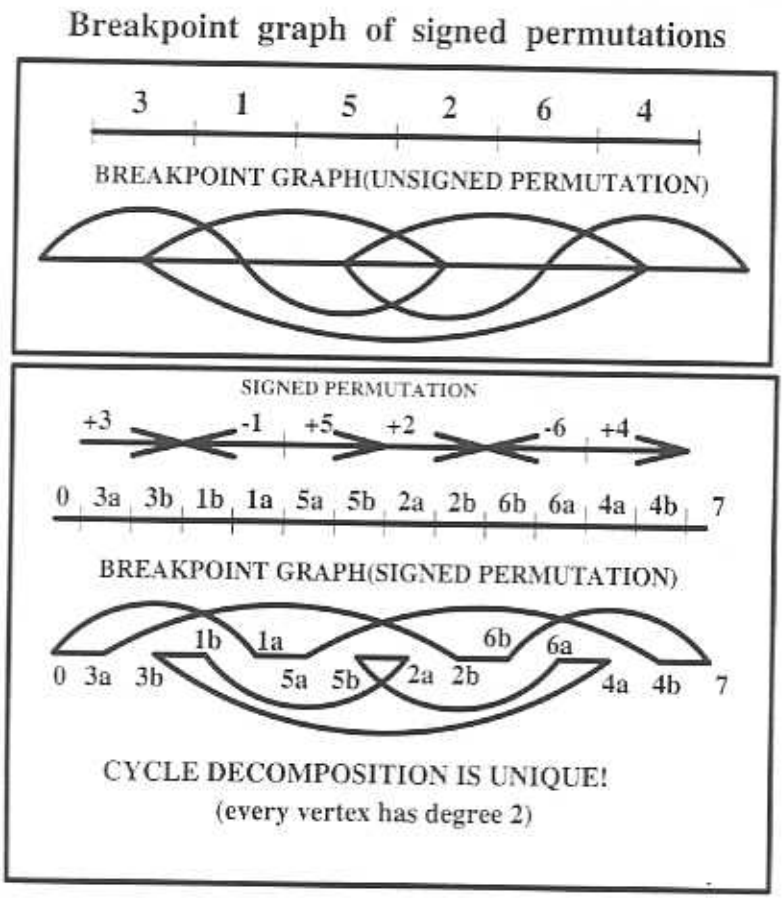

Figure 10.5: Modeling a signed permutation by an unsigned permutation.

Since the identity permutation has no breakpoints, sorting by reversals corresponds to eliminating breakpoints. From this perspective, it is natural to expect that for every permutation there exists an optimal sorting by reversals that never increases the number of breakpoints. Hannenhalli and Pevzner, 1996 [155] proved both the "reversals do not cut long strips" and the "reversals do not increase the number of breakpoints" conjectures by using the duality theorem for signed permutations.

Biologists derive gene orders either by sequencing entire genomes or by using comparative physical mapping. Sequencing provides information about the directions of genes and allows one to represent a genome by a signed permutation. However, sequencing of entire genomes is still expensive, and most currently avail-

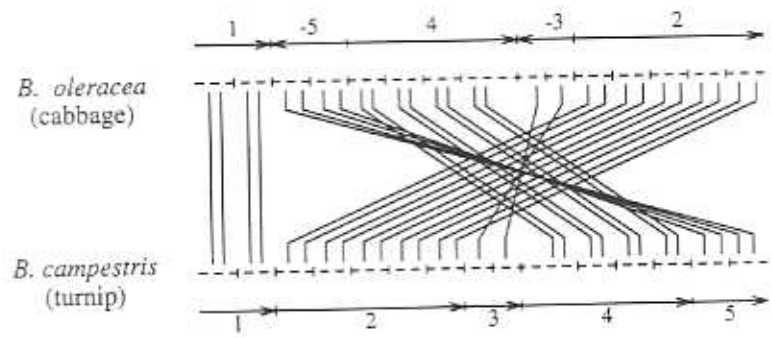

Figure 10.6: Comparative physical map of cabbage and turnip (unsigned permutation) and corresponding signed permutation.

able experimental data on gene orders are based on comparative physical maps. Physical maps usually do not provide information about the directions of genes, and therefore lead to representation of a genome as an *unsigned* permutation π. Biologists try to derive a signed permutation from this representation by assigning a positive (negative) sign to increasing (decreasing) strips of π (Figure 10.6). The "reversals do not cut long strips" property provides a theoretical substantiation for such a procedure in the case of long strips. At the same time, for 2-strips this procedure might fail to find an optimal rearrangement scenario. Hannenhalli and Pevzner, 1996 [155] pointed to a biological example for which this procedure fails and described an algorithm fixing this problem.

Permutations without singletons are called *singleton-free* permutations. The difficulty in analyzing such permutations is posed by an alternative, *"to cut or not to cut"* 2-strips. A characterization of a set of 2-strips *"to cut"* (Hannenhalli and Pevzner, 1996 [155]) leads to a polynomial algorithm for sorting singleton-free permutations and to a polynomial algorithm for sorting permutations with a small number of singletons. The algorithm can be applied to analyze rearrangement scenarios derived from comparative physical maps.

Low-resolution physical maps usually contain many singletons and, as a result, rearrangement scenarios for such maps are hard to analyze. The Hannenhalli and Pevzner, 1996 [155] algorithm runs in polynomial time if the number of singletons is $O(\log n)$. This suggests that $O(\log n)$ singletons is the desired trade-off of resolution for comparative physical mapping in molecular evolution studies. If the number of singletons is large, a biologist might choose additional experiments (i.e., sequencing of some areas) to resolve the ambiguities in gene directions.

Figure 10.7: Evolution of genome Π into genome Γ.

Rearrangements of Multichromosomal Genomes When the Brothers Grimm described a transformation of a man into a mouse in the fairy tale "Puss in Boots," they could hardly have anticipated that two centuries later humans and mice would be the most genetically studied mammals. Man-mouse comparative physical mapping started 20 years ago, and currently a few thousand pairs of homologous genes are mapped in these species. As a result, biologists have found that the related genes in man and mouse are not chaotically distributed over the genomes, but form "conserved blocks" instead. Current comparative mapping data indicate that both human and mouse genomes are comprised of approximately 150 blocks which are "shuffled" in humans as compared to mice (Copeland et al., 1993 [74]). For example, the chromosome 7 in the human can be viewed as a mosaic of different genes from chromosomes 2, 5, 6, 11, 12, and 13 in the mouse (Fig 1.4). Shuffling of blocks happens quite rarely (roughly once in a million years), thus giving biologists hope of reconstructing a rearrangement scenario of human-mouse evolution. In their pioneering paper, Nadeau and Taylor, 1984 [248] estimated that surprisingly few genomic rearrangements (178 ± 39) have happened since the divergence of human and mouse 80 million years ago.

In the model we consider, every gene is represented by an integer whose *sign* ("+" or "–") reflects 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, we are interested in a most parsimonious scenario of *evolution* of Π into Γ, i.e., the shortest sequence of rearrangement events (defined below) required to transform Π into Γ. In the following we assume that Π and Γ contain the same set of genes. Figure 10.7 illustrates four 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 (i.e., the *directions* of chromosomes are irrelevant). The four most common elementary rearrangement events in multichromosomal genomes are *reversals*, *translocations*, *fusions*, and *fissions*, defined below.

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

A translocation $\rho(\pi, \sigma, n+1, 1)$ concatenates the chromosomes π and σ, resulting in a chromosome $\pi_1 \dots \pi_n\sigma_1 \dots \sigma_m$ 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 $1 < i < n$ "breaks" a chromosome π into two chromosomes $(\pi_1 \dots \pi_{i-1})$ and $(\pi_i \dots \pi_n)$. This translocation, leading to an increase in the number of (non-empty) chromosomes, is known as a *fission*. Fusions and fissions are rather common in mammalian evolution; for example, the major difference in the overall genome organization of humans and chimpanzees is the fusion of two chimpanzee chromosomes into one human chromosome.

Kececioglu and Ravi, 1995 [191] made the first attempt to analyze rearrangements of multichromosomal genomes. Their approximation algorithm addresses the case in which both genomes contain the same number of chromosomes. This is a serious limitation, since different organisms (in particular humans and mice) have different numbers of chromosomes. From this perspective, every realistic model of genome rearrangements should include fusions and fissions. It turns out that fusions and fissions present a major difficulty in analyzing genome rearrangements. Hannenhalli and Pevzner, 1995 [153] proved the duality theorem for multichromosomal genomes, which computes genomic distance in terms of seven parameters reflecting different combinatorial properties of sets of strings. Based on this result they found a polynomial algorithm for this problem.

The idea of the analysis is to concatenate N chromosomes of Π and Γ into permutations π and γ, respectively, and to mimic genomic sorting of Π into Γ by sorting π into γ by reversals. The difficulty with this approach is that there exist $N!2^N$ different concatenates for Π and Γ, and only some of them, called *optimal concatenates*, mimic an *optimal* sorting of Π into Γ. Hannenhalli and Pevzner, 1995 [153] introduced techniques called *flipping* and *capping* of chromosomes that allow one to find an optimal concatenate.

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 many mammalian species (O'Brien and Graves, 1991 [254]). However, the resolution of these maps is significantly lower than the resolution of the human-mouse map. Since comparative physical mapping is rather laborious, one can hardly expect that the tremendous effort involved in obtaining the human-mouse map will be repeated for other mammalian genomes. However, an experimental technique called *chromosome painting* allows one to derive gene order without actually building an accurate "gene-based" map. In the past, the applications of chromosome painting were limited to primates (Jauch et al., 1992 [178]); attempts to extend this approach to other mammals were not successful because of the DNA sequence diversity between distantly related species. Later, Scherthan et al., 1994 [307] developed an improved version of chromosome painting, called *ZOO-FISH*, that is capable of detecting homologous chromosome fragments in distant mammalian species. Using ZOO-FISH, Rettenberger et al., 1995 [284] quickly completed the human-pig chromosome painting project and identified 47 conserved blocks common to human and pig. The success of the human-pig chromosome painting project indicates that gene orders of many mammalian species can be generated with ZOO-FISH inexpensively, thus providing an invaluable new source of data to attack the 100-year-old problem of mammalian evolution.