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

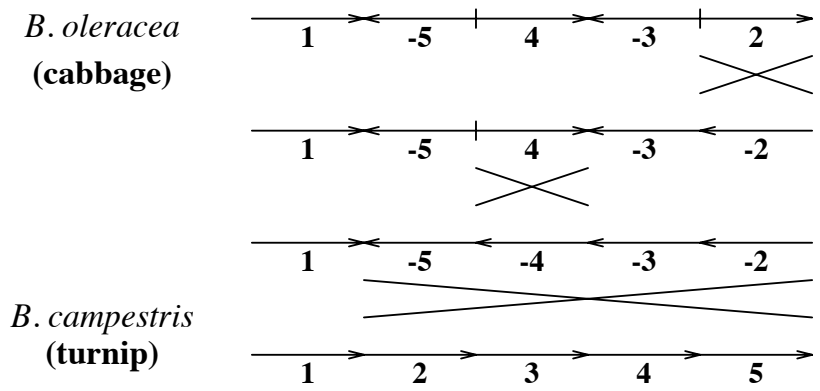

Figure 10.1: "Transformation" of cabbage into turnip.

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

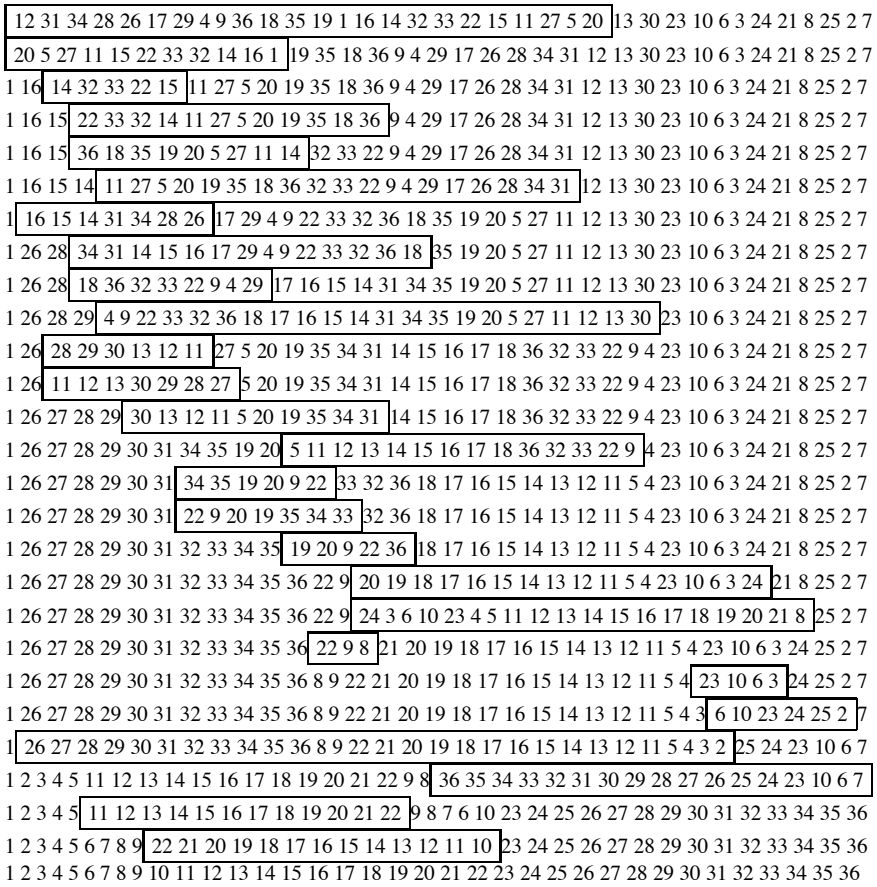

Figure 10.2: A most parsimonious rearrangement scenario for transformation of worm *Ascaris Suum* mitochondrial DNA into human mitochondrial DNA (26 reversals).

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.

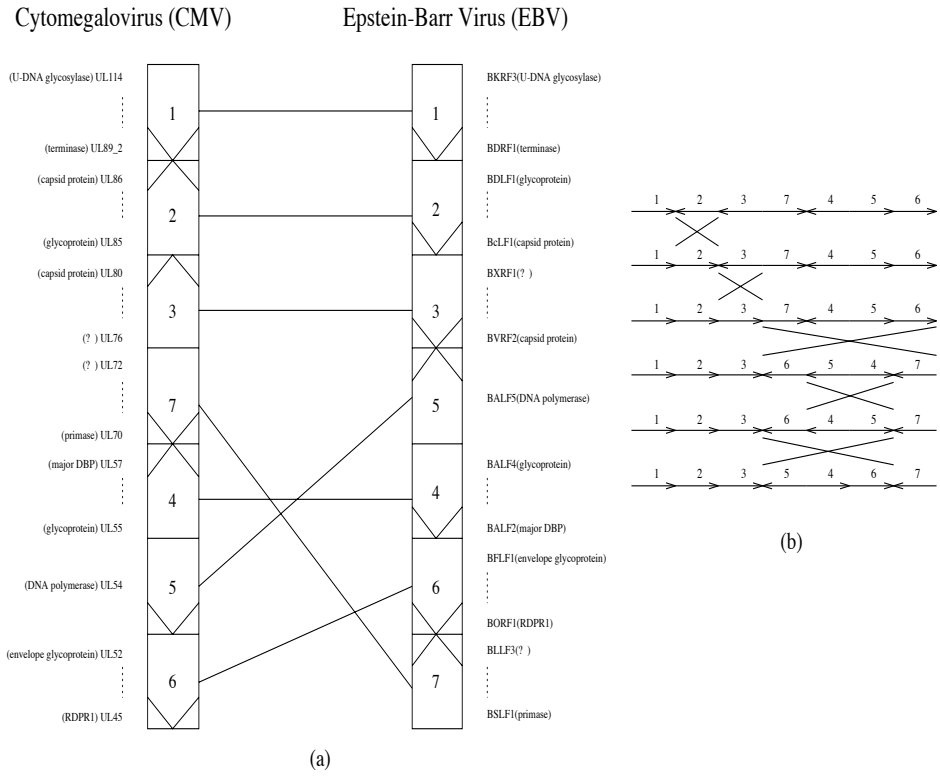

Figure 10.3: Comparative genome organization (a) and the shortest series of rearrangements transforming CMV gene order into EBV gene order (b).

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,

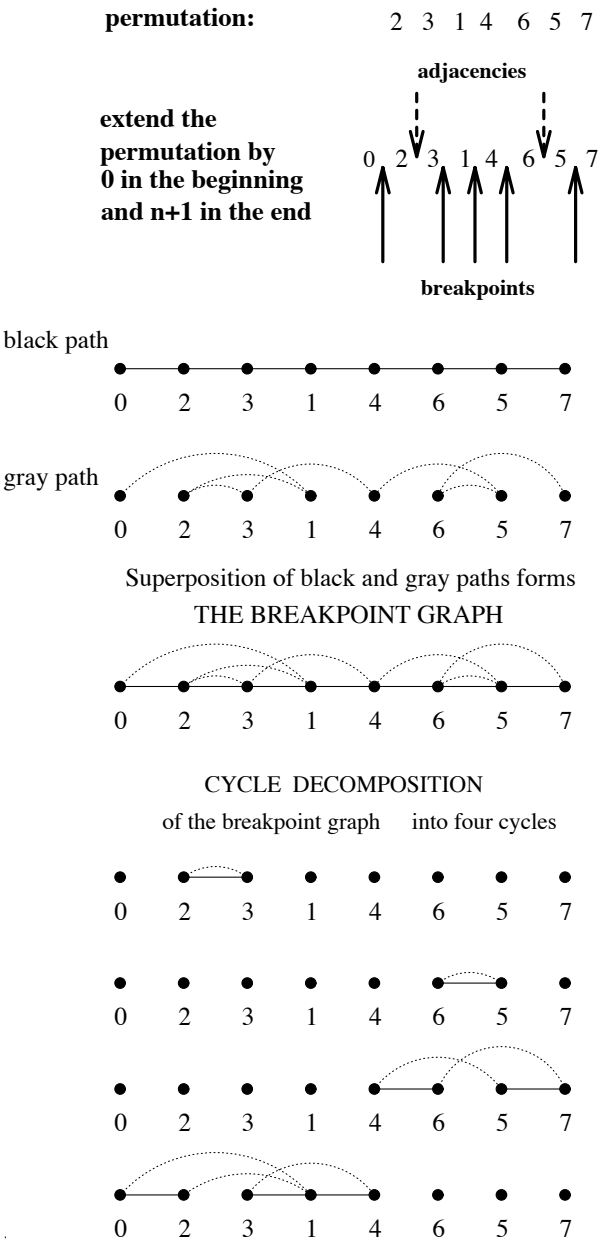

Figure 10.4: Breakpoints, breakpoint graph, and maximum cycle decomposition.