Combinatorics of Genome Rearrangements

Guillaume Fertin, Anthony Labarre, Irena Rusu, Éric Tannier and Stéphane Vialette

The MIT Press
Cambridge, Massachusetts
London, England

For information about special quantity discounts, please email special_sales@mitpress.mit.edu

This book was set in Times New Roman and Syntax on 3B2 by Asco Typesetters, Hong Kong.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Combinatorics of genome rearrangements / Guillaume Fertin . . . [et al.].
p. cm. — (Computational molecular biology)
Includes bibliographical references and index.
ISBN 978-0-262-06282-4 (hardcover : alk. paper) 1. Translocation (Genetics)—Mathematical models.
2. Translocation (Genetics)—Data processing. 3. Combinatorial analysis. 4. Genomics—Mathematics.
I. Fertin, Guillaume, 1972– II. Series.
[DNLM: 1. Gene Rearrangement. 2. Genome. 3. Models, Genetic. QU 470 C731 2009]
QH462.T7C66 2009
572.8′77—dc22 2008042152

10 9 8 7 6 5 4 3 2 1

1 Introduction

Although this book is combinatorially inclined and does not devote much discussion to the biological issues, we will start with a short introduction to molecular evolution, for conceptual and historical purposes; indeed, this is where the combinatorial study of genome rearrangements originates, and the invention of most variants of genome rearrangement problems are still driven by biological constraints. This introduction is not necessary to understand the combinatorial problems and their solutions, but it allows us to place them in their context and explain why they are important, independently of their mathematical value.

1.1 A Minimalist Introduction to Molecular Evolution

The "molecules of heredity," the support of genetic information, are present in every cell of all living organisms (bacteria, plants, animals, etc.). Each molecule is called a *chromosome*, and the set of all chromosomes is what we will call the *genome*. Chromosomes are made of *DNA* (**d**eoxyribo**n**ucleic **a**cid), a double-stranded molecule in which each *strand* is a long succession of *nucleotides* (a *sequence*). Nucleotides can be of four types—*A*, *C*, *G*, and *T*—and the two DNA strands are coupled in such a way that an *A* on one strand is always coupled with a *T* on the other strand, and a *C* on one strand is always coupled with a *G* on the other strand. Those strands are said to be *complementary*: the sequence on one strand determines the sequence on the other one. Figure 1.1 shows a representation of the above concepts.

Because of complementarity, a DNA molecule is usually represented as a single sequence (one arbitrary strand), but the organization in two strands will often be crucial for our purpose. Chromosomes can be either circular (the sequence forms a circle and has no ends), which is often the case in bacteria, or linear (the sequence has two ends, called *telomeres*), which is often the case in animals and plants. A *segment* of DNA is a part of this molecule made of consecutive nucleotides. A *gene* is a segment of DNA that contains the information needed to construct the other molecules in the cell.

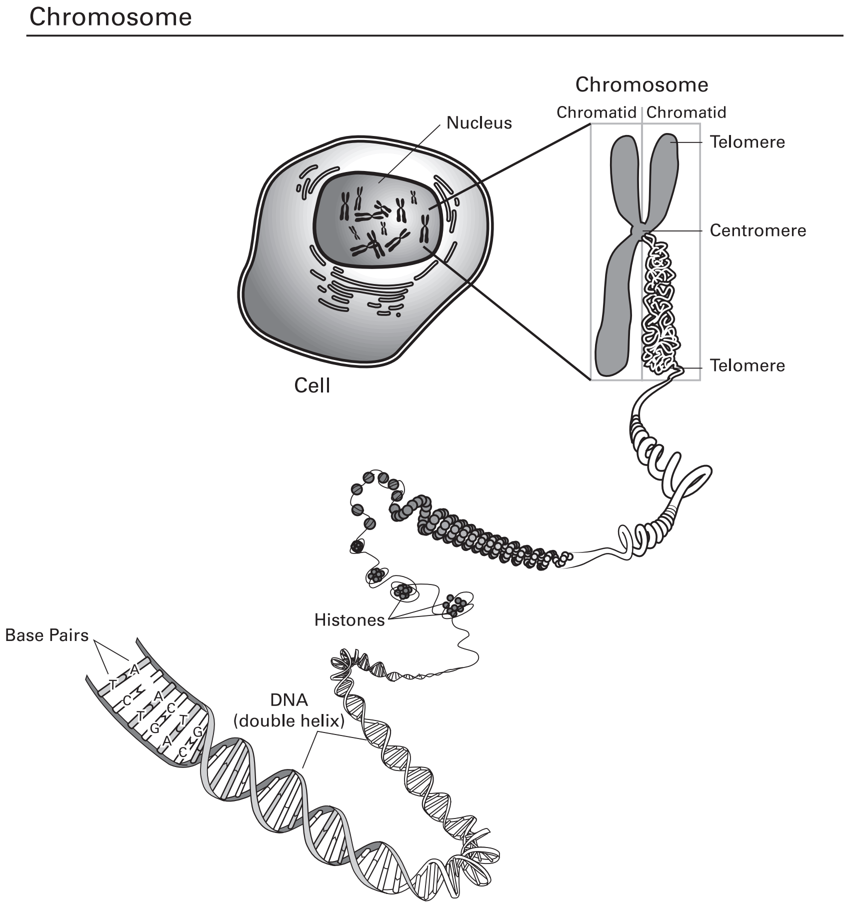

Figure 1.1
A chromosome and a fragment of a DNA molecule
Source: National Institutes of Health, National Human Genome Research Institute, Division of Intramural Research

What accounts for the diversity of living organisms is the possibility for DNA to *replicate* itself with some inaccuracy: one genome is used to produce another, almost identical genome. This inaccuracy is the principle of molecular *evolution*.

A DNA molecule may evolve by *point mutations* (i.e., mutations at the level of nucleotides). There are three different kinds of point mutations: substitutions (one nucleotide is replaced with another), insertions (a nucleotide is added to the sequence), and deletions (a nucleotide is removed from the sequence). Detecting these events is the goal of *sequence alignment* (for a presentation of this topic, see, for example, Setubal and Meidanis [333] or Jones and Pevzner [224]).

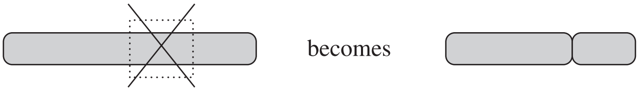

Figure 1.2
Deletion of the dotted region in a chromosome

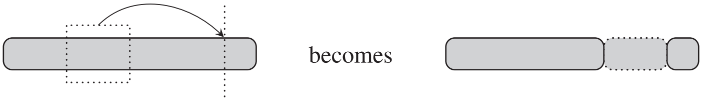

Figure 1.3
Transposition of the dotted region in a chromosome

$\text{CCGT}\underline{\text{GCGTAC}}\text{ACTGC}$ becomes $\text{CCGT}\boxed{\text{GTACGC}}\text{ACTGC}$

Figure 1.4
Reversal of the underlined segment, resulting in the boxed segment

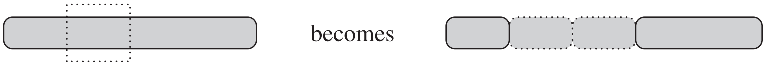

Figure 1.5
Tandem duplication of the dotted region in a chromosome

However, a sequence may also evolve by modifying its organization at a larger scale. These large-scale mutations are called *rearrangements*, or *structural variations*, and detecting them is the goal of *genome rearrangement* problems. The main rearrangements include the following:

- *Deletions*. A segment of the genome is lost (see figure 1.2).
- *Transpositions*. A segment of the genome moves to another location (see figure 1.3). Transpositions are sometimes referred to as translocations or insertions, but transposition is well adopted in the field of combinatorics of genome rearrangements.
- *Inversions* or *reversals*. A segment of the genome is reversed and the strands are exchanged (see figure 1.4).
- *Duplications*. A segment of DNA is copied and inserted in the genome. There are three main standard types of duplications: *tandem duplications*, illustrated by figure 1.5, which insert the copy next to the original; *retrotranspositions*, which insert a copy of a gene at an arbitrary location in the genome; and *whole genome duplications*, which copy either the whole genome or some of its chromosomes.

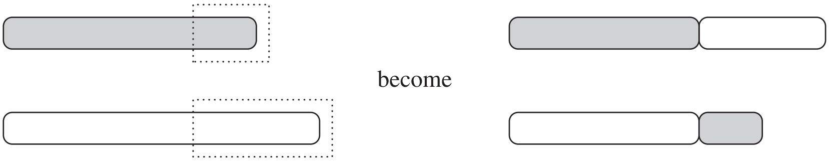

Figure 1.6
Reciprocal translocation of the dotted regions in two chromosomes

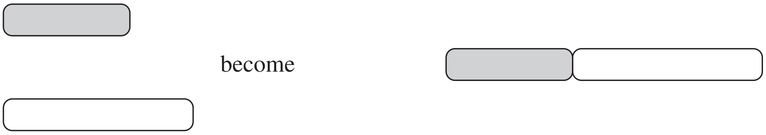

Figure 1.7
Fusion of two chromosomes

- *Reciprocal translocation*. A segment of a chromosome that contains a telomere is exchanged with a segment of another chromosome that also contains a telomere (see figure 1.6).
- *Fusion*. Two chromosomes are joined into one (see figure 1.7).
- *Fission*. One chromosome splits into two (this is the inverse of a fusion).
- *Horizontal*, or *lateral*, *transfer*. A segment of the genome is copied from one genome to another. This is common mainly in unicellular organisms.

All these operations act on a genome at the level of DNA segments rather than on nucleotides. This is why a genome is often represented by a sequence of segments in that setting: they are the segments that are found in an almost identical state in several species, not cut by rearrangements. Two segments are said to be *homologous* if they derive from a common ancestor and are distinguished by a replication event (they end up in two different genomes) or by a duplication event (they both belong to the same genome).

Genes are often taken as those homologous segments because, due to their functional utility, they are less subject to small mutations and are rarely cut by rearrangements, which is not the case for other parts of the genome.

1.2 Birth of the Combinatorics of Genome Rearrangements

In 1936 two renowned biologists, Dobzhansky (the inventor of the synthetic theory of evolution) and Sturtevant (the discoverer of rearrangement processes in genomes at the beginning of the twentieth century) proposed for the first time to use the degree of disorder between the organization of genes in two different genomes as an indicator of an evolutionary distance between organisms (see Dobzhansky and Sturtevant [145, 146]). They proposed a scenario of inversions to explain chromosomal differences between 17 groups of flies, as well as a reconstruction of putative ancestral gene arrangements and species histories from the observation of the gene order along the chromosomes.

Since rearrangements are relatively rare events, scenarios minimizing their number are more likely to be close to reality. In 1941 Sturtevant and Novitski [343] formulated the problem of minimizing the number of inversions that may explain the differences in arrangements between two species: "...for each such sequence there was determined the minimum number of successive inversions required to reduce it to the ordinal sequence chosen as 'standard.' For numbers of loci above nine the determination of this minimum number proved too laborious, and too uncertain, to be carried out...."

The reconstruction of genome rearrangements from the examination of chromosomes, using techniques such as "chromosome banding" or "in-situ hybridisation" [301] were numerous, all focusing on relatively close species, so that the number of rearrangements was small. All these studies were based on the parsimony criterion, which makes molecular biologists often prefer explanations of differences between genomes that involve as few mutations as possible. This principle makes the connection with combinatorial optimization possible, because the optimization principle meets the parsimony criterion.

As we entered the genome sequencing era, the importance of rearrangements in evolution or illnesses was pointed out by several biologists, such as Palmer and Herbon [289], who examined the differences in the gene order of the mitochondrial genomes of cabbage and turnip, which are very similar in sequence but dramatically different in structure. It was not until 1982 that some researchers working in combinatorial optimization started to formalize and become involved in this problem, in order to overcome the limit of nine genes stated by Sturtevant and Novitski [343]. Watterson et al. [369] proposed to represent the relative positions of genes in different genomes as *permutations*. In order to propose an evolutionary scenario between two species, one had to solve the problem of transforming one circular permutation into another with a minimum number of inversions. The problem was far from being solved after this first article, but it was well stated.

Transforming one permutation into another by means of a minimum number of allowed operations is often equivalent to *sorting* a permutation by means of the same operations (see page 17). Though it took a decisive start in a biological context, the problem of sorting permutations with constraints was not new: a few mathematicians and computer scientists had already tackled that kind of problem in the past. Those problems were not, however, motivated by biology: constraints were related to data structures as stacks, or were simply introduced as games that later turned out to be particular cases of genome rearrangement problems, or found uses in other fields.

New models were later proposed to handle more operations, duplicated segments, and several chromosomes. Shortly after Watterson et al. [369], the field started its dramatic expansion.

1.3 Statement of the Problem

The *genome rearrangement problem* is formulated in its most general form as follows: given a set of genomes and a set of possible evolutionary events, find a shortest set of events transforming those genomes into one another.

What "genome" means here, and what events are, makes the diversity of the problem. Miscellaneous models have been proposed, depending on various parameters, and we briefly review them in section 1.5. "Shortest" usually refers to the number of events, but it may also mean "of least weight" if events are weighted (e.g., according to their probability of occurrence).

The length (or weight) of an optimal sequence of events transforming one genome into another is called the *distance* between the two genomes. We will often require that this distance be a metric on the set of genomes, in the mathematical sense, and we recall its definition here.

Definition 1.1 A **metric** d on a set S is an application

$$d : S \times S \to \mathbb{R} : (s, t) \mapsto r$$

satisfying the following three axioms:

1. For all $s, t \in S$, $d(s, t) \geq 0$ and $d(s, t) = 0$ if and only if $s = t$ (positivity).
2. For all $s, t \in S$, $d(s, t) = d(t, s)$ (symmetry).
3. For all $s, t, u \in S$, $d(s, u) \leq d(s, t) + d(t, u)$ (triangular inequality).

A set S equipped with a metric d is called a **metric space** and is denoted (S, d). Finding an optimal sequence of events between two genomes of course yields the distance between the two genomes, but the converse is not always true. Therefore, most of the time we will examine both aspects of the problem. A related problem we will be interested in is that of determining the *diameter* of a metric space.

Definition 1.2 The **diameter** of a metric space (S, d) is the maximal value the distance can reach, that is,

$$\max_{s, t \in S} d(s, t).$$

1.4 Scope of This Survey

This survey is restricted to the algorithmic and combinatorial aspects of genome rearrangements, but it also encompasses a few problems that are similar in spirit, even if they were not motivated by biology in the first place. The motivation for this is twofold: first, those problems deserve at least to be mentioned here, since they are closely related to genome rearrangement problems; and second, the study of related problems or variants of our problems may provide insight on the original problems we are interested in.

There has been a lot of work on probabilistic models and statistical studies of genome rearrangement problems, which we will not consider in this work. We refer the reader to the surveys of Eriksson [165] and Durrett [150]. As the reader may have guessed by reading section 1.1, we also will not delve much into the biological aspects, and will focus on the mathematical aspects of genome rearrangements, though we will mention applications and biological contexts where appropriate.

Some partial surveys have been published in earlier articles or book chapters. In 1995, Hannenhalli and Pevzner [197] wrote the first survey on the combinatorics of genome rearrangements, mainly based on their success at sorting signed permutations by reversals (see sections 3.3 and 4.2). The chapters by Pevzner [296] and Setubal and Meidanis [333] dedicated to genome rearrangements mainly focus on sorting permutations by reversals. The books edited by Gascuel [182], Sankoff and Nadeau [322], Böckenhauer and Bongartz [74], Jiang et al. [222], and Tseng and Zelkowitz [360] contain chapters that survey part of the field or try to give a quick overview of it. A survey article by Li et al. [247] reveals the importance and popularity of the field, and this book intends to be a more developed version of it.

1.5 Overview of the Models

Depending on the assumptions that are made on the data, or the events we want to study, different models can be used. The basic objects will be *homologous markers* (i.e., segments of genomes that can be found in several species, leading to the belief that they belonged to the common ancestor of these species). Genes are a good example of such markers, though they are not the only ones; but since genes were historically first used as markers for genome rearrangement studies, we often say "genes" for "homologous markers," as a simplification.

We will start with the simplest possible model, and progressively extend it by dropping restrictions. In the case where

1. the order of genes in each genome is known,
2. all genomes share the same set of genes,

3. all genomes contain a single copy of each gene, and
4. all genomes consist of a single chromosome,

genomes are modeled by *permutations* (see page 13): each gene can be assigned a unique number and is found exactly once in each genome.

As explained in section 1.1, a DNA molecule has two strands, and some rearrangements may change the strand that a segment belongs to. Therefore, each segment may be assigned a + or a − sign (+ is omitted most of the time) to indicate the strand it resides on, leading to the model of *signed permutations* (see page 15). We have also seen in section 1.1 that chromosomes can be linear or circular, and the latter case can be modeled using circular permutations rather than linear (classical) ones.

In spite of all the technical progress that has been made over the last decades and the large number of genomes that have been completely sequenced, many genomes have been only partially sequenced, which means that we cannot model them using permutations because genes are not totally ordered. In that case, genomes can nevertheless be modeled by *partially ordered sets*, and some studies can still be conducted using that model, as we will see in chapter 5.

In general, however, genes do not appear exactly once in each genome: due to duplications and deletions, there can be several copies of a gene in a given genome, or no copy at all. In that case, genomes are modeled by *strings* (on the alphabet of genes, see page 91) rather than permutations. Of course, it is possible to sign the elements of the string or to deal with circular strings.

A great part of living organisms have a genome that consists of several chromosomes (in a variable number, which can lie between 1 and 100), as is the case for all animals, and permutations as we have presented them are no longer a realistic model in that case. One can use the *disjoint cycle decomposition* of permutations (see page 14) to represent each chromosome using a cycle, in the case where chromosomes are circular, but this concept does not extend to linear chromosomes or strings since it cannot model duplicated genes. We may therefore want to extend our model to disjoint sets of paths and cycles (page 159), where each path or cycle models a chromosome.

Finally, one may not care about or simply not know the order of genes in each chromosome, and care only about whether two genes are in *synteny* (i.e., whether they belong to the same chromosome). In that case, genomes are modeled by *collections* of *sets* of genes (see chapter 12).

1.6 Organization of the Book

The first three parts of this book are organized according to the models presented in the previous section, each part being devoted to a mathematical object that has been used to construct genome rearrangement problems. Part IV is devoted to an important generalization of the basic genome rearrangement problem, known as the *median problem*, which aims at considering more than two genomes at the same time and inferring their common ancestors. It surveys the attempts to reconstruct the kin relationships between genomes by drawing *phylogenetic trees* in which nodes are ancestral configurations and branches (edges) account for evolutionary events.

Part V is a collection of summaries that provide useful additional information on the field, such as a list of available software based on the algorithms that we describe in the book and a list of open problems. This book also includes two appendices: appendix A is devoted to basic concepts of graph theory, and appendix B recalls the basics of the algorithmic theory of complexity, as well as a few **NP**-complete problems.