

Bioinformatics

Genomics and post-genomics

Frédéric Dardel and François Képès

translated by
Noah Hardy

This work has been published with the help of the French Ministère de la
Culture – Centre national du livre



John Wiley & Sons, Ltd

Bioinformatics

Bioinformatics

Genomics and post-genomics

Frédéric Dardel and François Képès

translated by
Noah Hardy

This work has been published with the help of the French Ministère de la
Culture – Centre national du livre



John Wiley & Sons, Ltd

First published in French as *Bioinformatique. Génomique et post-génomique* © 2002 École Polytechnique

Translated into English by Noah Hardy

English language translation copyright © 2006 John Wiley & Sons Ltd
The Atrium, Southern Gate, Chichester,
West Sussex PO19 8SQ, England
Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk

Visit our Home Page on www.wileyeurope.com or www.wiley.com

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The Publisher is not associated with any product or vendor mentioned in this book.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Blvd, Mississauga, Ontario, Canada L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Bioinformatique. English

Bioinformatics : genomics and post-genomics / edited by Frédéric Dardel and François Képès; translated into English by Noah Hardy.

p. cm.

Includes bibliographical references and index.

ISBN-13: 978-0-470-02001-2 (cloth : alk. paper)

ISBN-10: 0-470-02001-6 (cloth : alk. paper)

1. Bioinformatics. 2. Genomics. I. Dardel, Frédéric. II. Képès, François. III. Title.

QH324.2.B558 2006

572.80285 – dc22

2006011225

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-02001-2 ISBN 0-470-02001-6

Typeset in 10½/12½pt Sabon by SNP Best-set Typesetter Ltd., Hong Kong

Printed and bound in Great Britain by Antony Rowe Ltd., Chippenham, Wilts

This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

Preface to the French edition	vii
Preface to the English edition	ix
1 Genome sequencing	1
1.1 Automatic sequencing	1
1.2 Sequencing strategies	4
1.3 Fragmentation strategies	8
1.4 Sequence assembly	12
1.5 Filling gaps	14
1.6 Obstacles to reconstruction	16
1.7 Utilizing a complementary 'large' clone library	18
1.8 The first large-scale sequencing project: The <i>Haemophilus influenzae</i> genome	19
1.9 cDNA and EST	20
2 Sequence comparisons	25
2.1 Introduction: Comparison as a sequence prediction method	25
2.2 A sample molecule: the human androsterone receptor	26
2.3 Sequence homologies – functional homologies	27
2.4 Comparison matrices	28
2.5 The problem of insertions and deletions	33
2.6 Optimal alignment: the dynamic programming method	34
2.7 Fast heuristic methods	38
2.8 Sensitivity, specificity, and confidence level	46
2.9 Multiple alignments	50
3 Comparative genomics	61
3.1 General properties of genomes	61
3.2 Genome comparisons	67
3.3 Gene evolution and phylogeny: applications to annotation	75
4 Genetic information and biological sequences	85
4.1 Introduction: Coding levels	85
4.2 Genes and the genetic code	85
4.3 Expression signals	87

4.4	Specific sites	91
4.5	Sites located on DNA	91
4.6	Sites present on RNA	96
4.7	Pattern detection methods	96
5	Statistics and sequences	107
5.1	Introduction	107
5.2	Nucleotide base and amino acid distribution	107
5.3	The biological basis of codon bias	112
5.4	Using statistical bias for prediction	113
5.5	Modeling DNA sequences	116
5.6	Complex models	120
5.7	Sequencing errors and hidden Markov models	123
5.8	Hidden Markov processes: a general sequence analysis tool	127
5.9	The search for genes – a difficult art	127
6	Structure prediction	131
6.1	The structure of RNA	131
6.2	Properties of the RNA molecule	132
6.3	Secondary RNA structures	134
6.4	Thermodynamic stability of RNA structures	138
6.5	Finding the most stable structure	144
6.6	Validation of predicted secondary structures	149
6.7	Using chemical and enzymatic probing to analyze folding	150
6.8	Long-distance interactions and three-dimensional structure prediction	152
6.9	Protein structure	155
6.10	Secondary structure prediction	158
6.11	Three-dimensional modeling based on homologous protein structure	161
6.12	Predicting folding	166
7	Transcriptome and proteome: macromolecular networks	169
7.1	Introduction	169
7.2	Post-genomic methods	170
7.3	Macromolecular networks	182
7.4	Topology of macromolecular networks	193
7.5	Modularity and dynamics of macromolecular networks	199
7.6	Inference of regulatory networks	206
8	Simulation of biological processes in the genome context	211
8.1	Types of simulations	213
8.2	Prediction and explanation	213
8.3	Simulation of molecular networks	215
8.4	Generic post-genomic simulators	226
	Index	233

Preface to the French edition

This book is directly based on a course that we teach at the *École Polytechnique*. We thank all our colleagues and friends there who have made its existence and publication possible:

Sylvain Blanquet, Chairman of the Biology Department, who first thought of creating this interdisciplinary course some ten years ago, when such a project was highly innovative. He has followed and supported the development of the project in the context of the genomic revolution.

Jean-Marc Steyaert, our colleague at the Computer Science Department, where he initiated the teaching of bioinformatics, and in which he remains active. His critical attention, constant theoretical and methodological contributions, and increasing involvement in biological problematics, have contributed in an essential way to bringing this book about, as well as influencing its contents.

Philippe Dessen, who participated in some of the very early stages in the teaching of bioinformatics at the *École Polytechnique* while he was there; our contacts with him over the years have been invaluable.

Finally, going from the stage of course-notes to a published book would not have been possible without the decisive contributions of Jean-Paul Coard, of *Éditions de l'École Polytechnique*, as well of those of Jean-Claude Mathieu, Véronique Lecointe, Martine Maguer, and Frédéric Zantonio, involved in the technical production.

Frédéric Dardel
François Képès

Preface to the English edition

At the suggestion of Vincent Schachter, to whom we are very grateful, we decided to produce an English version of our book. We have worked closely with Noah Hardy, to ensure the accuracy of the translation, and have updated all chapters with new material where necessary. Chapter 8 was rewritten entirely in English. We hope that this edition will enable many more readers to enjoy our book. We would like to thank Joan Marsh and her colleagues at Wiley for their help in producing this edition.

Frédéric Dardel
François Képès

1

Genome sequencing

1.1 Automatic sequencing

The dideoxyribonucleotide method, developed during the early 1980s in England, at the Cambridge University laboratory of Fred Sanger, is today universally employed to sequence DNA fragments. It is based on the use of DNA polymerase to elongate a single strand of DNA, starting from a primer, utilizing another DNA strand as the template. The DNA polymerase elongates the strand in the presence of four deoxyribonucleotide monomers (dATP, dTTP, dGTP, and dCTP) and a dideoxyribonucleotide analog (ddNTP), which acts as the chain terminator (Figure 1.1). Specific incorporation of the analog by DNA polymerase yields a mixture of fragments that selectively terminate at positions corresponding to each nucleotide (As, in the example below).

The principle of the dideoxyribonucleotide ('dideoxy') method is illustrated in Figure 1.2. Four parallel reactions are carried out, one with each ddNTP, the DNA fragments obtained being separated by electrophoresis. A fluorescent tracer is used to identify fragments synthesized by the polymerase so as to distinguish them from template DNA. The tracer is attached to one extremity of the fragment, either at the 5'-end of the sequencing primer or at the 3'-end of the dideoxynucleotide terminator. Modern automatic sequencers utilize an *in situ* detection system during electrophoresis, in which a laser beam emitting in the fluorophore absorption spectrum is passed through the gel (Figure 1.3). A migrating DNA fragment in the path of the laser beam then emits a fluorescent signal detected by a photodiode on the other side of the gel. The signal is amplified and transmitted to a computer programmed with special software for analyzing it.

Under favorable conditions, this technique permits reading up to 1,000 nucleotides per sequenced fragment, and an average of **500 to 800 nucleotides during routine experiments**. Two dideoxy methodologies coexist at present: one employs a single photophore, and the other uses four, each with a distinct emission spectrum. In the first system, the four mixtures, corresponding to the four ddNTPs, are introduced into different electrophoresis gel wells. Analysis is

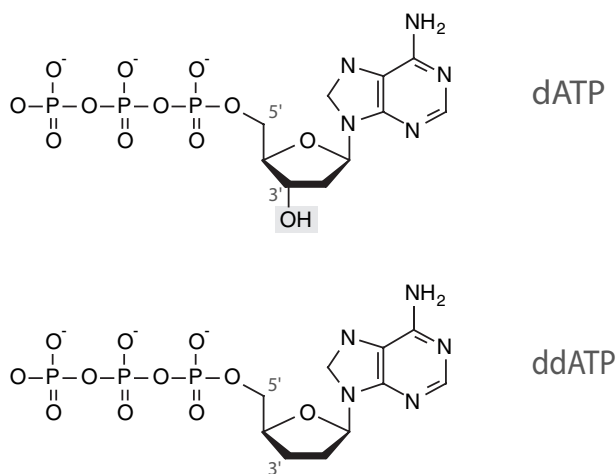


Figure 1.1 Dideoxynucleotide structure

Replacement of the 3'-OH group in the dideoxynucleotide (ddNTP) by a 3'-H group prevents formation of a phosphodiester link at its 3'-end. The modified nucleotides have a normal 5'-triphosphate side, thus may be incorporated into the chain by DNA polymerase. Since A-T and G-C pairing rules are followed during ddNTP incorporation, the ddATP will be incorporated wherever there is a T facing it on the template strand.

accomplished by comparing the migration rates of the fragments in the four resulting lanes.

In the second system, each of the four sequencing reactions uses a different fluorophore that modifies the corresponding ddNTP. After the four polymerization reactions have taken place, the resulting DNA fragments are mixed and introduced into the same gel well. Constituent nucleotides are identified according to the emission properties of the fluorescent tracers exposed to the laser beam using selective color filters, after which a single gel lane is analyzed.

The four-fluorophore technique is a bit more expensive, since it requires somewhat more varied chemistry. However, it has the advantage of being better adapted to high-throughput systems, since more samples are analyzed on the same gel. In the latest generation of automatic sequencers, the classical rectangular polyacrylamide gel is replaced by a reusable capillary tube, whereas the separation and detection principles remain unchanged. This technique reduces the time required for an experiment from several hours to a few minutes, also minimizing preparation time. In principle, the highest-performance multi-capillary machines can process up to 1,000 samples per day, equivalent to 0.5 Mbases of raw sequence per day per machine.

Massive high-throughput sequencing centers today often use several dozen such machines with robots that control the sequencing reactions automatically executing pipetting, mixing, and incubation steps, thereby minimizing the risk

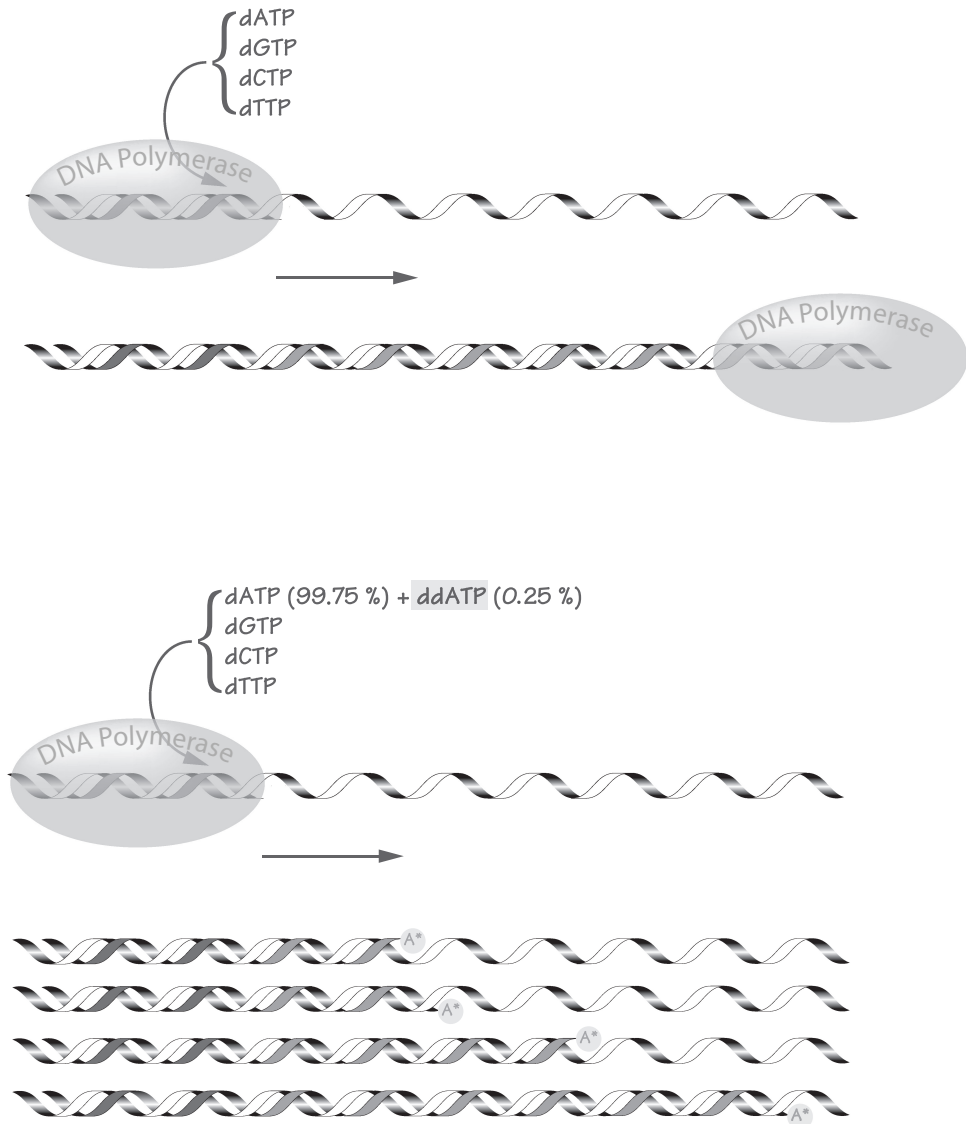


Figure 1.2 Principle of the Sanger sequencing method

In the presence of a template DNA strand and the four dNTPs, DNA polymerase can elongate a complementary DNA strand starting from an **oligonucleotide primer**, which hybridizes to the template strand. When a dideoxynucleotide is incorporated by the polymerase, it acts as a chain terminator, blocking further elongation. This incorporation is entirely random, proceeding at a rate that is a function of the ratio of the dideoxynucleotide concentration to that of the corresponding deoxynucleotide (here it is $[ddATP] / [dATP] = 1 / 400$).

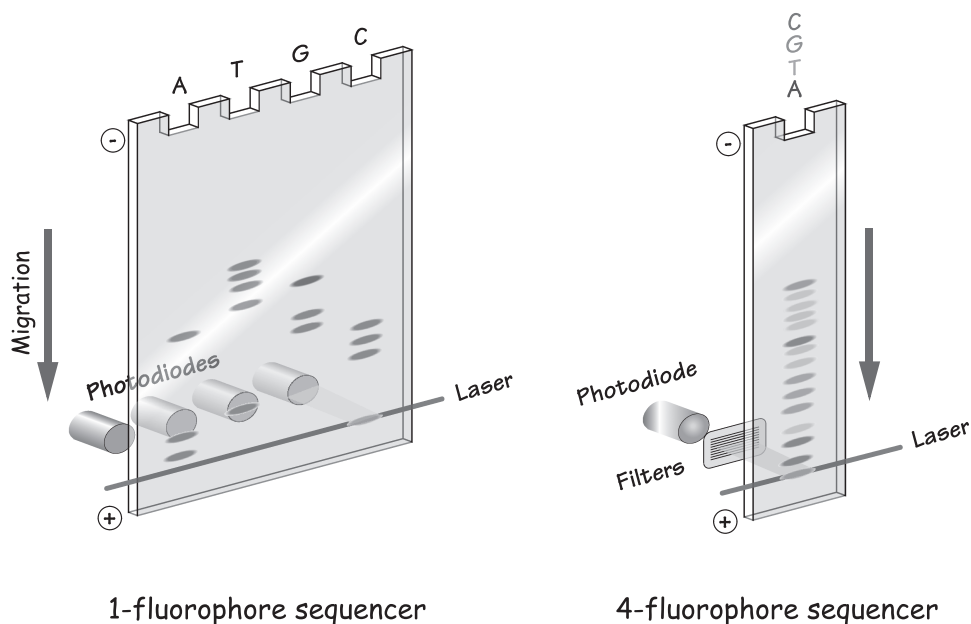


Figure 1.3 Automatic sequencing using 1- and 4-fluorophore sequencers. Samples introduced into the wells (top) are separated by electrophoresis on a polyacrylamide-urea gel. The 5'-CAATCCGGATGTTT sequence is read from bottom to top.

of human error (see Figure 1.4). The preparation of DNA templates remains the most difficult step to automate, although significant progress has been made in this respect.

1.2 Sequencing strategies

The sequencing methodologies described above fail to address major difficulties that need to be considered when operating a large-scale sequencing program:

- Only DNA fragments of between 500 and 1,000 nucleotides may be sequenced;
- A sequencing primer that is complementary to the template is required for the DNA polymerase to begin synthesizing.

Fortunately, these two obstacles may be simultaneously overcome by fragmenting the DNA that is to be sequenced into segments of size compatible to

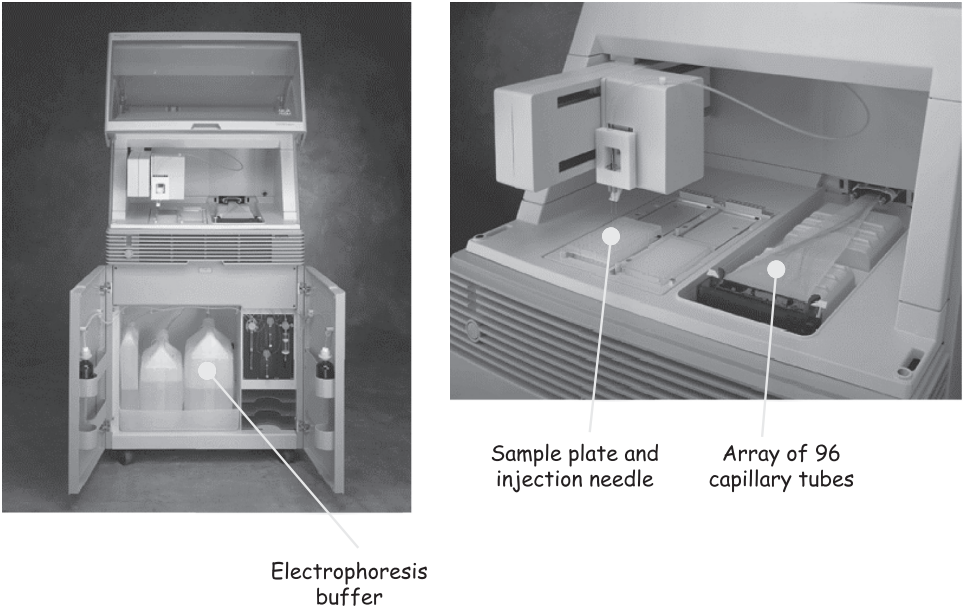


Figure 1.4 Advanced automatic multicapillary DNA sequencer for simultaneous sequencing of 96 samples. An automatic injection system executes several consecutive separations without manual intervention (®Applied Biosystems).

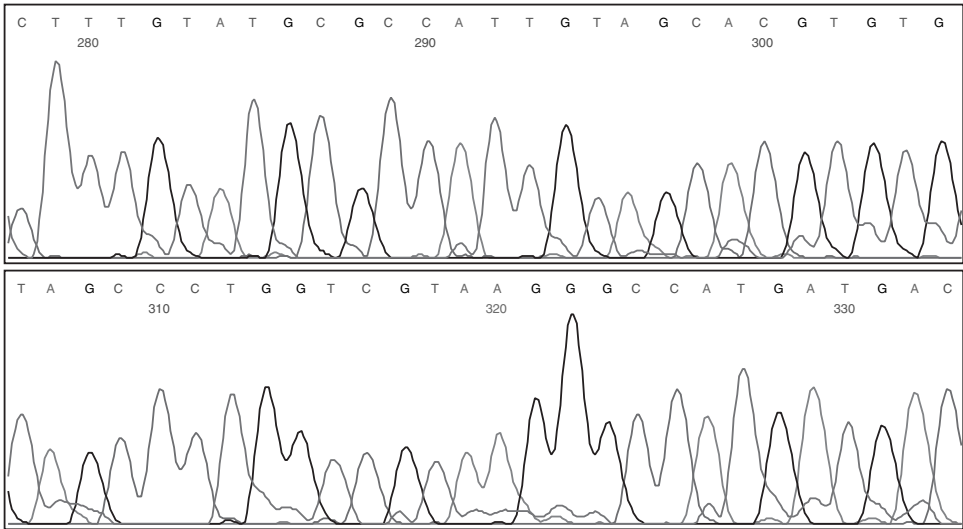


Figure 1.5 Example of a sequencing profile
Intensity of the signal detected by the photodiodes as a function of separation time. Each color is associated with one of the four separation reactions (A, G, C, and T).

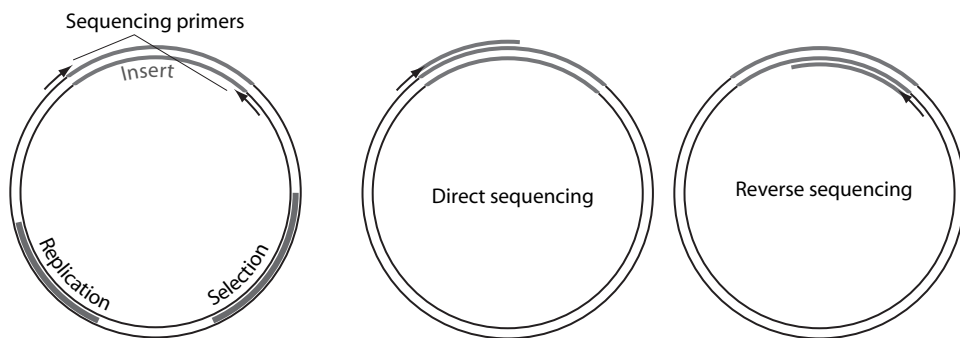


Figure 1.6 Sequencing in a vector starting with universal primers.

that of the sequencing system yield ($\sim 10^3$ base pairs) and by inserting them into an appropriate vector (plasmid or virus). The vector is selected according to several criteria:

- It must be able to replicate autonomously in a convenient host cell (usually *Escherichia coli*);
- It must bear one or several gene markers that permit selection of cells that contain it (antibiotic resistance, for example);
- Its nucleotide sequence must be known;
- It must contain restriction endonuclease sites that permit cloning by insertion of foreign DNA fragments.

In practice, small bacterial plasmids are generally used. The DNA to be sequenced is fragmented and ligated into the vector, which is then propagated in host cells. The clone cell lines (derived from a single initial cell by successive division), each containing a different recombinant vector with the same inserted DNA fragment are then isolated. A library of DNA fragments may thus be constituted by collecting a large number of these clone cell lines, and used for further study (see Figure 1.7).

In order to determine the DNA sequence of such a fragment, the corresponding cell line is cultured and its DNA extracted for sequencing by the dideoxynucleotide method. Since the nucleotide sequences located on each side of the vector clone site are known, they are used as the primer (see Figure 1.6). These primers are independent of the DNA inserted into the vector and may be used to sequence any fragment; they are therefore called *universal primers*. Because

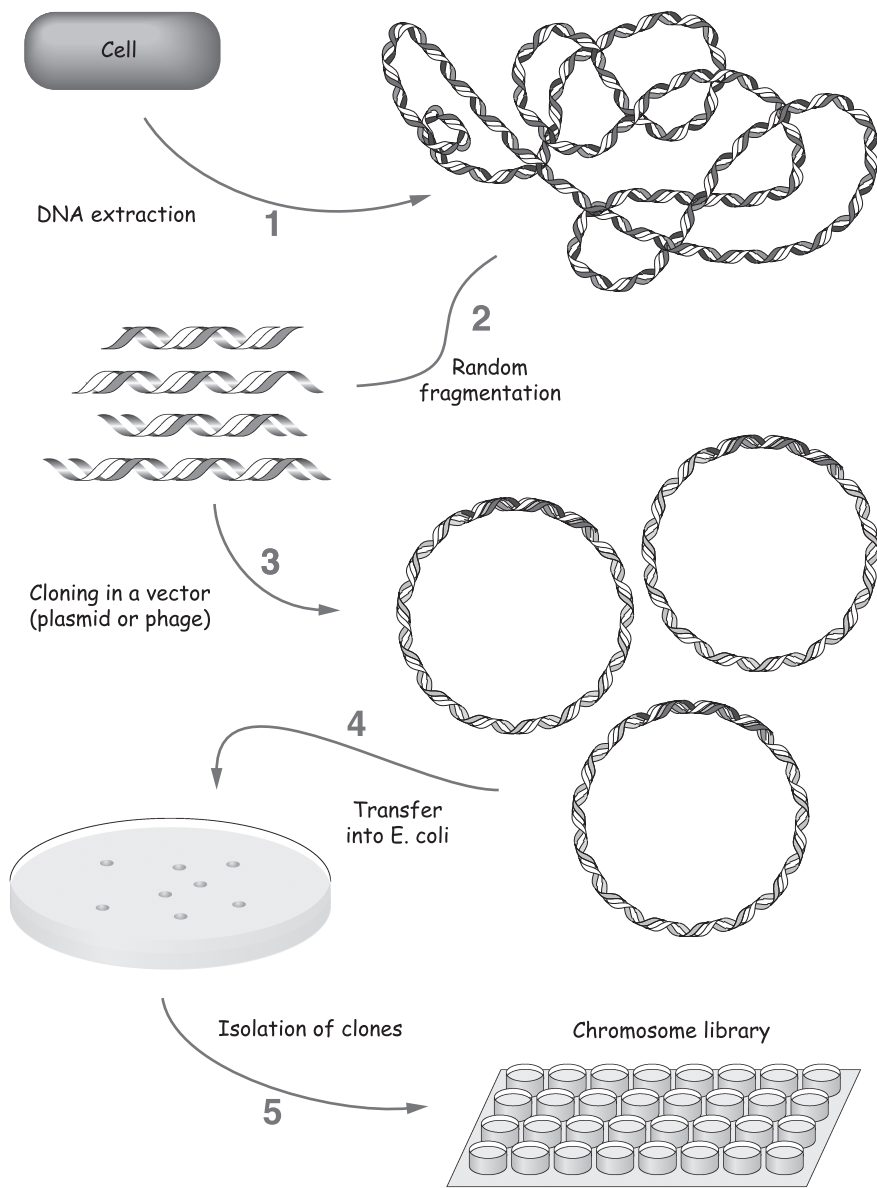


Figure 1.7 Strategy for constructing a DNA library.

such primers are constant, it is very easy to incorporate fluorescent tracers needed during oligonucleotide synthesis into them. The fluorescent primers thus produced may be used in most sequencing procedures.

1.3 Fragmentation strategies

In sequencing a long stretch of DNA – especially a complete genome – it is essential that it be cut into fragments of a size compatible with the sequencing technology. This poses two additional questions:

- Which cutting strategy should be employed?
- How can the complete sequence be reconstituted from the pieces?

These two questions are intricately related, since the reassembly method is sensitively dependent on how the fragmentation is accomplished. Two approaches have mainly been used: **random fragmentation** and **segmentation after mapping**.

Random fragmentation

In random fragmentation, the full length of the DNA to be sequenced is cut into small pieces of optimal sequencing size (~1,000 base pairs). A high cutting frequency (one site per 200–250 bp) restriction enzyme may be used for this purpose, under conditions of partial digestion (10–20 percent) in order to generate 1,000- to 2,000-bp fragments. Alternatively, ultrasound may be used to break the DNA into small pieces, since the mechanical constraints induced by ultrasonic vibrations in DNA in solution are sufficient to rupture the long phosphodiester chain.

The mechanical (ultrasound) method results in more random breaks than the enzymatic method but necessitates an additional step to repair the extremities of the DNA fragments produced, since breaks produced by ultrasound treatment do not occur at the same level in the two DNA strands. This may require paring the extended extremities of single strands, so that the resulting fragments may be inserted into blunt cloning sites in the sequencing vector.

The basic postulate of the random or *shotgun* method is that if enough clones are analyzed, the entire original DNA sequence will be covered. Assuming that fragmentation and cloning are really random processes and that the DNA sequence is sufficiently large compared with that of individual clones (which is generally the case for a full genome), the probability that a given DNA nucleotide studied **not be covered** by random sequencing is a Poisson distribution:

$$p_0 = e^{-N/L},$$

where N is the total number of nucleotides sequenced in the set of clones and L the total length of the DNA studied. N/L is the coverage rate, which is the rate of data redundancy. In order to obtain a 99 percent sequencing rate, that is, $p_0 = 0.01$, it is necessary to sequence a number of clones equal to 4.6 times ($\log 0.01 \approx -4.6$) the length of the DNA studied.

In the case of a genome or a very long DNA fragment, it is thus practically inevitable that gaps remain in the sequence, which must be filled using some approach other than the random *shotgun* method. It is also possible to statistically evaluate the length and average number of such gaps:

$$\text{Total length of gaps} = Le^{-N/L}$$

$$\text{Average length of each gap} = Ln/N$$

$$\text{Number of gaps} = N/ne^{-N/L},$$

where n is the average length of each fragment sequenced (~ 500 nucleotides). The following is an example of the results for a bacterial genome ($L \approx 10^6$ bp) and for the genome of a higher organism, such as a mammal or a plant ($L \approx 10^9$ bp) with a coverage rate of factor 6 (an average value for this type of project), which yields 99.75 percent of the sequenced nucleotides:

The random approach raises two important points:

- It is impossible to cover the entire genome without greatly increasing the number of clones to be sequenced; to be nearly certain of covering the entire bacterial genome in the above example would require that $p_0 \ll 10^{-6}$, at least 14 times the coverage rate. From the practical point of view, it is more economical to accept a coverage rate of between 4 and 6 and then fill the few dozen remaining gaps *ad hoc* (see Table 1.1 below).
- Assembly of the puzzle of the set of fragments may require systematic side-by-side comparison of all the sequences obtained. For k sequences, this

Table 1.1

	Bacteria (1 Mbp)	Mammals (1 Gbp)
Number of sequences	12,000	12,000,000
Number of remaining gaps	30	29,750
Average gap size	200	200

amounts to $k(k - 1)/2$, which is around 10^8 comparisons for a bacterial genome and 10^{14} for a mammalian genome. While today's computers can handle the first number of comparisons, the second number remains a formidable challenge.

Segmentation after mapping

The complexity of reconstruction becomes seriously difficult with very large genomes, which is why some groups have resorted to a two-level approach in such cases. The *shotgun* method is used only to reduce the number of clones required to cover the genome. A DNA library is then established in the following three types of vector, which can accept much larger fragments:

- **Cosmids** are hybrids of bacterial and bacteriophage plasmids grown in *E. coli*; they accept up to 30–40 kbp of inserted DNA.
- **BACs** (bacterial artificial chromosomes) are very large plasmids constructed starting from the replication origin of the bacterial chromosome. They can accept inserted DNA fragments of around 100–130 kbp.
- **YACs** (yeast artificial chromosomes) are analogs of the preceding two types of vectors, but derive from chromosomes of a lower eukaryote cell, such as brewer's yeast, *Saccharomyces cerevisiae*. YACs accept DNA fragments of the order of 1,000 kbp. Unlike the two other kinds of vector, which use *E. coli* as the host, YACs must be maintained in yeast cells.

These three types of vector can be employed to construct practically complete DNA libraries from given genomes, using a much smaller number of clones than plasmids intended for sequencing use. A human genome library consisting of 33,000 YACs that can accept inserted DNA fragments of average length around 1 Mbp has been constructed in a joint project between the Généthon (located in the Paris suburb of Evry) and the CEPH (*Centre d'étude du polymorphisme humain*, in Paris). The coverage rate of this clone library is thus practically 10 times the human genome, which consists of around 3.5×10^9 bp.

However, these large clone libraries cannot be directly exploited for sequencing purposes. A map must be constructed that positions the YAC, BAC, and cosmid vectors used in appropriate order on the chromosomes of the genome being studied. YAC positioning was accomplished in 1995 at Généthon/CEPH. The International Human Genome sequencing consortium later produced and mapped a library of 350,000 BACs, which provided similar tenfold coverage of the human genome. Since they bear smaller inserts, more BACs were required.

This map presents two advantages:

1. Used in conjunction with a genetic map, it allows isolation of genes associated with genetic diseases (Généthon's main goal) by localizing them on a given YAC.
2. It can be used as a framework in a sequencing program. Once YACs or BACs are positioned, they just have to be sequenced in an orderly manner, using the random strategy for each. The complexity of reconstruction is then analogous to that of a bacterial genome, since the construct size is on the order of one mega base pair. Considering the coverage rate (10 $\times$) of the YAC or BAC libraries, it is 'enough' to sequence a fraction of them to cover the entire human genome. This approach is a direct transposition of the classical 'divide and conquer' strategy used in computer science. For the human genome, the International Consortium sequenced 30,000 of the 350,000 BACs that made up its original library, enough to provide a 'tiling pathway' (see Figure 1.8) that completely covered the chromosomes. In the end, BACs were preferred to YACs, since they are smaller and easier to propagate in bacterial cells.

Mapping or positioning the primary library (cosmids, BACs, and YACs) is rather time-consuming. It is carried out using a combination of techniques, such

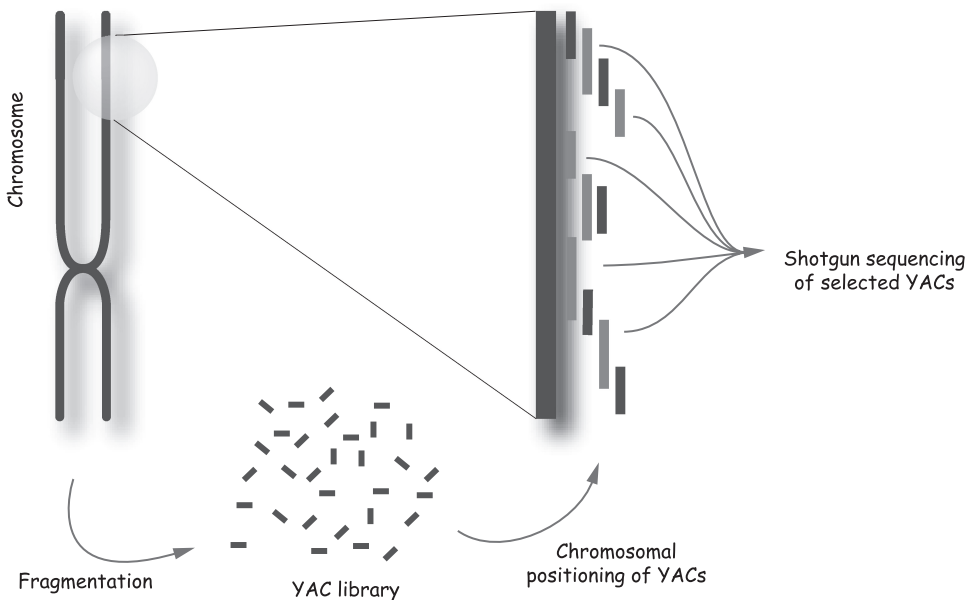


Figure 1.8 Two-step sequencing strategy with intermediate mapping.

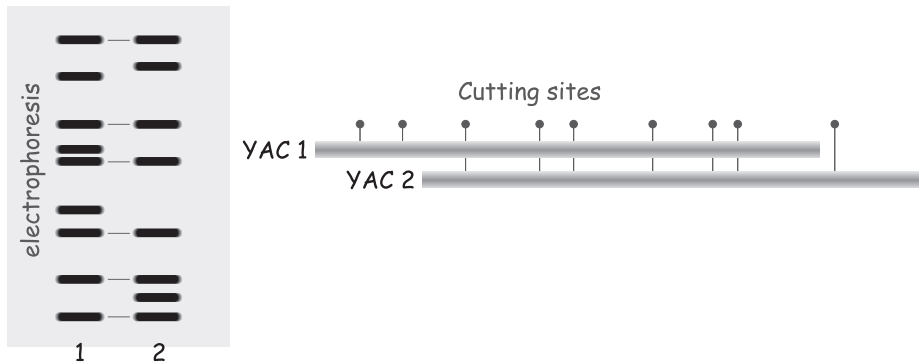


Figure 1.9 Example of the comparison of restriction enzyme digestion profiles of two overlapping YAC clones.

as comparing restriction enzyme digestion profiles, DNA-DNA hybridization, and genetic marker identification. For example, the restriction profiles of overlapping YACs contain common fragments. Systematic search for similarities among profiles throughout the entire YAC library identifies candidates.

It is then possible to verify the overlapping of two YACs by DNA-DNA hybridization. After the YAC① DNA restriction fragments have been separated by electrophoresis, they are transferred from the gel onto a membrane, where they are immobilized by cross-linking. The two strands are then separated by denaturing in an alkaline pH. The membrane is then incubated in the presence of the denatured YAC② DNA marked with a radioactive or fluorescent tracer. If the two YACs have common sequences, the marked YAC② DNA locally pairs with that of the YAC① DNA, in which case *hybridization* is said to have taken place between the two YACs. The tracer is seen to be fixed on the membrane, thereby revealing DNA fragments common to both YACs. If the restriction enzyme cutting profiles coincide accidentally and do not correspond to an overlap, there is no hybridization, thus no tracer fixation. Systematically conducted, this analysis results in a gene map.

1.4 Sequence assembly

The **sequenced fragment assembly method** requires that possible overlaps be identified first, so as to detect clones with common DNA sequences. If two clones are found to overlap, they are merged to form what is called a *contig*, a term designating a set of fragments connected to each other by overlapping sequences that are either identical or very similar (within the limits of sequenc-

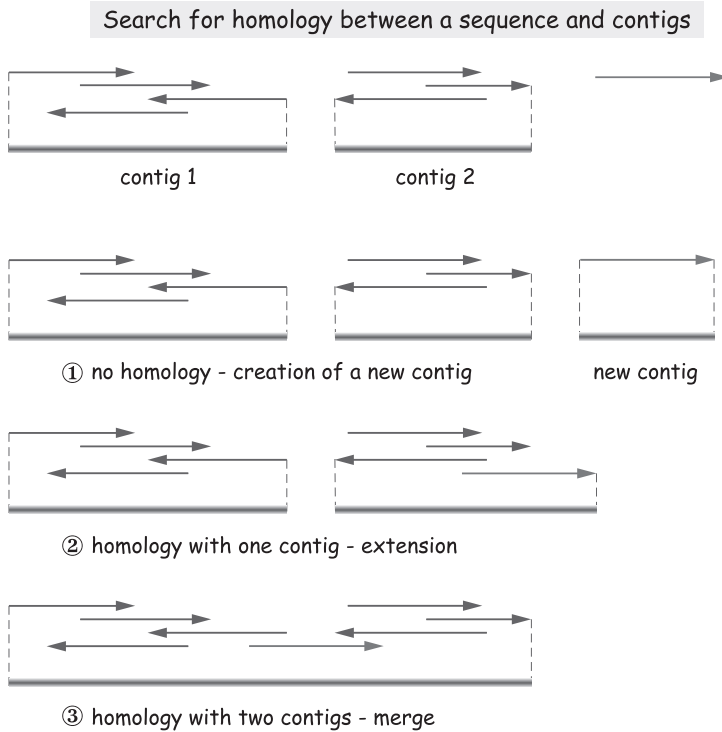


Figure 1.10 Iterative contig assembly

The arrows indicate the orientation of sequenced DNA strands.

ing error). The next step consists in step-by-step comparison of each new fragment with contigs that have already been identified (see Figure 1.10). This comparison must take into account the two possible relative orientations of the two sequences: If the fragment overlaps one contig, that contig is extended; otherwise a new contig consisting of this fragment alone is created. When a fragment simultaneously overlaps two contigs, both are fused with the fragment. At any time during a major sequencing project, the data correspond to a set of several contigs, whereas ideally, only one contig covering the whole sequence, remains at the end of the project.

While a contig is being assembled, a **consensus sequence** associated with it is defined according to the alignment of its constituent fragments. The consensus sequence compares the positions being read and checks their agreement, revealing any differences or ambiguity due to data errors (unread or incorrect nucleotide interpretation). Differences and interpretation ambiguities may be resolved by further data analysis and if necessary, by additional sequencing. This verification step is indispensable but time-consuming, since it is at least partly manual.