

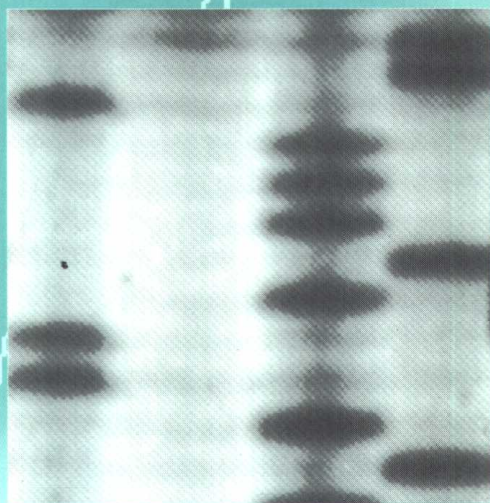
INTRODUCTION TO BIOTECHNIQUES

LUKE ALPHEY

DNA Sequencing

FROM EXPERIMENTAL METHODS TO BIOINFORMATICS

ACCGGCATGCCGAGCAARTG
470 480



Springer



世界图书出版公司

Bios
SCIENTIFIC
PUBLISHERS

Abbreviations

BLAST	Basic Local Alignment Search Tool
BSA	bovine serum albumin
CASP	critical assessment of structure prediction
CCD	charge-coupled device
cDNA	complementary DNA
DDBJ	DNA Databank of Japan
DDGE	double-strand denaturing gel electrophoresis
ddNTP	2', 3'-dideoxynucleotide
DMSO	dimethylsulfoxide
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EBI	European Bioinformatics Institute
EDTA	ethylenediamine tetraacetic acid
EMBL	European Molecular Biology Laboratory
EPD	Eukaryotic Promoter Database
EST	expressed sequence tag
ExoIII	exonuclease III
ftp	file transfer protocol
GCG	Genetics Computing Group
HPLC	high-performance liquid chromatography
HSSP	homology-derived structures of proteins
IUB	International Union of Biochemistry
IUPAC	International Union of Pure and Applied Chemistry
MCS	multiple cloning site
5-MeC	5-methylcytosine
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NMR	nuclear magnetic resonance
NP-40	Nonidet P-40
OMIM	Online Mendelian Inheritance in Man
ORF	open reading frame
PC	personal computer
PCR	polymerase chain reaction
PDB	protein databank
PEG	polyethyleneglycol
PNK	poynucleotide kinase
REBASE	Restriction Enzyme Database
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RT-PCR	reverse transcriptase-polymerase chain reaction
SCOP	structural classification of proteins
SRS	Sequence Retrieval System

SSCP	single-strand conformation polymorphism
STS	sequence tagged site
TBE	Tris-borate-EDTA
TE	Tris-EDTA
TEMED	N,N,N',N'-tetramethylethylenediamine
TES	2-[Tris(hydroxymethyl)methylamino]-1-ethanesulfonic acid
TREMBL	Translated EMBL
TTE	Tris-aurine-EDTA

Preface

In the 20 years since the current methods were first introduced, DNA sequencing has been at the heart of modern molecular biology. The sequence databases have been growing at an exponential rate, and even that rate of increase is improving, with doubling time down from about 22 months to 9 months. Whole new areas of research have been opened up by this technology, from molecular genetics to molecular taxonomy. With the advent of whole genome sequencing, exciting new vistas are emerging.

This book is intended as a practical guide, particularly at the strategic level. It aims to explain the options available and their relative merits, to allow the reader to decide which is most suitable for their application. The book covers the whole process of DNA sequencing, from planning the approach, through data acquisition, to extracting useful biological information from the data.

The book is aimed primarily at those new to DNA sequencing, but I hope that it will also prove a useful text for more experienced sequencers and that the information provided will be useful as a source of further information on familiar techniques and as a reference for less common ones. Part 1 describes the basic methods in detail, including manual and automated sequencing and the various pitfalls that may be encountered on the way. The equipment required is discussed, together with the advantages and disadvantages of each option.

Part 2 details the major applications of DNA sequencing: confirmatory sequencing to check a particular construct or mutant; sequencing PCR products; and strategies for sequencing large fragments of uncharacterized DNA. Part 3 covers Bioinformatics – the analysis of the sequence data to extract useful information. This section was contributed by Dr Andy Brass, Senior Lecturer in Bioinformatics at the University of Manchester, UK. It covers sequence analysis from checking and compiling the raw data through to homology searches and structural predictions.

Luke Alphey

Acknowledgements

First of all, I would like to thank Andy Brass for contributing the Bioinformatics section of the book. Jane Hewitt provided most of the gel examples for *Table 8.1* and Lawrence Hall provided the data for *Figure 7.2*. Eaton Publishing (*Figure 7.4*), VCH Verlagsgesellschaft mbH (*Figure 7.6*) and PE-Applied Biosystems (*Figures 10.3–10.6* and *10.8*) all generously permitted the reproduction of their copyright material. I am also grateful to Jane Hewitt, Lawrence Hall and Nina Nicholls for their critical reading of the manuscript. Finally, I would like to thank N.N. and B.B. for their constant encouragement and support.

Safety

Certain reagents indicated for use in this book are chemically hazardous or radioactive. The researcher is cautioned to exercise care with these reagents and with the equipment (e.g. electrophoresis equipment) used in these procedures, strictly following the manufacturer's safety recommendations. Disposal of waste (including waste chemicals and radioactive materials), must comply with all local, national and other applicable regulations. These procedures may also be governed by other relevant regulations, for example those covering the containment and use of genetically modified micro-organisms. While every care has been taken to ensure that the experimental details discussed in this book are accurate and safe, the author accepts no liability for any loss or injury howsoever caused.

Many of the procedures discussed in this book are protected by patents or other legal protection. The reader is hereby notified that the purchase of this book does not convey any license or authorization to practise any of these procedures.

Contents

Abbreviations	xi
Preface	xiii

PART 1: BASIC PRINCIPLES AND METHODS

1. What is DNA Sequencing?	1
An introduction	1
Nucleic acid structure	2
DNA sequencing	5
References	9
2. Chemical Degradation (Maxam and Gilbert) Method	11
A description of the method	11
References	13
3. Chain Termination (Sanger Dideoxy) Method	15
Introduction	15
Cycle sequencing	19
References	25
4. Instrumentation and Reagents	27
Getting started – sequencing kits	27
Oligonucleotide primers	28
Primer design	29
Primer design for cycle sequencing	31
DNA polymerase	31
Label	32
dNTPs and ddNTPs	35
dITP and 7-deaza-dGTP	37
Pyrophosphatase	38
References	39
5. Template Preparation	41
Introduction	41
Preparing single-stranded DNA templates	41

Preparing double-stranded DNA templates (plasmids)	46
PCR products	46
Single-stranded DNA templates from PCR products	47
Large templates (lambda, cosmids, P1)	49
Templates for semi-automated sequencing	50
References	50
6. Gel Electrophoresis	53
Introduction	53
Overview	53
Reading a sequence autoradiogram	54
Gel systems	55
Safety	55
Gel plates	55
Combs	55
Width	56
Thickness	56
Length	56
Temperature control	56
Reagents	57
Long Ranger™	57
Glycerol-tolerant gels	60
Formamide gels	60
Capillary electrophoresis	60
References	61
7. Nonradioactive Methods	63
Introduction	63
Semi-automated sequencers	64
ABI 377	65
Dye terminator chemistry	68
Dye primer chemistry	70
Optimizing sequencing on the ABI 377	71
Template quality	72
Primer quality	73
Template and primer concentrations	74
Removing unincorporated label	74
Future developments	75
Brighter dyes	75
Better electrophoretic resolution	76
Better software	76
Uniform peak heights	76
Increased throughput	76

LI-COR	77
References	78
8. Troubleshooting	81
Introduction	81
Co-termination	84
Secondary structure	84
Dirty template	85
Sequencing near to the primer	85
Incorrect dNTP incorporation	85
Reaction conditions	86
dITP	86
Compressions	86
Base analogs	87
Formamide gels	88
Reference	88
PART 2: APPLICATIONS	
9. Confirmatory Sequencing	89
Introduction	89
Checking constructs	89
Sequencing allelic variants	90
Alternatives to DNA sequencing	91
Using restriction endonucleases	91
Using oligonucleotide hybridization	92
Using PCR	93
References	96
10. Sequencing PCR Products	97
Introduction	97
Sequence information from PCR products	99
Sequence analysis of PCR products	99
Fidelity of other polymerases	100
Mutant detection by sequencing PCR products	101
Tailed primers	103
Custom dye primers	103
Dye terminators	105
Confirming the presence of heterozygotes	107
Sequencing methylated DNA	108
References	110

11. Strategies for New Sequence Determination	111
Introduction	111
Directed versus nondirected strategies	112
Primer walking	113
Restriction endonuclease digestion and subcloning	114
'Shotgun' methods	116
Frequently cutting restriction endonucleases	117
Sonication	117
DNase I digestion	117
Transposon-facilitated sequencing	117
Deletion series	118
Exonuclease digests too fast or too slow	120
DNA is completely degraded by exonuclease	121
Difficulty in cloning deletion products	121
Deletions using $\gamma\delta$ transposon	122
References	123

PART 3: SEQUENCE ANALYSIS

12. Introduction to Bioinformatics and the Internet – A. Brass	125
Introduction	125
Bioinformatics is a knowledge-based theoretical discipline	125
Access to bioinformatics tools	126
Getting access to tools on the Web	126
Navigating the Web – or how do I find what I want?	127
Using Web-based tools	129
E-mail servers	130
Accessing remote computers to get useful software – anonymous ftp	130
Good and bad practice	132
13. Sequence Databases – A. Brass	133
Background	133
Primary databases	134
DNA databases	134
Genome databases	135
Protein sequence databases	135
Protein structure databases	137
Primary sequence database annotation	138
Information retrieval systems	141
Submitting a sequence to a database	142

14. Sequence Alignment and Database Searches – A. Brass	145
Introduction	145
Scoring matrices	145
Gap penalties	147
Pairwise sequence alignments	148
Multiple sequence alignments	149
Comparing sequences against a database	150
When is a hit significant?	156
References	156
15. Sequencing Projects and Contig Analysis – A. Brass	159
Introduction	159
Analyzing clones	159
Removing the sequence vector	160
Removing other cloning sequence artifacts	160
Contig assembly	161
Predicting protein-coding regions	162
Coding regions in cDNA	162
Coding regions in genomic DNA	163
DNA analysis	164
Restriction enzyme maps	164
Promoters and other DNA control sites	165
RNA secondary structure prediction	166
References	166
16. Protein Function Prediction – A. Brass	167
Introduction	167
Comparing a protein sequence against a sequence database to determine function	167
Hydrophobicity, transmembrane helices, leader sequences and sorting	170
Calculating hydrophobicity profiles	170
Predicting transmembrane helices	170
Leader sequences and protein localization	172
Coiled-coils	172
Comparing a protein sequence against motif and profile databases to determine function	173
Motif databases – PROSITE	174
Profile databases	175
References	176

17. Protein Structure Prediction – A. Brass	179
Introduction	179
Protein structure resources	181
Secondary structure prediction	182
Tertiary structure prediction	183
Comparison against sequences of known structure	183
Homology modeling	184
Threading algorithms and fold recognition	184
Critical assessment of structure prediction (CASP)	186
References	187
Appendices	189
Appendix A: Glossary	189
Appendix B: Amino acid and nucleotide codes	195
Appendix C: Suppliers	197
Index	203

1 What is DNA Sequencing?

1.1 An introduction

DNA sequencing is the determination of all or part of the nucleotide sequence of a specific deoxyribonucleic acid (DNA) molecule. The ability to sequence DNA lies at the heart of the molecular biology revolution. Techniques to sequence DNA were developed only quite recently; the original papers describing the modern methods were published in 1977 [1, 2]. The rate at which new sequence information is determined has increased rapidly over the last 20 years. It is still accelerating, to the extent that the entire human genome sequence of approximately 3×10^9 base pairs will be determined within the next few years, as will the genome sequences of a considerable number of other organisms of medical, agricultural or scientific importance.

The fundamental reasons for wishing to know the sequence of a DNA molecule are:

- to make predictions about its function;
- to facilitate manipulation of the molecule.

The aim of this book is to show how DNA sequence information is obtained and analyzed, and some of the major reasons for doing so. Chapters 2–8 describe the sequencing methods in common usage, with particular emphasis on the relative merits and pitfalls of each approach. Chapters 9–11 describe the major applications of these methods. Subsequent chapters cover the computer-based analysis of sequence data.

Before discussing the principles behind DNA sequencing, we must first consider the structure of a DNA molecule.

1.2 Nucleic acid structure

The normal conformation of DNA is as a double helix (see *Figure 1.1*). This helix comprises two DNA strands running antiparallel to each other, each strand being a chain of bases, each base covalently linked to the next. The bases are each attached to deoxyribose, a sugar molecule, and each sugar molecule is linked to the adjacent sugar molecule via a phosphate group. The basic repeat unit of DNA therefore comprises a base, a sugar and a phosphate group, and is known as a nucleotide (see *Figures 1.2–1.5*).

The structure of a four-nucleotide segment of DNA is shown in *Figure 1.6*. Note that only one strand is shown. Note also the numbering of the carbon atoms in the deoxyribose (sugar) part of the molecule. These each have a 'prime', for example 5' and 3', to distinguish them from the atoms of the bases. It is the 5' and 3' carbons of adjacent sugars that are linked via the phosphate groups, so each covalently linked DNA strand will have a 5' end and a 3' end, as shown in

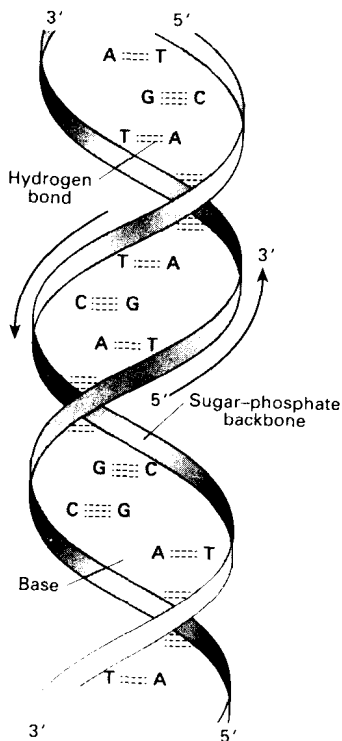


FIGURE 1.1: The DNA double helix. Reproduced from Williams et al. (1993) Genetic Engineering, BIOS Scientific Publishers Ltd.

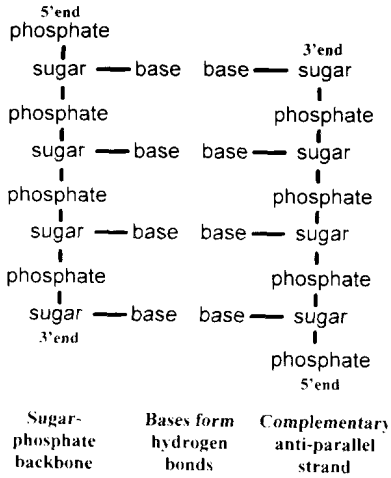


FIGURE 1.2: Components of a DNA helix. A single strand of nucleic acid has a sugar-phosphate backbone to which the bases are attached. These linkages are all covalent. The other strand runs antiparallel. The two strands are held together by hydrogen bonds formed between complementary bases.

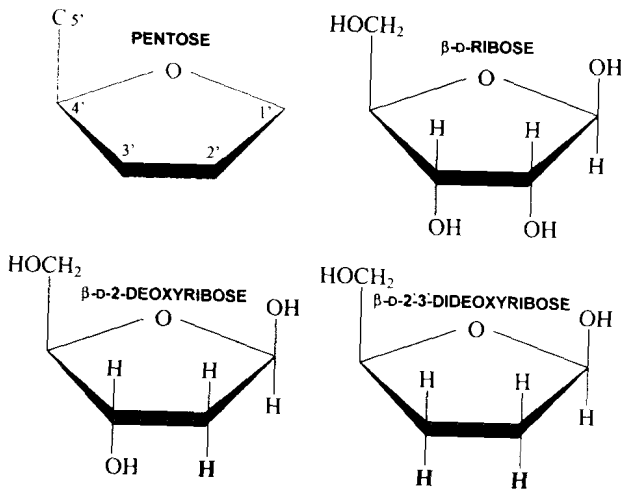


FIGURE 1.3: Sugar structures of rNTPs and ddNTPs. The sugar-phosphate backbone of RNA contains the 5-carbon sugar ribose, whereas that of DNA contains 2'-deoxyribose. ddNTPs which are used in DNA sequencing by the chain termination method (Chapter 3) contain the synthetic analog 2', 3'-dideoxyribose. The standard numbering system for the carbon atoms in the sugar part of nucleotides are designated 1', 2', etc. to distinguish them from the atoms in the base.

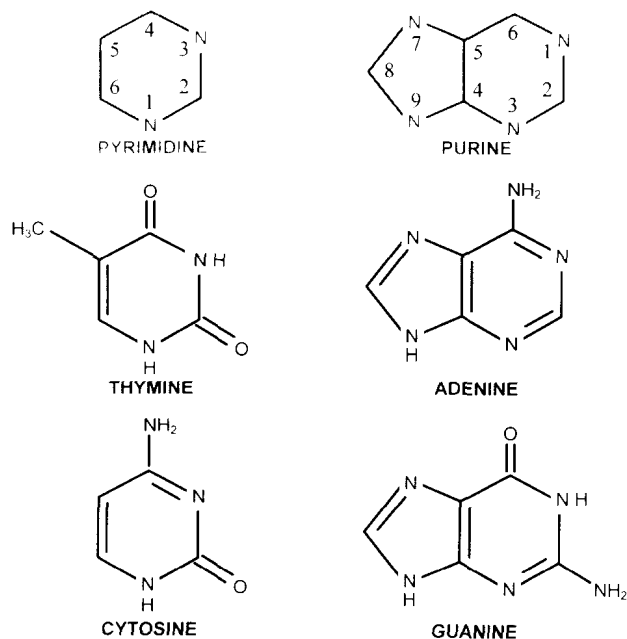


FIGURE 1.4: The structure of the bases found in DNA. Thymine and cytosine are pyrimidines, adenine and guanine are purines. The numbering system for the ring atoms is shown. Pyrimidines are linked to the sugar at N_1 , purines at N_9 .

Figure 1.6. In a linear double-stranded molecule, the 5' end of one strand is complementary to the 3' end of the other strand.

The two strands of the double helix are held together noncovalently by hydrogen bonds. The hydrogen bonds form between the complementary bases: adenine (A) pairs with thymine (T) and guanine (G) with cytosine (C) (see Figure 1.7).

Nucleoside = base + sugar

base	nucleoside
A adenine	adenosine
C cytosine	cytidine
G guanine	guanosine
T thymine	thymidine
U uracil	uridine

Nucleotide = base + sugar + phosphate

ATP	adenosine triphosphate
dATP	deoxyadenosine triphosphate
ddATP	dideoxyadenosine triphosphate
dGTP	deoxyguanosine triphosphate
dNTP	deoxynucleoside triphosphate
ddNTP	dideoxynucleoside triphosphate

FIGURE 1.5: Nomenclature of nucleic acid precursors. The abbreviations A, C, G, T, etc. are usually clearer than the full names. dNTP and ddNTP can contain any base and are often used to refer to an equimolar mix of all four (di)deoxynucleoside triphosphates.

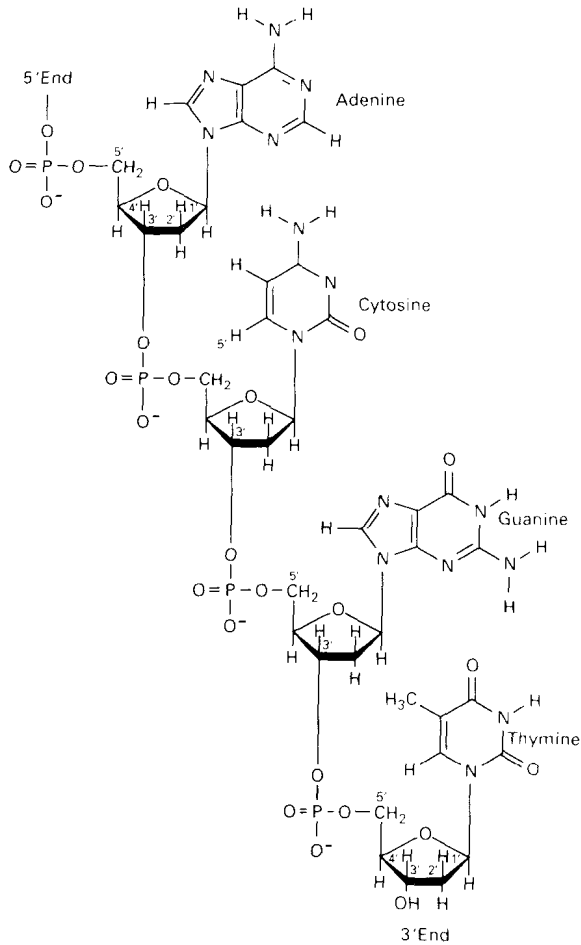


FIGURE 1.6: A single-stranded DNA molecule four nucleotides in length. Reproduced from Newton and Graham (1994) PCR, BIOS Scientific Publishers Ltd.

The related nucleic acid RNA (ribonucleic acid) differs from DNA in that the sugar in the sugar-phosphate backbone is ribose, rather than deoxyribose (see *Figure 1.3*), and uracil (U) is used in place of T.

1.3 DNA sequencing

Methods for sequencing RNA were developed earlier than for DNA, but now RNA is rarely sequenced directly. Instead, a complementary

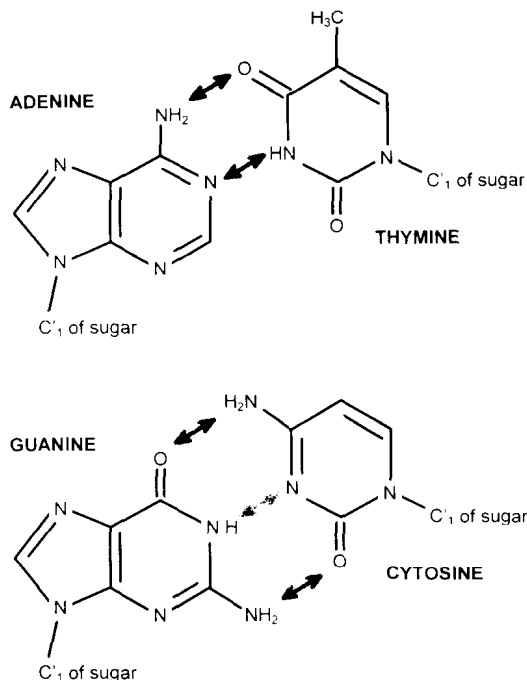


FIGURE 1.7: Base pairing in DNA. Adenine (A) pairs with its complementary base thymine (T) and guanine (G) with cytosine (C). In RNA, uracil (U), replaces thymine. Note that the separation between the glycosidic bonds and the sugars are exactly the same (10.85 Å) for each base pair.

DNA (cDNA) copy is synthesized. This cDNA is then sequenced, and the sequence of the original RNA deduced from this.

DNA sequencing is the determination of the base sequence of all or part of a DNA molecule. In the case of the molecule shown in *Figure 1.6*, DNA sequencing would advance our knowledge from 'a DNA fragment about four bases long' to 'a DNA fragment whose sequence is ACGT' (see *Figure 1.8*). Of course most, DNA molecules of biological interest are considerably longer than this!

The informational content of DNA is encoded in the order of the bases (A, C, G and T) in much the same way as binary information is stored in a computer as a string of 1s and 0s (*Figure 1.9*). The purpose of DNA sequencing is to determine the order (sequence) of these bases in a given DNA molecule. However, knowing the DNA sequence of a gene does not necessarily tell us what that gene does, any more than knowing the binary code of a computer program will necessarily tell