LABORATORY TECHNIQUES IN BIOCHEMISTRY AND MOLECULAR BIOLOGY

Vol. 10

Edited by

T.S. WORK

R.H. BURDON

DNA SEQUENCING

J. Hindley

R. Staden

DNA SEQUENCING

J. Hindley

Department of Biochemistry,
Medical School,
University of Bristol,
University Walk,
Bristol BS8 1TD, England

with a contribution by

R. Staden
MRC Laboratory of Molecular Biology,
Cambridge, England

1983

ELSEVIER BIOMEDICAL PRESS
AMSTERDAM · NEW YORK · OXFORD

ISBN – series: 0 7204 4200 I
– vol. 10: 0 444 80497 8
– vol. 10, paperback: 0444 80385 8

Published by:

ELSEVIER BIOMEDICAL PRESS
P.O. BOX 211
1000 AE AMSTERDAM, THE NETHERLANDS

Sole distributors for the U.S.A. and Canada:

ELSEVIER SCIENCE PUBLISHING CO., INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017

Printed in The Netherlands

Acknowledgements

The writing of this book was only made possible in its present form by the generous help afforded by Dr. F. Sanger and his colleagues, both past and present, at the M.R.C. Laboratory of Molecular Biology, Cambridge and Dr. Allan Maxam of Harvard University, both in supplying me with unpublished data and in granting permission to make extensive use of their procedures and observations. I am also indebed to Dr. Andrew J.H. Smith for permission to make use of material in his Ph.D. thesis (University of Cambridge, 1980) in connection with the description of the exonuclease III procedure for making single-stranded template DNA, to Drs. Jan Maat, Isabelle Seif and B. Barrell for autoradiographs of sequencing gels and to Dr. N.L. Brown of this Department for the detailed description of the depurination and ribosubstitution reactions and the accompanying autoradiographs.

Finally I wish to thank Miss G.A. Phear for checking the manuscript and several colleagues for discussions concerning particular points of technique and experimental observations.

Contents

Chapter 3. Chain terminator sequencing 72

CHAPTER 1

Introduction

1.1. Preliminary remarks

The present art of DNA sequencing has its origins in a variety of different fields of nucleic acid enzymology and chemistry. Indeed as early as 1970 our knowledge and understanding of these fields was, in theory, sufficiently far advanced to anticipate the development of the modern rapid methods but two obstacles had first to be overcome to convert these ideas into reality. The first was the problem of separating the oligonucleotides, generated in the sequencing reactions, in a rapid convenient and reproducible manner and displaying them as an ordered set of fragments according to their chain length. While the technique of homochromatography, in which a random mixture of polynucleotides of all possible chain lengths is used to develop a chromatogram (Brownlee and Sanger, 1969), was an important step in this direction, it was through the development of gel electrophoretic techniques that this problem was finally solved. All the methods to be described rely on the extraordinary resolving power of polyacrylamide gels run under denaturing conditions to achieve the final separations; much effort has gone into perfecting such systems so as to optimise their resolving properties. There is no doubt that, despite the power of this system, it remains the limiting step as far as sequence determination is concerned and further technical advances can be expected.

The second obstacle in thinking about new methods was conceptual rather than practical. Most biochemists with an eye on DNA sequencing had a background in either RNA or protein sequencing and it is clear that the earlier methods developed for DNA sequencing were essentially an extrapolation of the methods used for RNA sequencing (cf. Brownlee, 1972). A radical rethinking of the entrenched 'divide, fractionate and analyse' approach can, in retrospect, be seen as the other prerequisite for development of modern methods. This was, in a sense, anticipated in some of the later developments in RNA sequencing such as the end-labelling of polynucleotides with polynucleotide kinase (Szekeley and Sanger, 1969) and the use of sequential degradation methods.

The real breakthrough was the appreciation of the limitations inherent in trying to modify RNA sequencing methods to suit the DNA problem. In this respect it is perhaps fortunate that enzymes with the type of base specificity exhibited by the RNA endonucleases are not available for degrading DNA and this also played its part in driving the development of the modern chemical and chain termination procedures for DNA sequence analysis.

As with any other type of chemical or biochemical analysis of biological macromolecules the first hurdle is the isolation of the molecular species in sufficient quantity and purity. DNAs isolated from bacterial plasmids, bacteriophages and small animal viruses e.g. polyoma and SV40, are usually homogeneous single species and the preparation of a few hundred micrograms (equivalent to approx. 100 pmol of these DNAs) is a routine procedure. Since DNA sequencing procedures depend on radiochemical labelling, achieved by one means or another, to exhibit and relate the fractionated products, the sensitivity of the methods can be remarkably high; thus about 300 μg of DNA from a typical small plasmid such as pBR322 is sufficient for the chemical sequence analysis of one of the antibiotic resistance genes. However, as one moves up the evolutionary scale to progressively more complex organisms the isolation of particular genomic regions becomes a problem in its own right and it is usually necessary to employ molecular cloning methods to both isolate, and identify, and subsequently amplify the DNA fragment in question. *E. coui* has a haploid DNA content corresponding to a DNA of molecular weight 2.4×10^9.

This is equivalent to about 4×10^6 base pairs, about 1000 times larger than that of a small plasmid or virus. The DNA content of a typical mammalian cell nucleus is in turn about 1000 times greater than that in the *E. coli* cell. Moreover the DNA segments of such macromolecules are for all practical purposes indistinguishable from each other. Thus to isolate physically 100 pmol of a gene (10^3–10^4 base-pairs) from the total human genome (10^9 base pairs) would require in excess of 100 g of DNA from 100 kg of tissue—even supposing physical and chemical methods were somehow capable of resolving a single gene-sized piece from several hundred thousand others. In theory, and now with increasing frequency in practice, this problem has been solved by recombinant DNA techniques and nowadays DNA sequencing is inextricably linked with gene cloning. However, it is not the purpose of this book to discuss methods for gene cloning per se as a technique for isolating the DNA segment in question and we shall assume that as the starting material we already have the DNA either as a pure sample or cloned in phage or plasmid vector. The size of the DNA segment is the most important factor in determining the strategy of our sequencing approach. A fragment of a few hundred nucleotides may be directly amenable to sequencing whereas a more complex molecule of several thousand base pairs will usually require some sub-cloning of its restriction fragments to obtain defined sequenceable pieces.

The revolution in DNA sequencing technology has also dramatically affected our thinking and approach to RNA sequencing. The traditional RNA technology perhaps reached its culmination in the sequence determination of the genome of phage MS-2, a small coliphage containing a single stranded RNA genome of 3500 nucleotides. However, the effort needed to carry through this undertaking made it unlikely that similar approaches would form the basis of routine methods for studying the sequence of large mRNA species, or the genomes of influenza or the type 'C' tumour viruses. Application of the principles used for DNA sequencing have now made such problems much less daunting and a brief description of the methods currently used for RNA sequencing as based on DNA technology, have been included in this monograph.

One final point, which needs emphasis, is that at the time of writing the methods documented are the most up to date in current use. However, judging by past experience, it is likely that these will be further improved and amended as sequencing objectives become more ambitious and complex.

1.2. Organisation of the book

The modern rapid methods for DNA sequencing fall into two broad groups depending on the procedures used to generate and relate sets of labelled oligonucleotides which, after resolution by gel electrophoresis, permit the DNA sequence to be deduced.

The first group of methods employs a primed synthesis approach in which a single stranded template, containing or comprising the sequence of interest, is copied to produce the radioactively labelled complementary strand. By using chain-terminating inhibitors, and also by other methods, sets of partially elongated molecules are produced which can be fractionated on denaturing gels and the patterns of labelled bands obtained used to deduce the sequence. This procedure is very adaptable and though originally devised for sequencing naturally occurring single stranded DNAs, highly efficient procedures have now been developed for generating single stranded templates from any duplex DNA.

The forerunner of the primed synthesis methods was the 'plus and minus' method developed by Sanger and Coulson in 1975 and this marked the first real breakthrough in the search for new and efficient ways of sequencing DNA. This method is the subject of Chapter 2. By itself this procedure does have distinct limitations and additional back-up procedures had to be employed to confirm regions of the sequences deduced. One of these, the depurination method originally introduced by Burton and Petersen, proved a particularly useful adjunct and a description of this method is included as applied to the analysis of pyrimidine clusters in the products of a primed synthesis reaction. Another development which made the method more flexible was the single site ribosubstitution reaction and examples of the use of this technique are discussed. Chapter 2 is completed by a description of the classical 'wandering-spot' method for analysing short DNA sequences (Sanger et al., 1973). This method depends on the identification, by mobility changes in a two-dimensional fractionation procedure, of sets of oligonucleotides produced by the stepwise degradation of a labelled DNA with spleen phosphodiesterase (a 5′ → 3′ exonuclease) and venom phosphodiesterase (a 3′ → 5′ exonuclease). Though this method is more usefully applied to end-labelled, or uniformly labelled DNA, it is included here since, with the depurination method, it was the mainstay of the earlier DNA sequencing procedures and is still widely used for particular purposes.

Chapter 3 is concerned with the further developments in primed synthesis methods made possible by the introduction of the chain terminating dideoxy-nucleotides as specific inhibitors. A description of the background of this method is followed by detailed experimental methods and an account of the application of this method o DNA sequencing using the different primer–template combinations that can be obtained, starting from duplex DNA by taking advantage of the remarkable versatility of exonuclease III. This enzyme can be used to prepare either single stranded templates or primers, and in conjunction with the use of restriction enzymes to cleave out and select particular fragments the procedure has been developed into an extremely powerful and generally applicable sequencing procedure. DNA polymerase in the presence of dideoxynucleoside triphosphates can also be used to generate specific sets of fragments from 5′-end-labelled duplex DNA which has been nicked with DNAaseI, and this has been developed into a set of useful and versatile sequencing methods. These are also considered in Chapter 3.

Chapter 4 describes the application of primed synthesis methods to sequencing single stranded DNAs prepared by cloning restriction fragments into derivatives of the single stranded DNA phage

M13. The development of these phage vectors solved the general problem of preparing single stranded template from any double stranded DNA. The further development of new vectors with a variety of cloning sites together with the use of universal flanking primers which anneal adjacent to the cloned sequence have made this into the currently first choice primed-sequencing method. At the time of writing, this is under active development and new procedures in which the random selection of recombinant clones is replaced by a more structured approach can soon be expected. Sequencing via cloning into M13 is probably the most rapid and versatile procedure at the present. One way in which this can be structured by using exonuclease III treated restriction fragments as primers is discussed in this chapter. Current approaches to RNA sequencing, by analysis of cDNA transcripts synthesized by the use of reverse transcriptase are also reviewed in Chapter 4 with emphasis on the most recent developments.

Chapter 5 is devoted to the Maxam–Gilbert chemical method for sequencing 5′ or 3′ end-labelled DNA. The description and discussion is based on their most recent recommendations and a detailed experimental protocol is included. The final section is a discussion of the strategies which may be used for sequencing double stranded DNA by combining the Maxam–Gilbert approach with methods which permit the ordered sequencing and restriction mapping of DNA and ways for comparing and pinpointing regions of sequence divergence in related DNAs.

With this wealth of techniques available the main problem is likely to be one of selection. As far as possible the advantages of the different methods are discussed to help in choosing the most appropriate method for a particular problem.

At the end of this volume a series of appendices give useful information regarding sources of enzymes etc.

Chapter 6 describes the computer methods developed by Dr. Roger Staden at the MRC Laboratory of Molecular Biology, Cambridge, for handling the sequence data produced by the rapid 'shotgun' sequencing techniques described in Chapter 4. Since sequence data can be accumulated at a rate of up to 1,000 nucleotides a day the major problem soon becomes one of handling and keeping track of the sequences obtained. A computer can store the data and with the aid of suitable programmes can edit, analyse and print out the data in different forms. Sequences can be compared, matched up, modified and searched for overlaps or other common subsequences as each new piece of data is entered from the gel reading. Searches for repeated sequences, palindromes and restriction enzyme recognition sites can be carried out and additional programmes can search for regions of secondary structure such as hairpin loops, and translate DNA sequences into amino acid sequences. As the complexity of sequencing objectives increase so does the value of programmes capable of handling the immense amount of data and analysing it for particular features. Chapter 6 describes these programmes and the appendix contains a step by step description of how to run the programmes.

1.3. General background to DNA sequencing

The procedures described in this book make extensive use of enzymes which can synthesize, modify or degrade DNA. All the procedures rely on the incorporation of one or more [^{32}P]-labelled nucleotide residues into the DNA molecule to be sequenced and the sequence is finally deduced by the examination of autoradiographs of electrophoretic separations on polyacrylamide gels in which the different labelled fragments are ordered in a size-dependent manner. The ability to dissect DNA sequences into sequenceable sized fragments in a defined and specific fashion, the separation and identification of these fragments and their cloning into a vector DNA are all techniques which are inextricably involved in the modern sequencing methods. For the reader not familiar with this field the remainder of this chapter describes in a general way, the scientific background to the various enzymatic reactions used and their application to sequencing problems. In addition a description is given of the restriction endonucleases (restriction enzymes) for cleaving DNA sequences to produce fragments with defined termini for sequencing or cloning, or as a means of obtaining a 'restriction map' of the DNA sequence in question.

1.3.1. DNA synthesis with DNA polymerase I

DNA Polymerase I (DNA Pol I) from *E. coli* is the best characterized of all the DNA polymerases and, as far as DNA sequencing is concerned, is by far the most widely used. The purified enzyme is a single polypeptide chain of molecular weight 109,000 daltons containing approximately 1000 amino acid residues. The turnover number is about 667 nucleotides polymerized per molecule of enzyme per min at 37°C. One atom of zinc per molecule of enzyme appears to be required for the activity of the enzyme.

DNA polymerases (like RNA polymerases) are unique among enzymes in that the choice of substrate is determined by the template. *E. coli* polymerase can copy eukaryotic DNA and animal polymerases can copy bacterial DNA sequences given the appropriate DNA template. This means that DNA sequences, irrespective of their origin, can be accurately copied using *E. coli* DNA Pol I and this is the basis of the 'primed-synthesis' DNA sequencing methods.

The synthesis of DNA has two basic features. First, a phosphodiester bond is formed between the 3′-hydroxyl group (3′-OH) at the growing end of a DNA chain (the *primer* DNA strand) and the 5′-phosphate group of the incoming deoxynucleotide. The reaction involves a nucleophilic attack by the 3′-OH group at the innermost (α)-phosphate of the incoming deoxynucleoside triphosphate with the release of pyrophosphate. The direction of chain growth is $5' \rightarrow 3'$ (Fig. 1.1.). Secondly, each deoxynucleotide residue added to the growing end of the primer is selected by its ability to base-pair with the complementary nucleotide on the DNA *template* strand. The sequence of nucleotides added to the primer is therefore exactly complementary to the sequence of the template strand (Fig. 1.1.). DNA polymerase is unable to initiate

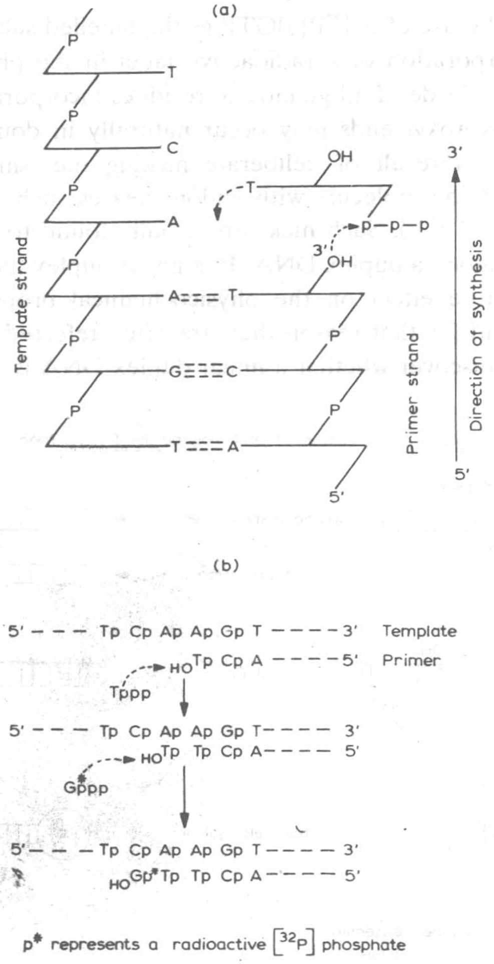
(a)
P
T
P
C
OH
3'
T
P
A
p—p—p
3'
OH
Template strand
P
A≡≡≡T
P
P
G≡≡≡C
P
P
T≡≡≡A
5'
Primer strand
Direction of synthesis
5'
(b)
5' — — — Tp Cp Ap Ap Gp T — — — — 3' Template
HO Tp Cp A — — — — 5' Primer
Tppp
5' — — — Tp Cp Ap Ap Gp T — — — — 3'
HO Tp Tp Cp A — — — — 5'
Gp*pp
5' — — — Tp Cp Ap Ap Gp T — — — — 3'
HO Gp* Tp Tp Cp A — — — — 5'
p* represents a radioactive [32P] phosphate

Fig. 1.1. Mechanism of chain extension catalysed by DNA polymerase on a primed template. (For discussion see text.)

the synthesis of new DNA strands in the absence of a free 3′-OH group. If one (or more) of the deoxynucleotide triphosphates are labelled with [^{32}P] in the innermost phosphate then the resulting synthesized DNA strand will be radioactively labelled. Thus, in