A Biologist's Guide to Principles and Techniques of Practical Biochemistry

Third Edition

Edited by
Keith Wilson

B.Sc., Ph.D.
Head of Division of Biological and Environmental Sciences,
The Hatfield Polytechnic

and
Kenneth H. Goulding

M.Sc., Ph.D.
Head of School of Applied Biology,
Lancashire Polytechnic

Edward Arnold

First published in Great Britain 1975 by
Edward Arnold (Publishers) Ltd, 41 Bedford Square, London WC1B 3DQ

Edward Arnold (Australia) Pty Ltd, 80 Waverley Road, Caulfield East,
Victoria 3145, Australia

Edward Arnold, 3 East Read Street, Baltimore, Maryland 21202, U.S.A.
Reprinted 1976, 1979
Second edition 1981
Reprinted with corrections 1983
Reprinted 1984
Third edition 1986

British Library Cataloguing in Publication Data

A Biologist's guide to principles and
techniques of practical biochemistry.——
3rd ed.——(Contemporary biology)
1. Biological chemistry——Technique
I. Wilson, Keith, *1936–* II. Goulding,
Kenneth H. III. Series
574.19'2'028 QP519.7

ISBN 0-7131-2942-5

Text set in 10/11pt Times Compugraphic
by Colset Pte. Ltd., Singapore
Printed and bound in Great Britain by Richard Clay Ltd., Bungay, Suffolk

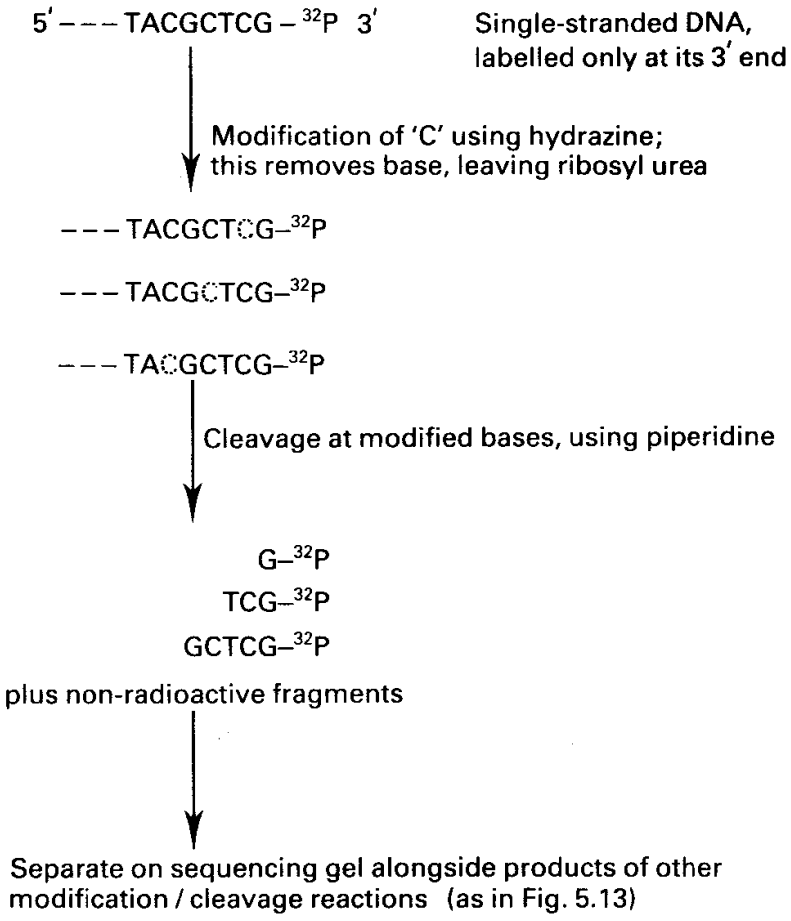

Fig. 5.14 Maxam and Gilbert sequencing of DNA. Only modification and cleavage of deoxycytidine is shown, but three more aliquots of the end-labelled DNA would be modified and cleaved at G, G + A, and T + C, and the products would be separated on the sequencing gel alongside those from the 'C' reactions.

to that produced by the Sanger method, since each sample now contains radioactive molecules of various lengths, all with one end in common (the labelled end), and with the other end cut at the same type of base. Analysis of the reaction products by electrophoresis is as described for the Sanger method.

Because the Sanger method produces oligonucleotides which are radioactively labelled throughout their lengths, rather than only at one end, the molecules can be made a lot more radioactive, and therefore easier to detect; so less DNA is needed for sequencing. Once M13 cloning has been set up in a laboratory, it provides a very convenient and rapid way to obtain single-stranded DNA. For these reasons, dideoxy sequencing of M13-cloned DNA is probably the most commonly used sequencing method, though the chemical procedure is still used by many laboratories.

5.5.3 Protein sequencing

Although protein sequencing may seem out of place in a section dealing with the analysis of DNA, the molecular biologist can often make use of a

knowledge of protein sequences when manipulating DNA. If the sequence of a protein is known, a gene coding for it can be synthesised chemically (though this is usually only worth doing for small polypeptides), or an *oligonucleotide probe* can be synthesised for use in recovering the gene for that protein from a *gene library* (Section 5.9.5).

Since it is currently impossible to sequence a polypeptide longer than about 100 amino acids, pure proteins must be fragmented to give polypeptides of a length which can be sequenced, and these polypeptides must be separated from each other prior to sequencing. Fairly specific and limited cleavage can be obtained by chemical means. For example, cyanogen bromide cleaves only at (rare) methionine residues, BNPS-skatole cleaves at tryptophan, and hydroxylamine breaks the linkage between asparagine and glycine. Similarly, several proteolytic enzymes, such as trypsin and V8-Protease, have a fairly specific site of action, and will therefore generate relatively few cleavage products.

The polypeptides so produced are separated from each other prior to sequencing, using such techniques as exclusion chromatography (Section 6.6) or HPLC (Section 6.8). Relative positions of the polypeptides within a protein can be found by looking for overlaps in the sequences of polypeptides generated by different means, and in this way the entire protein sequence may be deduced.

All protein sequencing methods are based on the *Edman degradation* of polypeptides, in which the *N*-terminal amino acid is specifically removed, leaving a polypeptide one amino acid residue shorter. Variations arise in the method of identifying the removed amino acid or the newly exposed *N*-terminal amino acid. By repeated cycles of Edman degradation and identification of product, the polypeptide can be sequenced.

In the *Edman reaction* (Fig. 5.15) the polypeptide is treated with phenylisothiocyanate (PITC), which reacts with the *N*-terminal amino acid to form a phenylthiocarbamyl (PTC) derivative of the polypeptide. Anhydrous trifluoroacetic acid is then used to cleave the molecule, giving the 2-anilino-5-thiazolinone derivative of the *N*-terminal amino acid and also the polypeptide shortened by one residue. The thiazolinone derivative is separated from the polypeptide and converted into the more stable 3-phenyl-2-thiohydantoin (PTH) derivative, which is then identified by HPLC or TLC. By repeating this cycle the polypeptide can be sequenced from its *N*-terminal end. The process has been automated, either by immobilising the protein on an inert, solid support (*solid-phase sequencers*), or by keeping the protein spread out in a thin film for maximum exposure to reagents (*spinning cup sequencers*). Such instruments can, under ideal conditions, sequence up to 100 residues of a protein.

The alternative *Dansyl-Edman* procedure (Fig. 5.16) is highly sensitive, allowing as little as 1 nmole of polypeptide to be sequenced, and it is therefore well suited to manual determination of sequences. It uses cycles of the Edman reaction to remove *N*-terminal amino acids sequentially, but, instead of identifying the released PTH derivatives, it identifies the newly exposed *N*-terminal amino acids. This is achieved by adding a dansyl group to the *N*-terminal of a very small sample of the polypeptide after each cycle of the

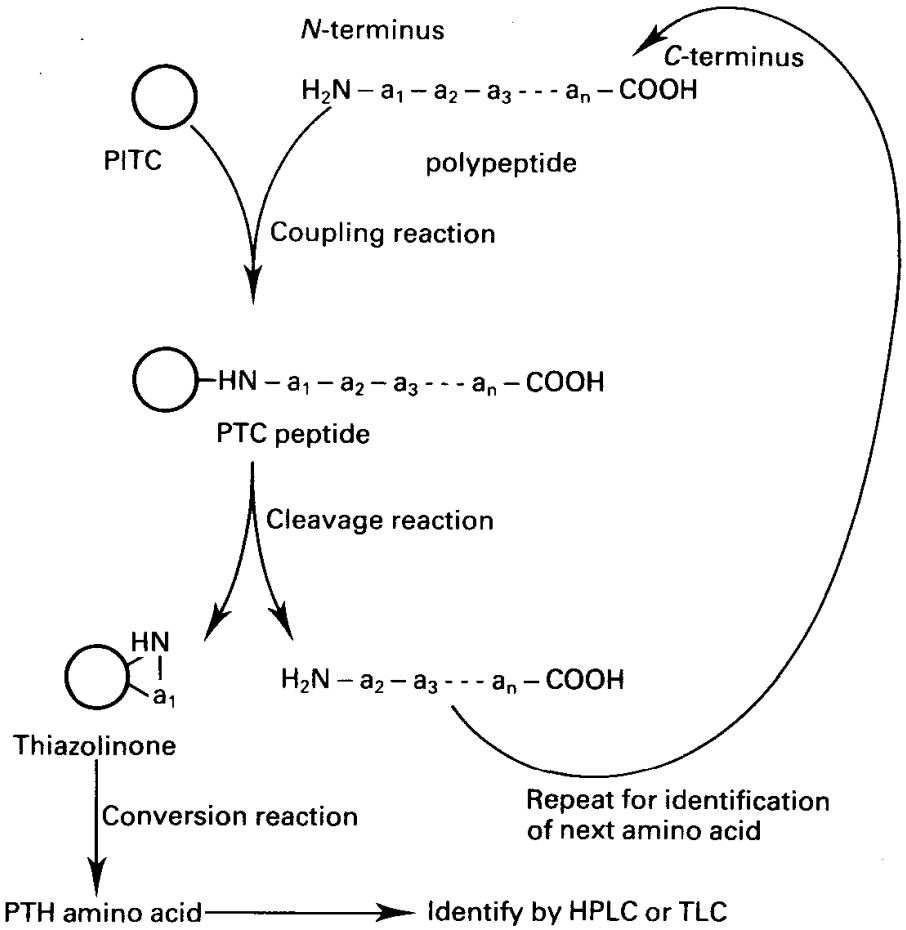

Fig. 5.15 Edman reactions. PITC phenylisothiocyanate; PTC, phenylthiocarbamyl; PTH 3-phenyl-2-thiohydantoin. Note that each cycle of reactions removes one amino acid from the *N*-terminus of the polypeptide.

Edman reaction, followed by cleavage with hydrochloric acid to release a dansyl amino acid plus free amino acids. The dansyl derivative can be identified by two-dimensional TLC on polyamide plates (Section 6.1.3). Up to about 15 amino acids can be sequenced before the cumulative effects of incomplete reactions and side reactions make impossible the unambiguous identification of the dansyl amino acid.

Given the nucleotide sequence of a gene, and our knowledge of the genetic code, it is easy to read off the amino acid sequence for which the gene codes, provided the correct reading frame is used, and the sequence is not interrupted by introns. Ironically, DNA sequencing, rather than protein sequencing, has sometimes been used to obtain amino acid sequences of proteins, especially when the pure protein has not been obtainable in sufficient quantities for direct sequencing. However, it should be remembered that a lot of effort is involved in the isolation of a specific gene, and this may more than offset the rapidity of DNA sequencing. The pace of sequencing is such that computers are now used by some laboratories for the analysis of sequencing gels, and sequence data banks have been set up to cope with the massive flow of information. In spite of this, it will be some time before the human genome

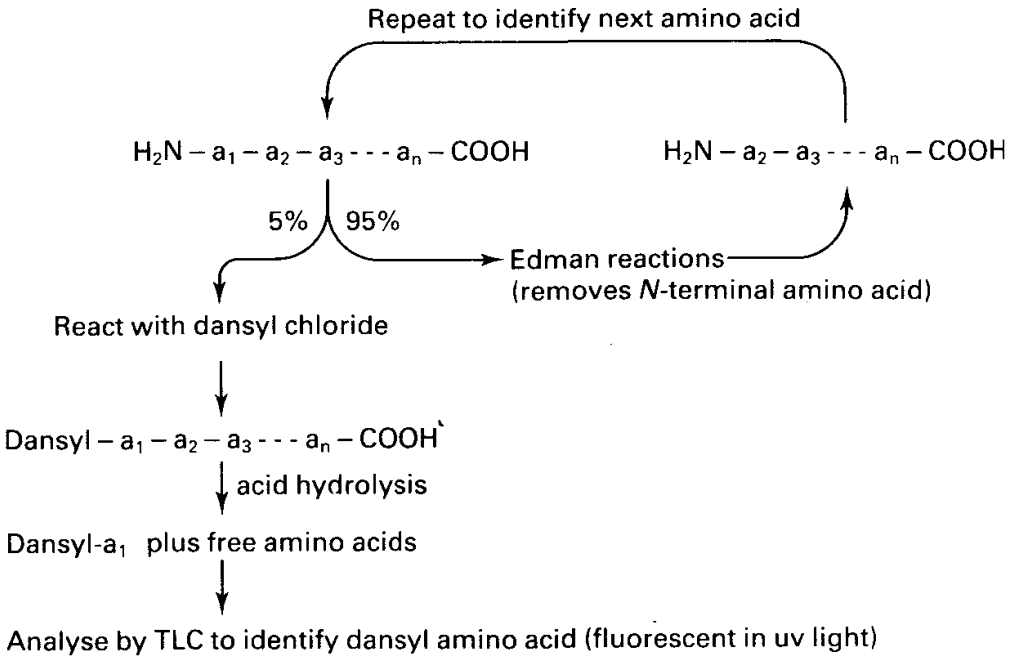

Fig. 5.16 Dansyl-Edman procedure. Only the *N*-terminal amino acid becomes dansylated, and can therefore be identified by TLC. The Edman degradation is used to remove *N*-terminal amino acids one-by-one from the polypeptide, and dansylation allows the identification of each newly exposed *N*-terminal residue.

is completely sequenced. Even at a rate of one base per second, the 3×10^6 kb of the haploid genome would take more than 100 years to be sequenced.

5.5.4 Renaturation kinetics

When preparations of double-stranded DNA are denatured by heat or alkali, and then allowed to renature, measurement of the *rate of renaturation* can give valuable information about the *complexity* of the DNA, i.e. how much information it contains (measured in base-pairs). The complexity of a molecule may be much less than its total length if some sequences are *repetitive*, but complexity will equal total length if all sequences are *unique*, appearing only once in the genome. In practice, the DNA is first cut randomly into fragments about 1 kb in length (Section 5.9.2), and is then completely denatured by heating above its T_m. Renaturation at a temperature about 10°C below the T_m is monitored either by decrease in absorbance at 260 nm (the *hypochromic effect*), or by passing samples at intervals through a column of hydroxyapatite, which will adsorb only double-stranded DNA, and measuring how much of the sample is bound. The degree of renaturation after a given time will depend on C_0, the concentration (in nucleotides per unit volume) of double-stranded DNA prior to denaturation, and t, the duration of the renaturation.

For a given C_0, it should be evident that a preparation of λ DNA (genome size 49 kb) will contain many more copies of the same sequence per unit