laboratory techniques

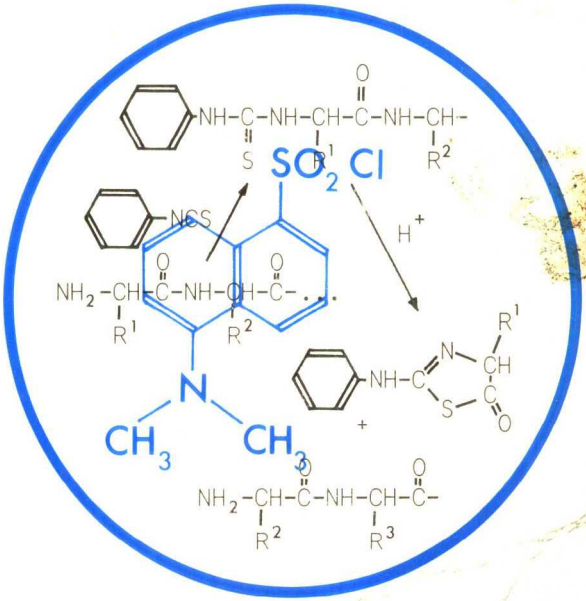
in biochemistry and molecular biology

general editors: R.H. BURDON and P.H. van KNIPPENBERG

sequencing of proteins and peptides

Second revised edition

G. ALLEN



SEQUENCING OF PROTEINS AND PEPTIDES

Second revised edition

G. ALLEN

Wellcome Biotech
Langley Court
Beckenham
Kent BR3 3BS
England

1989
ELSEVIER
AMSTERDAM · NEW YORK · OXFORD

ISBN 0-444-81021-8 (pocket edition)
ISBN 0-444-81022-6 (library edition)
ISBN 0-7204-4200-1 (series)

1st edition	*1981*
2nd printing	*1983*
3rd printing	*1986*
4th printing	*1986*
5th printing	*1986*
2nd edition	*1989*

Published by:
ELSEVIER SCIENCE PUBLISHERS B.V. (BIOMEDICAL DIVISION)
P.O. BOX 211
1000 AE AMSTERDAM
THE NETHERLANDS

Sole distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
655 AVENUE OF THE AMERICAS
NEW YORK, NY 10010
U.S.A.

Library of Congress Card No. 85-647011

Printed in The Netherlands

Preface

In the preface to the first edition of this book it was suggested that the ease with which DNA sequences could now be determined could increase the demand for analysis of protein primary structure. While it is true that the large majority of new protein sequences are now deduced from nucleotide sequences of cloned cDNA (or, in the case of prokaryotic proteins, from cloned genomic DNA fragments) there is a continuing high level of activity in direct studies on protein primary structures. With the availability of efficient methods for the synthesis of oligodeoxynucleotides, the most direct route to the isolation of cDNA or genomic DNA clones, a prerequisite for DNA sequence determination, is the use of specific probes designed from partial sequences already determined by direct protein sequencing methods. This requirement for partial sequences, often obtained at high sensitivity from the amino-terminus of the protein, has provided a major impetus for the development and dissemination of highly sensitive techniques. A further part of this activity is directed at the determination of structures introduced into proteins by post-translational modifications, such as glycosylation, methylation, phosphorylation, proteolysis and disulphide bonding, among many structural features that DNA sequences alone cannot reveal. The sequence of a protein can not in general be predicted from genomic DNA, at least in higher eukaryotes, since messenger RNAs are frequently generated by splicing of longer precursors with deletion

of intervening sequences (introns); while splicing sites have some common features, the rule is not completely universal, direct protein sequencing work is also required for confirmation of sequences predicted from DNA sequences when proteins from unusual sources are being studied. A further example where deductions from DNA sequences are misleading is the occurrence of frameshifting during the translation of some mRNAs. Chemical modification studies of proteins often require partial sequencing to reveal sites of specific reaction; the structures of 'active-site' peptides may be of particular interest. Again, the DNA sequence alone cannot provide the necessary data. The requirement, not least by regulatory authorities, for complete characterisation of protein molecules intended for prophylactic or therapeutic use in humans or other animals, and available through the new technology of genetic engineering, has also led to an increased demand for protein sequencing. The recent advances in peptide synthesis methods could also be expected to generate the need for degradative work to characterise the synthetic products, although the accuracy and reliability of the synthetic approach, together with quantitative monitoring during the course of the synthesis make such sequencing work unnecessary in most cases.

In the present decade significant changes have taken place in the techniques for protein sequence determination. There has been a pronounced shift towards the use of more sophisticated automated high-performance chromatographic techniques for purification of proteins and peptides, analysis of purity, and, almost to the exclusion of other methods such as thin-layer or gas-liquid chromatography, for the analysis of phenylthiohydantoin derivatives of amino acids released during the Edman degradation. The gas-phase automated sequencer has come to dominate the field of high-sensitivity sequence determination, although significant improvements in the longer-established automated liquid and solid-phase technologies have also been made. Significant advances in mass spectrometric techniques, such as fast atom bombardment mass spectrometry, have also begun to have an impact on protein sequence determination, likely to become more widespread than the previously used tech-

niques. The shift away from the simpler manual techniques is reflected in this second edition. There is still a need for the use of inexpensive methods, but it must be recognised that for work at high sensitivities, and for speed in accumulating results, at least one HPLC instrument is essential in a protein sequencing laboratory.

Detailed descriptions of the use of automated sequencers are still not included here, however, since the information provided by the manufacturers and available in the specialist literature is far more extensive than can be presented in this volume.

As in the first edition, it is impossible to cover the field of protein sequence determination fully in a laboratory manual of this type, and a great deal of selectivity has been applied, some of it subjective. Detailed descriptions are given of techniques that have been found in many laboratories to be reliable in yielding the required sequence data, and references to many alternative techniques preferred by some authors are also provided. The choice between equivalent techniques will depend to some extent upon the equipment available and the techniques already established in the investigator's laboratory.

I gratefully acknowledge the assistance of Mr. T.J. Warren of Applied Biosystems Ltd. for supplying a bibliography of the use of the gas phase sequencer.

List of abbreviations

Atz-amino acid	anilinothiazolinone derivative on an amino acid
BNPS-skatole	2-(2-nitrophenylsulfenyl)-3-methyl-3-bromoindolenine
butyl-PBD	2-(4′-*t*-butylphenyl)-5-(4″-biphenylyl)-1,3,4-oxadiazole
CM-	carboxymethyl-
Dabitc	4,4-dimethylaminoazobenzene-4′-isothio cyanate
Dabtc-	4,4-dimethylaminoazobenzene-4′-thiocarbamyl-
Dabth-amino acid	4,4-dimethylaminoazobenzene-4′-thiohydantoin derivative of an amino acid
DEAE-	diethylaminoethyl-
Dnp-	2,4-dinitrophenyl-
Dns-	1-dimethylaminonaphthalene-5-sulphonyl-
Dns-Cl	1-dimethylaminonaphthalene-5-sulphonyl-chloride (dansyl chloride)
DTNB	5,5′-dithio*bis*(2-nitrobenzoic acid)
E. coli	*Escherichia coli*
EDTA	ethylenediamine tetraacetic acid
FABMS	fast-atom bombardment mass spectrometry

FPLC	fast protein liquid chromatography
GLC	gas-liquid chromatography
HPLC	high-performance liquid chromatograp
Pitc	phenylisothiocyanate
Ptc-	phenylthiocarbamyl-
Pth-amino acid	phenylthiohydantoin derivative of an amino acid
SDS	sodium dodecyl sulphate
TLC	thin-layer chromatography
TLCK	3-tosylamido-7-amino-1-chloroheptan-2-one
TPCK	1-chloro-3-tosylamido-4-phenylbutan-2-one
ϵ	molar extinction coefficient in litre mol^{-1} cm^{-1}

Contents