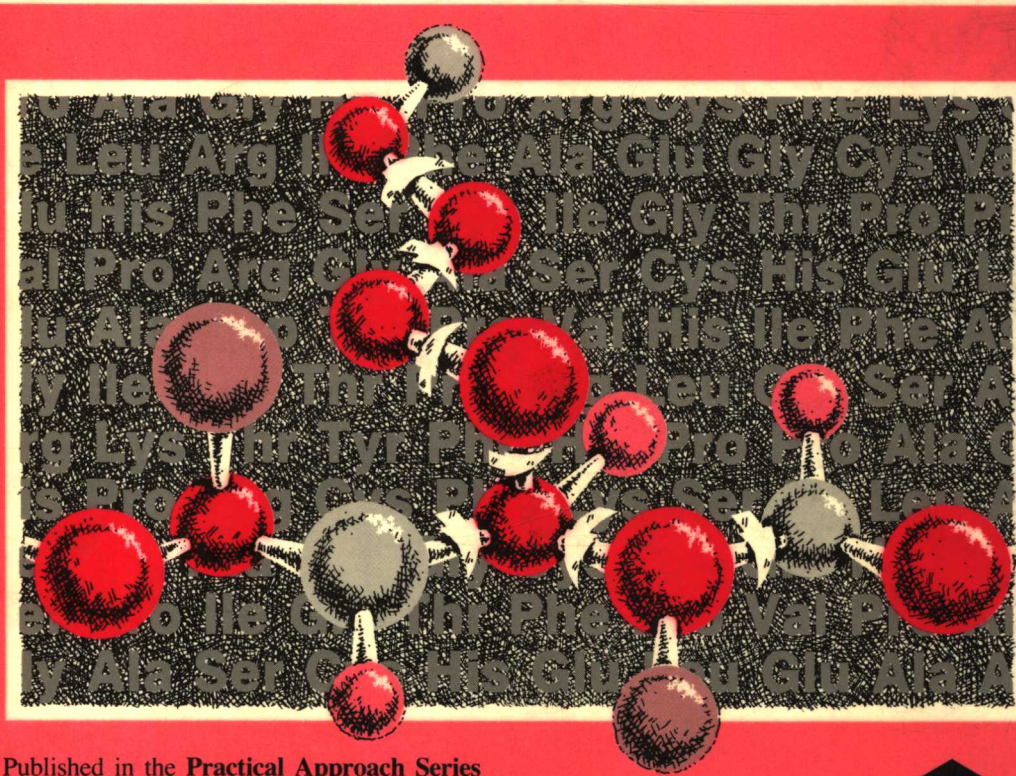
Protein sequencing

a practical approach

Edited by
J B C Findlay & M J Geisow



Published in the **Practical Approach Series**
Series editors: D.Rickwood & B.D.Hames

IRL PRESS
at
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

Protein sequencing

a practical approach

Edited by

J B C Findlay

Department of Biochemistry, University of Leeds, Leeds LS2 9JT, UK

M J Geisow

Delta Biotechnology, Castle Court, Castle Boulevard, Nottingham NG7 1FD, UK

IRL PRESS
at
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

IRL Press
Eynsham
Oxford
England

First published 1989

British Library Cataloguing in Publication Data

Protein sequencing.
1. Peptides & proteins. Sequences. Chemical analysis
I. Findlay, J.B.C. II. Geisow, Michael J. III. Series
547.7'5046

Library of Congress Cataloging-in-Publication Data

Findlay, J. B. C. (John B. C.)
Protein sequencing.

Includes bibliographies and index.
1. Amino acid sequence—Congresses. I. Geisow, Michael J. II. Title.
QP551.F65 1988 547.7'5 88-32025
ISBN 0 19 963012 7 (hardbound)
ISBN 0 19 963013 5 (softbound)

Previously announced as:
ISBN 1 85221 090 7 (hardbound)
ISBN 1 85221 091 5 (softbound)

Typeset and printed by Information Printing Ltd, Oxford, England.

Preface

The ability to determine the primary structure of a polypeptide was a major advance in biochemical sciences and sequencing work since then has had an important impact on our understanding of protein structure and function. At present, protein sequence analysis occupies a position at the forefront of modern molecular biology. Partial sequence determinations using microgram amounts of protein has enabled many genes with low levels of transcription to be cloned and their complete primary structures to be determined. This approach is also the only means available for determining the sites of post-translational modification.

This book is aimed at those wishing to isolate proteins or peptides and to carry out subsequent sequence analysis. It is written for both professional and inexperienced research workers and for those with and without access to high technology instrumentation.

At the start, much space is given over to the protein and peptide purification with particular emphasis on the important techniques of microscale HPLC and electrophoretic/electroelution techniques. The next chapter details a wide range of methods available for amino acid analysis, cleavage and detection of polypeptides and protein modification (also included as an appendix). The protocols described have been carefully selected for their general usefulness.

The sensitivity and efficacy of sequencing procedures have improved dramatically over recent years due to the development of new reagents, new protocols, new strategies and new instrumentation. The most effective modern techniques are detailed in sections dealing with manual, gas-, liquid- and solid-phase sequencing and with the rather different approach utilizing mass spectrometry.

The information for the folding and three-dimensional structure of a protein is contained in its amino acid sequence and considerable attention is being paid to the difficult task of deciphering this code. The final chapter of the book illustrates the principle and approaches to structure prediction and provides methods which can be used without recourse to sophisticated computer programs and hardware.

As in all the 'Practical Approach' volumes, as much detail as possible has been given together with critical references which illustrate and extend the methods described. Sequence analysis is in a very active growth phase and the competition, particularly in the commercial sector, to improve chemical approaches, separation systems and sequence instrumentation is very keen. We would be very interested to hear, therefore, of advances or corrections which would improve subsequent editions of this text.

J.B.C.Findlay
M.J.Geisow

Contributors

A.Aitken
National Institute of Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

K.Biemann
Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02137, USA

J.B.C.Findlay
Department of Biochemistry, University of Leeds, Leeds LS2 9JT, UK

M.J.Geisow
Delta Biotechnology, Castle Court, Castle Boulevard, Nottingham NG7 1FD, UK

C.Holmes
Department of Biochemistry, University of Dundee, Dundee DD1 4HN, UK

J.N.Keen
Department of Biochemistry, University of Leeds, Leeds LS2 9JT, UK

D.J.C.Pappin
Department of Biochemistry, University of Leeds, Leeds LS2 9JT, UK

W.R.Taylor
National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

J.M.Thornton
Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

K.J.Wilson
Applied Biosystems, 850 Lincoln Center Drive, Foster City, CA 94404, USA

A.Yarwood
Department of Botany, Science Laboratories, University of Durham, South Road, Durham, DH1 3LE, UK

P.M.Yuan
Applied Biosystems, 850 Lincoln Center Drive, Foster City, CA 94404, USA

Abbreviations

ACTH	adrenocorticotropin
AEAPG	*N*-(2-aminoethyl)-3-aminopropyl-glass
AP	3-aminopropyltriethoxysilane
APG	aminopropyl-glass
ATZ	anilinothiazolinone
CAP	catabolite activator protein
CID	collision-induced decomposition
CPG	controlled pore glass
DABITC	4,*N*;*N*-dimethylaminoazobenzene-4′-isothiocyanate
DABTH	dimethylaminoazobenzene thiohydantoin
dansyl	see DNS
DITC	*p*-phenylene diisothiocyanate
DMA	dimethylamine
DMF	dimethylformamide
DMPTU	*N*-dimethyl-*N*′-phenylthiourea
DNS	5-dimethylaminonaphthalene-1-sulphonyl
DPTU	*N*,*N*′-diphenylthiourea
DPU	*N*,*N*-diphenylurea
DTNB	5,5′-dithionitrobenzoic acid
DTT	dithiothreitol
EDC	1-ethyl-3(3-dimethylaminopropyl)-carbodiimide
EDTA	ethylenediamine tetraacetic acid
EGF	epidermal growth factor
FAB	fast atom bombardment
FMOC	*N*-(9-fluorenylmethoxycarbonyl)
GLC	gas–liquid chromatography
GOR	Garnier–Osguthorpe–Robson
HPLC	high-performance liquid chromatography
MHC	major histocompatibility complex
MS	mass spectrometry
NBS	*N*-bromosuccinimide
NEM	*N*-ethylmorpholine
NTCB	2-nitro-5-thiocyanobenzoic acid
OPA	*o*-phthalaldehyde
PAGE	polyacrylamide gel electrophoresis
PD	plasma desorption
PIR	protein identification resource
PITC	phenylisothiocyanate
PMSF	phenylmethylsulphonyl fluoride
PSQ	protein sequence query
PTC	phenylthiocarbamyl
PTH	phenylthiohydantoin
PVDF	polyvinylidene difluoride
QA	*N*-trimethoxysilyl-propyl-*N*,*N*,*N*-trimethylammonium chloride
RP	reverse-phase
SDS	sodium dodecyl sulphate
TCA	trichloroacetic acid

TEA	triethylamine
TETA	triethylenetetramine
TFA	trifluoroacetic acid
THF	tetrahydrofuran
TLC	thin-layer chromatography
TMA	trimethylamine
TNBS	2,4,6-trinitrobenzenesulphonate

Contents

ABBREVIATIONS xi

1. PROTEIN AND PEPTIDE PURIFICATION 1
K.J.Wilson and P.M.Yuan

Introduction 1
Microbore HPLC Instrumentation 3
Hardware considerations 3
Column parameters 4
Buffer systems 5
Detection enhancement 5
Effects of column diameter and column length 8
Flow-rate effect on separation 9
Flow-rate effect on peak elution volume 9
Electrophoresis Equipment 9
General concerns of chemical modification 9
Reagents 11
Equipment 11
Preparation of gel solutions, buffers and slab gels 13
Sample preparation, loading and electrophoresis 14
Electroelution–electrodialysis recovery method 15
Electroblotting, glass fibre derivatives 16
Electroblotting, polyvinylidene difluoride 17
Sequence analysis 18
HPLC Application Examples 20
Comparison of chromatographic methods 20
Micro-level alkylation and sample recovery 21
Peptide mapping 22
Selective methods for amino acid detection 24
Selectivity modulation through buffer changes 27
Two-dimensional chromatographic isolation of albumin 29
Electrophoresis Application Examples 29
Expected recovery yields 29
Two-dimensional chromatographic/electrophoretic isolation of albumin 35
Amino-terminal sequencing yields 37
Summary 37
Acknowledgements 40
References 41

2. PEPTIDE PREPARATION AND CHARACTERIZATION 43
A.Aitken, M.J.Geisow, J.B.C.Findlay, C.Holmes and A.Yarwood

Amino Acid Composition Analysis 43

Ion-exchange systems 43
High-performance liquid chromatography 47
Location and Isolation of Particular Peptides or Peptides Containing Specific Amino Acids 50
Detection of specific amino acids in peptides 51
Determination of disulphide bonds 52
Selective isolation by hydrophobicity modulation 52
Radiolabelling Proteins for Peptide Mapping 55
Iodination of protein in gel slices 56
Digestion with trypsin 56
Two-dimensional mapping 56
Cleavage Procedures for Proteins and Peptides 57
Enzymatic methods 57
Chemical cleavage methods 61
References 67

3. AUTOMATED SOLID-PHASE MICROSEQUENCING 69
J.B.C.Findlay, D.J.C.Pappin and J.N.Keen

Introduction 69
Advantages 69
Disadvantages 70
Sample Preparation 70
Methodology 71
Preparation of supports 71
Coupling procedures 74
Instrumentation 78
Sequencing Protocols 79
Residue Identification 83
Conversion of ATZ- to PTH-derivatives 83
PTH identification 83
Quantitation 83
Future Prospects 83
References 84

4. GAS- OR PULSED LIQUID-PHASE SEQUENCE ANALYSIS 85
M.J.Geisow and A.Aitken

Introduction 85
Sequencer Mechanics and Chemistry 85
Design of gas–liquid-phase sequencer 86
Sequencer chemistry 87
Sequencer optimization 88
Sample Application, Pre-treatment and Special Programmes 91
Sequence analysis of resin-bound synthetic peptides 92
Solid-phase coupling of peptides 92

Sample de-salting and clean-up 92
Reduction and S-alkylation of proteins 93
OPA blocking 93
Special programmes for optimization of particular amino acids 93
PTH-amino Acid Analysis 94
Separation 94
Phosphoamino Acids 95
Phosphoserine-containing peptides 95
Suppliers 97
Sequencer reagents 97
Associated equipment 97
General chemicals 97
HPLC solvents 97
Sequencing support media 97
Fraction collector tubes 97
Instruments 98
References 98

5. SEQUENCING BY MASS SPECTROMETRY 99
K.Biemann

Introduction 99
Methodology 99
Sample preparation 100
Choice of mass spectrometers 101
Applications 101
Structure problems requiring only molecular weight information 101
Problems that require sequence information 109
Cost effectiveness of tandem mass spectrometry for peptide sequencing 117
Conclusion and Outlook 117
References 117

6. MANUAL METHODS OF PROTEIN SEQUENCING 119
A.Yarwood

Introduction 119
N-Terminal Sequencing Methods: A Perspective 120
DABITC/PITC Double-coupling Method 123
Materials and reagents 123
Liquid-phase manual sequencing 126
Solid-phase manual sequencing 128
Identification of DABTH-amino acid derivatives 130
Dansyl–Edman Sequencing Method 138
Dansylation 138
Edman degradation 138

Chromatography 139
Identification of DNS-amino acids 141
C-Terminal Sequencing Methods 141
Other Methods 143
Reduction and sulphydryl modification 143
Citraconylation, succinylation and maleylation 143
Semi-quantitative determination of the amino acid composition of peptides 144
Acknowledgements 145
References 145

7. STRUCTURE PREDICTION **147**
J.M.Thornton and W.R.Taylor

Introduction 147
Sequence Data Banks and Search Techniques 148
Secondary Structure Predictions 149
Methods based on statistics 150
Methods based on sequence homology 153
Accuracy 153
Prediction of structural class 154
Hydrophobicity Profiles 155
Methods 155
Applications 158
Hydrophobic moment 159
General Sequence Patterns 161
Secondary structure patterns 161
Supersecondary structure patterns 164
Specific Functional Patterns 167
Nucleotide-binding patterns 169
Calcium-binding patterns 176
DNA-binding proteins 178
Cysteine-rich patterns 180
Immunoglobulin domain 182
Specific Family Patterns 183
Remote sequence alignment 184
Building a molecular model 186
Conclusions 186
References 188

APPENDIX **191**

INDEX **195**

CHAPTER 1

Protein and peptide purification

KENNETH J.WILSON and PAU M.YUAN

1. INTRODUCTION

The role of the protein chemist in the characterization of either natural or recombinant peptides or proteins has steadily gained in importance. Whereas some reviewers at the end of the 1970s questioned the need of continuing efforts in structural determination, one now suspects that a shortage of well-trained specialists in this field might well exist.

Molecular biologists currently isolate extremely small amounts of RNA or DNA that code for proteins or enzymes within cells or organisms. Accurate localization of a specific nucleotide sequence frequently depends upon the availability of sequence information. The protein sequence is translated into a nucleotide sequence which is synthesized and subsequently employed in hybridization studies, that is identification of that piece(s) of DNA (or RNA) coding for the desired protein product. The subsequent isolation, cloning, sequencing and expression of the coding sequence then produces the desired protein or peptide product. This simplified overview is the basis of the growing biotechnology industry.

Central to this identification process is the successful isolation and characterization of proteins. This characterization involves not only purification to homogeneity and primary sequence determination, but also the localization, and frequently structural determination, of the positions of post-translational modification. More than 100 different modifications are now recognized. As the repertoire of analytical techniques and instrument capabilities improve, the protein chemist will clearly discover more.

As more complex or poorly expressed peptides and proteins began to be characterized, a serious need for improved isolation and chemical methodologies has developed. For the most part, the proteins considered 'interesting' today are available only in small quantities, often at the microgram level, from sources which are either difficult to obtain in large quantities or prohibitively expensive.

High-performance liquid chromatography (HPLC) and gel electrophoresis are the two techniques which have found extensive use in both the initial isolation and characterization processes of these substances.

To isolate samples in such small amounts, new methodologies and increased detection sensitivities were needed, but not at the expense of significant chemical alteration of the desired sample. Amino-terminal blockage, proteolysis, or amino acid side-chain alteration, for example, the addition or removal of constituents are unacceptable. Consequently, exacting requirements were placed both on the chemicals which directly 'contacted' the sample on the 'wetted' surfaces of instruments, and on the recovery characteristics from surfaces. From these demands, a growing industry for the production

Table 1. Micro-chromatography and gel electrophoresis application areas in protein isolation/characterization.

Purification
Purity estimation
Sample concentration/de-salting
Assessing extent of chemical and enzymatic cleavages
Protein/peptide comparative mapping
Sequencing
Amino acid analysis

Table 2. Physiochemical properties of proteins affecting isolation.

Separation method	*Physical property*				
	Charge	*Hydrophobicity*	*Shape*	*Size*	*Solubility*
Centrifugation			X	X	
Chromatography	X	X	X	X	X
Electrophoresis	X		X	X	
Extraction					X
Gel filtration			X	X	

of high-purity chemicals and 'inert' instrumentation has arisen. *Table 1* lists the general applications in which either HPLC or electrophoresis, or both, prove useful to the protein chemist. These methods have been optimized for analytical and preparative applications on smaller and smaller amounts of material.

The physiochemical properties of samples which enable one to differentiate and thereby separate proteins and peptides are compared in *Table 2*. These properties also form the basis upon which most characterization techniques are dependent. Of the techniques listed in *Table 2*, chromatography is the only one which utilizes each of the different physical properties. The solubility differences, for example, among proteins in the various reagents and solvents used during Edman sequencing prevent them from washing out. Similarly, hydrophobic interaction chromatography relies on the 'salting-in and -out' of proteins in high-versus-low salt concentrations.

Ion-exchange columns separate on the basis of charge; hydrophobic differences are the basis of reverse-phase separations. The size exclusion packings differentiate by molecular shape and size and hydrophobic interaction packings separate by solubility. Electrophoresis achieves its characteristically high resolution through charge and/or size differences. The use of these two separate techniques, HPLC and electrophoresis, either singly or in combination, currently provide the optimal means of isolating small quantities of polypeptides.

Before proceeding to the following narrative, a few definitions are appropriate. 'Micro'-levels are defined as $1-5$ μg for proteins and at the microgram, or less, for peptides. These quantities correspond to the subnanomole level, and frequently the tens of picomoles, for proteins in the 50-kd (or under) molecular weight range. Currently, state-of-the-art instrumentation allows even those researchers with limited experience to determine extended ($20-40$ residues) amino acid sequences on such small amounts. Isolating, as well as handling, such minute quantities requires, however, specialized knowledge.

In the following text, where 'recovered by lyophilization' is mentioned, it implies that a liquid sample has been lyophilized under vacuum (< 100 mTorr) while centrifugation was simultaneously performed. The sample might have been frozen before vacuum introduction, depending on the volume and nature of the solvent. Samples containing relatively concentrated acids, and some organic solvents, require dilution with water prior to freezing.

Whenever possible, liquid samples are collected into polypropylene tubes (Eppendorf™, or the like) and stored frozen at −20°C until use. When samples remain frozen under these conditions oxidation reactions are reduced, sample losses minimized, concentrations remain known (assuming zero evaporation) and re-solubilization is not a potential problem. Sample storage in a dried/lyophilized state is not encouraged. Samples recovered from gels by electroblotting onto the appropriate glass surfaces should be stored at −20°C in an inert atmosphere, argon or nitrogen, minimally 99.998% pure.

2. MICROBORE HPLC INSTRUMENTATION

2.1 **Hardware considerations**

The need to develop techniques for micro-isolation required increased detection sensitivities and then the appropriate sequencing techniques with which to characterize an isolate. The preparation of a few micrograms, rather than tens or hundreds of

Table 3. HPLC hardware considerations in narrow and microbore chromatography.

Pumps	Variable flows, 1 − 1000 μl/min
	Flow-rate resolution, 1 μl/min
	High compositional accuracies (< 2% RSD) and minimum pulsation (< 5 p.s.i.)
	Operation independent of sqlvent viscosity and compressibility
	Pressure limits of 2000 − 5000 p.s.i.
Gradients	Control at the 1-μl level
	Compositional accuracy ± 1%
	Minimized gradient distortion during dynamic mixing
	Matched mixer volume for desired sensitivities
Injectors	Volume selectable through loop changes (μl to ml)
	Zero dead-volume without cross-contamination
	Accuracy and precision maximized
Columns	Commonly stainless-steel with polished, regular inner surfaces
	Zero dead-volume end assemblies
	Stainless-steel frits with minimized absorption characteristics
	Simplified column-to-column change, cartridge design optimal
Connections	Corrosion-resistant, high tolerances on diameters and concentricity
	Appropriately end-cut and cleaned
	Minimized band broadening or extra-column dispersion
	Stainless steel nuts, ferrules and unions; alternatively, finger-tight fittings of Kel-F or another inert material
Solvents	High purity − low background, low levels of trace impurities, minimized residue and particulate content
	Inert or non-reactive with LC components and chromatographed substances
	Appropriate solvent strength, UV cut-off, boiling point and viscosity
	Minimal toxicity and flammability

micrograms, from liquids (cell culture supernatants, urine, fermentation media) or solids (tumours, tissue extracts) is currently achievable within hours, not days or weeks. The cost of the starting material and the time required for work-up have been proportionately reduced. Chromatography, in particular HPLC, is the one technique readily adaptable to these varied prerequisites.

HPLC components are specifically designed and optimized to achieve the necessary performance. The important hardware considerations are listed in *Table 3*.

Because gradient elution and small diameter columns are basic requirements for successful HPLC isolations, the delivery of liquids in precise volumes and compositions is of paramount importance (1,2). Not only must the 'dead' volume of the system be minimized (3,4), but the mixing of the two (or more) requisite solvents/buffers must be efficient (5). An optimized mixing method that filters 'noise' from the pumps at very high sensitivities is required. Injectors for analytical and preparative analysis at micro-levels are most often manually operated. Fraction collection is also performed manually because peak volumes are commonly less than 100 μl. While these operations could be automated, current instrumentation has not been optimized for minimal 'dead' volumes, to maximize the percentage of sample injected, or to collect microlitre amounts into relatively small vials. Thus, neither a 'smart' fraction collector capable of peak detection and isolation nor a miniaturized autosampler is currently available.

2.2 **Column parameters**

The important column variables are listed in *Table 4*. Physical size (inner diameter and length), the support material and its chemically modified surface and the conditions used for chromatography are given (6,7). As will be explained later, microbore (1 mm i.d.) and narrowbore ($\sim$2 mm i.d.) columns of minimal lengths are employed for most applications (8–11).

Peptide and amino acid derivative HPLC is still optimally performed on silica supports. These supports are, however, now often replaced by polymeric materials. Isolations, for example, requiring buffers of higher pH (>8) or minimized denaturation and subsequently higher activity yields, are often conducted on these substitutes.

Column supports with particle sizes of 5–7 μm, 300 Å pore sizes, and either C4 or C8 surface bondings seem equally applicable (12,13). While maintaining reasonable

Table 4. Most commonly used column parameters for protein/peptide HPLC.

Parameters	*Ranges*
Columns size	
Diameter (mm)	1–4.6
Length (mm)	1.5–250
Column supports	
Particle size (μm)	5–20
Pore size (Å)	100–300
Surface bonding	Varies with chromatographic mode
Separation conditions	
Temperature range (°C)	5–40
Flow-rates (μl)	10–2000

Table 5. Most commonly used buffers for protein/peptide HPLC

Reverse-phase	*Acidic pH (2–4)*	*Neutral pH (4–7)*
	0.1% TEA, HFBA or H_3PO_4 5–60% formic acid	10–100 mM NH_4, TFA or sodium acetate 50–100 mM NH_4 or NaH_2PO_4 10–50 mM Tris, NH_4HCO_3 or KH_2PO_4
	Acetonitrile, propanol or some organic mixtures are used as the eluant.	
Ion-exchange	*Buffer (10–50 mM)*	*Salts (up to 1 M)*
anion	Tris, bis-Tris or phosphate	Chloride, acetate or phosphate
cation	Sodium acetate or phosphate	Sodium chloride
Hydrophobic	*Buffer (0.1–0.5 M pH 6–7)*	*Salts (up to 3 M)*
	Ammonium acetate	Ammonium acetate
	Sodium phosphate	Ammonium sulphate
	Potassium phosphate	Sodium sulphate

Abbreviations: HFBA—heptafluorobutyric acid; TFA—trifluoroacetic acid.

operating pressures, these column parameters optimize recoveries, roughly 50 and 80% for proteins and peptides, respectively.

Recently, some construction materials in chromatographic equipment have come under suspicion for chemically altering the samples. The end frits that secure the support material in the column have been reported to absorb proteinaceous material and to release it either slowly or not at all (14–16). Trace components, such as certain metal ions, are also known to leach out of stainless steel frits, tubing and wetted surfaces of the pump under conditions of high salts and/or extremes of pH. Documented effects of these trace compounds on isolates are lacking, but researchers should be aware of the possible phenomenon. Given the mild conditions which are normally employed for an isolation, one should probably first suspect the trace metal contaminants in water, solvents and/or buffers.

2.3 **Buffer systems**

Table 5 describes the buffer combinations employed for the three most useful chromatographic modes. The choice of one or another mode depends upon the desired application and properties of the sample. For example, the use of a low pH buffer system and an organic eluant for the isolation of an enzyme that is denatured under such conditions is ludicrous. Simple incubation experiments in vials will indicate the appropriateness of a chosen set of conditions.

Mass recoveries can only be estimated by chromatography with preliminary studies on small quantities. Frequently these results are misleading because recoveries often improve when larger amounts are chromatographed.

2.4 **Detection enhancement**

To improve detection for an application one may increase detector or recorder sensitivity settings, modify the proteins with more easily detectable reporter groups or simply reduce column inner diameter and flow-rate.

Reducing the column diameter to 2 mm, an adjustment possible with most modern HPLC equipment, and flow by a factor of five increases mass sensitivity by 5-fold.

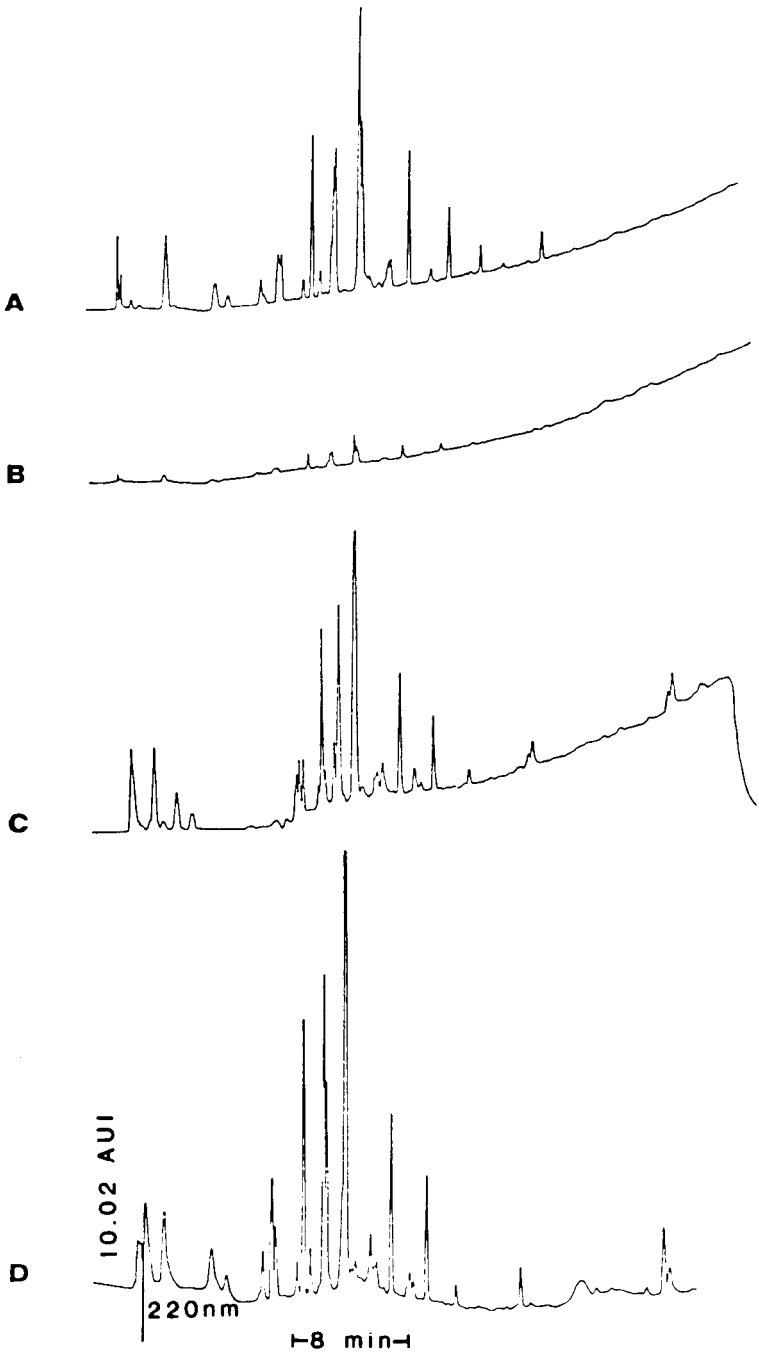


Figure 1. Sensitivity enhancements of narrower column diameters. A tryptic digest of apomyoglobin was chromatographed on a series of reverse-phase columns packed with an identical support (Aquapore^{TM} RP-300) using a 45-min linear gradient from 0% B buffer (0.1% TFA) to 100% B (60% A in 0.1% TFA). **(A)** 4.6 × 250 mm at 1 ml/min and 1000 pmol; **(B)** as in **A** except 100 pmol material injected; **(C)** 2.1 × 220 mm at 200 μl/min and 200 pmol; **(D)** 1 × 250 mm at 50 μl and 100 pmol.