STRUCTURE ELUCIDATION OF BIOPOLYMERS BY MASS SPECTROMETRY

A dissertation submitted for the degree of

Doctor of Philosophy

of the University of London

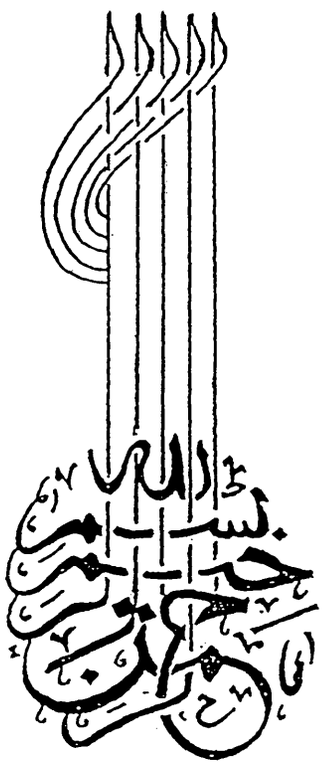
by

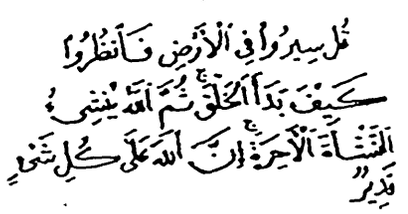
ASHRAF ANNE KUBLAWI CHATTERJEE

Department of Biochemistry

Imperial College of Science Technology and Medicine

November 1989.





A translation: Say " Travel through the earth and see how God did originate creation; so will God produce a later creation: For God has power over all things."

Al-Ankabut, v.20, Al-Quran.

ABSTRACT

A fundamental requirement for understanding the function and activity of biopolymers is a knowledge and accurate definition of their structures, in particular their primary sequences. In cases where data from nucleotide sequencing is available, it is often of limited use, particularly in describing the important post-translational modifications of expressed proteins i.e. whether they are glycosylated, proteolytically processed, covalently modified or disulphide bridged. Therefore, nucleotide sequencing techniques must be complemented by the use of classical and non-classical protein sequencing methods. One such non-classical technology that has contributed greatly to the elucidation of structures of biologically significant compounds, is mass spectrometry.

The work described in this thesis deals predominantly with Fast Atom Bombardment mass spectrometry (FAB-MS), involving methodology development with subsequent application for the solution of structural problems. In combination with established protein chemistry techniques the methods developed represent a significant contribution to modern protein chemistry.

The first methodology development presented in this thesis involves a novel chemical procedure, namely partial **concentrated** acid hydrolysis, combined with a physical method of analysis, FAB-MS. Under controlled conditions total sequencing of peptides of 2000

a.m.u. is possible. This method may also be applied to peptides in excess of 2000 a.m.u. to produce short stretches of sequence up to 10 amino acid residues. The implications for general use in larger molecules are examined and discussed.

Recent technological advances have made linked scanning a viable mass spectrometric method for analysis of small quantities of sample. Presented in this thesis is the first application of linked scanning to direct protein mixture analysis in which the rate-limiting step of purification is removed. Subsequently this technique was applied in an investigation of the primary sequence of the enzyme pyruvate decarboxylase. Linked scan analysis of an **intact** digest mixture of the enzyme revealed a single base difference to that in the nucleotide sequence as well as a previously unrecognised post-translational modification, a blocked N-terminus, the identity of which was characterised.

The methods developed were applied in major studies in the primary sequence analysis of the industrially important enzyme D-xylose isomerase, and in the structure determination of pyruvate decarboxylases.

For My Mother

ACKNOWLEDGEMENTS

I would like to express my sincere graditude to Professor Howard R. Morris for his invaluable supervision, help and advice throughout my three years research in his group. It has been a pleasure to work for Professor Morris with his unfailing sense of humour.

I also would like to express my thanks to Dr. Maria Panico, whom I have had the pleasure to work with upon numerous occasions. I am grateful to Roy McDowell, Dr. Anne Dell, Dr. Jane Thomas-Oates, David Featherby and Tony Etienne for their help and encouragement during this research work and the writing of this thesis. I would also like to extend my thanks to Ian Blench who operated the automated gas-phase sequencer. Thanks is also due to my fellow research students for their tireless support and lively discussions. Sharing a desk with my colleague, Anna Chan, during my research was an experience in itself!

To my friends with their endless supply of patience and understanding I thank you. Finally, to my Mother, Grandmother, Helen, Andrew and Kamie I am eternally bound in gratitude for your steadfast support, encouragement and love.

This work was supported by a Medical Research Council studentship.

CONTENTS

PAGE

Acknowledgments 6

List of Figures 9

List of Tables 15

Useful Data 18

CHAPTER ONE: Introduction 19

CHAPTER TWO: FAB-Mapping Techniques Applied to the Primary Sequence Analysis of D-Xylose Isomerase, *Arthrobacter* Strain B3728. 51

2.1 Introduction 51

2.2 Results and Discussion 56

PART I

2.2.1 Section A: Primary Sequence Analysis 56

2.2.2 Section B: Verification of the Gene Sequence. 80

PART II

2.3 Section C: Structure-Activity Relationship of D-Xylose Isomerase, *Arthrobacter* Strain B3728. 94

2.3.1 Introduction 94

2.3.2 Results and Discussion 96

2.4 Conclusion 103

CHAPTER THREE: Methodology Development I-Partial Acid Hydrolysis for Peptide Sequencing. 105

3.1 Introduction 105

3.2 Results and Discussion 109

3.2.1 Optimum Conditions of Hydrolysis for Peptide Sequencing 110

3.2.2 Investigation of Peptide Fragment Formation 117

3.2.3 Analysis of the Peptide Components of Acid Hydrolysates 126

3.2.4 Rationalising Suppression of Peptide Components in a Mixture, by FAB-MS Analysis, in Terms of Hydrophobicity 127

3.2.5 Recent Applications of Partial Acid Hydrolysis 140

3.2.6 Spectral Patterns 146

3.2.7 Partial Acid Hydrolysis of Disulphide Bridge Containing Peptides 146

3.2.8 Application of Partial Acid Hydrolysis to Large Peptides 147

3.3 Conclusion 156

		PAGE
CHAPTER FOUR:	Methodology Development II-Linked Scan Analyis of Peptides and Proteins; Direct Mixture Analysis	158
	4.1 Introduction	158
	4.2 Results and Discussion	166
	PART I - Linked Scanning of Pure Peptides	167
	PART II - Direct Mixture Analysis	177
	4.4 Variation of CAD Conditions	191
	4.5 Conclusion	194
CHAPTER FIVE:	Analysis of the Primary Sequence of Pyruvate Decarboxylase, Utilising a Number of Mass Spectrometric Techniques	198
	5.1 Introduction	198
	5.2 Results and Discussion	204
	5.2.1 FAB-Mapping	205
	5.2.2 B/E Linked Scan/CAD	231
	5.2.3 High Mass MS	234
	5.2.4 Electrospray Ionisation Mass Spectrometry	250
	5.3 Conclusion	262
CHAPTER SIX:	Overview	266
CHAPTER SEVEN:	Methods and Materials	270
REFERENCES		279

LIST OF FIGURES

		PAGE
Figure 1.1	Edman Degradation.	22
Figure 1.2	Cyanogen Bromide reaction.	25
Figure 1.3	Schematic Representation of a Reverse Geometry, Two-sector Mass Spectrometer (ZAB-HF).	29
Figure 1.4	Schematic Diagram of a PMA 3000 Electrospray Mass Spectrometer.	49
Figure 2.1	CNBr/tryptic FAB-Map of D-Xylose Isomerase.	59
Figure 2.1.1	FAB-Map of CNBr/tryptic Digest of D-Xylose Isomerase after ONE step of Manual Edman Degradation of the Intact Digest Mixture.	61
Figure 2.1.2	FAB-Map of CNBr/tryptic Digest of D-Xylose Isomerase after TWO steps of Manual Edman Degradation of the Intact Digest Mixture.	62
Figure 2.1.3	CNBr/tryptic/CARBOXYPEPTIDASE B FAB-Map of D-Xylose Isomerase.	64
Figure 2.2	Sequence (A) is the MS/Electron Density, Primary Sequence of D-Xylose Isomerase; Sequence (B) is the Initial Electron Density Map of D-Xylose Isomerase.	78
Figure 2.3	The Gene-derived Sequence of D-Xylose Isomerase *Arthrobacter* strain B3728.	81
Figure 2.4	Sequence Alignment Between Various Bacterial Strains of D-Xylose Isomerase.	82
Figure 2.5	D-Xylose Isomerase Subunit Structure.	95

		PAGE
Figure 2.6	HPLC Chromatograph of a CNBr Digest Mixture of the Thermolytic Product (37.5 kDa species) of D-Xylose Isomerase.	100
Figure 2.6.1	HPLC Chromatograph of a CNBr Digest Mixture of Native D-Xylose Isomerase.	101
Figure 3.1	FAB-MS of Acid Hydrolysed Neurotensin (500 pico -moles).	112
Figure 3.1.1	The Primary Sequence of Neurotensin, $[M+H]^+$ 1672.	113
Figure 3.2	FAB-MS of Acid Hydrolysed human Calcitonin Gene Related Peptide (hCGRP), 500 picomoles.	115
Figure 3.2.1	The Primary Sequence of human Calcitonin Gene Related Peptide (hCGRP), $[M+H]^+$ 3787.	116
Figure 3.3	FAB-MS of Native Neurotensin (500 picomoles).	119
Figure 3.4	FAB-MS of Acid Hydrolysed alpha-Malanocyte Stimulating Hormone (α-MSH).	123
Figure 3.4.1	FAB-MS of Acid Hydrolysed and <u>then</u> Acetylated alpha-Melanocyte Stimulating Hormone (α-MSH).	124
Figure 3.4.2	The Primary Sequence of alpha-Melanocyte Stimulating Hormone (α-MSH).	125
Figure 3.5	Primary Sequence of Substance P, $[M+H]^+$ 1347.	132
Figure 3.6	Primary Sequence of Corticotropin-like Intermediate-lobe peptide (CLIP), $[M+H]^+$ 2490.	135
Figure 3.7	The Primary Sequence of beta-Endorphin, $[M+H]^+$ 3461.	137
Figure 3.8	Signals m/z 1042 and m/z 1156 Enlarged so as to Illustrate the Change in Cluster Pattern on	144

PAGE

Partial Deamidation.

Figure 3.9 N→O acyl Transformation, Corresponding to Cleavage at the N-terminus of Serine Residues. 145

Figure 3.10 FAB-MS of Acid Hydrolysed Growth Hormone Releasing Factor 1-44 (GHRF), 500 picomoles. 150

Figure 3.10.1 Primary Sequence of Growth Hormone Releasing Factor (GHRF) 1-44. 153

Figure 3.11 Primary Sequence of Bovine Insulin. 154

Figure 4.1 Two-sector B/E Linked Scan/CAD (He) Spectrum of Leu-enkephalin amide. 168

Figure 4.1.1 FAB-MS of Leu-enkephalin amide. 169

Figure 4.2 Two-sector B/E Linked Scan/CAD (He) Spectrum of Angiotensin II. 171

Figure 4.3 Two-sector B/E Linked Scan/CAD Spectrum of Renin substrate (ZAB2E). 173

Figure 4.3.1 Four-sector CAD Spectrum of Renin substrate (ZABSE-4F). 174

Figure 4.4 Tryptic FAB-Map of Pyruvate Decarboxylase (A1). 179

Figure 4.5 Two-sector B/E Linked Scan/CAD (He) Spectrum of the Component m/z 1046 (Angiotensin II) in the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase. 181

Figure 4.6 Two-sector B/E Linked Scan/CAD (He) Spectrum of the Unknown Peptide Component m/z 789 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase. 183

Figure 4.7 Two-sector B/E Linked Scan/CAD (He) Spectrum of 187

		PAGE
	the Unknown Peptide Component m/z 1016 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase	
Figure 4.8	Two-sector B/E Linked Scan/CAD (He) Spectrum of the m/z 1778 Component of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.	189
Figure 4.9	Two-sector B/E Linked Scan Spectrum of the Unknown Component m/z 1016 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.	192
Figure 4.10	Two-sector B/E Linked Scan/CAD (air) Spectrum of the m/z 1778 Component of the Tryptic Digest Mixture of Pyruvate Decarboxylase.	193
Figure 5.0	Terminal Steps in Alcoholic Fermentation.	199
Figure 5.1	Tryptic FAB-Map of PDC subunit A1.	207
Figure 5.1.1	Tryptic FAB-Map of PDC subunit A2	208
Figure 5.1.2	Tryptic FAB-Map of PDC subunit B.	209
Figure 5.1.3	Nucleotide and Predicted Amino Acid Sequence of Pyruvate Decarboxylase.	210
Figure 5.1.4	Chymotryptic FAB-Map of PDC subunit A1.	221
Figure 5.1.5	Amino Acid Residues able to be Assigned for PDCA1 from the FAB-Mapping Experiments.	224
Figure 5.1.6	Amino Acid Residues able to be Assigned for PDCA2 from the FAB-Mapping Experiments.	225
Figure 5.1.7	Amino Acid Residues able to be Assigned for for PDCBfrom the FAB-Mapping Experiments.	226
Figure 5.2	FAB-MS of CNBr Digest of PDCA1 between Mass Range 300-2000 Da (ZAB-2SE).	236

PAGE

Figure 5.2.1 FAB-MS of CNBr Digest of PDCA1 between Mass Range 1000-3500 Da (ZAB-2SE). 237

Figure 5.2.2 FAB-MS of CNBr Digest of PDCA1 between Mass Range 3500-6000 Da (ZAB-2SE). 239

Figure 5.2.3 FAB-MS of CNBr Digest of PDCA1 between Mass Range 5000-10,000 Da (ZAB-2SE). 241

Figure 5.2.4 FAB-MS of Carboxymethylated (20% eluate) and then Digested (CNBr) PDCA1 between Mass Range 300-2000 Da. 243

Figure 5.2.5 FAB-MS of Carboxymethylated (20% eluate) and then Digested (CNBr) PDCA1 between Mass Range 1800-6000 Da. 245

Figure 5.2.6 FAB-MS of Carboxymethylated (40% eluate) and then Digested (CNBr) PDCA1 between Mass Range 5000-10,000 Da. 247

Figure 5.3 ESI Mass Spectrum of PDCA2 Carrying Multiple Positive Charges. 252

Figure 5.3.1 ESI-MS of PDCA1, Illustrating an Expanded Region of its Bell-shape Distribution. 254

Figure 5.3.2 ESI-MS of PDCA2, Illustrating an Expanded Region of its Bell-shape Distribution. 255

Figure 5.3.3 ESI-MS of PDCB, Illustrating an Expanded Region of its Bell-shape Distribution. 256

Figure 5.3.4 ESI-MS Containing the centroids of Each Molecular Ion Envelope for PDCA1 Observed in Fig. 5.3.1. 257

Figure 5.3.5 ESI-MS Containing the Centroids of Each Molecular Ion Envelope for PDCA2 Observed in Fig. 5.3.2. 258

PAGE

Figure 5.3.6 ESI-MS Containing the Centroids of Each Molecular Ion Envelope for PDCB observed in Fig. 5.3.3. 259

LIST OF TABLES

		PAGE
Table 2.1	Intense Quasi-molecular ions observed in; (i) CNBr/tryptic FAB-Map, (ii) CNBr/tryptic FAB-Map after ONE step of Edman Degradation, (iii) CNBr/ tryptic FAB-Map after TWO steps of Edman Degradation, and (iv) CNBr/tryptic/CpB FAB-Map of D-Xylose Isomerase.	62
Table 2.2	Intense Quasi-molecular ions observed in the FAB-Map of a tryptic/CNBr Digest of D-Xylose Isomerase.	84
Table 2.3	Mass Signals Observed in the CNBr/tryptic FAB-Map able to be assigned on to the Gene-derived Sequence.	85
Table 2.4	Intense Quasi-molecular Ions Observed in the tryptic/CNBr FAB-Map which may be Assigned on to the Gene-derived Sequence.	85
Table 2.4.1	Combined FAB-Mapping Data from both Digests for which Residue Assignment on to the Gene-derived Sequence could be made.	90
Table 2.5	A Comparison of Amino Acid Compositions of Different Bacterial Strains of D-Xylose Isomerase.	92
Table 2.6	CNBr Peptides Fragments Predicted to be Generated by CNBr Cleavage of D-Xylose Isomerase.	98
Table 3.1	Hydrophobicity Calculations for Peptide Fragments Identified, by FAB-MS, to be Present in the	128

		PAGE
	Purified Hydrolysate Mixture of Neurotensin.	
Table 3.2	Hydrophobicity Indices for the Peptide Observed in the FAB Spectrum of Acid Hydrolysed Substance P.	133
Table 3.3	Hydrophobicity Indices for (a) the Observed C- to N-terminal Peptide Fragments Interpreted from the FAB-MS Spectrum of Acid Hydrolysed ∝-MSH, and (b) the Reverse Amino Acid Sequence.	133
Table 3.4	Hydrophobicity Indices of Peptide Fragments Identified in the FAB-MS of the Acid Hydrolysate of CLIP.	138
Table 3.5	Hydrophobicity Indices of Peptide Fragments Identified in the FAB-MS of -Endorphin.	138
Table 3.6	Hydrophobicity Indices of Peptide Fragments Identified in the FAB-MS of Acid Hydrolysed hCGRP.	139
Table 3.7	Hydrophobicity Indices of Peptide Fragments Identified in the FAB-MS of Acid Hydrolysed GHRF, 1-44.	155
Table 3.8	Hydrophobicity Indices of Peptide Fragments Identified in the FAB-MS of Acid Hydrolysed Bovine Insulin.	155
Table 5.1	Quasi-molecular Ions Observed in the Tryptic FAB-Maps of Subunit A1, A2 and B, together with the Peptide Assignments on to the Gene-derived Sequence of PDC.	212
Table 5.1.1	Those Signals in the Tryptic FAB-Map Spectra	213

		PAGE
	which were unable to be Assigned on to the Gene -derived Sequence.	
Table 5.1.2	Additional Interpretable Intense Signals Observed in the Tryptic/V8 FAB-Maps of Subunit A1, A2 and B.	218
Table 5.1.3	Intense Signals Observed in the Chymotryptic FAB-Map of Subunit A1, together with Signal Assignments on to the Gene-derived Sequence.	218
Table 5.1.4	Peptide Fragments Identified after HPLC of the Chymotryptic Digest of A1, by FAB-MS, with Residue Assignment on to the Gene-derived Sequence.	222
Table 5.1.5	The Predicted Peptide Fragments, Based on the Gene Sequence, after Cyanogen Bromide Cleavage.	228
Table 5.2	High Mass FAB-MS Analysis of CNBr Digest Mixtures of Oxidised (native) and Carboxymethylated Samples of PDC A1.	248
Table 5.3	Molecular Weight Calculations Determined from Electrospray Ionisation Mass Spectra.	260

TABLE OF ABBREVIATIONS

MS	Mass Spectrometry
FAB	Fast Atom Bombardment
EI	Electron Impact
CI	Chemical Ionisation
FD	Field Desorption
PD	Plasma Desorption
ESI	Electrospray Ionisation
MIKES	Mass-Analysed Ion Kinetic Energy Spectroscopy
CAD	Collisionally Activated Decomposition
HF	High Field
Da	Dalton
a.m.u.	atomic mass unit
HPLC	High Performance Liquid Chromatography
MNBA	Mono-nitrobenzyl alcohol
PITC	Phenylisothiocyanate
PTH	Phenylthiohydantoin
CNBr	Cyanogen Bromide
CpB	Carboxypeptidase B
PDC	Pyruvate Decarboxylase
∝-MSH	alpha-Melanocyte Stimulating Hormone
CLIP	Corticotropin-like Intermediate-lobe Peptide
GHRF	Growth Hormone Releasing Factor
TFA	Trifluoroacetic acid
B	Magnetic Field Strength
E	Electric Sector Voltage
V	Accelerating Voltage

USEFUL DATA

1. The Masses of the Commonly Occuring Amino Acid Residues

AMINO ACID	SYMBOL		Residue Mass (a.m.u.)	Average Mass (a.m.u)
Alanine	Ala	A	71	71.0788
Arginine	Arg	R	156	156.1875
Asparagine	Asn	N	114	114.1038
Aspartic Acid	Asp	D	115	115.0886
Cysteine	Cys	C	103	103.1388
Glutamic Acid	Glu	E	129	129.1155
Glutamine	Gln	Q	128	128.1307
Glycine	Gly	G	57	57.0519
Histidine	His	H	137	137.1411
Isoleucine	Ile	I	113	113.1594
Leucine	Leu	L	113	113.1594
Lysine	Lys	K	128	128.1741
Methionine	Met	M	131	131.1926
Phenylalanine	Phe	F	147	147.1766
Proline	Pro	P	97	97.1167
Serine	Ser	S	87	87.0782
Threonine	Thr	T	101	101.1051
Tryptophan	Trp	W	186	186.2132
Tyrosine	Tyr	Y	163	163.1760
Valine	Val	V	99	99.1326

2. FAB-MS matrices

Glycerol Peaks mass (a.m.u.)	MNBA mass (a.m.u.)
93	154
185	307
277	460
365	613
461	766
553	919
645	
737	
829	

CHAPTER ONE

INTRODUCTION

A knowledge of the fundamental chemical and physical properties of cellular constituents is required for understanding fully the life processes of the body. Proteins participate in such diverse and necessary processes such as enzymatic catalysis, transport and storage, co-ordinated motions, mechanical support, immunological defence, generation and transmission of nerve impulses, and control of growth and differentiation. It was during the early research work of the last century, with attempts to unravel the mechanisms of specific physiological processes, that it was realised that proteins are ubiquitous throughout nature and essential to life itself. It is highly appropriate therefore that the name protein means "first" or "foremost". This name was designated by G.J. Mulder in 1838, who made the observation:

> "There is present in plants and animals a substance which is without doubt the most important of the known substances in living matter, and without it, life would be impossible on our planet. This material has been named Protein."

The great diversity in protein function arises from diversity in protein structure. The key to the structure of the thousands of different proteins is the group of relatively simple building blocks, amino acids, from which proteins are built. All proteins

from those of the most simple prokaryotic cells, to those of the highest form of life are constructed from the same set of twenty amino acids covalently linked in characteristic sequences. In addition to these twenty amino acids which are commonly found in proteins there are some which are less frequent, such as hydroxyproline, hydroxylysine and more complicated structures such as desmosine, a derivative of lysine, found in the fibrous protein elastin. However, it is the number and sequential order of these amino acids which lends a protein its chemical and biological individuality, with each protein posesessing its own combination of amino acids. The importance of sequence to function can be illustrated by a number of examples in which a change in the amino acid sequence of a protein, of possibly just one residue, results in partial or total loss of activity in the biological system in which it is acting. The molecular recognition of the cause of sickle-cell anaemia, a human hereditary blood disease, was one of the most successful scientific pioneering efforts, and which alerted the scientific world to the genetic origin of many diseases. It was found that the cause of sickle-cell anaemia lies with genetically altered haemoglobin molecules. This genetic defect in sickle-cell haemoglobin is the result of a mutation in the beta-chain of the haemoglobin, in which the glutamic acid residue present in position 6 of the normal haemoglobin beta-chain, is replaced by a valine residue. This single mutation to the beta-chain gene results in clinical symptoms such as shortness of breath and heart murmurs brought on by physical exertion. Similarly, recent research into connective tissue diseases has identified a single base mutation of the alpha-1 chain of type I

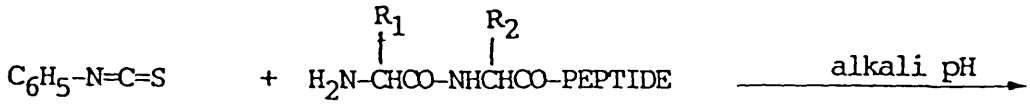
procollagen, that replaced glycine with cysteine [1]. This single amino acid change has been found in a lethal variant of *osteogenesis imperfecta*, a hereditable connective tissue disorder.

1.1 Classical Protein Sequencing Methods

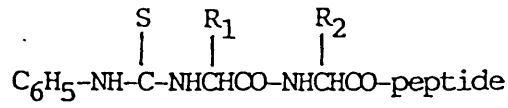
With the realisation that it is the amino acid sequence that identifies a protein and distinguishes it from all other proteins a concerted effort was made for the development of sequence analysis methods. In 1950 Edman [2] introduced a method for the sequential degradation of polypeptides from the N-terminus, which probably still represents one of the single most important classical sequencing techniques developed. The degradation is a cyclic procedure, by which amino acid residues are cleaved one at a time from the N-terminus of the peptide and identified as phenylthiohydantoin derivatives. There are three steps in each cycle, namely, coupling of phenylisothiocyanate (PITC) to the amino-terminal residue, which forms a phenylthiocarbamyl derivative, cleavage of the amino-terminal residue via cyclisation to the anilinothiazolinone and conversion of the thiazolinone derivative formed to the more stable phenylthiohydantoin (PTH) derivative, which may be identified chromatographically. The Edman chemistry is summarised in Figure 1.1. For the coupling of PITC to proceed, the N-terminal amino acid must bear a free primary amine.

However, it was found that the manual Edman degradation method had a number of limiting factors associated with it such as, the number of repetitive cycles permissible was restricted; identification of these limitations initiated the development of an automated

1st Step: Coupling

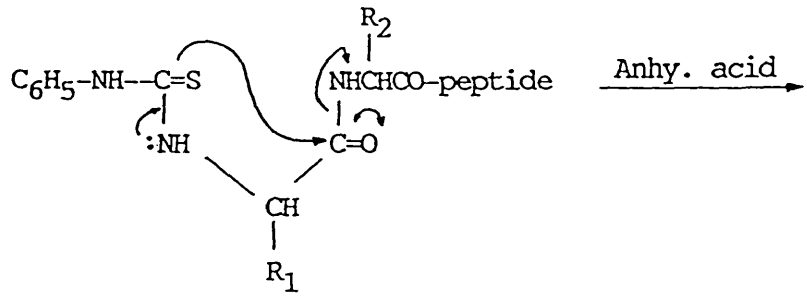


Phenylisothiocyanate

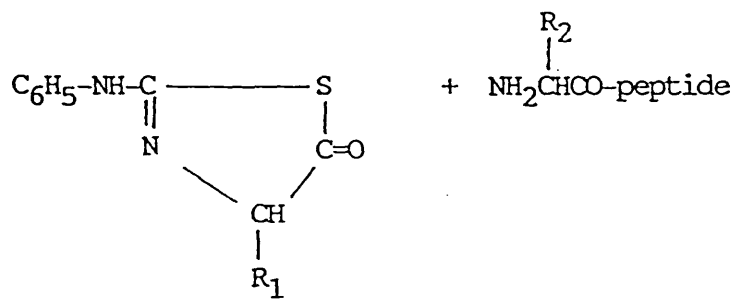


Phenylthiocarbamyl peptide (PTC-peptide)

2nd Step: Cyclisation



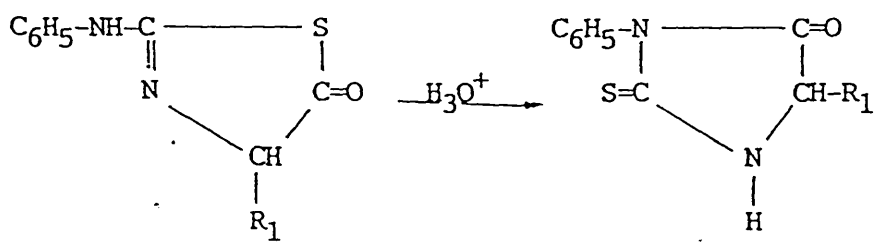
PTC-peptide



Anilino-thiazolinone derivative (ATZ)

Peptide residue

3rd Step: Conversion



ATZ-amino acid

Phenylthiohydantoin A.A. (PTH)

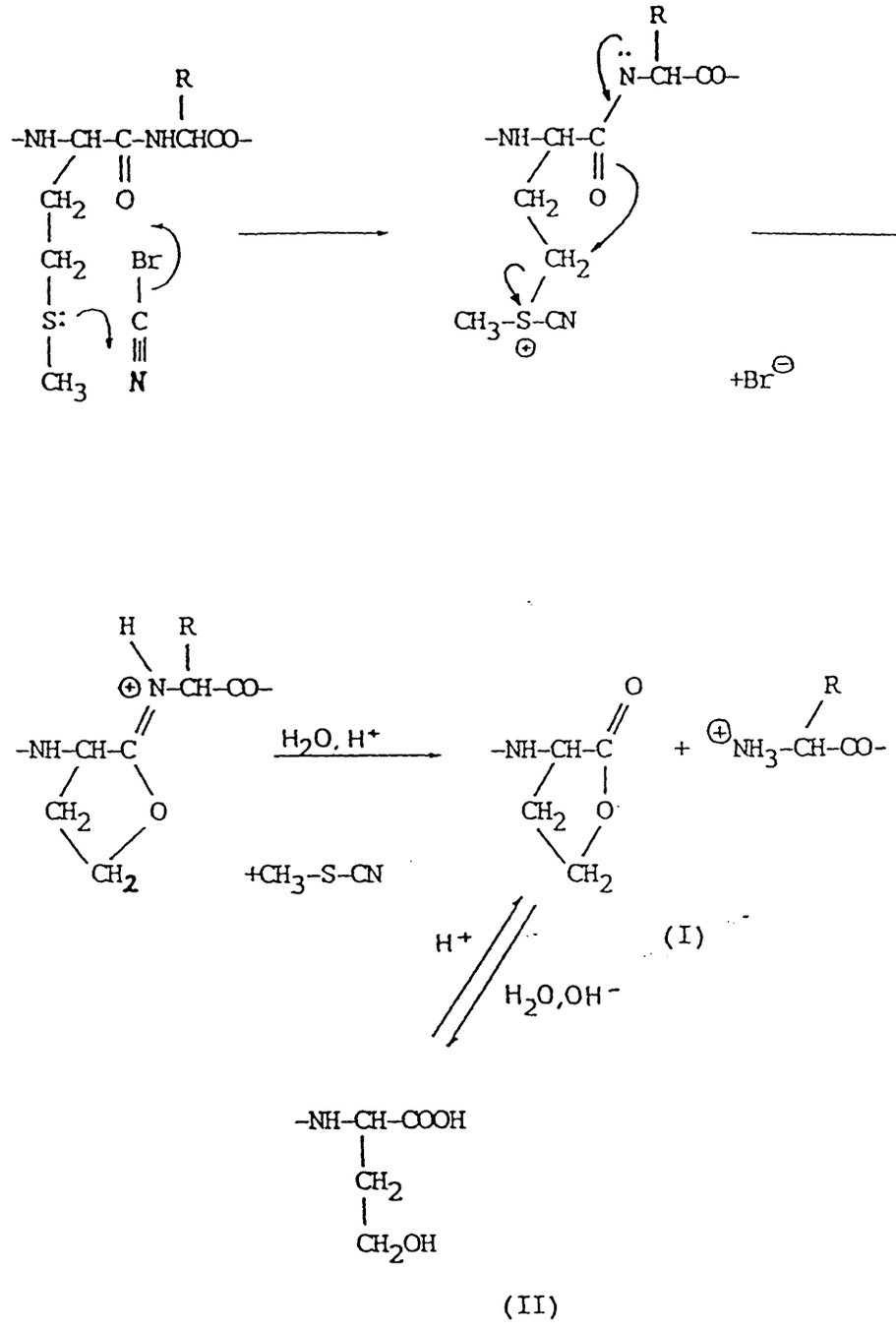
Figure 1.1 EDMAN DEGRADATION

instrument. In 1967, Edman and Begg [3] published a description of an automated protein sequenator, capable of carrying out degradations of up to sixty cycles in a single run. This achievement revolutionised protein sequencing, since the Edman degradation could be repeated with high efficiency. The number of cycles that can be accomplished is limited by the efficiency of coupling, approximately 90%, by the purity of the reagents used and the PTH detection methods available. Since 1967 instrument design has improved greatly such that modern protein sequencers allow sequence determination from as little as 10 picomoles of material. The technical problems found manually such as, difficulties of maintaining an effective barrier against atmospheric oxygen and performing effective extractions are overcome in the automated gas-phase sequencer. The reaction is carried out in an atmosphere of argon, with reagents transported to the reaction vessel in a vapour (the sample is loaded on a glass fibre disc), however, the PITC is delivered as a liquid. Also the anhydrous trifluoroacetic acid though transported as a vapour condenses on arrival, hence the nomenclature to a certain extent is misleading with reactions taking place in the liquid phase. However, all manipulations are precise and exact allowing high sensitivity with efficient removal of by-products. By-products are either flushed or evaporated away allowing "clean" products to be isolated. One of the main disadvantages of the manual Edman degradation method is the build up of by-products. However, even with these conditions satisfied it is not possible to indefinitely sequence a large protein by this method because of less than perfect repetitive yields. Hence other techniques are required to complement Edman chemistry, such as

those which cleave a large protein into small peptide fragments, which after purification are amenable to gas-phase sequencing.

1.2 Specific Cleavages of Proteins

Since the early 1960s there has been a dramatic increase in the methods available for cleaving a peptide chain, especially at specific amino acid residues. When a particular amino acid occurs infrequently in a peptide chain, specific cleavage at these residues results in a comparatively small number of fragments, which can be readily separated and purified. A large number of chemical treatments has been developed over the years, however, this introduction will be restricted to the methods employed in the investigations presented in the chapters following. A frequently used chemical procedure is the cyanogen bromide (CNBr) reaction of Gross and Witkop [4] for specific cleavage of methionine residues. Methionine residues, in a protein, are selectively converted to homoserine (and/or its lactone), with simultaneous cleavage of the peptide bond on the carboxyl side of the methionine, see Figure 1.2. In addition to the chemical methods, Edman degradation and CNBr cleavage, discussed there are many enzymes which exhibit a high specificity for peptide bond cleavage. Two such enzymes, which have been extensively employed in protein sequence studies in this thesis are trypsin and alpha-chymotrypsin. Both these enzymes are serine endopeptidases, i.e. serine is found at the active site. Sequence homology is observed at the active site of enzymes of this class. Alpha-chymotrypsin is relatively specific for peptide bond cleavage on the carboxyl side of amino acids, tryptophan, tyrosine, phenylalanine and leucine, which possess aromatic and hydrophobic



Methionine residues are converted to a mixture of C-terminal homoserine residues (II) and homoserine lactone residues (I).

Figure 1.2 CYANOGEN BROMIDE reaction.

bulky side-chains; peptide bond cleavage is also occasionally observed at other sites but with a slower rate of hydrolysis. Trypsin is more specific than alpha-chymotrypsin, with cleavage on the C-terminal side of lysine and arginine peptide bonds only. Another enzyme with useful bond cleavage specificity is *Staphylococcal protease*, (from the V8 strain), an extracellular protease from *Staphylococcus aureus*. This enzyme is specific for cleaving glutamyl peptide bonds, and under the correct conditions aspartyl bonds also. There are a variety of enzymes employed for protein sequencing with relatively little bond specificity e.g. thermolysin and pepsin. However, the enzymes widely employed in this thesis reasearch work have been trypsin, alpha-chymotrpsin and V8 protease, chosen for their specificity, since when dealing with proteins of unknown sequence, random peptide cleavage serves only to complicate the sequence analysis.

The chemical and enzymic digestions discussed deal with either N-terminal amino acid determination or peptide bond cleavage of a polypeptide chain into fragments. The methods available for C-terminal analysis are less satisfactory than those for N-terminal residue determination. One of the most conv enient methods for C-terminal analysis is by the use of carboxypeptidases. An example of such an enzyme is carboxypeptidase B which is highly specific for C-terminal cleavage of lysine and arginine residues only. This enzyme is ideal for use in conjunction with trypsin in certain strategies (see later), and has been widely implemented in studies reported upon in this thesis.

1.3 Mass Spectrometry

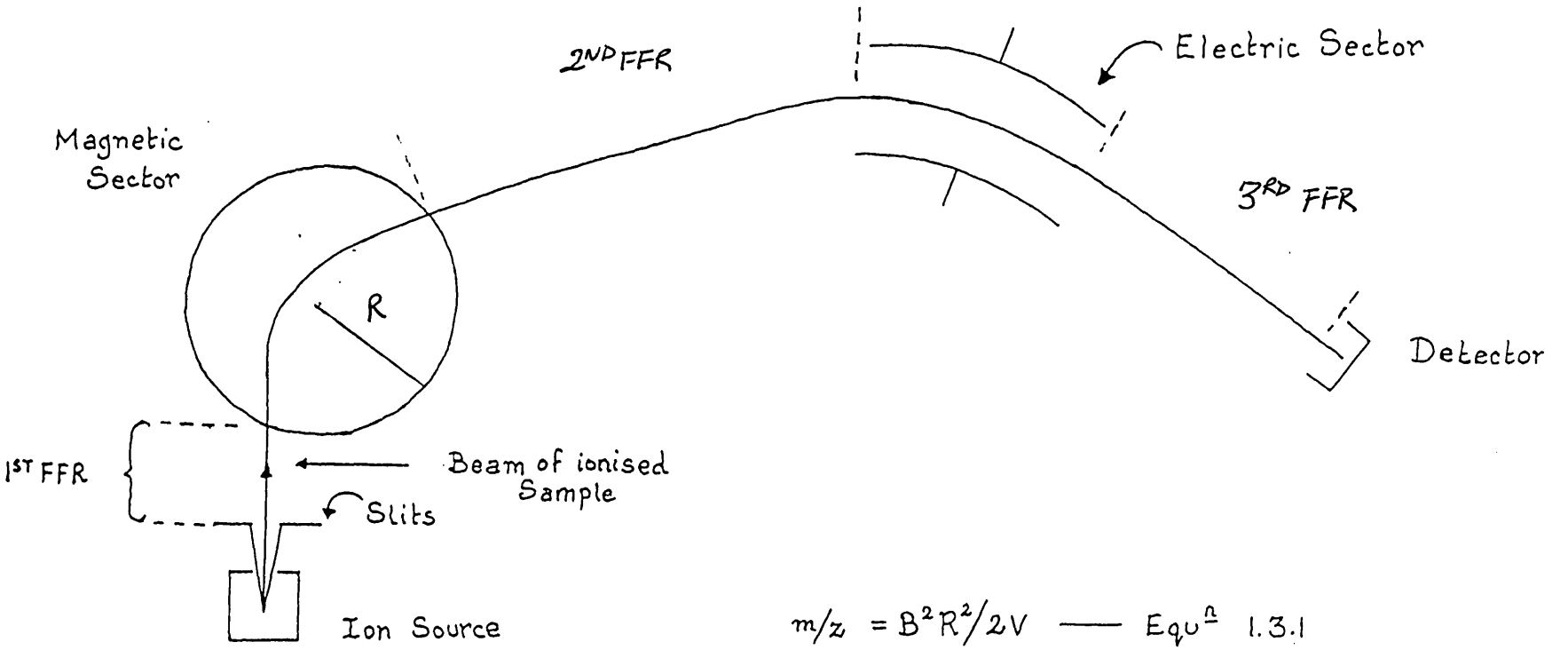
1.3.1 The Historical Role of Mass Spectrometry in Sequence Analysis

Mass spectrometry is a technique which is gaining acceptance in the field of micro-protein sequencing studies, sometimes as the method of choice. As well as playing an important role in sequence analysis investigations, other highly significant characteristics of mass spectrometry lie in its a) sensitivity, b) accurate molecular weight determination, c) mixture handling capabilities and d) detection and characterisation of post-translational modifications. It is now some thirty years since studies on small synthetic peptides indicated the possibility of sequence analysis by mass spectrometry [5,6] and twenty years since the first successful analysis of natural protein-derived peptides and peptide mixtures was demonstrated using N-acetyl N,O permethyl derivatives [7,8]. Mass spectrometry has thus played an important role in biopolymer analysis for more than two decades and has been demonstrated to be a valuable complementary technique to classical protein sequencing methods, such as Edman degradation. In addition, it is the primary method for the analysis of post-translational modifications, for example, glycosylation [9], phosphorylation [10] and carboxylation [11,12], where classical methods cannot give data. The power of mass spectrometry in identifying and determining the presence of post-translational modifications is illustrated by the first characterisation of a hormone of insect neuroendocrine origin. The structure of locust adipokinetic hormone, an N-terminally blocked decapeptide, which could not be sequenced classically was determined [7] by mass spectrometry. Mass

spectrometric methods have been instrumental in many discoveries of biological importance. Examples in which mass spectrometry has played a crucial role are the discovery of the γ-carboxylation of glutamic acid in the blood coagulation zymogens [11,12] and in the characterisation of leukotriene D_4 [13], a compound of great biological significance, associated with a variety of allergen-induced hypersensitivity reactions, for example asthma. However, there are a numerous strategies and methods by which mass spectrometry maybe applied to a problem, and not all possible combinations of pre-treatment, mass spectrometric type, ionisation methods etc., will solve particular problems. The development of sensible strategies for solving different types of problem is a key feature of current research in mass spectrometry applied to protein chemistry, and several chapters in this thesis will concentrate on these objectives.

1.3.2 **Mass Spectrometric Analysis**

A molecule is analysed mass spectrometrically by using some means, such as a beam of electrons, to ionise it and impart internal energy, which leads to fragmentation. The way in which the molecule breaks up is diagnostic of its original structure. Mass spectrometry, (MS), is the production of ions from neutral compounds, by a variety of different ionisation methods e.g. Chemical Ionisation (CI), Electron Impact (EI), and Fast Atom Bombardment (FAB). A mass spectrometer is designed to separate and measure the masses of ions produced, by their mass-to-charge ratio. In Figure 1.3 a reverse-geometry double focussing mass spectrometer is depicted. When a molecule is ionised a molecular ion, $M^{\cdot+}$, is



$$m/z = B^2R^2/2V \quad \text{—— Equ}^{n} \ 1.3.1$$

Figure 1.3 Schematic Representation of a Reverse-Geometry, Two Sector Mass Spectrometer (ZAB-HF)

produced which may contain sufficient internal energy to fragment by the ejection of a neutral particle, N, with the formation of fragment ion A^+. The ejected neutral particle maybe a radical or a neutral i.e.

$$M \xrightarrow{-e} M^{+\cdot} \longrightarrow A^+ + N$$

Subsequently if A^+ has sufficient internal energy then further decomposition may occur with the formation of B^+, C^+....., until there is not sufficient internal energy in any one ion for further reaction. This stepwise decomposition is a fragmentation pathway, which often allows the original structure of the compound to be determined. The extent of fragmentation is therefore determined by the amount of internal energy originally imparted to the molecular ion and by the time allowed between ion formation and detection. Hence, the mass spectrum represents the probabilities of any particular fragment ion being formed. Any neutrals or radicals produced are not recorded mass spectrometrically since only charged species are deflected in a magnetic field. The equation of motion of an ion passing through a mass spectrometer, with both magnetic and electric sectors, equation 1.3.1, is included in Figure 1.3.

After the beam of ions has been accelerated through a potential of 4-8 keV away from the source it is divergent and slits are used to reduce the spread. In a reverse-geometry instrument, as shown in Figure 1.3, the ion beam on leaving the source passes through a magnetic sector, which separates the ions according to their mass/charge ratio. On passing through the magnetic field the ions

experience a centripetal force from the magnet (balanced by their own centrifugal force) and follow a radius of curvature according to their mass/charge relationship. The focussed ion beam then enters an electric sector which may simply be regarded as an energy focussing device. This energy focussed beam then falls at the collector. By considering equation 1.3.1 it can be seen that a spectrum may be generated by scanning either the magnet or the accelerating voltage. Magnetic scanning is preferable over varying the accelerating voltage, V, (with the magnetic field strength, B, constant) as the sensitivity of the ion detection decreases with the decreasing accelerating voltage and sensitivity is a key factor of any methodology to be applied to biological samples. Any fragment ion formed outside the ion source will possess less kinetic energy than originally imparted and thus, will not be focussed by the electric sector and hence will not be detected. This type of fragment ion is termed "metastable" and numerous techniques have been devised to detect these daughter ions, such as linked scanning, which is discussed below.

1.3.3 **Methods of Sample Ionisation**

When discussing ionisation techniques it is important to distinguish between "hard" and "soft" ionisation methods. These two techniques differ by the amount of internal energy which is imparted to the sample. Up until the late 1970s/early 1980s electron impact (EI) mass spectrometry, a hard ionisation technique, was the most widely used mass spectrometric method for sequence analysis. Electron impact mass spectrometry has been instrumental in defining the structures of a number of biologically

significant compounds such as, Leu- and Met-enkephalin [14] and 1,25 dihydroxycholecalciferol [15]. This ionisation technique involves the production of a stream of electrons from a heated filament in vacuo, which is accelerated through a voltage, V, across an ion chamber, so that the electrons have an energy, eV, where "e" is the electronic charge. A volatilised sample, M, and an electron of energy greater than the ionisation energy of M will react if they pass close enough for the electron to impart sufficient of its energy to the neutral species. Two main sequence analysis strategies, utilising EI-MS, were developed;

(i) the direct probe analysis method [16], and

(ii) gas chromatography/mass spectrometry, (g.c./ms), [17,18,19].

During this period, the concept of analysing mixtures of peptides generated from protein/glycoprotein digests was first introduced utilising the direct probe method [8]. The great advantage of this approach is that the rate-determing step of purification is by-passed [8,9,20,21]. The principle is one of fractional distillation of the individual peptides from the mixture, using either the probe or ion-source heater. This method is best suited to mixtures of two to five peptides. A requirement for sample ionisation is that it must be volatile and hence derivatisation of peptides/proteins was necessary leading to the development of "fast reaction" high yield acetyl-permethyl derivatives [22,23]. The g.c./ms strategy was also developed to examine mixtures of small peptides which are suitably derivatised.

However, there were several drawbacks of EI-MS arising from;

(a) the need to derivatise samples,

(b) the relatively large quantities of sample required (greater than 5 nanomoles), and

(c) the extensive fragmentation, due to the transfer of high internal energy, so that often EI mass spectra do not extend to the molecular ion, and thus valuable information is lost, i.e. molecular weight. To overcome the disadvantages associated with "hard" ionisation techniques alternative procedures were developed, promoting the study of peptides by "soft" ionisation methods, the most significant of which has been fast atom bombardment (FAB), discussed below. This method of ionisation has been that employed throughout this thesis research work.

1.3.4 **Soft Ionisation Techniques**

One of the most successful soft-ionisation techniques is field desorption (FD), introduced by Beckey in 1969 [24]. The principle of FD-MS involves the ionic desorption of atoms or molecules in a high electric field. The analysis of peptides by this technique usually allows observation of M^+ and $[M+H]^+$, and may also give rise to adduct ions such as $[M+Na]^+$ and $[M+K]^+$. Field desorption mass spectrometry, despite having proven to be a valuable mass spectrometric method not only in the peptide field but also for a wide range of biologically significant compounds [25], was found to have some disadvantages; most spectra were devoid of fragmentation data and the fragile filaments were easily destroyed. However, by the late 1970s "soft" ionisation techniques were widely recognised as powerful methods for biopolymer analysis.

It was probably the introduction of secondary ion mass spectrometry

(SIMS) that innovated the development of fast atom bombardment mass spectrometry, (FAB), in 1981 by Barber and co-workers [26,27,28]. Secondary ion mass spectrometry involves the bombardment of a sample with a beam of ions resulting in the formation of sample ions which are then mass analysed. However, despite the applicability of this technique there are some difficulties which limit the usefulness of this method such as, charging effects at the surface bombarded. To overcome this problem the surface is bombarded by "hot" (fast) atoms instead of ions, this modification of sample ionisation was termed fast atom bombardment mass spectrometry (FAB-MS).

An advantage of FAB-MS over hard ionisation methods such as EI-MS is that the need for peptide derivitisation is circumvented, the sample can be directly loaded on to the target in the liquid phase. This ionisation method together with the introduction of High Field magnet technology [29,30], developed earlier, and other high mass instruments [31], allowed machine design with an extension of the resolvable mass range beyond 3000 Daltons, coupled with high sensitivity. Using these techniques it is now possible for sample analysis at the picomole-nanomole level. This ionisation method allows the observation of an intense molecular ion, and hence calculation of the molecular weight of compounds studied. Once a molecular weight of an unknown has been determined, immediately many of the possibilities as to the nature of the compound can be eliminated.

1.3.5 **FAB Ionisation**

In FAB-MS, ionisation is achieved by bombarding the liquid phase sample with a beam of high energy atoms; in the original experiments "fast" argon atoms were used. A beam of ions, Ar^+, can be produced by ionising argon atoms and accelerating the resulting ions through an electric field. The resulting (fast) ions are directed through an argon gas chamber where charge-exchange occurs to give fast atoms, with excess fast argon ions deflected.

$$Ar^+_{(fast)} + Ar_{(thermal)} \longrightarrow Ar_{(fast)} + Ar^+_{(thermal)}$$

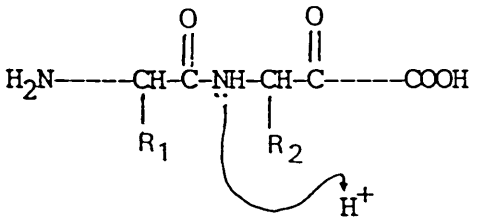
Other inert gases such as xenon were first used in this laboratory [32] to produce a beam of fast atoms and ions, with xenon giving three times the momentum of argon and resulting in an approximate three-fold increase in sensitivity of the resultant secondary beam. In fact, research here also led, in 1981, to the discovery that atom generation is not necessary for FAB, and the fast ions themselves give identical results. The FAB system used for the work in this thesis is a fast ion system designed and manufactured by M-Scan Ltd to allow close positioning of the ion gun and sample target, thus maximising ion flux and hence, secondary ion sensitivity. When a beam of fast atoms bombard a metal target coated with a sample in a liquid matrix, the large amount of kinetic energy in the atoms is dissipated in various ways, some of which lead to volatilisation and ionisation of the sample. Best results are usually obtainable by coating the target with a relatively involatile polar liquid; commonly used matrices are glycerol/thioglycerol and mono-nitrobenzylalcohol (MNBA). By maintaining a suitable voltage gradient from the target either

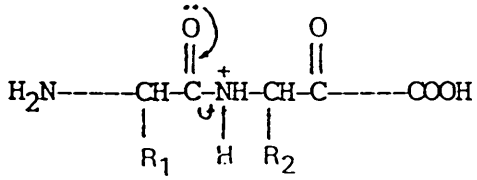
positive $[M+H]^+$ or negative $[M-H]^-$ ions can be directed towards the magnet of the mass spectrometer.

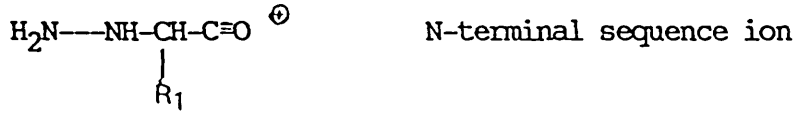
1.3.6 Peptide Fragmentation by FAB-MS

The way in which a peptide ionises on bombardment by fast atoms is not fully understood, but the fragmentation mechanisms of peptide bond fission were recognised by Morris and co-workers as being identical to those produced by chemical ionisation [33]. Morris *et al* postulated the protonation of the amide nitrogen of the peptide bond, giving rise to several types of fragmentation [27].

(i) N-terminal acylium ions: in which cleavage occurs across N-C$_\alpha$ bond



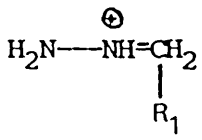




N-terminal sequence ion

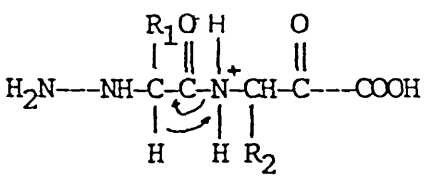
This acylium species may undergo further fragmentation with cleavage across the C$_\alpha$-CO bond to produce an aldimine fragment

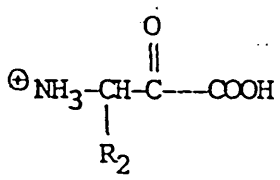
ion;



Aldimine fragment ions are often observed in FAB mass spectra, found a characteristic 28 a.m.u. (-CO) lower than the N-terminal (acylium) fragment ion from which it was derived.

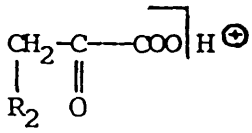
(ii) C-terminal (ammonium) fragment ions; peptide bond cleavage by the C_{α} re-arranged hydrogen;





C-terminal sequence ion

Further fragmentation of the N-C_{α} bond can occur with proton transfer and charge retention, with the formation of a C-terminal alkyl sequence ion;



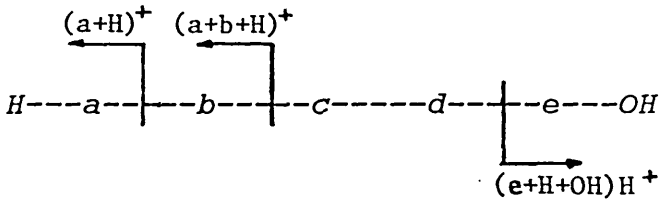
This alkyl species is observed a characteristic 15 a.m.u lower in

mass than its parent quaternary ammonium (C-terminal) fragment ion.

Thus the presence of aldimine and/or alkyl fragment ions in a FAB mass spectrum provides a means of distinguishing between those signals corresponding to N- or C-terminal fragment ions.

The way in which a peptide sequence may be determined, from a FAB mass spectrum containing N- and C-terminal fragment ions, is illustrated by considering the simple schematic representation of a peptide below;

N-terminal sequence ions



C-terminal sequence ions

where a, b, c, d and e, represent peptide residue masses. Thus, if fragmentation has occured at each peptide bond then by the mass differences between each series of C-terminal fragment ions or the series of N-terminal fragment ions the amino acid sequence can be determined.

FAB-MS has been used to solve structural problems of a wide range of compounds of biological interest, for example, the structure elucidation of a cardioactive nonapeptide with a blocked C-terminus, isolated from Aplysia brasiliana [34]. The recognition and identification of this blocked C-terminus by mass spectrometry

was of great importance since, in the absence of the blocking group the peptide has no activity. FAB-MS studies on the human Calcitonin Gene related peptide (hCGRP), [35], also provide an excellent illustration of FAB-MS work.

Fragmentation of peptides is not always observed in FAB mass spectra, which is the major drawback of this technique over EI-MS. However, advantages associated with this method such as the increase in sensitivity, the presence of a molecular ion species (not usually a feature of EI) and the need for derivatisation of peptides circumvented, are together of sufficient importance to compensate for the lack of fragmentation. The fragmentation of a peptide is dependent on probabilities, with no predictions possible, and thus, every peptide examined by FAB-MS has its own unique fragmentation pattern.

1.4 Strategies Developed in This Thesis Research Work for Peptide Sequence Analysis

Despite all the advantages of FAB-MS which make it preferable, in most structural investigations, to EI-MS, the need for sequence analysis is of the utmost importance thus, strategies have been developed for application in conjunction with FAB-MS for sequence studies. With this objective in mind, two methodologies have been developed in this thesis research work for peptide sequence analysis. The first method is a chemical procedure for the production of peptide fragments from a parent molecule with analysis of these fragments by FAB-MS. The second procedure is a method of mass spectrometry for enhancing fragment ion formation

and analysis, which has been applied for the first time to a genuine mixture analysis problem, related to the study of an enzyme sequence.

1.4.1 Partial Acid Hydrolysis for Peptide Sequencing

The first methodology, presented in Chapter Three of this thesis, addresses primarily the production of peptide fragments for sequence determination using partial acid hydrolysis. Partial acid hydrolysis, as a method for investigating peptide sequences, has been widely implemented over the past fifty years. Sanger, in his Nobel prize winning sequence analysis of insulin used partial acid hydrolysis to obtain structurally overlapping peptides, from which he determined the protein sequence. However, a general criticism of partial acid hydrolysis as a method for sequence analysis has been its relative lack of specificity for peptide bond cleavage, which arises from the susceptibilty of all peptide bonds to hydrolysis. There are many studies reported in the literature evaluating bond hydrolysis [36,37] which conclude that certain peptide bonds hydrolyse at a faster rate than others [38]. Another well established disadvantage associated with this technique is the deamidation of asparagine and glutamine residues by acid treatment. However, a fundamental difference between previous methods of partial acid hydrolysis and that presented in this thesis work is that traditionally dilute mineral or concentrated organic acids are used, whereas in the method developed here a **concentrated** mineral acid is employed. The method presented in this thesis exploits not only the lack of specificity of partial acid hydrolysis for bond cleavage but also the deamidation of asparagine

and glutamine residues is utilised as a positive feature of the technique. However, it is the method of analysis, FAB-MS, of the peptide acid hydrolysates, that determines those fragments observed in the FAB mass spectrum. This method for peptide sequence analysis, as developed here, has been applied to a wide range of peptides with molecular weight varying from 1346 (substance P) to 40,000 daltons (alcohol dehydrogenase). This method was devised to be applicable to the micro-sequencing of peptides, i.e. it is a highly sensitive technique, rapidly providing valuable sequence information which alone may solve the peptide structure or contribute seminal sequence data, which combined with information from other sequence methodologies will enable the complete peptide structure to be elucidated.

1.4.2 B/E Linked Scanning/CAD

The formation of daughter ions by fragmentation of the parent ion results in a fragmenation pathway, often allowing structure determination. If this fragmentation occurs after the ions have left the source and before they reach the detector then these fragment ions are termed "metastables". There are three regions where metastable ions can be formed, denoted the first, second and third field-free regions, which are depicted in the diagram of a reverse geometry mass spectrometer, Figure 1.3. Soon after the introduction of FAB-MS in 1981, "linked scans" were investigated as methods for studying fragmentation products formed outside the source. The results at this time were not encouraging for further development. Reported in this thesis however, is work investigating B/E linked scans for (a) pure peptide analysis, and more

importantly, (b) **direct mixture analysis** [39]. The most significant component of this work is the novel application of two-sector linked scanning to the sequence analysis of unknown components of an unfractionated protein digest mixture, thus, the rate-determining step of purification is removed.

There are a variety of linked scan methods for fragment ion analysis. Focussing and detection of fragment ions may be achieved by simultaneously varying combinations of the magnetic field strength, B, the accelerating voltage, V, and the electric field strength, E, of the electric sector, as discussed by Boyd and Beynon [40]. The type of scanning used in this second methodology for sequencing studies, is a B/E linked scan with the accelerating voltage kept constant. This method of scanning was selected as it enabled daughter ions (m_2^+) formed from a particular precursor ion (m_1^+) to be studied. Collisionally activated decompositions (CAD), using helium as the collision gas in the first field-free region, enhance fragment ion production. When initially examined as a potential method for sequencing, a major drawback was the sample quantity required (greater than 5 nanomoles) combined with the manual operation of the linked scan performed. However, with the current availabilty of new improved instrumentation and with the development of sophisticated computer soft-ware for summed scans (written at the request of this group) linked scanning/CAD as illustrated in the study presented here, has been shown to be a rapid and highly sensitive technique, ideal for fragmentation enhancement, allowing sequence analysis of pure peptides as well as those components of an unfractionated digest mixture.

1.5 **FAB-Mapping**

The methodologies developed in this thesis have been for sequence analysis with an emphasis on fragment ion generation. However, some methodologies utilise the lack of fragmentation in FAB mass spectra as an advantage, for example, FAB-Mapping [41]. The concept of the FAB-mapping strategy is to obtain a spectrum or "map" of a peptide generated by chemical/enzymic digestion of a protein. The digestion of the protein produces a number of peptide fragments which fall within the resolvable mass range of a modern mass spectrometer and hence, are amenable to examination. Thus, very complex mixtures can be directly examined without fear of the FAB mass spectrum becoming uninterpretable, as the absence of fragment ions results in a spectrum containing only molecular ions. It is therefore possible to rapidly screen either natural proteins or the products of recombinant DNA technology by implementing FAB-Mapping procedures. Since the probability of two peptides from any digest of a protein having the same molecular weight is low, the FAB-Map is thus an accurate diagnostic, determined by the presence or absence of certain peptide components, of the fidelity of the gene-derived sequence. Each ion assigned in the spectrum can be confirmed by chemically or enzymically manipulating the digest mixture in a predictable manner e.g. Edman degradation, carboxypeptidases, acetylation etc., followed by additional FAB-Maps after each treatment [42,43,44,45]. Point mutations are usually visualised as mass shifts with the numerical value of this shift revealing the identity of the replacement amino acid. DNA sequencing is an extremely elegant and rapid procedure for the structure elucidation

of large proteins. However, positive identification of initiation, termination and intron codons requires comparison with amino acid sequences so as to prevent misinterpretation of the DNA reading frame. A non-terminating, sometimes non-sensical amino acid sequence may be derived by simply missing or incorrectly assigning one base when reading the DNA-sequence. The erroneous observation of a termination codon is a second such error. Furthermore, it is not possible for post-translational modifications such as glycosylation and phosphorylation to be determined from the DNA sequence. Therefore, an emphasis lies with methodology development for providing a means for rapid checking of partial peptide sequence derived solely from the DNA coding information, from which it is possible to interpret incorrect protein structure.

Conversely, FAB-Mapping techniques can be implemented for the generation of small stretches of peptide sequence, typically between six to twelve amino acid residues, from a protein of unknown sequence, to employ in the production of synthetic oligonucleotide probes to screen a gene library. Methionine and tryptophan containing peptides are favoured for probe generation, because unlike other amino acids they do not have degenerate codes. Methionine residues, in general occur infrequently in protein sequences hence, protein cleavage with cyanogen bromide (see Figure 1.2) has the advantages of (a) generating large peptide fragments which are reduced in secondary and tertiary structure which can be further digested with enzymes, and (b) producing methionine containing peptides.

FAB-Mapping techniques are ideal for generating peptide fragments

from large proteins and have been applied to complex problems such as human low density lipoprotein APO B [46,47]. As well as being a highly sensitive method involving direct mixture analysis, a definite advantage of the FAB-Mapping procedure is that there is an equal probability of observing C-terminal as well as N-terminal regions of the protein, unlike alternate methods of analysis such as automated gas-phase sequencing.

1.5.1 **Applications of FAB-Mapping**

Two major studies in this thesis, employing FAB-Mapping techniques, have been the primary sequence analysis of two proteins D-xylose isomerase and pyruvate decarboxylase. D-xylose isomerase *Arthrobacter* strain B3728, molecular weight 43 kDa [48], is an enzyme of industrial importance for catalysing the conversion of D-glucose to its sweeter isomer D-fructose. This reaction is the basis for the modern industrial production of high fructose syrups. An important feature of this study was that it allowed the otherwise uninterpretable X-ray crystallography data to be solved. This study elucidated approximately 80% of the primary sequence of D-xylose isomerase prior to the determination of the gene sequence, which was itself corrected using this data. This is believed to be the first study in which medium resolution X-ray crystallography data has been interpreted by utilising mass spectrometric data to define sequence and molecular weight. The FAB-Mapping strategy was also successfully implemented, in conjunction with other methodologies, some developed here, to the primary sequence analysis of pyruvate decarboxylase, isolated from wild type yeast, *Saccharomyces cerevisiae*. Pyruvate decarboxylase, molecular

weight 60 kDa [49,50], is responsible for the production of ethanol from pyruvate by anaerobic glycolysis (alcoholic fermentation) in organisms such as yeast, this reaction is of great industrial interest.

Due to the phenomenon of supression [27] of certain peptide components in a digest mixture by others, on FAB-MS analysis of the mixture, the need for purification methods such as High Performance Liquid Chromatography (HPLC), remain invaluable for sequence studies. Classical sequencing techniques such as automated gas-phase sequencing may be used in combination with mass spectrometric methods for the sequence determination of purified peptide components. Reverse-phase HPLC has been the main method of fractionation implemented where needed in the FAB-Mapping studies reported in this thesis. Reverse-phase chromatography involves a stationary phase in which long-chain alkyl groups are immobilised on silica microspheres, and a mobile phase which consists of a mixed organic/aqueous solvent. By applying a solvent gradient the larger and more hydrophobic peptides are retarded on the column, eluting in high percentages of organic solvent.

1.6 **High Mass Analysis**

Classically molecular weight estimations of large proteins are obtained by gel filtration or SDS-polyacrylamide gel electrophoresis, with cal ibration against a standard set of proteins of known molecular weight [51]. Techniques such as mass spectrometry provide an accurate and highly sensitive method for molecular weight determinations. However, before 1977 it was

difficult to analyse molecular weights in excess of 1000 a.m.u. by MS, though the mass range amenable to mass spectrometry has increased rapidly during the past decade. From the equation of motion of an ion passing through a two-sector mass spectrometer (see Figure 1.3) it can be determined that mass is directly proportional to both magnetic field strength (B) and the radius of the magnet (R), thus, there are two options for extension of the mass range of the mass spectrometer, i.e. an increase of either B or R. Most instruments currently working at high mass are high field (HF) mass spectrometers. The advent of HF-MS [29,30] and then desorption-ionisation techniques such as FAB-MS [26,27,28], allowed the mass range to be considerably extended, with resolvable molecular ions being observed beyond 3000 a.m.u. Studies into the application of FAB-MS to molecules in the mass range 5000-8000 Da, showed that small proteins such as insulin and monellin [52] are amenable to direct analysis.

A second option for increasing the mass range is to relinquish the use of magnets, as for example with plasma desorption mass spectrometry (PD-MS). This approach utilises radionuclide ionisation (^{252}Cf plasma desorption) for rapid heating. A fission fragment (^{252}Cf decays to two fission products) on impact with the sample causes a localised "hot spot", of about 10,000°K. This large amount of energy in the sample results in rapid volatilisation before decomposition can occur; ionisation occurs through ion-pair formation and ion/molecule reactions, with analysis of ions by a time-of-flight mass spectrometer. Plasma desorption has proved successful for the analysis of proteins, with molecular weight

determinations up to 23 kDa [53].

Recently mass spectrometry has been extended for the study of high molecular weight biopolymers, based upon **electrospray ionisation.** This technique has been applied in this thesis, and has been instrumental in determining the molecular weights of subunits A1, A2 and B of the enzyme pyruvate decarboxylase. Electrospray ionisation, (ESI), was initially introduced in 1968 by Dole *et al* [54] but has been recently revived by Fenn *et al* [55,56] and others. ESI is a soft-ionisation method capable of producing multiply charged molecular ions and hence, according to the mass/charge relationship, the charged species is detected within the resolvable mass range of the instrument. The signal cluster is characterised by its bell-shape formation related to the number of charges carried. The optimum instrument design for ESI is not a conventional two-sector mass spectrometer, but a quadrupole. A quadrupole consists of four precisely arranged parallel rods. Between each pair of opposite and electrically connected rods a dc voltage and a superimposed radio-frequency (r.f.) potential are applied. Under the influence of this combination of fields, ions in the analyser undergo complex trajectories [57,58,59]. With the r.f./dc voltage ratio constant, the voltages are varied to effect separation of ions according to mass. The mass range of new quadrupoles is approximately 2000-3000 Da, so that a large protein carrying multiple charges, e.g. a protein mass 100 kDa with 100 positive charges will be observed in the spectrum at m/z 1000 thus, lying within the resolvable mass range of a quadrupole. A simplistic diagram of an ESI-MS design is shown in Figure 1.4.

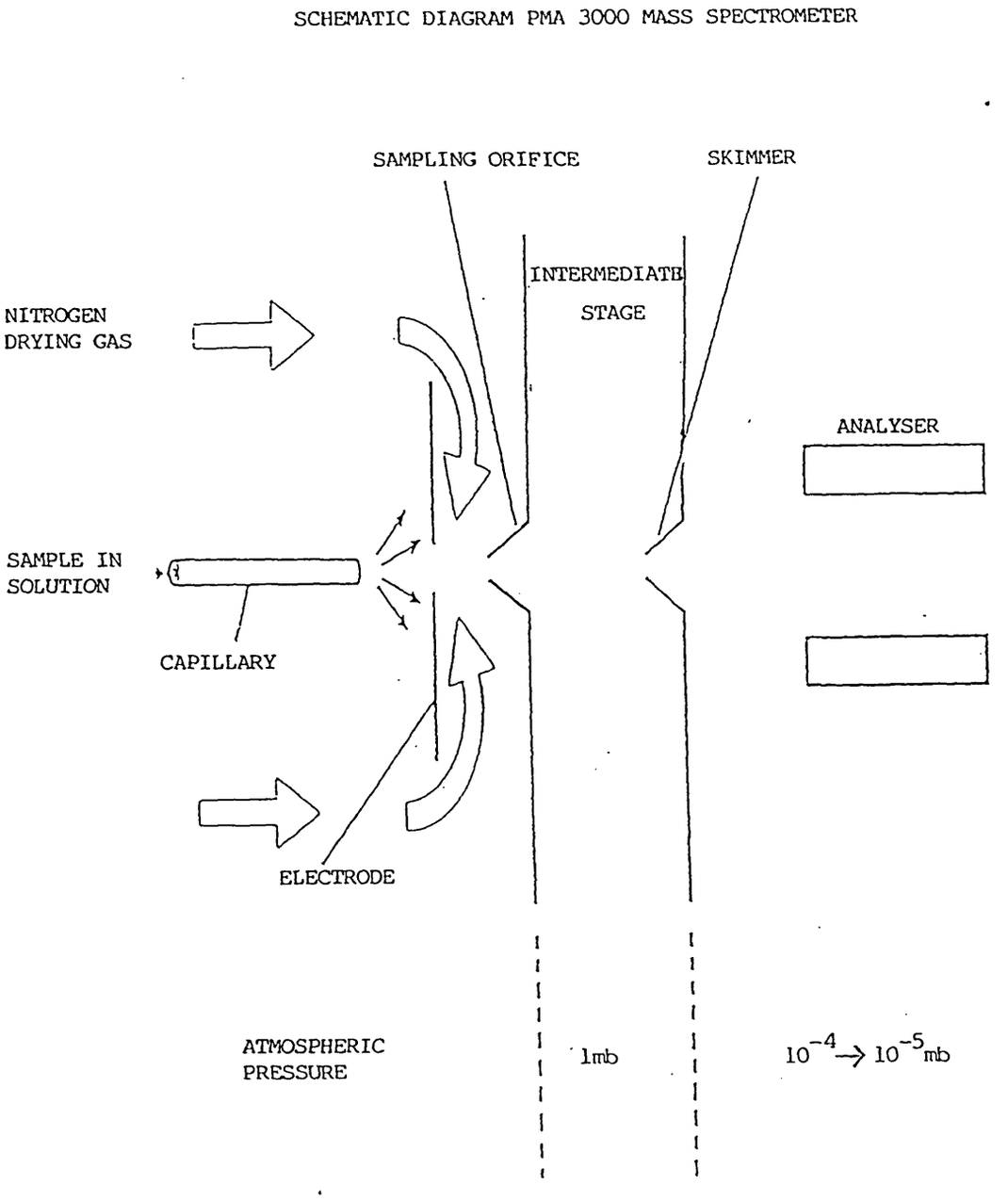


Figure 1.4 Schematic Diagram of a PMA 3000 Electrospray Mass Spectrometer

ESI-MS as applied in this thesis has allowed a molecular weight determination, for the first time, for each of the isolated monomer units of pyruvate decarboxylase, previously estimated within ±3000 Da, to a value within ±5-20 Da (+0.02%). The results obtained were from less than a nanomole of sample demonstrating the sensitivity of the procedure.

The work reported in this thesis demonstrates the broad applicability of mass spectrometry as an analytical tool, when used in a concerted strategy in combination with other complementary techniques for sequence analysis of proteins and peptides of biological significance. Recognising the current needs for advanced sequencing technologies, two methods, linked scanning of enzyme digests and concentrated mineral acid partial hydrolysis, have been developed and subsequently applied to problems of relevance in the field of protein chemistry. A third component of this thesis has been the study of two enzymes, D-xylose isomerase and pyruvate decarboxylase, which as well as investigating their primary sequences, has demonstrated how mass spectrometry can significantly extend and complement the data yielded by other methodologies such as X-ray Crystallography and DNA-sequencing techniques.

CHAPTER TWO

FAB-MAPPING TECHNIQUES APPLIED TO THE PRIMARY SEQUENCE ANALYSIS OF D-XYLOSE ISOMERASE, *Arthrobacter* Strain B3728

2.1 Introduction

This chapter describes FAB-Mapping research which was essential in determining the sequence and crystal structure of the industrially important enzyme D-xylose isomerase, *Arthrobacter* strain B3728 [48]. A number of bacteria contain D-xylose isomerases that catalyse in vivo the isomerisation of D-xylose to D-xylulose and also convert D-glucose to D-fructose under certain conditions [60,61]. The conversion of D-glucose to D-fructose is the basis for the current industrial production of large quantities of high fructose syrups. Commercially used in an immobilised form D-xylose isomerase converts glucose to its sweeter isomer fructose, to give a product virtually identical with "invert sugar" produced from beet and cane. A second major commercial interest lies in the production of ethanol from xylose occurring in several yeast strains, in particular, those belonging to the genera *Pachysolen* and *Candida*. However, the industrially used yeasts of the genus *Saccharomyces* are unable to produce ethanol in this way. Work is currently directed towards introduction of a bacterial D-xylose isomerase gene into a suitable yeast host [62,63,64,65]. Due to the commercial importance of this enzyme a

protein engineering programme has been initiated on D-xylose isomerase, *Arthrobacter* strain B3728, investigating the primary, secondary, tertiary and quaternary structures.

In the absence of an initial primary structure a tentative sequence was built up, combining information derived from X-ray crystallographic techniques [66], the N-terminal sequence, 15 residues, determined by gas-phase sequencing in this group, and by probable sequence homology studies with other D-xylose isomerase enzymes; *Bacillus Subtilis* [67], *E.Coli* [68], *Streptomyces violaceoniger* [69] and *Ampuriella sp* [70] (see Figure 2.4, page 82). However, it proved impossible to fully interpret the X-ray crystallographic data, and thus the crystal structure, by these means, and Professor Blow (Biophysics) sought collaboration from this group on the primary structure determination. A mass spectrometric approach was employed for the sequence analysis investigation, reported here.

Fast atom bombardment mass spectrometry, FAB-MS, introduced in 1981 [26,27,28], has played an important role in biopolymer analysis. The development of FAB-MS, combined with High Field magnet technology, HF-MS [29,30] developed earlier, has allowed large molecules of 3000 Da or more to be examined and characterised. Today, commercially available mass spectrometers can determine molecular weights ranging from 10,000 to 15,000 Da at full accelerating voltages, 8-10 keV. However, many biopolymers of interest have molecular weights which greatly exceed this mass range. The concept of FAB-Mapping [41] was devised in 1983

specifically to address questions of primary structure in large biopolymers, including enzymes, receptors and recombinant proteins. The FAB-Mapping strategy has been applied to a variety of significant biological problems such as glucose transporter protein [71] and apolipoprotein B [46,47]. Conversely, FAB-Mapping techniques can also be used as part of a strategy for the elucidation of short peptide sequences. Typically, from a protein of unknown structure, between six to twelve amino acid residues are required, to employ for the generation of synthetic oligonucleotide probes to screen a gene library.

The FAB-Mapping procedure involves a combination of chemical and or enzymic techniques for the generation of small peptides falling within the resolveable mass range of a modern mass spectrometer. The procedure developed in this thesis, for the study of D-xylose isomerase, has been chemical treatment followed by enzymic digestion. In many cases the three dimensional conformation of a protein sterically hinders the approach of a large enzyme molecule, and so potential cleavage sites in the interior of the folded protein substrate remain untouched by the enzyme, which results in limited proteolysis. In contrast, the effects of steric hinderance on the approach of a small molecule such as cyanogen bromide (CNBr) are not usually significant. Thus, by employing the appropriate experimental conditions specific bonds can be chemically cleaved to produce large peptide fragments with less secondary and tertiary structure. These peptides are now suitable for enzymic proteolysis. The resultant digest mixture is then analysed by FAB-MS. The FAB mass spectrum is termed a "FAB-Map". A typical FAB-Map contains a

number of signals of varying intensity. The mass signals, $[M+H]^+$, should correspond to peptides derived from the sample, which can be mapped on to the anticipated structure if known, demonstrating that the primary sequence is correct. The larger the protein the more complicated is the mapping of the various mass signals on to the sequence, in such cases a computer programme may be employed (M-Scan Ltd., Ascot, and R. McDowell, Dept. Biochem., Imperial College, London). Briefly, this computer programme consists of two main sections, a protein sequence databank and a search programme. The protein sequence under investigation is stored in the databank, with any modifications of the sequence, for example, substitution of methionine residues for the homoserine derivative, incorporated. The search programme is able to locate peptides along the protein sequence on the basis of their molecular weights. For each mass value there often exists a number of sequences, derived from the predicted sequence however, by utilising the knowledge of the digest specificty, e.g. peptides generated from a trypsin digestion would be expected to terminate with either a lysine or arginine residue, many of the sequence possibilities can be dissmissed as not satisfying the digest conditions. Signal assignments are subsequently validated by predictable chemical and enzymic manipulations of the digest, for example Edman degradation [2] and carboxypeptidase B, followed by additional FAB-Maps after each treatment. Signals in the first FAB-Map should shift in a predictable manner, with unexpected mass shifts often indicative of post-translational modifications which sometimes may be deduced from the numerical value of the mass difference. Carboxypeptidase B is a very useful enzyme (specific for cleaving C-terminal arginine

and lysine residues only) for use in conjunction with trypsin. A comparison of the tryptic FAB-Map with the FAB-Map produced after an aliquot of the intact tryptic digest mixture has been digested with carboxypeptidase B is very useful for identifying the true carboxy terminus of the protein; the peptide corresponding to the C-terminus will not shift in mass after CpB treatment, unless the C-terminal amino acid of the protein is lysine or arginine.

The strategy of FAB-Mapping is also applicable to proteins of unknown sequence. Emphasis lies on the provision of short stretches of sequence, typically six to twelve amino acid residues, for the production of synthetic oligonucleotide probes for "fishing" in a gene pool for the gene encoding the protein. The FAB-Mapping procedure remains identical to that described above. However, the role of Edman degradation and CpB digestion of the intact double digest mixture, are to provide valuable sequence information from unknowns for N- and C-terminal amino acid residue determinations. The next step would be purification of the digest mixture by techniques such as High Performance Liquid Chromatography, HPLC, with identification of each isolated peptide by mass spectrometry complemented by classical protein sequencing methods such as automated gas-phase Edman degradation. This chapter describes the successful application of FAB-Mapping techniques with appropriate strategies devised for elucidating valuable sequence information which made an important contribution to the primary sequence analysis of D-xylose isomerase.

There were initially two objectives of this study;

1) the primary sequence analysis of D-xylose isomerase, by a novel approach utilising mass spectrometric data for the interpretation of the otherwise inconclusive X-ray crystallographic data, and
2) the provision of definitive stretches of amino acid sequence for the subsequent production of complementary synthetic oligonucleotide probes.

Utilising the data from this study the gene sequence was later generated and corrected. As discussed previously errors of translation or some other modification of the sequence unanticipated from the nucleotide template, for example post-translational modifications, are common place and thus, there is a growing demand for complementary protein sequencing techniques to determine the fidelity of a sequence derived solely by DNA technologies. Therefore, a third objective of this study was to;
(a) check the gene sequence, and
(b) elucidate a number of residues unable to be translated from the DNA template.

The sample of D-xylose isomerase used in the biochemical structure work reported here was supplied as a crystal from our colleagues in the Department of Biophysics.

2.2 Results and Discussion

PART I

2.2.1 Section A: PRIMARY SEQUENCE ANALYSIS of D-xylose isomerase, *Arthrobacter* strain B3728

The objectives of this first section of study were, (a) the primary sequence analysis of D-xylose isomerase by FAB-Mapping techniques, (b) interpretation of the otherwise uninterpretable crystallography data, and (c) the provision of sequence data suitable to employ for the generation of oligonucleotide probes.

The N-terminus of each identical subunit of D-xylose isomerase, Ser-Val-Gln-Pro-Ala-Asp-His-Phe-Thr-Phe-Gly-Leu-Trp, had been previously determined by automated gas-phase sequencing in this laboratory. This N-terminal sequence represented the only stretch of amino acid residues, in the initial electron density map sequence, not determined by X-ray crystallographic methods.

The strategy of FAB-Mapping, as described above, was applied to a crystal of D-xylose isomerase. Chemical treatment followed by enzymic digestion was selected as a suitable approach for cleaving this enzyme. The protein was initially cleaved with cyanogen bromide (CNBr) and then sub-digested with trypsin. Cyanogen bromide cleaves on the C-terminal side of methionine residues, with modification of the methionine to its homoserine and/or lactone derivative, see Figure 1.2 (Chapter One). When observed mass spectrometrically the derivatisation of methionine (at the C-terminus of a CNBr generated peptide fragment) is often discernible by the presence of two related signals 18 a.m.u. apart which correspond to the homoserine and its lactone. Double digestion of the protein, D-xylose isomerase, generated a large number of peptides with masses amenable to detection and analysis

by FAB-MS. The mass spectrometer used throughout this study (ZAB-HF) had a resolvable mass range of approximately 3300 Da at full accelerating voltages. FAB-MS analysis of an aliquot of this CNBr/Tryptic digest produced the FAB-Map shown in Figure 2.1. This spectrum contains a number of intense signals corresponding to peptide fragments generated from the double digestion of D-xylose isomerase, the observed quasi-molecular ions, $[M+H]^+$ are summarised in Table 2.1. However, the amino acid sequence of these peptides was unknown. As described above, the next step in the FAB-Mapping strategy in the absence of a sequence is to chemically and/or enzymically manipulate the digest mixture in a predictable manner, prior to purification, with analysis of the products after each step by FAB-MS. Though a tentative sequence had been assigned by X-ray crystallographic methods this data was inconclusive and therefore, the FAB-Mapping strategy was applied to the primary sequence analysis of D-xylose isomerase as if it were unknown, with data from this study utilised for interpretation of the electron density data. An aliquot of the double digest (CNBr/tryptic) was submitted to two cycles of manual Edman degradation. The Edman chemistry is illustrated in Figure 1.1 (Chapter One). Edman degradation is a chemical method for the removal of the N-terminal amino acid of a peptide/protein; for the reaction to proceed the N-terminus must be the free amine. After each step of Edman chemistry the products were analysed by FAB-MS, the resulting FAB-Maps are illustrated in Figures 2.1.1 and 2.1.2, respectively. The mass differences between those signals seen in the initial FAB-Map (Figure 2.1) and those observed after one step of Edman degradation, Figure 2.1.1, allow determination of the N-terminal

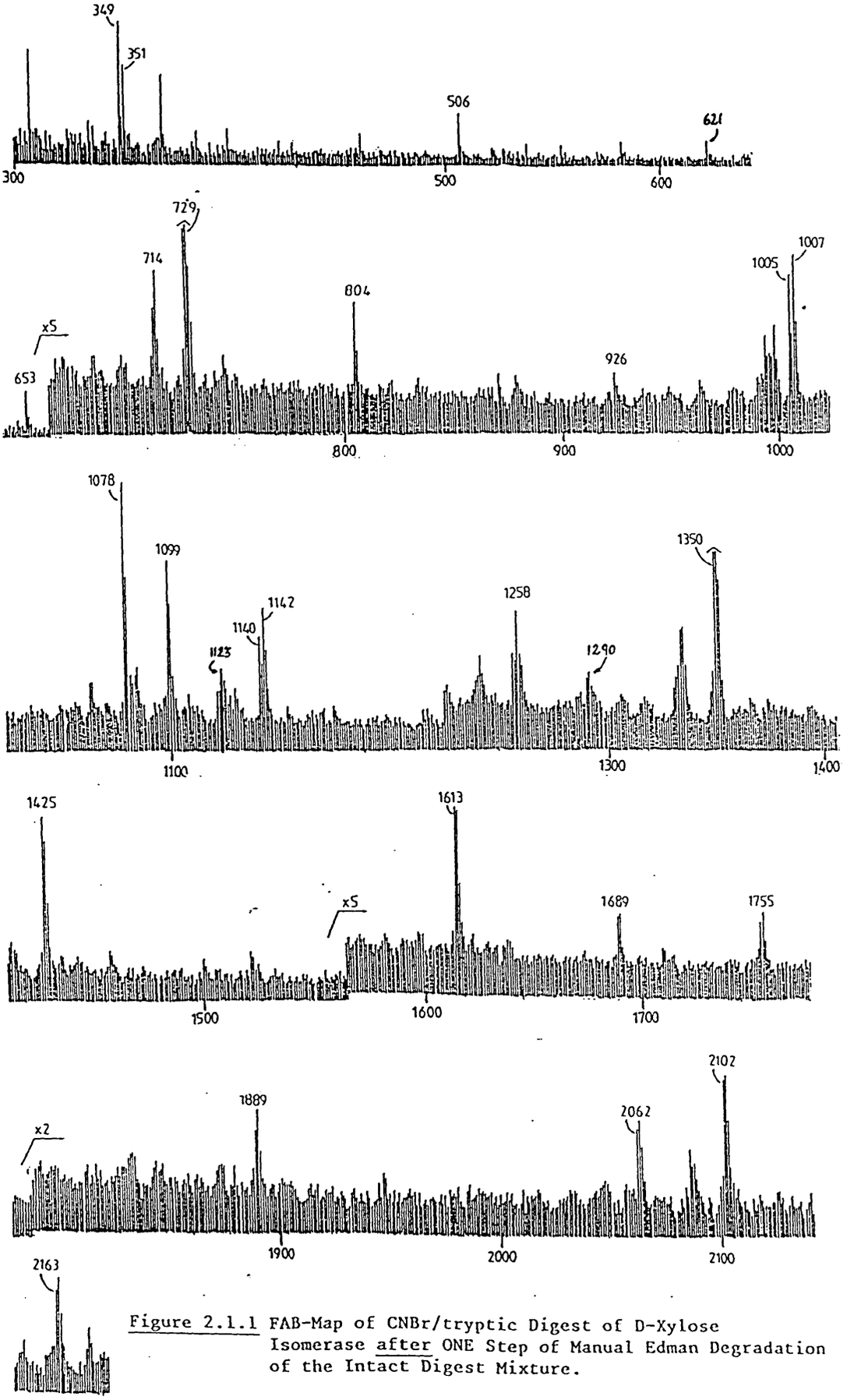


Figure 2.1.1 FAB-Map of CNBr/tryptic Digest of D-Xylose Isomerase after ONE Step of Manual Edman Degradation of the Intact Digest Mixture.

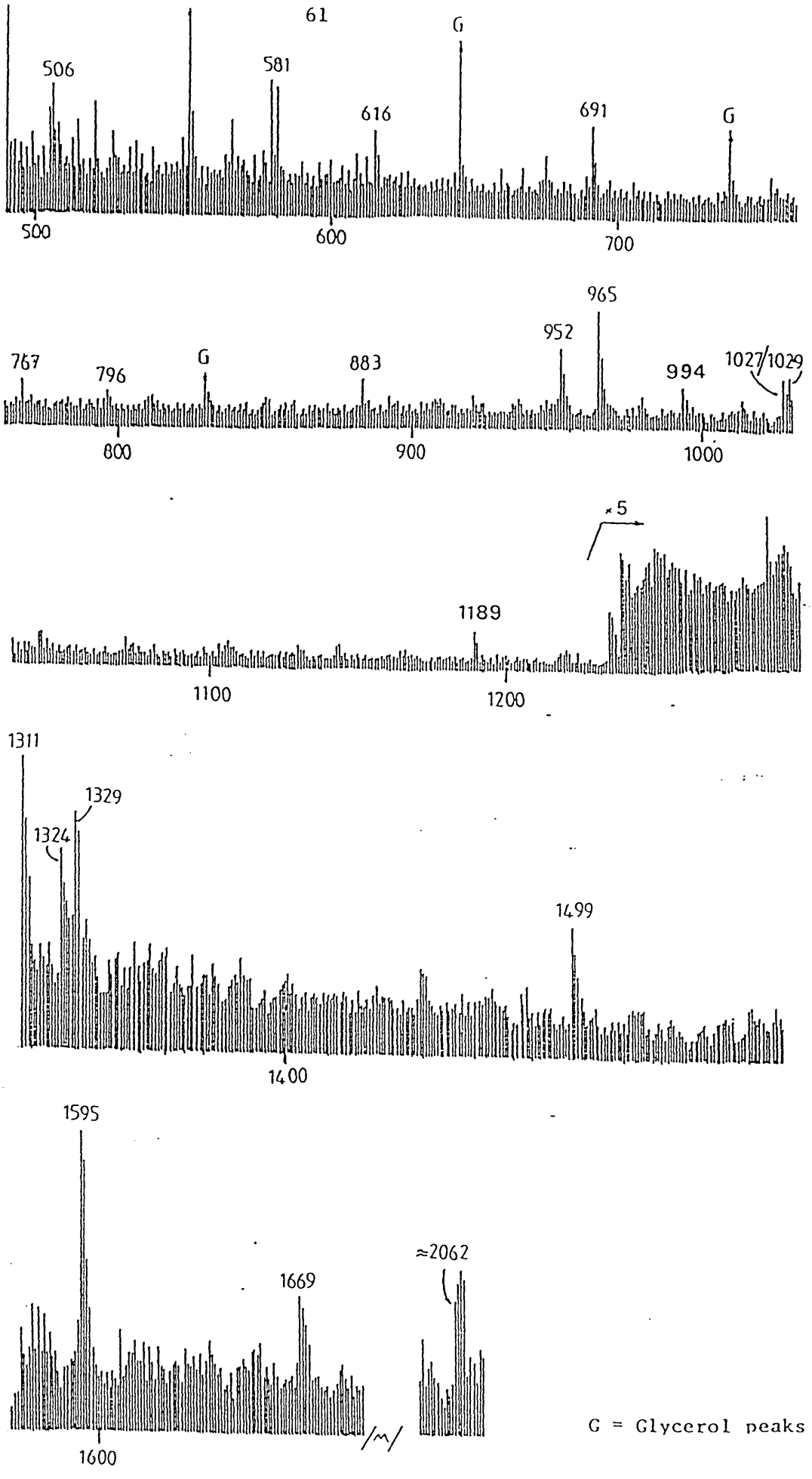


Figure 2.1.2 FAB-Map of CNBr/tryptic Digest of D-Xylose Isomerase after TWO steps of Manual Edman Degradation of the Intact Digest Mixture.

CNBr/Tryptic digest $[M+H]^+$	EDMAN step 1 $[M+H]^+$	EDMAN step 2 $[M+H]^+$	CNBr/Tryptic/CpB: $[M+H]^+$
397	349	506	389
399	506	581	461
413	581	616	553
417	621	675	573
545	653	691	583
548	729	883	645
573	804	952	688
576	924	965	829
590	1005/1007	994	833
634	1078	1027/1029	846
649	1099	1040	856
670	1123	1072	921
677	1140	1102/1105	974
721	1142	1129	990
736	1258	1142	1055
766	1290	1189	1056
809	1350	1220	1114
828	1425	1236	1128
844	1613	1311	1151
854	1689	1324	1177
898	1755	1329	1195
942	1889	1446	1253
974	2062	1449	1261
984	2163	1595	1263
1055	~2630	1669	1336
1118	~2750		1378
1177			1381
1195			1557
1212			1572
1213			1587
1263			1711
1279			~2037
1389			
1463			
1698			
1826			
1930			
~2160			

Table 2.1 Intense Quasi-Molecular Ions, $[M+H]^+$, Observed in; (i) CNBr/tryptic FAB-Map (Fig. 2.1), (ii) CNBr/tryptic FAB-Map after One Step of Edman Degradation, (iii) CNBr/tryptic FAB-Map after Two Steps of Edman degradation, and (iv) CNBr/tryptic/CpB FAB-Map, of D-Xylose Isomerase.

amino acid of each peptide observed. Similarly, by comparing the FAB-Maps generated after one step and then after two steps of manual Edman degradation (Figures 2.1.1 and 2.1.2) and by calculation of signal mass shifts the second N-terminal amino acid in each peptide sequence can be determined. Hence, this procedure is generally referred to as "subtractive Edman" [43,44] (observing peptides rather than PTH derivatives). From these two steps of subtractive Edman degradation the first two amino acids of peptide fragments observed in the FAB-Map of the double digest mixture were interpreted. To determine those fragments which had been generated by trypsin digestion and those which were derived from CNBr cleavage of the protein a method for C-terminal analysis was implemented, carboxypeptidase B, CpB, digestion. As discussed CpB is specific for cleaving C-terminal lysine and arginine residues only and thus, those peptides generated by trypsin digestion are expected to terminate with either a lysine or arginine residue. Those peptides formed from C-terminal CNBr cleavage will terminate with a chemically modified methionine residue. FAB-MS analysis of a CpB digestion of an aliquot of the CNBr/tryptic digest mixture will allow visualisation of those peptides with C-terminal lysine or arginine residues by mass losses of 128 and 156 a.m.u. respectively. Those signals in the CNBr/tryptic/CpB FAB-Map, Figure 2.1.3, that have not changed in mass can be assigned as peptides terminating with a modified methionine residue or to the natural C-terminus of D-xylose isomerase. The experiments described above very rapidly allowed both N- and C-terminal residue analysis of the generated peptide fragments. However, it should be noted that not all fragments generated by the digests are observed by FAB-MS

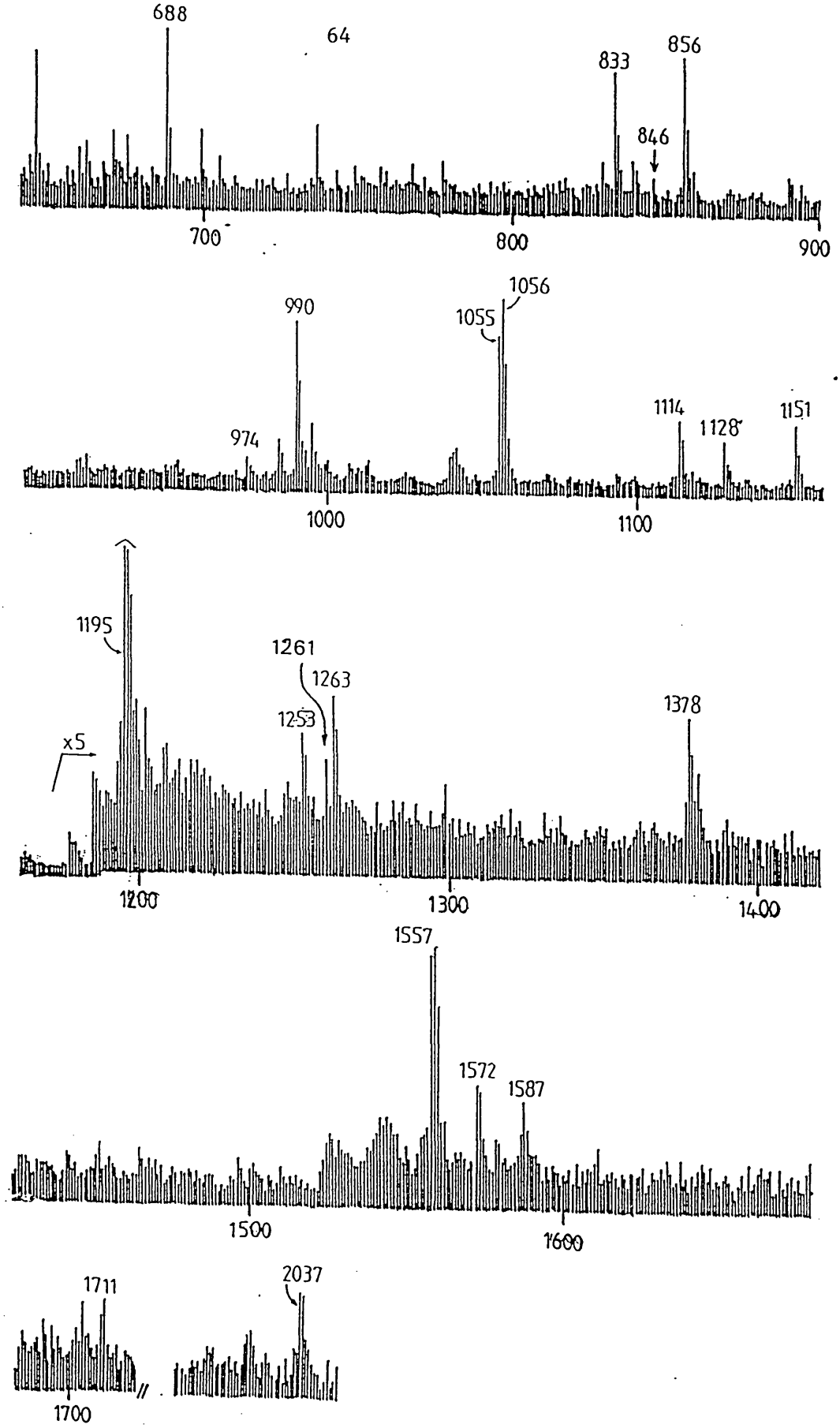


Figure 2.1.3 CNBr/tryptic/CARBOXYPEPTIDASE B FAB-Map of D-Xylose Isomerase.

analysis of the digest mixtures, or after Edman degradation and/or carboxypeptidase B treatment. This is rationalised by the phenomenon of suppression [27] of certain peptides by others in a mixture when the intact mixture is directly analysed by FAB-MS. Therefore, the absence of any expected peptides in a FAB-Map of a mixture and/or those expected after steps of Edman degradation or CpB digestion is inconclusive. In order to observe these suppressed peptides the digest mixture must be purified by techniques such as High Performance Liquid Chromatography, HPLC.

The intense mass signals, $[M+H]^+$, in the CNBr/tryptic FAB-Map (Figure 2.1) of D-xylose isomerase, are listed together with the Edman degradation and carboxypeptidase B data, in Table 2.1 (see Page 60). Despite the phenomenon of suppression a large number of signals are observed in each FAB-Map. As discussed, those signals which are not shifted after CpB digestion, such as m/z 1177 and 1195, correspond to peptides with C-terminal modified methionine residues or to the natural C-terminus of D-xylose isomerase. The pair of ions m/z 1177 and 1195 have a mass difference of 18 a.m.u. indicating that they correspond to a peptide terminating with a methionine residue chemically modified to its lactone and homoserine by CNBr digestion. This conclusion was supported by their presence in the CNBr/tryptic/CpB FAB-Map, Table 2.1.3.

Two computer programmes were used in this study; the function of one programme (K. Hendrick, Dept. Biophys., Imperial College, London) is to compute all combinations of the possible three dimensional amino acid structures in the electron density map. This

programme was in fact based upon that available in the Department of Biochemistry, Imperial College, London (R. McDowell and M-Scan Ltd., Silwood Park, Ascot). This latter programme consists of a protein databank (in which the primary sequence of proteins are stored) and a search programme. The function of the search programme is to locate stretches of peptide sequence on the basis of their mass within the primary sequence of the protein from which they are derived. This programme was therefore implemented in the second part of this study to check the fidelity of the gene-derived sequence. To distinguish between the two reference will be made to programme I (Dept. Biochem.) and programme IA (Dept. Biophys.).

From the mass spectrometric/chemical data summarised in Table 2.1, two signals could be assigned on to the sequence in the electron density map, $[M+H]^+$ 1177/1195 and $[M+H]^+$ 1212.

Interpretation of $[M+H]^+$ 1177/1195: These two signals correspond to a peptide fragment bearing a chemically modified C-terminal methionine residue as discussed above. The presence of signals m/z 1078 and m/z 965 in Figures 2.1.1 and 2.1.2 respectively, allowed valine and leucine/isoleucine (respectively) to be determined as the N-terminal amino acid residues of this peptide. Furthermore, it was inferred that the residue preceding valine in the intact protein was either arginine, lysine or methionine, i.e. a digest cleavage site. The location of this peptide within the original protein sequence was sought by employing the computer programme, IA, available in the Department of Biophysics (K. Hendrick). The function of this programme was to

search for the best mass fit of the peptide at m/z 1195/1177 within the tentative sequence of the electron density map, and then to compute all possible three dimensional amino acid structures lying within the defined parameters of the electron density map. The programme was able to mass assign the peptide on to residues 118 to 128 in the electron density map, with the sequence, Val-Leu/Ile-His-Asn-Ile-Asp-Leu-Ala-Ala-Glu-Met. However, previous interpretation of the X-ray cystallographic data had assigned glutamic acid and glycine to positions 120 and 127 respectively. These residues were revised to agree with the mass spectrometric data.

Interpretation of $[M+H]^+$ 1212: Two steps of subtractive Edman degradation identified leucine/isoleucine and phenylalanine respectively, by the mass differences between the signals m/z 1212 (Figure 2.1), m/z 1099 (Figure 2.1.1) and m/z 952 (Figure 2.1.2), as the N-terminal amino acid residues of this peptide and the C-terminal residue as arginine (see Table 2.1). Using the computer search programme (IA) for a best fit led to the assignment of the peptide on to the electron density map between positions 240 to 249, correcting an error in the electron density data at position 241. This residue had been assigned as tyrosine from the crystallographic data not phenylalanine as determined from this mass spectrometric study. This residue, 241, was subsequently changed to agree with the Edman/MS data which proved the second residue of this peptide at m/z 1212 to be phenylalanine. The sequence of this peptide was determined as Leu/Ile-Phe-His-Ile-Asp-Leu-Thr-Gly-Gln-Arg. The residue preceding

leucine in the intact protein sequence was logically assigned as either arginine, lysine or methionine.

The remaining signals from the FAB-Map not assigned, were run through the computer search programme (IA) for a best mass fit. However, few convincing possibilities were produced. Thus, it became apparent that the sequence in the electron density map contained many errors in residue assignment. Therefore, the peptides generated from the CNBr/trypsin digestion of D-xylose isomerase needed to be analysed by definitive sequencing methods rather than attempting to mass assign them on to the sequence derived by X-ray crystallographic methods. Thus, the next logical step in this study was to purify the digest mixture by reverse-phase HPLC, thereby isolating the observed peptides and also those peptides not previously observed due to suppression in the digest mixture. An aliquot of each HPLC fraction was analysed by FAB-MS. A second aliquot of each isolated and mass assigned peptide fragment was submitted for automated gas-phase sequencing. Presented below are seven peptide sequences identified and elucidated, after HPLC purification of the digest mixture, by a combination of mass spectrometric and gas-phase sequencing techniques.

Interpretation of $[M+H]^+$ 1930: This signal was observed in the FAB-Map of the double digest mixture, Figure 2.1. However, no previous sequence assignment could be made, with no Edman or CpB data interpretable from previous digests. Purification of the digest mixture allowed isolation of this peptide, identified by

FAB-MS. An aliquot of the HPLC fraction was examined by automated gas-phase sequencing with several cycles yielding the partial sequence, Leu-Ala-Glu-Leu-Gly. It was found that this pentapeptide could be matched with the primary sequence of the *Streptomyces violaceoniger* (Svn) strain of D-xylose isomerase between positions 42 and 46, see Figure 2.4 (see Page 82). However, the amino acid sequence subsequent to residue 46 in the Svn strain of D-xylose isomerase did not agree for a mass fit of 1930. Correlation of the electron density derived sequence between residues 42 and 46 with the Svn primary sequence showed that there was partial sequence homology in this region. However, the electron density data had assigned residue 44 as serine and not glutamine. In the sequence derived by automated gas-phase Edman degradation and in that of the Svn sequence residue 44 was assigned as glutamine therefore, in view of these sequences residue 44 in the electron density map was changed to read glutamine. However, even with this change to the electron density data no mass fit for m/z 1930 could be made on to the sequence in this region indicating that the amino acid residues after residue 46, in the electron density map, were incorrect. There also remained the possiblity that this pentapeptide was related to another peptide mass eluting in the same HPLC fraction as the peptide $[M+H]^+$ 1930, but which was not observed on FAB-MS analysis due to suppression. As this study progressed this matter was resolved and is reported later in the chapter.

Interpretation of $[M+H]^+$ 1389: This signal was observed in the FAB-Map of the double digest mixture (CNBr/tryptic), Figure

2.1. Two cycles of Edman degradation allowed identification of the N-terminal amino acid residues as valine and threonine, by the presence of signals m/z 1290 and 1189 in the FAB-Maps presented in Figures 2.1.1 and 2.1.2 respectively. FAB-MS analysis of the CNBr/tryptic/CpB digest mixture found a signal at m/z 1261 with m/z 1389 absent. Thus, the C-terminal residue of the peptide at m/z 1389 was determined to be lysine. However, no mass fit for m/z 1389 on to the electron density map could be made. Using the computer search programme (IA) the closest assignment of this signal corresponded to residues 88 to 99 in the electron density map of the intact protein. However, two amendments to the crystallographic data had to be made to positions 88 and 94 to agree with this mass spectrometric data. In the first instance to agree with the molecular weight and in the second to agree with the N-terminal residue valine, determined by subtractive Edman degradation. Residues 88 and 94 were tentatively assigned from the crystallographic data in the electron density map as threonine and alanine respectively. From the Edman/MS data threonine was shown to be an incorrect assignment. Eight cycles of automated gas-phase sequencing of the purified peptide determined the sequence, Val-Thr-Thr-Asn-Leu-Phe-Ser-His, corresponding to residues 88 to 95, which agreed with the mass spectrometric data. On receipt of this information confident mapping of the peptide at m/z 1389 between residues 88 to 99 in the intact protein sequence was possible, with two changes made to the elctron density map. The sequence of the peptide at m/z 1389 was therefore determined as, Val-Thr-Thr-Asn-Leu-Phe-Ser-His-Pro-Val-Phe-Lys. The residue preceding the N-terminal valine was deduced to be either lysine,

arginine or methionine.

Interpretation of [M+H]$^+$ 844: From mass signals at m/z 729 and m/z 616 observed in the FAB-Maps (Figures 2.1.1 and 2.1.2 respectively) generated after each step of Edman degradation, the N-terminal amino acid residues of this peptide were determined as aspartic acid and leucine/isoleucine. Carboxypeptidase B digestion of the double digest mixture identified the C-terminal residue as arginine, by the presence of the mass signal m/z 688 (Figure 2.1.3) corresponding to a mass difference of 156 a.m.u. between this signal and m/z 844. However, no mass assignment of this peptide, m/z 844, on to the electron density map of D-xylose isomerase could be made. The computer search programme (IA) failed to generate any convin cing sequence possibilities for this peptide mass. Purification of the double digest mixture allowed isolation by HPLC, and identification of this peptide, [M+H]$^+$ 844, by FAB-MS. An aliquot of this HPLC fraction was submitted for automated gas-phase sequencing, with seven cycles yielding the sequence, Asp-Leu-Ala-Ala-Ala-Leu-Asp. As the first two N-terminal residues of this sequence were in agreement with the Edman/MS data it was confidently concluded that this sequence corresponded to the peptide at m/z 844 and not some other component in the HPLC fraction, suppressed on FAB-MS analysis. However this heptapeptide, [M+H]$^+$ 688, obviosly does not correspond to the complete sequence of the peptide at m/z 844. Interpretation of the CpB/MS data allowed the remaining sequence to be determined. The mass signal at m/z 688 was observed in the FAB-Map of the CNBr/tryptic/CpB digest (Figure 2.1.3) and interpreted as the peptide formed as a result of

a C-terminal cleavage of arginine from the peptide at m/z 844. Thus, the sequence of the peptide at m/z 844 was determined as, Asp-Leu-Ala-Ala-Ala-Leu-Asp-Arg. Once this sequence had been elucidated it was compared with the electron density map to locate a stretch of amino acids which closely resembled this derived sequence. The best correlation was found between residues 149 to 156. The sequence between these residues was Gln-Val-Lys-Glu-Ala-Leu-Asp-Arg. Thus, the crystallographic study had incorrectly assigned residues 149 to 152. On receipt of the gas-phase/MS data the electron density map was amended between residues 149 and 152 to agree with the sequence derived from a combination of automated gas-phase sequencing and mass spectrometric techniques.

Interpretation of $[M+H]^+$ 1279: This signal was observed in the original FAB-Map of the double digest mixture. Two steps of Edman degradation allowed identification of the N-terminal amino acid residues as arginine and glutamic acid calculated from the mass differences between signals m/z 1279, 1123 and 994, observed in the FAB-Maps presented in Figures 2.1, 2.1.1 and 2.1.2 repectively. The C-terminal residue was determined by the mass difference between m/z 1279 and m/z 1151 (Figure 2.1.3) as lysine. However, no mass assignment of this peptide at m/z 1279 could be made on to the electron density map. This peptide was isolated by HPLC and identified by FAB-MS. An aliquot of the HPLC fraction containing this peptide was partially sequenced by automated gas-phase Edman degradation, with ten cycles determining the sequence, Arg-Glu-Gly-Val-Asp-Thr-Ala-Ala-Gly-Tyr. This sequence

was in agreement with the Edman degradation/MS data. However, the mass of this decapeptide was calculated as 1038. Utilising the CpB data, i.e. the C-terminal residue of this peptide was lysine allowed the total sequence to be determined. A simple calculation involving addition of the residue mass of lysine to the decapeptide mass with subtraction of this sum from 1279, arrived with a mass difference of 113 a.m.u.; this mass value clearly corresponds to a leucine/isoleucine residue. Hence, the sequence of the peptide at m/z 1279 was elucidated as,
Arg-Glu-Gly-Val-Asp-Thr-Ala-Ala-Gly-Tyr-Leu/Ile-Lys. The electron density map was searched for this sequence, with a close match of this peptide found between residues 158 to 169. However, two changes needed to be made to the electron density sequence in positions 159 and 167 to agree with the derived sequence; these residues had been assigned as glutamine and lysine respectively. However, the Edman degradation/MS data determined residue 159 as glutamic acid and not glutamine, with the assignment of residue 167 as lysine also shown to be incorrect by the automated gas-phase sequence data. On receipt of the MS/gas-phase sequence information the electron density map was changed in positions 159 and 167 to agree with the above sequence. An unusual feature of this peptide sequence is that the N-terminal amino acid residue was found to be arginine, a potential tryptic cleavage site. It was concluded that the residue preceding arginine was either a CNBr or tryptic cleavage site, i.e. methionine, arginine or lysine. However, the finding that Arg-158 had not been cleaved was rationalised by consideration of the adjacent amino acid residue, glutamic acid. The presence of this acidic residue next to the arginine is thought

to repel the active site of the trypsin and hence, the rate of hydrolysis of this peptide bond is reasoned to be slower.

Several HPLC fractions thought to contain pure peptides (analysed by FAB-MS) were examined by automated gas phase sequencing yielding partial sequence information, but sequence data which did not correlate with the mass spectrometric data. This observation suggested that certain HPLC fractions did not contain a pure peptide and that separation was incomplete. This incomplete separation was not initially recognised due the method of fraction analysis, FAB-MS. As mentioned earlier all the components of a peptide mixture, when FAB-MS analysed, may not be detected due to the suppression of certain components in the mixture by others. Purification of a complex digest mixture allows identification of previously suppressed peptides in the mixture however, if two peptides were to elute in similar positions then an HPLC fraction may not contain just one peptide. Classical gas-phase sequencing may only yield one interpretable sequence if one component is present in a higher concentration than the other. The signals analysed and discussed below were of the nature described, i.e. the sequence determined from automated gas phase sequencing did not correlate to the quasi-molecular ion determined by FAB-MS analysis of the HPLC fraction.

Interpretation of $[M+H]^+$ 1118: This signal was observed in the FAB-Map of the double digest mixture, Figure 2.1. Only one step of manual Edman degradation was interpretable, assigning leucine/isoleucine as the N-terminal amino acid residue, determined

by the presence of the mass signal m/z 1005 in Figure 2.1.1. The C-terminal residue was deduced to be lysine by the mass difference between the signal m/z 1118 and m/z 990 observed after CpB digestion of the double digest (Figure 2.1.3). However, no mass fit on to the electron density sequence could be made. Purification of the double digest mixture allowed isolation of this peptide, detected by FAB-MS. An aliquot was examined by automated gas-phase Edman degradation which yielded the sequence, Ile-Leu-Gly-Asp-Phe-Asn, with the N-terminal amino acid residue agreeing with that determined by manual Edman degradation/MS. The electron density map was screened for this hexapeptide sequence. It was found that the sequence between residues 70 to 75, in the electron density map closely resembled the gas-phase/MS derived sequence. However, residues 70 and 73 had been assigned as asparagine and arginine, respectively, in the electron density sequence. In view of the gas-phase/MS sequence data, the electron density map was amended in these two positions to agree with the above sequence. However, even with these changes to the electron density sequence the peptide at m/z 1118 could not be mass matched, in this region, with the crystallography data. This suggested that the amino acid sequence after residue 75 in the electron density map was incorrect. Another possible explanation could have been that the sequence of the derived hexapeptide was not related to the peptide $[M+H]^+$ 1118. However, as the N-terminal amino acid of this peptide at m/z 1118 had been proved, as described above, to be leucine/isoleucine agreeing with the above sequence this latter possibility appeared unlikely. The inference that the amino acid sequence after residue 75, in the electron density map, was

incorrect was shown later in this study to be true. The sequence of the peptide at m/z 1118 is reported below.

Interpretation of $[M+H]^+$ 736: Two steps of subtractive Edman degradation of the double digest mixture determined the N-terminal amino acid residues of this peptide as two adjacent aspartic acid residues, by the presence of signals m/z 621 and 506 in Figures 2.1.1 and 2.1.2 respectively. In the FAB-Map of the CNBr/tryptic/CpB digest mixture the signal m/z 736 was not observed but neither was a peak which corresponded to a loss of lysine or arginine from this peptide mass. This can probably be explained by supression effects, with either the peptide at m/z 736 not having been digested but suppressed by other components in the CNBr/tryptic/CpB digest mixture or that the newly formed peptide (C-terminal cleavage of the peptide at m/z 736) has been suppressed. A computer search (IA) of mass 736 in the electron density sequence did not provide any convincing sequence assignments. However, after purification of the double digest mixture by HPLC this peptide was isolated. An aliquot of the HPLC fraction containing the peptide $[M+H]^+$ 736 was examined by automated gas-phase sequencing which elucidated six residues, Thr-Asp-Gly-Tyr-Asp-Gly. This sequence was found to correspond to residues 298 to 303 in the electron density sequence data. However, no mass fit beyond residue 303, for the peptide mass 736, could be made. Gas-phase sequence analysis of the HPLC fraction containing the peptide at m/z 736 determined two other partial sequences (though weak), Leu-Ser and Trp-Ala-His. This finding indicated that the HPLC fraction was not pure containing possibly three or more

different peptides. However, as none of the derived sequences could be correlated with the Edman/MS data, for the peptide at m/z 736, it suggested that none were related to this peptide. The inability to mass match this signal with an amino acid sequence in the electron density map indicated that there were still sequence errors in the crystallographic data.

Interpretation of $[M+H]^+$ 1121: This mass value was not observed in the FAB-Map of the double digest mixture and only detected after purification of the digest mixture. As this signal was not observed in the original FAB-Map no N- or C-terminal information was available from the Edman or CpB digestion. The electron density sequence was screened for a mass fit for m/z 1121 however, no match with the amino acid sequence could be made. Automated gas-phase sequencing of an aliquot of the HPLC fraction containing the peptide at m/z 1121 yielded the sequence, Gly-Asp-Ile-Phe-Leu-Pro. This sequence was found to correspond to residues 188 to 193 in the electron density map. However, no mass correlation beyond residue 193 could be made for m/z 1121. This suggested that either (a) the amino acid sequence folowing residue 193 in the electron density map was incorrect, or (b) the sequence did not correspond to the peptide at m/z 1121 and that the HPLC fraction contained a mixture of peptides; this is answered later in the chapter.

The primary sequence of D-xylose isomerase determined by a combination of mass spectrometric and classical sequencing techniques is presented in Figure 2.2, Sequence A, with the

```
1    SVQPTPADHFTFGLWTVGWNGSDPFGTATRASLDPVEATQSLAELGSYGITFHDDDLVP   A
1    SVQPTPADHFTFGLWAIGLLADPFGDATRPLLTPVAAQALAAAGAYGVTFHDGDLVPFS   B
                    ** * * **** **** ******   * **********  ***

60   FGATAAERSSILGDFNQALADTGDIVVMVTINLFSHPVFKDGGFTSNDRAIRRFALGKV   A
60   GAAEREKDIEWQGPARALAATLTAAPMIGINLFKHGGFVATGTAGLGRYALAESIGTIS   B
     *  * ***   * ****** ***  * ** ********** ******* ***** ***

119  LHNIDLAAEMGASTYVMWGGREGSEWDGSKDIAAALDRMREGVDTAAGYIKSQGYDLRL   A
119  LASAEGAYAFRGRMGAIFDAAHAADAAVTEYASSDVAAKLHSSAKVARLAIEPKPNEPR   B
      ***** ***** * ********** ******************** ************

178  AIEPKPNEPRGDIFLPTVGHGLAFIEQLAEGDIVGLNPEVGHEQMAGLNFIHGIAQALW   A
178  GDIMLPTVGRALAFSERTSLDAAGVNPEVGHSQMAGLNFSHGIAQALWAGKLFHIDLSG   B
     ***** *** *** ** ***** *****  ** ****** ***** ************

237  AGKLFHIDLNGQRGIKYDQDLITFGEGDLITSAFWISDLLEDGFPDGGPKYTGPRHFDYHP   A
237  NKGIKYDQDAARGVAGDFQAAFWLVDLLESGYPSGGPAFAGPRFLDFDGVWPSAEGARM   B
     ****************** ** ** ** * ** * ****  ** ******* ***** *

296  PRTDGYDGVWEDAIACMIMYLILKARSDAFRANPEVQTAESTSGIFQLGQTTHSSGGTA   A
296  AAGLMANYVIVKGRANPQVAVLSASAGQAVAGPMADAKTATVAAAAAGPKADAGFPAFD   B
      ******* ***** ************* ******** *  *** * ** **      *

355  GSLESSGGSHSGHDAKAAADRLWAFIRLDMLAIEELLGPT                     A
355  GAAASKTRKIWVSDVGIMALVGPA                                     B
       *     *  *  * **  *
```

Figure 2.2 Sequence (A) is the Mass Spectrometric/Electron Density Determined Primary Sequence of D-Xylose Isomerase; Sequence (B) is the Initial Electron Density Map, Compiled from Strain Homology and Crystallographic Studies, of D-Xylose Isomerase.

Key: Underlined residues are those sequenced by MS and gas-phase sequencing, and residues in sequence (B) marked with * are those which were changed on interpretation of the MS data.

stretches of amino acid residues elucidated by these methods underlined. Sequence B is the original electron density map of D-xylose isomerase derived from the crystallography data and sequence homology studies between different bacterial strains of the enzyme, as discussed earlier. It is interesting to note that this sequence was incomplete with the C-terminal region missing. This tentative sequence was constantly under revision on receipt of data from this mass spectrometric study, which allowed correct interpretation of numerous amino acid residues. It can be observed that this sequence (B) contains an abundance of alanine residues. Alanine, in general, was assigned to positions in the electron density map where the crystallographic data was uninterpretable. From the data presented here it is apparent that major changes have been made to this tentative sequence. The inability to assign a number of peptide masses on to the sequence, e.g. m/z 736, 1118 and 1121, suggested that there were still some errors in the electron density map in the regions not highlighted.

Only in very few modern protein chemical studies of proteins the size of D-xylose isomerase or larger would a complete sequence be required. Emphasis lies on the provision of 6 to 12 residues of sequence, to employ for the production of synthetic complementary oligonucleotide probes to screen a gene library. FAB-Mapping procedures supplemented by automated gas phase sequencing techniques, as described above, have provided an excess of such stretches of sequence. This data was supplied to our colleagues in the Department for Biotechnology, to aid the generation of the gene sequence of D-xylose isomerase. However, as discussed previously,

the nucleotide sequence is of limited use without information on the nature of the expressed protein, i.e. if it is proteolytically processed, or there are unrecognised post-translational modifications such as glcosylation. Therefore nucleotide sequencing techniques must be complemented by the use of classical and non-classical protein sequencing techniques. The correct FAB-Mapping strategy was employed to determine the fidelity of the gene sequence which was generated and later corrected by utilisation of the sequence data elucidated above. The results of this study are reported below.

2.2.2 Section B: **Verification of The Gene Sequence**

When the gene-derived sequence, Figure 2.3, became available the FAB-Mapping strategy was applied to a second sample of D-xylose isomerase. This sample was initially digested with trypsin and then cyanogen bromide and a sample of the double digest mixture analysed by FAB-MS. It is interesting to note that by reversal of the previously employed digest conditions, a number of mass signals, corresponding to peptide fragments, are generated and observed in the FAB-Map of this double digest which were not found previously in the FAB-Map of the CNBr/tryptic digest, Figure 2.1. This illustrates the earlier supposition that the three dimensional conformation of a protein sterically hinders, to varying extents, the approach of an enzyme and hence, not all and/or different bonds, that were previously available for cleavage, are digested. The change in order of the digest conditions has in fact provided

1	ser	val	gln	pro	thr	pro	ala	asp	his	phe	thr	phe	gly	leu	trp	thr	val	gly	trp
20	thr	gly	ala	asp	pro	phe	gly	val	ala	thr	arg	lys	asn	leu	asp	pro	val	glu	ala
39	val	his	lys	leu	ala	glu	leu	gly	ser	tyr	gly	ile	thr	phe	his	asp	asn	asp	leu
58	ile	pro	phe	asp	ala	thr	glu	ala	glu	arg	glu	lys	ile	leu	gly	asp	phe	asn	gln
77	ala	leu	lys	asp	thr	gly	leu	lys	val	pro	met	val	thr	thr	asn	leu	phe	ser	his
96	pro	val	phe	lys	asp	gly	gly	phe	thr	ser	asn	asp	arg	ser	ile	arg	arg	phe	ala
115	leu	ala	lys	val	leu	his	asn	ile	asp	leu	ala	ala	glu	met	gly	ala	glu	thr	phe
134	val	met	trp	gly	gly	arg	glu	gly	ser	glu	tyr	asp	gly	ser	lys	asp	leu	ala	ala
153	ala	leu	asp	arg	met	arg	glu	gly	val	asp	thr	ala	ala	gly	tyr	ile	lys	asp	lys
172	gly	tyr	asn	leu	arg	ile	ala	leu	glu	pro	lys	pro	asn	glu	pro	arg	gly	asp	ile
191	phe	leu	pro	thr	val	gly	his	gly	leu	ala	phe	ile	glu	gln	leu	glu	his	gly	asp
210	ile	val	gly	leu	asn	pro	glu	thr	gly	his	glu	gln	met	ala	gly	leu	asn	phe	thr
229	his	gly	ile	ala	gln	ala	leu	trp	ala	glu	lys	leu	phe	his	ile	asp	leu	asn	gly
248	gln	arg	gly	ile	lys	tyr	asp	gln	asp	leu	val	phe	gly	his	gly	asp	leu	thr	ser
267	ala	phe	phe	thr	val	asp	leu	leu	glu	asn	gly	phe	pro	asn	gly	gly	pro	lys	tyr
286	thr	gly	pro	arg	his	phe	asp	tyr	lys	pro	ser	arg	thr	asp	gly	tyr	asp	gly	val
305	trp	asp	ser	ala	lys	ala	asn	met	ser	met	tyr	leu	leu	leu	lys	glu	arg	ala	leu
324	ala	phe	arg	ala	asp	pro	glu	val	gln	glu	ala	met	lys	thr	ser	gly	val	phe	glu
343	leu	gly	glu	thr	thr	leu	asn	ala	gly	glu	ser	ala	gly	asp	leu	met	asn	asp	ser
362	ala	ser	phe	ala	gly	phe	asp	ala	glu	ala	ala	ala	glu	arg	asn	phe	ala	phe	ile
381	arg	leu	asn	gln	leu	ala	ile	glu	his	leu	leu	gly	ser	arg					

Figure 2.3 The Gene-Derived Sequence of D-Xylose Isomerase, Arthrobacter strain B3728 (Hartley et al, unpublished results).

Highlighted amino acid sequences correspond to those residues mapped by the mass spectrometric study presented in this chapter (see Table 2.4.1)

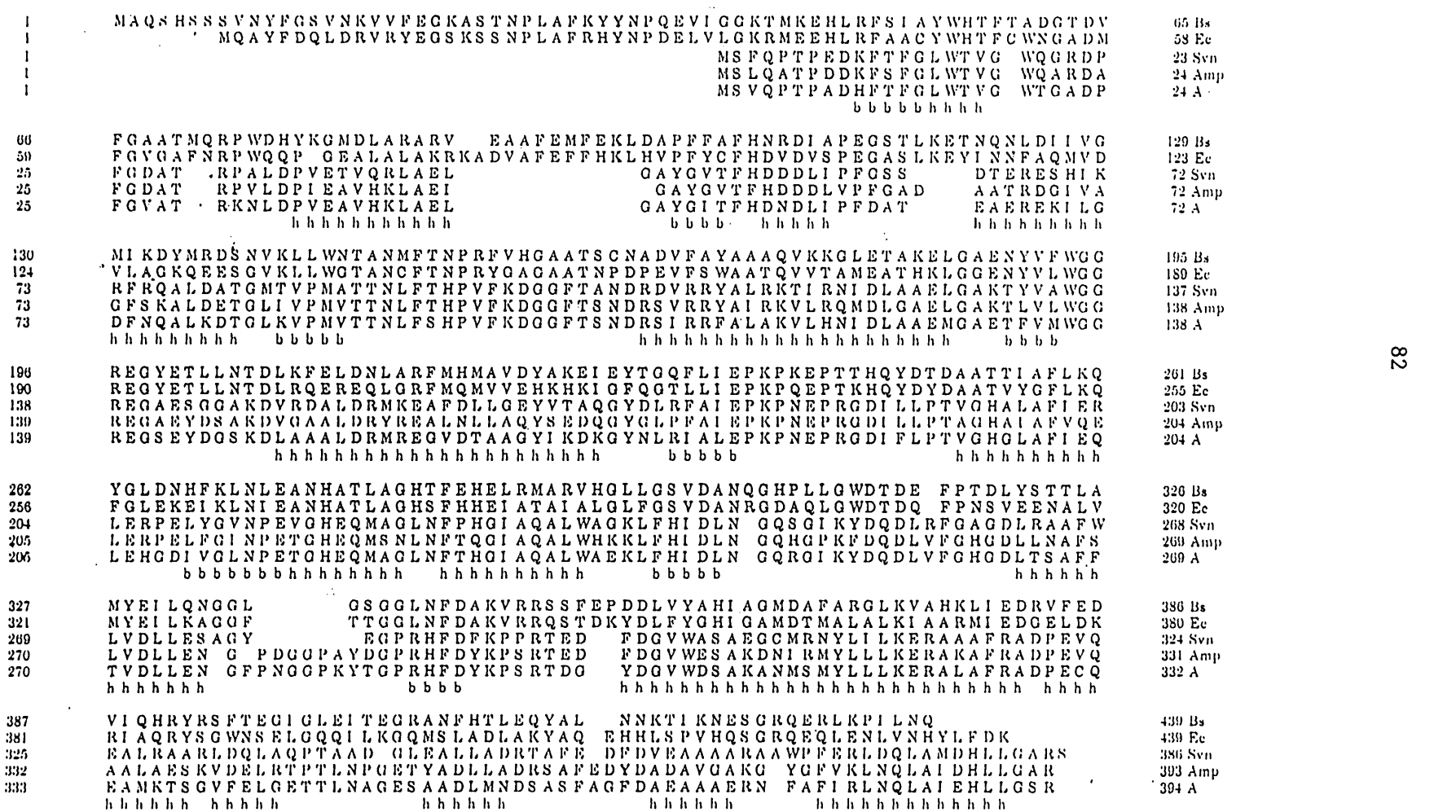


Figure 2.4 Sequence Alignment Between Various Bacterial Strains of D-Xylose Isomerase. Bs=Bacillus Subtilis (67), Ec=Escherichia coli (68), Svn=Streptomyces violaceoniger (69), Amp=Ampullariella sp. strain 3876 (70), A=Arthrobacter strain B3728 (Hartley et al, unpublished results). h=helices; b=beta-pleated sheet

additional data to that previously obtained, see Table 2.1. The intense quasi-molecular ions, $[M+H]^+$, observed in the FAB-Map of the tryptic/CNBr digest together with those signals which could be mass assigned on to stretches of amino acids in the gene-derived sequence are listed in Table 2.2. Since the FAB-Mapping strategy was employed to check the fidelity of the gene sequence N- and C-terminal amino acid analysis of each of the observed peptide fragments was unecessary. If the gene sequence was correct then the peptide fragments observed in the FAB-Map, theoretically, should be readily mass assigned on to the gene-derived sequence. However, N- and C-terminal amino acid determination methods such as Edman degradation and carboxypeptidase B, respectively, are useful for confirmation of sequence assignments.

There were a number of peptide masses from the initial digest which could not be assigned to a stretch of amino acid sequence, for example m/z 1121 and m/z 1118, discussed in section A of this chapter. Utilising the computer programme available in the Department of Biochemistry, Imperial College, London, (I) the gene-derived sequence was loaded into the protein databank and the previously uninterpretable peptide masses were subsequently screened for a sensible mass fit on to this protein sequence. Listed in Table 2.3 are those signals observed in the initial FAB-Map (Figure 2.1) for which no previous sequence assignment could be made, due to errors in the electron density data, but which were able to be mapped on to the gene-derived sequence. Of particular importance was the peptide at m/z 1463 which was determined to correspond to the C-terminus, (382-394), of the

Intense Signals in tryptic/CNBr FAB-Map	Residue Assignment on to Gene-derived Sequence	Signals NOT Observed in CNBr/tryptic FAB Map (Fig. 2.1) (*)
314		
326		
398		
486		
533	(80-84)	*
549	(113-117)	*
577	(322-326)	*
593	(285-289)	*
599		*
622	(172-176)	*
659		*
670		
692		*
706	(129-135)	*
728/9		*
736		
767	(376-381)	*
828		
844	(149-156)	
872		*
974		
995/6		*
1055		
1118	(70-79)	
1121	(32-41)	*
1123		*
1197		*
1212	(240-249)	
1235		*
1249		*
1263	(177-187)	
1313	(298-309)	*
1389	(88-99)	
1463	(382-394)	
1534		*
1653		*
1728	(359-375)	*
1750		*
1756		*
1826	(223-239)	
2160	(42-67)	
2415/7		*
~2862		*

Table 2.2 Intense Quasi-molecular Ions, $[M+H]^+$, Observed in the FAB-Map of a Tryptic/CNBr Digest of D-Xylose Isomerase.

m/z	residue assignment
1121	(32-41)
1263	(177-187)
1118	(70-79)
649	(315-319)
1463	(382-394)

Table 2.3 Mass Signals Observed in the CNBr/tryptic FAB-Map able to be Assigned on to the Gene-Derived Sequence (Fig. 2.3).

m/z	residue assignment
533	(80-84)
549	(113-117)
577	(322-326)
593	(285-289)
622	(170-176)
706	(129-134)
767	(376-381)
1313	(298-309)
1728	(359-375)
1826	(223-239)

Table 2.4 Intense Quasi-Molecular Ions Observed in FAB-Map of tryptic/CNBr Digest of D-Xylose Isomerase (not observed in the CNBr/tryptic FAB-Map) which may be Assigned on to the Gene-derived Sequence.

gene-derived sequence of D-xylose isomerase. Often, due to pre-processing (naturally or otherwise), or difficulties in interpreting the DNA-sequence, the C-terminus of a protein can be difficult to characterise. Hence, the identification of the C-terminus of D-xylose isomerase was a highly significant contribution to the overall primary sequence investigation. This assignment was unable to be made previously due to incorrect assignment of amino acid residues by crystallographic techniques. There were four positions of error in the sequence, three of which, 385, 390 and 394, once corrected were observed to be homologous with the primary sequence of bacterial strains *Streptomyces violaceoniger* and *Ampullariella sp* (Figure 2.4).

Previous analysis of the purified CNBr/tryptic digest mixture had identified the peptide at m/z 1121. Automated gas-phase sequencing of an aliquot of the HPLC fraction containing this peptide had determined the sequence, Gly-Asp-Ile-Phe-Leu-Pro, with this sequence mapping on to residues 188 to 193. However, as reported earlier, no mass fit for peptide $[M+H]^+$ 1121 could be made in this region. This led to the question as to whether this sequence corresponded to the peptide at m/z 1121 or some other component not observed on FAB-MS analysis due to suppression effects. Generation of the gene sequence allowed clarification of this sequence assignment. Screening the gene-derived sequence for mass 1121 mapped this peptide on to the sequence between residues 32 and 41, Asn-Leu-Asp-Pro-Val-Glu-Ala-Val-His-Lys. This assignment could not previously be made due to five residues having been incorrectly assigned in the electron density map in this region, see Figure

2.3. The sequence of the hexapeptide, 188 to 193, elucidated by automated gas-phase was also found in the gene-derived protein sequence. The fragment expected to be generated from this region (188-193) after a CNBr/trypsin cleavage is a peptide at m/z 3739. It can be therefore concluded that the HPLC fraction containing the peptide at m/z 1121 was not pure.

Previous investigations of the peptide at m/z 1118 had determined a partial sequence between residues 70 and 75 in the electron density map. The mass spectrometric data was instrumental in revising the sequence in the electron density map however, even with these corrections no sensible mass fit could be made beyond residue 75. This indicated that either there were further errors in the electron density map after residue 75, or that the HPLC fraction contained another peptide fragment. However, the former possibilty was inferred to be correct as the Edman/MS data of the peptide at m/z 1118 was in agreement with the derived sequence between residues 70 and 75. The gene-derived sequence was screened for mass 1118. A mass match was made on to amino acid residues between 70 to 79 with the sequence, Ile-Leu-Gly-Asp-Phe-Asn-Gln-Ala-Leu-Lys. Therefore, this data proved that the residues following Asn-75 in the electron density map were incorrect as suggested earlier.

Previously no sequence assignment could be made for the peptide $[M+H]^+$ 1930, discussed in section A. Automated gas-phase sequencing of the HPLC fraction containing this peptide yielded a sequence which was found to map on to the primary sequence of the bacterial strain *Streptomyces violaceoniger* (Svn) of D-xylose

isomerase between residues 42 and 46. As described above the electron density data was changed in this region to agree with both the automated Edman and Svn strain sequence. However, no mass assignment of this peptide at m/z 1930 could be made in this region. The generation of the gene-derived sequence showed residues 42 to 46 to agree the automated Edman data. The peptide expected to be generated by a CNBr/tryptic digest of this protein would be from residues 42 to 67, $[M+H]^+$ 2160. This mass signal, m/z 2160, was in fact observed in the CNBr/tryptic FAB-Map, Figure 2.1, but could not be previously assigned due to numerous errors in residue assignment by crystallographic methods in the electron density map. Therefore, it is concluded that the HPLC fraction contained a mixture of peptides and the sequence derived by automated gas-phase is not related to the peptide at m/z 1930. Screening the gene-derived sequence for mass 1930 did not locate a peptide sequence corresponding to both the mass value and the digest specificity.

Another ambiguity arose in the assignment of m/z 736, with the manual Edman degradation data determing the first two N-terminal residues of this peptide as two adjacent aspartic acid residues. However, this data conflicted with the sequence derived from automated gas-phase sequencing, Thr-Asp-Gly-Tyr-Asp-Gly, mapping on to residues 298 to 303 in the primary sequence of D-xylose isomerase. The N-terminal residues of this peptide do not correlate with those determined by manual Edman degradation/MS with no mass fit for this peptide at m/z 736 possible in this region. This finding again suggested that there was incomplete separation of the

digest mixture by HPLC and that this sequence corresponded to some other component present in the HPLC fraction. This was supported by the elucidation of two other partial sequences (though very weak) by automated gas phase sequencing. Utilising the computer search programme (I) the gene-derived sequence was screened for mass 736. No mass match on to the protein sequence could be made. However, the predicted peptide fragment derived from the region in the protein sequence, partially elucidated by MS/gas-phase techniques, (298-303) would be one formed from residues 298 to 309, $[M+H]^+$ 1313. This peptide is observed in the FAB-Map of the tryptic/CNBr digest of D-xylose isomerase, see Table 2.2.

The inability to mass assign peptides m/z 736 and 1930 indicated that there were still some errors in the primary sequence, either an error in the gene-derived sequence or a post-translational modification of the gene sequence e.g. glycosylation. The residues able to be assigned from FAB-MS analysis of the tryptic/CNBr digest are listed in Table 2.4 (see Page 85), these mass signals were not previously observed in the FAB-Map of the CNBr/tryptic digest mixture, i.e. reversal of the digest order generated extra data. Together, with the already determined N-terminal sequence from automated gas-phase sequencing and the data from both FAB-Maps 55% of the gene-derived sequence was confirmed. A list of all sequence assignments is presented in Table 2.4.1. A comparison of the mass spectrometric/electron density determined primary sequence of D-xylose isomerase, Figure 2.2, and the gene-derived sequence later produced, Figure 2.3. showed that there is a large conservation of sequence, approximately 80%. However, the nucleotide sequence was

Quasi-Molecular ions $[M+H]^+$	Residue Assignment on to Gene-derived Sequence
533	(80-84)
549	(113-117)
577	(322-326
593	(285-289)
622	(172-176)
649	(315-319)
706	(129-135)
767	(376-381)
844	(149-156)
1118	(70-79)
1121	(32-41)
1177/1195	(118-128)
1212	(240-249)
1263	(177-187)
1279	(159-169)
1313	(298-309)
1389	(88-99)
1463	(382-394)
1728	(359-375)
1826	(223-239)
2160	(42-67)

Table 2.4.1 Combined FAB-Mapping data from both digests, i.e. CNBr/tryptic and tryptic/CNBr, for which residue assignment on to the gene-derived sequence could be made.

incomplete with several positions of ambiguity. No amino acid assignment could be made at residues; 46, 47, 175, 176, 195, 196, 356 and 357 in the gene-derived sequence, due to an inability to translate the DNA bases in these positions. However, it is apparent from the information in Table 2.4.1 that some of these gaps in the gene-derived sequence were able to be assigned utilising the mass spectrometric/gas-phase studies. Residues 175 and 176 are encompassed by the peptide at m/z 622 and can be assigned as leucine and arginine respectively. Similarly residues 356 and 357 which are included in the peptide at m/z 1728 and thus, are determined to be glycine and aspartic acid respectively.

The primary sequence is of the utmost importance for understanding not only the secondary, tertiary and quaternary structures but also the physical and chemical properties of the molecule. Illustrated in Table 2.5 is a comparison of amino acid compositions of the *Ampullariella sp* strain 3876, *E.coli*, and the *Arthrobacter* strain B3728 of D-xylose isomerase. A feature common to all is the low cysteine content which is characteristic of thermostable proteins. Enzyme kinetic studies have shown that the *Ampullariella sp* strain of D-xylose isomerase retains its activity up to 75 Celcius for 24h and similarly the *E.Coli* strain possesses activity up to 60 Celcius [72]. The *Arthrobacter* strain of D-xylose isomerase is stable over a temperature range of 50-80 degrees [48]. The immobilised *Arthrobacter* enzyme serves as an excellent industrial D-xylose isomerase. The temperature used in the industrial process is limited to 60 degrees because of undesirable side reactions that

AMINO ACID	NO. OF EACH AMINO ACID RESIDUES		
	Amp	Ec	A
ALA	44	44	41
CYS	0	4	0
ASP	38	24	32
GLU	24	33	27
PHE	22	25	26
GLY	38	38	41
HIS	11	19	13
ILE	13	13	15
LYS	18	24	20
MET	4	12	9
ASN	13	20	18
PRO	21	14	16
GLN	13	16	16
ARG	21	19	18
SER	13	16	16
THR	18	19	25
VAL	21	23	16
TRP	5	8	5
TYR	12	18	9

Table 2.5 A Comparison of Amino Acid Composition of Different Bacterial Strains of D-Xylose Isomerase.

Amp = Ampullariella sp strain 3876 [70]

Ec = Escherichia coli [68]

A = Arthrobacter strain B3728 (Hartley et al, unpublished results)

take place at high temperature and alkaline pH. If the pH optimum could be lowered below 6.5 and thermostability retained, more rapid an equilibrium could be attained at higher temperature. More over the isomerase equilibrium [73] gives 53.5% fructose at 60 degrees but 58.6% fructose at 80 degrees. Additional benefits would be reduced viscosity and elimination of pH adjustments in the syrups used. Therefore, the knowledge of the primary structure is vital for further studies to propose targets for site directed mutagenesis. The elucidation of the primary sequence of D-xylose isomerase, *Arthrobacter* strain B3728, is industrially of the utmost importance. FAB-Mapping procedures have played a vital role in determining the amino acid sequence. The FAB-Mapping strategy employed here was crucial to the interpretation of the crystal structure, previously uninterpretable, with the data provided from these studies also supplied for aiding the production of the gene sequence. There were however, difficulties in obtaining an independant partial amino acid sequence by interpretation of the gene sequence. Two regions of this inferred amino acid sequence were infact corrected from the data generated by this mass spectrometric study. Furthermore, a debate arose over the nature of a residue in the C-terminal region of the D-xylose isomerase enzyme sequence which was clearly assigned as arginine by the mass spectrometric and crystallographic data however, this residue was inferred from the gene sequence as alanine. Re-examination of the gene sequence experimental work showed the mass spectrometric data to be correct. Thus, this mass spectrometric study has played a key role not only for providing data for use for the production of the gene sequence but also for correcting and finally checking the

fidelity of the gene-derived sequence.

PART II

2.3 Section C: STRUCTURE-ACTIVITY relationships of D-xylose isomerase, Arthrobacter strain B3728

2.3.1 Introduction

The X-ray crystallographic studies [66], which utilised the data and interpretation presented in this chapter, Part I, subsequently demonstrated that the molecule is a tetramer comprised of four identical subunits. The secondary structure of each subunit is composed of two structural domains; the main domain contains an inner cylinder of eight strands of parallel beta-pleated sheet. Adjacent strands are connected by helical segments running anti-parallel to the strands of sheet. The C-terminal domain is a loop structure consisting of five helical segments and is involved in intermolecular contacts that comprise the tetramer. Hendrick et al [66] also proposed that a number of salt-bridges are responsible for stabilising the tertiary and quaternary structure. The active site has been found to lie in a deep pocket near the C-terminal ends of the beta-strands of the barrel domain including residues from from a second subunit hence, the active site is composed of residues from both subunits. Shown in Figure 2.5 is a schematic representation of the subunit structure [74]. The enzyme is dependant on a divalent cation, Mg^{2+}, for catalytic activity, with each monomer requiring two metal ions.

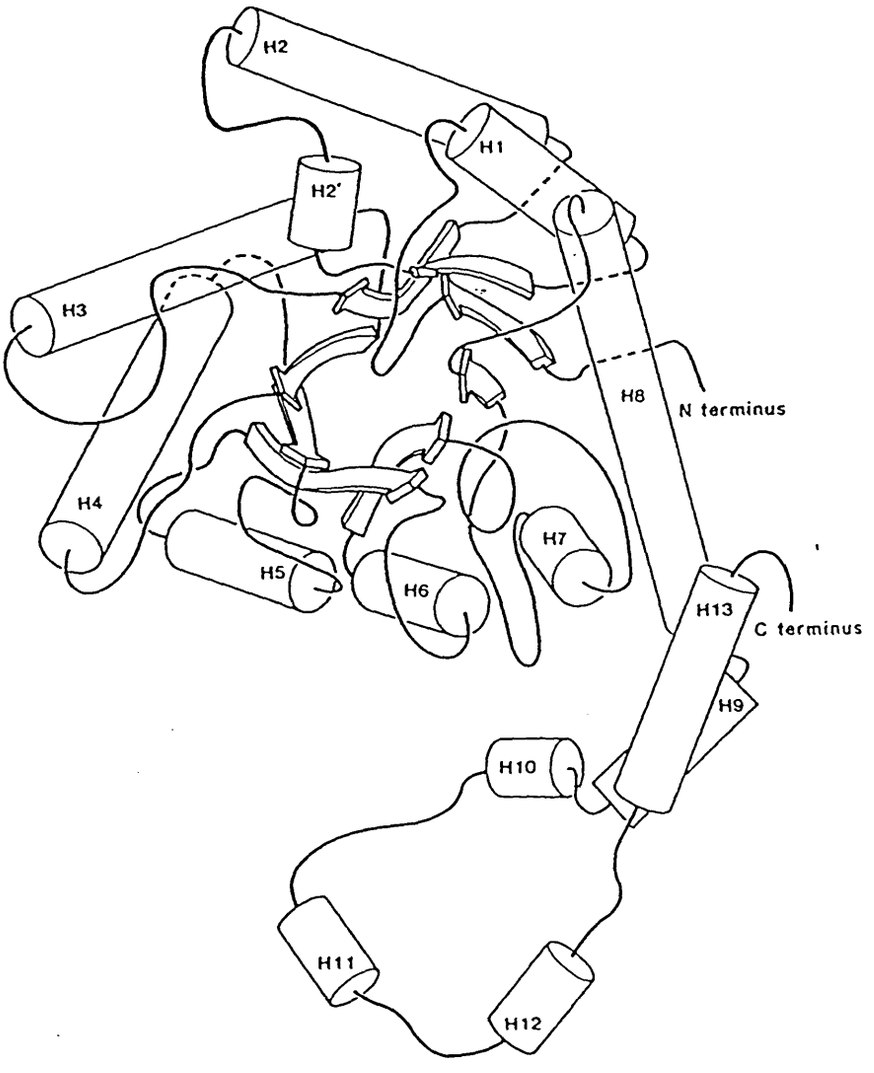


Figure 2.5 D-Xylose Isomerase Subunit Structure; Helices are Represented as Cylinders and Strands as Twisted Arrows (74).

Proteins in general and in particular globular proteins, often prove resistant to proteolysis. However, many proteins possess one or more peptide bonds which are unusually susceptible to particular proteases. An example of limited proteolysis is the conversion of the inactive zymogen pepsinogen to the active enzyme pepsin, which occurs by the cleavage of specific peptide bonds. In other proteins these sensitive peptide bonds may link globular domains together or simply lie on the surface of a tightly folded molecule. Occasionally the amino or carboxyl termini may be accessible and limited proteolysis may merely result in a shortening of the polypeptide chain by a few residues. Limited proteolysis of D-xylose isomerase with thermolysin resulted in a mass loss of approximately 6000 Da from the D-xylose isomerase molecule. The important aspect of this limited proteolysis experiment was that the enzyme was reported, after proteolysis, to retain full enzymic activity (Hartley *et al*, personal communication, 1989).

The sample, derived from limited proteolysis of native D-xylose isomerase, studied here was supplied by the Department for Biotechnology, Imperial College, London.

2.3.2 Results and Discussion

As discussed above it had been determined that after limited proteolysis, of D-xylose isomerase, with thermolysin for 120 hours at 45°C, full enzymic activity was reported. Analysis of the

thermolytic product by SDS gel-electrophoresis showed a band at approximately 37.5 kDa with native D-xylose isomerase (43.5 kDa) used as a standard. It appeared that this "nicking" experiment had cleaved off approximately 6,000 Da. A sample of this thermolytic species (37.5 kDa) was examined by automated gas-phase sequencing in this group. A sequence corresponding to the N-terminus of native D-xylose isomerase was yielded by this method. Thus, it was concluded that the thermolysin digestion had not cleaved the N-terminus. As enzymic activity was retained, despite an apparent loss of approximately 6000 Da from the protein, it was important to determine where this 6 kDa had been cleaved from in the protein sequence. Reported below is the elucidation of the thermolysin cleavage site(s) utilising mass spectrometric techniques.

A suitable strategy was devised to determine the site(s) of thermolysin cleavage of D-xylose isomerase. This approach involved the comparison of those peptide fragments derived from a digest of native D-xylose isomerase and those formed from an identical digest of a sample of the thermolytic species (37.5 kDa). This would generate a mixture of peptides from each sample which should be identical however, the 37.5 kDa sample would be expected to contain less fragments, with those "lost" peptide fragments corresponding to the 6 kDa cleaved by thermolysin. A sample of both the native and 37.5 kDa D-xylose isomerase was chemically cleaved with cyanogen bromide (CNBr). This type of digest was specifically selected due to the location of the methionine residues in the primary sequence, the masses of the peptide fragments generated and the resolvable mass range of the mass spectrometer used (ZAB-HF).

Predicted Fragments	Expected $[M+H]^+$
1 - 87	>
88 - 128	>
129 - 135	706/724
136 - 158	2335/2553
159 - 222	>
223 - 313	>
316 - 336	2414/2432
337 - 359	2278/2296
360 - 394	>

Table 2.6 CNBr Peptide Fragments Predicted to be Generated by CNBr Cleavage of D-Xylose Isomerase.

Key: (>) Indicates the mass of the peptide fragment lies outside the mass range (3300 Da) of the mass spectrometer at full accelerating voltage.

The predicted CNBr peptide fragments, from a CNBr digest are listed in Table 2.6. From this list it can be seen that there are four peptides which fall within the resolvable mass range of the mass spectrometer at full accelerating voltages. It is interesting to note that cleavage at Met-336 with residues 336 to 394 removed would produce a molecule of mass approximately 37.5 kDa, (the approximate molecular weight of the thermolytic species determined by SDS-gel electrophoresis). However, the number and position of the peptide bonds cleaved by thermolysin had yet to be determined. Each of the CNBr digest mixtures was purified by HPLC, using identical elution gradients. The chromatographs of the CNBr digest of the thermolytic product and native D-xylose isomerase are presented in Figures 2.6 and 2.6.1 respectively. Each major peak in the chromatograph of the CNBr digest of the thermolytic product, Figure 2.6, was analysed by FAB-MS, identifying peptide fragments at **m/z 706 (129-135), m/z 2335 (136-158), and m/z 2414 (316-336).** All these peptide masses were calculated to correspond to the lactone derivatives of methionine residues of the peptide fragments formed by CNBr cleavage. However, an unpredicted intense signal, m/z 1268, was identified by FAB-MS analysis of HPLC fraction number 29. The computer search programme (I) gave six sequence possibilities for $[M+H]^+$ 1268. However, only one of these sequences was determined to arise from realistic cleavages between residues 336 to 347, on the basis of this species being a thermolytic product digested by CNBr. The sequence between these residues is, Lys-Thr-Ser-Gly-Val-Phe-Glu-Leu-Gly-Glu-Thr-Thr; the amino acid preceding the N-terminal lysine of this peptide, from the

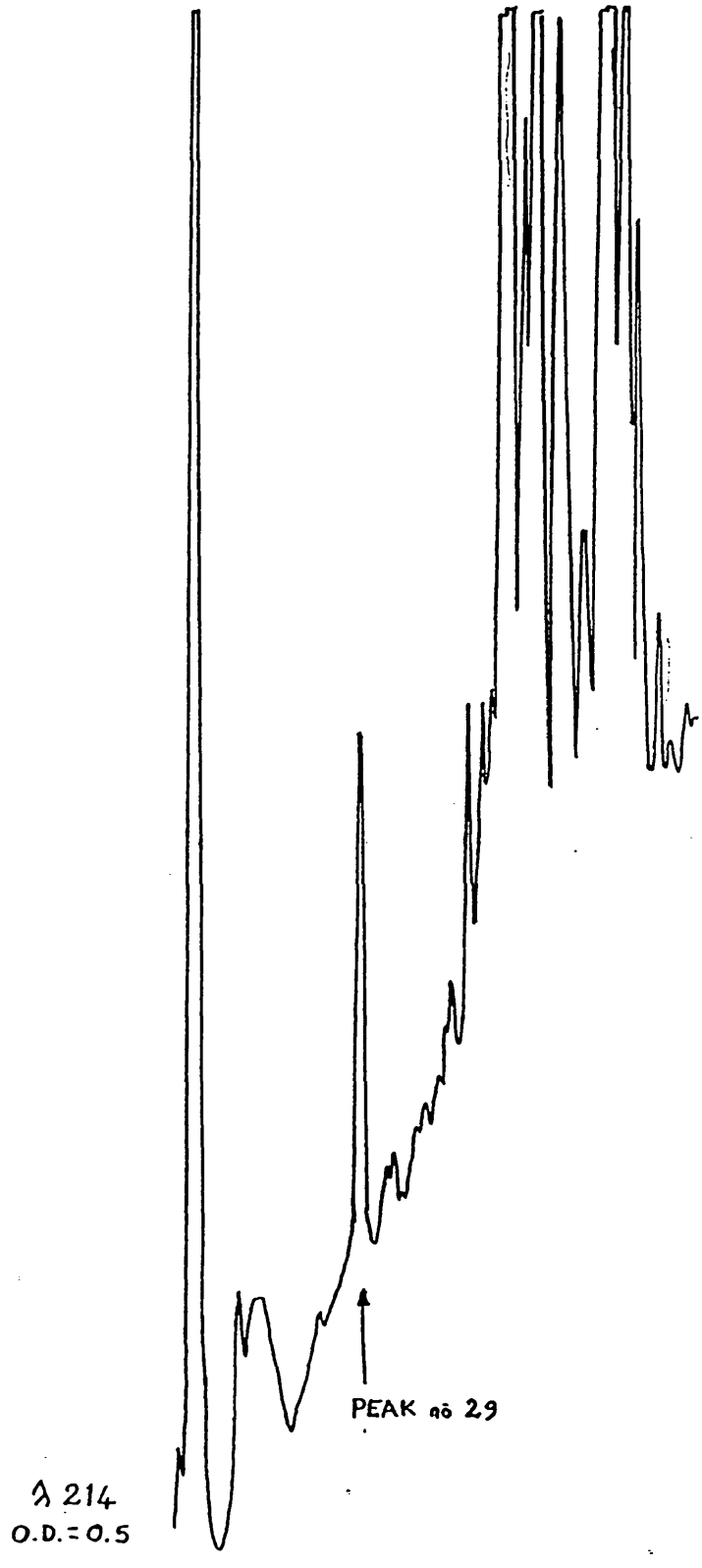


Figure 2.6 HPLC Chromatograph of a CNBr Digest Mixture of the Thermolytic Product ("37.5" kDa species) of D-Xylose Isomerase.

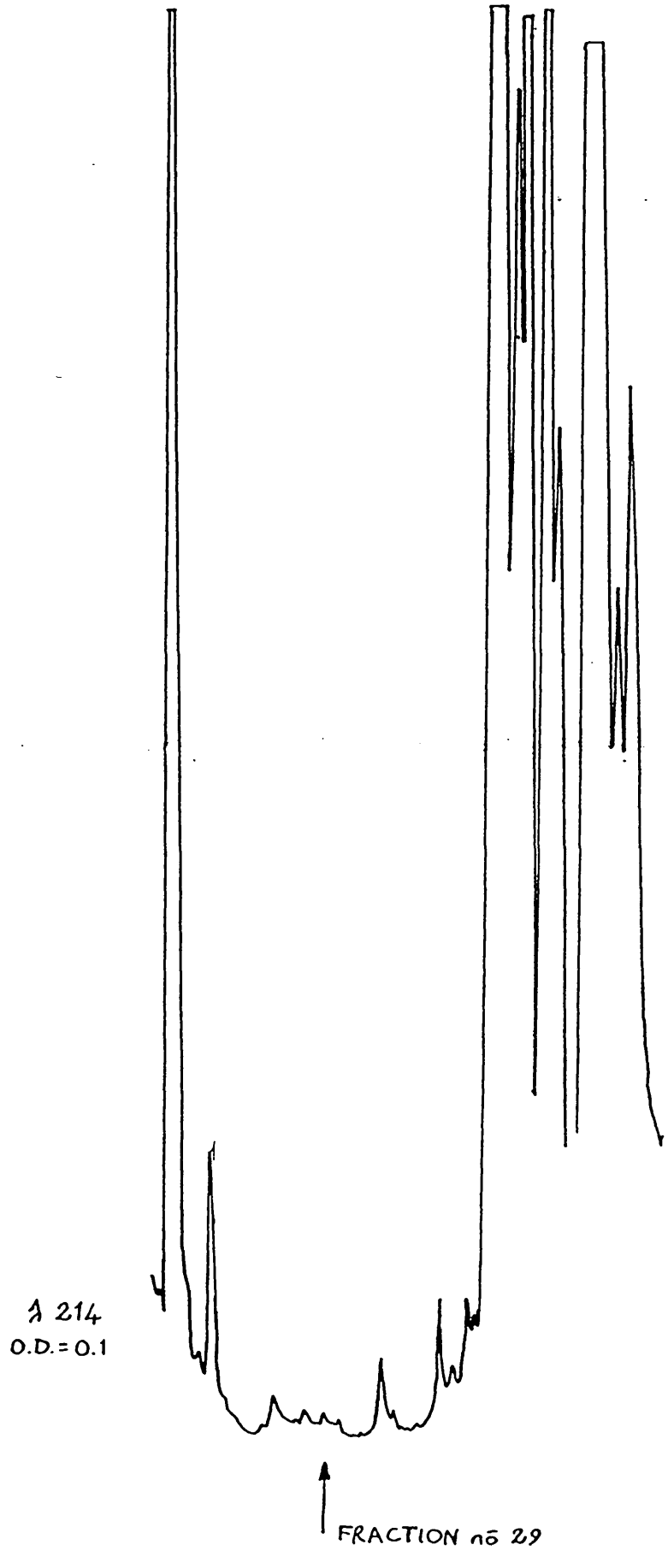


Figure 2.6.1 HPLC Chromatograph of a CNBr Digest Mixture of Native D-Xylose Isomerase.

gene-derived sequence, is methionine, a CNBr cleavage site. This sequence assignment was confirmed by a step of subtractive Edman degradation after which a new signal m/z 1140, was observed corresponding to an 128 a.m.u. mass loss (lysine) from m/z 1268. The other peptide bond necessary to be cleaved for the generation of this peptide is between Thr-347 and Leu-348, which corresponds to a thermolytic cleavage site.

Screening of the HPLC fractions of the purified digest of native D-xylose isomerase, by FAB-MS, did not identify this peptide, with no significant peak in the HPLC chromatograph in the region of the 29th fraction as detected above. It was concluded that this "nicking" experiment with thermolysin had cleaved between residues Thr-347 and Leu-348. This series of experiments elucidated the cleavage site and allowed determination of the molecular weight of the "37.5 kDa" species as 38,138 Da (based on the gene-derived sequence). Hence, with knowledge of this cleavage site it was possible to relate this to the secondary, tertiary and quaternary structure. This prelimary study has indicated that the C-terminal loop has been partially cleaved, with this species, 38,138 Da, reported to retain full enzymic activity. This finding is most interesting as it has been proposed that the C-terminal domain, Figure 2.5, consisting of five helical segments, is involved in intermolecular contacts between the subunits that comprise the tetramer. This data suggests that not all the C-terminal portion is necessary for enzyme activity. Hendrick *et al* [66] proposed that helices H10 (340-344) and H11 (354-359), present in the C-terminal arm, are reponsible for locking the tetramer molecule together. It

is possible to infer from this study that the helix H11 is not necessary for enzyme activity, and therefore, does not play a significant role in locking the monomer units together.

2.4 Conclusion

Presented in this chapter is what is believed to be the first study in which medium resolution X-ray crystallographic data has been interpreted by utilising mass spectrometric data. The FAB-Mapping data has also been instrumental in providing numerous stretches of sequence data aiding the production of the gene sequence. Before the gene sequence became available 80% of the primary sequence of D-xylose isomerase was determined from mass spectrometric techniques and interpretation of the electron density map. The FAB-Mapping strategy was subsequently applied to determine the fidelity of the gene-derived sequence with this study confirming 55% of the primary sequence. Another important contribution of this mass spectrometric study was for the elucidation of four positions of ambiguity in the gene-derived sequence, i.e. bases which were not able to be translated from the DNA-sequence. However, signals such as m/z 736 and 1930 for which no sequence assignment could be made may possibly be explained by errors in the gene sequence and/or the presence of post-translational modifications. Evidence for glycosylation is still circumspect but if true it may account for those signals unassigned.

The second component of this study was an investigation of

structure-activity relationships. Limited proteolysis of D-xylose isomerase with thermolysin generated a product which was reported to possess full enzymic activity. Identification of the cleavage site(s) was necessary to relate activity to structure. A suitable strategy was developed which allowed the hydrolysed bond(s) to be determined by chemical and mass spectrometric techniques which revealed a **single** cleavage between residues 347 and 348, i.e. the C-terminal loop had been cleaved. This suggests that this region is not necessary for enzyme activity and hence, for subunit interactions forming the tetramer. This finding is of great relevance for relating secondary, tertiary and quaternary interactions to enzymic activity, and will also be of considerable interest from an industrial standpoint.

CHAPTER THREE

METHODOLOGY DEVELOPMENT I-PARTIAL ACID HYDROLYSIS FOR PEPTIDE SEQUENCING

3.1 Introduction

The key to fully understanding the function and mechanism of action of a molecule in a biological system is the knowledge of its structure. More specifically, in proteins a definition of the amino acid sequence is of primary importance. Proteins in which one or more amino acid substitutions have been introduced may be rendered partially or completely inactive. For example, a point mutation in antithrombin results in inactivity and impaired interaction of the variant antithrombin with thrombin [75]. This has been shown to be responsible for thrombosis in patients carrying the pathologically defective gene. This example clearly illustrates the importance of structure to function and, more precisely, in proteins the amino acid sequence.

The importance of protein structure has long been recognised and consequently methods for sequence analysis of proteins have been developed over several decades. One such widely used method for sequence analysis is partial dilute acid hydrolysis. Partial acid hydrolysis has been extensively employed in early sequence analysis studies, for example, Sanger in his Nobel prize winning sequence

determination of insulin used partial acid hydrolysis to degrade each of the performic-acid oxidised insulin chains to di- and tri-peptides which were then subjected to sequence analysis. Partial acid hydrolysis was the primary method for peptide chain cleavage used by Sanger, as it generated a large number of small overlapping peptide fragments that could be sequenced. However, the introduction of Edman degradation [2] made the sequence analysis of large peptide fragments much easier, and thus this application of partial acid hydrolysis for the generation of small fragments became less desirable. However, the relative lack of specificity limited the usefulness of this method. The lack of specificity of partial acid hydrolysis arises from the susceptibility of all peptide bonds in a protein molecule to hydrolysis by acid. The kinetics of peptide bond hydrolysis by dilute acid have been widely investigated with the rate of hydrolysis of a particular bond demonstrated to be dependant on various properties, including electrostatic and steric factors [36] which affect the accessibility of the bond to approach by protons. For a great number of years partial acid hydrolysis, using dilute mineral (dilute hydrochloric or sulphuric acid) or concentrated organic acids (glacial acetic acid), has been investigated for protocol development for the production of sequence data from peptides/proteins. From these past investigations it is apparent that partial dilute acid hydrolysis is not completely non-specific. For example, it has been found that on hydrolysis there is preferential release of free aspartic acid from proteins [37] and that the peptide bonds hydrolysed most rapidly are those involving the amino groups of serine and threonine [76]. Piskievicz *et al*

[38] also observed that Asp-Pro bonds were readily cleaved under mild dilute acidic conditions. This is a consequence of the relatively high basicity of the prolyl nitrogen atom coupled with neighbouring group catalysis by the aspartyl side-chain carboxyl group.

The recognition that some peptide bonds are more labile than others has proved particularly interesting especially with respect to membrane proteins where sequence determination methods, devised for water-soluble proteins, are rendered difficult due to the extreme hydrophobic nature of many of these proteins in aqueous solution [77,78]. Partial acid hydrolysis may be useful since particularly sensitive bonds involving asparagine, serine, and threonine may be infrequent in the hydrophobic sequences hence, large peptide fragments would be expected to be generated instead of a complex mixture of small peptides. Gerber and co-workers demonstrated, by their investigation of the primary structure of the membrane protein bacteriorhodopsin [79], that partial acid hydrolysis was suitable for application to membrane proteins for the production of peptide fragments. Separation and analysis of these fragments was achieved by gas chromatography/mass spectrometry.

During the course of this thesis work De Angelis and co-workers [80] published a study, combining partial dilute acid hydrolysis and FAB-MS for peptide sequencing. This method involves taking spectra of peptide acid hydrolysates generated over a range of hydrolysis times. The signals in the mass spectra correspond to peptide fragments formed on partial hydrolysis of the parent

molecule. Sequence determination by this method is shown to be a relatively lengthy exercise involving analysis of the sequence of the structurally overlapping fragments. In response to this published study of De Angelis and co-workers Biemann *et al* [81] made a number of comments regarding this application of partial dilute acid hydrolysis and the role of four sector instrumentation. Biemann's first criticism of De Angelis's method lay with the relatively large sample quantity required, approximately 5 nanomoles, for sequence determination. Other points were in reference to the number of spectra required to sequence one relatively small peptide combined with the total deamidation of asparagine, glutamine and blocked C-terminal amide residues. There was nothing novel in these criticisms since they are well recognised deficiences of dilute acid hydrolysis protocols.

The strategy developed in this thesis, in contrast, is demonstrated to be a sensitive method with respect to the sample quantity required for obtaining sequencing information from peptides, being comparable to that needed for a tandem mass spectrometric study. Similarly, for sequence analysis using this method only one FAB spectrum per peptide is necessary from which the peptide sequence may be determined. The deamidation of asparagine and glutamine residues is exploited in the methodology presented in this chapter; by controlling the extent of deamidation it is possible to identify the positions of susceptible residues, asparagine and glutamine, in the primary sequence. This method of partial acid hydrolysis has been applied to a wide range of peptides/proteins differing in composition and molecular weight. However, the key difference in

the strategy presented here compared to that not only of De Angelis but also to previous methodologies employing partial acid hydrolysis, is that the method described in this chapter involves hydrolysis with a **concentrated** mineral acid, 12M hydrochloric acid, with peptide acid hydrolysates directly analysed by FAB-MS. This method of acid hydrolysis as presented here represents an important contribution to peptide sequence analysis, possessing the key features necessary for application to biological samples, i.e. rapid generation of high quality sequence data at high sensitivity.

3.2 Results and Discussion

Acid hydrolysis has been widely implemented for structural analysis of proteins and peptides, as discussed above. The conditions of the hydrolysis procedure are often indicated by the nomenclature of the technique, for example, amino acid analysis is a method of acid hydrolysis. However, as implied, the application of this technique is for the determination of the amino acid composition of a protein, i.e. all peptide bonds are hydrolysed. For hydrolysis of all bonds the protein is usually incubated with 6M HCl for 24h at 110°C. The individual amino acids generated by this hydrolysis procedure are separated, quantified and recorded by an amino acid analyser. Hence, the main difference between partial acid hydrolysis and the conditions of hydrolysis for amino acid analysis is that only certain peptide bonds are hydrolysed, generating peptide fragments as opposed to individual amino acids from the parent molecule. The amino acid composition of a protein is the

first step towards characterisation of the chemical structure of the molecule. However, as previously discussed, it is the arrangement of these amino acids which lends a protein its functionality. Hence, as methods for separation, detection and analysis of peptide fragments were developed the technique of partial acid hydrolysis for the generation of small fragments amenable to sequence analysis became an attractive procedure, in certain strategies.

3.2.1 Optimum Conditions for Hydrolysis for Peptide Sequencing

The partial acid hydrolysis method presented in this thesis differs fundamentally to techniques developed prior to this study in that concentrated hydrochloric acid (12M HCl) is used, unlike others employing dilute mineral (1-3M) or concentrated organic acids. In a series of experiments designed to examine the variables of the hydrolysis kinetics, time, temperature, acid concentration and handling, a set of conditions were determined as optimum for generating a series of sequence related peptides, with those fragments not fitting into a series referred to as random. Neurotensin and human calcitonin gene related peptide (hCGRP) were employed as standards in these kinetic studies. The results of these experiments determined the conditions for maximising the generation of sequence data, to involve incubation of the peptide sample with an excess of 12M HCl at 65°C for forty-five minutes, with termination of the reaction by the addition of water, to aid freezing, followed by lyophilisation.

3.2.1.1 Partial Acid Hydrolysis of Neurotensin

Presented in Figure 3.1 is the FAB mass spectrum of acid hydrolysed neurotensin (500 picomoles), with hydrolysis under the optimum conditions determined above. The spectrum is characterised by intense signals at m/z; 245, 408, 505, 661, 817, 914, 1042, 1156, 1285, 1448, 1561, and 1672. The complete sequence of neurotensin, $[M+H]^+$ 1672, can be determined by the mass differences between each of these intense peaks which correspond to amino acid residue masses. The sequence of neurotensin was determined, from this mass spectrum, as

pyro-Glu-Leu-Tyr-Glu-**Asn**-Lys-Pro-Arg-Arg-Pro-Tyr-Ile-Leu, (it should be noted that Leu cannot routinely be distinguished from Ile by conventional mass spectrometry except when in a terminal sequence position). The difference between m/z 1672 and m/z 1561 is 111 a.m.u. which corresponds to a cyclised glutamic acid residue, pyro-glutamic acid, and so fixes the N-terminal amino acid. It is worth noting that classical sequencing techniques such as Edman degradation would not be able to generate a sequence from such a peptide due to the presence of the cyclised N-terminal residue, i.e. the N-terminus is blocked. The molecular ion $[M+H]^+$ 245 corresponds to the di-peptide, Ile-Leu. The structure of this di-peptide is supported by the presence of the signal m/z 132, in the FAB mass spectrum, corresponding to a single leucine residue. The signal m/z 1448 is much weaker than the other signals in the spectrum which are sequence related. This is to be expected since the peptide bond between leucine and tyrosine which must hydrolyse to produce this peptide signal is less susceptible to acid hydrolysis, for steric reasons, as both residues possess bulky side-chains. The sequence of neurotensin and the predicted N- and

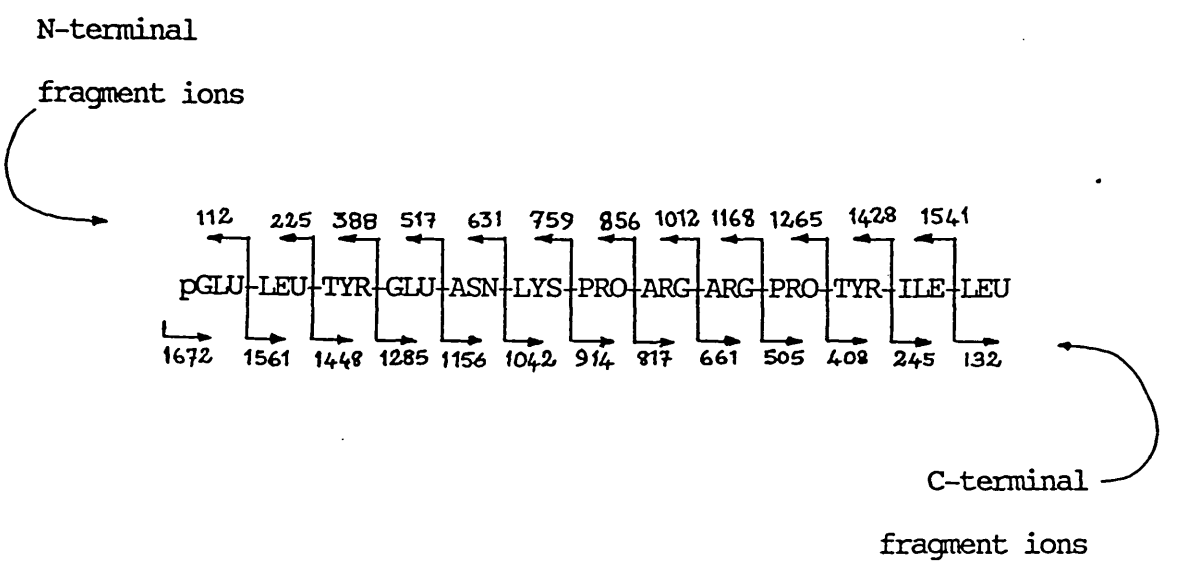


Figure 3.1.1 The Primary Sequence of Neurotensin, $[M+H]^+$ 1672

C-terminal fragment ions are depicted in Figure 3.1.1. It can be seen from the spectrum of acid hydrolysed neurotensin, Figure 3.1, that the most intense signal in each ion cluster after the signal at m/z 1042 is 1 a.m.u. higher in mass than the m/z values quoted above. This difference arises from partial deamidation of asparagine to aspartic acid. Experimental conditions have been devised so as to control the extent of deamidation and so that it is recognisable from the signal cluster shape. This is described fully in section 3.2.5.

3.2.1.2 **Partial Acid Hydrolysis of hCGRP**

Illustrated in Figure 3.2 is the FAB mass spectrum of acid hydrolysed human calcitonin gene related peptide (hCGRP, 500 picomoles), a 37 amino acid residue peptide, molecular weight 3786. The hydrolysis conditions employed were identical to those used to generate the FAB mass spectrum of acid hydrolysed neurotensin in Figure 3.1. Interpretation of the spectrum of the acid hydrolysate of hCGRP, Figure 3.2, determines that total sequencing of the molecule has not been achieved since insufficient sequence related signals are observed in the FAB mass spectrum. However, a stretch of sequence corresponding to residues 6 to 16 in the hCGRP molecule can be deduced from the mass differences between the intense signals at m/z; 245, 302, 373, 486, 642, 779, 880, 979, 1082 and 1183, from which the sequence;
Thr-Cys-Val-Thr-His-Arg-Leu-Ala-Gly-Leu-Leu, is determined. The complete sequence of hCGRP [35] is shown in Figure 3.2.1 with the peptide bonds hydrolysed by acid, that give rise to the sequence determined above, highlighted on the hCGRP sequence. An interesting

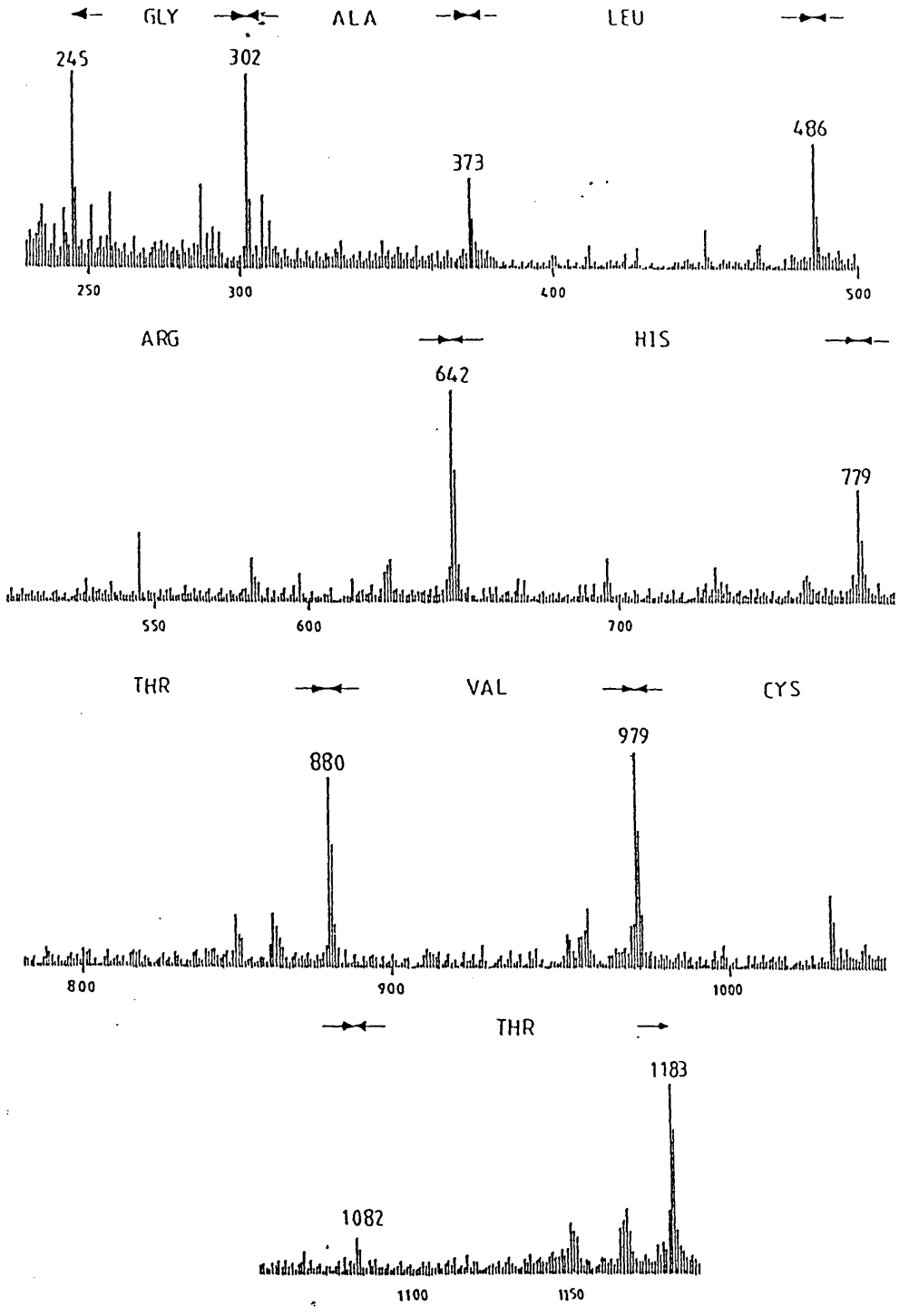
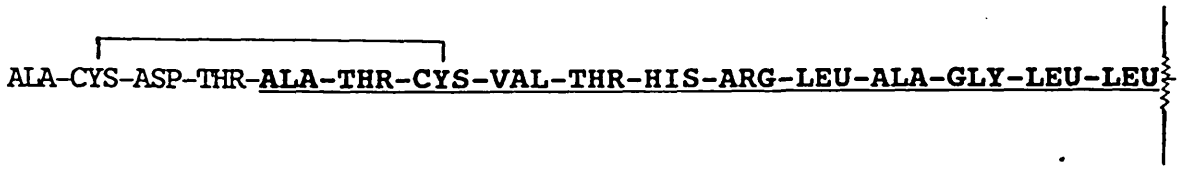


Figure 3.2 FAB-MS of Acid Hydrolysed human Calcitonin Gene Related Peptide (hCGRP), 500 picomoles.



SER-ARG-SER-GLY-GLY-VAL-VAL-LYS-ASN-ASN-PHE-VAL-PRO-THR-ASN-VAL-GLY-

SER-LYS-ALA-PHE

Figure 3.2.1 The Primary Sequence of human Calcitonin Gene Related Peptide (hCGRP), $[M+H]^+$ 3787.

The highlighted amino acids correspond to the residues sequenced by partial **concentrated** acid hydrolysis/MS.

feature of the partial acid hydrolysis of hCGRP is that the disulphide bridge between cysteine residues 2 and 7 in the sequence, despite acidic conditions which may be expected to preserve disulphide bridges, appears to have been reduced. Reported later is this chapter is an investigation to explain this observation, determining whether destruction of disulphide bridging by partial acid hydrolysis is common to all disulphide containing peptides.

One possible drawback of this partial acid hydrolysis method, realised by the analysis of the FAB mass spectrum of acid hydrolysed hCGRP, is that if the sequence of hCGRP were unknown then it would not be possible to determine from the FAB spectrum whether the deduced sequence ran in a C-terminal to N-terminal direction or vice versa. Peptides such as neurotensin pose no such problem since, the presence of the blocked N-terminal amino acid residue establishes the direction of sequence. However, if a blocking group was absent then this potential problem could be resolved for peptides, from which total sequencing was possible, by one step of manual Edman degradation of the native sample observing the mass shift mass spectrometrically; the mass shift corresponding to the N-terminal amino acid residue cleaved.

3.2.2 Investigation of Peptide Fragment Formation

The next step in this study was to investigate the generation of these sequence related peptide fragments, which are apparently formed in preference to other possible and anticipated random fragments. It was necessary therefore, to determine the mechanism

of formation of these peptide fragments, i.e. if they are generated by fragmentation in the mass spectrometer or by acid treatment of the native peptide under investigation. Peptide fragmentation may also result from a combination of both techniques. A preliminary investigation involved a comparative study of an equal quantity of intact and hydrolysed sample by FAB-MS under identical conditions. Figure 3.3 shows the mass spectrum of intact neurotensin. The aim of this experiment was to demonstrate that the fragment ions are not generated solely by a mass spectrometric event. As expected the most intense signal in the spectrum of intact neurotensin, Fig. 3.3, is the molecular ion, $[M+H]^+$, at m/z 1672. This spectrum also contains various other signals at, m/z; 1446, 1140, 913/5, 836.5 and 661. Interpretation of this FAB mass spectrum determined that the signal m/z 1446 corresponded to the molecular ion (1-11) formed by peptide bond fragmentation of the neurotensin molecule. Similarly, m/z 661 was assigned as a C-terminal fragment ion generated from amino acid residues (9-13), m/z 1140 as an N-terminal aldimine fragment ion (1-9) and the signal cluster at m/z 836.5 as an intact neurotensin molecule carrying two positive charges. In accordance with the equation of motion of an ion deflected through a two sector mass spectrometer (see Figure 1.3, Chapter One) a molecule of neurotensin carrying two positive charges, $[M+2H]^{2+}$, is observed at half its molecular weight, plus hydrogen. The signal cluster at m/z 836.5 is comprised of peaks separated by half mass units. The signals in the spectrum (Fig. 3.3) are interpreted to be N- and C- terminal fragment ions generated mass spectrometrically. The proposed mechanisms for N- and C-terminal fragment ion formation [27] are described in Chapter

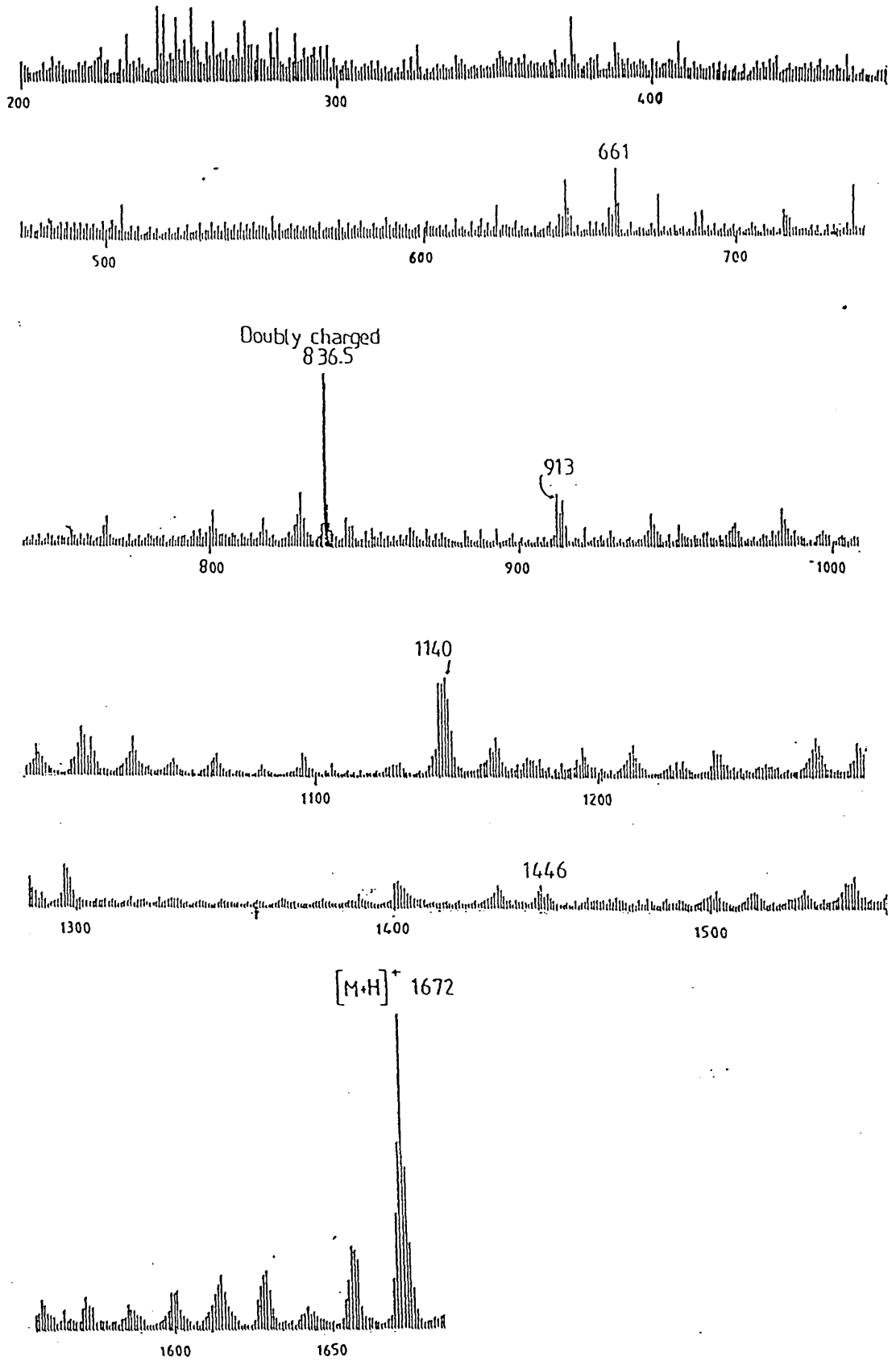


Figure 3.3 FAB-MS of Native Neurotensin (500 picomoles)

One of this thesis. Peptide fragmentation by FAB-MS is unpredictable and varies with every peptide examined. One of the disadvantages of FAB-MS as opposed to EI-MS for peptide sequence analysis studies, is the relative lack of fragmentation observed in FAB mass spectra. However, the positive features of FAB-MS such as sensitivity and the need for derivatisation of peptides removed, outweigh this lack of fragmentation which in some cases proves to be an advantage e.g. FAB-Mapping [41]. Thus, complementary techniques to FAB-MS which enhance or generate fragment ions are invaluable for sequencing studies. On comparison of the mass spectra of intact and hydrolysed neurotensin it is clear that despite some signals being common to both spectra (Figures 3.1 and 3.3), acid hydrolysed samples invariably produce intense and easily interpretable sequence related ions, on FAB-MS analysis, allowing sequence determination. Thus, it is reasonable to conclude that the intense signals in the spectrum, Figure 3.1, of acid hydrolysed neurotensin are not fully attributable to the fragmentation processes in the ion source of the mass spectrometer. Therefore, it was established from these experiments that the partial acid hydrolysis method developed here generates valuable sequence information not available from FAB-MS analysis of an unhydrolysed sample.

The results obtained from partial acid hydrolsis of neurotensin and hCGRP were most encouraging for further application of this method of sequence analysis, to a series of peptides of differing composition and sequence. The peptides examined were substance P $[M+H]^+$ 1347, alpha-melanocyte-stimulating hormone (α-MSH)

$[M+H]^+$ 1664, renin substrate $[M+H]^+$ 1758, corticotropin-like intermediate-lobe peptide (CLIP) $[M+H]^+$ 2490 and beta-endorphin $[M+H]^+$ 3461. Virtually complete sequence determination of the peptide sequences of substance P, alpha-MSH and renin substrate was achieved by the application of this partial acid hydrolysis/MS method. Partial sequence data was derived by acid hydrolysis of beta-endorphin and CLIP; twelve amino acids corresponding to residues 16 to 27 were elucidated from the FAB mass spectrum of acid hydrolysed beta-endorphin, and ten residues 13 to 22 determined from the FAB mass spectrum of acid hydrolysed CLIP. In parallel with each peptide hydrolysate studied a sample of intact peptide was also analysed by FAB-MS for comparison.

3.2.2.1 **Acetylation of Peptide Acid Hydrolysates**

It has been demonstrated that partial acid hydrolysis gives additional information to the normal FAB-MS fragmentation of intact samples. However, the extent of the mass spectrometrically generated fragmentation contribution to the observed spectra had yet to be determined. This was resolved by chemically tagging the N-terminus of the peptide under investigation which would elucidate where chemical cleavage stopped and where mass spectrometric fragmentation started. The acid hydrolysis reaction was terminated by dilution (to aid freezing) and lyophilisation with subsequent acetylation of the peptide hydrolysate under correct conditions for acetylation of primary alpha-amino groups only. A fast acetylation reaction [82,83] was employed, which acetylated all fragments or intact molecules with a free N-terminus, therefore, if carried out and terminated before FAB-MS analysis but after hydrolysis those

fragments produced mass spectrometrically would not carry the acetyl group on their N-termini. This experiment was designed so as to allow differentiation between those peptide fragments formed by hydrolysis prior to FAB-MS analysis and those formed as a result of fragmentation in the mass spectrometer ion source. Interpretation of the spectrum of the acetylated hydrolysate was aided by using a mixture of acetic and deuteroacetic anhydride, ($^1H_6$) and ($^2H_6$), in a ratio of 1:1, so that the generated FAB-MS spectrum is characterised by doublets arisng from the incorporation of the $^1H_6$ and $^2H_6$ labels. Since, this reaction, when carried out in an excess of methanol, is specific for primary amines each chemically distinct amino group could be expected to be derivatised. In a mixture consisting of a large number of smaller peptides from the hydrolysis of a larger peptide each fragment would have its own amino terminus which would be acetylated accordingly. Presented in Figures 3.4 and 3.4.1 are the FAB mass spectra of (a) acid hydrolysed, and (b) acid hydrolysed and then acetylated, alpha-MSH. The spectrum in Figure 3.4 is dominated by signals m/z; 214, 342, 399, 585, 741, 888, 1025, 1154, 1285, 1372 and 1535, which correspond to C-terminal molecular ions from which the sequence, by calculating the mass differences between these intense signals, is deduced as,
Tyr-Ser-Met-Glu-His-Phe-Arg-Trp-Gly-Lys-. Illustrated in Figure 3.4.2 is the sequence of alpha-MSH together with the calculated N- and C-terminal fragment ions. The FAB-MS spectrum (Figure 3.4.1) of acid hydrolysed and then N-acetylated alpha-MSH was generated from an identical quantity of native sample as that used for partial acid hydrolysis (the FAB mass spectrum shown in Figure 3.4). By

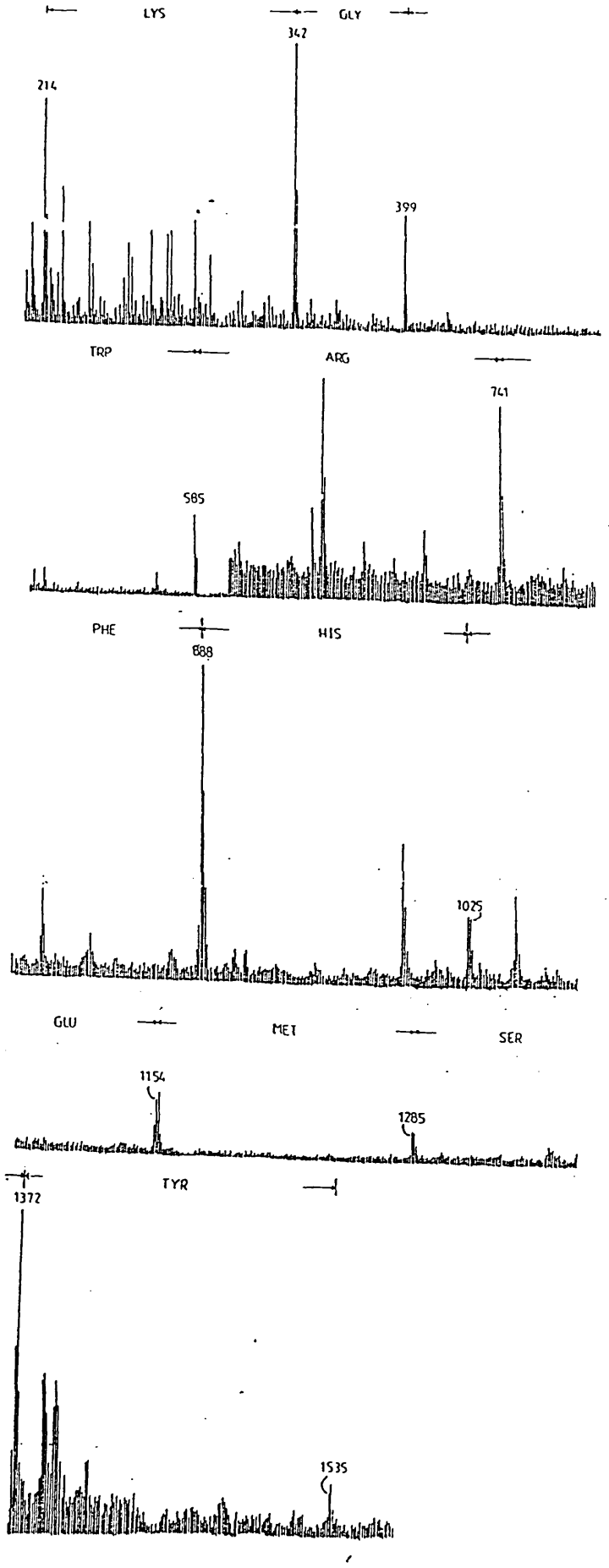


Figure 3.4 FAB-MS of Acid Hydrolysed alpha-Melanocyte Stimulating Hormone (∝-MSH)

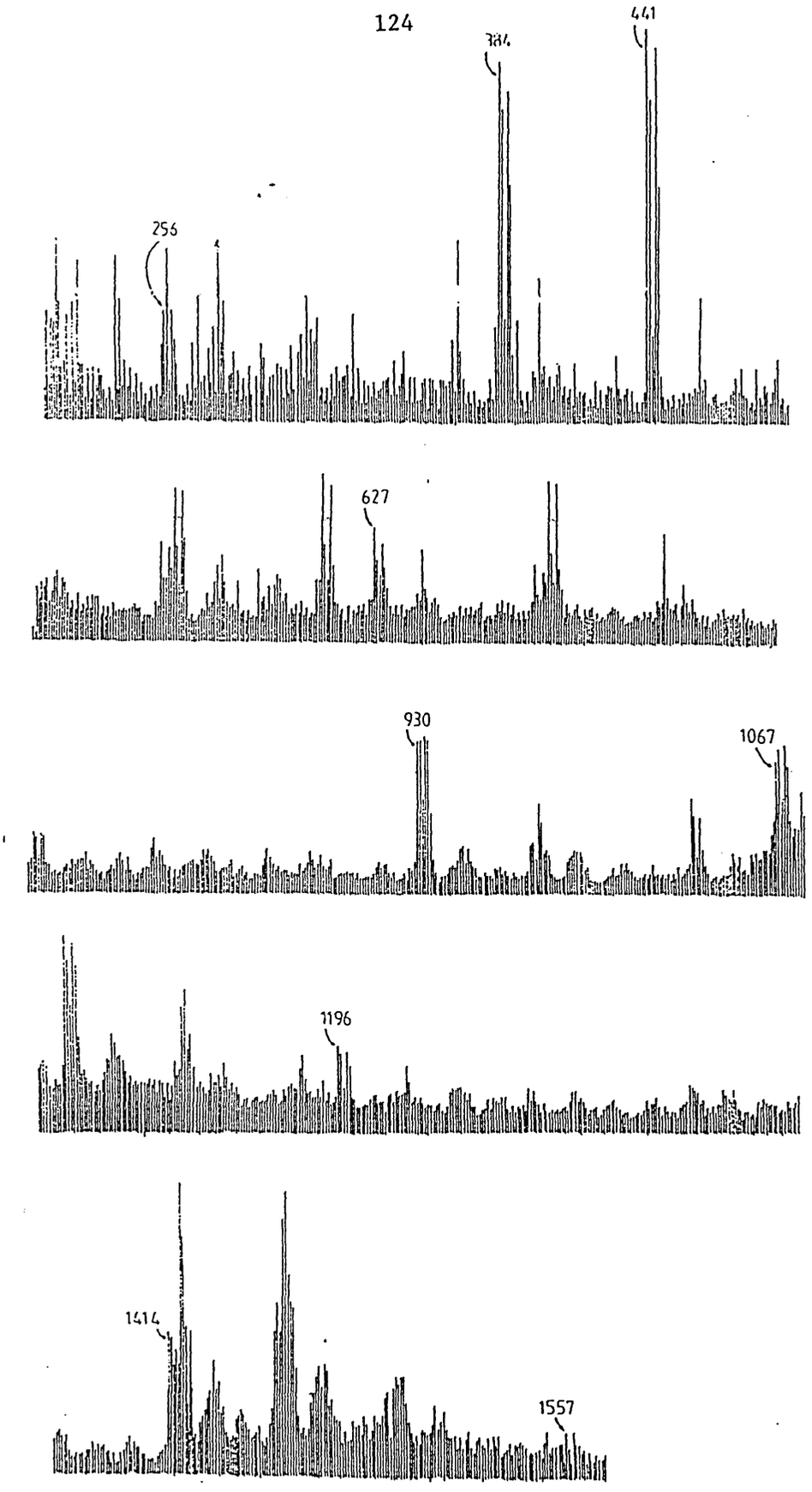


Figure 3.4.1 FAB-MS of Acid Hydrolysed and then Acetylated alpha-Melanocyte Stimulating Hormone (∝-MSH)

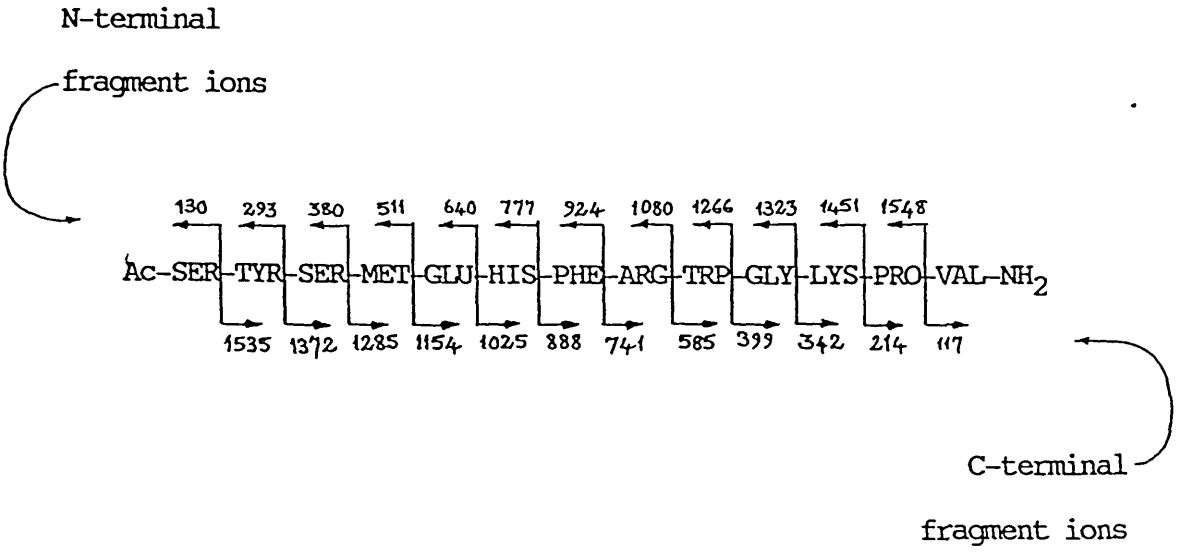


Figure 3.4.2 The Primary Sequence of alpha-Melanocyte Stimulating Hormone (MSH), $[M+H]^+$ 1664

comparison of these two spectra it is apparent that each of the C-terminal molecular fragment ions, generated by acid hydrolysis and observed in Figure 3.4, have shifted after acetylation and are now observable as characteristic doublets indicative of acetylation. The signals in this spectrum, Figure 3.4.1, are observed 42/45 a.m.u. higher than the original C-terminal molecular ions, i.e. m/z; 256, 384, 441, 627, 783, 930, 1067, 1196, 1414 and 1577. This increase in mass of 42/45 a.m.u. corresponds to the addition of a mixed acetyl group, CH_3CO/CD_3CO, to the free amino terminus of each peptide generated with the corresponding loss of a proton on substitution.

This experiment clearly demonstrates that the peptide fragments, identified by FAB-MS and that allow sequence analysis of the parent compound, are chemically produced. This was confirmed on repetition of this experiment with the other peptides studied. The actual process of peptide fragment formation however, was as yet unanswered, i.e. are the observed fragments generated by a single peptide bond cleavage on hydrolysis or by the sequential hydrolysis of peptide bonds. If the former is the method of fragment formation then the question arises as to why the remaining, complementary, peptide portions have not been observed by FAB-MS analysis of the hydrolysate mixture, for example if a fragment 7 to 13 (of a 13 residue peptide) is observed why is the signal corresponding to the peptide 1 to 6 not found in the FAB mass spectrum?

3.2.3 **Analysis of the Peptide Components of Acid Hydrolysates**

In order to determine which peptides are generated by concentrated partial acid hydrolysis the acid hydrolysate mixture of neurotensin was purified. The lyophilised reaction mixture was purified by reverse-phase High Performance Liquid Chromatography, (HPLC), with analysis of the HPLC fractions by FAB-MS. From partial acid hydrolysis experiments a large number of peaks corresponding to peptide fragments were recorded on the HPLC chromatograph. Subsequent screening by FAB-MS of the HPLC fractions to analyse which peptide fragments had been generated, identified a large number of peptides many of which were not previously observed by direct FAB-MS analysis of the reaction mixture, prior to purification. The fractionated peptide fragments identified by FAB-MS analysis are listed in Table 3.1. Purification and identification of a large number of peptide fragments, previously undetected, allows inference that the formation of sequence related peptide fragments is not by sequential hydrolysis. Hence, the question arises as to why many of the peptides identified to be present in the acid hydrolysate mixture are not observed on FAB-MS analysis of the intact reaction mixture; this can probably be answered mass spectrometrically.

3.2.4 Rationalising Suppression of Peptide Components in a Mixtures by FAB-MS Analysis in Terms of Hydrophobicity

Suppression of certain components in a mixture by others on FAB-MS analysis of the intact mixture is a phenomenon extensively reported upon [27], with recent correlations between peptide hydrophilicity/hydrophobicity drawn [84]. This recently proposed explanation of suppression is that the more hydrophobic components

MOLECULAR ION	SEQUENCE	$\Sigma f/n$	OBSERVED IN THE FAB-MS OF THE INTACT MIXTURE
245 (12-13)	I-L	-1550	+
408 (11-13)	Y-I-L	-1510	+
505 (10-13)	P-Y-I-L	-1175	+
538 (2-5)	L-Y-E-N	-420	-
548 (9-12)	R-P-Y-I	-590	-
553 (1-4)	pE-L-Y-E	-515	-
554 (3-6)	Y-E-N-K	+38	-
661 (9-13)	R-P-Y-I-L	-802	+
667 (2-6)	L-Y-E-D-K	-300	-
688 (7-11)	P-R-R-P-Y	-78	-
764 (2-7)	L-Y-E-D-K-P	-278	-
801 (7-12)	P-R-R-P-Y-I	-307	+/-
817 (8-13)	R-R-P-Y-I-L	-533	+
914 (7-13)	P-R-R-P-Y-I-L	-499	+
962 (3-9)	Y-E-N-K-P-R-R	+234	-
1042 (6-13)	K-P-R-R-P-Y-I-L	-379	+
1156/7 (5-13)	N/D-K-P-R-R-P-Y-I-L	-238/-269	+
1285/6 (4-13)	E-N/D-K-P-R-R-P-Y-I-L	-163/-191	+
1448/9 (3-13)	Y-E-N/D-K-P-R-R-P-Y-I-L	-278/-304	+
1561/2 (2-13)	L-Y-E-N/D-K-P-R-R-P-Y-I-L	-273/-296	+
1672/3 (1-13)	pE-L-Y-E-N/D-K-P-R-R-P-Y-I-L	-310/-345	+

Table 3.1 Hydrophobicity Calculations (based on Bull & Breese values) for the Peptide Fragments Identified, by FAB-MS to be Present in the Purified Hydrolysate Mixture of Neurotensin.

Key: (+)=seen on FAB-MS analysis of intact hydrolysate mixture
(-)=supressed, i.e. not observed.

The highlighted peptide fragments correpond to those observed in the FAB-MS of acid hydrolysed Neurotensin and from which the sequence was deduced.

of a mixture are expected to lie on the surface of the matrix from which they are ionised and desorbed, whereas the hydrophilic peptides do not as readily occupy the surface of the matrix and hence, are less readily ionised and desorbed and thus, are not observed in FAB spectra. This hypothesis proposed by Williams *et al* (1986)[84] has recently been corroborated by an independant study by Naylor and co-workers [85]. Their study has extended this hydrophobicity theory to large peptides/small proteins and have reported a striking correlation between the hydrophobic character and the molecular ion abundance observed in FAB mass spectra. This hypothesis was tested in this thesis work by calculating the hydrophobicity index, using the Bull and Breese scale [86] of hydrophobicity, for each petide fragment derived from neurotensin and identified to be present in the acid hydrolysate mixture. The Bull and Breese scale is described as "the preference or reluctance of an amino acid to transfer from aqueous solution to an air-water interface, with a positive index indicative of hydrophilicity and with a negative index hydrophobicity is associated", therefore the more negative the index the more hydrophobic is the amino acid. In calculating the overall hydrophobicity/hydrophilicity of a peptide the indices of the individual amino acids comprising the peptide sequence are summed and the average calculated; $\Sigma f/n$ where "f" is the free energy of transfer to surface and "n" the number of amino acid residues in the peptide. Summarised in Table 3.1 are the calculated hydrophobicity indices for each peptide fragment identified to be present by FAB-MS (after purification) in the acid hydrolysate of neurotensin. All the fragments observed which fall in a series of sequence-related peptides, namely 12-13, 11-13

.....1-13, have negative values, supporting the proposed theory. However, some of the fragments identified to be present after purification of the hydrolysate mixture have negative indices, but are <u>not</u> usually observed by FAB-MS analysis of the intact mixture. For example, the fragment (9-12), $[M+H]^+$ 548, has a calculated negative value of, -590, which according to the hydrophobicity theory should be observed, but is not. This value should be compared with peptides at m/z 817, 914, 1042, 1156/7, 1285/6, 1448/9, 1561/2, and 1672/3, listed in Table 3.1, which are observed in the FAB mass spectrum of acid hydrolysed neurotensin, with hydrophobicity indices of, -533, -499, -379, -238/-269, -163/-191, -278/-304, -273/-296, and -310/-345 respectively. Similarly, there are other fragments with calculated negative indices which again are not seen in the FAB-MS spectrum of the unfractionated hydrolysate mixture. It is recognised that di- and tri-peptides despite having calculated negative indices may not be observed in the FAB spectrum due to the polar influences of the amino and carboxyl termini which are expected to weigh heavily in short peptides [84]. It is not possible though, to rationalise observed suppression solely in terms of the hydrophobicity and hydrophilicity concept proposed by these authors. These observations question the conclusions of Williams *et al* and the more recent publication from that laboratory suggesting derivatisation of hydrophilic peptides [87] in efforts to lend a hydrophobic character to overcome suppression effects on FAB-MS analysis. Naylor *et al* also observed from their results, that in certain cases the simple hydrophobic model was not a sufficient expla nation. This conclusion was drawn from their study of the

protein phospholipase A_2. The hydrophobicity index of this protein is +69 and therefore should not be observed if for example mixed with bovine insulin (hydrophobicity index -130). However, Naylor and co-workers did observe a signal in the FAB mass spectrum corresponding to this protein. From this observation they concluded that other factors such as secondary and tertiary structure may also be of importance for determining the surface activity of large peptides.

In an attempt to correlate this hydrophobicity concept to the other peptides examined here, for sequence analysis by the partial acid hydrolysis procedure developed in this thesis research work, hydrophobicity calculations were made for each stretch of sequence elucidated. Substance P was completely sequenced by this partial acid hydrolysis method, with the sequence related peptide fragments running from the C-terminus to the N-terminus. The amino acid sequence of substance P is shown in Figure 3.5, with the calculated hydrophobicity indices for the peptide fragments illustrated in Table 3.2. These sequence related peptides are found to have negative indices and thus, agree with the hydrophobicity theory of Williams *et al*. Discrepancies are observed when the hydrophobicity indices are calculated for the C-terminal sequence related peptides of alpha-MSH, shown in Table 3.3. The FAB mass spectrum of acid hydrolysed alpha-MSH is dominated by intense signals corresponding to sequence related peptide fragments, with the sequence running in a C- to N-terminal direction. The hydrophobicity indices of these peptides are mostly negative and thus correlate with the hydrophobicity theory. However, there are a

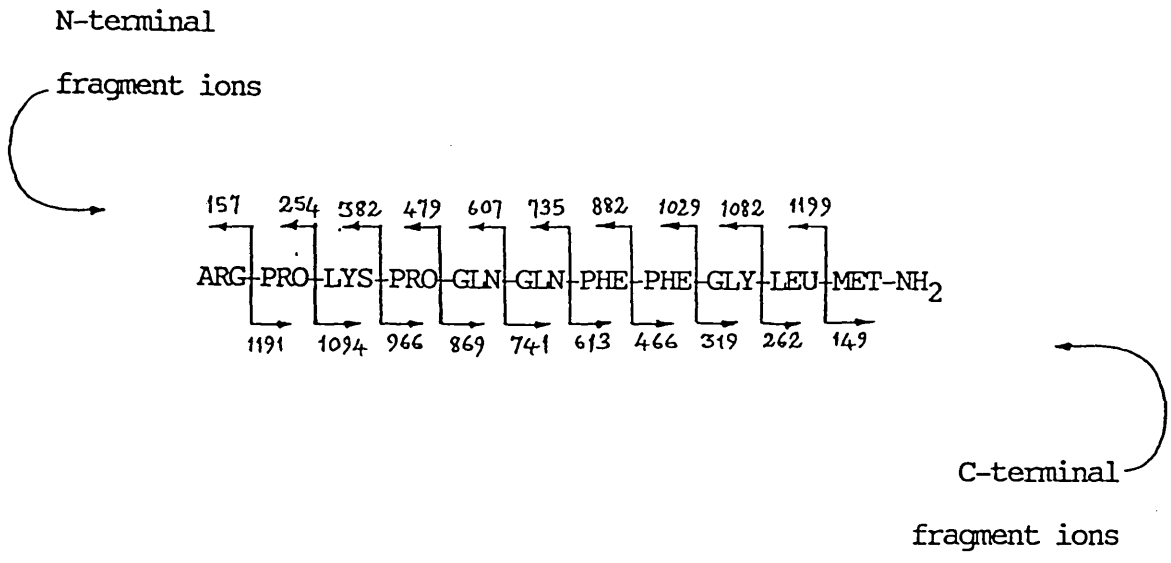


Figure 3.5 Primary Sequence of Substance P, $[M+H]^+$ 1347.

Peptide Fragments in FAB-MS of Acid Hydrolysate	$\Sigma f/n$
L-M	-1155
G-L-M	-500
F-G-L-M	-755
F-F-G-L-M	-908
Q-F-F-G-L-M	-595
Q-Q-F-F-G-L-M	-371
P-Q-Q-F-F-G-L-M	-346
K-P-Q-Q-F-F-G-L-M	-256
P-K-P-Q-Q-F-F-G-L-M	-248
P-P-K-P-Q-Q-F-F-G-L-M	-163

Table 3.2 Hydrophobicity Indices for the Peptides Observed in the FAB Spectrum of Acid Hydrolysed Substance P.

Peptides Fragments in FAB-MS of Acid Hydrolysate	$\Sigma f/n$	Reverse Sequence N- to C-terminal	$\Sigma f/n$
PV	-460	SY	-505
KPV	-153	SYS	-197
GKPV	+88	SYSM	-313
WGKPV	-170	SYSME	-148
RWGKPV	-27	SYSMEH	-8
FRWGKPV	-240	SYSMEHF	-224
HFRWGKPV	-124	SYSMEHFR	-110
EHFRWGKPV	-53	SYSMEHFRW	-231
MEHFRWGKPV	-114	SYSMEHFRWG	-127
SMEHFRWGKPV	-66	SYSMEHFRWGK	-74
YSMEHFRWGKPV	-179	SYSMEHFRWGKP	-82
SYSMEHFRWGKPV	-133	SYSMEHFRWGKPV	-133

Table 3.3 Hydrophobicity Indices for (a) the Observed C- to N-terminal Peptide Fragments Interpreted from the FAB-MS Spectrum of Acid Hydrolysed alpha-MSH, and (b) the Reverse (incorrect) Amino Acid Sequence.

number of fragments (observed in the FAB-MS of acid hydrolysed alpha-MSH) with very low negative indices with one peptide fragment, 10-13 $[M+H]^+$ 399, possessing a positive index, +88. Also listed in Table 3.3 are the hydrophobicity indices of sequence related peptides from an N-terminal to C-terminal direction, these can be seen to also possess negative indices. Therefore, as the peptide bonds that are hydrolysed for the formation of the C- to N-terminal sequence related peptides are those which would allow formation of the converse, N- to C-terminal sequence related fragments, then in accordance with the hydrophobicity concept, if formed, these peptide masses have a high probability of being observed in the FAB spectrum. However, no N-terminal sequence related peptides were observed in the FAB-MS of acid hydrolysed alpha-MSH, see Figure 3.4. It was demonstrated by the purification of the hydrolysate of neurotensin that a number of other peptide fragments, not observed in the FAB spectrum of the intact hydrolysate mixture, were formed but were suppressed (despite possessing negative hydrophobicty indices). If this is reasoned to be true for all peptides treated by partial concentrated acid hydrolysis then a number of peptide fragments from both substance P and alpha-MSH, if generated by acid treatment, are calculated to possess negative hydrophobicty indices, however, none are observed in the FAB-MS of the acid hydrolysate mixtures. The hydrophobicity model was subsequently applied to those peptides from which partial sequence information was generated by acid hydrolysis. The amino acid sequence of CLIP is depicted in Figure 3.6, with the partial sequence 13-22, determined by interpretation of the FAB mass spectrum of acid hydrolysed CLIP, higlighted. The intense signals

ARG-PRO-VAL-LYS-VAL-TYR-PRO-ASN-GLY-ALA-GLU-**ASP-GLU-LEU-ALA-**

GLU-ALA-PHE-PRO-LEU-GLU-PHE

Figure 3.6 Primary Sequence of Corticotropin-like Intermediate-lobe peptide (CLIP), $[M+H]^+$ 2490.

The highlighted amino acids correspond to the residues sequenced by partial **concentrated** acid hydrolysis/MS

in the FAB mass spectrum were found to correspond to a set of sequence related peptides with the sequence running in a C-terminal to N-terminal direction. The calculated hydrophobicity indices for these peptide fragments, listed in Table 3.4, are found to be negative. Similarly, partial acid hydrolysis of beta-endorphin allowed elucidation of the sequence between residues 15 and 27. Unlike the sequence determined from CLIP the direction of this stretch of amino acids was found to run in an N-terminal to C-terminal direction. The sequence of beta-endorphin is illustrated in Figure 3.7; the hydrophobicity indices calculated for the peptides fragments from which the sequence was deduced are listed in Table 3.5 which are found to be negative. However, it has been established that partial acid hydrolysis generates a number of fragments from each peptide studied, many of which are not observed in FAB spectra of the intact hydrolysates, thus, what cannot be completely explained by the proposed hydrophobicity concept is the absence of peptides, derived from other regions in the primary sequences which have negative hydrophobic indices, in the FAB spectra. For example, beta-endorphin peptide fragments such as 1-2, 1-4, 1-5, and 1-6, all have calculated negative indices (-310, -333, -398 and -380 respectively) but no mass signals corresponding to any of these peptides are observed in the FAB mass spectrum of acid hydrolysed beta-endorphin. The probability of peptide fragments 1-5 and 1-6, in particular, having been formed by partial acid hydrolysis is relatively high as residues 6 and 7 are threonine and serine respectively which are readily cleaved by acid treatment. Presented in Figure 3.2 is the FAB mass spectrum of the acid hydrolysate mixture of hCGRP, which is characterised by

TYR-GLY-GLY-PHE-MET-THR-SER-GLU-LYS-SER-GLN-*THR-PRO-LEU-VAL-*

THR-LEU-PHE-LYS-ASN-ALA-ILE-ILE-LYS-ASN-ALA-TYR-LYS-LYS-GLY-

GLU

Figure 3.7 The Primary Sequence of beta-Endorphin, $[M+H]^+$ 3461

The highlighted amino acids correspond to the residues sequenced by partial **concentrated** acid hydrolysis/MS.

Peptide Fragments in FAB-MS Acid Hydrolysate	$\Sigma f/n$
EF	-505
LEF	-887
PLEF	-708
FPLEF	-870
AFPLEF	-623
EAPLEF	-461
AEAPLEF	-328
LAEAPLEF	-474
ELAEAPLEF	-376
DELAEAPLEF	-286

Table 3.4 Hydrophobicity Indices of Peptide Fragments Identified in the FAB Mass Spectrum of the Acid Hydrolysate of Corticotropin-like Intermediate-lobe Peptide (CLIP)

Underlined residues elucidated from FAB-MS of intact acid hydrolysate

Peptide Fragments in FAB-MS of Acid Hydrolysate	$\Sigma f/n$
TPLV-T	-398
TPLV-TL	-607
TPLV-TLF	-737
TPLV-TLFK	-588
TPLV-TLFKN	-424
TPLV-TLFKNA	-320
TPLV-TLFKNAI	-423
TPLV-TLFKNAII	-509
TPLV-TLFKNAIIK	-434
TPLV-TLFKNAIIKN	-34
TPLV-TLFKNAIIKNA	-276
TPLV-TLFKNAIIKNAY	-348

Table 3.5 Hydrophobicty Indices of Peptide Fragments Identified in the FAB Mass Spectrum of the Acid Hydrolysate of beta-Endorphin

Underlined residues correspond to those elucidated from the FAB-MS of the intact acid hydrolysate.

Peptide Fragments in FAB-MS of Acid Hydrolysate	Σf/n
LL	-1650
GLL	-830
AGLL	-470
LAGLL	-706
RLAGLL	-473
HRLAGLL	-307
THRLAGLL	-233
VTHRLAGLL	-290
CVTHRLAGLL	-225
TCVTHRLAGLL	-178

Table 3.6 Hydrophobicty Indices of Peptide Fragments Identified in the FAB Mass Spectrum of Acid Hydrolysed human Calcitonin Gene Related Peptide (hCGRP)

intense signals from which a sequence, in a C-terminal direction, is deduced between residues 6 and 16. Listed in Table 3.6 are the hydrophobicity indices for the sequence related peptide fragments highlighted in Figure 3.2.1. From Table 3.6 it can be seen that these values are all negative and thus, the presence of intense mass signals in the FAB spectrum of the hydrolysate mixture correlates with the hydrophobicity hypothesis. Thus, it appears that the hydrophobicity model proposed by Williams *et al* does play a role, to a certain extent, in determining which peptide fragments are observed by FAB-MS analysis of peptide mixtures, but cannot explain every case. However, the data generated from the study presented here illustrates that the supposition of Naylor and co-workers for small proteins is not true since the small peptides generated here would not be expected to hold secondary or tertiary structures. It can be concluded that the hydrophobicity model is too simplistic, even involving secondary structure features, to explain the observations from this partial acid hyrolysis mixture data.

3.2.5 **Recent Application of Partial Acid Hydrolysis**

Subsequent to this study De Angelis and co-workers reported their investigations into the application of partial dilute acid hydrolysis for peptide sequencing. They demonstrated their application of partial acid hydrolysis using two model peptides [2-D-Ala,5-Leu]enkephalin amide, (structure YAGFL-NH_2) $[M+H]^+$ 569, and substance P, $[M+H]^+$ 1347. The strategy involved following the hydrolytic process by taking several FAB spectra of the crude hydrolysis mixtures at several time points. A series of

FAB-MS spectra were obtained which were characterised by intense molecular ions of the oligomers arising from the hydrolysis, with an expected decrease in high-mass fragments as the hydrolysis proceded. By subsequent interpretation of the complex spectra involving the alignment of overlapping sequences, structure determination was possible. This methodology is a relatively lengthy procedure in that it is necessary for seven FAB mass spectra to be taken, over a period of 6 hours, involving analysis of 39 out of a possible 64 hydrolysis products, with reassembly of their molecular masses for the original sequence of the undecapeptide to be compiled. This procedure would become more complex if it were applied to peptides larger than the example quoted of substance P, molecular weight 1346 a.m.u. However, two major differences between De Angelis's method and the one presented in this thesis are that De Angelis employed, as with the previous applications of partial acid hydrolysis, dilute acid, aqueous 5% HCl, with seven FAB spectra required to determine the structure of substance P, whereas the method developed and presented in this thesis utilises **concentrated** hydrochloric acid with only **one** spectrum per peptide examined required, i.e. a rapid method (reaction time 45 minutes) for sequence generation. The method presented by De Angelis requires 5 nanomoles of peptide, which was one of the main criticisms of Biemann who published a response to De Angelis's proposed methodology. The combination of partial acid hydrolysis, using concentrated acid, with peptide hydrolysate analysis by FAB-MS as reported in this thesis is demonstrated to be readily applicable to less than 500 picomoles without computer summing of spectra on a nine year-old ZAB instrument. This

sensitivity is illustrated by the high quality FAB mass spectra of the acid hydrolysates of neurotensin and hOGRP (Figures 3.1 and 3.2 respectively), both generated from 500 picomoles. Hence, the method presented here requires a comparable quantity of sample to that needed for sequence analysis by tandem MS, an alternative technique suggested by Biemann. In addition, to the quantitative limitation of De Angelis's method complete deamidation of asparagine, glutamine and C-terminal amide residues was observed. The complete deamidation of these amino acid residues was also initially observed with the method developed here, however, this was solved by omitting to add water at the end of the reaction to assist freezing prior to lyophilisation, with the method of direct lyophilisation of the reaction mixture implemented instead. The deamidation of asparagine and glutamine is still observed but to a lesser extent and is readily recognisable from signal cluster shapes therefore, it is no longer a disadvantage. The change and interpretation of signal clusters can be illustrated by comparing the cluster shapes of signals m/z 1042 and m/z 1157 in the neurotensin acid hydrolysed FAB mass spectrum, Figure 3.1. This change in cluster shape is the result of deamidation of the asparagine residue to aspartic acid. Despite m/z 1157 being the most intense signal it is determined that m/z 1156 is the correct mass corresponding to fragment 5 to 13 in the neurotensin molecule with a mass difference of 114 a.m.u. (Asn) between m/z 1156 and m/z 1042. Thus, the signal m/z 1157 is comprised of a combination of the ^{13}C contribution from the peptide fragment only containing asparagine, and the aspartic acid peptide component in the mixture, which together produce an unusually intense signal. In sequence

analysis investigations where amino acid residues are assigned purely on the basis of their residue masses interpreted from a FAB mass spectrum it is difficult to distinguish between glutamine and lysine residues, as both have the same residue mass of 128 a.m.u. Thus, partial deamidation of glutamine residues provides a means of differentiating between lysine and glutamine as lysine would not be deaminated on hydrolysis and hence, would not produce a change in signal cluster shape. Partial deamidation of glutamine and asparagine residues still occurs, despite the ommission of water at the end of the reaction, as hydrochloric acid itself is composed of approximately 37% acid and 63% water. As discussed above positions of partial deamidation are readily detected by a change in the signal cluster shape, see Figure 3.8, and thus, partial deamidation is exploited as a positive feature of the technique. However, it is noted that the C-terminal amide blocking group of substance P is completely deamidated by this acid hydrolysis method. This may possibly be attributed to the exposed position of the blocking group in the overall peptide structure, with very little steric hinderance. This observation is found to vary for each peptide studied, as for example sequence analysis of alpha-MSH (a C-terminal amide blocked peptide) by this acid hydrolysis/MS method finds that the amide blocking group is not completely deamidated. However, in general, a sample of native peptide would be examined by FAB-MS prior to hydrolysis and hence, the presence of a blocked C-terminus would be readily deduced from the molecular weight of the peptide. Therefore deamidation of C-terminally amide blocked peptides would not prove to be a significant disadvantage.

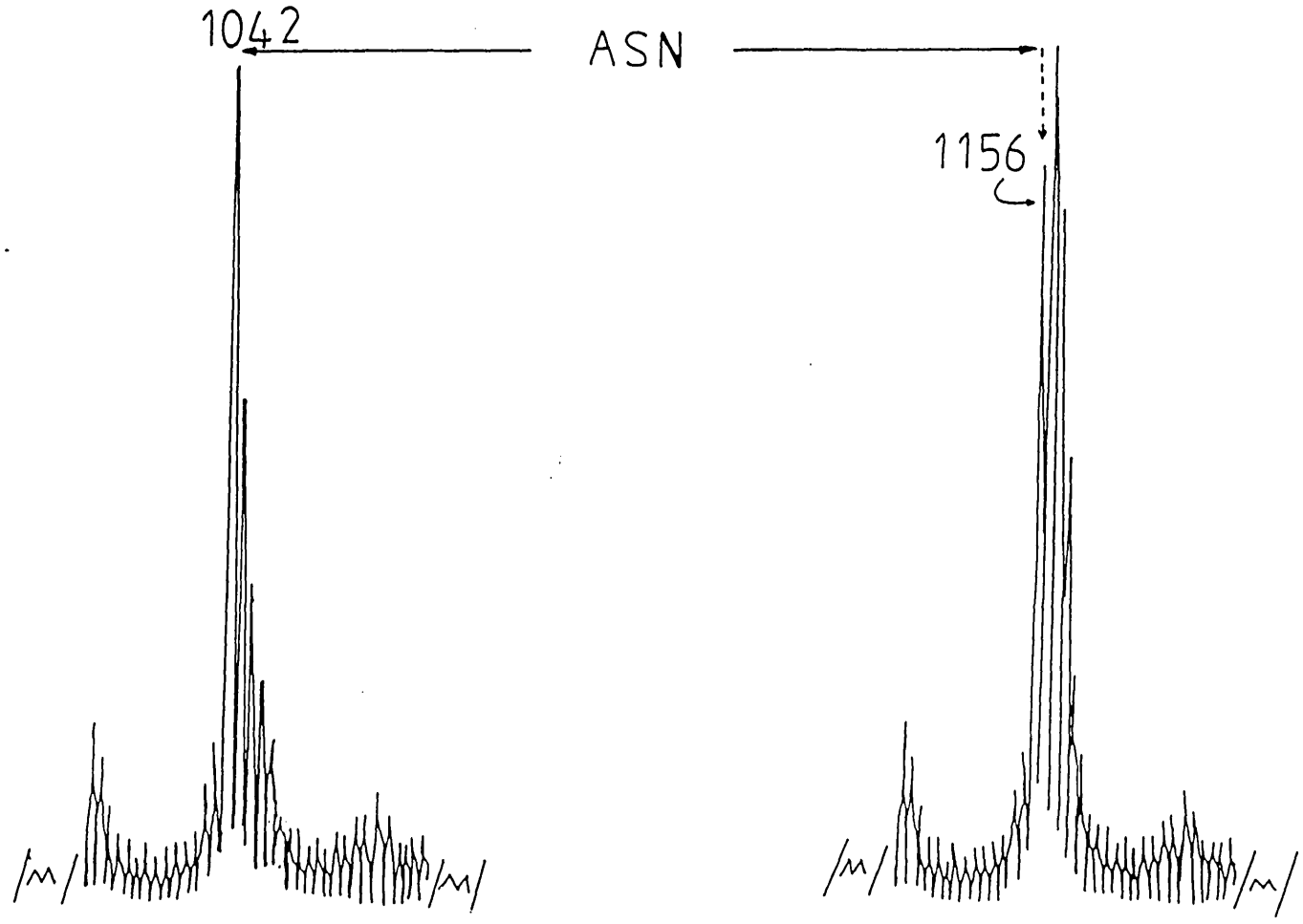


Figure 3.8 Signals m/z 1042 and m/z 1156 (observed in the FAB-MS of acid hydrolysed Neurotensin, Figure 3.1) Enlarged so as to Illustrate the Change in Signal Cluster Shape on Partial Deamidation.

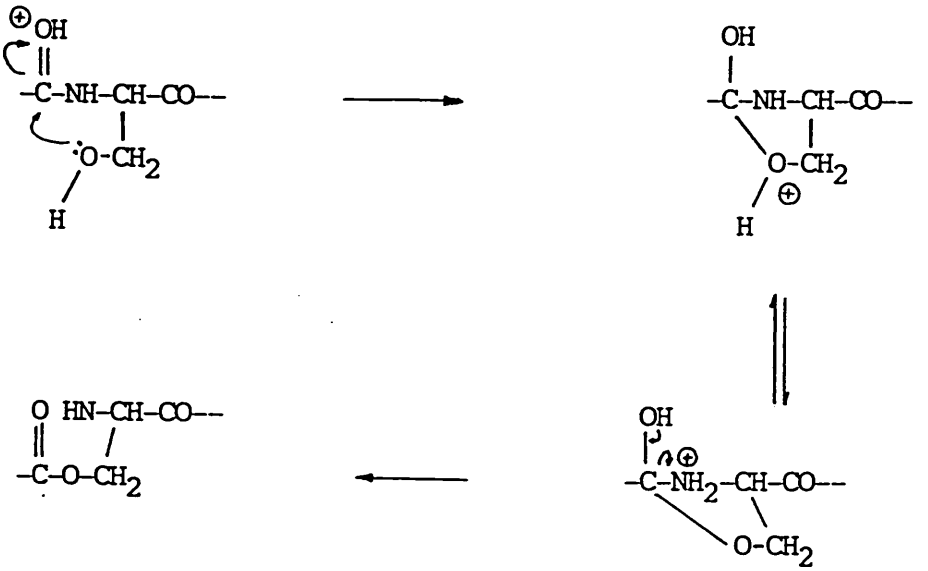


Figure 3.9 N•O acyl Transformation, Corresponding to Cleavage at the N-terminus of Serine Residues.

3.2.6 Spectral Patterns

In addition to the change in the cluster shape pattern observed on deamidation, as discussed, other such spectral changes have been noted which also aid interpretation. For example, a less intense signal observed 18 a.m.u. lower than the main signal is presumably indicative of a loss of H_2O, since it can be correlated with the presence of serine or threonine in the component peptide sequence. The resultant ion would be due to the generation of dehydro-Ser/Thr from a serine or threonine residue. This signal pattern can be observed in the acid hydrolysed sample of hCGRP, Figure 3.2. It has also been observed that, due to the lability of serine residues on acid hydrolysis, if the N-terminal or C-terminal residue is serine the molecular ion of the serine containing peptide will not be seen. The cleavage on the N-terminal side of serine peptide bonds as for example in hCGRP can be mechanistically explained as an N→O acyl transformation [88]. The mechanism of this cleavage reaction is shown Figure 3.9. Another curious observation is that there is evidence for a condensation reaction between juxtaposed carboxyl and amine groups resulting in a loss of 18 a.m.u. This was first noticed in the acid hydrolysed spectrum of neurotensin, between asparagine and glutamic acid residues, with the apearance of a single peak 18 a.m.u. lower than the main signals (see Figure 3.1).

3.2.7 Partial Acid Hydrolysis of Disulphide Bridge Containing Peptides

The earlier observation that by partial acid hydrolysis of hCGRP the disulphide bridge appeared to be reduced was investigated by appplying this method to other peptides, namely, vasopressin,

somatostatin and insulin which are disulphide bridge containing. On FAB-MS analysis of the peptide acid hydrolysate of vasopressin and somatostatin it was determined that the disulphide bridges in both remained intact and as a consequence less sequence data was generated. However, analysis of the FAB mass spectrum of the acid hydrolysate of insulin elucidated a peptide sequence from the B chain between residues 12 and 20. This peptide sequence contains a cysteine residue which in the intact molecule forms a disulphide bridge with a cysteine residue in the A chain (B_{C-19} to A_{C-20}). For the peptide fragments, from which the sequence 12 to 20 was deduced, to be observed this disulphide bridge must have been broken. It is proposed therefore, that the formation of certain peptide fragments by acid hydrolysis of large peptides places a strain upon disulphide bridges which eventually are destroyed. Thus, it is proposed that the susceptibility of disulphide bridge disruption increases as the peptide size increases. Therefore, this acid hydrolysis/FAB-MS protocol may not be so suitable for the analysis of oxidised, small, disulphide bridge containing peptides and that reduction of this tertiary structure prior to hydrolysis may be necessary for the applicability of this method for attaining optimum sequence data.

3.2.8 **Application of Partial Acid Hydrolysis to Large Proteins**

Having demonstrated that virtually complete sequence can be obtained by partial acid hydrolysis/FAB-MS of peptides with a molecular weight less than 2000 Da, and partial sequence from molecules with molecular weight between 2-3.5 kDa, this study was extended to include peptides with molecular weight exceding 3.5

kDa. The advent of efficient DNA-sequencing techniques has meant that it is often more time effective to sequence a protein at the DNA level rather than using classical protein sequencing techniques. However, protein sequencing methods are invaluable for generating short stretches of amino acid sequence to employ for the production of synthetic oligonucleotide probes used to identify the structural gene. Therefore, the aim in a true biological research study would be to produce a stretch of six to ten amino acid residues for subsequent oligonucleotide probe generation. Hence, if the partial acid hydrolysis/FAB-MS method, developed here, was applicable to large proteins for the generation of stretches of sequence data, as observed for hCGRP, insulin, and beta-endorphin, then this method would be highly suitable for the rapid production of sequence data for DNA-sequencing techniques.

Initially a large protein molecule alcohol dehydrogenase, molecular weight 40 kDa, was studied. Partial acid hydrolysis under the experimental conditions described above produced a variety of signals on analysis of the reaction mixture by FAB-MS. However, no relationship could be found between the signals for interpretion of an amino acid sequence. To determine if the hydrolysis conditions for maximising sequence information in a large protein, in which factors such as increased secondary and tertiary structure may be involved, were different, a time course reaction, with analysis of the hydrolysates at each time point by FAB-MS, was carried out. From this time course experiment it became apparent that it was too ambitious to apply this method to such a large molecule. Attention was subsequently focussed on a smaller protein cytochrome C,

molecular weight 11 kDa, with the aim of replicating the type of results obtained from smaller molecules. However, partial acid hydrolysis of cytochrome C, with experimental conditions identical to those devised for small peptides, did not generate data from which partial sequence could be derived. Mass differences between a few signals in the FAB spectrum did correspond to amino acid residues, possibly sequence related, but not sufficient for confident assignment or subsequent sequence analysis. Furthermore, time course reactions did not provide any alternate optimum conditions for sequencing. In order to rationalise this finding a sample of cytochrome C was reduced and carboxymethylated; reduction of a protein's disulphide bridges result in disruption of the secondary and tertiary structure of the molecule. However, the partial acid hydrolysis/FAB spectrum of carboxymethylated cytochrome C did not provide any additional information from which a stretch of sequence could be deduced. These experiments tenatively indicate that partial acid hydrolysis may not be readily applicable to large proteins.

However, more success was achieved from the application of this partial acid hydrolysis method to proteins with molecular weights between 5000 and 6000 a.m.u. namely, growth hormone releasing factor 1-44, (GHRF), molecular weight 5103 Da, adrenocorticotropic hormone (ACTH-Human), molecular weight 4536 Da and insulin, molecular weight 5730 Da. FAB-MS analysis of the acid hydrolysate of GHRF, employing the same experimental conditions determined previously for application to small molecules, generated high quality data, presented in Figure 3.10. From the most intense

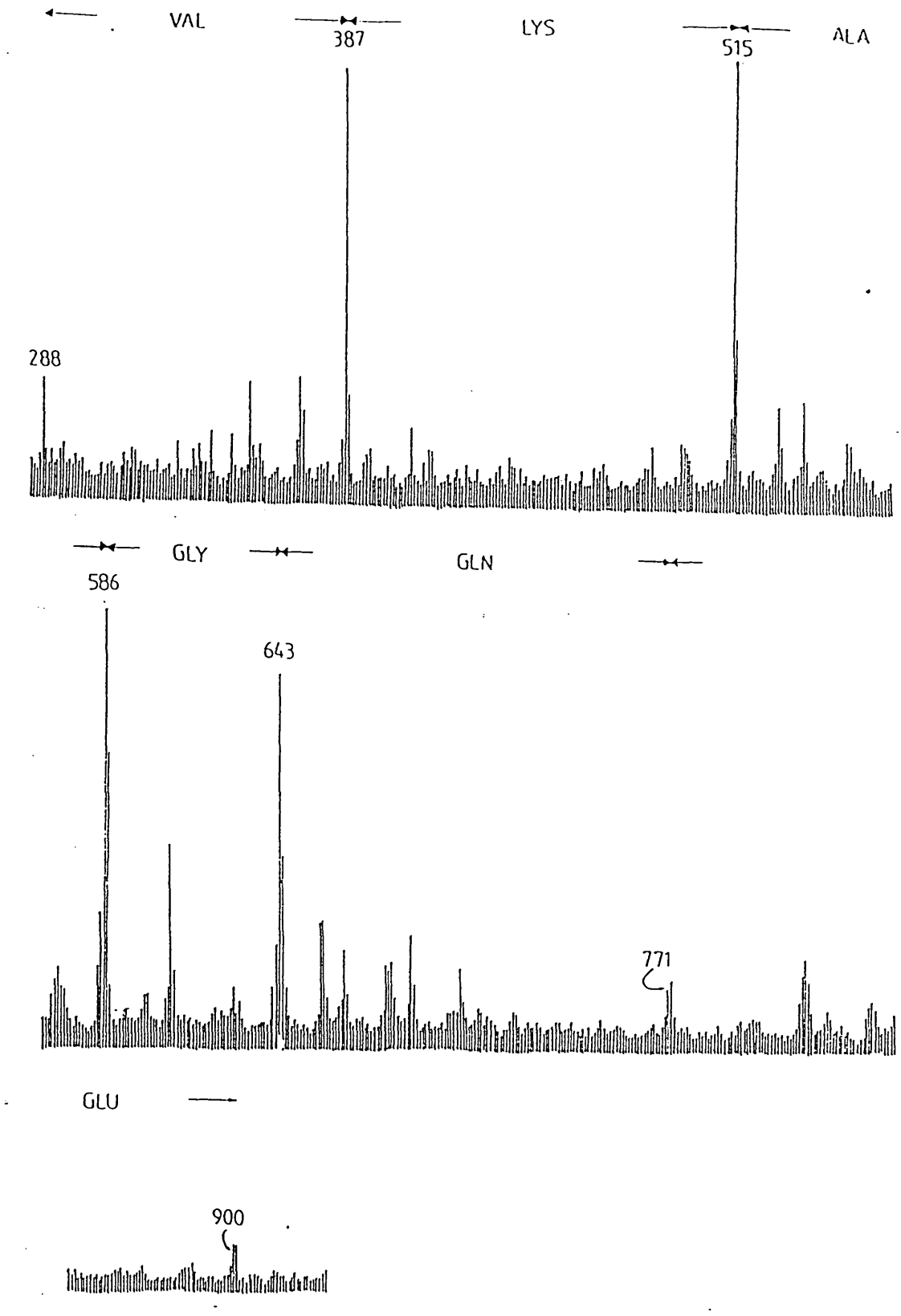


Figure 3.10 FAB-MS of Acid Hydrolysed Growth Hormone Releasing Factor 1-44 (GHRF), 500 picomoles.

signals in this spectrum m/z; 288, 387, 515, 586 and 643 , and by calculating the mass differences between these signals, the sequence, -Val-Lys-Ala-Gly-, is determined. Lysine was assigned in preference to glutamine since there was no evidence for partial deamidation from the signal cluster pattern. The mass signal m/z 288 was interpreted to correspond to residues Arg-Leu/Ile, the C-terminus of the molecule. Applying the same logic to the high mass end of the spectrum the next signal which gives a sensible residue mass from m/z 643 is m/z 771/772. There is an obvious change in the isotopic ratio at this point highlighting a position of deamidation. The mass difference between m/z 643 and 771 is 128 a.m.u., corresponding to either a glutamine or lysine residue. However, the change in the isotope pattern so that m/z 772 is slighlty more intense than m/z 771 indicates that deamidation has occured. Thus, it can be confidently concluded that this mass difference corresponds to a glutamine residue. The next signal corresponding to an amino acid mass difference is m/z 900/901 with the cluster shape reflecting the contribution of the previous deamidation. The mass difference of 129 a.m.u. allowed the next residue in the sequence to be assigned as glutamic acid. The sequence derived, by this acid hydrolysis method, of 8 amino acid residues corresponding to the C-terminus of the protein GHRF is thus; Glu-Gln-Gly-Ala-Lys-Val-Arg-Leu.

Interpretation of the FAB mass spectrum of acid hydrolysed insulin allowed eight residues between positions 12 to 20 in the B chain to be determined. This sequence runs in a N- to C-terminal direction. Similarly, the result of partial acid hydrolysis of ACTH allowed

the sequence of nine amino acids from the C-terminus of the peptide to be assigned. It is interesting to note that although ACTH is twice as large as CLIP, which represents its C-terminal portion, both molecules yield sequence information from their C-termini by the partial acid hydrolysis procedure presented here.

The research work presented in this thesis chapter suggests that this partial acid hydrolysis methodology would not be directly applicable to peptides with a molecular weight greater than 6000 Da. However, if a large protein of interest were digested and suitable sized fragments purified, then this method could be applied to rapidly provide stretches of sequence from each fragment of appropriate size.

3.2.8.1 **Correlation of the Hydrophobicity Model to the Partial Sequence Derived from Large Peptides**

The hydrophobicity indices for the sequences elucidated from the small proteins, GHRF and insulin were calculated, and are presented in Tables 3.7 and 3.8 respectively (the hydrophobicity indices for ACTH will be the same as those calculated for CLIP, (ACTH 18-39), see Table 3.4, as the same region has been sequenced). The complete amino acid sequences are shown in Figures 3.10.1 and 3.11 respectively. The hydrophobicity indices calculated for GHRF are a combination of negative and positive values, which clearly do not agree with the hydrophobicity model proposed by Williams *et al*, as being those peptide fragments preferentially observed. The hydrophobicity indices of the peptide fragments from which the sequence B (12-20) of insulin was determined (by intense signals in

TYR-ALA-ASP-ALA-ILE-PHE-THR-ASN-SER-TYR-ARG-LYS-VAL-LEU-GLY-GLN-LEU-

SER-ALA-ARG-LYS-LEU-LEU-GLN-ASP-ILE-MET-ASN-ARG-GLN-GLN-GLY-GLU-ARG-

ASN-GLN-**GLU-GLN-GLY-ALA-LYS-VAL-ARG-LEU**-NH_2

Figure 3.10.1 Primary Sequence of Growth Hormone Releasing Factor (GHRF), 1-44.

The highlighted amino acids correspond to those residues elucidated from the FAB mass spectrum of acid hydrolysed GHRF.

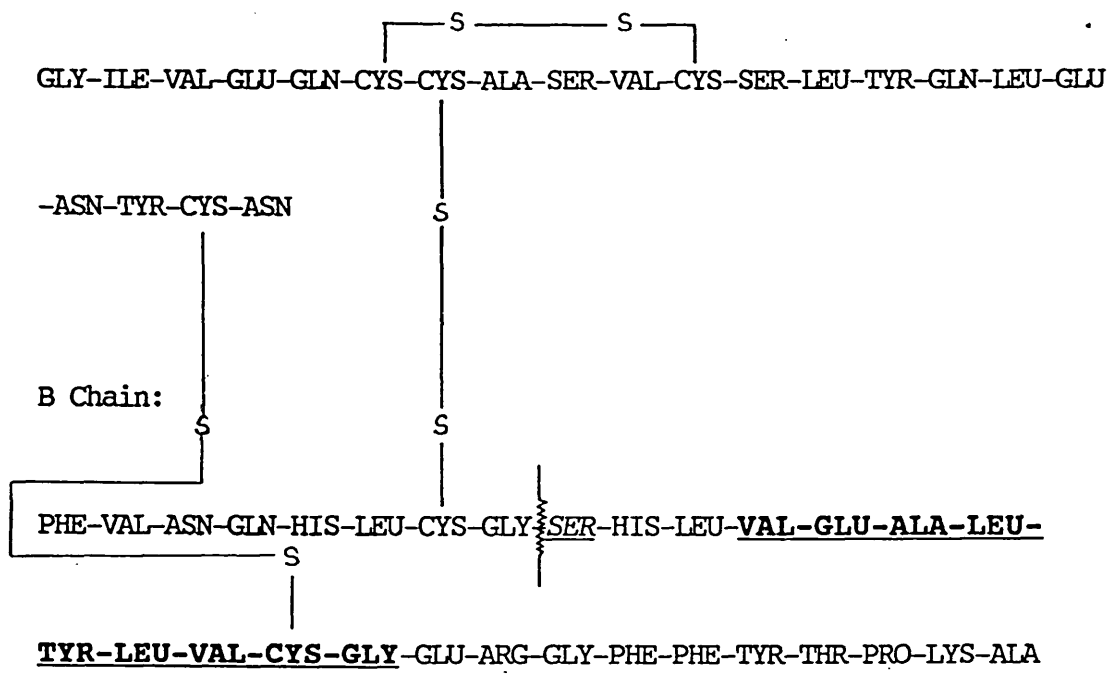


Figure 3.11 Primary Sequence of Bovine Insulin

Highlighted amino acids correspond to those elucidated from the FAB-MS of acid hydrolysed insulin.

Peptide Fragments in FAB-MS Acid Hydrolysate	$\Sigma f/n$
RL	-480
VRL	-570
KVRL	-313
AKVRL	-128
GAKVRL	+28
QGAKVRL	+163
EQGAKVRL	+206

Table 3.7 Hydrophobicity Indices of Peptide Fragemnts Identified in the FAB Mass Spectrum of Acid Hydrolysed Growth Hormone Releasing Factor, 1-44, (GHRF).

Peptide Fragments in FAB-MS Acid Hydrolysate	$\Sigma f/n$
SHL-V	-323
SHL-VE	-156
SHL-VEA	-28
SHL-VEAL	-260
SHL-VEALY	-406
SHL-VEALYL	-544
SHL-VEALYLV	-565
SHL-VEALYLVC	-481
SHL-VEALYLVCG	-373

Table 3.8 Hydrophobicity Indices of Peptide Fragments Identified in the FAB Mass Spectrum of Acid Hydrolysed Bovine Insulin

Underlined residues correspond to those elucidated from the signal mass differences in the FAB spectra of intact acid hydrolysates.

the FAB-MS of the acid hydrolysate of insulin) are found to all be negative. However, what cannot be explained is why the peptide fragments which allow the sequence between residues 12 and 19 from the A chain of insulin to be determined are not observed as these peptides have high negative hydrophobicity indices, higher than many of those peptide fragments observed in the FAB spectrum. Furthermore, such a sequence (or at least part of it) would be derived by cleavage at serine, which is readily hydrolysed by acid treatment.

3.3 Conclusion

Presented here is the novel application of partial acid hydrolysis using a concentrated mineral acid, hydrochloric acid. Analysis of peptide acid hydrolysate mixtures by FAB-MS has allowed full sequence analysis of peptides with masses lower than 2000 a.m.u. and partial sequence data from those peptides exceeding this mass value. High quality data has been generated from less than 500 picomoles of material indicating the sensitivity of the technique. Previous applications of partial acid hydrolysis methods have employed dilute mineral or concentrated organic acids, with this being the first report of partial acid hydrolysis utilising a concentrated mineral acid instead. The exact nature of this sequencing phenomenon is not completly understood with experiments indicating that numerous peptide fragments are generated by the acid hydrolysis procedure. However, it is the physical method of analysis, FAB-MS, which determines which fragments are seen. From

the intense signals observed in the FAB mass spectra it has been shown that the sequence or partial sequence of the peptide may be deduced. A time-course experiment allowed the determination of forty-five minutes at 65°C as the optimum time for producing a series of related peptides with random fragmentation occuring before and after this reaction time.

It is concluded from the data presented in this chapter that many fragments are generated by acid hydrolysis but which are not seen, due to suppression effects incurred by FAB-MS analysis of the hydrolysate mixture. It has also been demonstrated that the hydrophobicity theory presented by Williams *et al* is not sufficient to explain every case and hence, the recent proposal from this laboratory to derivatise hydrophilic peptides so as to lend a hydrophobic character may not be the answer for certain peptides where other contributory factors come into play. There have been many other suggestions for overcoming suppression effects for example, using different matrices and improving detection of supressed peptides by employing a continuous flow probe [89].

The method of partial acid hydrolysis reported in this chapter has been demonstrated to be a highly sensitive technique allowing sequence analysis of peptides, rapidly providing high quality data by FAB-MS analysis of peptide acid hydrolysates.

CHAPTER FOUR

METHODOLOGY DEVELOPMENT II - LINKED SCAN ANALYSIS OF PEPTIDES AND PROTEINS; DIRECT MIXTURE ANALYSIS

4.1 Introduction

Mass spectrometry has played an important role in the field of biomolecular structure determination over the last two decades. The most widely used mass spectrometric method for the generation of amino acid sequence information from peptides, until the early 1980s, was electron impact mass spectrometry, EI-MS. Although EI-MS has played an important role in biopolymer studies there are some disadvantages associated with this technique for sample analysis, such as:

(a) the need for derivatisation of samples to their volatile N-acetyl N,O permethyl derivatives [22,23],

(b) relatively large amounts of sample (greater than 5 nanomoles) are required, and

(c) the amount of internal energy transferred to the molecular ion is such that typical EI spectra are devoid of molecular ions.

The inception of fast atom bombardment (FAB) as a method of sample ionisation, FAB-MS, in 1981 [26,27,28] by Barber and co-workers has made a significant contribution to the field of structural analysis of biopolymers. Combined with High Field mass spectrometry, HF-MS,

introduced by this group [29,30], high quality molecular weight data is generated from less than 1 nanomole of material. Another advantage of the technique was that the need for the derivatisation of peptides/proteins was circumvented. However, a noticeable drawback of this method was that the fragmentation of peptides was minimal [27] resulting in a reduction of sequence information available. This observation promoted an investigation by Morris *et al*, 1981, into the application of linked scans for the analysis of fragment ions enhanced by collisionally activated decompositions, CAD. At that time, the data produced were not encouraging, a factor which no doubt led to an emphasis on development of four sector mass spectrometry for CAD/sequencing studies. The study presented below describes the recent re-evaluation of two sector linked scan methods for some of the structural problems encountered in the research work reported in this thesis, the results of which are presented in this chapter.

4.1.2 **Metastable Ion Formation and Detection**

When a substance is ionised, in the FAB source, ions are formed possessing a range of internal energies. The smaller the excess of internal energy the slower the rate of fragmentation. Some ions will possess sufficient internal energy to fragment further whilst still in the ion source, while ions with a small excess of energy will be able to reach the detector before they have time to fragment and are detected as normal M^+ ions. Those ions which possess the internal energy to fragment after leaving the ion source but before reaching the detector are termed "metastable" ions. Detection and analysis of metastables can contribute valuable

information towards the elucidation of the original structure of the molecule under study.

Metastable ions are those ions in a mass spectrometer which at the detector have less than the full kinetic energy originally imparted at the ion source. The actual place in the mass spectrometer where metastables are formed determines how they may be detected. There are three defined areas of formation the first, second and third field free regions (see Figure 1.3). In a conventional double-focussing instrument the first field free region (1st FFR) lies between the ion source and electric sector. Metastables formed in this region have the wrong energy to pass through the electric sector (an energy focussing device) and hence, will not be observed in the mass spectrum. A number of methods have been introduced for focussing metastable ions, such as maintaining a constant ion-accelerating voltage with variation of the electric sector voltage. Another method employed involves variation of the accelerating voltage with the voltage of the electric sector kept constant. Both these methods allow metastable ions to be focussed but which prevent the passage of normal ions. As normal ions are defocussed, [90,91,92], these methods are termed "defocussing" techniques.

4.1.3 Linked Scan Methods for Metastable Ion Analysis

There are a number of methods available for scanning mass spectrometers to detect metastables without interference from normal ions, thereby reducing the ambiguity of interpretation of the mass spectrum. These techniques are collectively called "linked

scanning" and have been automated so that these complex modes of scanning can be computer controlled. The term "linked scanning" refers to the simultaneous variation of the voltage (E) of the electric sector with either the accelerating voltage (V) or the magnetic field (B); a specified relationship is maintained between them throughout the scan. The various combinations of B, V and E are discussed by Boyd and Beynon [40].

One of the first methods of linked scan developed involves a static magnetic field, B, with variation of the accelerating and electric sector voltages so that E^2/V remains constant [93]. By focussing initially on a selected precursor ion, m^+, and then performing this linked scan, metastable ions which are products of fragmentation of m^+ in the 1st FFR of a conventional mass spectrometer can be determined. However, both this method and the defocussing method suffer from the limitations incurred by varying the accelerating voltage; the mass range is limited and the efficiency of the ion source changes during the scan.

A B/E linked scan method was introduced, [94], in which the accelerating voltage is kept constant with linked scanning of the magnetic field and electric sector. A B/E linked scan allows the generation of a metastable ion spectrum analogous to that obtained from an E^2/V scan in that all product ions of a chosen precursor are observed [95]. The mathematical derivation of the B/E relationship is shown below.

An ion m^+ in the magnetic field experiences an equal and opposite

centripetal and centrifugal force:

$$\text{Centripetal} = \text{Centrifugal}$$

$$Bev = mv^2/r$$

$$Be = mv/r \qquad (1)$$

where m=mass of the ion, m^+; v=velocity of the ion; e=electronic charge; B=magnetic field strength; r=magnet radius

In the electric sector

$$\text{Kinetic energy} = \text{Translational energy}$$

$$mv^2/2 = eV \qquad (2)$$

where V=accelerating voltage

The equation for the transmission of m^+ through the magnetic sector is given by;

rearranging (1) $\qquad mv = Ber \qquad (3)$

rearranging (2) $\qquad mv^2 = 2eV \quad \text{hence } v = (2eV/m)^{1/2} \qquad (4)$

substituting (4) into (3)

$$m(2eV/m)^{1/2} = Ber \qquad (5)$$

rearranging (3)

$$B = mv/er \qquad (6)$$

substituting (6) in (5)

$$mv = (2eVm)^{1/2}$$

and

$$v = (2eVm)^{1/2}/m \qquad (7)$$

The condition for the transmission of m^+, through an electric field, E, of radius R and plate separation d is given by

$$mv^2/R = Ee/d \quad \text{or} \quad mv^2/2 = EeR/2d = eV$$

Considering the fragmentation $m_1^+ \longrightarrow m_2^+ + m_3$ in the 1st FFR. Assuming that the velocity of mass m_1 is the same as the daughter ion mass, m_2 i.e. v_1 then;

energy of $m_2^+ = 1/2\ m_2 v_1^2$

substituting for v_1 from (7)

energy of $m_2^+ = m_2/m_1 eV$

momentum of $m_2^+ = m_2 v_1$ from (7) $= m_2/m_1 (2eVm_1)^{1/2}$

$= m_2/m_1 m_1 v_1$

Energy and momentum of m_2^+

$= m_2/m_1$ (energy or momentum of m_1^+)

Therefore in order for transmission of m_2^+ B is changed to $B(m_2/m_1)$ and similarly E is changed to $E(m_2/m_1)$. Thus, for the transmission of any fragment ion B and E must be varied so that the ratio B/E remains constant throughout the scan.

There are other methods of linked scanning such as B^2/E, in which it is possible to study the precursor ions leading to a particular daughter ion. Another linked scan method is $B/E(1-E)^{1/2}$, which selectively detects all metastables resulting from fragmentation. In addition to those techniques discussed, there are many methods of linked scanning available for the detection of metastables formed in the other field free regions e.g. mass-analysed ion kinetic energy spectroscopy, MIKES [96,97]. However, an in depth discussion of all these techniques would be outside the scope of this introduction.

4.1.3 B/E Linked Scanning

The B/E linked scan method was selected as appropriate for studying daughter ions generated in the 1st FFR of a reverse geometry two-sector instrument, ZAB-2SE, enhanced by collisionally activated decompositions, CAD. This relationship is preferred over variation of the accelerating voltage due to the limitations which are incurred, as previously disscussed.

The data generated from the early 1981 work (H.R. Morris, personal communication), investigating the potential of linked scanning/CAD for enhancing fragment ions and analysis of small peptides i.e. less than 1000 a.m.u., was of low intensity and largely devoid of sequence information. Factors such as the large amounts of sample required, greater than 10 nanomoles, the limitations associated with instrument design parameters at this time, particularly with respect to the width of the mass window examined, the accuracy of the scan-law generation and manual mass calibration did not encourage the development of this method as a suitable procedure for studying high mass peptides or peptide mixtures.

Developments in instrument sensitivity and improved computer software for computer averaging of summed scans (MCA), a development requested by this group, encouraged the re-examination of the potential of linked scanning as a method for the detection and enhancement of peptide (and carbohydrate) fragmentation for sequence studies. Presented in this thesis work is an investigation of the application of the B/E linked scan/CAD method for peptide

sequencing and more importantly this is the first work in which linked scanning has been directly and successfully applied to protein mixture analysis [39]. A B/E linked scan allows the transmission of one parent ion and examines the daughter products. Hence, for the analysis of a pure peptide the computer is programmed to calculate the B/E scan-law relationship for that particular molecular ion, with subsequent detection of the daughter ion products from the molecular ion, $[M+H]^+$, only. Therefore, for each parent ion transmitted there is calculated B/E scan-law relationship which differs for each peptide and its daughter fragments examined. The success reported below with pure peptides promoted the application of this method to a true biological research problem investigating peptide components of an **unresolved** proteolytic digest mixture. The application of linked scanning as described here represents a method for providing valuable sequence data from peptides and components of an unfractionated digest mixture using a two sector mass spectrometer. The data generated from this linked scan study of pure peptides has been shown to be fully comparable with data obtained from a four sector instrument. This novel application of linked scan/CAD to a complex digest mixture of a large protein has been demonstrated in this thesis work to be a rapid and sensitive technique. An important aspect to this approach to **direct mixture analysis** is that the rate-limiting step of purification has been removed.

Coinciding with this investigation into the applicability of linked scanning/CAD to unfractionated digest mixtures was the analysis of the primary sequence of the enzyme pyruvate decarboxylase. The

investigation of the primary sequence analysis is reported in Chapter Five of this thesis. The linked scan analysis of an intact proteolytic digest of pyruvate decarboxylase, (PDC), allowed the sequence of two unknown components to be determined.

In addition to the investigation of the applicability of B/E linked scanning/CAD for peptide sequence analysis, it was important to examine, using the same peptides, a range of CAD conditions. A set of experiments which compares spectra produced in the absence of a collision gas (normal metastables) with those generated using collision gases (helium and air), is reported later in this chapter.

The data presented below illustrates that the linked scanning/CAD method, as applied here, will contribute valuable information to modern sequence analysis studies of biopolymers.

4.2 Results and Discussion

With the increased sensitivity, of currently available mass spectrometers, and improved software for computer averaging of summed scans (MCA), B/E linked scanning/CAD was evaluated, on a ZAB-2SE HF two sector mass spectrometer for;

(a) sequence analysis of pure peptides, and

(b) **direct** mixture analysis of unknown peptide components of a complex proteolytic digest mixture.

For each peptide examined a B/E scan-law is computed to allow

transmission and detection of the daughter ions of each peptide, which are enhanced by collisionally activated decompositions (CAD). The daughter fragments of peptides are postulated to arise from two main types of fragmentation mechanisms [27], N- and C-terminal described fully in Chapter One. These species may undergo further fragmentation with an N-terminal fragment ion (acylium) forming an aldimine fragment ion by loss of carbon monoxide. Similarly, C-terminal fragment ions may also fragment further to form the alkyl species.

The daughter ions are enhanced, in the first field free region (1st FFR), by collisionally activated decompositions, CAD. The collision gas used (unless stated otherwise) was helium at a pressure (1.5 x 10^{-7} Torr) designed to reduce parent ion intensity to approximately 30% of its original intensity. It should be noted that all spectra in this chapter were recorded by computer, hence, each peak is assigned an accurate mass. The difference between accurate and nominal mass becomes apparent for mass values above 1000 a.m.u. This mass difference arises from certain elements such as hydrogen being mass sufficient (accurate mass 1.00785) and others such as oxygen being mass deficient (15.994915 a.m.u.). In general an increase of approximately half a mass unit per 1000 a.m.u. can be taken as a conversion factor for nominal to accurate mass.

PART I - Linked Scanning of Pure Peptides

4.2.1 B/E Linked Scan/CAD Analysis of Leu-enkephalin amide

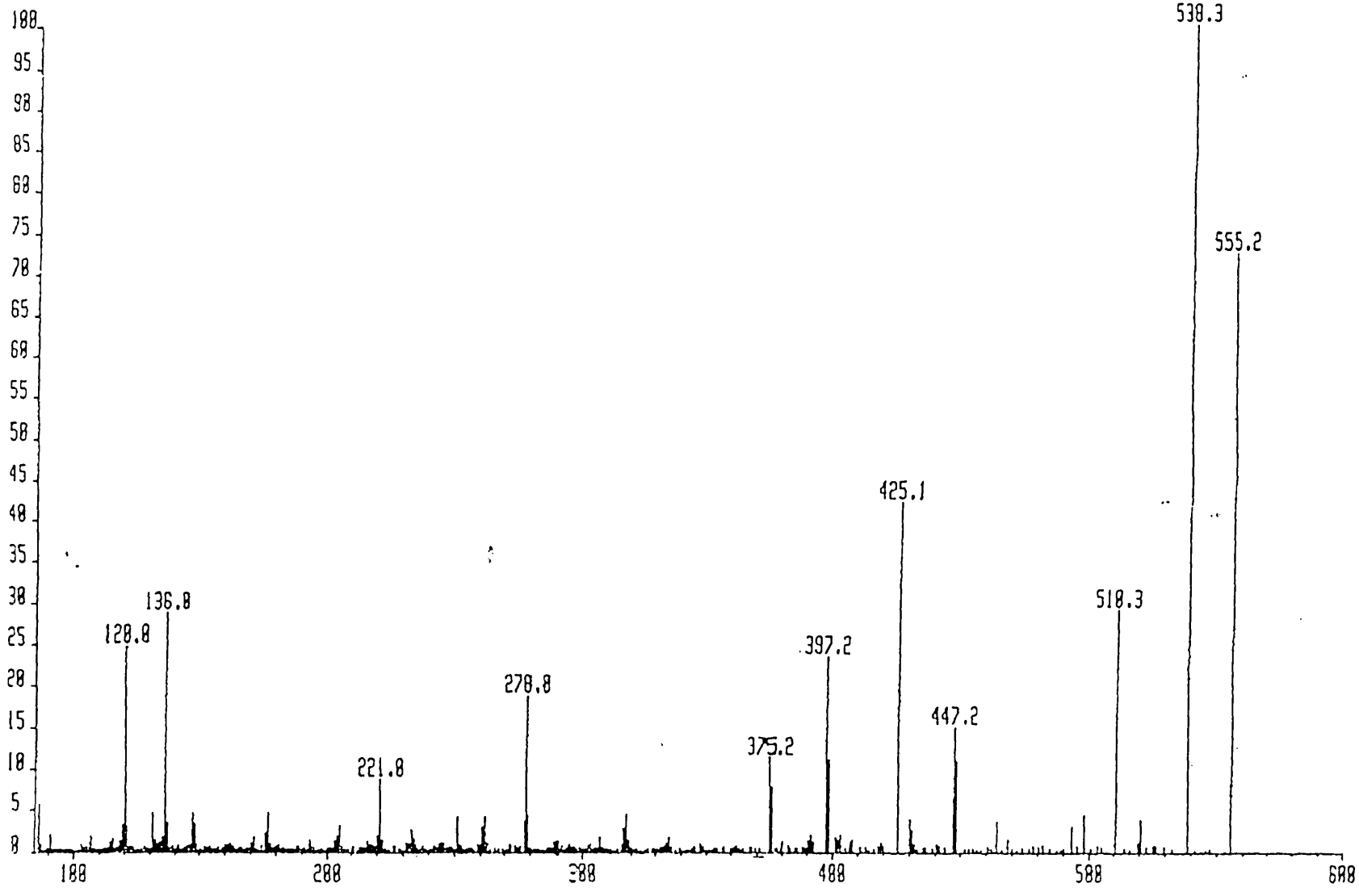


Figure 4.1 Two-sector B/E Linked Scan/CAD (He) Spectrum of Leu-enkephalin amide.

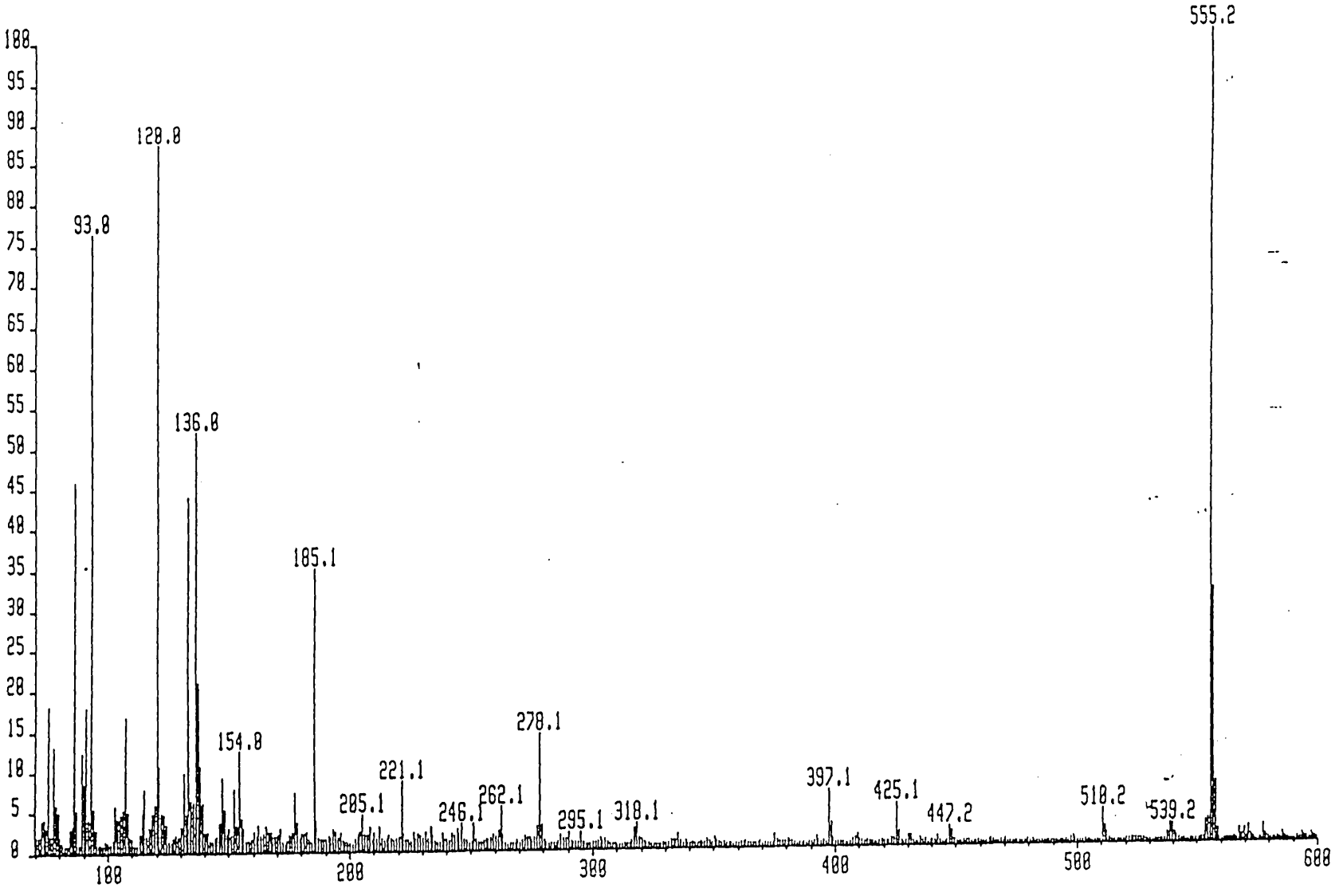


Figure 4.1.1 FAB-MS of Leu-enkephalin amide (ZAB2E).

The B/E linked scan/CAD spectrum of Leu-enkephalin amide, $[M+H]^+$ 555, with helium gas (1.5×10^{-7} Torr) in the collision chamber, is presented in Figure 4.1. From intense signals; m/z 136, 221, 278, 397, 425 and 510, determined to be a mixture of acylium and aldimine fragment ions, the sequence of Leu-enkephalin amide can be deduced by the mass differences between these fragment ions. The sequence is confirmed as Tyr-Gly-Gly-Phe-Leu-amide. It is interesting to note that the normal FAB spectrum of Leu-enkephalin amide shown in Figure 4.1.1 contains virtually the same fragments as those produced from the B/E linked scan/CAD. However, on comparison of the two spectra it is observed that the fragment ions in the linked scan spectrum are of much higher intensity. The sequence of enkephalin was, in fact, originally deduced to be a mixture of two pentapeptides by electron impact MS, [14], and found to consist of two related peptides differing in their C-termini, and hence are referred to as Leu- and Met-enkephalin. The extent of fragmentation observed from FAB-MS analysis of peptides is an unpredictable occurence, with the FAB spectra of some peptides containing sufficient fragmentation data for sequence analysis or partial sequence determination. The presence of interpretable fragment ions in the normal FAB mass spectrum of Leu-enkephalin amide is a fortuitous example and in general, FAB-generated spectra exhibit little fragmentation data. Thus, as sequence analysis is of the utmost importance complementary methods to FAB-MS, combining high sensitivity and rapid production of interpretable data, for structure elucidation are in great demand.

With the promising results obtained from the B/E linked scan

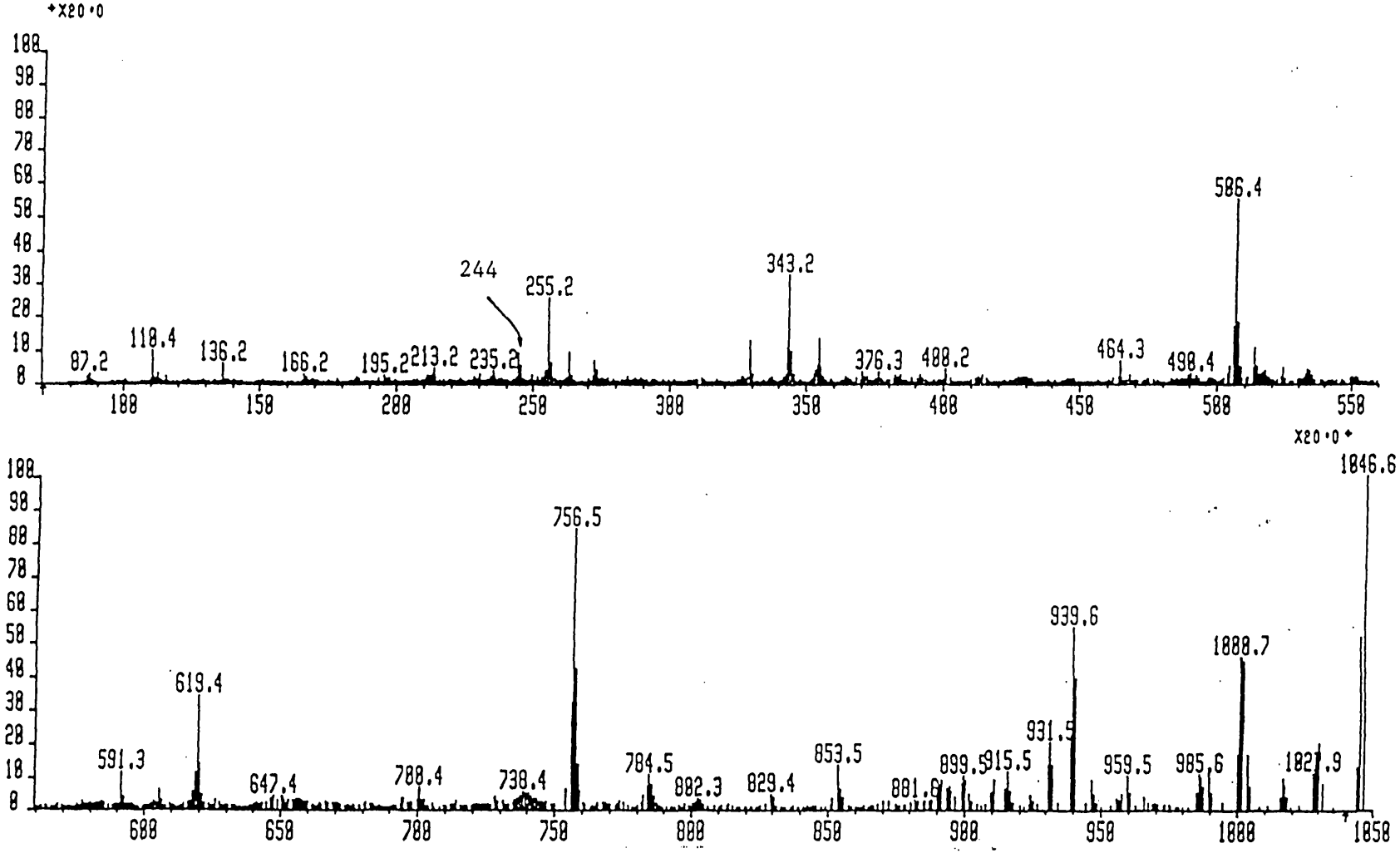


Figure 4.2 Two-sector B/E Linked Scan/CAD (He) Spectrum of Angiotensin II.

analysis of the small peptide, Leu-enkephalin amide, this study was subsequently extended to test the applicability of this method to larger peptides for sequence determination.

4.2.2 B/E Linked Scan/CAD Analysis of Angiotensin II

The linked scan/CAD spectrum of angiotensin II, $[M+H]^+$ 1046, is shown in Figure 4.2. The intense signals at m/z; 88, 244, 343, 506, 619, 756, 853 and 1000 correspond to N-terminal aldimine fragment ions. The sequence, Asp-Arg-Val-Tyr-Ile-His-Pro-Glu-Phe, is deduced from the mass differences between these intense signals, corresponding to amino acid residue masses. This spectrum is characterised by intense fragment ions with a very low signal to background noise ratio. There is also a relatively weak partial C-terminal fragment ion sequence m/z; 166, 263, 400, 513, 931 and 1046. The presence of these C-terminal fragment ions, sufficiently weak so as not to confuse sequence assignment from the N-terminal aldimine fragment ions, serve as being confirmatory.

4.2.3 B/E Linked Scan/CAD Analysis of Renin Substrate

The B/E linked scan/CAD spectrum of renin substrate, $[M+H]^+$ 1758, is presented in Figure 4.3. By calculating the mass differences between each of the intense signals, m/z; 343, 506, 619, 756, 853, 1000, 1137, 1250, 1363, 1462 and 1625, together with the quasi-molecular ion, $[M+H]^+$ 1758, observed in the linked scan spectrum, the sequence, Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu-Leu-Val-Tyr-Ser, is confirmed. These intense signals are determined to correspond to aldimine fragment ions. It should be noted that the mass values of

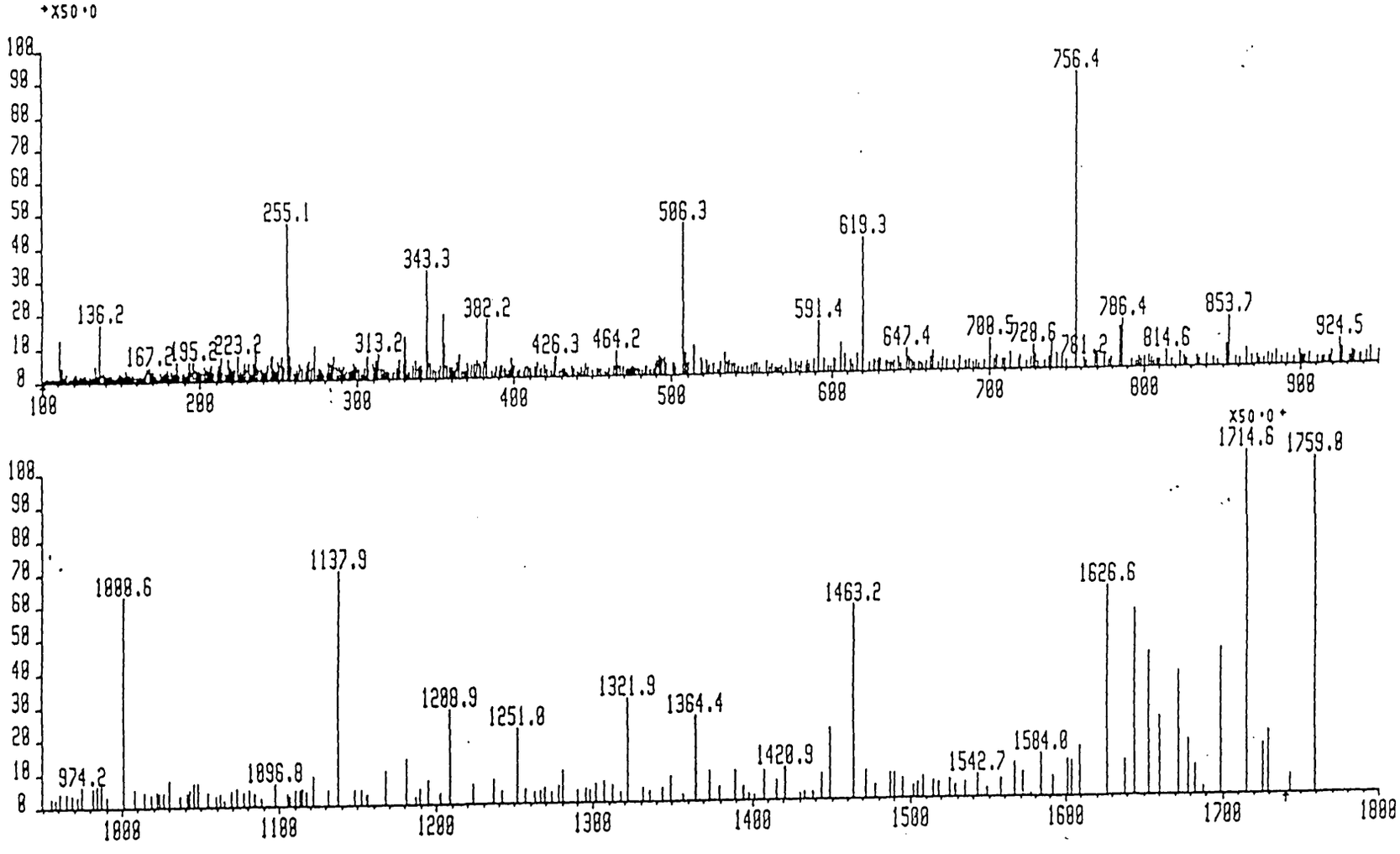


Figure 4.3 Two-sector B/E Linked Scan/CAD (He) Spectrum of Renin Substrate (ZAB2E).

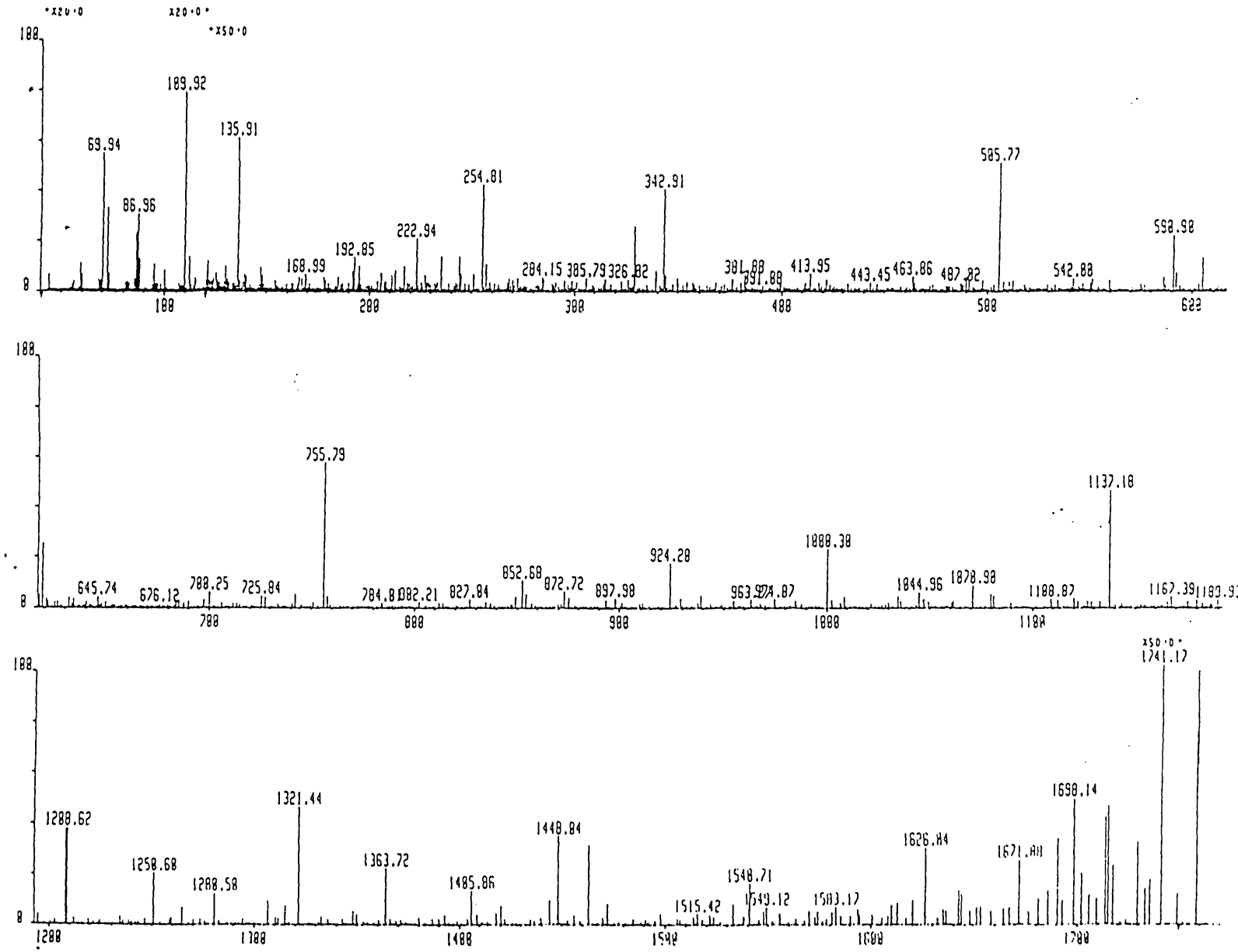


Figure 4.3.1 Four-sector CAD Spectrum of Renin Substrate (ZABSE-4F).

the fragment ions listed are nominal values whereas those in the spectrum, Figure 4.3, are the accurate masses. Figure 4.3.1 shows a spectrum remarkably similar to that in Figure 4.3. This is the spectrum of renin substrate on a four sector MS/MS instrument. The quality of the linked scan spectrum of the tetradecapeptide is fully comparable with that generated from a four sector instrument, with the background signals for pure peptides shown to be of similar intensity for both two sector linked scan and tandem mass spectrometry. The two sector B/E linked scan/CAD spectrum of renin substrate, Figure 4.3, containing easily interpretable intense mass signals, has been generated from less than 1 nanomole of material, indicating that this is a highly sensitive procedure for peptide sequence analysis.

The data presented above illustrates the applicability of B/E linked scan/CAD analysis for rapid sequence analysis of pure peptides at high sensitivity. To determine if this technique has a mass limitation for attaining sequence ions in sufficient abundance for the entire amino acid sequence to be deduced, as with the above examples, this B/E linked scanning/CAD study was subsequently extended to examine a large peptide, human calcitonin gene related peptide (hCGRP).

A B/E linked scan/CAD spectrum of the 37 amino acid residue peptide, human calcitonin gene related peptide (hCGRP) with a molecular weight 3786 a.m.u., indicated a possible limitation for the applicability of this method to peptides with a mass exceeding approximately 2000 a.m.u. The linked scan spectrum contained a

limited selection of N-terminal acylium and aldimine fragment ions. However, it was not possible for a partial sequence to be deduced as these fragments were random. Despite this apparent limitation, linked scanning may still be useful for application to large peptides such a hCGRP in sequence analysis studies where the sequence, or partial sequence, has been determined by other techniques. The role of linked scanning would be to efficiently provide a highly sensitive and rapid procedure for partial sequence confirmation by the provision of random fragment ions which can be correlated to the elucidated sequence. A definite advantage of observing random fragment ions is that there is an equal probability of observing signals corresponding fragments derived from the C-terminal as well as the N-terminal region of a large peptide.

During the course of this thesis work investigating the applicability of B/E linked scanning/CAD on a two sector MS for peptide sequencing Nibbering *et al* [98] recently and independently investigated tandem mass spectrometry and collisional activation, FAB/MS/CA/MS, for the structure determination of several beta-endorphin fragments and two adrenocorticotropic hormone peptides by linked scanning of their daughter products. From their study they observed that (i) N-terminal fragment ions were slightly more abundant than C-terminal however, they determined this feature to be compound dependant, (ii) other collision gases such as argon, xenon and hexafluoride were used but did not generate a noticeable enhancement of the fragmentation of the protonated species, and (iii) the diminishing utility of CA of

molecules as molecular mass increases.

The study presented in this thesis demonstrates that B/E linked scanning/CAD, on a two sector instrument, of pure peptides (with masses comparable to those used by Nibbering and co-workers) generates high quality data suitable for peptide sequence analysis. A similar mass limitation for the applicability of this method has also been reported here, and in addition an alternate use of this method for those peptides with molecular weight exceeding 2000 a.m.u. has been proposed. However, this investigation was not restricted to pure peptide analysis only. The main research objective once the applicability of linked scanning had been established for pure peptide analysis was the study of linked scan methods for **direct** mixture analysis using two sector instrumentation.

4.3 PART II- Direct Mixture Analysis by Linked Scanning

Purification procedures such as HPLC, for the fractionation of mixtures for the provision of pure components for subsequent analysis by classical protein sequencing techniques, have proven invaluable for protein sequence studies. This has been shown to be particularly so when a generated protein digest is initially analysed by FAB-MS. Due to the phenomenon of suppression [27] of certain components in the mixture by others, on FAB-MS analysis of the unfractionated digest, many of the components are not observed in the FAB mass spectrum hence, the need for purification

procedures is apparent. However, with advances in DNA-sequencing techniques the need to sequence all components of a protein digest mixture, in certain structural analysis studies, has been removed. Instead the emphasis lies on the provision of a stretch of amino acid sequence, typically between six to twelve residues, to employ for generation of complementary synthetic oligonucleotide probes to screen a gene library. Therefore, a method with the ability to rapidly screen for a sequence or partial sequence of unseparated components in a peptide mixture would be of great value with the rate-determining step of purification removed. Presented in this chapter is the first data demonstrating the novel application of two-sector B/E linked scanning/CAD to a complex mixture analysis problem, for solving the structures of two unknown components of a proteolytic digest of the enzyme pyruvate decarboxylase, molecular weight 60,000 Da.

Coinciding with this investigation of linked scans was the analysis of the primary structure of the enzyme pyruvate decarboxylase, the gene sequence of which had recently been published [98]. This study of the primary structure of the enzyme pyruvate decarboxylase, isolated from wild type yeast, *Saccharomyces cerevisiae*, is reported in Chapter Five. A tryptic digestion of pyruvate decarboxylase resulted in the formation of a complex peptide mixture. An aliquot of this proteolytic digest mixture was analysed by FAB-MS. The resultant FAB-Map is shown in Figure 4.4. A simple computer programme was employed (R. McDowell, Dept. Biochemistry, Imperial College, London. SW7 2AZ) into which the gene-derived sequence of the enzyme was read. For each mass, in the FAB

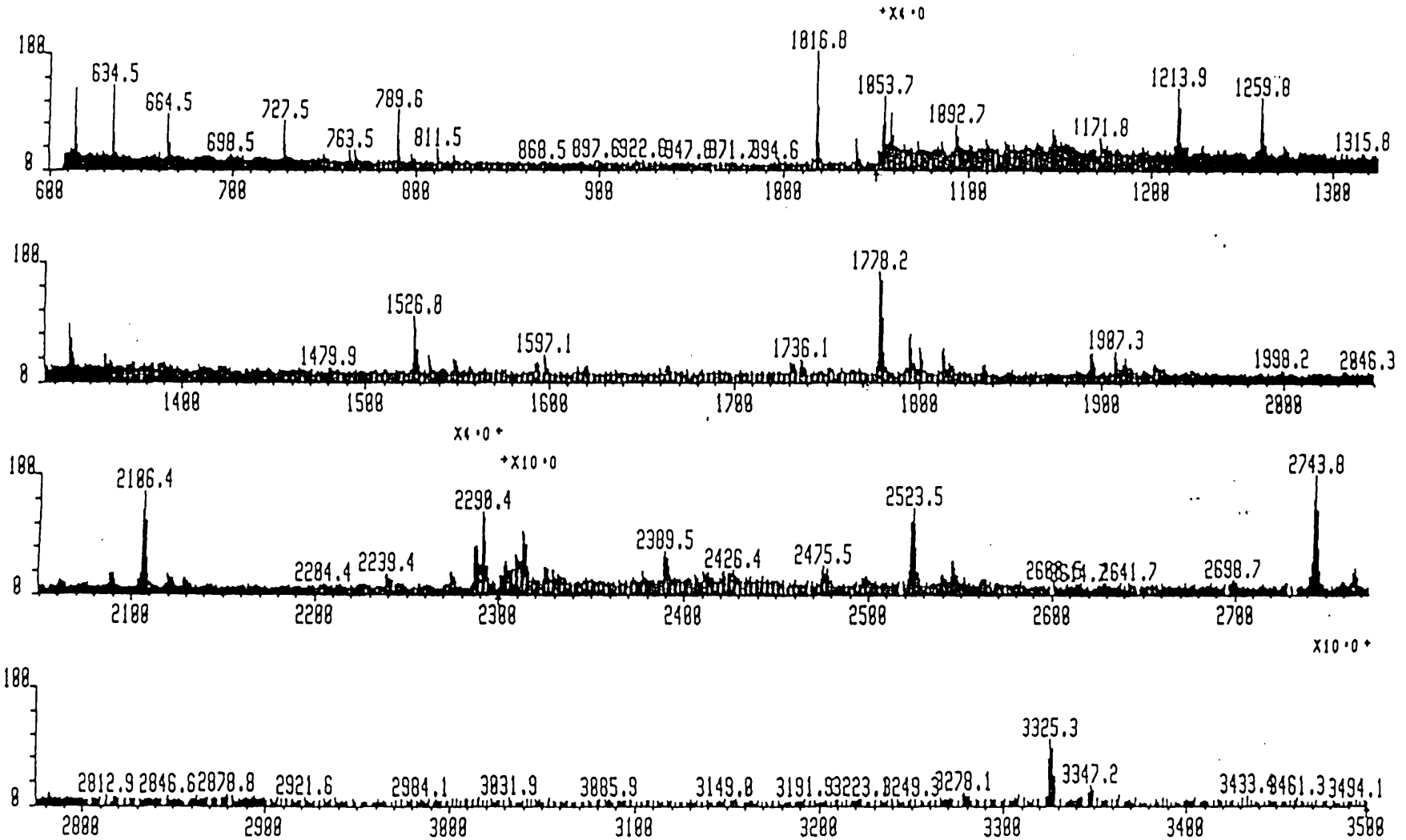


Figure 4.4 Tryptic FAB-Map of Pyruvate Decarboxylase (A1).

spectrum, the computer calculates all possible peptide sequences arising from the gene-derived sequence for a particular mass. Utilising the knowledge of the type of digestion employed and hence predicted cleavage sites, most of the possible sequences for each mass value can be discounted, with only one peptide sequence satisfying the mass value and enzyme specificity. It was discovered that on checking the observed masses in the FAB-Map, Figure 4.4., against those predicted from the gene-derived sequence, by trypsin digestion, that some signals could not be assigned to the predicted protein structure. In particular, intense signals m/z 789 and m/z 1016, could not be mapped on to residues in the gene-derived protein sequence, suggesting either a post-translational modification and/or a difference between the actual protein sequence and that deduced from the gene sequence. The study presented here demonstrates for the first time that B/E linked scanning/CAD is a highly appropriate method for direct mixture analysis as illustrated by the application of this method to the solution of two components of a complex proteolytic digest of the enzyme pyruvate decarboxylase, PDC.

Before approaching a real sequence analysis problem to determine unknown components of the PDC proteolytic digest mixture a known "contaminant", angiotensin II, $[M+H]^+$ 1046, was mixed with an aliquot of the tryptic digest mixture of PDC. This peptide was selected on the basis of its molecular weight, as it is similar in mass to one of the unidentified signals m/z 1016, in the FAB-Map of the digest (Figure 4.4). Since the width of the mass window one can tune to on a two-sector instrument is approximately 5 a.m.u

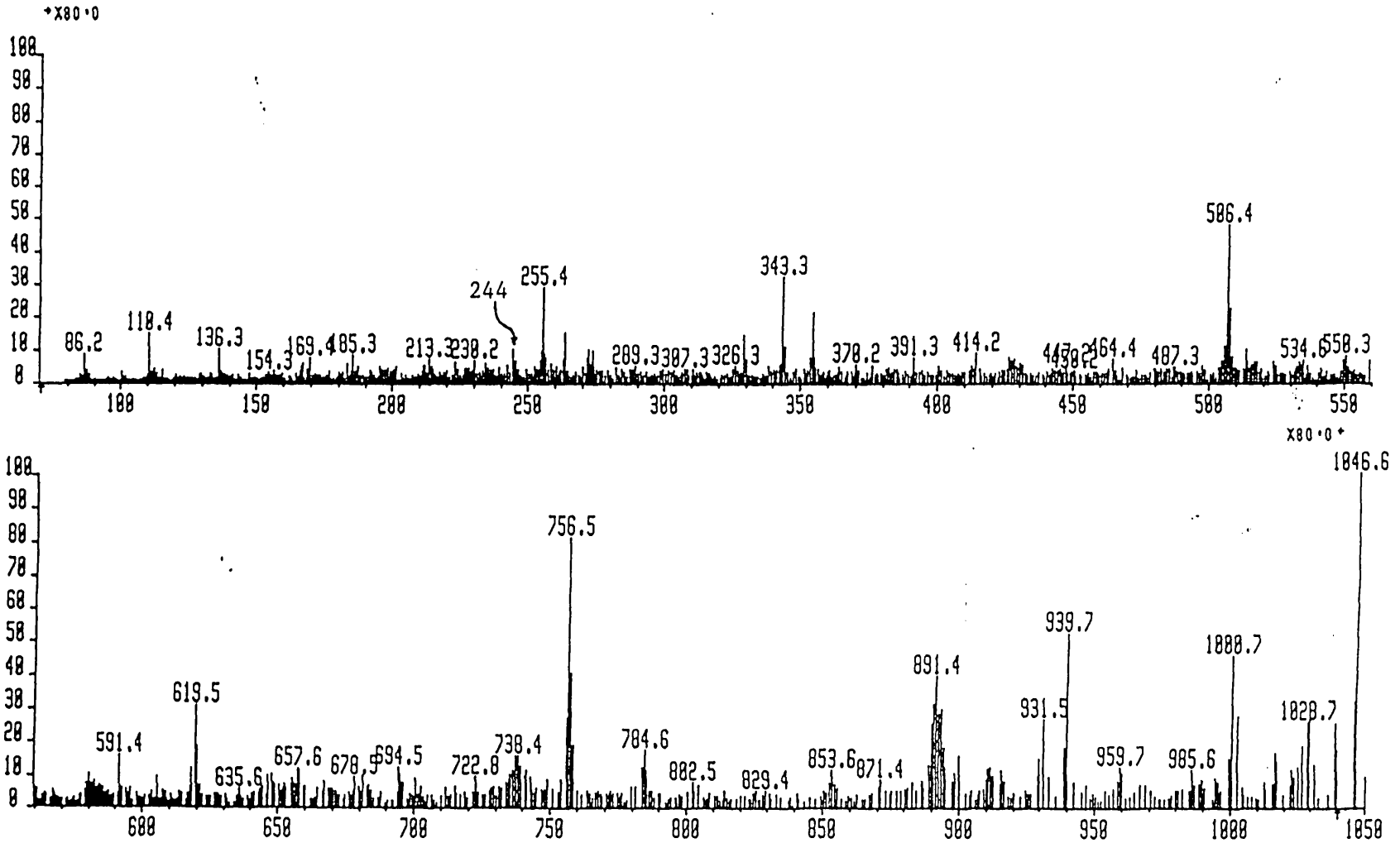


Figure 4.5 Two-sector B/E Linked Scan/CAD (He) Spectrum of the Component m/z 1046 in the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.

angiotensin II (mol.wt. 1045) is fully resolvable from m/z 1016. The amount of angiotensin II added (approximately 1 microgramme, 950 picomoles) was estimated so as to be of similar concentration as the other components of the digest. On FAB-MS analysis of an aliquot of this mixture the signal intensity of the angiotensin II molecular ion was shown to be approximately 80% of that of the peptide component at m/z 1016. A linked scan/CAD spectrum, Figure 4.5, of the angiotensin II component of the artificial mixture generated is shown to be of high quality. By the mass differences beteeen the intense N-terminal aldimine fragment ions at m/z; 88, 244, 343, 506, 619, 756, 853, and 1000, the complete sequence of angiotensin II was determined. The quality of this spectrum is remarkably similar to the B/E linked scan/CAD spectrum of the pure sample of angiotensin II as shown in Figure 4.2 with only a slight increase in the background noise observed. The data from this experiment was most encouraging for the application of B/E linked scanning/CAD to the sequence analysis of the unknown peptide components $[M+H]^+$ 789 and $[M+H]^+$ 1016, formed by trypsin digestion of pyruvate decarboxylase (PDC).

4.3.1 **Linked Scan Analysis of m/z 789**

The linked scan/CAD spectrum of m/z 789, shown in Figure 4.6, was generated from a single loading of an aliquot of the unfractionated tryptic digest mixture. From the mixture of N- and C-terminal fragment ions at m/z; 130, 204, 259, 317, 372, 418, 473, 531, 643, 660 and 789, the sequence of this peptide was determined. Interpretation of the spectrum initially allowed a relationship between the following set of intense signals to be defined, m/z

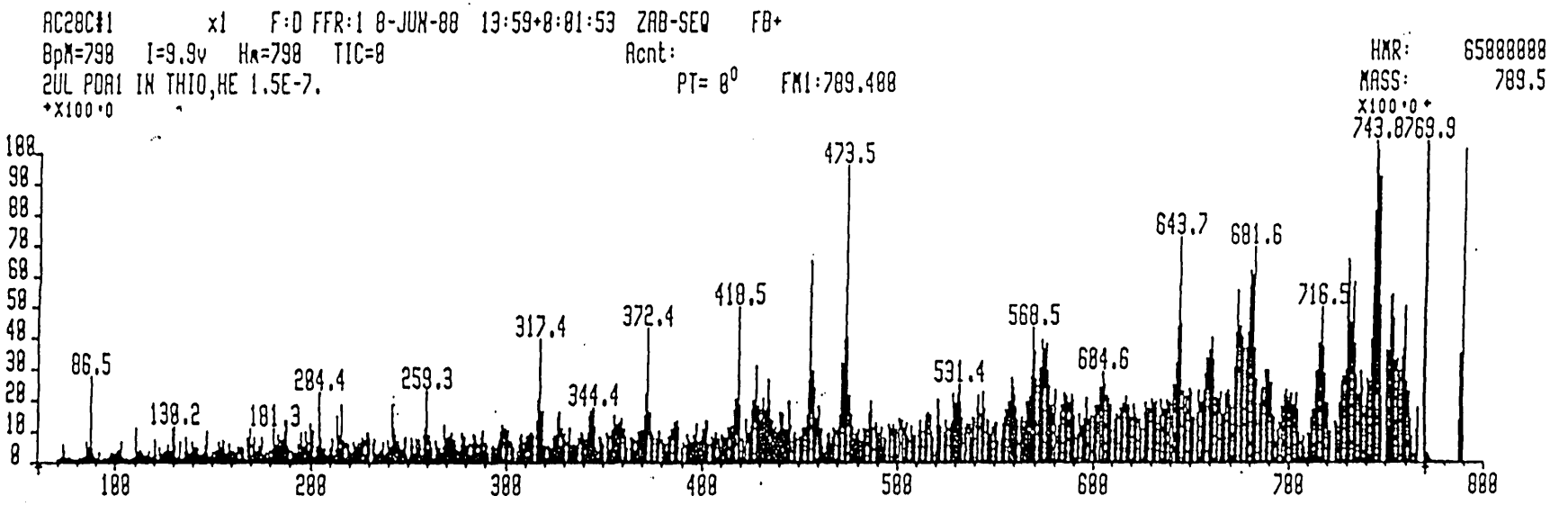
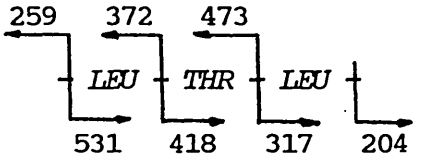


Figure 4.6 Two-sector B/E Linked Scan/CAD (He) Spectrum of the Unknown Peptide Component m/z 789 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.

204, 259, 317, 372, 418, 473 and 531. These signals corresponded to a mixture of N- and C-terminal fragment ions as illustrated below;



It was necessary to establish which set of signals were N-terminal fragment ions and those which were C-terminal, in order for the full amino acid sequence of the peptide to be determined. Utilising the knowledge that the peptide at m/z 789 was derived from a tryptic digest it was logical to assume that the C-terminal amino acid residue of this peptide was either a lysine or arginine residue (tryptic cleavage sites). Mass values 204 and 259 were both tested for C-terminal lysine or arginine, with only m/z 204 giving a sensible fit for *Gly-Lys*. Therefore, the signals at m/z; 259, 372, and 473 were assigned as N-terminal fragment ions. This inference was correlated with the spectrum signals, with the intense signal at m/z 643 now identifiable as the subsequent N-terminal fragment ion in the sequence series. Thus, the amino acid sequence able to be determined was; *Leu-Thr-Leu-Gly-Lys* (with leucine indistinguishable from isoleucine). This pentapeptide, LTLGK, was found to correspond to residues 4 to 8 in the gene-derived sequence (Chapter Five, Figure 5.1.3) i.e., the N-terminal region of the intact enzyme. However, the predicted N-terminal tryptic peptide expected to be formed, by trypsin digestion, from the protein sequence is an octapeptide, residues 1 to 8. The sequence of this predicted octapeptide is,

Met-Ser-Glu-Ile-Thr-Leu-Gly-Lys, $[M+H]^+$ 878. The molecular weight of this peptide clearly does not agree with the observed mass of the partially identified peptide at m/z 789. There is an obvious mass difference between these two peptides of 89 a.m.u. which suggests that the difference is not merely a simple truncation of the N-terminus of the expressed protein, as 89 a.m.u. does not correspond to an amino acid residue mass value. This raised the posibilities of either a difference in the amino acid sequence preceding the pentapeptide and/or that there is a post-translational modification. The presence of the signal m/z 130 in the linked scan spectrum proves the residue preceding isoleucine as glutamic acid. However, the residue mass of serine (preceding glutamic acid) is 87 a.m.u. and thus, there remains 43 a.m.u. unaccounted for. A test for possible blocking groups on the serine residue provided an answer by suggesting an N-acetyl blocked serine residue.

Thus, the structure of the peptide at m/z 789, determined to correspond to the N-terminus of the protein pyruvate decarboxylase, was determined from the B/E linked scan/CAD spectrum, of the unfractionated peptide component m/z 789, to be; *Ac-Ser-Glu-Ile-Thr-Leu-Gly-Lys*. The blocked N-terminus of this peptide was confirmed later by a single step of subtractive Edman degradation [2], of the intact digest mixture, followed by FAB-MS analysis. The FAB-Map after one step of Edman showed no shift in the signal m/z 789, confirming the presence of the blocked N-terminus. Hence, a processing event and a post-translational modification have been identified from a B/E linked scan/CAD

analysis of an unseparated component of a complex proteolytic digest mixture.

4.3.2 Linked Scan Analysis of m/z 1016

Another intense signal from the FAB-Map, Figure 4.4 of the tryptic digest of PDC, for which no mass match with the gene-derived sequence of pyruvate decarboxylase could be made was m/z 1016. The B/E linked scan/CAD spectrum of m/z 1016 in the intact tryptic digest mixture is shown in Figure 4.7. This spectrum contains a number of intense signals which were interpreted in the same manner as before. From the intense signals observed at m/z; 218, 289, 404, 475, 588, 689, 790 and 903, the sequence of this peptide, $[M+H]^+$ 1016 was able to be deduced. The mass differences between the intense fragment ions at m/z 404, 475, 588, 689, 790 and 903 allowed the sequence, *Leu/Ile-Thr-Thr-Leu/Ile-Ala*, to be determined. Analysis of the lower end of the spectrum enabled assignment of signals at m/z 289 and 218, with mass differences allowing the sequence *Asp-Ala* to be eluicidated. It was not known whether these signals corresponded to N- or C-terminal fragment ions. However, by applying the knowledge of the digest specificity (trypsin) both signals m/z 903 and 218 were tested for a lysine or arginine fit. The signal at m/z 218 agreed for the C-terminal fragment ion, Ala-Lys, thereby, determining the mass signals as C-terminal and allowing the complete sequence of the peptide at m/z 1016 to be determined as;
Leu-Leu-Thr-Thr-Ile-Ala-Asp-Ala-Ala-Lys. This sequence was deduced solely from the linked scan spectrum, and found to closely agree with the gene-derived sequence between residues 332 to 341.

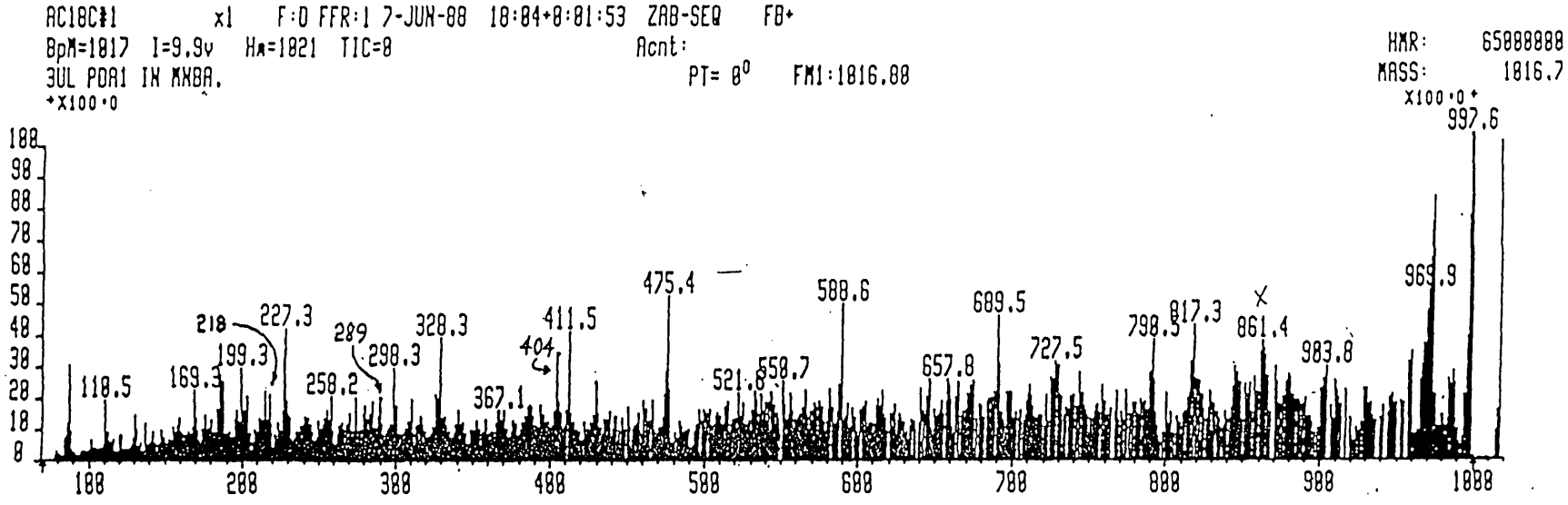


Figure 4.7 Two-sector B/E Linked Scan/CAD (He) Spectrum of the Unknown Component m/z 1016 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.

However, the peptide sequence between residues 332 and 341 differs to the sequence determined by linked scanning in one position, residue 335. Residue 335 has been translated from the gene sequence as asparagine and not threonine, as clearly determined from the linked scan spectrum. Hence, the linked scan data has allowed identification of an amino acid difference to that of the gene-derived sequence. This finding was subsequently confirmed by several steps of subtractive Edman degradation.

Reported earlier in this chapter was the limitation encountered for attaining total sequence information from pure peptides by this method. It was found that sequence data was generated from peptides of mass greater than 2000 a.m.u. however, this data was not sufficient for complete sequence determination. This observation was shown to be true for unseparated components of a digest mixture with molecular weight of approximately 2000 a.m.u. or greater. However, rapid generation of partial fragmentation data can be of great use for confirmation of sequence, as discussed above. This can be illustrated by the linked scan spectrum of the peptide observed at m/z 1778 (nominal mass 1777).

4.3.4 **Linked Scan Analysis of m/z 1778**

The results obtained from from the linked scan analysis of components m/z 789 and 1016 enabled two differences to the recently published gene-derived sequence to be identified. These findings underline the need for independant checks of stretches of amino acid residues in cases where the sequence is solely derived by DNA-sequencing methods. However, in situations where a signal from

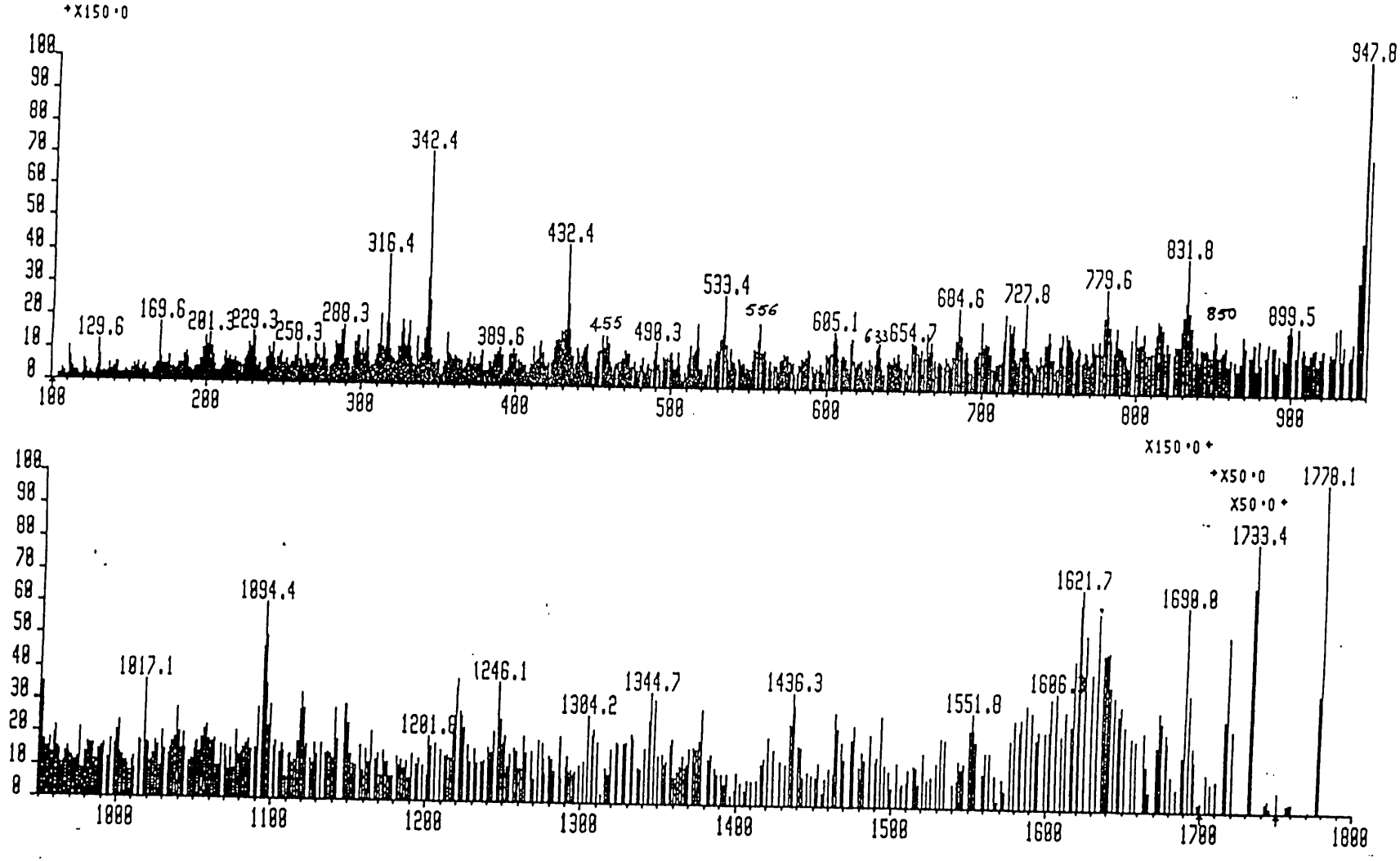


Figure 4.8 Two-sector B/E Linked Scan/CAD (He) Spectrum of the m/z 1778 Component of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.

a FAB-Map of a digest mixture can be assigned to a stretch of amino acid sequence it is not necessary to analyse the fragmentation data exhaustively, but rather to check that some of the anticipated fragment ions are observed. The intense signal m/z 1778 observed in the tryptic FAB-Map (Figure 4.4) of the enzyme PDC was able to be mass matched, by the computer search programme, onto the gene-derived sequence between residues 233 and 248. The predicted amino acid sequence between these residues is; Leu-Ile-Asp-Leu-Thr-Gln-Phe-Pro-Ala-Phe-Val-Thr-Pro-Met-Gly-Lys, (*LIDLTQFPAFVTPMGK*). The B/E linked scan/CAD spectrum of m/z 1778 in the intact tryptic digest mixture is presented in Figure 4.8. The intense signals in the linked scan/CAD spectrum are more than sufficient to confirm the assignment of this signal to the peptide originating from PDC, residues 233 to 248. The sequence is confirmed by the N-terminal sequence ions at m/z; 342 (*LID*), 831 (*LIDLTQF*) and 1346 (*LIDLTQFPAFVT*), and the C-terminal fragment ions, m/z; 432 (*PMGK*), 533 (*TPMGK*), 947 (*PAFVTPMGK*) and 1094 (*FPAFVTPMGK*).

This application of linked scanning, as previously suggested, provides rapid sequence data at high sensitivity which enables, as demonstrated above, the peptide sequence assigned to be a particular mass signal to be confirmed with confidence. Another advantage of this technique, in addition to the high sensitivity and the rapid generation of data, noted earlier when discussing this application to larger peptides, is that there is an equal probability of observing mass signals corresponding to peptide fragments derived from the N- and C-terminal regions of the peptide

sequence.

All mixture analysis spectra were generated from a single 1-2 nanomole loading of the intact tryptic digest mixture.

4.4 Variation of CAD Conditions

Linked scanning allows detection of the daughter products of a particular parent ion with collisionally activated decompositions (CAD) enhancing fragment ion formation. All spectra presented above have been generated using helium as the collision gas. Another feature of this research work was to investigate fragment ion enhancement using a range of CAD conditions.

4.4.1 B/E Linked Scanning in the Absence of a Collision Gas

Illustrated in Figure 4.9 is the B/E linked scan spectrum of the peptide at m/z 1016 in the intact tryptic digest mixture without the use of a collision gas i.e. the non-activated daughter spectrum. Though the sequence data is still useable it is clearly not as high quality as that previously generated utilising a collision gas, Figure 4.7. The major difference between the two spectra in Figures 4.7 and 4.9, is that without a collision gas, the lower end of the spectrum is not as intense due the lack of high energy collisions which enhances the fragmentation. However, interpretation of the non-activated linked scan spectrum, Figure 4.9, still allows the sequence of peptide $[M+H]^+$ 1016 to be readily determined. This example illustrates that interpretable linked scan data can generated even when a collision gas is not

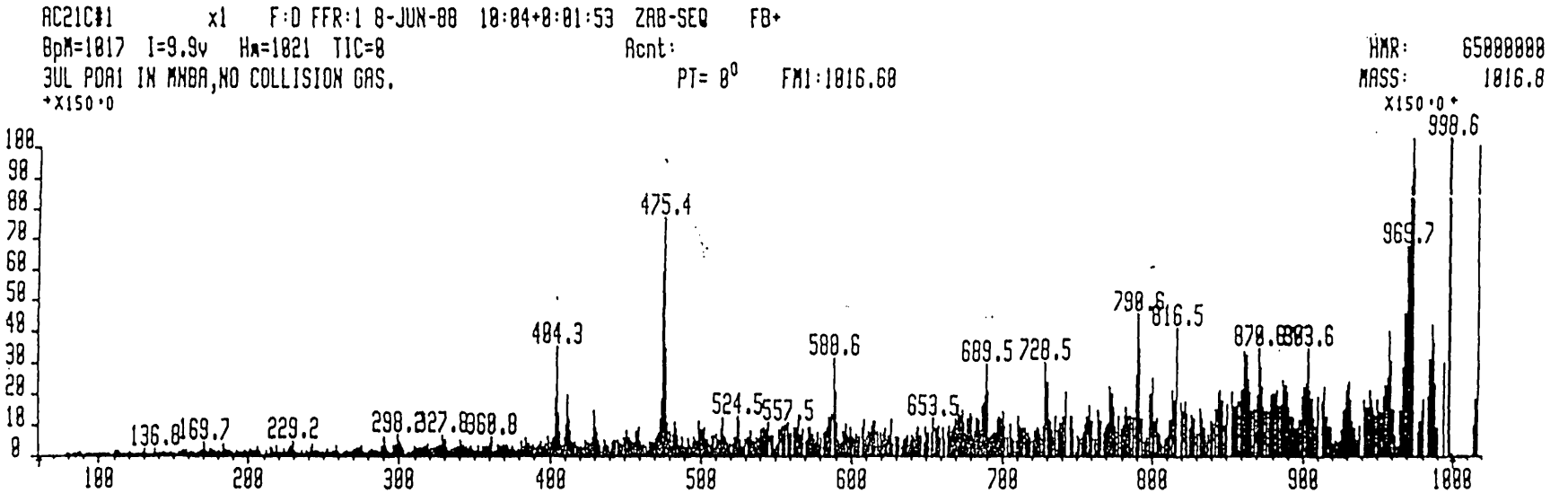


Figure 4.9 Two-sector B/E Linked Scan Spectrum of the Unknown Component m/z 1016 of the Intact Tryptic Digest Mixture of Pyruvate Decarboxylase.

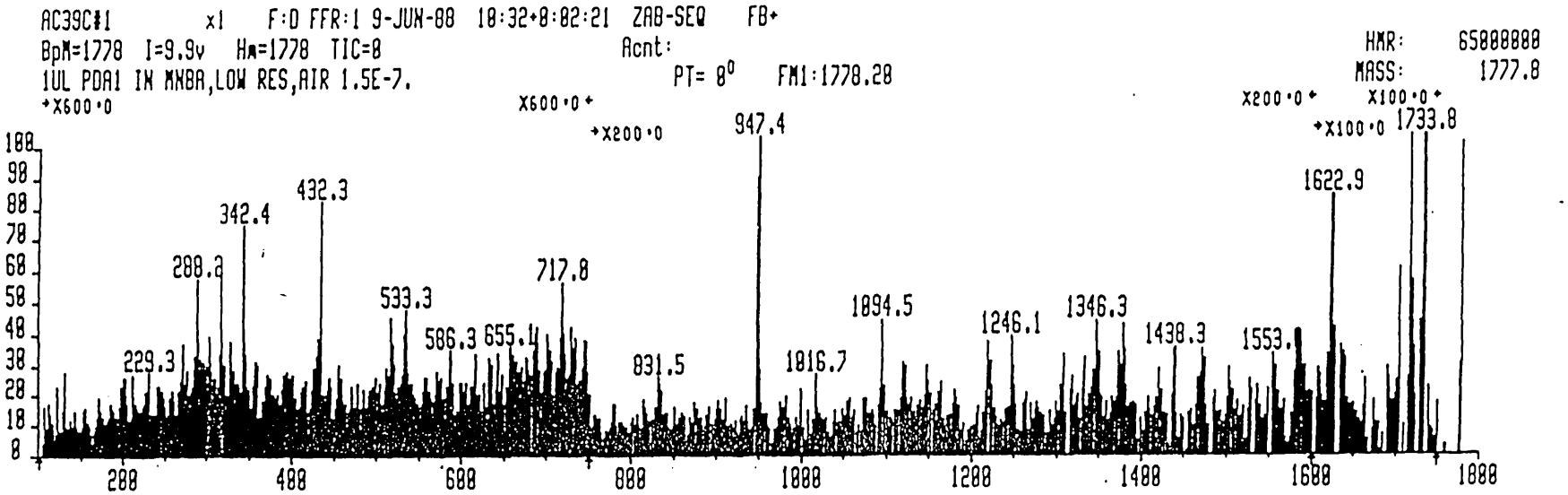


Figure 4.10 Two-sector B/E Linked Scan/CAD (Air) Spectrum of the m/z 1778 Component of the Tryptic Digest Mixture of Pyruvate Decarboxylase.

used.

4.4.2 B/E Linked Scanning/CAD (Air)

The B/E linked scan/CAD spectrum of m/z 1778 in the intact proteolytic digest mixture, utilising air as the collision gas is shown in Figure 4.10. This spectrum contains intense signals comparable to those observed in the linked scan/CAD spectrum generated with helium as the collision gas. However, the spectrum (Figure 4.10) with CAD by air can be observed to have a slightly higher signal to background noise ratio than when helium is the collision gas, i.e. the spectrum is not as "clean".

Therefore, the presence of a collision gas would be preferable to a vacuum, with a significant enhancement of fragment ions in the low mass end of the spectrum observed, aiding sequence analysis.

It was determined that using mono-nitrobenzyl alcohol,(MNBA), as a matrix, enhanced the sample life-time, with better quality spectra obtained by MCA, in comparison to those samples loaded in glycerol/thioglycerol matrix. The sample quantity used ranged from 100 picomoles (for pure peptide analysis) to 2 nanomoles (for mixture components).

4.5 Conclusion

This research work has investigated the application of B/E linked scanning/CAD to the sequence analysis of

(i) pure peptides, such as those obtained after purification by HPLC, and more significantly the novel approach to the analysis of, (ii) unfractionated components of a complex proteolytic digest mixture.

High quality data has been generated from pure peptides, with the linked scan/CAD spectrum of renin substrate being fully comparable to that obtained from a tandem mass spectrometer. This method has been proven to be sensitive with some spectra generated from as little as 200 picomoles of material. Good fragmentation data are obtainable to masses of 2000 a.m.u. which allow full sequence analysis for both pure peptides examined and those unseparated components of a digest mixture. It had been anticipated that the background contribution from the matrix and other decomposition products, in B/E linked scan spectra, would be significant and considerably more than that from a four-sector instrument since the width of the mass window is larger than on a four-sector instrument. However, the mixture analysis data indicates that the background contribution is negligible and is not limiting up to mass 2000 since, despite a higher background than that which would be present in a four-sector study, the data are clearly useable. Peptides exceeding mass 2000 are still amenable to linked scan analysis as despite the reduction in fragment ions observed, as has been demonstrated, those present are useful in a confirmatory capacity in cases where the sequence has been derived by other sequencing techniques.

The most significant aspect of this study has been the analysis of components of an unfractionated digest mixture by linked scan/CAD.

Linked scanning of a tryptic digest mixture of the enzyme pyruvate decarboxylase elucidated two unknown components which determined an amino acid difference to that in the gene-derived sequence at residue 335, together with a previously unreported processing event and post-translational modification of the N-terminus of the protein. Data such as these support the need for complementary protein sequence analysis when the amino acid sequence has been derived solely from nucleotide analysis, and to examine the fidelity of a recombinant protein produced in an abnormal cell line. Ordinarily to detect these previously unrecognised errors of the published gene sequence, purification procedures such as HPLC would be required with subsequent isolation and analysis involving complex microchemical manipulations.

A possible limitation of the linked scan method lies with the width of the mass window examined, some 5 a.m.u. If two signals of interest were to fall within the same mass window it would be impossible to resolve the two. However, when the protein sequence is not available it is not necessary for the entire protein sequence to be analysed by classical protein chemistry procedures or by non-classical methods such as linked scanning. Rather, with the advances in molecular biology emphasis lies on the provision of several short stretches of sequence for the production of complementary synthetic oligonucleotide probes for screening a gene library, so that the complete amino acid sequence of the protein is not required.

Linked scanning/CAD as applied in this chapter has been shown to be

a highly sensitive and rapid method for protein sequence analysis. This data represents a significant contribution to modern protein structure analysis methods and provides an important extension of two sector instrumentation for biopolymer sequencing studies.

CHAPTER FIVE

THE PRIMARY SEQUENCE INVESTIGATION OF PYRUVATE DECARBOXYLASE APPLYING A COMBINATION OF MASS SPECTROMETRIC TECHNIQUES

5.1 Introduction

Glycolysis is a central pathway, in most living organisms, for the recovery of chemical energy as adenosine 5'-triphosphate, ATP. The glycolytic pathway differs between species only in how the rate of glycolysis is regulated and in the subsequent metabolic fate of the pyruvate formed from glucose. Some animal tissues e.g. skeletal muscle, when subjected to vigorous exercise cannot oxidise pyruvate due to a lack of molecular oxygen, instead the pyruvate is converted by the enzyme lactate dehydrogenase to lactate. This process, anaerobic glycolysis, is an important source of ATP energy in muscles undergoing very intense physical activity. In organisms such as yeast, which contain no lactate dehydrogenase, pyruvate is metabolised by two alternate enzymatic reactions; this process is termed alcoholic fermentation. Alcoholic fermentation is identical to anaerobic glycolysis except for the final stages, which produce different products, see Figure 5.0. The first step of alcoholic fermentation involves decarboxylation of pyruvate to acetaldehyde and carbon dioxide in a reaction catalysed by the enzyme pyruvate decarboxylase, PDC. The acetaldehyde is then reduced to ethanol by alcohol dehydrogenase, which utilises nicotinamide adenine

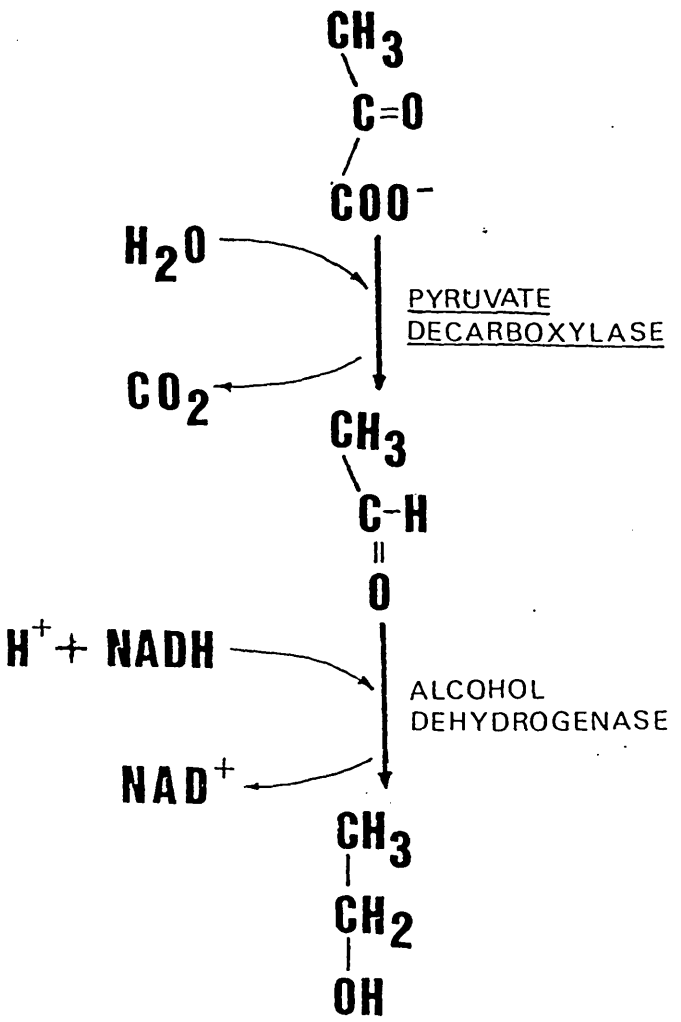


Figure 5.0 Terminal Steps in Alcoholic Fermentation

dinucleotide, NADH as a coenzyme.

The overall equation of alcoholic fermentation is;

$$GLUCOSE + 2P_{inorganic} + 2ADP \longrightarrow 2ETHANOL + CO_2 + 2ATP + 2H_2O$$

The loss of the carboxyl group from pyruvate by the action pyruvate decarboxylase, PDC, is a simple decarboxylation and does not involve the net oxidation of the pyruvate. This reaction is irreversible in the cell with PDC requiring Mg^{2+} and a tightly bound coenzyme, thiamine pyrophosphate for activity. The coenzyme, thiamine pyrophosphate, serves as a transient carrier of the acetaldehyde group which is covalently attached to the thiazole ring, i.e. the carboxyl group of pyruvate is lost as carbon dioxide and the rest of the pyruvate molecule, often referred to as the "active aldehyde", is simultaneously transferred to the "2" position of the thiazole ring of the pyrophosphate to yield its hydroxyethyl derivative. This intermediate exists only transiently since the hydroxy-ethyl group is quickly cleaved from the coenzyme to yield free acetaldehyde.

Pyruvate decarboxylase is a key enzyme in the fermentation of glucose to ethanol in *Saccharomyces cerevisiae*. In the presence of high glucose concentrations, the cells contain high levels of PDC and produce ethanol at a high rate. Conversely, at very low glucose concentrations the levels of PDC are decreased by a factor of 10-20 and mainly oxidative metabolism takes place [100].

The molecular weight of the active PDC molecule has been estimated, by SDS-gel electrophoresis, as 230 kDa, [49,50]. It has been demonstrated that under the same conditions for removal of the cofactors Mg^{2+} and thiamine pyrophosphate, the active oligomer of PDC reversibly dissociates into subunits of one half the molecular weight [49], with thiamine pyrophosphate being an obligatory co-factor for reconstitution of the enzymically active holoenzyme. In addition it was determined that these subunits were infact dimers and thus, it was proposed that pyruvate decarboxylase is a tetramer comprised of four identical or nearly identical subunits. Gounaris *et al* in 1975 presented data which indicated that the molecular weight of a monomer unit is 60,000 ±3,000 Da. Furthermore it was found that the dissociation of the holoenzyme into dimers, and the dimers into monomers, was brought about under different conditions and thus it was proposed that interchain interactions that form the dimers are not the same as those leading to formation of the tetramer. The two possibilities offered by Gounaris and co-workers with regards to the structural nature of the dimers were; (a) that the dimers are composed of two different chains of similar molecules, as for example in the case of haemoglobin, and (b) that the monomer units are identical and in the absence of a co-factor form the dimer.

High resolution SDS-polyacrylamide gel electrophoresis of PDC from Brewers yeast [101,102] and wheat germ [103], by Ullrich *et al*, 1982, [104], separated two types of polypeptide chains differing slightly in size, $M_r(\alpha)$ = 63,000 - 65,000 Da; $M_r(\beta)$ = 61,000 - 62,000 Da. The isolation of two polypeptide chains of differing

molecular weight led Ullrich to suggest that beta chains may be produced from alpha chains by limited proteolysis, either physiologically or as a preparation artifact.

Recently the PDC1 gene of *Saccharomyces cerevisiae* encoding pyruvate decarboxylase, E.C.4.1.1.1., was sequenced by Kellermann and co-workers [99]. From the DNA sequence a polypeptide of 549 amino acids with a molecular weight of 60,008 Daltons was determined. This molecular weight was in agreement with that previously reported, for a monomer unit of PDC. A comparison of the amino acid compositions of both the alpha and beta chains with that of the DNA sequence led Kellermann *et al* to conclude that it was highly unlikely that proteolytic cleavage was responsible for the conversion of the alpha polypeptide chain to the beta. However, there was not sufficient evidence to conclude whether the PDC1 gene encoded the hypothesised alpha or beta chain. Since no information regarding the pyruvate decarboxylase primary sequence was available from other technologies it was also not possible for Kellermann and co-workers to definitively identify the sequenced DNA fragment as the structural gene.

The objectives of the study presented here were;

(a) the primary sequence analysis of three separable monomer units of PDC, termed A1, A2 and B, isolated from wild type yeast, *Saccharomyces cerevisiae* MC16,

(b) the confirmation of the published gene sequence, PDC1, as the structural gene, and

(c) the investigation of sequence differences, if any, between the

isolated subunits, A1, A2 and B of pyruvate decarboxylase.

An appropriate strategy or combination of strategies was selected to tackle the objectives listed above with respect to primary sequence analysis, and to relate the results of this thesis study to secondary, tertiary and quaternary interactions. The FAB-Mapping strategy [41] was selected as a method highly suited for the primary sequence analysis of a large protein such as pyruvate decarboxylase. The principles of FAB-Mapping have been described in Chapter Two of this thesis.

A second technique applied to the sequence analysis of PDC was a method of producing sequence fragment ions by mass spectrometry, linked scanning. Chapter Four describes in detail the novel application of linked scanning/CAD in an approach to direct mixture analysis of complex proteolytic digests. Linked scanning/CAD as applied in this thesis work has contributed valuable sequence information to the primary structure analysis of pyruvate decarboxylase. The data from these linked scan studies allowed identification of a post-translational modification and a processing event of the N-terminus of the protein and an amino acid difference to that in the published gene-derived sequence.

Molecular weight determinations by techniques such as SDS-polyacrylamide gel electrophoresis provide only an estimation of molecular weights. For example the molecular weight determined for a PDC subunit, by this method, was within ±3000 Da at 60,000 Da. The Electrospray Ionisation Mass Spectrometry, ESI-MS,

technique originally developed by Dole *et al* [54], has been demonstrated by this thesis work and others [55,56] to provide accurate molecular weight determinations for high mass biopolymers. ESI-MS produces gas-phase multiply charged molecular ions of molecules from highly charged liquid droplets by an electric field. The highly charged droplets are formed by electrically induced nebulisation in a dry nitrogen atmosphere. The hot nitrogen causes the droplets to shrink to their charge limit resulting in ion desorption or evaporation of intact molecular ions. Molecular weights greater than 130 kDa have been reported [105] to be determined by an ESI quadrupole mass spectrometer. A simplified schematic representation of an ESI mass spectrometer is depicted in Figure 1.4, Chapter One. In this thesis study a molecular weight determination for each of the isolated subunits A1, A2 and B was achieved by ESI-MS. Furthermore, these molecular weights have been determined within 5-20 daltons (±0.02%). The molecular weight data, determined from ESI-MS, has contributed valuable information for elucidating the existence of one or more monomer units and the extent of sequence differences between these monomers.

The preparations of subunits A1, A2 and B studied in this thesis work were supplied by our colleagues in the Department for Biotechnology.

5.2 Results and Discussion

The PDC samples, A1, A2 and B investigated in this thesis work were

isolated from wild type yeast, *Saccharomyces cerevisiae*, MC16. Purification of the crude extract by ion-exchange chromatography (DE-52 ion-exchange column) yielded two active peaks termed A and B, with B eluting from the column later than A, supporting earlier literature reports pertaining to the existence of two monomer units. However, a second purification step by ion-exchange (DEAE-Sephacel column) of both A and B showed peak A to subdivide to a second peak, A2, which also exhibited enzymic activity. Gel-filtration (AcA-22 gel-column) of A1, A2 and B, showed A1 and A2 to elute from the column in approximately the same position, with B slightly retarded. Thus, three separable monomer units were isolated from this preparation; the existence of a third subunit had not previously been reported.

This investigation into (a) the primary sequence analysis of each PDC monomer unit and hence, the fidelity of the gene sequence, (b) the possible existence of a third subunit of PDC and, (c) the elucidation of any sequence differences between the three subunits, can be divided into four sections:

5.2.1 FAB-Mapping.

5.2.2 B/E Link Scanning/CAD.

5.2.3 High-Mass MS studies.

5.2.4 Electrospray Ionisation MS.

5.2.1 **FAB-Mapping**

5.2.1.1 Trypsin Digestion

Utilising the FAB-Mapping strategy a sample of each subunit, A1, A2 and B, was enzymically digested by trypsin, with analysis of the resultant digest mixtures by FAB-MS. The spectra generated, FAB-Maps, are presented in Figures 5.1, 5.1.1 and 5.1.2 respectively. It should be noted that the spectra are recorded by computer and hence, signals are assigned accurate masses, as discussed in Chapter Four. A conversion factor of approximately half a mass unit per 1000 a.m.u. can be subtracted from an accurate mass to convert it to nominal. This difference between accurate and nominal mass arises from certain elements being mass sufficient, e.g. hydrogen, whilst others are mass deficient e.g. oxygen. The FAB-Maps in Figures 5.1, 5.1.1 and 5.1.2, are complex each characterised by a large number of intense quasi-molecular ions. These mass signals correspond to the tryptic peptide fragments formed from trypsin digestion of the protein subunits. By employing the computer search programme (M-Scan Ltd., Ascot, and R. McDowell, Department of Biochemistry, Imperial College, London, SW7 2AZ) used in Chapter Two, each of the quasi-molecular ions observed in the FAB-Maps was tested for a mass fit on to a stretch of amino acid residues in the gene-derived sequence (Figure 5.1.3). For each mass value the computer generates a number of sequences corresponding to a particular molecular ion. Knowing the specificity of the enzyme used to generate the digest mixture most of the possible peptide sequences corresponding to a particular mass value can be readily discounted as not agreeing with the enzyme specificity requirements. The type of digest employed here was trypsin, specific for cleaving on the C-terminal side of lysine and arginine residues. Thus, only one of the peptide sequences, given by the

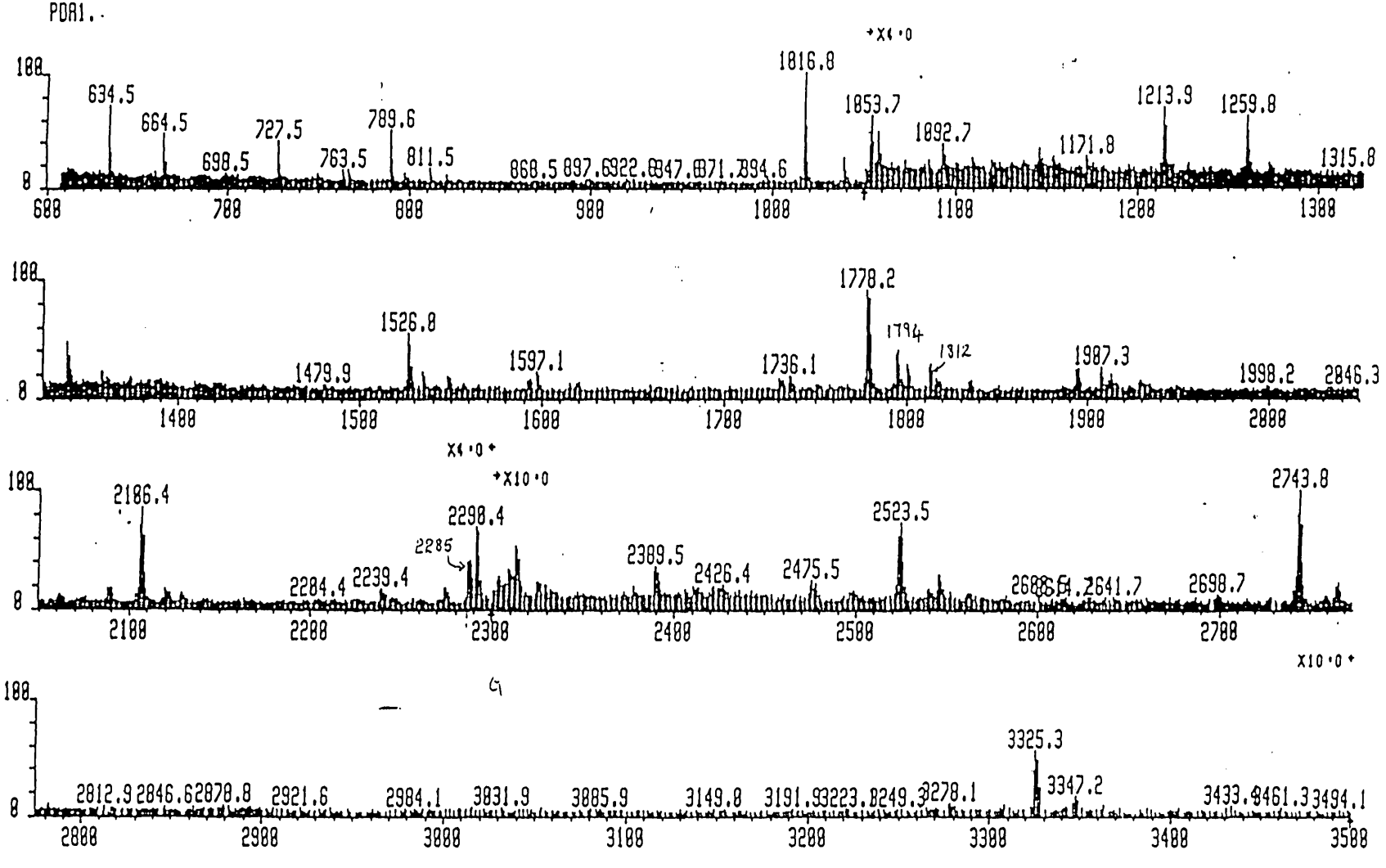


Figure 5.1 Tryptic FAB-Map of PDC subunit A1

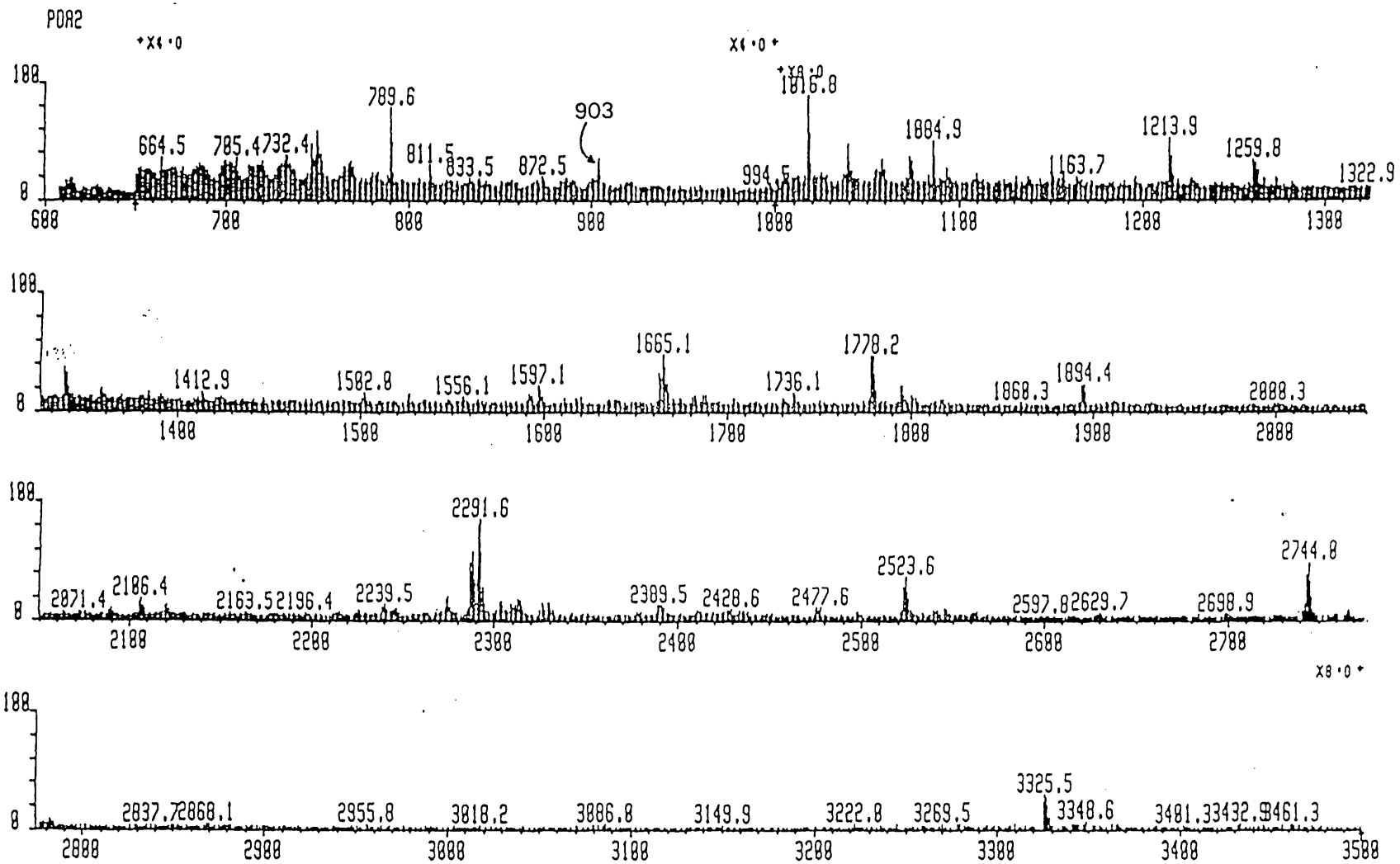


Figure 5.1.1 Tryptic FAB-Map of PDC subunit A2.

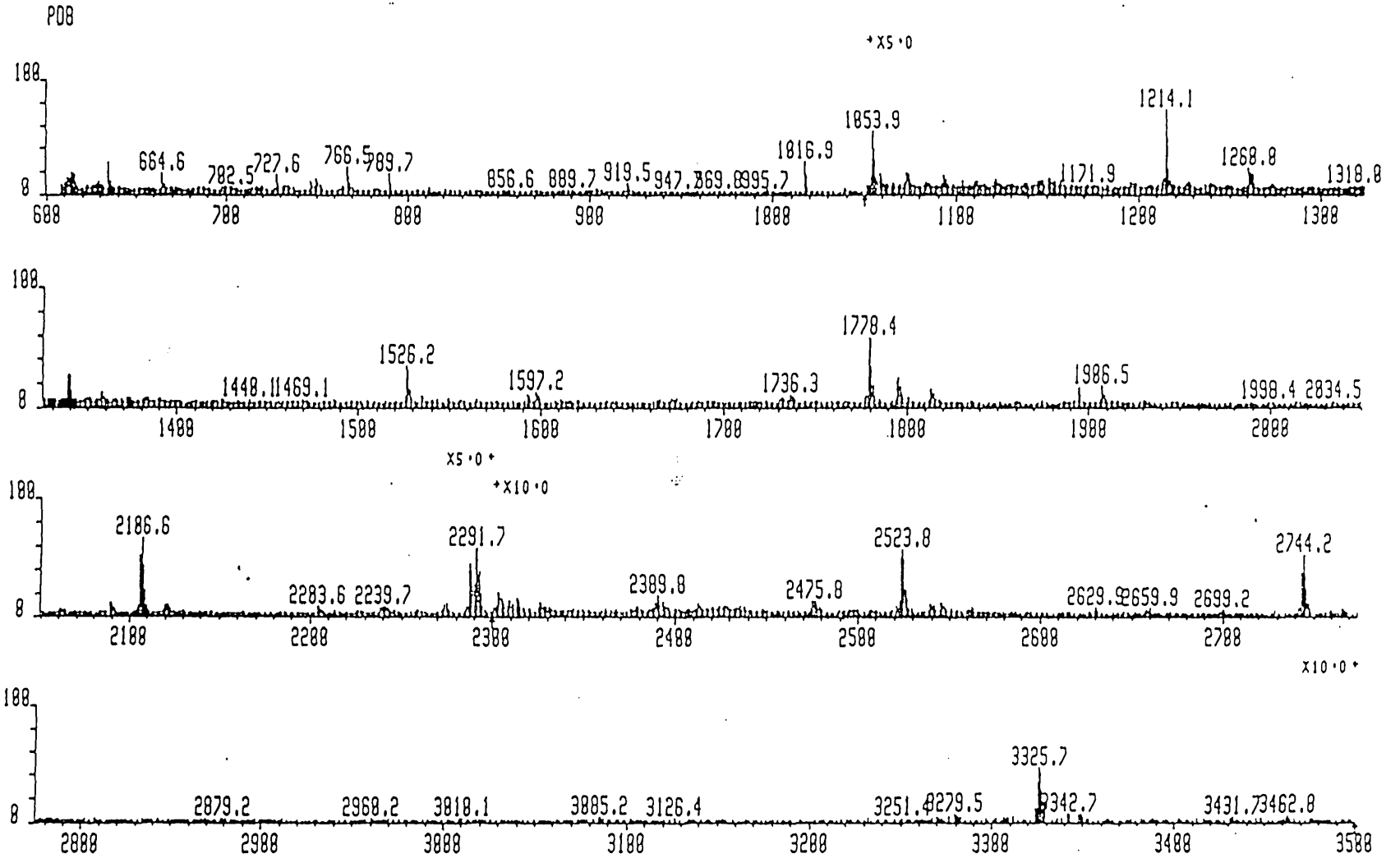


Figure 5.1.2 Tryptic FAB-Map of PDC subunit B.

-467 GGCTGGTGG AAAAAATGAA TAATTTATGA ATTTGAGAAC AATTTTGTGT TGTTACGGTA TTTTACTATG GAATAATCAA TCAATTGAGG ATTTTATCGC

-368 AAATATCGTT TGAATATTTT CCGACCCCTT GAGTACTTTT CTTCATAATT GCATAATATT GTCCGCTGCC CGTTTTTCTG TTAGACCGGTG TCTTGATCTA

-268 CTTGCTATCG TTCAACACCA CCTTATTTTC TAACTATTTT TTTTTTAGCT CATTTGAATC AGCTTATGGT GATGGCACAT TTTTGCATAA ACCTAGCTGT

-168 CCTCGTTGAA CATAGGAAAA AAAAATATAT AAACAAGGCT CTTTCACTCT CCTTGGAATC AGATTTGGGT TTGTTCCCTT TATTTTCATA TTTCTTGTCA

- 68 TATTCTTTTC TCAATTATTA TCTTCTACTC ATAACCTCAC GCAAAATAAC ACAGTCAAAT CAATCAAA ATG TCT GAA ATT ACT TTG GGT AAA TAT
Met Ser Glu Ile Thr Leu Gly Lys Tyr

28 TTG TTC GAA AGA TTA AAG CAA GTC AAC GTT AAC ACC GTT TTC GGT TTG CCA GGT GAC TTC AAC TTG TCC TTG TTG GAC AAG ATC
Leu Phe Glu Arg Leu Lys Gln Val Asn Val Asn Thr Val Phe Gly Leu Pro Gly Asp Phe Asn Leu Ser Leu Leu Asp Lys Ile

112 TAC GAA GTT GAA GGT ATG AGA TGG GCT GGT AAC GCC AAC GAA TTG AAC GCT GCT TAC GCC GCT GAT GGT TAC GCT CGT ATC AAG
Tyr Glu Val Glu Gly Met Arg Trp Ala Gly Asn Ala Asn Glu Leu Asn Ala Ala Tyr Ala Ala Asp Gly Tyr Ala Arg Ile Lys

196 GGT ATG TCT TGT ATC ATC ACC ACC TTC GGT GTC GGT GAA TTG TCT GCT TTG AAC GGT ATT GCC GGT TCT TAC GCT GAA CAC GTC
Gly Met Ser Cys Ile Ile Thr Thr Phe Gly Val Gly Glu Leu Ser Ala Leu Asn Gly Ile Ala Gly Ser Tyr Ala Glu His Val

280 GGT GTT TTG CAC GTT GTT GGT GTC CCA TCC ATC TCT TCT CAA GCT AAG CAA TTG TTG TTG CAC ACC TTG GGT AAC GGT GAC TTC
Gly Val Leu His Val Val Gly Val Pro Ser Ile Ser Ser Gln Ala Lys Gln Leu Leu Leu His Thr Leu Gly Asn Gly Asp Phe

364 ACT GTT TTC CAC AGA ATG TCT GCC AAC ATT TCT GAA ACC ACT GCT ATG ATC ACT GAC ATC TGT ACG CCC CAG GCT GAA ATT GAC
Thr Val Phe His Arg Met Ser Ala Asn Ile Ser Glu Thr Thr Ala Met Ile Thr Asp Ile Cys Thr Pro Gln Ala Glu Ile Asp

448 AGA TGT ATC AGA ACC ACT TAC GTC ACC CAA AGA CCA GTC TAC TTA GGT TTG CCA GCT AAC TTG GTC GAC TTG AAC GTC CCA GCT
Arg Cys Ile Arg Thr Thr Tyr Val Thr Gln Arg Pro Val Tyr Leu Gly Leu Pro Ala Asn Leu Val Asp Leu Asn Val Pro Ala

532 AAG TTG TTG CAA ACT CCA ATT GAC ATG TCT TTG AAG CCA AAC GAT GCT GAA TTC GAA AAG GAA GTC ATT GAC ACC ATC TTG GTC
Lys Leu Leu Gln Thr Pro Ile Asp Met Ser Leu Lys Pro Asn Asp Ala Glu Ser Glu Lys Glu Val Ile Asp Thr Ile Leu Val

616 TTG GTC AAG GAT GCT AAG AAC CCA GTT ATT CTG GCT GAT GCT TGT TGT TCC AGA CAC GAC GTC AAG GCT GAA ACT AAG AAG TTG
Leu Val Lys Asp Ala Lys Asn Pro Val Ile Leu Ala Asp Ala Cys Cys Ser Arg His Asp Val Lys Ala Glu Thr Lys Lys Leu

700 ATT GAC TTG ACT CAA TTC CCA GCT TTC GTC ACC CCA ATG GGT AAG GGT TCC ATT AGC GAA CAA CAC CCA AGA TAC GGT GGT GTT
Ile Asp Leu Thr Gln Phe Pro Ala Phe Val Thr Pro Met Gly Lys Gly Ser Ile Ser Glu Gln His Pro Arg Tyr Gly Gly Val

784 TAC GTC GGT ACC TTG TCC AAG CCA GAA GTT AAG GAA GCC GTT GAA TCT GCT GAC TTG ATT TTG TCT GTC GGT GCT TTG TTG TCT
Tyr Val Gly Thr Leu Ser Lys Pro Glu Val Lys Glu Ala Val Glu Ser Ala Asp Leu Ile Leu Ser Val Gly Ala Leu Leu Ser

868 GAT TTC AAC ACC GGT TCT TTC TCT TAC TCT TAC AAG ACC AAG AAC ATT GTC GAA TTC CAC TCC GAC CAC ATG AAG ATC AGA AAC
Asp Phe Asn Thr Gly Ser Phe Ser Tyr Ser Tyr Lys Thr Lys Asn Ile Val Glu Phe His Ser Asp His Met Lys Ile Arg Asn

952 GCC ACT TTC CCA GGT GTC CAA ATG AAA TTC GTT TTG CAA AAG TTG TTG ACC AAT ATT GCT GAC GCC GCT AAG GGT TAC AAG CCA
Ala Thr Phe Pro Gly Val Gln Met Lys Phe Val Leu Gln Lys Leu Leu Thr Asn Ile Ala Asp Ala Ala Lys Gly Tyr Lys Pro

1036 GTT GCT GTC CCA GCT AGA ACT CCA GCT AAC GCT GCT GTC CCA GCT TCT ACC CCA TTG AAG CAA GAA TGG ATG TGG AAC CAA TTG
Val Ala Val Pro Ala Arg Thr Pro Ala Asn Ala Ala Val Pro Ala Ser Thr Pro Leu Lys Gln Glu Trp Met Trp Asn Gln Leu

1120 GGT AAC TTC TTG CAA GAA GGT GAT GTT GTC ATT GCT GAA ACC GGT ACC TCC GCT TTC GGT ATC AAC CAA ACC ACT TTC CCA AAC
Gly Asn Phe Leu Gln Glu Gly Asp Val Val Ile Ala Glu Thr Gly Thr Ser Ala Phe Gly Ile Asn Gln Thr Thr Phe Pro Asn

1204 AAC ACC TAC GGT ATC TCT CAA GTC TTA TGG GGT TCC ATT GGT TTC ACC ACT GGT GCT ACC TTG GGT GCT GCT TTC GCT GCT GAA
Asn Thr Tyr Gly Ile Ser Gln Val Leu Trp Gly Ser Ile Gly Phe Thr Thr Gly Ala Thr Leu Gly Ala Ala Phe Ala Ala Glu

1288 GAA ATT GAT CCA AAG AAG AGA GTT ATC TTA TTC ATT GGT GAC GGT TCT TTG CAA TTG ACT GTT CAA GAA ATC TCT ACC ATG ATC
Glu Ile Asp Pro Lys Lys Arg Val Ile Leu Phe Ile Gly Asp Gly Ser Leu Gln Leu Thr Val Gln Glu Ile Ser Thr Met Ile

1372 AGA TGG GGC TTG AAG CCA TAC TTG TTC GTC TTG AAC AAC GAT GGT TAC ACC ATT GAA AAG TTG ATT CAC GGT CCA AAG GCT CAA
Arg Trp Gly Leu Lys Pro Tyr Leu Phe Val Leu Asn Asn Asp Gly Tyr Thr Ile Glu Lys Leu Ile His Gly Pro Lys Ala Gln

1456 TAC AAC GAA ATT CAA GGT TGG GAC CAC CTA TCC TTG TTG CCA ACT TTC GGT GCT AAG GAC TAC GAA ACC CAC AGA GTC GCT ACC
Tyr Asn Glu Ile Gln Gly Trp Asp His Leu Ser Leu Leu Pro Thr Phe Gly Ala Lys Asp Tyr Glu Thr His Arg Val Ala Thr

1540 ACC GGT GAA TGG GAC AAG TTG ACC CAA GAC AAG TCT TTC AAC GAC AAC TCT AAG ATC AGA ATG ATT GAG GTT ATG TTG CCA GTC
Thr Gly Glu Trp Asp Lys Leu Thr Gln Asp Lys Ser Phe Asn Asp Asn Ser Lys Ile Arg Met Ile Glu Val Met Leu Pro Val

1624 TTC GAT TGC TCC ACA AAA CTT GGT TGA ACAAGCTAAG TTGACTGCTG CTACCAACGC TAAGCAATAA GCGATTTAAT CTCTAATTAT AGTTAAAGTT
Phe Asp Cys Ser Thr Lys Leu Gly -o-

1721 TTATAAGCAT TTTTATGTAA CGAAAAATAA ATTGGTTCAT ATTATTACTG CACTGTCACT TACCATGGAA AGACCAGACA AGAAGTTGCC GACACGACAG

1821 TCTGTTGAAT TGGCTTAAGT CTGGGTCCCC TTCTTTACAA ATTTGAAGAA TTTCTCTTAA ACGATATGTA TATTCTTTCG TTGGAAAACA TGTCTTCCAA

1921 AAAAAAAAAC CGACGAATTA GTGGAACCAA GGAAAAAAAA AGAGGTATCC TTGATTAAGG AACACTGTTT AAACAGTGTG GTTTCCAAAA CCCTAGAACT

2021 GCATTAGCAT GATAAGAGAC TAGACACCTC GATACAAATA ATGGTTACTC AA

Figure 5.1.3 Nucleotide and Predicted Amino Acid Sequence of Pyruvate Decarboxylase (99).

computer for a mass fit for a quasi-molecular ion, will normally meet the digest conditions terminating with a lysine or arginine residue as well as the residue preceding the N-terminus of the peptide sequence corresponding to a tryptic cleavage site hence, all other sequence possibilities, except that one, may be discounted. The intense mass signals observed in each FAB-Map of the tryptic digest mixtures, A1, A2 and B, are listed in Table 5.1 together with the sequence assignments on to the gene-derived sequence. It is apparent that there are a number of signals in each of the FAB-Maps which cannot be assigned on to the gene-derived sequence namely, m/z 789, 1016, 1213, 1664, 2104, 2285, 2521 and 3322. The data from this digest tentatively suggests the possible existence of either post-translational modifications, sequence mutations or sequence deletions, in the gene sequence. It is interesting to note that initially the peptide at m/z 1259 could not be mass matched on to the gene-derived sequence; however, it was later determined to correspond to the peptide, 212 to 223. This assignment was initially not made due the presence of an intra-molecular disulphide bridge, which was expected to have been broken, formed between two adjacent cysteine residues, 220 and 221. Sterically, this is thought not to be possible, being energetically unfavourable.

A comparison of the three tryptic FAB-Maps, Figures 5.1, 5.1.1 and 5.1.2, shows that the signal m/z 1664 (for which no peptide assignment can be made) is unique to the tryptic FAB-Map of subunit A2. Summarised in Table 5.1.1 are those signals observed in the tryptic FAB-Maps (Figures 5.1, 5.1.1 and 5.1.2) which were not able

PDCA1		PDCA2		PDCB	
$[M+H]^+$	Residue	$[M+H]^+$	Residue	$[M+H]^+$	Residue
634	327-331	664	478-483	634	327-331
664	478-483	789		664	478-483
727	9-13	903		727	9-13
789		1016	525-531	789	
811	525-531	1038		811	525-531
1016		1057	342-351	1016	
1038		1213		1053	
1053		1259	213-224	1213	
1057	342-351	1337	352-365	1259	213-224
1092	317-326	1356	304-314	1337	352-365
1213		1596	258-272	1356	304-314
1259	213-224	1664		1525	465-477
1337	352-365	1777	233-248	1533	164-178
1356	304-314	1893		1596	258-272
1525	465-477	2104		1777	233-248
1533	164-178	2285		1794	
1596	258-272	2289	16-36	1811	
1777	233-248	2521		1893	
1794		2741	154-178	1905	232-248
1811		3322		2104	
1893				2285	
1905	232-248			2389	16-36
2104				2387	484-505
2285				2521	
2289	16-36			2741	154-178
2387	484-505			3322	
2521					
2741	154-178				
3322					

Table 5.1 Quasi-molecular Ions, $[M+H]^+$, Observed in the Tryptic FAB-Maps of Subunits A1, A2 and B, together with the Peptide Assignments on to the Gene-derived Sequence of PDC [99].

Signals not able to be Assigned	Observed in tryptic FAB-Maps		
	A1	A2	B
m/z			
789	+	+	+
903	-	+	-
1016	+	+	+
1213	+	+	+
1664	-	+	-
1794	+	+	+
1893	+	+	+
2104	+	+	+
2285	+	+	+
2521	+	+	+
3322	+	+	+

Table 5.1.1 Those Signals in the Tryptic FAB-Map Spectra which were unable to be Assigned on to the Gene-derived Sequence.

Key: (-) signal not observed in the FAB-Map of that particualr subunit and,

(+) signal observed in FAB-Map.

to be mass matched on to a stretch of amino acid residues in the gene-derived sequence, together with those signals which are unique to one or two tryptic FAB mass spectra.

The next step in the FAB-Mapping strategy is to confirm the peptide assignments of the various mass values listed in Table 5.1, and to obtain if possible information regarding the unassigned mass values from each spectrum. An aliquot of each of the tryptic digest mixtures of monomer units A1, A2 and B was manipulated by predictable chemical and enzymic digestions, with the products after each step examined by FAB-MS. An aliquot of each digest mixture was submitted for two cycles of manual Edman degradation, with FAB-MS analysis after each step. The mass difference between a signal mass value prior to Edman degradation and that after (determined by comparison of the two FAB-Maps) corresponds to the N-terminal amino acid residue cleaved by the Edman degradation, as discussed in Chapter Two. Two cycles of Edman degradation/MS of the intact digest mixture allowed identification of the first two N-terminal amino acid residues of most of the tryptic peptides observed in the initial FAB-Maps to be determined, which confirmed previous sequence assignments. The peptide at m/z 1525 was initially not able to be mapped on to the gene-derived sequence as a peptide arising from tryptic cleavage however, two steps of subtractive Edman degradation proved leucine and phenylalanine as the N-terminal first and second residues of this peptide. This data enabled the peptide at m/z 1525 to be assigned on to the gene-derived sequence, on the basis of its mass and the N-terminal amino acids elucidated by Edman/MS, between residues 465 to 477.

However, this assignment does not correlate for an N-terminal trypsin cleavage site, i.e. the residue in the protein sequence preceding the N-terminus of this peptide is not lysine or arginine, but tyrosine. Hence, this peptide though terminating with a lysine residue is the product of an N-terminal chymotryptic and C-terminal tryptic cleavage. The detection of chymotryptic cleavage sites was most unexpected. However, the presence of chymotrypsin is explained as impurity in the trypsin. Similarly, the weak signal m/z 1533, with only one step of Edman degradation interpretable, mapped on to the gene-derived sequence between positions 164 to 178. For a mass match this peptide was found to have been formed by a chymotryptic cleavage between residue 163 (tyrosine) and 164 (leucine).

The presence of chymotrypsin cleavages substantiates the need not only for N-terminal analysis (Edman degradation) but also for C-terminal amino acid determination. An enzyme specific for C-terminal lysine and arginine residue cleavage only is carboxypeptidase B, CpB. This enzyme is highly appropriate for use with tryptically derived peptides, since one expects a tryptic digest to be composed of peptides terminating with either lysine or arginine residues. An aliquot of each of the tryptic digest mixtures A1, A2 and B was sub-digested with carboxypeptidase B. The double digest mixtures, tryptic/CpB, were FAB-MS analysed. A comparison of the tryptic FAB-Maps and the tryptic/CpB FAB-Maps showed signals to have shifted by predictable mass increments of either 128 or 156 a.m.u, corresponding to C-terminal cleavage, by CpB, of lysine and arginine residues respectively. Analysis of each of the mass shifts confirmed the previous mass assignments on to

the gene-derived sequence. However, it is interesting to note that the signals at m/z 664, 1525 and 2289 did not change in mass value after digestion with carboxypeptidase B, even though all three assigned peptide sequences terminate with a C-terminal lysine residue. This observation was rationalised in each case by consideration of the amino acid residue preceding each lysine. The penultimate amino acid residue in the peptide sequence $[M+H]^+$ 664 is proline; it has been well established that proline peptide bonds are hydrolysed very slowly. The residues preceding lysine in both peptides at m/z 1525 and 2289 are acidic, glutamic and aspartic acid respectively. The peptide bond between an acidic residue and lysine/arginine is hydrolysed at a slower rate due the repulsion of the active site of the enzyme, CpB, by the acidic residue. However, with the exceptions of the peptides discussed, it was anticipated that all other peptides in the tryptic digest mixture would be C-terminally digested, with the exception of the natural C-terminus of the protein, unless the C-terminal residue of the intact protein was in fact lysine or arginine. Analysis of the FAB-Map of the tryptic/CpB digest allowed the elucidation of the C-terminal residue of both peptides at m/z 2285 and 3322, as lysine. Neither of these signals could be mass assigned on to the gene-derived sequence even though the CpB data proved that these peptides were tryptically derived. This finding substantiated the earlier supposition of the existence of possible sequence errors in the gene-derived sequence. Both these relatively large peptides were found to be common to all three subunits. From this prelimary study it was shown that many of the unassigned intense signals were common to A1, A2 and B. The extent of sequence differences between

the isolated subunits are investiagted and discussed later in this research study.

5.2.1.2 Trypsin/V8 protease Digestion

An extension of this investigation of the primary sequence analysis of pyruvate decarboxylase involved sub-digestion of an aliquot of each tryptic digest mixture with V8 protease (an extracellular enzyme isolated from *Staphylococcus aureus*). This enzyme is specific for cleaving on the C-terminal side of glutamyl and aspartyl peptide bonds only. The experimental conditions employed here were suitable for glutamyl bond cleavage. Glutamic acid residues are frequent in the gene-derived sequence, hence, it was expected that most of the tryptic peptides would be digested into two or more peptides and with those tryptic peptides not containing glutamic acid not expected to change in mass. An aliquot of each of the tryptic/V8 digest mixtures was analysed by FAB-MS. The additional sequence information interpreted from the tryptic/V8 FAB-Maps is presented in Table 5.1.2. The summary of this tryptic/V8 FAB-Mapping data, in Table 5.1.2, indicates that little additional sequence information has been gleaned. This is rationalised by the phenomenon of suppression [27] on FAB-MS analysis of a complex digest mixture. It has been recognised, since the introduction of FAB-MS, that certain components in a digest mixture suppress others and hence, are not detected in the FAB mass spectrum. Therefore, the sub-digestion of the tryptic digest is thought to have generated too many peptides fragments, with some possessing the physico-chemical properties which suppress the other components. It was therefore not possible to draw any conclusions

Subunit A1		Subunit A2		Subunit B	
$[M+H]^+$	Residues	$[M+H]^+$	Residues	$[M+H]^+$	Residues
403	148-150	474	304-307m	474	304-307m
474	304-307m	488	232-235m	720	453-458
488	232-235m	531	4-8	1369	258-270
531	4-8	1584	179-192		
545	190-194				
720	453-458				

Table 5.1.2 Additional Interpretable Intense Signals Observed in the Tryptic/V8 FAB-Maps of Subunits A1, A2 and B.
Key: m=residues partially mapped from the previous tryptic FAB-Map data.

$[M+H]^+$	Residue Assignment
474	75-99
978	301-308
1010	334-343
1096	207-216
1233	243-254
1361	313-324
1501	267-280
2027	344-364

Table 5.1.3 Intense Signals Observed in the Chymotryptic FAB-Map of Subunit A1, able to be Assigned on to stretches of Amino Acid Residues in the Gene-derived Sequence.

concerning the absence of any particular peptide previously observed in the tryptic FAB-Maps, as it may be absent due to digestion by V8 protease or simply that in the presence of an increased number of peptides it was suppressed on FAB-MS analysis.

5.2.1.3 Trypsin/CNBr Digestion

From the previous double digest experiment the lack of sequence information interpreted from the FAB-Maps was rationalised by suppression effects on FAB-MS analysis of the digest mixture, as discussed. Hence, the next manipulative digest was carefully selected for high specificty and with the amino acid residue type preferentially cleaved appearing infrequently in the protein sequence. In the gene-derived sequence there are only twelve methionine residues shown to be present thus, cleavage at these residues would not generate an abundance of peptide fragments in addition to the already present tryptic peptides. An aliquot of each tryptic digest mixture was chemically digested with cyanogen bromide, CNBr, specific for cleavage on the C-terminal side of methionine residues. The resultant tryptic/CNBr digest mixtures were each analysed by FAB-MS. Interpretation of the FAB mass spectra allowed confirmation of numerous peptide sequences previously mapped from the tryptic digest data alone. However, in addition a peptide $[M+H]^+$ 882 was mapped on to the gene-derived sequence between positions 179 to 186, a region previously unassigned. The data from the analysis of this double digest, assigned five of the twelve methionine residues in the gene-derived sequence. The presence of signals m/z; 1213, 1794 and 1811 in the tryptic/CNBr FAB-Maps indicated that the peptide sequences

corresponding to these masses do not contain methionine residues, as these signals were observed in the original tryptic FAB-Map spectra. However, as noted previously due to suppression of some components of a mixture by others on FAB-MS analysis the absence of signals previously observed in the tryptic FAB-Map and now absent in the tryptic/CNBr FAB-Map cannot be solely interpreted as methionine containing peptides which have been digested.

The peptides masses not able to be assigned on to the gene-derived sequence were greater in number than those mass signals unique to one or two subunits. Therefore, a priority was placed on determining the extent of sequence differences between the published gene-derived sequence and that elucidated practically in this FAB-Mapping study. Reported later in this chapter is research work enabling characterisation of the sequence differences between the various subunits, see section 5.2.4.

5.2.1.4 Chymotrypsin Digestion of Intact PDC Subunit A1

A sample of intact protein, A1, was enzymically digested with chymotrypsin. This enzyme is specific for cleaving on the C-terminal side of tryptophan, tyrosine, phenylalanine and leucine peptide bonds. Other residues are also hydrolysed but at a slower rate. An aliquot of this chymotryptic digest mixture was FAB-MS analysed. The chymotryptic FAB-Map is presented in Figure 5.1.4. By employing the computer programme, as previously described, the most intense signals from the chymotryptic FAB-Map were able to be assigned to peptides in the protein sequence. Seven intense signals were able to be confidently mass matched on to the gene-derived

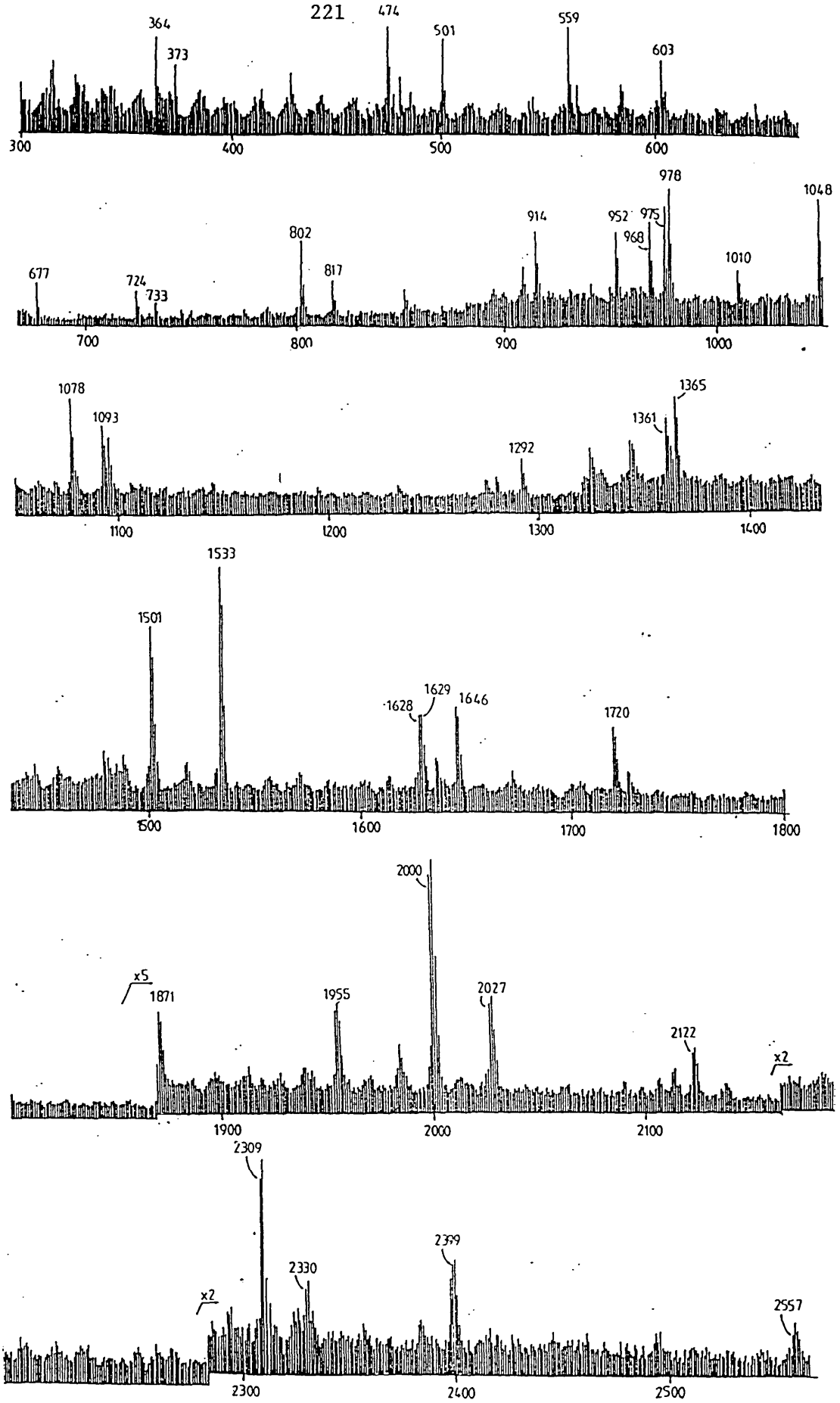


Figure 5.1.4 Chymotryptic FAB-Map of PDC subunit A1.

$[M+H]^+$	Residue Assignment
336	369-370
337	374-376
365	423-426
389	263-266
446	283-287
417	12-14
477	498-501
480	412-416
503	487-490
538	35-38
553	502-506
561	441-446
605	24-29
641	174-179
681	83-89
688	46-52
862	217-224
906	39-45
919	510-517
1031	225-233
1118	207-216
1481	427-439

Table 5.1.4 Peptide Fragments Identified after HPLC of the Chymotryptic Digest of A1, by FAB-MS, with Residue Assignments on to the Gene-derived Sequence.

sequence, which are listed together with their sequence assignments in Table 5.1.3 (see page 218). To isolate and identify any chymotryptic fragments generated, but not observed by FAB-MS analysis of the intact digest mixture due to suppression, the chymotryptic digest was purified by reverse-phase High Performance Liquid Chromatography, HPLC.

Analysis of the HPLC fractions by FAB-MS allowed detection of a number of peptide fragments not previously observed in the FAB-Map of the unfractionated digest mixture (Figure 5.1.4). FAB-MS analysis of the HPLC fractions allowed identification of twenty two peptides which could be mass assigned on to the gene sequence, these are presented in Table 5.1.4 together with their corresponding sequence assignments.

The amino acid residues which were confirmed by the FAB-Mapping experiments (trypsin, V8 protease, chymotrypsin and CNBr digestion) are highlighted on the gene sequence for each subunit A1, A2 and B in Figures 5.1.5, 5.1.6 and 5.1.7 respectively. It should be noted that less sequence is mapped for subunits A2 and B since, they were not chymotryptically digested. From the sequence highlighted in Figures 5.1.5, 5.1.6 and 5.1.7 it is apparent that there are a number of regions in the gene-derived sequence, for example between residues 90-127 and 377-422, for which very little sequence has been assigned. The C-terminal region of the protein has not been assigned from the above "mapping" experiments. It had been observed on screening the gene-derived sequence that methionine residues were infrequent. Furthermore, the regions of interest, i.e. where very little of the sequence had been assigned, were found to be

1 met ser glu ile thr leu gly lys tyr leu phe glu arg leu lys gln val asn val
20 asn thr val phe gly leu pro gly asp phe asn leu ser leu leu asp lys ile tyr
39 glu val glu gly met arg trp ala gly asn ala asn glu leu asn ala ala tyr ala
58 ala asp gly tyr ala arg ile lys gly met ser cys ile ile thr thr phe gly val
77 gly glu leu ser ala leu asn gly ile ala gly ser tyr ala glu his val gly val
96 leu his val val gly val pro ser ile ser ser gln ala lys gln leu leu leu his
115 thr leu gly asn gly asp phe thr val phe his arg met ser ala asn ile ser glu
134 thr thr ala met ile thr asp ile cys thr pro gln ala glu ile asp arg cys ile
153 arg thr thr tyr val thr gln arg pro val tyr leu gly leu pro ala asn leu val
172 asp leu asn val pro ala lys leu leu gln thr pro ile asp met ser leu lys pro
191 asn asp ala glu ser glu lys glu val ile asp thr ile leu val leu val lys asp
210 ala lys asn pro val ile leu ala asp ala cys cys ser arg his asp val lys ala
229 glu thr lys lys leu ile asp leu thr gln phe pro ala phe val thr pro met gly
248 lys gly ser ile ser glu gln his pro arg tyr gly gly val tyr val gly thr leu
267 ser lys pro glu val lys glu ala val glu ser ala asp leu ile leu ser val gly
286 ala leu leu ser asp phe asn thr gly ser phe ser tyr ser tyr lys thr lys asn
305 ile val glu phe his ser asp his met lys ile arg asn ala thr phe pro gly val
324 gln met lys phe val leu gln lys leu leu thr thr ile ala asp ala ala lys gly
343 tyr lys pro val ala val pro ala arg thr pro ala asn ala ala val pro ala ser
362 thr pro leu lys gln glu trp met trp asn gln leu gly asn phe leu gln glu gly
381 asp val val ile ala glu thr gly thr ser ala phe gly ile asn gln thr thr phe
400 pro asn asn thr tyr gly ile ser gln val leu trp gly ser ile gly phe tyr tyr
419 gly ala thr leu gly ala ala phe ala ala glu glu ile asp pro lys lys arg val
438 ile leu phe ile gly asp gly ser leu gln leu thr val gln glu ile ser thr met
457 ile arg trp gly leu lys pro tyr leu phe val leu asn asn asp gly tyr thr ile
476 glu lys leu ile his gly pro lys ala gln tyr asn glu ile gln gly trp asp his
495 leu ser leu leu pro thr phe gly ala lys asp tyr glu thr his arg val ala thr
514 thr gly glu trp asp lys leu thr gln asp lys ser phe asn asp asn ser lys ile
533 arg met ile glu val met leu pro val phe asp cys ser thr lys leu gly

Figure 5.1.5 Amino Acid Residues able to be Assigned for PDC A1 from the FAB-Mapping Experiments.

1	met	ser	glu	ile	thr	leu	gly	lys	tyr	leu	phe	glu	arg	leu	lys	gln	val	asn	val
20	asn	thr	val	phe	gly	leu	pro	gly	asp	phe	asn	leu	ser	leu	leu	asp	lys	ile	tyr
39	glu	val	glu	gly	met	arg	trp	ala	gly	asn	ala	asn	glu	leu	asn	ala	ala	tyr	ala
58	ala	asp	gly	tyr	ala	arg	ile	lys	gly	met	ser	cys	ile	ile	thr	thr	phe	gly	val
77	gly	glu	leu	ser	ala	leu	asn	gly	ile	ala	gly	ser	tyr	ala	glu	his	val	gly	val
96	leu	his	val	val	gly	val	pro	ser	ile	ser	ser	gln	ala	lys	gln	leu	leu	leu	his
115	thr	leu	gly	asn	gly	asp	phe	thr	val	phe	his	arg	met	ser	ala	asn	ile	ser	glu
134	thr	thr	ala	met	ile	thr	asp	ile	cys	thr	pro	gln	ala	glu	ile	asp	arg	cys	ile
153	arg	thr	thr	tyr	val	thr	gln	arg	pro	val	tyr	leu	gly	leu	pro	ala	asn	leu	val
172	asp	leu	asn	val	pro	ala	lys	leu	leu	gln	thr	pro	ile	asp	met	ser	leu	lys	pro
191	asn	asp	ala	glu	ser	glu	lys	glu	val	ile	asp	thr	ile	leu	val	leu	val	lys	asp
210	ala	lys	asn	pro	val	ile	leu	ala	asp	ala	cys	cys	ser	arg	his	asp	val	lys	ala
229	glu	thr	lys	lys	leu	ile	asp	leu	thr	gln	phe	pro	ala	phe	val	thr	pro	met	gly
248	lys	gly	ser	ile	ser	glu	gln	his	pro	arg	tyr	gly	gly	val	tyr	val	gly	thr	leu
267	ser	lys	pro	glu	val	lys	glu	ala	val	glu	ser	ala	asp	leu	ile	leu	ser	val	gly
286	ala	leu	leu	ser	asp	phe	asn	thr	gly	ser	phe	ser	tyr	ser	tyr	lys	thr	lys	asn
305	ile	val	glu	phe	his	ser	asp	his	met	lys	ile	arg	asn	ala	thr	phe	pro	gly	val
324	gln	met	lys	phe	val	leu	gln	lys	leu	leu	thr	thr	ile	ala	asp	ala	ala	lys	gly
343	tyr	lys	pro	val	ala	val	pro	ala	arg	thr	pro	ala	asn	ala	ala	val	pro	ala	ser
362	thr	pro	leu	lys	gln	glu	trp	met	trp	asn	gln	leu	gly	asn	phe	leu	gln	glu	gly
381	asp	val	val	ile	ala	glu	thr	gly	thr	ser	ala	phe	gly	ile	asn	gln	thr	thr	phe
400	pro	asn	asn	thr	tyr	gly	ile	ser	gln	val	leu	trp	gly	ser	ile	gly	phe	tyr	tyr
419	gly	ala	thr	leu	gly	ala	ala	phe	ala	ala	glu	glu	ile	asp	pro	lys	lys	arg	val
438	ile	leu	phe	ile	gly	asp	gly	ser	leu	gln	leu	thr	val	gln	glu	ile	ser	thr	met
457	ile	arg	trp	gly	leu	lys	pro	tyr	leu	phe	val	leu	asn	asn	asp	gly	tyr	thr	ile
476	glu	lys	leu	ile	his	gly	pro	lys	ala	gln	tyr	asn	glu	ile	gln	gly	trp	asp	his
495	leu	ser	leu	leu	pro	thr	phe	gly	ala	lys	asp	tyr	glu	thr	his	arg	val	ala	thr
514	thr	gly	glu	trp	asp	lys	leu	thr	gln	asp	lys	ser	phe	asn	asp	asn	ser	lys	ile
533	arg	met	ile	glu	val	met	leu	pro	val	phe	asp	cys	ser	thr	lys	leu	gly		

Figure 5.1.6 Amino Acid Residues able to be Assigned for PDC A2 from the FAB-Mapping Experiments.

1	met	ser	glu	ile	thr	leu	gly	lys	tyr	leu	phe	glu	arg	leu	lys	gln	val	asn	val
20	asn	thr	val	phe	gly	leu	pro	gly	asp	phe	asn	leu	ser	leu	leu	asp	lys	ile	tyr
39	glu	val	glu	gly	met	arg	trp	ala	gly	asn	ala	asn	glu	leu	asn	ala	ala	tyr	ala
58	ala	asp	gly	tyr	ala	arg	ile	lys	gly	met	ser	cys	ile	ile	thr	thr	phe	gly	val
77	gly	glu	leu	ser	ala	leu	asn	gly	ile	ala	gly	ser	tyr	ala	glu	his	val	gly	val
96	leu	his	val	val	gly	val	pro	ser	ile	ser	ser	gln	ala	lys	gln	leu	leu	leu	his
115	thr	leu	gly	asn	gly	asp	phe	thr	val	phe	his	arg	met	ser	ala	asn	ile	ser	glu
134	thr	thr	ala	met	ile	thr	asp	ile	cys	thr	pro	gln	ala	glu	ile	asp	arg	cys	ile
153	arg	thr	thr	tyr	val	thr	gln	arg	pro	val	tyr	leu	gly	leu	pro	ala	asn	leu	val
172	asp	leu	asn	val	pro	ala	lys	leu	leu	gln	thr	pro	ile	asp	met	ser	leu	lys	pro
191	asn	asp	ala	glu	ser	glu	lys	glu	val	ile	asp	thr	ile	leu	val	leu	val	lys	asp
210	ala	lys	asn	pro	val	ile	leu	ala	asp	ala	cys	cys	ser	arg	his	asp	val	lys	ala
229	glu	thr	lys	lys	leu	ile	asp	leu	thr	gln	phe	pro	ala	phe	val	thr	pro	met	gly
248	lys	gly	ser	ile	ser	glu	gln	his	pro	arg	tyr	gly	gly	val	tyr	val	gly	thr	leu
267	ser	lys	pro	glu	val	lys	glu	ala	val	glu	ser	ala	asp	leu	ile	leu	ser	val	gly
286	ala	leu	leu	ser	asp	phe	asn	thr	gly	ser	phe	ser	tyr	ser	tyr	lys	thr	lys	asn
305	ile	val	glu	phe	his	ser	asp	his	met	lys	ile	arg	asn	ala	thr	phe	pro	gly	val
324	gln	met	lys	phe	val	leu	gln	lys	leu	leu	thr	thr	ile	ala	asp	ala	ala	lys	gly
343	tyr	lys	pro	val	ala	val	pro	ala	arg	thr	pro	ala	asn	ala	ala	val	pro	ala	ser
362	thr	pro	leu	lys	gln	glu	trp	met	trp	asn	gln	leu	gly	asn	phe	leu	gln	glu	gly
381	asp	val	val	ile	ala	glu	thr	gly	thr	ser	ala	phe	gly	ile	asn	gln	thr	thr	phe
400	pro	asn	asn	thr	tyr	gly	ile	ser	gln	val	leu	trp	gly	ser	ile	gly	phe	tyr	tyr
419	gly	ala	thr	leu	gly	ala	ala	phe	ala	ala	glu	glu	ile	asp	pro	lys	lys	arg	val
438	ile	leu	phe	ile	gly	asp	gly	ser	leu	gln	leu	thr	val	gln	glu	ile	ser	thr	met
457	ile	arg	trp	gly	leu	lys	pro	tyr	leu	phe	val	leu	asn	asn	asp	gly	tyr	thr	ile
476	glu	lys	leu	ile	his	gly	pro	lys	ala	gln	tyr	asn	glu	ile	gln	gly	trp	asp	his
495	leu	ser	leu	leu	pro	thr	phe	gly	ala	lys	asp	tyr	glu	thr	his	arg	val	ala	thr
514	thr	gly	glu	trp	asp	lys	leu	thr	gln	asp	lys	ser	phe	asn	asp	asn	ser	lys	ile
533	arg	met	ile	glu	val	met	leu	pro	val	phe	asp	cys	ser	thr	lys	leu	gly		

Figure 5.1.7 Amino Acid Residues able to be Assigned for PDC B from the FAB-Mapping Experiments.

methionine containing. Hence, a CNBr chemical cleavage of the intact protein subunits was selected as an appropriate strategy to investigate the regions, in the primary sequence, which had not been assigned.

5.2.1.5 CNBr Cleavage of Intact Subunits A1, A2 and B of PDC

Samples of intact protein subunits, A1, A2 and B, were chemically cleaved with CNBr, and an aliquot of each digest mixture analysed by FAB-MS. The masses of the predicted CNBr generated peptides are listed in Table 5.1.5. For each peptide there are two theoretical masses corresponding to the lactone and homoserine derivatives of methionine, a characteristic 18 a.m.u. apart. The chemistry of the CNBr cleavage of a peptide bond is shown in Chapter One, Figure 1.2. FAB-MS analysis of each CNBr digest mixture allowed identification of three predicted peptides at m/z; 976, 1313 and 2552. The peptide at m/z 976 was found to map on to the gene-derived sequence between residues 128 to 137, a region previously not assigned, and to correspond to the lactone modification of the C-terminal methionine residue. Furthermore, the peptide at m/z 2552 was assigned to residues 44 to 67 in the gene-derived sequence; this peptide mass corresponds to the homoserine derivative of methionine. Identification of this peptide allowed twenty two amino acid residues, for which no sequence data had been previously assigned, to be mapped. The CNBr fragment m/z 1313 (lactone modification) was found to be derived from the protein sequence between residues 314 and 325, a stretch of sequence previously mapped. These peptides masses were observed in each of the CNBr FAB-Maps of all three subunits. In order to

Predicted CNBr Fragments	$[M+H]^+$ Lactone	Homoserine
1-43	4781	4799
44-67	2364	2552
68-127	6197	6215
128-137	976	994
138-186	5431	5449
187-246	6556	6574
247-313	7225	7243
314-325	1313	1331
326-369	4641	4659
370-456	9336	9355
457-534	9068	9086
535-538	443	461
539-549	1179	

Table 5.1.5 The Predicted Peptide Fragments, Based on The Gene Sequence, After Cyanogen Bromide Cleavage.

investigate whether the other two predicted low mass CNBr peptides, (535-538, $[M+H]^+$ 461, and the C-terminus 539-549, $[M+H]^+$ 1179), see Table 5.1.5, were formed, but not observed due to suppression effects in the mass spectrometer, the CNBr digest mixture was purified.

FAB-MS analysis of the HPLC fractions identified peptides at m/z 976, 1313 and 2552, previously observed in the FAB-Maps of the unfractionated CNBr digest mixtures. A signal m/z 2568 was also detected in the same HPLC fraction as the peptide at m/z 2552. This signal was explained by the probable oxidation of a tryptophan residue found in the sequence of the peptide at m/z 2552 which would result in a mass increase of this peptide by 16 a.m.u. However, a curious observation was the presence of two intense signals m/z 1072/1090 (18 a.m.u. apart) detected in the same HPLC fraction as the peptide at m/z 976. The presence of the signal m/z 1090, i.e 18 a.m.u. higher than m/z 1072 is indicative of a peptide with a C-terminally modified methionine residue. Several steps of manual Edman degradation of an aliquot of the HPLC fraction, with analysis after each step by FAB-MS, showed the peptide at m/z 1072 to possess the same N-terminal sequence as that of m/z 976. This inferred that these peptides were related, with the possibility of a post-translational modification of peptide $[M+H]^+$ 976. The mass difference between these peptides was calculated as 96 a.m.u., a value not corresponding to a known post-translational modification, such as γ-carboxyglutamic acid, phophorylation and not large enough for glycosylation. However, a review of the fractionation procedures resolved this matter with the conclusion that this mass

difference is the result of the attachment of a molecule of trifluroacetic acid, TFA, (added to the HPLC buffers) to the peptide $[M+H]^+$ 976. An aliquot of the HPLC fraction containing the peptide at m/z 1072/1090 only was examined by automated gas-phase sequencing. The peptide at m/z 976 had been found to correspond to the sequence, Ser-Ala-Asn-Ile-Ser-Glu-Thr-Thr-Ala-Met (128-137). Automated gas-phase Edman degradation of the peptide at m/z 1072/1090 yielded the partial sequence, Ser-Ala-Asn-Ile-Ser however, beyond Ser-132 the data was not interpretable. It is suggested therefore that the TFA molecule is attached to the glutamic acid residue 133, or threonine 134 or 135. The attachment of a TFA molecule to this peptide suggests that this an extremely active part of the PDC molecule.

Further analysis of the remaining HPLC fractions, by FAB-MS, revealed intense signals m/z; 457 and 913, both of which could not be mass matched on to the gene-derived sequence. Other signals observed were relatively weak and could not be relied on as having been derived from PDC

The sequence data elucidated from the CNBr digests above has been recorded on the gene-derived sequence for each subunit with the previously determined FAB-Mapping data in Figures 5.1.5, 5.1.6 and 5.1.7. From this FAB-Mapping study over 50% of the gene sequence has been confirmed. However, data has been generated to suggest that there are some differences in the PDC protein sequence investigated here, and that determined from the published gene sequence. Furthermore, there is evidence to suggest that the

differences between the subunits A1, A2 and B are relatively small and that limited proteolysis of one polypeptide chain to form the other is unlikely.

5.2.2 B/E Linked Scanning/CAD

Coinciding with this investigation into the primary sequence analysis of pyruvate decarboxylase was a parallel study evaluating B/E linked scanning as a viable method for direct mixture analysis, removing the rate-determining step of purification. Briefly, linked scanning as reported in this thesis involves a study of daughter ion fragments generated from a peptide and enhanced by collisionally activated decompositions, CAD. The term B/E linked scan refers to the simultaneous variation of the magnetic field strength, B, and the electric sector voltage, E, so as a constant relationship is maintained between the two throughout the scan. Collisionally activated decompositions are brought about by the location of a helium containing collision cell in the first field free region of the mass spectrometer (Chapter One, Figure 1.3). When initially introduced as a method for sample ionisation, FAB was shown to possess many advantages over previous techniques employed. However, a drawback of FAB is the relative lack of fragment ions observed in the mass spectrum, though when compared to the positive advantages of the method this disadvantage is greatly outweighed. In fact strategies such as FAB-Mapping as described in section 5.1 of this chapter utilise the lack of fragmenation as a positive feature of FAB, previously discussed in

Chapter Two. Linked scanning/CAD of pure peptides (Chapter Four) has shown that it is possible to deduce the amino acid sequence of the peptide under study from the mass differences between the enhanced fragment ions. The research work evaluating the applicability of linked scanning/CAD for sequence analysis of pure peptides and more importantly sequence analysis of a peptide of an **unfractionated** complex digest mixture is described in Chapter Four of this thesis.

In order to assign several of the unknown signals in the complex tryptic digest mixture of A1, prepared for the FAB-Mapping studies in section 5.1, linked scanning techniques were applied. The tryptic FAB-Map of subunit A1 is shown in Figure 5.1. From the previous FAB-Mapping experiments, it was found that no assignment on to the gene-derived sequence could be made for peptides at m/z 789 and m/z 1016, observed in the tryptic FAB-Maps of all three subunits. An aliquot of the tryptic digest of A1 was examined by linked scan/CAD initially for the peptide at m/z 789. The computer was programmed for a linked scan of m/z 789 with the resulting linked scan/CAD spectrum shown in Chapter Four, Figure 4.6. From the mass differences between the intense fragment ions the sequence, **Acetyl-Ser-Glu-Ile-Thr-Leu-Gly-Lys**, was elucidated. A detailed explanation of the spectrum interpretation is presented in full in Chapter Four of this thesis. This peptide was found to correspond to the N-terminus of the gene-derived sequence. However, the data revealed a processing event of the N-terminus and a post-translational modification of the gene-derived sequence. The blocked N-terminus was confirmed by a single step of Edman

degradation after which no mass shift of the signal m/z 789 was observed on FAB-MS analysis.

The high quality data from the linked scan/CAD analysis of an unknown component m/z 789, of the intact tryptic digest of the PDC subunit A1, encouraged the application of this linked scanning method to another unknown peptide component, m/z 1016. The computer was programmed to calculate the B/E scan-law for this peptide mass. The linked scan/CAD spectrum of m/z 1016 is shown in Chapter Four, Figure 4.7. Similarly, from the intense fragment ions in the linked scan spectrum the sequence,
Leu-Leu-Thr-Thr-Ile-Ala-Asp-Ala-Ala-Lys, was determined. A detailed interpretation of this linked scan/CAD spectrum is presented in Chapter Four of this thesis. This sequence was found to correspond closely but not identically to residues 332 to 341 in the gene-derived sequence. This signal was previously unable to be mass matched on to the gene-derived sequence, and the linked scan data shows that the reason is the finding of threonine at position 335 in the sequence not asparagine, as predicted from the gene sequence. As described in Chapter Four, the finding that the linked scan method was limited in applicability to peptides less than 2000 Da pre-empted its use in studying the unknowns at m/z 2104, 2285, 2287, 2521, 2741 and 3322.

This study has illustrated that;
(a) the novel application of B/E linked scan/CAD to direct mixture analysis is ideal for parallel investigations with FAB-Mapping studies for primary sequence analysis, particularly for the

determination of unknown components of complex proteolyic digests, and

(b) a processing and post-translational modification of the N-terminus of the gene-derived sequence has occurred together with the detection of a base difference in the gene sequence, leading to a difference in amino acid sequence [39].

5.2.3 High Mass MS

Access to an instrument with a resolveable mass range up to 15,000 Da at full accelerating voltages allowed those peptide fragments, generated by CNBr cleavage, previously outside the resolvable mass range of the mass spectrometer (ZAB-HF) used, section 5.2.1.5, to be examined. A sample of intact protein, subunit A1, was chemically digested with CNBr. A second sample of native PDC A1 was carboxymethylated prior to CNBr digestion. This carboxymethylation reaction involves the reduction of the protein's disulphide bridges with subsequent substitution of a carboxymethyl group at each free cysteine residue to stop the disulphide bridges reforming. Thus, reduction and carboxymethylation of disulphides disrupts the tertiary structure of the protein. However, the carboxymethylated subunit A1 contained salts and thus needed to be purified. Salts are present due to the reaction conditions of the carboxymethylation which require the protein sample to be dissolved in an alkaline buffer, TRIS-(hydroxymethyl)-methylamine, "TRIS buffer". Salts suppress peptide signals on FAB-MS analysis often forming salt adducts. The salt was removed by loading the sample on

to a C-18 Sep-pak; the protein was eluted by two solvent concentrations, 20% and 40% 1-propanol in 5% acetic acid. Both fractions after lyophilisation appeared (detected visually) to contain protein and hence, the contents of both the 20% and 40% eluates were digested with CNBr. Reference will be made to A1-20% and A1-40% to distinguish between the two eluted protein fractions. An aliquot of the CNBr digests of native A1, 20%-A1 and 40%-A1 were examined by FAB-MS. The predicted CNBr fragments from the gene-derived sequence are listed in Table 5.1.5. The FAB mass spectra generated are shown in Figures 5.2, 5.2.1, 5.2.2, 5.2.3 (disulphide bridged PDC A1), 5.2.4, 5.2.5, (A1-20%) and 5.2.6 (A1-40%). The spectra were computer recorded through a variety of different mass ranges, the absence of spectra covering a certain mass range indicates that there were no significant peaks in these regions. It should be noted that accurate masses are recorded in the spectra and hence an approximate conversion factor substracting half a mass unit per 1000 daltons from the quasi-molecular ions, $[M+H]^+$, in the spectra, may be employed to convert accurate to nominal mass, as previously discussed.

Interpretation of the spectrum in Figure 5.2.

This spectrum is characterised by signals m/z; 457, 977, 1167 and 1314. With the exception of the signal m/z 1167 (though it is noted that this signal is weak) the other mass signals have been previously observed, with the peptide at m/z 977 (nominal mass 976) and at m/z 1314 (nominal mass 1313) having been assigned between positions (128-137) and (314-325) respectively, on to the gene-derived sequence (Figure 5.1.3). The signal m/z 1167 could not

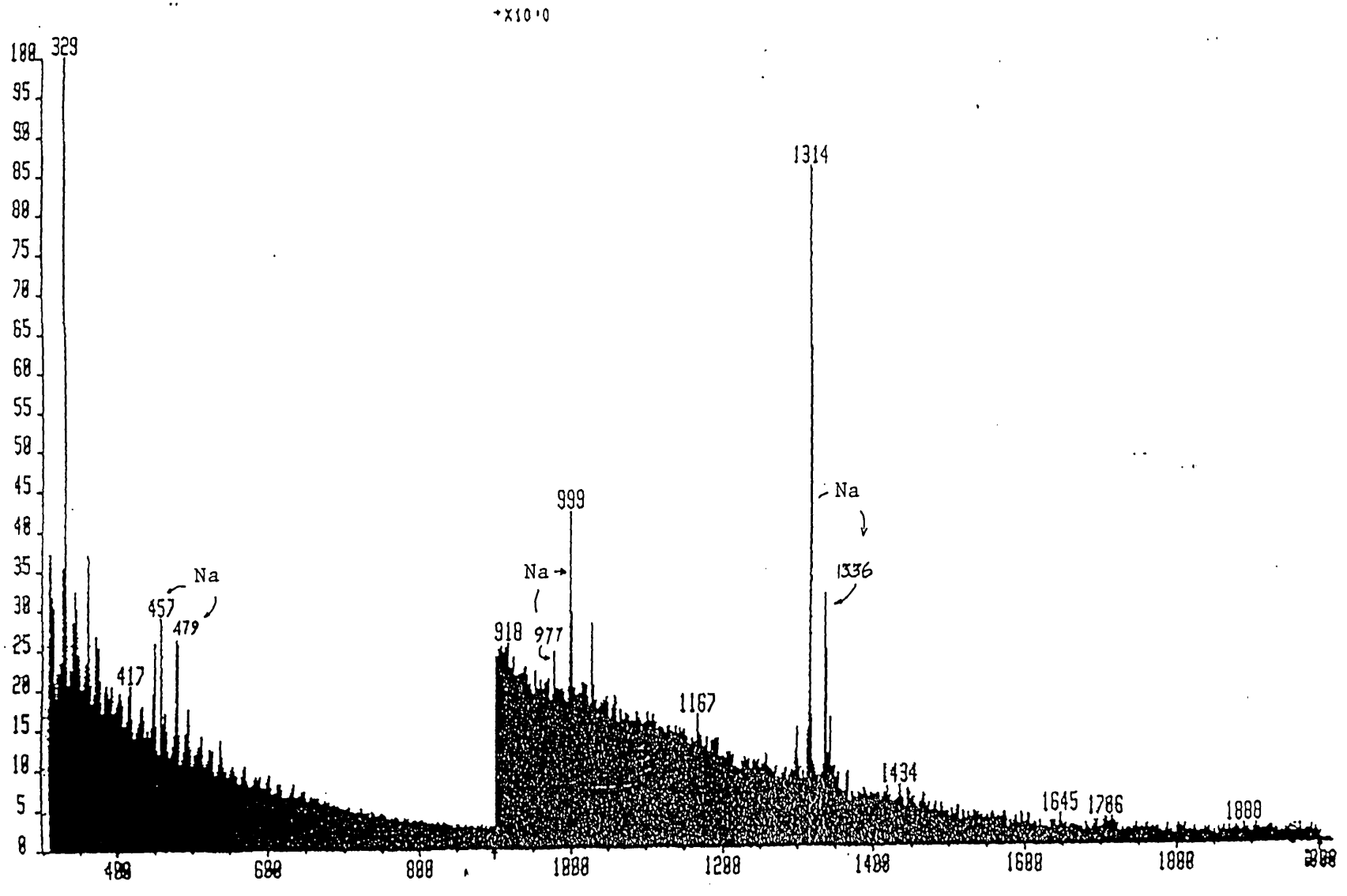


Figure 5.2 FAB-MS of CNBr Digest of PDC A1 between Mass Range 300-2000 Da (ZAB-2SE).

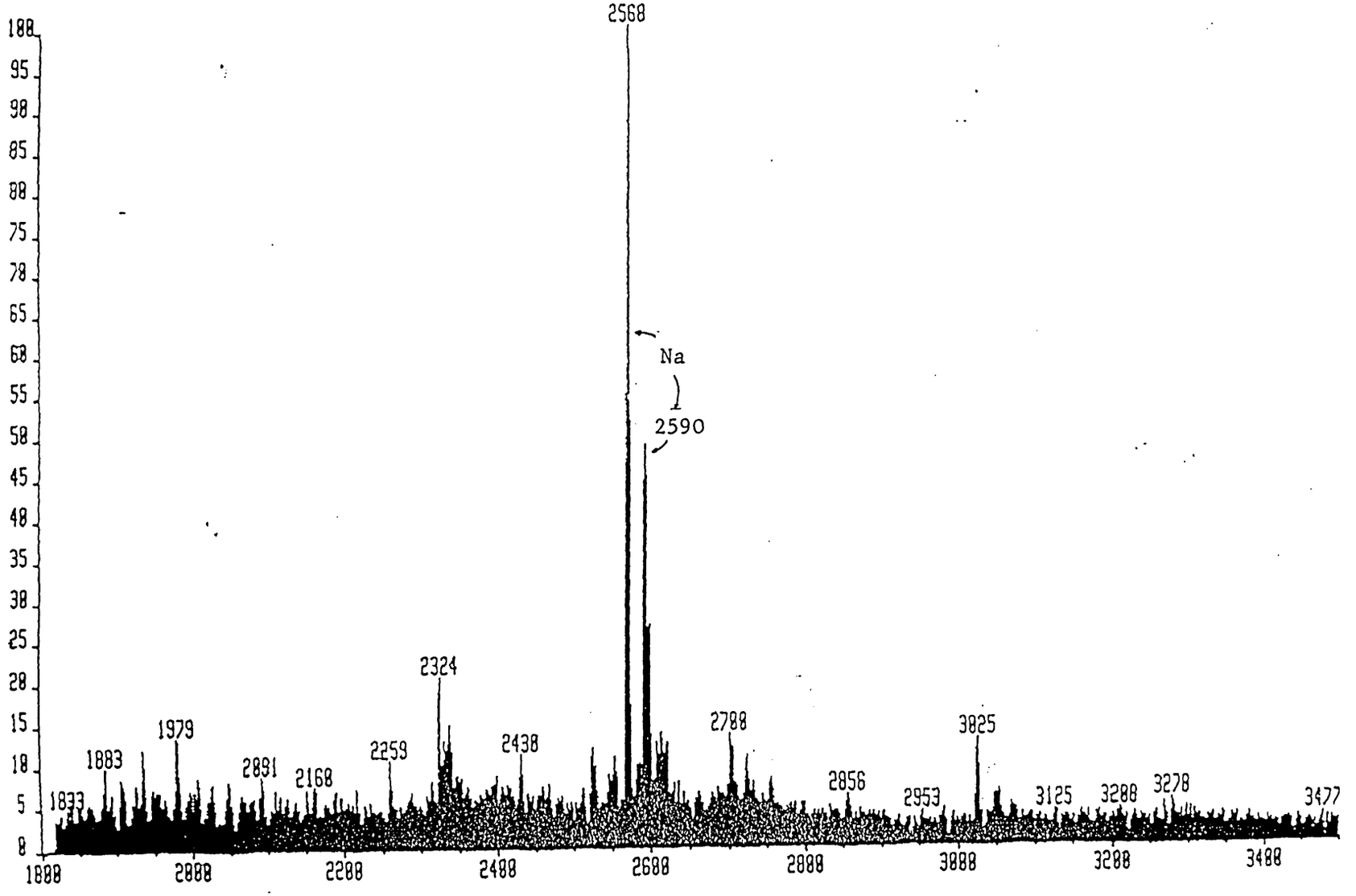


Figure 5.2.1 FAB-MS of CNBr Digest of PDC A1 between Mass Range 1000-3500 Da (ZAB-2SE).

be mass matched on to the gene-derived sequence. This peptide at m/z 1167 is similar in mass to the expected CNBr generated C-terminal fragment m/z 1179. However, the mass difference between the two is only 12 a.m.u. and thus, it was concluded that the peptide at m/z 1167 does not correspond to a truncated C-terminal fragment.

Interpretation of the spectrum in Figure 5.2.1

The most intense signal m/z 2568 in this spectrum was identified as the oxidised form of the peptide at m/z 2552 (44-67). This signal has previously been observed, section 5.1, and assigned as the product of oxidation of the peptide $[M+H]^+$ 2552. This spectrum also contains a number of signals m/z; 2324, 2700 and 3025. None of these mass signals could be mapped on to the gene-derived sequence, although it is noted that these signals are of relatively low intensity.

Interpretation of the spectrum in Figure 5.2.2

The intense signal m/z 4645 in this spectrum was assigned on to the gene-derived sequence as a CNBr peptide fragment, generated from between residues 326 and 369, with the methionine chemically modified to its lactone. The signal m/z 4674 (approximately 28 a.m.u. higher than mass 4645), of lesser intensity, is interpreted as the formyl derivative of the peptide at m/z 4645 (for CNBr cleavage to proceed the sample to be digested is dissolved in 70% formic acid). Another signal in this spectrum which could be assigned to a predicted CNBr fragment was m/z 4782 (nominal mass 4799) which was found to correspond to the peptide 1 to 43 in the

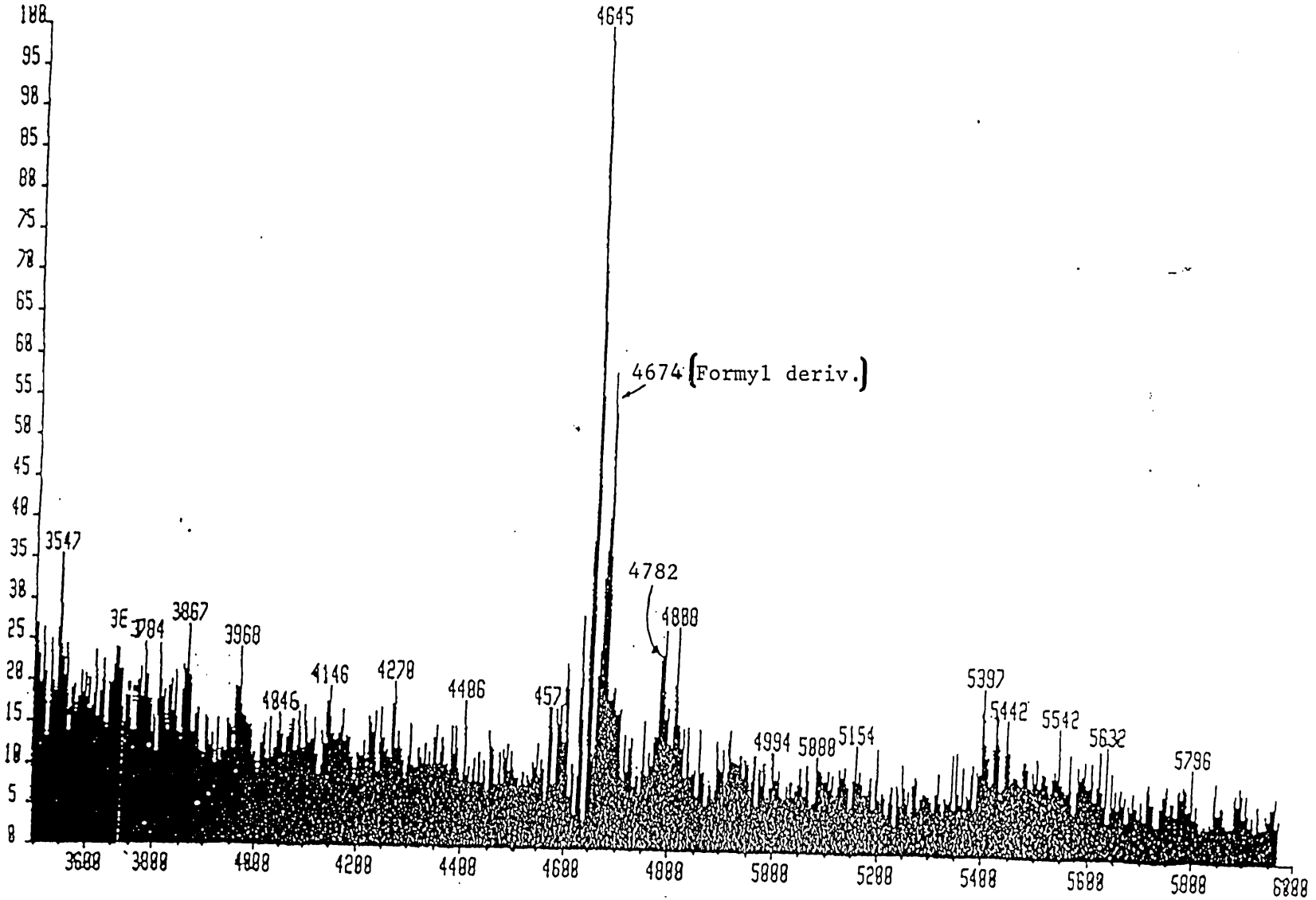


Figure 5.2.2 FAB-MS of CNBr Digest of PDC A1 between Mass Range 3500-6000 Da (ZAB-2SE).

gene-derived sequence, the post-translational modification and processing event at the N-terminus determined from the linked scan/CAD study, section 5.2.2, were incorporated in the mass calculation.

Interpretation of the spectrum in Figure 5.2.3.

In this spectrum intense signals at m/z; 5396, 5422, 6531 and 6558 are observed. Signals m/z 5422 and 6558 were assigned as the formyl derivatives of the peptides at m/z 5396 and 6531 respectively. The computer marks the highest isotope of the cluster and since this varies statistically expansions of the region are used to assign masses. However, neither of these signals m/z 5396 or m/z 6531 could be assigned to a predicted CNBr fragment. The calculated nominal mass for the CNBr fragment predicted to be generated between residues 187 and 246 is 6556 a.m.u. This mass value corresponds to the C-terminal methionine residue chemically modified to its lactone, it was noted that the peptides able to be assigned previously were, in general, detected as their lactone derivatives. The peptide generated from this region (187-246) in the gene-derived sequence contains a stretch of twenty two amino acid residues, 187 to 206, which have not been assigned by the previous FAB-Mapping study in section 5.1. Hence, this region is highlighted as one containing possible sequence changes. The mass difference between the observed mass, 6531 (accurate mass) and the predicted mass, residues 187 to 246, 6556 (nominal mass) is approximately 28 a.m.u. The region, 187 to 206, for which no sequence information was available was reviewed for the predicted tryptic fragments arising between these residues. A tryptic peptide

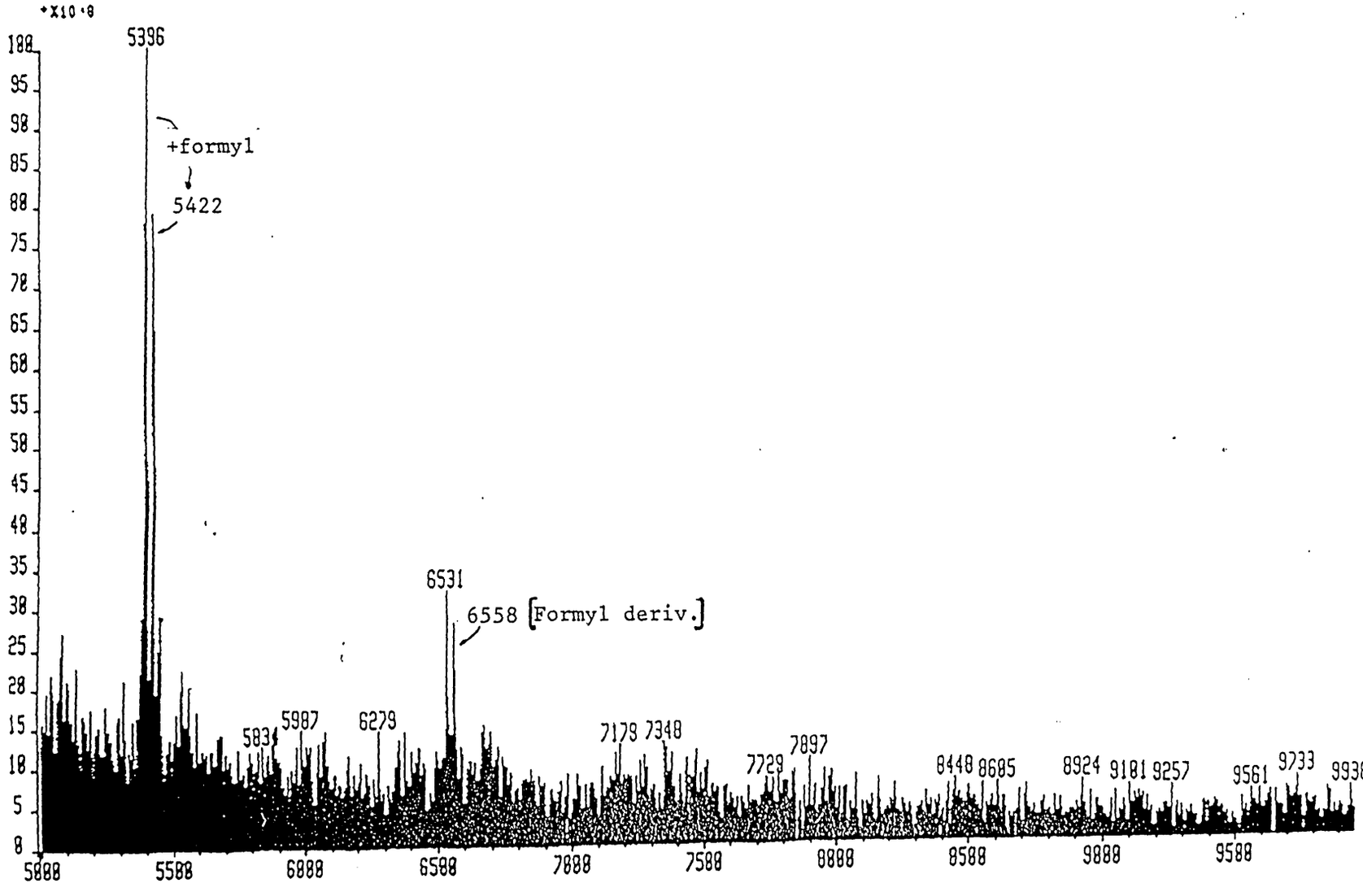


Figure 5.2.3 FAB-MS of CNBr Digest of PDC A1 between Mass Range 5000-10,000 Da (ZAB-2SE).

$[M+H]^+$ 1241 was expected to be generated between residues 198 and 208 however, no such mass signal was observed in the tryptic FAB-Map, Figure 5.1. There was nevertheless an intense signal observed in this mass region in the tryptic FAB-Map, m/z 1213, which was not able to be assigned on to the gene-derived sequence. Interestingly the mass difference between 1241 and 1213 is 28 a.m.u. It is proposed therefore, that the observed peptide at m/z 6531 (acc. mass) maps on to the gene-derived sequence between residues 187 and 246, and that between positions 198 and 208 there is a sequence error. This error may be the result of one or more amino acid changes. There is no other evidence e.g. from Edman/MS, to support this hypothesis though the identification of differences in the gene-derived sequence by linked scanning/CAD in section 5.2 demonstrates this proposal to be realistic.

Interpretation of the spectrum in Figure 5.2.4

This corresponds to a FAB mass spectrum of the carboxymethylated/CNBr digest of protein subunit A1 (A1-20%). scanned between masses 300-2000 Da. The presence of the signal at m/z 1314 (314-325) indicates that this peptide is not cysteine containing and confirms the sequence assignment. The absence of the signal m/z 1167, previously observed in Figure 5.2, with the appearance of the signal m/z 1285 in this spectrum suggests that they may be related with the peptide at m/z 1167 containing two cysteine residues which have been carboxymethylated (+118 a.m.u.). The gene-derived sequence was screened for a possible CNBr fragment containing two cysteine residues and mass 1167 Da however, no assignment could be made. As previously discussed the signal m/z

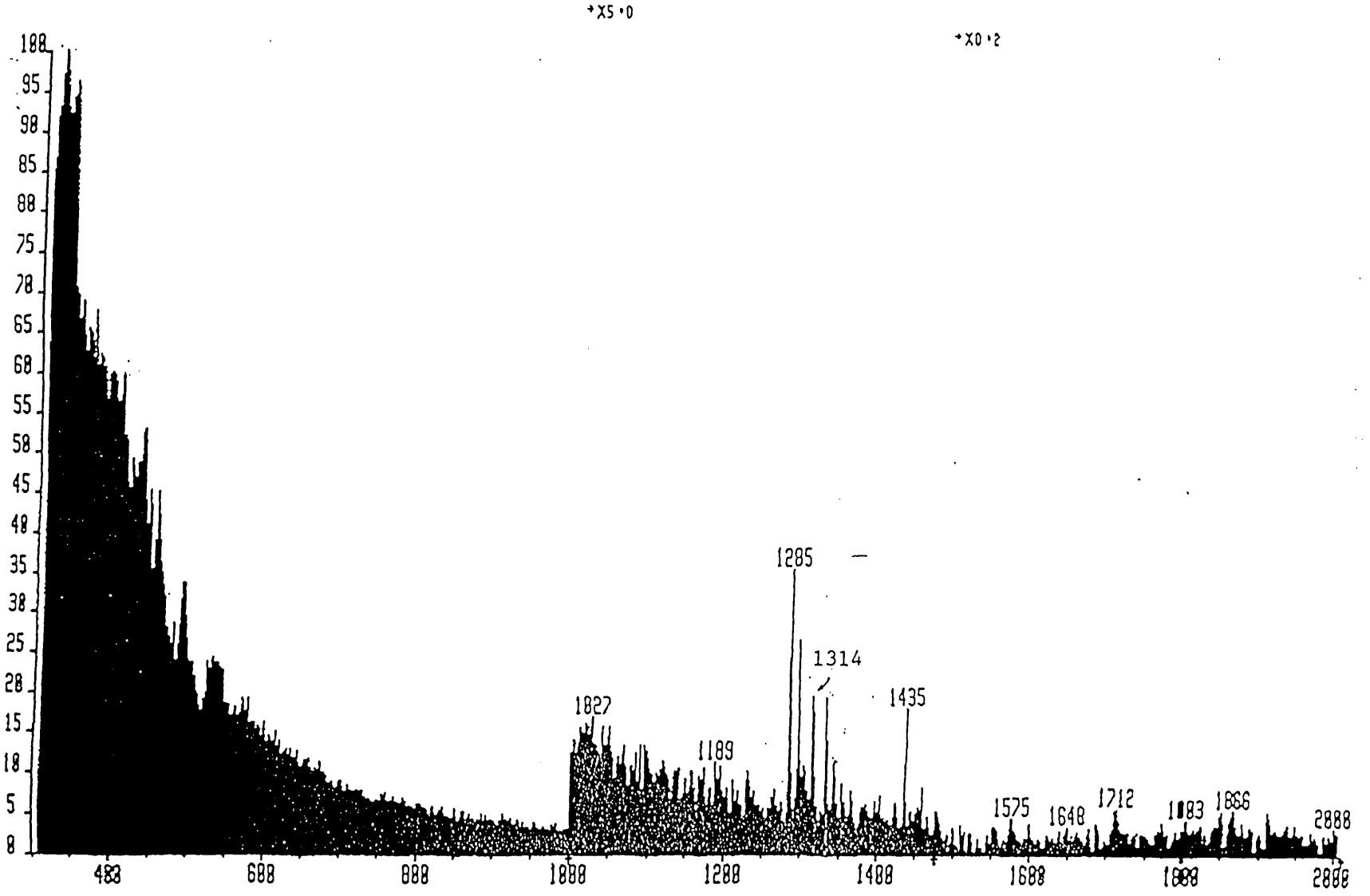


Figure 5.2.4 FAB-MS of Carboxymethylated (20% eluate) and then Digested (CNBr) PDC A1 between Mass Range 300-2000 Da (ZAB-2SE).

1167 is very close in value to the calculated C-terminal fragment m/z 1179, not detected, though this C-terminal sequence contains only one cysteine residue. A single amino acid substitution for example Asp-543 with a cysteine residue, would produce a C-terminal fragment with a mass 1167 a.m.u. with subsequent carboxymethylation generating the signal m/z 1285. However, for cysteine to be misread as aspartic acid at the DNA-level two bases must have been incorrectly interpreted, which would have a low probability of occurring. Furthermore, when this hypothesis was tested by reviewing the FAB-MS data of previous digests, section 5.2.1, no appropriate signal was found to support this proposal with no significant sequence information found which would prove or disprove this proposed explanation of signals m/z 1167 and 1285. Also observed in this spectrum, Figure 5.2.4, is a signal m/z 1435 for which no sequence assignment could be made.

Interpretation of the spectrum in Figure 5.2.5.

The most intense signal, in this spectrum of A1-20%, m/z 2569 corresponds to the peptide 44 to 67 ($[M+H]^+$ 2552) containing an oxidised tryptophan residue. This signal was previously observed in Figure 5.2.1, the presence of this peptide after carboxymethylation indicates that it does not contain cysteine, which confirms the earlier sequence assignment (44-67). Similarly, the presence of the peptide at m/z 3025 (previously observed in Fig. 5.2.1), for which no peptide assignment on to the protein sequence was possible, indicates that it does not contain any cysteine residues. Also observed in this spectrum is a very weak signal m/z 4644/5, this signal had been previously observed in Figure 5.2.2 and assigned on

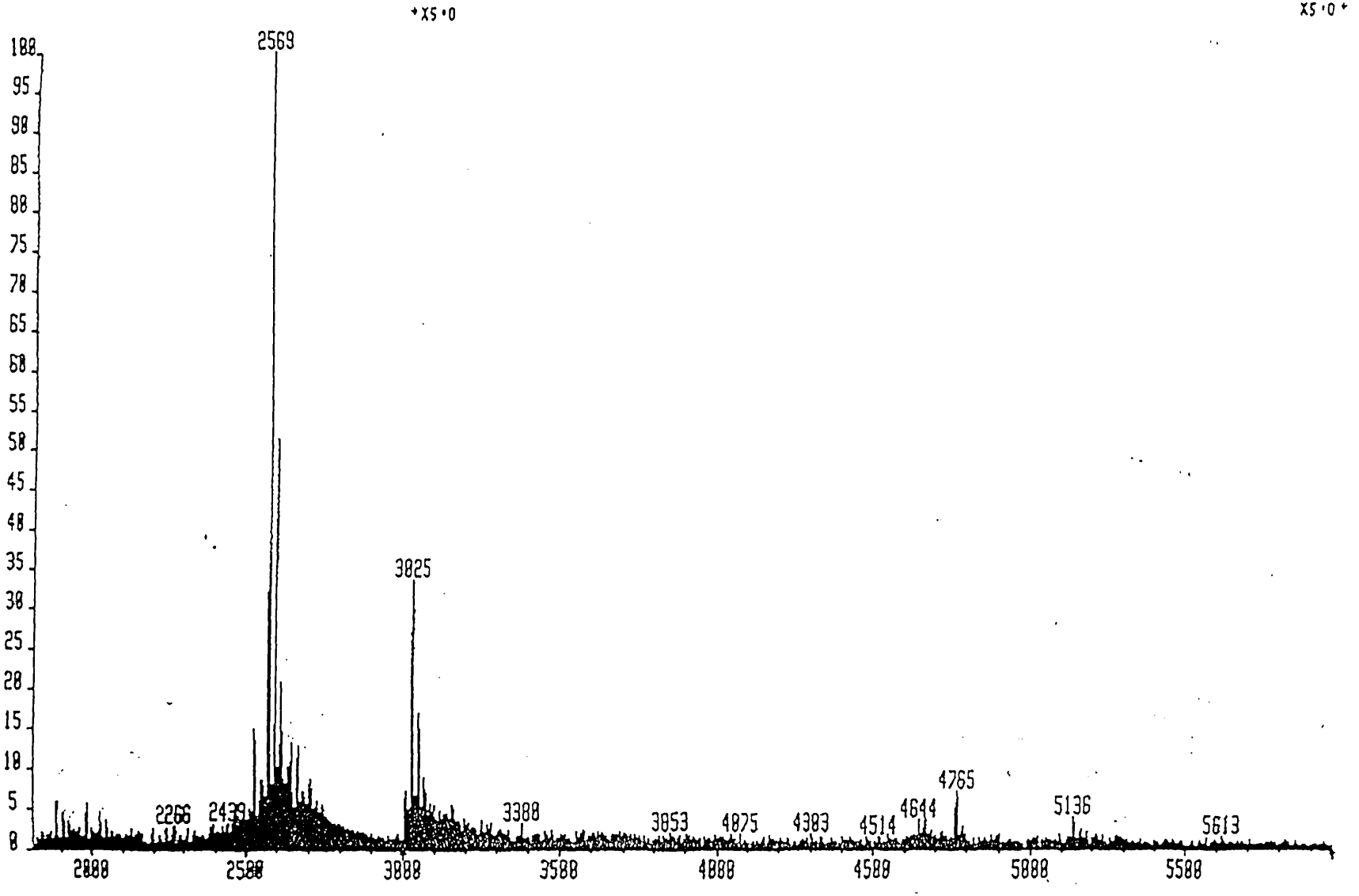


Figure 5.2.5 FAB-MS of Carboxymethylated (20% eluate) and then Digested (CNBr) PDC A1 between Mass Range 1800-6000 Da (ZAB-2SE).

to residues 326 to 369. In a similar mass region as the signal m/z 4644/5 there has appeared a signal m/z 4765 in Figure 5.2.5. It is interesting to note that the approximate mass difference between these two signals corresponds to the carboxymethylation of two cysteine residues. However, the peptide sequence between residues 326 and 369 does not contain any cysteine residues. Therefore, no sequence assignment of the peptide at m/z 4765 could be made. A weak signal at m/z 5136 is also observed in this spectrum however, this is thought to correspond to a dimer of the peptide at m/z 2568 $[2M+H]^+$.

Interpretation of the spectrum in Figure 5.2.6

This spectrum corresponds to a CNBr digest of the carboxymethylated sample eluted off the Sep-pak in 40% propan-1-ol (in 5% acetic acid), A1-40%. This spectrum is characterised by the intense signal cluster around m/z 5549. This approximate mass value would correspond to the carboxymethylated (containing two cysteine residues) predicted CNBr fragment at m/z 5431, 138-186, (lactone) generated by CNBr digestion. However, this peptide at m/z 5431 was not observed in the CNBr digest of oxidised PDC A1. The spectrum in which this signal would have been expected to be observed, Figure 5.2.3 contains only an intense signal m/z 5396 for which no sequence assignment could be made. Thus, the presence of this signal may be suppressing the predicted CNBr fragment (138-186) on FAB-MS analysis of the digest mixture.

The data from these six spectra are summarised in Table 5.2 together with the deduced sequence assignments. The absence of

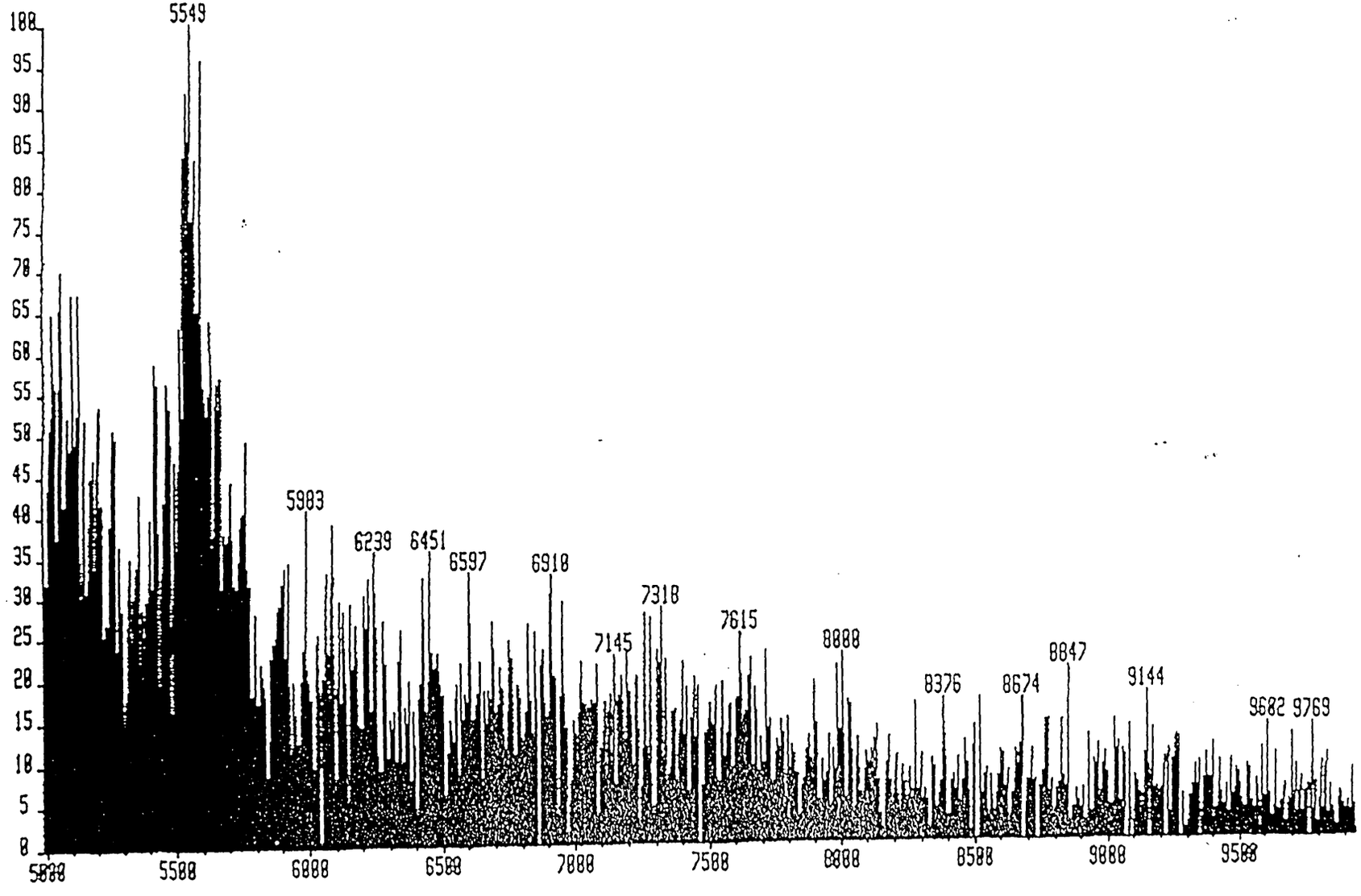


Figure 5.2.6 FAB-MS of Carboxymethylated (40% eluate) and then Digested (CNBr) PDC A1 between Mass Range 5000-10.000 Da (ZAB-2SE).

oxidised A1		A1 CMe-20%	A1 CMe-40%
$[M+H]^+$		$[M+H]^+$	$[M+H]^+$
457		457	
977	(128-137)		
1167	——— 2CYS ——→	1285	
1314	(314-325)	1314	
		1435	
2324			
2568	(44-67)	2569	2568
2700			
3025		3025	
4645	(326-369)	4765	4765
4674	Formyl		
4787			
4782	(1-43)		
		5136	
5396			
			5549 (138-186)
6531			

Table 5.2 High Mass FAB-MS Analysis of CNBr Digest Mixtures of Oxidised (native) and Carboxymethylated Samples of Pyruvate Decarboxylase, Subunit A1.

predicted signals such as m/z 7243, 9086 and 9984 cannot be interpreted as errors in the gene-derived sequence as they may have been generated by the digest mixture but suppressed on FAB-MS analysis. However, the presence of signals such m/z 3025 and 5396 for which no predicted CNBr generated peptide fragment exists indicates that there may be one or two extra methionine residues in this PDC sequence.

This data has allowed confirmation of numerous signals previously assigned to peptide sequences in the gene-derived sequence, for example m/z 1314 and 2568. Several high mass peptides have also been mapped on to large peptides in the protein sequence providing a rapid check of the gene-derived sequence, for example, the signal at m/z 4645 which was able to be assigned to residues 326 to 369. In addition, the post-translational modification and truncation of the PDC protein, determined previously from the linked scan study (section 5.2.2), has been confirmed by the presence of the signal at m/z 4782 (4799-nominal mass) which corresponds to residues 1 to 43 with these changes incorporated. Furthermore, a region of amino acid sequence has been identified as one containing possible errors by the presence of the signal m/z 6531 in spectrum 5.2.3 which is proposed to correspond to residues 187 to 246 (predicted fragment nominal mass, 6556). This difference of approximately 28 a.m.u. is supported by the previously observed signal m/z 1213 (unable to be assigned) in all three tryptic FAB-Maps (Figures 5.1, 5.1.1 and 5.1.2) as discussed. Another large region (138-186) also able to be assigned from this high mass analysis of CNBr fragments is from interpretation of the signal at m/z 5549 in spectrum 5.2.6. This

signal mass is determined to correspond to the carboxymethylated CNBr fragment (lactone modification of methionine) mass 5431 Da. Thus, the data obtained from these high mass studies of a CNBr digest has aided the primary sequence analysis of pyruvate decarboxylase, with interpretation of the FAB mass spectrum allowing determination of regions of possible sequence error. The presence of signals such as m/z 457, 1167 (1265), 2324, 3025, and 5396 (although it is noted that most of these signals are of weak intensity) indicate that there are further regions of sequence difference between the isolated enzyme subunits studied here and that of the gene sequence.

5.2.4 Electrospray Ionisation Mass Spectrometry

The data from sections 5.2.1, 5.2.2 and 5.2.3 indicated that there were sequence differences between the subunits of pyruvate decarboxylase studied here and the gene-derived sequence. The results from those studies reported above indicate that the hypothesis that proteolytic cleavage of one peptide chain generates the other is unlikely. Another component of this study was the investigation of possible sequence differences between each isolated protein subunit. A suitable method was required which would allow differentiation between the separable subunits A1, A2 and B.

Classically molecular weight determinations have employed methods such as SDS-gel electrophoresis. This method provides an

approximate molecular weight; for example, the molecular weight of a PDC monomer unit by this technique was determined as 60,000 ±3000 Da (±5%). A technique which would allow accurate molecular weight determinations would have a great impact upon the field of structural analysis of biopolymers. For example, a gene-derived sequence could be rapidly checked for sequence deletions or post-translational modifications such as glycosylation. The presence of such modifications could be inferred from the calculated molecular weight of the gene-derived protein sequence and that determined practically, by an accurate method. Electrospray ionisation mass spectrometry, ESI-MS, has recently been demonstrated, in this thesis work and by others [105], to allow molecular weight determinations within ±0.02% Thus a molecular weight determination, of a protein monomer unit of PDC, by this method would generate a value within ±10 daltons. An accurate molecular weight determination of subunits A1, A2 and B would aid

(a) correlation of the gene-derived sequence with the practically determined molecular weight, and

(b) an estimation of the extent of the sequence difference, if any, between the subunits.

An ESI spectrum is characterised by a bell shape distribution of multiply charged molecular ions. The multiple charging phenomenon reduces the effective molecular weight range by a factor equal to the number of charges and hence a quadrupole can be employed instead of a conventional two sector mass spectrometer. The resolution of a quadrupole mass spectrometer is low and thus, does

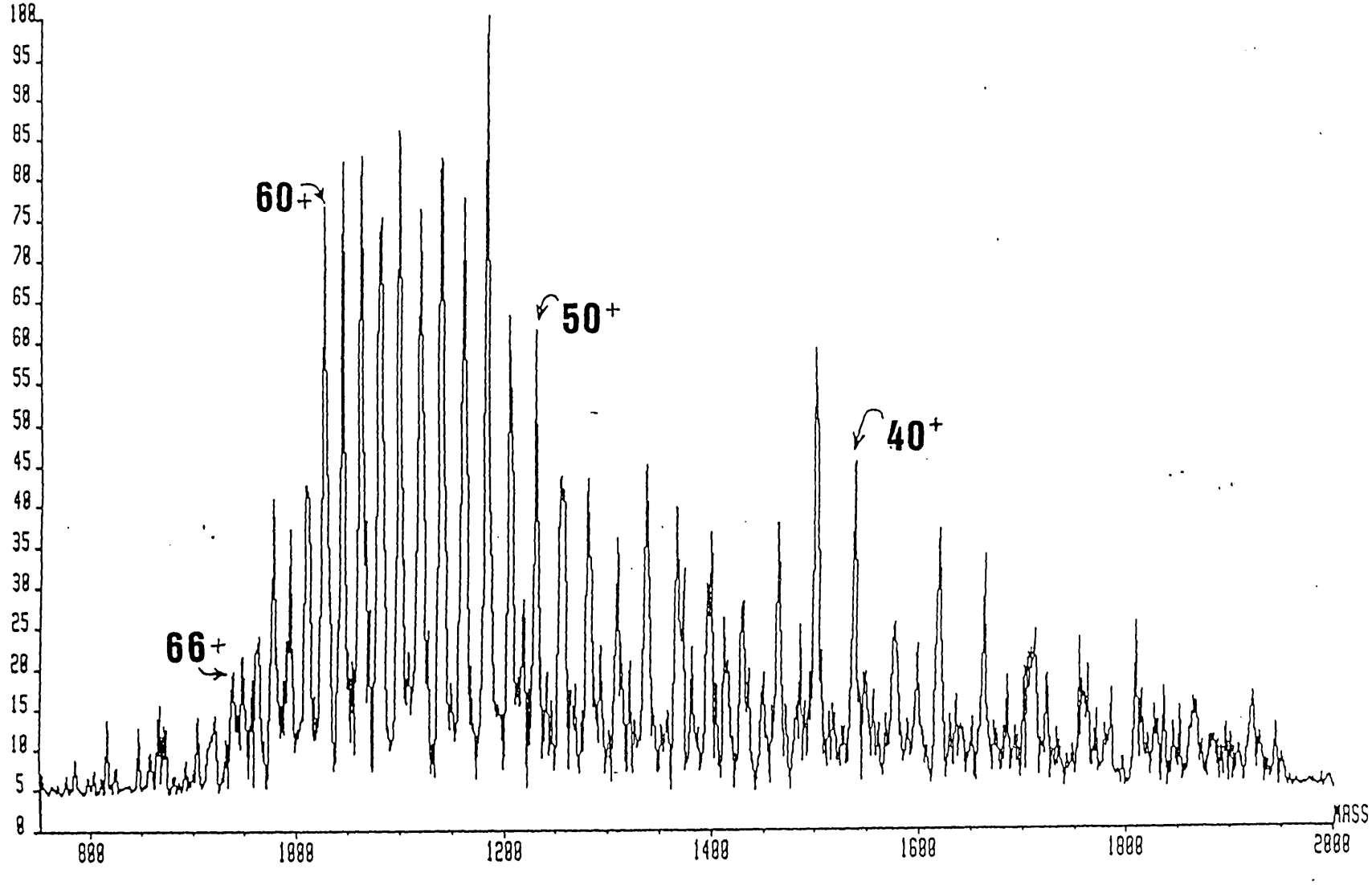


Figure 5.3 ESI Mass Spectrum of PDC A2 Carrying Multiple Positive Charges.

not allow resolution of each isotopic contribution of the molecular ion for high molecular weight substances. Therefore, methods of molecular weight determinations from the most abundant ion or average mass are based upon the centroid of the molecular ion envelope. Adjacent peaks differ by only one charge. The ESI spectrum of enzyme subunit A2 is presented in Figure 5.3 illustrating the characteristic bell-shape formation, as discussed. The expanded bell-shape distribution of each subunit A1, A2 and B are shown in Figures 5.3.1, 5.3.2 and 5.3.3 respectively. These peaks as discussed represent the unresolved isotopic contribution of each charged species and the centroids of these peaks are shown in 5.3.4, 5.3.5 and 5.3.6. It is from these masses assigned to each centroid that the molecular weight calculations are based upon.

Interpretation of the spectra 5.3.4, 5.3.5 and 5.3.6 is by consideration of the relationship between a multiply charged ion m_1 and the molecular weight of the molecule, i.e.

$m_1 n$ = mol. wt. + Hn where n=no. of charges

The molecular weight relationship between two adjacent multiply charged ions m_1 and m_2, where $m_2 > m_1$, is therefore;

$m_2(n-1)$ = mol. wt. + H(n-1)

These two equations may be solved for n

$n = (m_2 - H)/(m_2 - m_1)$

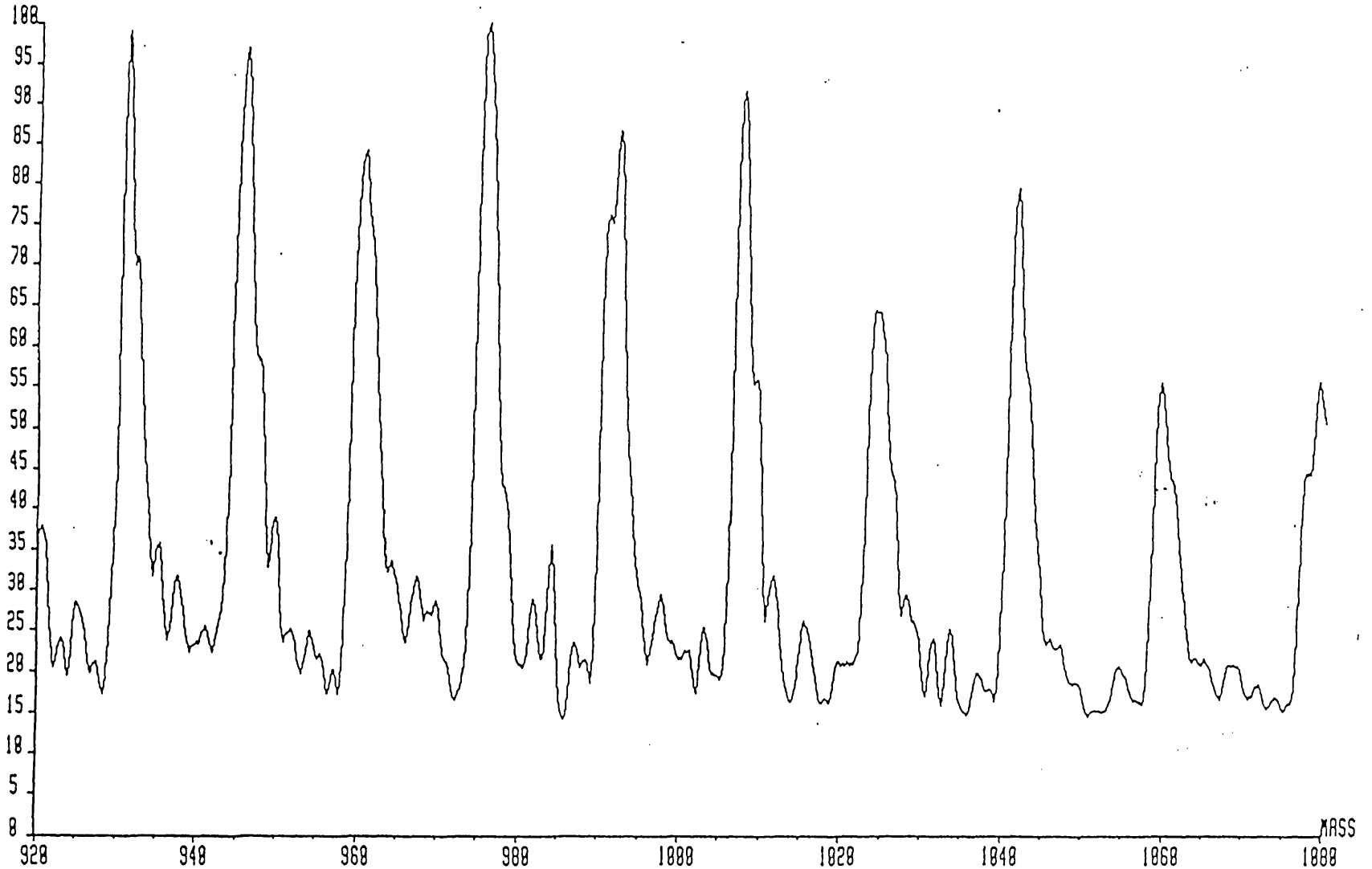


Figure 5.3.1 ESI Mass Spectrum of PDC A1, Illustrating an Expanded Region of its Bell-shape Distribution.

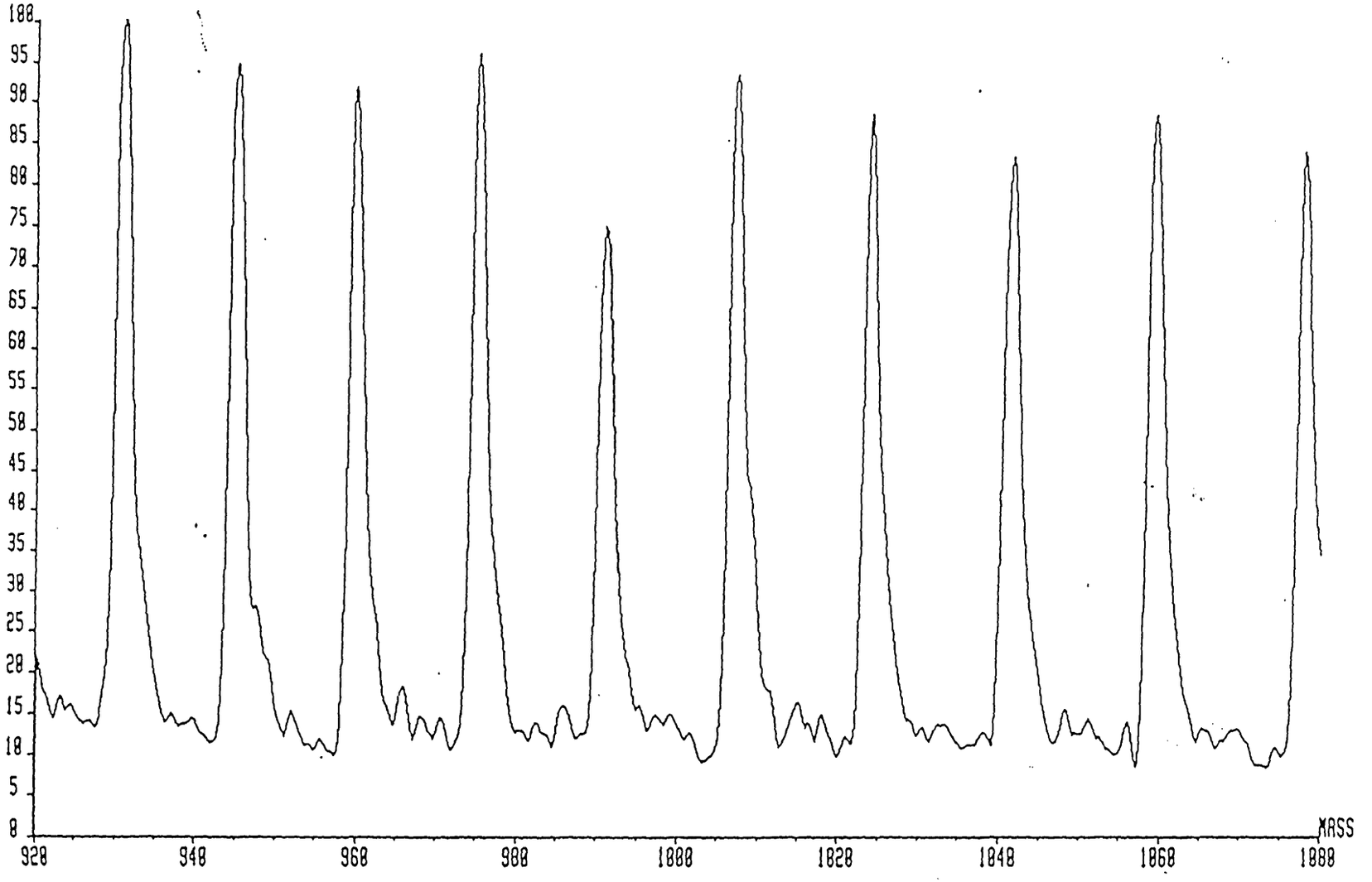


Figure 5.3.2 ESI Mass Spectrum of PDC A2, Illustrating an Expanded Region of its Bell-shape Distribution (Fig. 5.3).

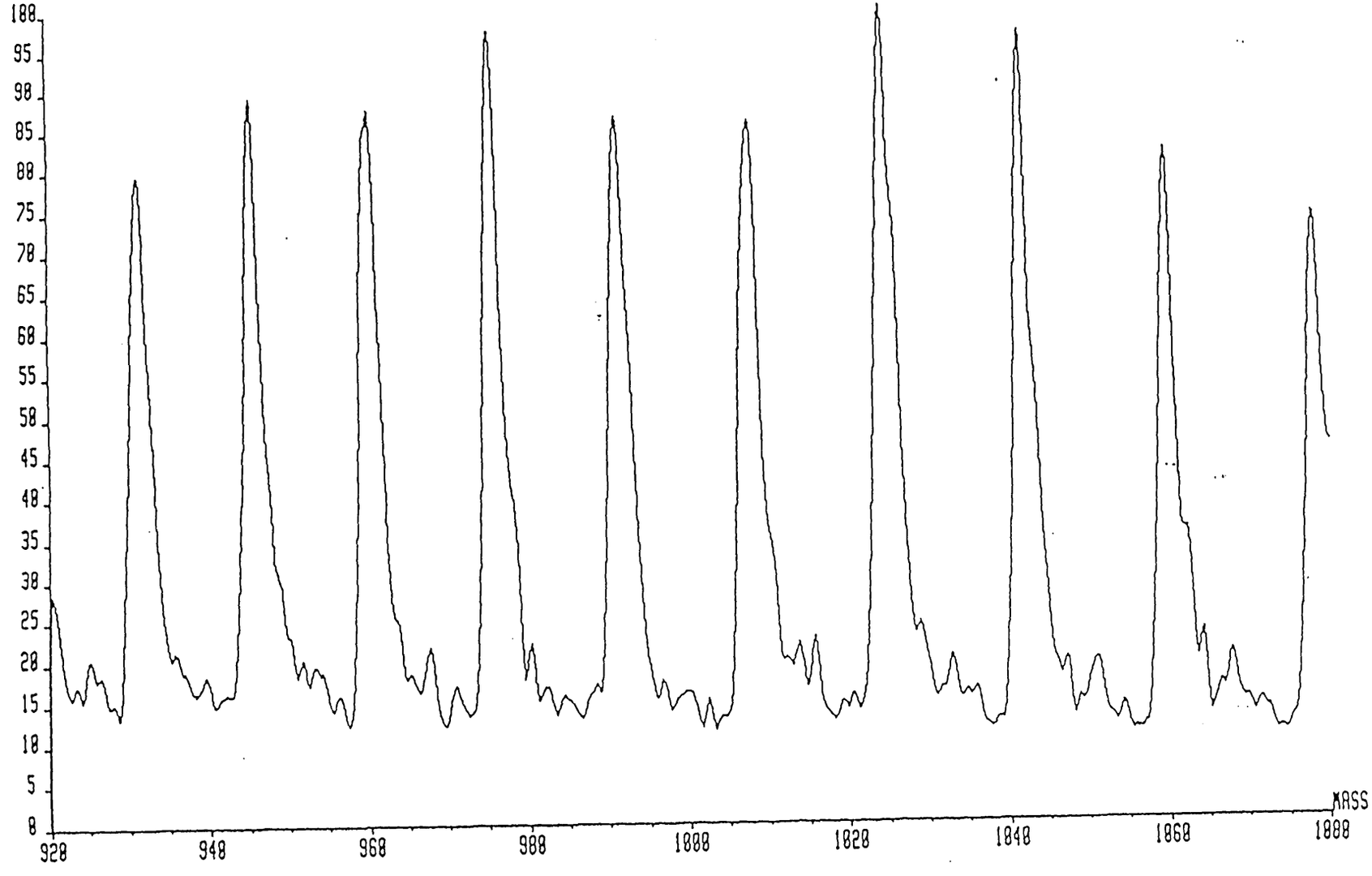


Figure 5.3.3 ESI Mass Spectrum of PDC B, Illustrating an Expanded Region of its Bell-shape Distribution.

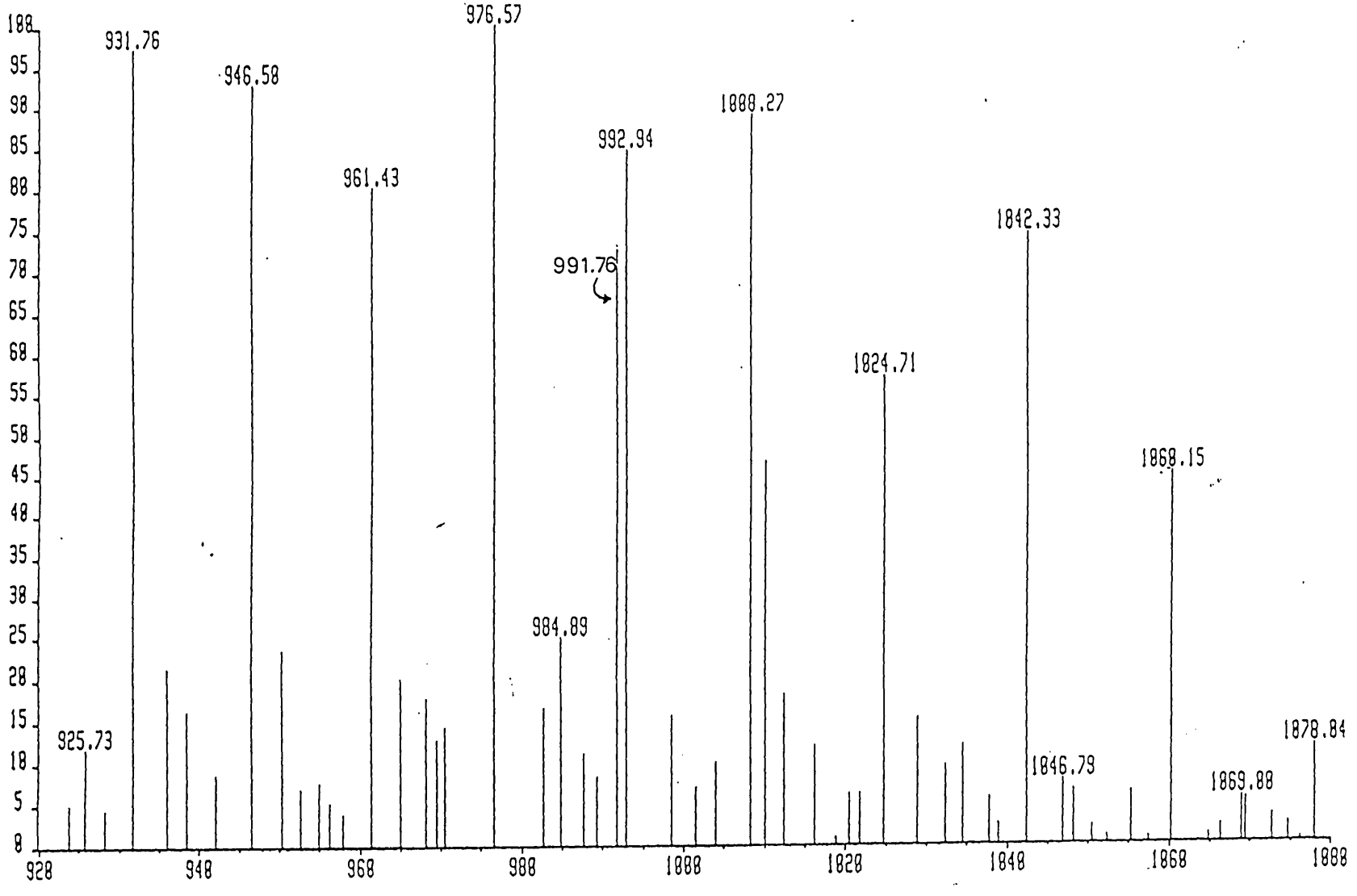


Figure 5.3.4 ESI-MS Containing the Centroids of each Molecular Ion Envelope for PDC A1 Observed in Fig. 5.3.1.

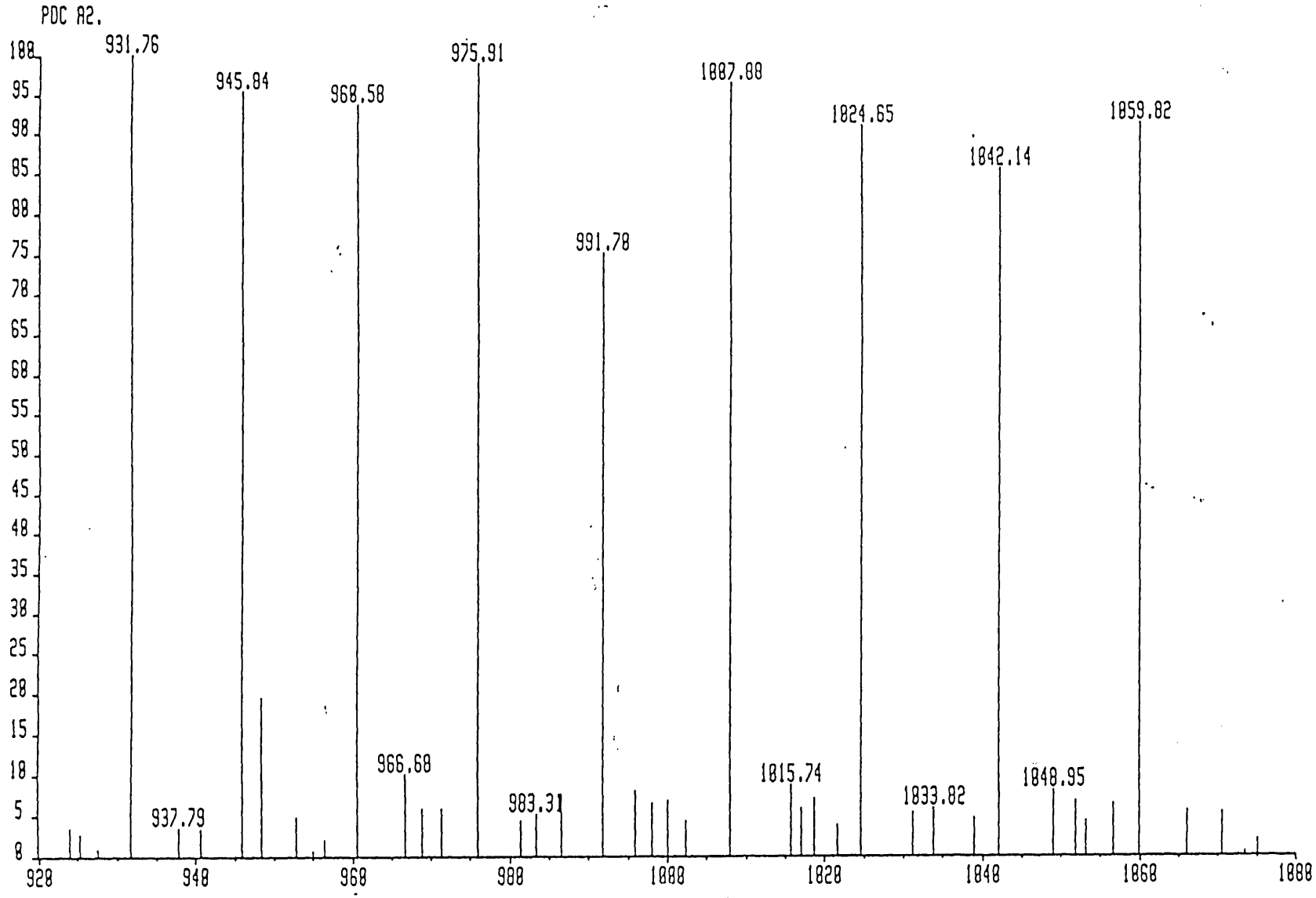


Figure 5.3.5 ESI-MS Containing the Centroids of Each Molecular Ion Envelope for PDC A2 Observed in Fig. 5.3.2.

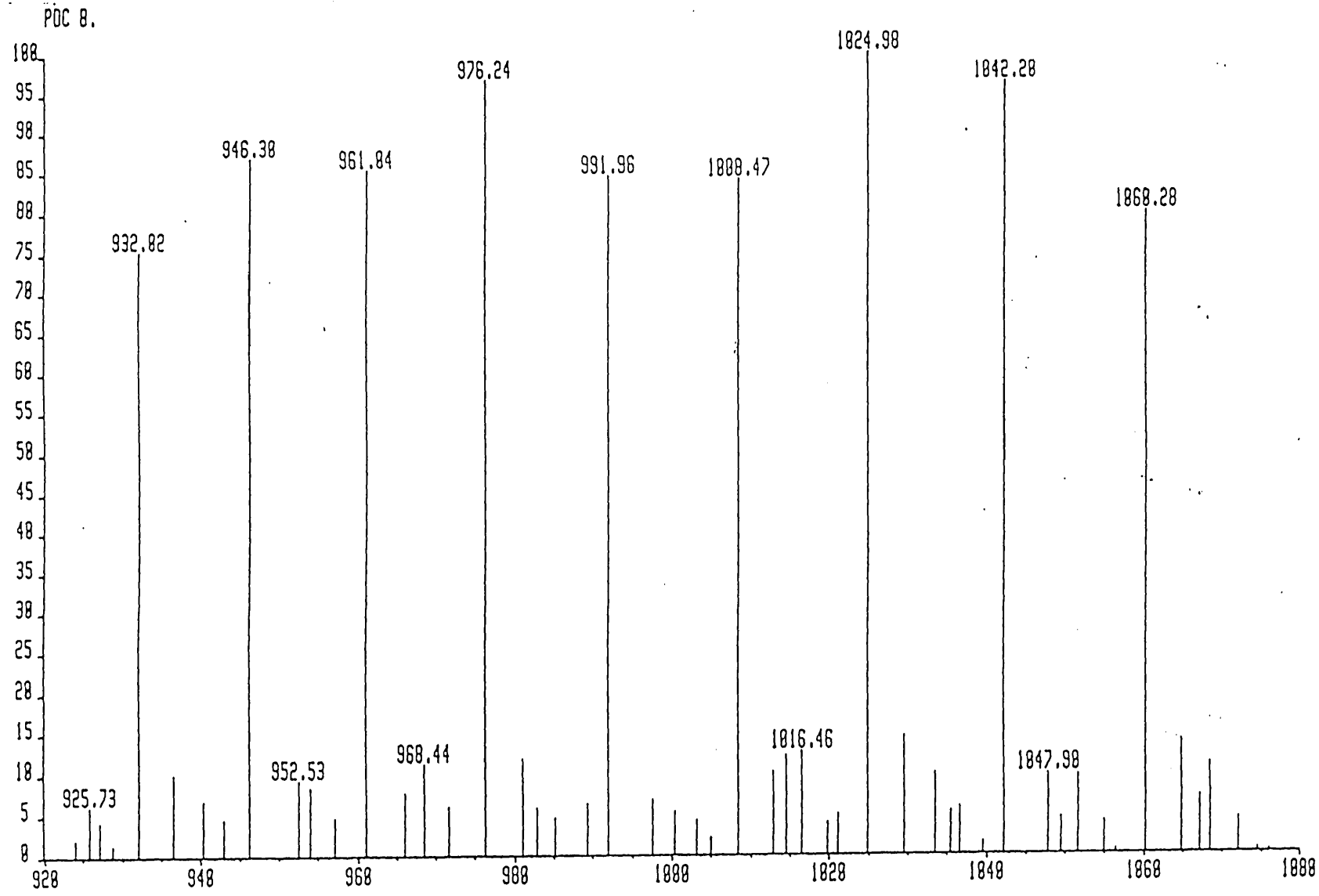


Figure 5.3.6 ESI-MS Containing the Centroids of each Molecular Ion Envelope for PDC B Observed in Fig. 5.3.3.

Molecular weight determination for subunit PDC A1

no. of charges, n	mass, m	mol. wt.
66	931.76	61,429.64
65	946.50	61,456.99
64	961.43	61,467.01
63	976.57	61,460.41
62	991.76	61,426.63
61	1008.27	61,442.99
60	1024.71	61,422.13
59	1042.33	61,438.00
58	1060.15	61,430.24

mean mol.wt. = 61,441.56
standard deviation = ±16.3

Molecular weight determination for subunit PDC A2

no. of charges, n	mass, m	mol.wt.
66	931.76	61,429.64
65	945.84	61,414.09
64	960.58	61,412.61
63	975.91	61,418.83
62	991.70	61,422.91
61	1007.88	61,419.20
60	1024.65	61,418.53
59	1042.14	61,426.79
58	1059.82	61,411.10

mean mol. wt. = 61,419.30
standard deviation = ±6.3

Molecular weight determination for subunit PDC B

no. of charges, n	mass, m	mol.wt.
66	932.02	61,446.80
65	946.30	61,443.99
64	961.04	61,442.05
63	976.24	61,439.62
62	991.96	61,439.03
61	1008.47	61,455.19
60	1024.98	61,438.33
59	1042.20	61,430.33
58	1060.28	61,437.78

mean mol.wt. = 61,441.46
standard deviation = ±6.9

Table 5.3 Molecular Weight Calculations Determined from Electrospray Ionisation Mass Spectra.

Table 5.3 contains the molecular weight determinations for each subunit A1, A2 and B respectively. The calculated molecular weights of each subunit determined from the ESI-MS data are;

A1 = 61,441.56 ±16.3 Da

A2 = 61,419.30 ±6.3 Da

B = 61,441.46 ±6.9 Da

These data indicate that A1 is indistinguishable from B and that A2 is 20-40 Da lower in mass than A1 and B. From the above molecular weights of each subunit it is inferred that there are in fact only two subunits. PDC B isolated in the enzyme preparation studied is suggested to be identical to PDC A1 but chromatographically separable due to the deamidation of certain residues and hence, is more negatively charged than PDC A1 and thus elutes from the ion-exchange column later. Thus, PDC B can probably be explained as a preparation artifact. Furthermore, the observation that PDC B elutes from the gel filtration column later than PDC A1 can be explained in terms of its increased acidity, on deamidation of certain residues, as it has been found that acidic proteins tend to be retarded slightly longer than one would expect. These accurate molecular weights conclusively illustrate that one subunit is not formed by limited proteolysis of the other, as the mass difference in molecular weights, 20-40 Da, does not correspond to a loss of even a single amino acid residue. This data also substantiates conclusions drawn from the FAB-Mapping experiments and High Mass analysis of CNBr fragments, in sections 5.2.1 and 5.2.3 respectively, in that there are regions of amino acid sequence not

coded for in the published gene sequence. The average mass calculated from this gene sequence is 60,095.83. This value is approximately 1345 Da lower in mass than the molecular weights of A1 and B and 1323 Da lower than that of A2 determined by ESI-MS.

The ESI-mass spectrometer was initially callibrated with myoglobin, average molecular weight 16,950.7 Da. The molecular weight determined from an ESI spectrum is 16,950.58 ±3.3 Da. This demonstrates the accuracy of the technique, with recent applications of ESI to large protein analysis of mass similar to that of PDC e.g. bovine albumin [105] giving data within ±20 daltons.

5.3 Conclusion

The study presented in this chapter has investigated and solved a number of important structural problems contributing towards an understanding of the genetic regulation of yeast glycolysis. The sequence data generated in the studies above prove that PDC A1 is encoded by the PDC1 gene, i.e. PDC1 is the structural gene of pyruvate decarboxylase. The strategies employed for the primary sequence analysis of PDC have been illustrated to be powerful techniques ideal for sequence confirmation, the detection of post-translational modifications and determining regions of sequence errors as well as processing events. The high mass study of CNBr fragments of PDC showed there to be a number of mass signals unable to be correlated with the gene-derived sequence.

From this investigation it can be inferred that there are stretches of amino acid sequence which differ from the published gene-derived sequence and that these numerous primary sequence differences observed here may possibly be explained as a putative strain difference between the MC16 strain (investigated in this thesis research work) and Nasmyth's and Reed's gene library.

The novel application of B/E linked scanning/CAD to direct mixture analysis of unknown components of an intact proteolytic digest of pyruvate decarboxylase has illustrated;
(a) the method of linked scanning/CAD to be highly appropriate for sequence analysis of large biopolymers with the rate-determining step of purification removed, and
(b) the elucidation of a processing event at the N-terminus and a post-translational modification of the pyruvate decarboxylase protein subunits.

In addition, this thesis research work has allowed clarification of the postulated existence of two subunits of pyruvate decarboxylase. There were a variety of proposals reported as to the nature of the differences between these subunits such as;
(i) two different polypeptides, as in the case of haemoglobin with an alpha and beta chain,
(ii) one polypeptide chain is the result of proteolytic cleavage of the other, and
(iii) the subunits forming the tetramer are identical.
However, the data generated here indicates that none of these suggestions are correct. This conclusion was elucidated in two ways

(1) FAB-Mapping, and (2) ESI-MS which allowed for the first time an accurate molecular weight, for each of the subunits to be determined within ± (5-20) Da, (±0.02%). From these molecular weights it was possible to determine subunit A2 as 20-40 Da lower in mass than A1 or B. With the molecular weights of A1 and B indistinguishable; the difference in elution positions from the ion-exchange columns during purification explained by the proposal that subunit B is identical to A1 except that it is partially or totally deamidated, hence more negatively charged.The observation that B eluted later from the gel-filtration column was explained by the observation that acidic proteins tend to be retarded, slightly longer than one would expect, on gel columns. The hypothesis that one subunit is derived from the other by proteolysis can determined as incorrect, from the molecular weight information, as the mass difference between subunit A2 and A1 (and B) is less than a single amino acid residue mass. Sequence differences between the subunits were determined to be virtually non-existant with the exception of one or two signals. Thus, it is concluded that the sequence differences between the subunits may be confined to only one or more amino acid changes. However, an alternative explanation of this mass difference of ±20 Da, between the subunits, is that subunit A1 (and B) still has the Mg^{2+} co-factor (at.wt. 24) or some other small molecule, tightly bound. This small molecule may even be a transition intermediate, for example bound acetaldehyde (44 a.m.u.). This proposal is supported by the observation, in the FAB-Mapping section of this chapter, that a molecule of trifluoroacetic acid was found attached to a peptide derived by CNBr cleavage of PDC. Thus, the subunits of pyruvate decarboxylase are

possibly identical.

In addition an accurate molecular weight provides a rapid means of checking the fidelity of the gene sequence. The molecular weights of the subunits as determined above are approximately 1350 Da greater than that calculated for the gene-derived sequence. It may be therefore that in this strain of PDC, the C-terminus of the protein is extended, or that there is an insertion sequence. This would also correlate with the FAB-Mapping data and other sequence studies presented here from which mass signals were observed but which could not be assigned on to the gene-derived sequence. This would also rationalise the inability to map the C-termini of the protein subunits despite the extensive sequence analysis methods employed here.

The genetic regulation of yeast glycolysis is a complex mechanism and not fully understood. It has been recently determined that there are two very homologous structural genes for pyruvate decarboxylase, PDC1 and PDC5. However, the transcription of PDC5 is seen only when PDC1 is deleted. Thus, proposed areas of study would be directed towards the primary sequence analysis of the protein encoded by the PDC5 gene for a comparison with that encoded by the PDC1 gene. The protocol developed and described in this chapter would be ideal for such a sequence analysis investigation. Another interesting aspect would be a study of sequence homology between the pyruvate decarboxylases isolated from different bacterial strains, but this work must await production of sufficient quantities of enzyme.

CHAPTER SIX

OVERVIEW

One of the most challenging areas of research lies with the structure elucidation of biologically active substances. Proteins and peptides probably represent one of the most difficult examples of structural analyses. In general they are polar, highly complex molecules whose complexity may be further compounded by post-translational modifications. Difficulties of structure elucidation are associated with the quantity available, involatility, combined with relatively high molecular weights.

Traditionally, mass spectrometry was not an obvious method of choice for the study of compounds such as proteins and peptides; it was normally associated with the analysis of low molecular weight volatile substances. However, the successful application of mass spectrometry to the stucture determination of numerous biologically significant compounds over the last thirty years, as discussed in this thesis, has revealed the potential of mass spectrometry so that often it is the preferred method for solving structural problems.

There are numerous methods and strategies by which mass spectrometry may be applied to a problem, however, not all combinations will solve particular problems. Thus, a key feature of

current research in structure elucidation studies is the development and correct implementation of sensible strategies for solving different types of problems. One of the most important aspects of this research work was the selection of the correct combination of methods used in a concerted strategy for the primary sequence analysis of two enzymes of industrial importance, D-xylose isomerase and pyruvate decarboxylase. In both these protein studies the strategy of FAB-Mapping was selected as highly appropriate for the primary sequence analysis. However, it was the determination of the best approach and application of this method which made these studies successful. For example, important decisions lay with the choice of digest, purification procedures and if they are compatible with mass spectrometry. Illustrated by both the study of D-xylose isomerase (Chapter Two) and pyruvate decarboxylase (Chapter Five) is the integration of FAB-MS protocols with other strategies encompassing the well established techniques of X-ray crystallography and DNA technologies. Furthermore, presented in Chapter Two is the novel application of mass spectrometric data for interpreting the otherwise uninterpretable crystallographic data. Similarly, the primary sequence analysis of pyruvate decarboxylase, in Chapter Five, demonstrates, by the discovery of sequence differences between the protein studied and that of the gene-derived sequence, the need for efficient methods allowing independant checks of an amino acid sequence which has been solely derived by DNA sequencing methods.

Realising the need for providing new, realistic methodologies for protein sequence analysis, two chapters in this thesis have been

devoted to such an aim. The key to the success of any devised methodology, which is to be applied to biological samples, is sensitivity. Both the techniques for sequence analysis presented here have been shown to possess this important feature. The partial acid hydrolysis/MS method has been demonstrated to be readily applicable to less than 500 picomoles of material. The novel aspect of this application of partial acid hydrolysis was that for the first time a concentrated mineral acid was used, unlike previous acid hydrolysis studies. Similarly, from the linked scan/CAD study high quality data has been generated from 100 picomoles to 1 nanomole of sample. This method of B/E linked scan/CAD has been applied for the first time to direct peptide mixture analysis and thus, has been demonstrated here not only to be well suited for the sequencing of pure peptides, such as those isolated by HPLC purification, but also has been shown to be ideal for studying components of intact complex digest mixtures. This novel application of linked scanning represents an important extension of two sector mass spectrometry, removing the rate-limiting step of purifcation when studying complex mixtures, and thus, presents an attractive alternative, in certain circumstances, to four sector mass spectrometry. Furthermore, from this linked scan study of two unknown components of an unfractionated tryptic digest mixture of pyruvate decarboxylase, a post-translational modification and truncation of the N-terminus of the protein was recognised and identified, illustrating that mass spectrometry is the method of choice for the analysis of post-translational modifications.

In summary, this thesis presents data illustrating the successful

development and application of two methodologies for peptide sequencing, and the selection and implementation of sensible strategies to address individual structure elucidation problems. With the anticipated introduction of mass spectrometers with design specifications so as to allow the detection and measurement of as little as 10 femtomoles of high mass samples, the combination of this instrumentation with the strategies devised herein promises a revolution for protein and carbohydrate analyses over the next decade.

CHAPTER SEVEN

MATERIALS AND METHODS

7.1 Mass Spectrometry

7.1.1 The Fast Atom Bombardment mass spectra of;

(a) digest mixtures of D-xylose isomerase studied in Chapter Two,

(b) acid hydrolysates and derivatives investigated in Chapter Three, and

(c) digest mixtures of monomer units A1, A2 and B of pyruvate decarboxylase analysed in Chapter Five,

were obtained on a VG ZAB-2HF instrument equiped with an M-Scan fast atom gun (any alternate instrument used is noted in the text). The gun was operated using xenon gas at a beam current of 10μA at 10keV. Mass spectra were recorded up to mass 3200 daltons at full accelerating voltage. Both pure peptides and digest mixtures examined were dissolved in 5% (v/v) acetic acid and 2μl aliquots loaded on to the FAB target coated with a glycerol/thioglycerol matrix. The instrument was scanned in a mass controlled mode with linear scans performed at a scan rate of 200 sec/decade.

7.1.2 The **linked scan** spectra, Chapter Four, and those from the high mass analysis of a cyanogen bromide (CNBr) digest of pyruvate decarboxylase subunit A1 (Chapter Five) were recorded on a VG ZAB-2SE mass spectrometer, with a resolveable mass range up to

15,000 daltons at full accelerating voltage. Samples were dissolved in 5% (v/v) acetic acid and 1-2µl loaded on to a FAB target coated with mono-nitrobenzyl alcohol (MNBA) matrix. Spectra were recorded using a 11250J data system and ionisation was achieved using a high energy caesium gun operating at 30keV. Linked scan spectra were recorded in the MCA mode by summing between 5-10 scans, depending upon the sample life-time, at 10-20 sec/decade scan speed. The optimum collision-gas (helium) pressure in the linked scan/CAD study was found at a pressure reading of 1.5×10^{-7} Torr on the analyser gauge. This pressure corresponded to the reduction of a tuned signal of B12 (m/z 1329) at the collector to one-third of its original height on the oscilloscope. The collision chamber was located in the first field free region, between the ion source and magnetic sector. Synthetic peptides Leu-enkephalin amide, Angiotensin and Renin substrate were purchased from Sigma (London) Chemical Co.

7.1.3 **Electrospray Ionisation Mass Spectrometry** (ESI-MS)

The ESI-MS quadrupole mass spectrometer used was a VG PMA 3000. The PDC subunits A1, A2 and B (less than 1 nanomole) were dissolved in a volatile solvent, e.g. methanol and injected into a capillary column at atmospheric pressure. The solution eluting from the capillary, at a flow-rate of 5µl/min, was passed between an electrode slit into a warm nitrogen atmosphere which is used to aid desolvation and eliminate cluster formation. The multiply charged sample is then directed into a quadrupole working at a pressure of 10^{-6} Torr.

The water used in the following methods was purified using a four-bowl Milli-Q system.

Specific Cleavages of Polypeptide Chains

7.2 ENZYMIC DIGESTIONS

7.2.1 **Trypsin**

Trypsin was purchased from Sigma (London) Chemical Co. The proteins D-xylose isomerase and pyruvate decarboxylase studied in Chapters Two and Five respectively were digested with trypsin. Enzymic digestion of both the proteins was carried out in pH 8.4 ammonium bicarbonate buffer (50mM) at an enzyme:substrate ratio of 1:50 (w/w) for 3h at 37°C. The reaction was terminated by freezing followed by lyophilisation.

7.2.2 **Chymotrypsin**

Chymotrypsin was purchased from Sigma (London) Chemical Co. Chymotrypsin digestion was carried out on a sample of pyruvate decarboxylase, subunit A1 (Chapter Five). The sample was incubated with chymotrypsin, at an enzyme:substrate ratio of 1:50 (w/w) in an ammonium bicarbonate buffer(50mM), pH 8.4, for 2h at 37°C. The reaction was terminated by freezing followed by lyophilisation.

7.2.3 *V8 protease*

V8 protease (from S.aureus) was purchased from ICN Immunobiologicals. V8 protease digestion was carried out on the mixture of peptides generated by trypsin digestion of PDC subunits A1, A2 and B, Chapter Five. Aliquots of each digest mixture (1-2

nanomoles) were dissolved in an ammonium bicarbonate buffer (0.1M), pH ~8, and incubated with the enzyme V8 protease with an enzyme:substrate ratio of 1:20 (w/w) for 6h at 40°C. The reaction was terminated by freezing followed by lyophilisation.

7.2.4 **Carboxypeptidase B**

Tryptic digest mixtures of D-xylose isomerase (Chapter Two) and subunits of pyruvate decarboxylase (Chapter Five) were incubated with carboxypeptidase B with an enzyme:substrate ratio of 1:50 (w/w). The digestion was carried out in an pH 8.4 ammonium bicarbonate buffer (50mM) for 1h at 37°C. The reaction was terminated by freezing followed by lyophilisation.

7.3 CHEMICAL CLEAVAGE METHODS

7.3.1 **Cyanogen Bromide**

Samples of D-xylose isomerase (Chapter Two) and pyruvate decarboxylase subunits (Chapter Five) were chemically cleaved with cyanogen bromide (CNBr). Samples (1-2 nanomoles) of each were dissolved in 70% formic acid with the addition of cyanogen bromide at a CNBr:protein ratio of 3:1 (w/w). This reaction was carried out at room temperature, in the dark, for 5h. The reaction was terminated by the addition of water (5-fold volume of formic acid used to dissolve the sample) followed by freezing and lyophilisation.

7.3.2 **Edman Degradation**

Sequential Edman degradation was carried out on the digest mixtures of D-xylose isomerase and subunits of pyruvate decarboxylase

(Chapters Two and Five respectively), using at each stage, 5% phenylisothiocyanate (PITC, Rathburn Chemicals Ltd., Walkerburn, Scotland), in dry pyridine (Sigma (London) Chemical Co.), for 1h, followed by cleavage with trifluroacetic acid (Rathburn Chemicals Ltd., Walkerburn, Scotland) for 15 minutes at 45°C. The reaction mixture was dried under vacuum.

7.3.3 **Partial Acid Hydrolysis**

Synthetic peptides Substance P, alpha-Melanocyte Stimulating hormone (MSH), Neurotensin, Renin substrate (porcine), beta-Endorphin, Calcitonin Gene related peptide (CGRP), human Adrenocorticotropic hormone (ACTH), Growth hormone releasing factor 1-44 (GHRF), Cytochrome C and Alcohol dehydrogenase were purchased from Sigma (London) Chemical Co. Samples (500 picomoles) of each were incubated with 12M hydrochloric acid (HCl, 100ul) for 45 minutes at 65°C. The reaction was terminated by lyophilisation of the acid hydrolysate mixture.

7.4 <u>Chemical Modification of Proteins</u>

7.4.1 **Acetylation of Primary Amine Groups**

Samples (1-5 nanomoles) of peptide acid hydrolysates (Chapter Three) were taken up in one drop of water, to which 200µl of an acetylation mixture was added, stoppered, and left at room temperature for 1 minute, the reaction mixture was dried under vacuum. The acetylation mixture used was composed of a 1:1 (v/v) $(CH_3CO)_2O$ and $(CD_3CO)_2O$ mixture of acetic and deuteroacetic anhydride added to methanol in a ratio of 1:4 (v/v).

7.4.2 **S-Carboxymethylation**

S-carboxymethylation was carried out under nitrogen in tris-(hydroxymethyl)-methylamine (TRIS)/HCl buffer (0.6M), pH 8.5, after initial reduction with 5-fold molar excess (over cysteine residues) of dithiothreitol (DTT, Sigma (London) Chemical Co.) followed by incubation at 37°C for 30 minutes. Iodoacetic acid (Sigma (London) Chemical Co.) was used, with a 20-fold molar excess (over cysteine residues), for alkylation at 37°C for 1h followed by acidification to pH 4 using glacial acetic acid. The S-carboxymethylated sample was then loaded on to a C_{18} sep-pak.

7.5 Methods of Protein and Peptide Mixture Purification

7.5.1 **Sep-pak**

Desalting of S-carboxymethylated protein and peptides was achieved by using a C_{18} reverse-phase Sep-pak (Waters). The Sep-pak was pre-washed using the following sequence of solvents, methanol (5ml), propan-1-ol (5ml) and 5% acetic acid (5ml). The S-carboxymethylated sample was loaded on to the Sep-pak and the protein/peptide eluted off the column with 5% acetic acid (5ml, this fraction is expected to contain the salts) 20% propan-1-ol (in 5% acetic acid, 5ml) and 40% propan-1-ol (in 5% acetic acid, 5ml). Each of these fractions was collected separately and dried down under vacuum.

7.5.2 **High Performance Liquid Chromatography (HPLC)**

A Waters Associate high pressure liquid chromatographic model was used equipped with a Universal Injector (UK66), a model 660

gradient programmer, two solvent delivery systems (Model 510), and two fixed wavelength ultra-violet (u.v.) absorbance detectors (254 and 280 nm). The same reverse-phase C_{18}-Ultrasphere^{TM}-ODS column of dimensions 4.6mm (ID) x 25cm was used throughout this thesis work. The solvents used for peptide elution were, water + 0.1% trifluroacetic acid (TFA, Sigma (London) Chemical Co.), solvent A, and 90% (in water) acetonitrile (Rathburns, Scotland) + 0.1% TFA, solvent B. A flow rate of 1ml/min was employed and fractions were collected at 1 minute intervals. Digest mixtures of D-xylose isomerase, peptide acid hydrolysates and digest mixtures of pyruvate decarboxylase (Chapters Two, Three and Five respectively) were purified by HPLC with the conditions utilised as described with variation only of elution gradient devised for fractionation for each different sample.

7.6 Tables of elution gradients

Solvent A = water + 0.1%TFA; Solvent B = 90% acetonitrile + 0.1%TFA

7.6.1 Elution gradient employed for purification of the CNBr/tryptic digest mixture of D-xylose isomerase (Chapter Two).

	TIME (min)	A (%)	B (%)	Gradient
Isocratic	0	90	10	1
	10	90	10	6
	40	80	20	6
	100	60	40	6

120	40	60	6
130	20	80	6

7.6.2 Elution gradient employed for fractionation of the CNBr digest of thermolytic digest product of D-xylose isomerase, (limited proteolysis experiment, Chapter Two).

	TIME (min)	A (%)	B (%)	Gradient
Isocratic	0	90	10	1
	20	80	20	6
	50	70	30	6
	70	60	40	6
	85	30	70	6
	100	10	90	6

7.6.3 Elution gradient for the isolation of the components of the acid hydrolysate of Neurotensin (Chapter Three).

	TIME (min)	A (%)	B (%)	Gradient
Isocratic	0	100	0	1
	10	90	10	6
	15	75	25	6
	95	30	70	6
	100	10	90	6

7.6.4 Elution gradient employed for the purification of the chymotryptic digest mixture of PDCA1 (Chapter Five).

	TIME (min)	A (%)	B (%)	Gradient
Isocratic	0	100	0	1
	15	90	10	6
	35	80	20	6
	80	60	40	6
	90	40	60	6
	100	0	100	6

7.6.5 Elution gradient used for the purification of CNBr digests of subunits A1, A2 and B of pyruvate decarboxylase (Chapter Five).

	TIME (min)	A (%)	B (%)	Gradient
Isocratic	0	90	10	1
	15	80	20	6
	75	55	45	6
	85	45	55	6
	95	35	65	6
	115	0	100	6

REFERENCES

1. B.E. Vogel, R.R. Minor, M. Freund and D.J. Prockop, (1987) J.Biol.Chem., 262, 14737
2. P. Edman, (1950) Acta.Chem.Scand., 4, 283
3. P. Edman and G.Begg, (1967) Eur.J.Biochem., 1, 80
4. E. Gross and B. Witkop, (1962) J.Biol.Chem., 237, 1856.
5. C.O. Anderson, (1958) Acta.Chem.Scand., 12, 1353
6. K. Biemann, F. Gapp and J. Siebl, (1959) J.Am.Chem.Soc., 81, 2274.
7. H.R. Morris, A.J. Geddes and G.N. Graham, (1969) Biochem.J., 111, 38
8. A.J. Geddes, G.N. Graham, H.R. Morris, F. Lucas, M. Barber and W.A. Wolstenholme, (1969) Biochem.J., 114, 695
9. R.J. Robb, R.M. Kutny, M. Panico, H.R. Morris, W.F. DeGrado and V. Chowdry, (1983) Biochem.Biophys.Res.Commun., 116, 1049.
10. P. Petrelli, P. Pucci, H.R. Morris and F. Addeo, (1986) Biochem.Biophys.Res.Commun., 140, 28.
11. S. Magnusson, L. Sottrup-Jensen, T.E. Petersen, H.R. Morris and A. Dell, (1974) FEBS Lett., 44, 189.
12. H.R. Morris, A. Dell, T.E. Petersen, L. Sottrup-Jensen and S. Magnusson, (1976) Biochem.J., 153, 663.
13. H.R. Morris, G.W. Taylor, P.J. Piper and J.R. Tippins, (1980) Nature, 285, 104.
14. J. Hughes, T.W. Smith, H.W. Kosterlitz, L. Fothergill, B.A. Morgan and H.R. Morris (1975), Nature, 258, 577.
15. D.E.M. Lawson, D.R. Fraser, E. Kodicek, H.R. Morris and D.H. Williams, (1971) Nature, 230, 228

16. H.R. Morris, (1979) Phil.Trans.R.Soc. A, 293, 39.

17. J.D. Priddle, K. Rose and R.E. Offord, (1976) Biochem.J., 157, 777.

18. G. Hudson and K. Biemann, (1976) Biochem.Biophys.Res.Commun., 71, 212.

19. H.G. Khorana, G.E. Gerber, W.C. Herlihy, C.P. Gray, R.J. Anderegg, K. Nihei and K. Biemann (1979) Proc.Natn.Acad.Sci. U.S.A., 76, 5046.

20 H.R. Morris, D.H. Williams, G.C. Midwinter and B.S. Hartley, (1974) Biochem.J., 141, 701

21. W.V. Shaw, L.C. Packman, B.D. Burghleigh, A. Dell, H.R. Morris and B.S. Hartley, (1979) Nature, 282, 870.

22. A.I. Hakamori, (1964) J.Biochem., 55, 205.

23. H.R. Morris, (1972) FEBS Lett., 22, 257

24. H.D. Beckey, (1969) Int.J.Mass Spectrom.Ion Phys., 2, 500.

25. H.R. Schulten, (1981) in "Soft Ionisation:Biological Mass Spectrometry" ed. H.R. Morris, Heyden U.K., p6.

26. M. Barber, R.S. Bordoli, R.D. Sedgwick and A.N. Tyler, (1981) J.Chem.Soc.Chem.Commun., 325.

27. H.R. Morris, M. Panico, M. Barber, R.S. Bordoli, R.D. Sedgwick and A.N. Tyler, (1981) Biochem.Biophys.Res.Commun., 101, 623.

28. D.H. Williams, C.V. Bradley, S. Santikarn and G. Bojesen, (1982) Biochem.J., 201, 105.

29. H.R. Morris, A.Dell, A.E. Banner, S. Evans, R. McDowell and D. Hazelby, (1977) Proceedings of the 25th Conference on Mass Spectrometry and Allied Topics, ASMS, p73.

30. H.R. Morris, A. Dell and R.A. McDowell, (1981) Biomed.Mass. Spectrom., 8, 463.

31. R.D. Macfarlane and D.F. Torgerson, (1976) Science, 191, 920

32. H.R. Morris, (1982) Int.J.Mass Spectrom.Ion Phys., 45, 331.

33. W.R. Gray, L.H. Wojcik and J.H. Futrell, (1970) Biochem.Biophys.Res.Commun., 41, 1111.

34. H.R. Morris, M. Panico, A. Karplus, P.E. Lloyd and B. Riniker, (1982) Nature, 300, 643.

35. H.R. Morris, M. Panico, A.T. Etienne, J. Tippins, S.I. Girgis and I. MacIntyre, (1984) Nature, 308, 746.

36. J.I. Harris, R.D. Cole and N.G. Pon (1956) Biochem.J., 62, 154.

37. S.M. Partridge and H.F. Davis (1950) Nature, 165, 62.

38. D. Piszkiewicz, M. Landon and E.L. Smith, (1970) Biochem.Biophys.Res.Commun., 40, 1173.

39. H.R. Morris, A. Chatterjee, M. Panico, B.N. Green, M.A.A. Almeida da Silva and B.S. Hartley, (1989) Rapid Commun. Mass Spectrom., 3, 110.

40. R.K. Boyd and J.H. Beynon, (1977) Organic Mass Spectrometry, 12, 163.

41. H.R. Morris, M. Panico and G.W. Taylor (1983) Biochem.Biophys.Res.Commun., 117, 299.

42. H.R. Morris, M. Panico and A.T. Etienne, (1983) in Proceedings of the 31st Conference on Mass Spectrometry and Allied Topics, Boston MA, p.683

43. H.R. Morris, G.W. Taylor, M. Panico, A. Dell, A.T. Etienne, R.A. McDowell and M.B. Judkins, (1982) in "Methods in Protein Sequence Analysis", ed. M. Elzinga, Humana Press, New Jersey, p.243

44. B.W. Gibson, L. Poulter, D.H. Williams and J.E. Maggio, (1986) J.Biol.Chem., 261, 5341.

45. S. Lemaire, L. Chouinard, A. Denis, M. Panico and H.R. Morris, (1982) Biochem.Biophys.Res.Commun., 105, 51.

46. T.J. Knott, S.C. Rall, Jr., T.L. Innerarity, S.F. Jacobson, M.S Urdea, B. Levy-Wilson, L.M. Powell, R.J. Pease, R. Eddy, H. Nakai, M. Beyers, L.M. Priestly, E. Robertson, L.B. Rall, C. Betsholtz, T.B. Shows, R.W. Mahley and J. Scott, (1985) Science, 230, 37.

47. T.J. Knott, R.J. Pease, L.M. Powell, S.C. Wallis, S.C. Rall, Jr., T.L. Innerarity, B. Blackhart, W.H. Taylor, Y. Marcel, R. Milne, D. Johnson, M. Fuller, A.J. Luis, B.J. McCarthy, R.W. Mahley, B. Levy-Wilson and J. Scott, (1986) Nature, 323, 734.

48. C.A. Smith, (1980) PhD Thesis, University of London.

49. A.D. Gounaris, I. Turkenkopf, L.L. Civerchia and J. Greenlie, (1975) Biochem.Biophys.Acta, 405, 492.

50. H. Zehender and J. Ullrich, (1985) FEBS Lett., 180, 51.

51. K. Weber and M. Osborn, (1969) J.Biol.Chem., 244, 4406.

52. A. Dell and H.R. Morris, (1982) Biochem.Biophys.Res.Commun., 106, 1456.

53. B. Sundqvist, P. Roepstorff, J. Fohlman, A. Hedin, P.Hakansson, I. Kamensky, M. Lindberg, M. Salehpour, G. Sawe, (1984) Science, 226, 696.

54. M. Dole, L.L. Mack, R.L. Hines, R.C. Mobley, L.D. Ferguson and M.B Alice, (1968) J.Chem.Phys., 49, 2240.

55. M. Yamashita and J.B. Fenn, (1984) J.Phys.Chem., 88, 4451.

56. C.M. Whitehouse, R.N. Dreyer, M. Yamashita and J.B. Fenn, (1985) Anal.Chem., 57, 675.

57. G. Lawson and J.F.J. Todd, (1972) Chemistry in Britain 8, 373.

58. J.F.J. Todd and G. Lawson, (1975) in MTP International Review

of Science, Physical Chemistry Two, vol. 5, ed. A. Maccoll, Butterworths, London, p.289.

59. J.E. Campana, (1980) Int.J.Mass Spec. Ion Phys., 33, 101

60. W.P. Chen, (1980) Process Biochem., 8/9, 30.

61. W.P. Chen, (1980) Process Biochem., 8/9, 36.

62. P.P. Ueng, K.J. Volpp, J.U. Tucker, C.S. Gong and L.F. Chen, (1985) Biotechnol.Letters, 7, 153.

63. M. Wilhelm and C.P. Hollenberg, (1984) EMBO J., 3, 2555.

64. V.B. Lawlis, M.S. Dennis, E.Y. Chen, D.H. Smith and D.J. Henner (1984) Appl.Envir.Microbiol., 47, 15.

65. J.J. Huang and N.W.Y. Ho, (1985) Biochem.Biophys.Rep., 126, 1154.

66. K. Hendrick, C.A. Colleyer and D.M. Blow, (1989) J.Mol.Biol., 208, 129.

67. M. Wilhelm and C.P. Hollenberg, (1985) Nucl.Acids.Res., 13, 5717.

68. G.D. Schellenberg, A. Sarthy, A.E. Larson, M.P. Backer, J.W. Crabb, M. Lidstrom, B.D. Hall and C.E. Furlong (1984) J.Biol.Chem., 259, 6826.

69. D. Drocourt, S. Bejar, T. Calmels, J. Reynes and G. Tiraby (1988) Nucl.Acids.Res., 16, 9337.

70. G.C. Saari, A.A. Kumar, G.H. Kawasaki, M.Y. Insley and P.S. O'Hara, (1987) J.Bacteriol., 169, 612.

71. M. Mueckler, C. Caruso, S.A. Baldwin, M. Panico, I. Blench, H.R. Morris, G.E. Lienhard, W.J. Allard and H.F. Lodish (1985) Science, 229, 941.

72. M.G. Wovcha, D.L. Steurwald and K.E. Brooks, (1983) Appl.Envir.Microbiol., 45, 1402.

73. Y. Takasaki, (1967) Agric.Biol.Chem., 31, 309.

74. A.M. Lesk and K.D. Hardman (1985) Methods Enzymol., 115, 381.

75. H. Erdjument, D.A. Lane, M. Panico, V. DiMarzo and H.R. Morris, (1988) J.Biol.Chem., 263, 5589.

76. A.H. Gordon, A.J.P. Martin and R.L.M. Synge, (1941) Biochem.J., 35, 1369.

77. L.M. Keefer and R.A. Bradshaw, (1977) Fed.Proc., 36, 1799.

78. J.E. Walker, A.F. Carne and H.W. Schmitt, (1979) Nature, 278, 653.

79. G.E. Gerber, R.J. Anderegg, W.C. Herlihy, C.P. Gray, K. Biemann and H.G. Khorana, (1979) Proc.Natl.Acad.Sci. U.S.A., 76, 227.

80. F. DeAngelis, M. Botta, S. Ceccarelli and R. Nicoletti, (1986) Biochem.J., 236, 609.

81. H.A. Scoble, S.A. Martin and K. Biemann, (1987) Biochem.J., 245, 621.

82. E. Hunt and H.R. Morris, (1973) Biochem.J., 135, 833.

83. H.R. Morris, (1974) Biochem.Soc.Trans., 2, 806.

84. S. Naylor, A.F. Findeis, B.W. Gibson and D.H. Williams, (1986) J.Am.Chem.Soc., 108, 6359.

85. S. Naylor, G. Moneti and S. Guyan, (1988) Biomed.Environ.Mass Spectrom., 17(5), 393.

86. H.B. Bull and K. Breese, (1974) Arch.Biochem.Biophys., 161, 665.

87. L. Poulter, S-G. Ang, D.H. Williams and P. Cohen, (1987) Biochem.Biophys.Acta, 929(3), 296.

88. H.R. Morris, (1970) PhD Thesis, University of Leeds.

89. R.M. Caprioli, W.t. Moore and T. Fan, (1987) Rapid Commun.Mass Spectrom., 1(1), 15.

90. M. Barber and R.M. Elliot, (1964) in Proc. 12th Ann.Conf. on Mass Spectrometry and Allied Topics, ASMS, Montreal.

91. J.H. Futrell, K. Ryan and L.W. Siek, (1965) J.Chem.Phys., 43, 1832

92. K.R. Jennings, (1965) J.Chem.Phys., 43, 4176.

93. A.F. Weston, K.R. Jennings, S. Evans and R.M. Elliot, (1976) Int.J.Mass Spectrom.Ion Phys., 20, 317

94. A.P. Bruins, K.R. Jennings and S. Evans, (1978) Int.J.Mass Spectrom.Ion Phys., 26, 395.

95. D.S. Millington and J.A. Smith, (1977) Org.Mass Spectrom., 12, 264.

96. J.H. Beynon, R.P. Morgan and A.G. Brenton, (1979) Phil.Trans.R.Soc. London A, 293, 157.

97. F.W. McLafferty, (1980) Accounts of Chemical Research 13, 33.

98. K.B. Tomer, M.L. Gross, H. Zappey, R.H. Fokkens and N.M.M. Nibbering, (1988) Biomed.Environ.Mass Spectrom., 15, 649.

99. E. Kellermann, P.G. Seeboth and C.P. Hollenberg, (1986) Nucl.Acids Res., 14, 8963.

100. H.D. Schmitt and F.K. Zimmermann, (1982) J.Bacteriol., 151, 1146.

101. J. Ullrich and C.H. Freisler, (1977) Hoppe-Seyler's Z.Physiol.Chem., 358, 318.

102. R.F.W. Hopmann, (1980) Eur.J.Biochem., 110, 311.

103. H. Zehender, (1984) Doctoral Thesis, Fak.f.Biologie d.Univ.Freidburg

104. J. Ullrich, (1982) Ann.NY.Acad.Sci., 378, 287.

105. J.A. Loo, H.R. Udseth and R.D. Smith, (1989) Anal.Biochem., 179, 404.