PROTEIN SEQUENCING: PAST AND PRESENT

A short historical account with present aspects illustrated by the structures of aspartate proteinases being an edited version of a lecture given at FEBS Advanced Course No 61, Prague, 6–12 July 1980.

B FOLTMANN

Institute of Biochemical Genetics
University of Copenhagen
Denmark

Introduction

In 1980 it is almost impossible for a young biochemist to imagine that any competent protein chemist could think of protein structure without the peptide bond as fundamental unit. However, when teaching modern protein chemistry it is worthwhile remembering that it was only after the advent of protein sequencing that the peptide nature of proteins was generally accepted as a fact.

Some students may regard such a historical view with scepticism. They think that their world began in the late 'sixties, and behind that are only dark ages of little or no interest at all. In some way it is difficult to argue against such an attitude — you may operate an ultracentrifuge without having heard of Svedberg, and you may carry out an automated amino acid analysis without knowing about all the difficulties the protein chemist had thirty or forty years ago. And yet I feel that our intellectual life becomes the poorer if we do not know about our background.

Early protein chemistry

The term protein was introduced by the Dutch chemist Mulder in 1838, after a suggestion from the Swedish chemist Berzelius. In a letter Berzelius wrote:

> The word protein that I propose to you for the organic oxide of fibrin and albumin, I would wish to derive from *proteios*, because it appears to be the first or principal substance of animal nutrition that plants prepare for the herbivores, and which the latter then furnish to the carnivores.[1]

However, it is interesting that in its origin the word protein did not designate a member of the large class of macromolecules which we today call proteins. Berzelius and Mulder thought that all these compounds were built of the same fundamental unit which they called 'protein', and in 1839 Mulder estimated the composition to be: $C_{40}H_{62}N_{10}O_{12}$.[1]

In the following decades several of the amino acids were identified as components of proteins. Although the amino acids were not yet recognized as the fundamental units of all proteins, Mulder's idea about a repeated unit was abandoned, and only the term protein remained.

Around the turn of the century the work of Emil Fischer deeply influenced our concepts of protein chemistry. In 1906 Fischer summarized his investigations in a lecture held at the German Chemical Society. In the introduction he said:

> Since the proteins participate in one way or another in all chemical processes in the living organism, one may expect highly significant information for biological chemistry from the elucidation of their structures and their transformations.[2]

This statement is still 100 per cent valid, and in the same lecture Fischer introduced the term 'peptides'. He emphasized similarities between synthetic peptides and proteins; from this he advocated the importance of the peptide bond, but he also added:

> . . . simple amide formation is not the only possible mode of linkage in the protein molecule. On the contrary, I consider it to be quite probable that it contains piperazine rings.

In the following years many biochemists were in favour of the diketopiperazine as the fundamental structural unit in proteins because little experimental evidence was found for the existence of peptide bonds in native globular proteins.

Even my old teacher Kaj Linderstrøm-Lang wrote as late as in 1938 that he had to give

> . . . a warning against the conclusion that genuine proteins contain peptide bonds because they are split by proteinases like trypsin. They give a certain indication that peptide bonds are formed or 'appear' (like SH–groups) upon denaturation, but they are not conclusive enough to decide whether or not some hydrolysable peptide bonds are preformed in the molecules of the genuine globular proteins.[3]

Seven years later Sanger started to sequence insulin and in 1952 the sequencing of insulin and other peptide hormones provided increasing evidence for the peptide nature of proteins. Linderstrøm-Lang could then conclude:

> We have now returned, I believe, to Fischer's theory and consider the proteins as being built up by peptide chains. At any rate, that is what I am going to assume to-day.[4]

Linderstrøm-Lang subsequently introduced the terminology: *primary structure* for the covalent (amino acid) sequence, *secondary structure* for the hydrogen bond stabilized structures, and *tertiary structure* for the rest of the folding of the peptide chain leading to the globular structure of the soluble proteins.

Beginning of sequencing

Sequencing of proteins has from the beginning been based on methods for separation of proteins and peptides, and methods for identification and quantification of amino acids, thus it may be of interest to survey when the prototypes for some of our present methods were introduced in protein chemistry.

Table 1 gives a summary of methods for purification and fractionation of peptides and amino acids. Looking back, it is in some way strange that paper chromatography was not used in biochemistry at a much earlier date. Paper chromatography could have been used since filter paper became common in the first half of the nineteenth century. As far as I understand a kind of paper chromatography was also used in the German dyestuff industry in the last half of the nineteenth century, but not in biochemistry.

Table 1
Fractionation of peptides and amino acids

Chromatography:	
Filter paper	Consden, Gordon and Martin (1944)[5]
Granular starch	Synge (1944)[6]
Granular starch (quantitative)	Moore and Stein (1948)[7]
Ion exchange (quantitative)	Moore and Stein (1951)[8]
Ion exchange (quantitative and automatic)	Spackman, Stein and Moore (1958) [9]
Gel filtration	Porath and Flodin (1959)[10]
Electrophoresis:	
High voltage paper	Michl (1952)[11]
Two-dimensional with chromatography (fingerprinting)	Ingram (1958)[12]
Identification of single amino acid residues:	
Acid stable derivatives of *N*-terminal residues:	
FDNB	Sanger (1945)[13]
Dns	Gray and Hartley (1963)[14]
Degradation and identification of liberated residues as PTH amino acids:	Edman (1950)[15]
Automated procedure	Edman and Begg (1967)[16]

Before 1945 a quantitative amino acid analysis was not available for one single protein. The first reliable analyses were obtained by Erwin Brand and co-workers, using amino acid-requiring mutants of acid-producing bacteria; but these methods were soon surpassed by the column chromatographic methods developed by Moore and Stein, in 1949 on cellulose columns and since 1951 on ion exchange resins. Compared to our present automatic amino acid analysers the first chromatographic methods were very time-consuming. To obtain one analysis we had to work hard for one week — and yet it was a very great progress.

Methods for purification of peptides and quantitative amino acid analyses are necessary for protein sequencing, but unless the question concerns very small peptides, a method for identification of at least the *N*-terminal residue is required. In 1945 Sanger solved this problem by reacting the free α-amino group with fluoro-dinitro-benzene (FDNB). This was a breakthrough, but today the method is superseded by fluorescent labelling with dansyl chloride (Dns). Another milestone in protein sequencing was the degradation with phenylisothiocyanate (PITC) and identification of liberated amino acid residues as phenylthiohydantoins (PTH). This method was introduced by Edman in 1950, and since then most protein sequencing has been carried out using the Edman degradation.

The start of protein sequencing was slow. In 1955 Sanger and co-workers[17] completed the structure of insulin, but we had to wait five years for the structure of the first enzyme (ribonuclease).[18] Since then amino acid sequences of proteins have been determined at an ever-increasing rate. In 1965 about twenty proteins with more than 100 amino acid residues had been sequenced. The number has later grown with a doubling time of about three years, and it is estimated that the primary structures of more than 1500 proteins now are known.

Sequencing today

The rapid increase in number of sequenced proteins reflects recent improvements in methodology (see reviews in references [19] and [20]). For quantitative amino acid analysis about 1 μmol of each amino acid was required in the first column chromatographic methods, whereas today most commercial amino acid analysers are·able to work at a nanomol scale and provide a complete analysis in less than two hours. The automated methods for protein sequencing have also been greatly improved. In the latest (non-commercial) version[21] the protein sequenator operates at a 10 pico-mol level, a 10^4-fold increase in sensitivity relative to the method described by Edman and Begg in 1967. The increase in sensitivity depends to a wide extent on identification of PTH amino acids through high performance liquid chromatography; however the manual sequential Edman degradation-dansylation may be carried out with 1–10 nanomol of peptide, and many problems may still be solved with this method. It is also worthwhile mentioning that the dansylation method with subsequent identification of Dns amino acids on thin-layer plates is very suitable for students' experiments.

Above it was pointed out that the vast majority of all protein sequences has been obtained with methods that involve the Edman degradation, but the application of mass-spectrometry has recently made considerable progress. In 1979 the first total protein sequence assignment was made by mass-spectrometry independent of classical methodology.[22] Even if one is not aiming for the primary structure of a complete protein by means of mass-spectrometry, this method offers advantages in the solution of special problems. Mass spectrometry is very useful for identification of postsynthetic modifications that may be lost with the conventional methods. Thus mass-spectrometry played a vital role in the discovery of a new amino acid (γ-carboxyglutamic acid) and its location in the *N*-terminal region of prothrombin.[23]

Conclusions from sequencing

As already emphasized the first fundamental result of protein sequencing was the experimental proof for the peptide theory, but our knowledge about the primary structures has also had a profound influence on our understanding of the relationships between protein structure and function and of the evolution of proteins.

From investigations on the primary structures it soon became clear that proteins which perform similar functions also have similarities in their amino acid sequences. Among such homologous proteins we may distinguish between orthologous proteins that perform the same function in different species, and paralogous proteins that perform different but related functions within one organism.

By comparison of the amino acid sequences of orthologous proteins it is possible to describe the evolutionary history at the molecular level. The amino acid sequence of cytochrome *c* has been determined for several organisms from bacteria and fungi to mammals. From these observations a phylogenetic tree has been constructed, and for the vertebrates this corresponds fairly well to that inferred from fossil records.[24] However, the molecular comparison has also revealed a relationship among animals, insects and plants — a relationship that could never be observed from the traditional sources.

Furthermore, the study of paralogous proteins has contributed to a deeper understanding of some of the molecular mechanisms of the evolution.[25] It is now generally accepted that each group of paralogous proteins arose from a single gene through repeated duplications. Each of the duplicated genes then underwent a separate evolution; through variation and adaptation new specificities evolved while the fundamentally-important parts of the structures were conserved.

A group of homologous proteins that have evolved from a common ancestral gene is called a superfamily, and the available knowledge about the tertiary structures indicate that within a protein superfamily the folding of the peptide chains is even more conservative than the amino acid sequences.

Aspartic proteinases as example of an enzyme superfamily

The known protein structures are classified in about 200 superfamilies, and to illustrate investigations on evolution and structure-function relationships among the proteins in one superfamily I choose examples from my own field of research, the acid or aspartic proteinases (EC 3.4.23). The name acid proteinases indicates that most of these enzymes show optimal proteolytic activity at acid pH, and by analogy with the serine proteinases (EC 3.4.21) the name aspartic proteinases expresses the fact that two aspartic acid residues participate in the catalytic mechanism.

The proteinases from the gastric juice of the vertebrates all belong to this group. Pig pepsin (EC 3.4.23.1) and calf chymosin (EC 3.4.23.4) have been fully sequenced, and from partial sequences of other gastric proteinases combined with immunochemical cross-reactions we are able to deduce a highly probable evolutionary relationship among these enzymes.

Fig 1 shows a schematic comparison between the primary structures of pig pepsin and calf chymosin, the two enzymes are clearly homologous with a degree of identity of about 50 per cent. When tested with antisera raised in rabbits, pepsin and chymosin show no immunochemical cross-reaction, whereas partial immunological identity has been observed among pepsins from man, pig, cattle, dog, cat, and horse.[26] Corresponding to this, fragmentary information about sequences of pepsins from man, cattle and horse indicates that the pepsins from different species have about 75 to 85 per cent identity. Immunological relationships have also been observed among calf

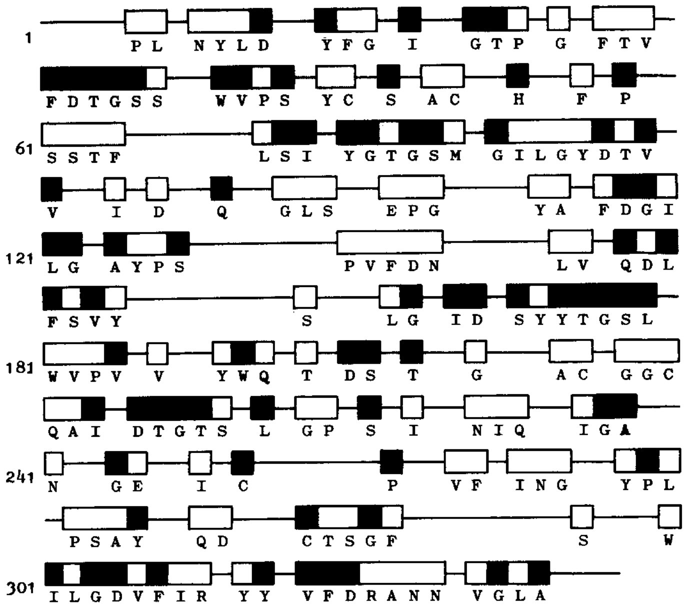
1
P L   N Y L D   Y F G   I   G T P   G   F T V
F D T G S S   W V P S   Y C   S   A C   H   F   P
61
S S T F   L S I   Y G T G S M   G I L G Y D T V
V   I   D   Q   G L S   E P G   Y A   F D G I
121
L G   A Y P S   P V F D N   L V   Q D L
F S V Y   S   L G   I D   S Y Y T G S L
181
W V P V   V   Y W Q   T   D S   T   G   A C   G G C
Q A I   D T G T S   L   G P   S   I   N I Q   I G A
241
N   G E   I   C   P   V F   I N G   Y P L
P S A Y   Q D   C T S G F   S   W
301
I L G D V F I R   Y Y   V F D R A N N   V G L A

Figure 1 *Schematic comparison among the primary structures of pig pepsin, calf chymosin and penicillopepsin.*[27] *Each line represents thirty amino acid residues. The boxes indicate identities between the two gastric proteinases, and the black areas are identities among all three enzymes. Common residues are indicated in the single letter code according to the recommendations of IUB–IUPAC. Numbering starts from the* N*-terminus of pig pepsin.*

chymosin and neonatal proteinases from the other species with the exception of man. The investigations have further demonstrated the presence of a third group of mammalian gastric proteinases (pepsin C or gastricsin, EC 3.4.23.3) immunologically different from both pepsin and chymosin, and comparison of amino acid sequences of pepsin, gastricsin and chymosin from cattle indicate that these proteinases have about 50 per cent of common residues. This means that from a structural point of view cattle pepsin is more related to, eg pig pepsin than to the other gastric proteinases from cattle itself.[27]

The relationship among the individual groups of mammalian gastric proteinases and those from other vertebrates are not yet elucidated, but they are obviously all derived from a common ancestor. All the gastric proteinases are secreted as inactive precursors that are converted into active enzymes through a limited proteolysis which removes the *N*-terminal part of the peptide chain. This peptide segment plays an important role by keeping the zymogens in inactive conformation, and sequence studies have shown a clear homology in the *N*-terminal parts of the zymogens from cartilaginous fishes to the mammals.[26]

From these observations we are now able to summarize the main features of the evolution of the gastric proteinases: A protozymogen for these proteinases was present before the evolution of vertebrates. This zymogen gave rise to several paralogous proteins and among these we observe at least three types of proteinases that apparently were present before divergence of the mammals.

Sequence studies have further shown that proteinases from this superfamily are present in other organs than the stomach and in other phyla as well. In Fig 1 the black areas represent identities between the two gastric proteinases and penicillopepsin. This is a proteinase that is produced by the fungus ***Penicillium janthinellum***. The identities among penicillopepsin and gastric proteinases amount to about 30 per cent, and these sequence studies lead us to the conclusion that all enzymes in this superfamily are derived from one ancestral gene that was evolved when the fungi and the mammals had a common ancestor. With the present results in mind we may predict that enzymes of this superfamily are present in all eukaryotes.

All this exemplifies how protein sequencing has contributed to a new understanding of the evolution of proteins.

Three-dimensional structure and protein sequencing

The folding of the peptide chain to a well-defined three-dimensional structure is determined by the amino acid sequence. Considerable progress has also been made in prediction of secondary structures from the amino acid sequences. However, prediction of the total tertiary structures appears to be very difficult, and optimistic expectations that both primary and tertiary structures may be solved at a stroke through X-ray crystallography have so far not been fulfilled. Hence we must assume that in the near future the elucidation of complete protein structures will depend on combined information from protein sequencing, X-ray crystallography and eventual computer refinements.

The problems may again be illustrated by examples from the aspartic proteinases. Excellent X-ray crystallographic investigations have been carried out for several years with aspartic proteinases from *Rhizopus* and *Endothia*, but detailed model building with orientation of amino acid side chains has not been possible because the primary structures are not yet determined. Among the other enzymes from this superfamily the most detailed information is available for penicillo-pepsin,[28] and the tertiary structures of pig pepsin and calf chymosin are almost completed.[29,30]

However, in spite of the shortcomings of detailed model building the results from the X-ray crystallography are very interesting when compared with the available information about the primary structures of these enzymes. The crystallographic investigations show that all these enzymes share a common folding of the peptide chain. Fig 2 shows an expanded and schematic illustration of the folding of the peptide chain. The molecule is bilobal with a deep cleft running almost perpendicular to the largest dimension of the molecule and the drawing is made so that we are looking down in the cleft which is opened.

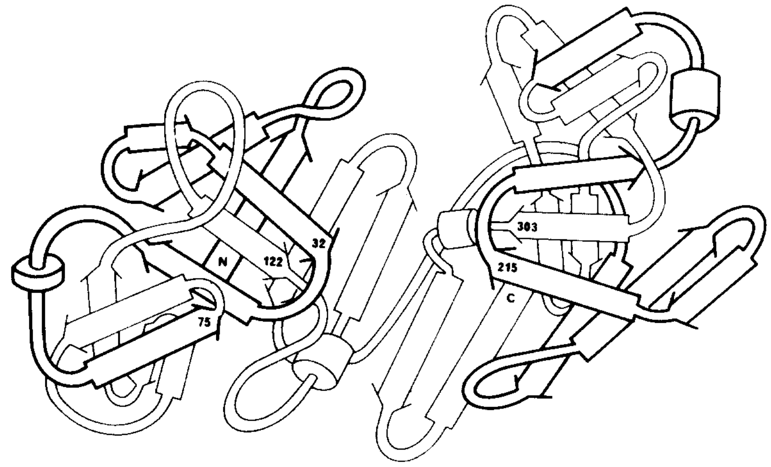

Figure 2 *Expanded and schematic illustration of the folding of the peptide chain of the aspartic proteinases. The arrows indicate parallel or anti-parallel* β*-structures, the cylinders indicate* α*-helices,* N*-terminus and* C*-terminus are marked by N and C, numbers refer to the sequence given in Fig 1. For further details see the text. The figure is drawn from models and data of Blundell* et al[29] *and Andreeva* et al.[30]

If we now compare the folding of the peptide chain with the amino acid sequence shown in Fig 1 important observations are made. Aspartic acid 32 and 215 are in the primary structure surrounded by clusters of residues that are identical in all the aspartic proteinases. In the tertiary structure these residues are located in U-shaped bends pointing towards the cleft, and in the real models Asp 32 and Asp 215 come so close that they are hydrogen-bonded to each other. Two other clusters of invariable residues are found around glycine 122 and 303, located where the peptide chain slips through the two U-shaped bends. A fifth cluster of identities is found around tyrosine 75 which in the real model comes near to the opening of the cleft. The catalytic mechanism has not yet been elucidated in details, but inhibitor studies indicate that Asp 32 and Asp 215 both participate in the catalytic mechanism. In addition to these we observe that many of the residues which have been found invariable by the sequencing also are located in the immediate surroundings of the cleft, so it is a sound guess that several of these residues some way or another participate in the catalytic mechanism.

The observations described here are by no means exceptional; on the contrary we may regard these results as typical for the way protein sequencing and X-ray crystallography support each other.

Postsynthetic modifications

With the recent progress in rapid nucleotide sequencing the question may be raised if the same results of primary structures could have been gained by nucleotide sequencing? — To some extent they probably could, but certain reservations have to be considered.

It is now generally accepted that most of the secreted proteins are synthesized with a prepart or signal peptide which enables the peptide chain to penetrate the membrane of the endoplasmic reticulum. The signal peptide is cleaved off by limited proteolysis, but from nucleotide sequencing we cannot predict where the signal peptide ends and where the normal protein starts. Limited proteolysis further plays an important role in formation of biologically-active proteins from inactive precursors like pro-hormones and zymogens. The activations of pepsinogen or prochymosin are again typical examples of such reactions, and through amino acid sequencing the pathways of the reactions have been elucidated.[31]

Further we must remember that all the postsynthetic modifications of amino acid side chains cannot be deducted from nucleotide sequencing. γ-carboxylation has been mentioned; phosphorylation, acetylation, glucosylation, or methylation are other examples of modifications that need protein chemistry in the structural analyses.

The future

This review has covered a wide field from Mulder's introduction of the term protein to examples of modern protein chemistry. Is this then 'the protein chemistry that was'? What will happen tomorrow? The answer may be found in the last part of this paper. In the future protein sequencing will not be the unique tool it was fifteen years ago, but there are many problems that only can be solved through sequencing proteins. In collaboration with immunochemistry, X-ray crystallography and nucleotide sequencing, protein chemistry and protein sequencing will still maintain its importance in the years to come.

References

[1] Quoted from Fruton, J S (1972) 'Molecules and Life', Wiley-Interscience, New York, NY
[2] Fischer, E (1906) *Ber Deut Chem Ges* **39**, 530–610
[3] Linderstrøm-Lang, K, Hotchkiss, R D and Johansen, G (1938) *Nature (London)* **142**, 996
[4] Linderstrøm-Lang, K (1952) *Lane Medical Lectures, Stanford Univ Publ Med Series* **6**, 1–115
[5] Consden, R, Gordon, A H and Martin, A J P (1944) *Biochem J* **38**, 224–232
[6] Synge, R L M (1944) *Biochem J* **38**, 285–294
[7] Moore, S and Stein, W H (1948) *Ann N Y Sci* **49**, 265–278
[8] Moore, S and Stein W H (1951) *J Biol Chem* **192**, 663–681
[9] Spackman, D H, Stein W H and Moore, S (1958) *Analyt Chem* **30**, 1190–1206
[10] Porath, J and Flodin, P (1959) *Nature (London)* **183**, 1657–1659
[11] Michl, H (1952) *Mh Chem* **83**, 737
[12] Ingram, V M (1958) *Biochim Biophys Acta* **28**, 539–545
[13] Sanger, F (1945) *Biochem J* **39**, 507–515
[14] Gray, W R and Hartley, B S (1963) *Biochem J* **89**, 59P
[15] Edman, P (1950) *Acta Chem Scand* **4**, 283–293
[16] Edman, P and Begg, G (1967) *Eur J Biochem* **1**, 80–91
[17] Ryle, A P, Sanger, F, Smith, L F, and Kitai, R (1955) *Biochem J* **60**, 541–556
[18] Hirs, C H W, Moore, S and Stein, W H (1960) *J Biol Chem* **235**, 633–647
[19] Hirs, C H W and Timasheff, S N (1977) (eds) 'Methods in Enzymology', **47**, Academic Press, New York
[20] Perham, R N (1975) (ed) 'Instrumentation in Amino Acid Sequence Analysis', Academic Press, London
[21] Hunkapiller, M W and Hood L E (1980) *Science* **207**, 523–525
[22] Morris, H R (1979) *Phil Trans Roy Soc London* **A**, **293**, 39–51
[23] Morris, H R, Dell, A, Petersen, T E, Sottrup-Jensen, L, and Magnusson, S (1976) *Biochem J* **153**, 663–679
[24] Dayhoff, M O (1978) (ed) 'Atlas of Protein Sequence and Structure' Vol 5, Suppl 3, Nat Biomed Res Found, Washington DC
[25] Hartley, B S (1979) *Proc Roy Soc London* **B**, **205**, 443–452
[26] Foltmann, B and Axelsen, N (1980) in 'Enzyme Regulation and Mechanism of Action', *FEBS Proceedings* **60**, 271–280
[27] Foltmann, B and Pedersen, V B (1977) *Adv Exp Med Biol* **95**, 3–22
[28] Hsu, I-N, Delbaere, L T J, James, M N J, and Hofmann, T (1977) *Nature (London)* **266**, 140–145
[29] Blundell, T L, Jones, H B, Khan, G. Taylor, G. Sewell, B T, Pearl, L H, and Wood, S P (1980) in 'Enzyme Regulation and Mechanism of Action', *FEBS Proceedings* **60**, 281–288
[30] Andreeva, N S and Gustchina, A E (1979) *Biochem Biophys Res Comm* **87**, 32–42
[31] Pedersen, V B, Christensen, K A and Foltmann, B (1979) *Eur J Biochem* **94**, 573–580

SIMPLE VISUAL DEMONSTRATIONS OF THE CATALYTIC ACTIVITY OF IMMOBILIZED CELLS AND ENZYMES

PETER S J CHEETHAM and CHRISTOPHER BUCKE

Tate & Lyle Ltd
Philip Lyle Memorial Research Laboratory
Reading, Berkshire, UK

Introduction

Recently it has become apparent that biotechnology has great economic potential. Much of the work carried out in this field uses various forms of immobilized enzymes as catalysts because of the great variety of reactions they catalyse, their high catalytic activities and stereospecificities, and the mild conditions under which they operate. However, before industrial application can be made enzymes or cells possessing enzyme activities must usually be immobilized so as to ensure their easy recovery after reaction has been completed. Then the immobilized biocatalyst may be employed in continuous reactors so allowing re-use of the enzyme(s), and preventing contamination of the product by the enzyme or cells.

Immobilization may be defined as any technique which severely limits the free diffusion of the cells or enzyme molecules. Methods of immobilization include covalent binding and adsorption to solid supports, entrapment or encapsulation in solid supports; or aggregation of the cells or enzymes.[1,2] Techniques for the immobilization of cells and enzymes to solid supports are of particular interest to biochemists wishing to study enzyme kinetics and the tertiary structure of proteins,[1] and are acquiring analytical applications especially in the form of enzyme electrodes.[1] However, methods of immobilization attract interest chiefly from biochemists, microbiologists and chemical engineers interested in the commercial application of enzymes.