MICROSEQUENCE ANALYSIS OF BIOLOGICALLY
ACTIVE PROTEINS AND PEPTIDES

Hugh D. Niall

Howard Florey Institute of Experimental Physiology & Medicine,
University of Melbourne, Parkville, Victoria, Australia 3052

Abstract - Despite impressive advances in techniques for sequence
analysis of nucleic acids, direct sequencing of proteins will continue
to be required in many areas of biological research.   Traditional
approaches are adequate, though still laborious,when carried out on
micromole $(10^{-6}M)$ amounts of material.   Recent advances employing
radioisotopic and carrier methodology allow sequencing in the nanomole
$(10^{-9}M)$ range.   A further increase in sensitivity to the level of 10^{-12}
- $10^{-15}M$ is required for certain problems to be solved (e.g. the
structure of polypeptide hormone receptors and other cell surface
membrane-bound proteins).   A biosynthetic approach, which will
incorporate ^{14}C or ^3H labelled amino acids directly into the protein
to be studied, seems the most promising line to follow, when coupled
with high efficiency isolation techniques which take advantage of the
stereospecificity of biological macromolecules, i.e. affinity
chromatography.

INTRODUCTION

Only a few years ago the determination of amino acid sequences of polypeptides repres-
ented the only practical approach to problems requiring decoding of genetically regulated
informational macromolecules.   Thus sequence analyses of numerous globins, cytochrome
C molecules, immunoglobulins, peptide hormones and enzymes provided insight into a
variety of genetic and cellular regulating systems as well as indicating the main lines of
molecular evolution.   At that time techniques for sequencing DNA were primitive, so that
mutations in the germ cell line had to be inferred from alterations found in the polypeptide
products of the genes under study.   Degeneracy in the genetic code necessarily introduces
a minor element of uncertainty into this kind of analysis since at least some mutational
events would not, in fact, result in a different polypeptide.   With the advent of modern
techniques for DNA sequencing, reviewed by Sanger at this meeting (Ref. 1) the tables have
almost been turned, and in some situations it may now prove easier to sequence the gene
rather than its polypeptide product.   However, there is little doubt that, these advances in
nucleic acid chemistry notwithstanding, direct microsequencing of biologically active
proteins and peptides will continue to be required on a major scale for a long time to come.

Successive editions of the "Atlas of Protein Sequences " edited by Margaret O. Dayhoff
(Ref. 2) attest to the rapid growth in our knowledge of protein structure.   Almost all the
proteins studied to date, however, have been those quite readily purified in relatively
abundant quantities (tens or hundreds of micromoles).   Thus traditional sequencing pro-
cedures developed in the 1950's and 1960's have been applied.   However, a whole series of
new problems now require much more sensitive sequencing techniques, in conjunction with
scaled-down isolation procedures.   The current interest in the structure and properties of
the membrane-bound proteins found at cell surfaces is to a great extent responsible for this.
A number of other polypeptides with hormonal or regulatory properties can only be
isolated in trace quantities.

A simple example gives some idea of the magnitude of the problem.   Cell surface receptors
for drugs and for peptide hormones represent a class of protein of great physiological
interest and potential therapeutic importance.   We know almost nothing of their chemistry

other than that they are of high molecular weight (at least 150,000 daltons), may consist of subunits, and probably have a carbohydrate component. Kinetic experiments using radio-actively labelled hormones reveal that these are of the order of only a few thousand horm-one receptor molecules per cell (Ref.3). Receptor protein represents 1 part by weight in 10^7 of the total membrane protein. Thus a ten millionfold purification is required. Tsushima and coworkers (Ref.4) report an overall yield of the growth hormone receptor isolated from pregnant rabbit liver of about 2 nanograms per 100 gms of tissue - i.e. fem-tomole (10^{-15} moles) amounts. Clearly extremely sensitive sequencing techniques of a kind not yet developed will be required to cope with this and similar problems.

APPROACHES TO MICROSEQUENCE ANALYSIS

Mass spectrometry
Initial studies were carried out in 1958 by Anderson (Ref.5) who first used mass spectro-metry for the analysis of small peptides after trifluoroacetylation and esterification to render them volatile. Stenhagen (Ref.6) has suggested a similar approach. Even less polar and more volatile derivatives were prepared by Biemann and coworkers (Ref.7) who reduced the carbonyl groups of the peptide backbone to the corresponding polyamino-alcohols. Das et al (Ref.8) used a permethylation technique. All of these manoeuvers are directed towards rendering non-polar the many polar functional groups of peptides, since these restrict volatility. This problem limits the direct application of mass spectro-metry in microsequencing, since, even with derivatization, peptides bigger than 15 - 20 residues usually cannot be analysed. The need for derivatization also introduces handling losses and restricts the sensitivity of the approach. A further problem is that peptides isolated from biological sources are often contaminated by trace amounts of products derived from chromatographic media. These introduce spurious peaks into the mass spectrometry analysis. For all these reasons mass spectrometry , despite almost two decades of effort by expert groups, has never really become established as a useful practical technique for peptide sequencing. Notable exceptions include the elegant analysis of gastrin by Kenner (Ref.9) and of the hypothalamic hormone releasing peptides LHRH and TRH.

A recent approach of promise combines enzymatic and chemical methods with mass spec-trometry (Refs. 10, 11) as summarized in Table 1. Peptides are reduced to 2 sets of dipeptides representing all possible dipeptide combinations which can be generated from the parent peptide. Initially arginyl and prolyl residues gave trouble as they may be resis-tant to the aminodipeptidases used. However, newer enzymes have now been found which can split bonds involving these residues. Ambiguities obviously can arise where there are repeating sequences, and often additional experiments must be done. Nevertheless, Krutzsch and Pisano (Ref.11) have convincingly demonstrated the practicality of the tech-nique at the level of a few nanomoles, on a variety of shorter peptides.

TABLE 1.   Sequence analysis by mass spectrometry

e.g.   A B C D E F G H I J K L M N O P

1.   Treat with aminodipeptidase

⟶ set of dipeptides (AB, CD, etc)

2.   Remove terminal amino acid by Edman degradation

⟶ B C D E F G H I J K L M N O P

Generate second set of dipeptides with aminodipeptidase (BC, DE, etc)

3.   Derivatize each set of dipeptides

(polyaminoalcohols, trimethylsilylation)

4.   GC — MS — Computer

Clearly at present mass spectrometry represents at best a partial answer to the problem of microsequencing, since only relatively short peptides can be handled and one does not escape the need for protein fragmentation and isolation. Thus the technique remains expensive in time, labour and, not least important, the cost of the sophisticated equipment required. In the long term, however, it may prove the method of choice, since it has the potential of analysing at high sensitivity complex peptide mixtures without necessarily isolating individual peptides. It is already the best way of analysing smaller peptides with unusual structural features (e.g. cyclic peptides, those containing D-amino acids) which resist standard chemical or enzymatic techniques.

COMPOSITIONAL APPROACH

Insulin was originally sequenced in Sanger's classical work (Ref. 12) by a process of generating a large number of different, overlapping peptide subfragments using different cleavage procedures. From the amino acid composition of these fragments and from end group analyses (by the DNP method) the structure could be put together. The development of a good sequential chemical method by Edman (Ref. 13) soon rendered the compositional method obsolete. However, it is worth remembering that techniques for end group analysis by the ^{14}C dansyl, ^{35}S phenylisothiocyanate or sulphoindonyl techniques (Refs. 14-16) now require only picomole (10^{-12}M) quantities. Amino acid composition of peptides can also be obtained at the picomole level using the double isotope dansyl chloride technique (Ref. 14). Thus the sensitivity of this "old-fashioned" sequencing approach is now limited only by the ability to isolate peptides at the picomole level. Thin layer mapping of ^{14}C dansylated peptide digests represents one possible avenue which is currently being explored in our laboratory.

SEQUENTIAL APPROACH

Sequential methods depend upon the application of a repetitive process which splits amino acids one by one from the N or C terminal end of a peptide chain. The amino acids are then identified either as such or as chemical derivatives. Obviously this approach fails in the special case of cyclic peptides like valinomycin and when the terminal residue has an attached substituent blocking the degradation (e.g. N terminal acetyl, formyl or pyroglutamyl groups ; C terminal amide). The repetitive process can be chemical or enzymatic, but in practice the sequential enzymatic approach is limited. This is because the process of amino acid release by amino- or carboxy-peptidases is continuous, and the order of release is soon obscured as different molecules are digested at differing rates.

Chemical degradations are available for both N- and C-terminal approaches. Of these the N-terminal method devised by Edman (Ref. 13) is by far the most successful. In fact it has dominated sequencing methodology for at least 20 years and there is so far no sign that it will be soon supplanted.

Edman degradation
The chemistry of this technique is now in every standard text book. However, it must be reviewed briefly here in the context of high sensitivity sequencing. The reagent phenylisothiocyanate reacts with the N-terminal α amino group of the peptide chain to give a phenylthiocarbamyl adduct. Under anhydrous acidic conditions (e.g. trifluoroacetic acid) the proximity of the sulfur atom to the carbonyl component of the first peptide bond allows the ready formation of a heterocyclic derivative of the first amino acid in the chain, with concomitant cleavage of the bond linking it to the next residue. The derivative formed is an anilinothiazolinone. This is extracted away from the residual peptide in an organic solvent. It is then exposed separately to aqueous acidic conditions (1N HCl, 10 minutes, 80°C) when it rearranges to form a more stable derivative, a phenylthiohydantoin (PTH), which is then identified by a chromatographic procedure. The shortened peptide can then be subjected to repetitions of this procedure, with successive removal of the second, third and subsequent derivatives as PTH amino acids. The coupling and cleavage steps under appropriate conditions proceed almost quantitatively, and the technique has been successfully automated (Ref.14). Degradations of up to 50 - 60 cycles are possible on suitable proteins in the automated instrument, termed a "sequenator". Provided enough material is available shorter peptides (up to 30 amino acids) can usually be sequenced completely by this procedure. Clearly if sensitivity of analysis is not a problem, entire proteins can be (and have been) sequenced virtually by this procedure alone, granting that some subfragments will still have to be isolated.

Since here our focus is on microsequencing, let us consider the limitations of the Edman procedure when only trace amounts (say 1 nM) of peptide are available (Table 2). Some imperfections in the degradation (e.g. incomplete coupling or cleavage) tend to occur regardless of the amount of material. Though they limit the length of a degradation, there is no special difficulty when nanomole quantities are being sequenced. Non-specific peptide bond cleavage is in the same category. However, certain side-reactions become particularly important as the quantity of peptide decreases. Competing reactions such as amino-terminal blocking or desulfuration of the phenylthiocarbamyl adduct are due to finite quantities of impurities including oxidants in the reagents and solvents used. In addition there are small mechanical losses of polypeptide during the solvent extraction steps. Finally, if the yield of PTH amino acids obtained is too low, there is insufficient for identification.

TABLE 2.   Limitations of Edman procedure

| Problem | Significance | Severity | |
		1 μM	1 nM
Incomplete coupling or cleavage	Out of phase	+	+
Competing reactions - α amino blocking - desulfuration of $-C = S$	Fall in yield	+	++++
Peptide bond cleavage	Noise	+	+
Mechanical losses	Fall in yield	-	++++
Solvent losses (short peptides only)	Fall in yield	++	++++
PTH amino acid losses	No identification	-	++++

Over the past several years my colleagues and I have developed several approaches which used together overcome these problems, at least to the extent of allowing useful sequence information to be obtained on subnanomole amounts of material. Degradations are carried out in the presence of a poly-amino acid carrier, ideally composed of non-natural amino acids. A copolymer of norleucine and arginine was used for initial evaluation of this approach (Ref. 18). Polyhydroxyproline was also found to be a suitable carrier for degradation of proteins excepting, of course, those derived from collagen. Silver and Hood (Ref. 19) have independently used a similar technique. This approach, which we have termed "protected Edman degradation", relies upon the mass of carrier polypeptide absorbing the impurities and oxidants referred to above, and thus preventing N-terminal blocking of the peptide. The carrier PTH amino acids protect the "unknown" derivatives during handling prior to identification. This is an advantage over the blocked carriers proposed by Silver and Hood, since these cannot themselves be degraded. The increased mass of material in the sequenator cup eliminates any important losses due to mechanical washout.

The limitation imposed by the sensitivity of the identification procedure is readily overcome through the use of ^{35}S-phenylisothiocyanate in the degradation (Ref. 20). The derivatives are identified after two-dimensional thin-layer chromatography. The labelled PTH amino acids are located by autoradiography, eluted, and quantitated by scintillation counting.

Some dozens of polypeptides have been sequenced using a combination of the carrier and radioisotope methods, in amounts ranging from 0.8 to 5 nanomoles. Since there is room for improvement in both the nature of the carrier and in the specific activity of the isotope, further increases in sensitivity should be obtainable.

An alternate approach is feasible when the protein to be sequenced can be obtained in vitro through biosynthesis. Radioactive amino acids (^{14}C and/or ^{3}H) can be incorporated using tissue slice or cell culture techniques or even in-cell-free systems. The protein must then be isolated free of other radiolabelled proteins. This is easier if the protein under study is the major biosynthetic product of the system. At times one can achieve this by appropriate manipulation of the conditions. For example parathyroid hormone is secreted under the stimulus of a falling or low calcium level. Thus parathyroid tissue slices incubated in a low calcium medium made parathyroid hormone in large amounts, thus facilitating the isolation of radiolabelled hormone. Standard isolation techniques (gel filtration, ion exchange chromatography, disc gel electrophoresis) may be adequate, but it is often advantageous to use an immunoprecipitation step employing an antiserum specific for the protein under study.

When the protein is isolated free of other radioactive components, it is mixed with a carrier (virtually any non-radioactive protein will do) and sequenced using automated Edman degradation. In our own studies reported previously (Ref. 20) the biosynthetic precursors to parathyroid hormone were isolated and sequenced. One of these precursors, termed "preproparathyroid hormone ", has only a transient existence in the cell of origin, as it is rapidly cleaved enzymatically. However, much of the sequence has now been completed, using material synthesised in a wheat germ system stimulated by parathyroid hormone-specific messenger RNA. Many laboratories are now using similar approaches for sequencing precursors to other secreted proteins (e. g. peptide hormones, albumin, immunoglobulins). Cell surface proteins such as transplantation antigens are also starting to yield structural information through this microsequencing approach.

The biosynthetic approach to sequence analysis is extremely sensitive, since under favourable circumstances sequences can be obtained on a picomole or less of material.

Unsolved problems.

The above microsequencing approaches based on the sequential Edman degradation are useful and quite sensitive even at their present stage of development. However, there is considerable room for improvement. A good method for extrinsic labelling of polypeptides is desirable when the biosynthetic approach is not applicable. One would like to be able to label all residues in the chain uniformly and to high specific activity. A promising technique which still needs further work is that of Hembree et al (Ref. 21). Proteins are labelled through tritium exchange by microwave activation in the presence of tritium gas in a specially constructed apparatus. Though labelling is not uniform, quite high specific activities have been obtained.

There is also need for a sensitive sequential method of degradation of peptides from the C terminal end. Renewed interest in this approach was recently shown at a specialist meeting (Refs.22 and 23). Finally, none of the methods discussed so far are satisfactory when the peptide chain being degraded is short - in practice less than 20-30 residues. This is because the solvent extractions necessary in sequential degradation lead to heavy losses of peptide. A great deal of work has been devoted to attempts to circumvent this problem by coupling the peptide to be degraded to a solid support (for a review, see Ref. 24). A related approach,which has great attraction, would be to link the peptide to a soluble polymer, thus making it protein-like in its solvent partition properties.

It is also important to emphasize that when biologically active polypeptides are being studied, a major limitation lies in the isolation and bioassay methods. Clearly the bioassay has to be used to monitor the fractionation steps. If it requires limiting amounts of material, then progress will be greatly impeded. Here sensitive methods for bioassay that use the amplification provided by enzymatic processes may be well worth developing, in conjunction with efficient purification by affinity chromatography.

REFERENCES

1.   F. Sanger, Proc. R. Soc. Lond.B. 191, 317 (1975).

2.   M.O. Dayhoff, Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Silver Spring, Maryland (1973).

3.   D. Cowburn and A. Fischman, Peptides : Chemistry, Structure and Biology. Ann Arbor Science, Ann Arbor, Michigan, pp 997-1012 (1975).

4.   T. Tsushima, Y. Imai, N. Sasaki and F. Matsuzaki, Vth International Congress of Endocrinology, Hamburg. Abstracts, p. 248, abstract no. 605 (1976).

5.   C.O. Andersson, Acta Chem.Scand. 12, 1353 (1958).

6.   E. Stenhagen, Fresenius' Z. Anal. Chem. 181, 462 (1961).

7.   K. Biemann, F. Gapp and J. Seibl, J. Am. Chem. Soc. 81, 2274 (1959).

8.   B. Das, S.D. Gero and E. Lederer, Biochem. Biophys. Res. Commun. 29, 211 (1967).

9.   G.W. Kenner and R.C. Sheppard, Frontiers in Gastrointestinal Hormone Research, Almquist and Wiksell, Stockholm (1973).

10. H. Nau, J.A. Kelley, H-J Forster and K. Biemann, Biochemistry, submitted (1975).

11. H.C. Krutzsch and J.J. Pisano, Peptides : Chemistry, Structure and Biology, Ann Arbor Science, Ann Arbor, Michigan, pp 985-990 (1975).

12. A.P. Ryle, F. Sanger, L.F. Smith and R. Kitai, Biochem.J. 60, 541 (1955).

13. P. Edman, Acta Chem. Scand. 4, 283 (1950).

14. J.P. Brown and R.N. Perham, Eur. J. Biochem. 39, 69 (1973).

15. J.W. Jacobs and H.D. Niall, J. Biol. Chem. 250, 3629 (1975).

16. C.P. Ivanov and Y. Vladovska-Yukhnovska, Biochim Biophys. Acta 194, 345 (1969).

17. P. Edman and G. Begg, Eur. J. Biochem. 1, 80 (1967).

18. H.D. Niall, J.W. Jacobs, J. van Rietschoten and G.W. Tregear, FEBS Letters 41, 62 (1974).

19. J. Silver and L. Hood, Anal. Biochem.60, 285 (1974).

20. H.D. Niall, Peptides : Chemistry Structure and Biology, Ann Arbor Science, Ann Arbor, Michigan, pp 975-984 (1975).

21. W.C. Hembree, R.E. Ehrenkaufer, S. Lieberman and A.P. Wolff, J. Biol. Chem. 248, 5532 (1973).

22. A. Darbre and M. Rangarajan, Solid Phase Methods in Protein Sequence Analysis, Pierce Chemical Company, Illinois, pp 131-138 (1975).

23. G.E. Tarr, ibid, pp 139-148.

24. R.A. Laursen, ibid, pp 3-10.