Application of Automated Sequence Analysis to the Understanding of Protein Structure and Function*

ROBERT L. HEINRIKSON, Ph.D.

Department of Biochemistry,
The University of Chicago,
Chicago, IL 60637

ABSTRACT

Application of automated Edman degradation to the covalent structural analysis of proteins has had an important impact upon the generation of structural information. The strategies involved in determining the complete sequences of proteins have been simplified such that studies of this kind can be carried out more rapidly and with less expenditure of material. Moreover, the high efficiency of the procedures in the analysis of intact polypeptides has allowed the convenient comparison of extensive regions of sequence at the amino-termini of proteins. Such structural studies have revealed a more detailed understanding of the function and post-translational modification of many proteins.

The study of structure-function relationships in biological macromolecules has become a major preoccupation of contemporary molecular biology. Although the functions of these substances have been understood, at least at a rudimentary level, for many years, it has not been until recently that details with regard to molecular architecture have come to light. These insights concerning the structure of macromolecules such as proteins and nucleic acids have been made possible by and large through the application of chromatographic and x-ray crystallographic techniques. The latter method has provided the spectacular three dimensional structures of many proteins and these models have, in turn, served as a basis for a detailed understanding of protein-ligand and protein-protein interactions. Chromatography has been crucial to the development of every phase of present-day structural biochemistry from the isolation of homogeneous preparations of macromolecules to the separation and sequence analysis of their component parts.

The present article is concerned with some applications of automated Edman degradative technology to the solution of problems related to protein covalent structural analysis. Proteins are the most diverse both in terms of structure and function of all the informational macromolecules in living systems. They in-
clude among their number enzymes, hormones, antibodies, transport and storage proteins, contractile elements and structural components and it is clear that they play key roles in all biological processes. Knowledge of their covalent structures is basic to any understanding in molecular terms of how proteins function. Detailed interpretation of the results of x-ray crystallographic analysis is impossible if the sequence is not known. The sequence provides the blueprint for synthesis of proteins and peptides and this is of great value both in medicine and in the study of structure function relationships.

Comparative sequence analysis has led to the construction of phylogenetic trees based upon chemistry rather than morphology. It has also revealed functionally essential regions of polypeptides including enzymes and antibodies. No better example of this may be found than in the "serine" proteases. Comparative sequence analysis of these enzymes has shown them to be highly homologous despite their diverse function in digestion, blood clotting, fertilization and a variety of systems important in regulation. Perhaps the most interesting development in our understanding of protein function which has resulted in recent years from the application of sequence technology has been the recognition of the highly complex nature of post-translational modifications which take place in the processing of active proteins.

It should be clear from this brief survey that protein sequencing has had far reaching implications with respect to function as well as structure. The vast majority of sequence information has been obtained by the successive application of a series of chemical reactions by means of which amino acids may be removed one at a time from the amino terminus of peptide chains. This method, first described over 20 years ago by Edman, is still the mainstay of modern day sequence analysis. Until very recently, Edman continued to devote his energies to the refinement of the method, finally developing an automated system which has changed the course of protein sequence analysis. Dr. Edman passed away earlier this year and this article is humbly dedicated to his memory.

**Edman Degradation**

The series of reactions shown in figure 1 constitute the Edman degradation. All amino groups in the protein or peptide are coupled at alkaline pH with phenylisothiocyanate (or derivatives in which R = -CH₃, -SO₃⁻ or other possible side chains). The resulting phenylthiocarbamyl peptide is then subjected to anhydrous acid whereupon the sulfur atom of the substituent on the αNH₂ group attacks the carbonyl carbon of the first amino acid. This cyclization reaction results in cleavage of the first amino acid in the form of an anilinothiazolinone leaving the residual peptide chain shorter by one residue. The anilinothiazolinone is extracted and converted to the more stable phenylthiohydantoin (PTH) derivative prior to identification and the residual peptide is subjected to as many cycles of the degradation as seems warranted. When performed manually the degradation is seldom successful beyond 15 residues.

The PTH amino acids may be identified directly by gas, thin layer or high pressure liquid chromatography or by amino acid analysis following hydrolysis to the parent amino acid in hydroiodic acid (figure 2). The course of the sequence analysis may also be followed subtractively either by amino acid analysis of the peptide after each cycle of the Edman procedure or by the more sensitive reaction of a portion of the residual peptide with dansyl chloride (figure 2).

In 1968, Edman and Begg reported the development of instrumentation capable of performing the coupling and cyclization phases of the degradation. The efficiency of the operation was such that as many as 60
cycles could be carried out on the protein myoglobin. Thus, the stage was set for the sequence analysis of intact proteins and polypeptides rather than smaller peptides. Up until that time protein sequencing involved the generation and isolation of many small peptides. The most time consuming aspect of the work involved peptide purification and the determination of where each peptide fit in the parent chain. Access to the more efficient automated method simplified the task in that methods could be employed to generate only a few large fragments, each of which could be extensively sequenced automatically. Moreover, the method could be employed on intact proteins so that NH$_2$-terminal sequences could be compared for a variety of related proteins without completing the whole sequence analysis.

The Edman and Begg procedure utilizes a spinning glass cup in which all of the reactions are carried out in an atmosphere as free as possible from oxygen. There is no covalent attachment of the protein sample to the cup so it can be lost extractively after several cycles. Other methods of automation have been developed in which the peptides are attached to insoluble polymers prior to degradation (solid phase Edman) but this article will consider only the spinning cup technology. The topics to be discussed in relation to automated Edman degradation are: (1) the strategy of complete sequence analysis; (2) comparative analysis of NH$_2$-terminal sequences; and (3) efforts toward increasing the sensitivity of sequence analysis. Most of the examples selected are taken from publications from our laboratory; it is hoped that they will emphasize the many ways in which this technology may be applied to problems of clinical interest.

**Strategies for Complete Sequence Analysis**

Since the automated Edman procedure is optimally applied to the analysis of larger fragments, the strategy for complete sequence determinations involves use of cleavage methods which are limited, specific and quantitative. Two amino acids which occur relatively rarely in proteins are methionine and arginine. Several years ago, a method was described in which cyanogen bromide was used for the selective and quantitative cleavage of proteins at methionine residues. Cyanogen bromide cleavage is one of the most important methods in protein sequencing by automated Edman procedures. Specific cleavage at arginine is possible by tryptic digestion of proteins in which all lysyl residues have been modified. A useful reagent for this purpose is citraconic anhydride since the modification can be

\[ R_1 - R_2 - R_3 \]

\[ \text{Citraconic anhydride} \]

**Figure 1.** Sequential liberation of amino acids from the NH$_2$-terminus of a peptide chain by the Edman degradation.

\[ \text{Amino acid analysis} \]

\[ \text{Edman cycle 1} \]

\[ \text{Edman cycle 2} \]

\[ \text{DNS-A} \]

\[ \text{DNS-B} \]

**Figure 2.** Methods for monitoring the course of protein sequence analysis.
reversed by exposure to acid. Thus lysine residues can be liberated for subsequent tryptic digestion.

The application of these tactics to the sequence analysis of human platelet factor 4 (HPF4) is shown in figure 3. This polypeptide of 70 residues contains no methionine but three residues of arginine (residues 20, 22 and 49). Since the arginine at residue 20 is followed by a proline (a bond not cleaved by trypsin) only three peptides (Tc-1, Tc-2 and Tc-3) would be expected by tryptic digestion of the citraconylated protein. Automated Edman degradation of the intact protein proceeded 30 residues into the chain, thus providing the alignment of the Tc fragments. Application of this method to purified Tc-2 and Tc-3 gave the complete sequences of the fragments and, hence, of HPF4 itself. The point of this example is that automated Edman degradation performed even under conditions in which there is no covalent attachment to the cup, may proceed to the end of a peptide. It has been found by us that this is the case especially with fragments terminating with the strongly positively charged arginy1 residues. Conventional sequence analysis of Tc-3 would have been a particularly tricky undertaking in view of the unique constellation of charged and hydrophobic amino acids in this segment (Leu-Tyr-Lys-Lys-Ile-Ile-Lys-Lys-Leu-Leu-).

The high efficiency of automated Edman degradation was also observed in the analysis of protease inhibitors from pineapple stem. These inhibitors are isolated as two chains (A and B, figure 4) connected by disulfide bonds and are extremely heterogeneous. Indeed, seven fractions containing bromelain inhibitory activity can be resolved by ion exchange chromatography. Even within one fraction chosen for sequence analysis, microheterogeneity was observed and it is doubtful that this would have been detected by conventional sequence analysis. The sequence shown in figure 4 was obtained by single analyses of each separated A and B chain. However, analysis of residues 1 and 8 of the A chain gave small yields of Glu and Thr, respectively. Similarly, analysis of the B chain revealed Arg as well as Glu at 11. Thus, the inhibitor shown in figure 4 was contaminated by a related protein containing substitutions at three positions. This example demonstrates the power of the automated method in allowing detection of minor amino acid replacements in peptide chains. Its application in the study of aberrant proteins associated with disease states such as sickle cell anemia is obvious.

**Comparative Analysis of Amino Terminal Sequences**

In many cases, interest is focused upon partial rather than complete sequence in-
formation. Automated Edman degradation is especially valuable in studies of this kind. Partial sequences may be desirable for several reasons. Because the sequence analysis provides yields of PTH amino acids in addition to the sequence itself, such studies can provide information regarding the purity and/or subunit structure of a protein. Yeast inorganic pyrophosphatase was a subject of some controversy insofar as various reports had been published regarding its subunit structure. The enzyme of molecular weight 64,000 was variously interpreted to be a single polypeptide chain, a dimer of identical subunits and a $\alpha_2\beta_3$-tetramer. Automated Edman degradation of a sample of the intact protein gave a single sequence with yields of PTH derivatives consistent with those expected from a chain of molecular weight about 30,000. These and other criteria, therefore, supported the dimer theory and the structural analysis of the subunit has since been completed by us.

Another example in which NH$_2$-terminal sequence analysis was of value is seen in the study of phytohemagglutinin in mitogenic proteins (PHAP). The PHAP from the red kidney bean comprise a class of five glycoproteins that are isomeric tetramers composed of varying proportions of L and R subunits. Within the native tetramer, the L subunit is a potent leukoagglutinin and mitogen that lacks hemagglutinating properties; the R subunit is a potent hemagglutinin with little mitogenic activity. Separation of the R and L subunits proved to be a difficult task. Automated Edman degradation showed that although they differed in sequence in the first seven residues, thereafter the R and L subunits were identical through residue 24 (figure 5). Further analysis indicated that the similarity between R and L extended throughout their respective chains. Therefore, the very distinctive biological properties of the R and L subunits must be a consequence of rather restricted differences in their covalent structures.

Amino terminal sequence analyses of a particular protein from many different sources may reveal regions of the molecule that are highly invariant or absolutely conserved in addition to sequences that are subject to wide variation. Such is the case in the variable region of immunoglobulins in which hypervariable domains have been identified. Comparative analysis of the NH$_2$-terminal sequences of phospholipases A

A Chain: H$_2$N-Asp-Glu-Tyr-Lys-Cys-Tyr-Cys-Ala-Asp-Thr-
(A-1) 1 5 10

Tyr-Ser-Asp-Cys-Pro-Gly-Phe-Cys-Lys-Lys-
15 20

Cys-Lys-Ala-Glu-Phe-Gly-Lys-Tyr-Ile-Cys-
25 30

Leu-Asp-Leu-Ile-Ser-Pro-Asn-Asp-Cys-Val-
35 40

Lys-COOH

B Chain: H$_2$N-Thr-Ala-Cys-Ser-Glu-Cys-Val-Cys-Pro-Leu-
(B-2) 1 5 10

Gln-COOH

FIGURE 4. Amino acid sequence of a bromelain inhibitor from pineapple stem. Variability was observed in positions 1 (Glu or PCA) and 8 (Thr) in the A chain and in position 11 (Arg) in the B chain, thus giving rise to the possibility of structural isomers. [From Reddy et al., with permission.]

| L Subunit |
| H$_2$N- Ser - Asn - Asp - Ile - Tyr - Phe - Asn - |
| 1 5 |

| R Subunit |
| H$_2$N- Ala - Ser - Glu - Thr - Ser - Phe - Ser - |
| 1 5 |

(Glu-NH$_2$)$_2$-(Man)$_4$- |

| -Phe - Glu - Arg - Phe - Asn - Glu - Thr - Asp - Leu - |
| 10 12 15 |

| -Ile - Leu - Glu - Arg - Asp - Ala - Ser - Val - |
| 20 |

FIGURE 5. Partial sequence analysis of the L and R subunits of the tetrameric mitogenic PHAP from the red kidney bean. Note: The residues in italics (residues 8-24) are shown here only once since they are identical for both L and R subunits. [From Miller et al., with permission.]
from various sources revealed two regions in which particular sequences were highly conserved. The conclusion that these regions might be important in catalysis has since been borne out by x-ray crystallographic studies.

It has been known now for some time that proteins are often synthesized as inactive or unfinished precursors. The pancreatic proteinases trypsin and chymotrypsin, for example, are elaborated as inactive zymogens trypsinogen and chymotrypsinogen. Insulin is derived from a single chain precursor, proinsulin, by proteolytic excision of a so-called C-peptide. This posttranslational processing of proteins has recently been shown to be even more extensive and the evidence for this has been obtained by automated Edman degradation on a micro scale. Sequence analysis of the protein product synthesized in cell-free systems containing appropriate messenger RNA has revealed that secretory proteins such as proinsulin, immunoglobulin light chains, parathyroid hormone, growth hormone, placental lactogen, insect promellitin and the pancreatic zymogens are made as precursors containing extensions of about 20 amino acids at the NH$_2$-termini. These pre-sequences are believed to be involved in the vectorial discharge and sequestration of the nascent polypeptide chains within the microsomal cisternae. It would follow that these peptide extensions or “signal” peptides would be cleaved soon after each cycle of synthesis by microsomal proteases. Thus far, the major attribute shared by all of the peptide extensions is their predominant content of hydrophobic amino acids. This may expedite their interaction with membranes.

The signal peptide hypothesis remains to be confirmed but it is important from the standpoint of this communication to stress that automated Edman degradation was applicable to the analysis of samples in the picomole range. This was done by incorporating particular radioactively labeled amino acids into the cell free synthetic product. The radioactive protein was then sequenced as a mixture with carrier protein of known structures and each cycle was analyzed for radioactivity.

Increasing the Sensitivity of the Edman Degradation

In the case of many proteins, including the messenger products referred to previously, it is difficult to obtain more than a few mg of material, if that much. Most of the sequence analysis to date has been performed on samples greater than 100 nmol. Efforts have been made to extend the sensitivity of the procedure by varying the programs employed so that quantities of protein in the 3 to 20 nmol range can be sequenced extensively. The use of a fluorescent isothiocyanate has also been explored. However, in order to extend the sensitivity to the picomole range, it would seem to be most appropriate to employ radioactivity. The studies described for the pre-protein sequence analyses were performed on numerous samples, each of which were synthesized with only one radiolabeled amino acid. Although this approach is laborious, the results thus obtained are perhaps more reliable than would be the case if a mixture of all of the radiolabeled amino acids were employed in the cell free synthesis. However, procedures have been described in which $[^{35}\text{S}]$phenylisothiocyanate is used in the degradation. The $[^{35}\text{S}]$PTH amino acids thus liberated during the analysis of sub-nanomole quantities of protein are added to a mixture of unlabeled PTH amino acids and the mixture is subjected to two-dimensional thin layer chromatography. After inspection of the plate, the chromatogram is subjected to autoradiography and the labeled PTH amino acid is identified. Quantitation can be achieved by scraping the spot and counting in a
liquid scintillation counter. Alternatively, samples can be analyzed by gas chromatography in a system fitted with a detector for radioactivity.

Conclusions

Few procedures used decades ago in biochemical analysis are employed today. A notable exception is the Edman degradation. Although the method has been refined and adaptations have been made, the chemistry is the same. New approaches to protein sequencing involving mass spectrometry are being evaluated and it may be that these will provide the capabilities for complete automation of the process with the sensitivity required for the many as yet uncharacterized proteins. Until then, this degradative scheme will continue to yield the bulk of the important sequence information yet to be derived.

References