UNWINDING THE PROTEIN

by

PEHR EDMAN

Max-Planck-Institute of Biochemistry
8033 Martinsried bei München
German Federal Republic

Key Words: protein, primary structure, sequence determination

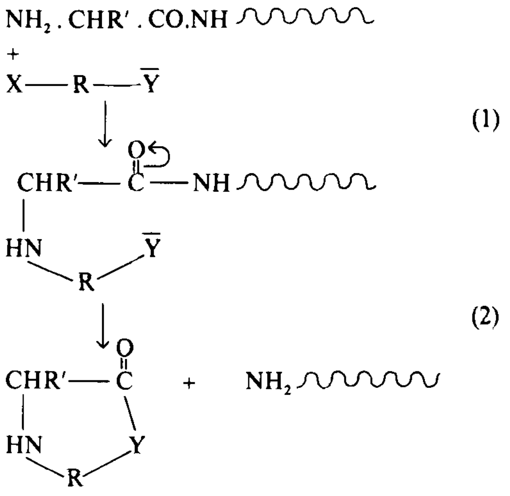
K. LINDERSTRØM-LANG lecture delivered at the Carlsberg Laboratory Centennial and the Inauguration of the Carlsberg Research Center on September 28th, 1976.

In his 1951 Lane Medical Lecture (15) on »Proteins and Enzymes« LINDERSTRØM-LANG introduced the terms *primary, secondary* and *tertiary* to represent different levels of spatial arrangements in a protein structure. The primary structure would then essentially be the amino acid sequence, whereas the secondary and tertiary structures represent higher spatial arrangements. How to determine a primary structure is actually the theme of this lecture.

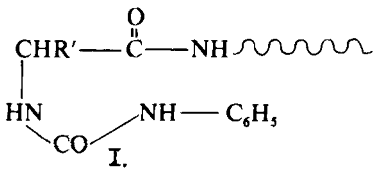
As is common knowledge, a protein consists of a peptide chain(s) built of 20 different amino acids bound to each other through amide (peptide) bonds. To know the primary structure of the protein one must determine the order in which the amino acids occur, i.e. the amino acid sequence. Here is to observe that the peptide chain has a direction caused by the amino-carboxyl polarity of the amino acids. As a consequence one end carries a free α-amino group, the N-terminal, and the other a free α-carboxyl group, the C-terminal.

That much we have known almost since the beginning of this century. However, the solution to the problem of determining the amino acid sequence was slow in coming. A start was made when the partial hydrolysis method was introduced in the mid-forties. Here the peptide bonds were partially and randomly hydrolysed, and a sufficient number of fragments then isolated and structurally determined. The data enabled a reconstruction of the original structure. This procedure was used very successfully by SANGER in his elucidation of insulin (16). However, it soon stood clear that this approach was not applicable to the generally much larger protein structures.

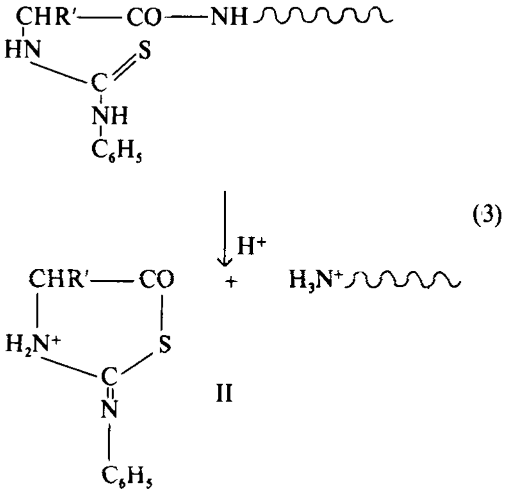
Around that time I became interested in the problem. A stepwise degradation starting from the N- or C-terminal end, respectively, with the removal and identification of one amino acid at a time seemed a reasonable approach. Two problems then immediately presented themselves. The first was how to break one peptide bond, and leave all the others unaffected. Peptide bonds are very stable, and to break them usually requires boiling with strong acids or alkalis for many hours. One approach would be to offer the -CO-group of the peptide bond an energetically more favourable reaction partner than the -NH-group. A nucleophile would be a likely candidate. Let us call the nucleophile $\overline{\text{Y}}$. Secondly, the nucleophilic attack had to be directed at the carbon of the first peptide bond and only there. This could be done by attaching the nucleophilic agent through a second group, X, to the terminal amino group. Our bifunctional reagent is now of the type X-R-$\overline{Y}$. A stepwise degradation could then be envisaged to proceed as shown schematically (Reactions 1 and 2).

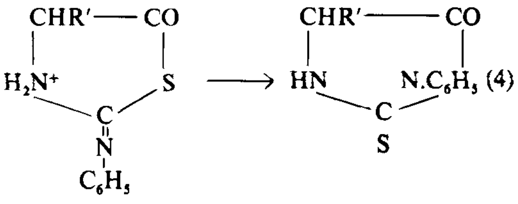


An old observation by ABDERHALDEN and BROCKMANN (1) seemed to support such a scheme. They had reacted a peptide with phenylisocyanate to form a phenylcarbamyl derivative (Formula I.). Following hydrolysis the terminal amino acid could, as was to be expected, be isolated as a phenylhydantoin. However, the rate of cleavage of the first peptide bond was increased more than was to be expected from a mere blocking of the terminal amino group. This could be interpreted in the terms of a nucleophilic attack as already indicated. Our interpretation however was not correct, since it has later been shown that the cleavage of the peptide bond *precedes* the formation of the hydantoin (11). In any event it led us to try phenylisothiocyanate in the expectation that it would be a more efficient reagent. So, indeed, it was. The peptide bond was split in few seconds at room temperature and in a anhydrous acid medium. The expected phenylthiohydantoin and the shortened peptide could be isolated from the reaction medium. In other words, everything seemed to fit neatly (5).


I.

Only later did the suspicion arise that this was not the whole story. It was discovered that the first compound to form during the cleavage reaction was not a phenylthiohydantoin but a 2-anilino-5-thiazolinone (Formula II, Reaction 3), and the latter readily converted to the isomeric phenylthiohydantoin (Reaction 4) (6).





The nucleophile was, in fact, the thioketonic sulphur, and here lies the explanation for the extreme facility of the reaction. An entirely new reaction had actually been discovered. Its generality has subsequently been demonstrated in many other stepwise degradation reactions using other reagents, but where the key reaction is always the formation of a thiazolinone (for a review see (8)).

The chemistry just described has been put to extensive use in primary structure determination of proteins, and has allowed, to this date, the positioning of more than 80,000 amino acids in sequence.

However, much had to be done before the chemistry could be made into an efficient method for sequence determination. I shall only touch lightly on this work. It was mainly a matter of identifying and eliminating side reactions. So was it, for example, discovered that the thiocarbamyl group (Reaction 3) was easily desulphurised through oxydation. The resulting carbamyl group was unreactive, and the degradation came to an end. This side reaction could be eliminated by performing the reaction in the absence of oxygen, e.g. in a nitrogen atmosphere. Further, reaction conditions for coupling, cleavage and conversion that embraced all amino acids had to be found. In fact, until recently I throught we knew these conditions, since they had been used successfully for many years in our own laboratory. Then it turned out that one of our peptides from fibrinogen gave rise to persistent overlaps, i.e. contamination from a preceding step, starting at a particular proline residue. Reports to the same effect had also appeared from NEURATH'S (14) and von HOLT'S (3) laboratories. The matter was studied jointly with von HOLT'S group (2). We found that proline was much more slowly released than other amino acids during the cleavage reaction (Figure 1). However, the position of the proline in the sequence was also of importance, since the rate of release could vary considerably (Figure 2).

Finally, identification methods had to be worked out for the phenylthiohydantoins. Originally it was done by paper chromatography and later by thin-layer chromatography. Today the arsenal has been extended to include gas chromatography, mass spectroscopy and recently also high performance liquid chromatography (for a review see (8)).

The next step in the development was the automation of the degradation. Two essential preconditions were already at hand. First, standard reaction conditions applicable to all amino acids were known. Secondly, since side reactions had largely been eliminated, a high repetitive yield[1] was possible. The importance of a high repetitive yield should be clear from

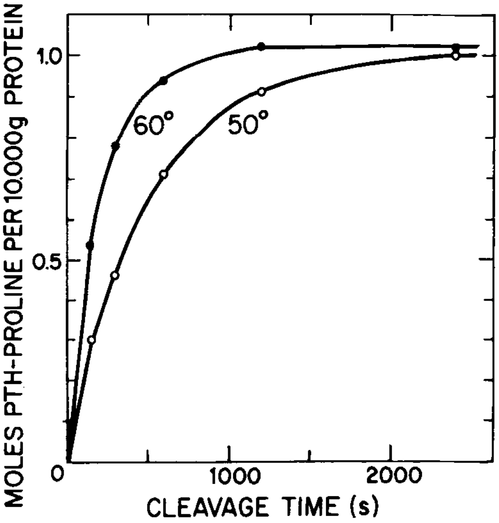


Figure 1. The time course of the cleavage of N-terminal proline in phenylthiocarbamylated salmine at 50°C (-○-○-) and at 60°C (-●-●-). Medium: heptafluorobutyric acid. Penultimate amino acid: arginine.
(Courtesy Walter de Gruyter & Co., Berlin)

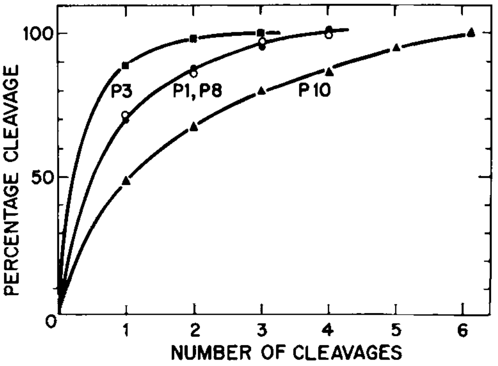


Figure 2. Sequenator analysis of histone H2B from chicken erythrocytes. Shown is the course of the cleavage of prolines at position 1 (-○-○-), position 3 (-■-■-), position 8 (-●-●-), and position 10 (-▲-▲-). Duration of each cleavage operation approximately 90 s. Temperature 55°C.
(Courtesy Walter de Gruyter & Co., Berlin)

[1] Repetitive yield is defined as the percentage yield from one degradation cycle to the next.

the fact that 97%, 98% and 99% make possible 30, 60 and 120 degradation cycles, respectively. It remained to find a suitable technical device in which to perform the reactions of a degradation cycle. We chose to spread out the reaction media in thin films inside a spinning cup (7). The film is very suited for carrying out extractions, dryings and so on (Figure 3). The whole degradation cycle may be programmed. The instrument is called a sequenator (Figure 4). This instrument allows extended degradations, in favourable cases up to 50 to 60 amino acids. The speed is about 25 amino acids a day in contrast to the one or two amino acids with the manual technique. There are still limitations to the sequenator technique, but there is good reason to think that in time they will be removed.

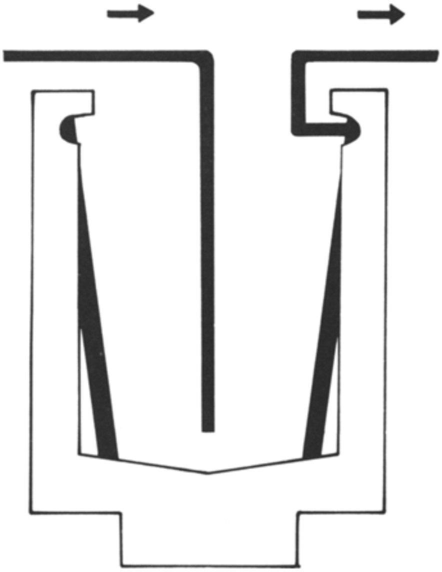


Figure 3. Schematic presentation of the operation of the sequenator cup. An aqueous solution (white) is being extracted by an organic solvent (black).

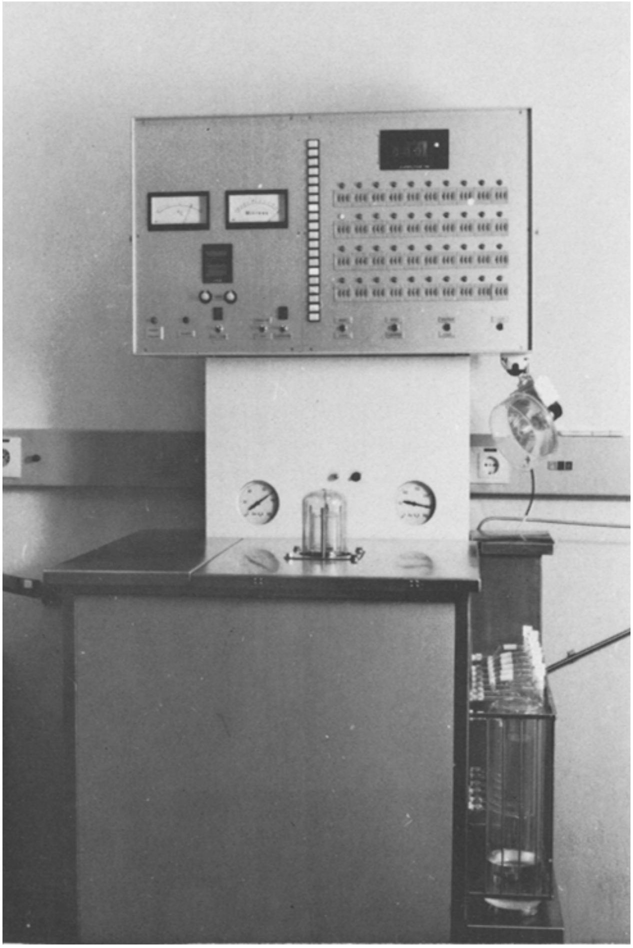


Figure 4. The sequenator.

Automation has changed the strategy of primary structure determination. It is no longer necessary to break up large structures in many small fragments. On the contrary, it is now desirable to have large fragments not only because they behave better in the sequenator, but above all because the time and labour involved in separating many small fragments (including numerous overlap fragments) are substantially reduced. Therefore automation increases the speed of sequencing in more than one way.

To illustrate how a structural elucidation is done today I am going to use some recent work from our own laboratory. This work was carried out by Dr. A. HENSCHEN in collaboration with two students, Mr. F. LOTTSPEICH and Miss R. WARBINEK, and Dr. T. SEKITA (13). The subject was the structure of human fibrinogen. I should also say that parts of structures I will show have also been reported by BLOMBÄCK's group in Stockholm (4) and DOOLITTLE's in La Jolla (17).

Fibrinogen has a molecular weight of 340,000, and its structure is known to contain six peptide chains, Aα, Bβ and γ, which are pairwise identical (Figure 5). The conversion of soluble fibrinogen into insoluble fibrin is preceded by the splitting off of two small peptides, fibrinopeptides A and B, by thrombin. The chains are held together by interchain disulphide bonds. Following reduction and alkylation of these bonds the α-, β- and γ-chains may be separated from each other. This is technically a difficult step as it has to be carried out in the presence of 8 M urea (Figure 6). Using the isothiocyanate method a quantitative N-terminal determination was made on the separated chains, and this immediately gave us their molecular weight, i.e. the α-chain 75,000, the β-chain 67,000 and the γ-chain 51,000.

The work began with the γ-chain. The molecular weight indicated that it contained about 400 amino acids. A structure this size is far to large to be sequenced directly, and has to be fragmented. A most elegant method, due to GROSS and WITKOP (10), permits the selective cleavage of the peptide chains C-terminally of the methionine residues. The reagent here is cyanogen bromide. The amino acid analysis showed eight methionines, and a cyanogen bromide cleavage should therefore give rise to nine fragments. The fragments were separated by various chromatographic techniques, which I shall leave out here. I shall only mention that the separation procedure was monitored with N-terminal determination. It is very easy to loose a fragment in the separation procedure, particularly if the fragment is small, and has no u.v. absorption. This danger is very much reduced through what I would call »N-terminal book-keeping«, i.e. all N-terminal amino acids present in the original digest should be accounted for by the N-terminal amino acids of the isolated fragments. The fragments, with short N-terminal sequences[2], are shown in Figure 7.

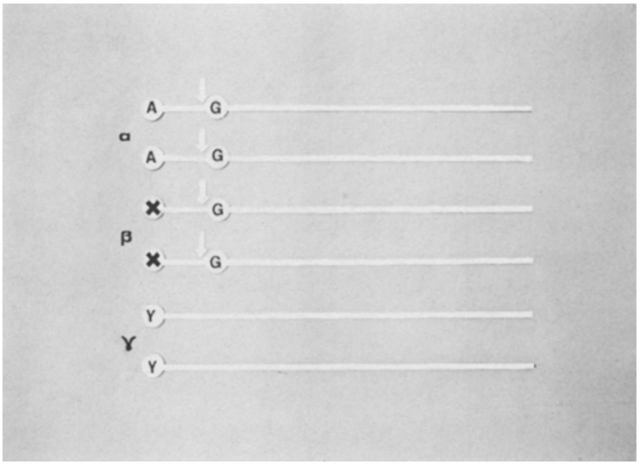


Figure 5. Schematic presentation of the fibrinogen structure. The capital letters represent N-terminal amino acids: A, alanine; X, pyroglutamic acid; Y, tyrosine; G, glycine. The arrows indicate where thrombin cleaves fibrinogen to form fibrin.

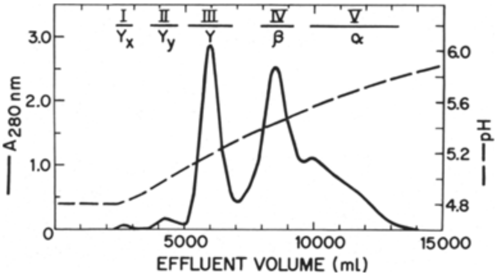


Figure 6. Chromatographic separation of the α-, β- and γ-chains of fibrin on CMC-cellulose using a pH gradient.

The true order of the fragments (apart from the N- and C-terminal fragments) must now be established, and for this we have to find the following overlap sequences:

MKK
MKI
MKY
MID

MLE
MFK
MQF
MNK

To this purpose the chain was fragmented in a different way. The procedure chosen was to split C-terminally of the arginine residues. This

[2] It is practice in our laboratory to name fragments after their N-terminal sequences. These are easy to memorize of one uses the one-letter notation for the amino acids, and simplify the verbal communication in the laboratory.

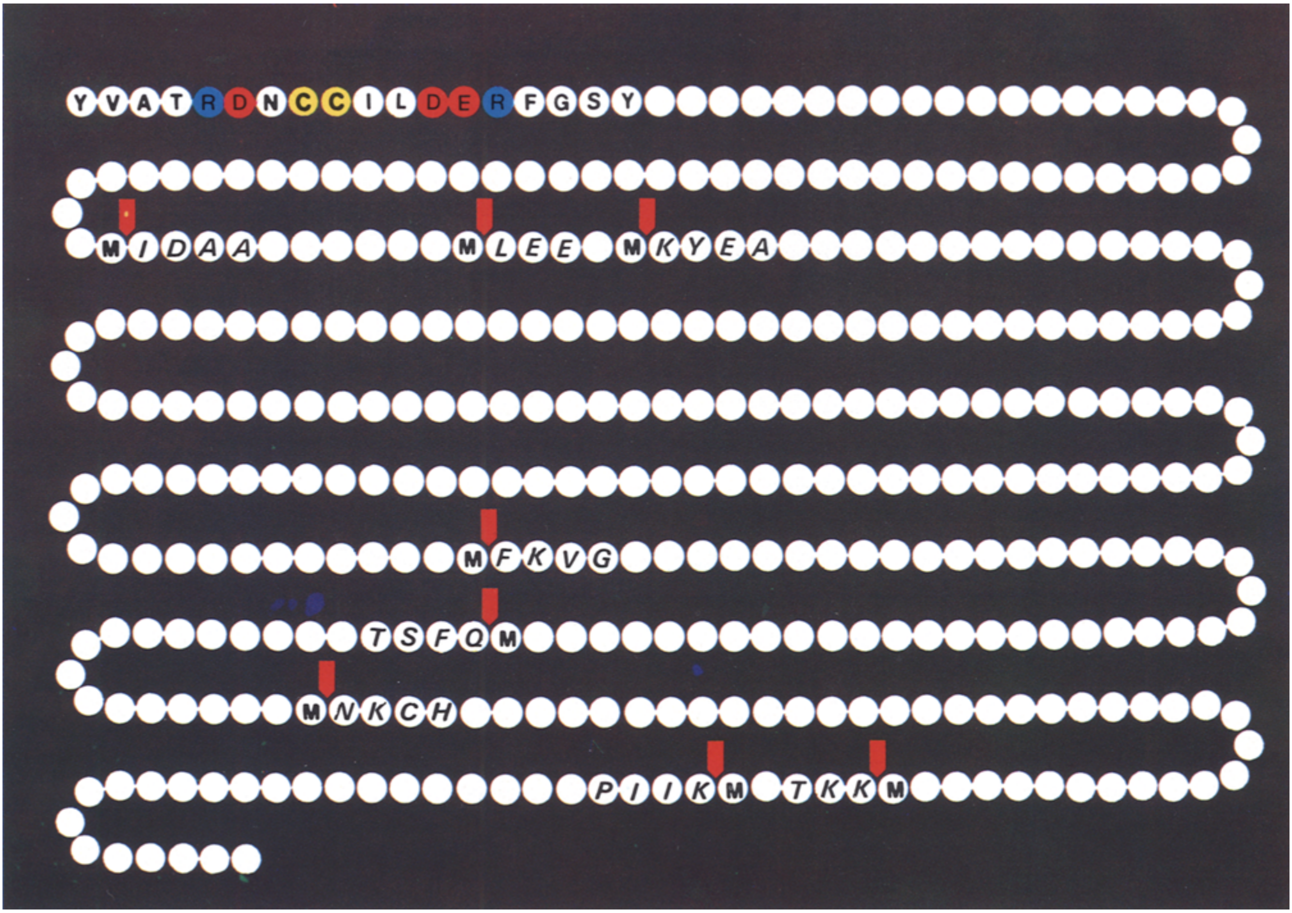


Figure 7. Cyanogen bromide cleavage of the fibrin γ-chain. The arrows indicate the peptide bonds cleaved by cyanogen bromide. The one-letter abbreviations for the amino acid residues are those recommended by the IUPAC-IUB Commission on Biochemical Nomenclature.

could be done with trypsin provided the amino function of the lysine residues was first blocked (9). The overlap sequences may be seen in Figure 8. it would obviously have been equally possible to carry out the procedure in the reverse order, i.e. to start with trypsin fragmentation and to use the cyanogen bromide fragments to order the trypsin fragments.

The broad principles followed in a primary structure determination should be sufficiently expounded, and I shall go directly to the final result. Figure 9 shows the stand of the work a while ago. I may add that the complete structure of the γ-chain is now known.

Some general comments may be in place here. The first concerns the time schedule of the undertaking. The actual sequencing work took less than six months. Far more time was spent in a) finding the right conditions for isolating the chains in pure form and sufficient amounts, i.e. the logistics, and b) in isolating the cyanogen bromide and trypsin fragments. The sequencing was therefore the lesser part of the task.

My second comment has perhaps a more personal note to it. People with no experience of structural work sometimes tell me that they believe it is a) tedious, b) with automation a routine. This is wide off the mark. It may have been tedious at the time when all sequencing

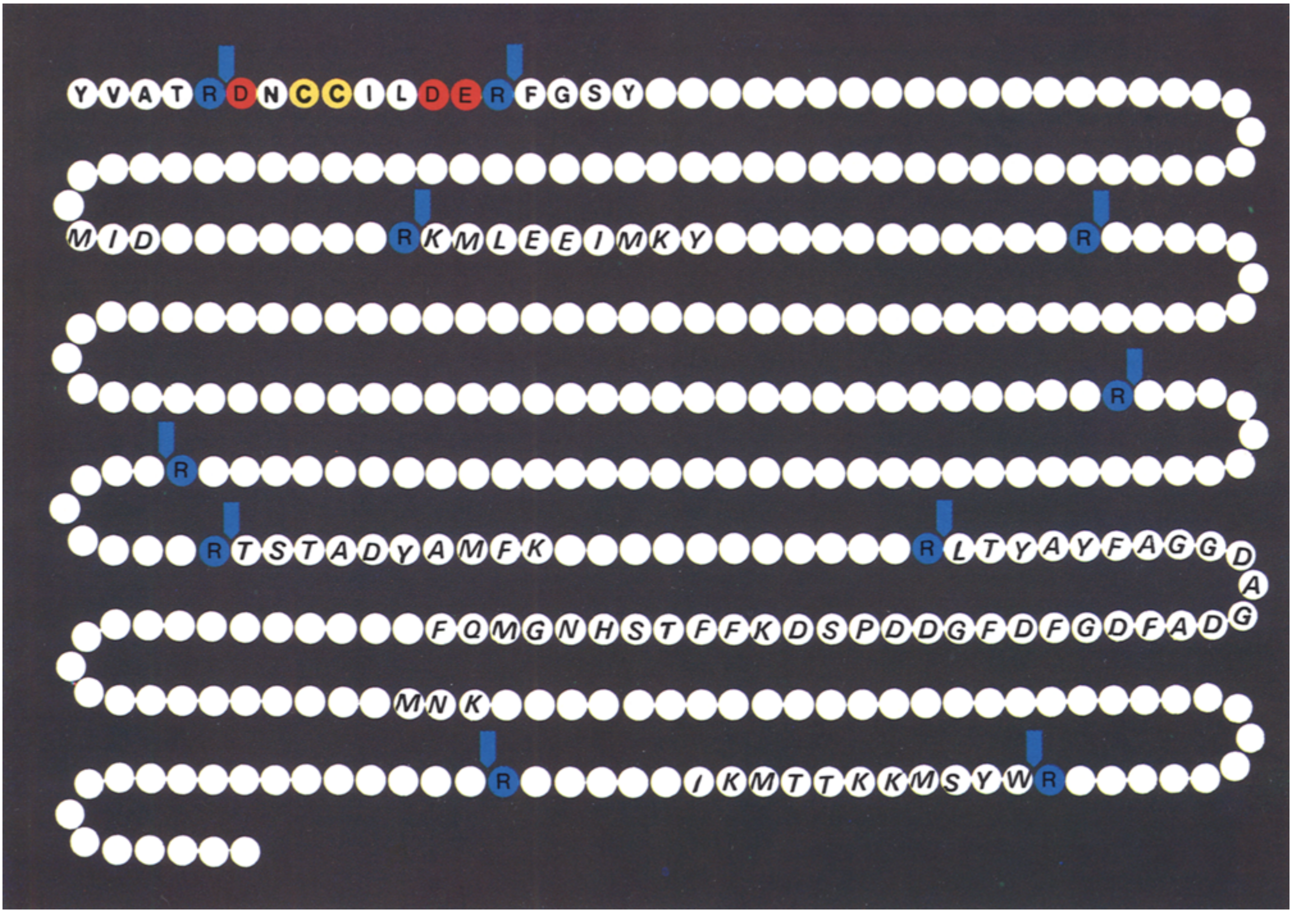

Y V A T R D N C C I L D E R F G S Y
M I D R K M L E E I M K Y R
R
R
R T S T A D Y A M F K R L T Y A Y F A G G D
A
F Q M G N H S T F F K D S P D D G F D F G D F A D G
M N K
R I K M T T K K M S Y W R

Figure 8. Trypsin cleavage of the fibrin γ-chain after blocking the free amino groups of the lysine residues. The sequences around the methionyl residues are shown.

had to be done by hand, but automation has done away with most of that. The belief that the work has become a routine is even more untrue. In fact, the solution of a large structure taxes the investigator's resources of skill and knowledge to the utmost, and luck is not an unessential factor. The comparison to the solution of a super-crossword puzzle is perhaps apt. Anyone, who has experienced the elation in the laboratory when a fragment known to be missing has been found, or a tantalizing inconsistency has been resolved, would know what I mean.

I indicated earlier that more than 80,000 amino acids have been placed in sequence. This figure may seem large, but is dwarfed when compared with the maximum coding capacity in a mammalian genome of 2 billion residues. Most likely only a very small fraction of this capacity is actually used for the coding of protein structures. Even so the number is likely to be large, and this means that the road before us will be long. At present there are several hundred automatic sequencing devices in operation. This will indoubtedly speed up the data accumulation, and it may reasonably be expected that in the next few years the 200,000 mark will be reached: This faces us with a problem of increasing urgency and importance, namely that of data storage, data retrieval and data processing. Every structure has, apart from its interest per se, also an interest in its

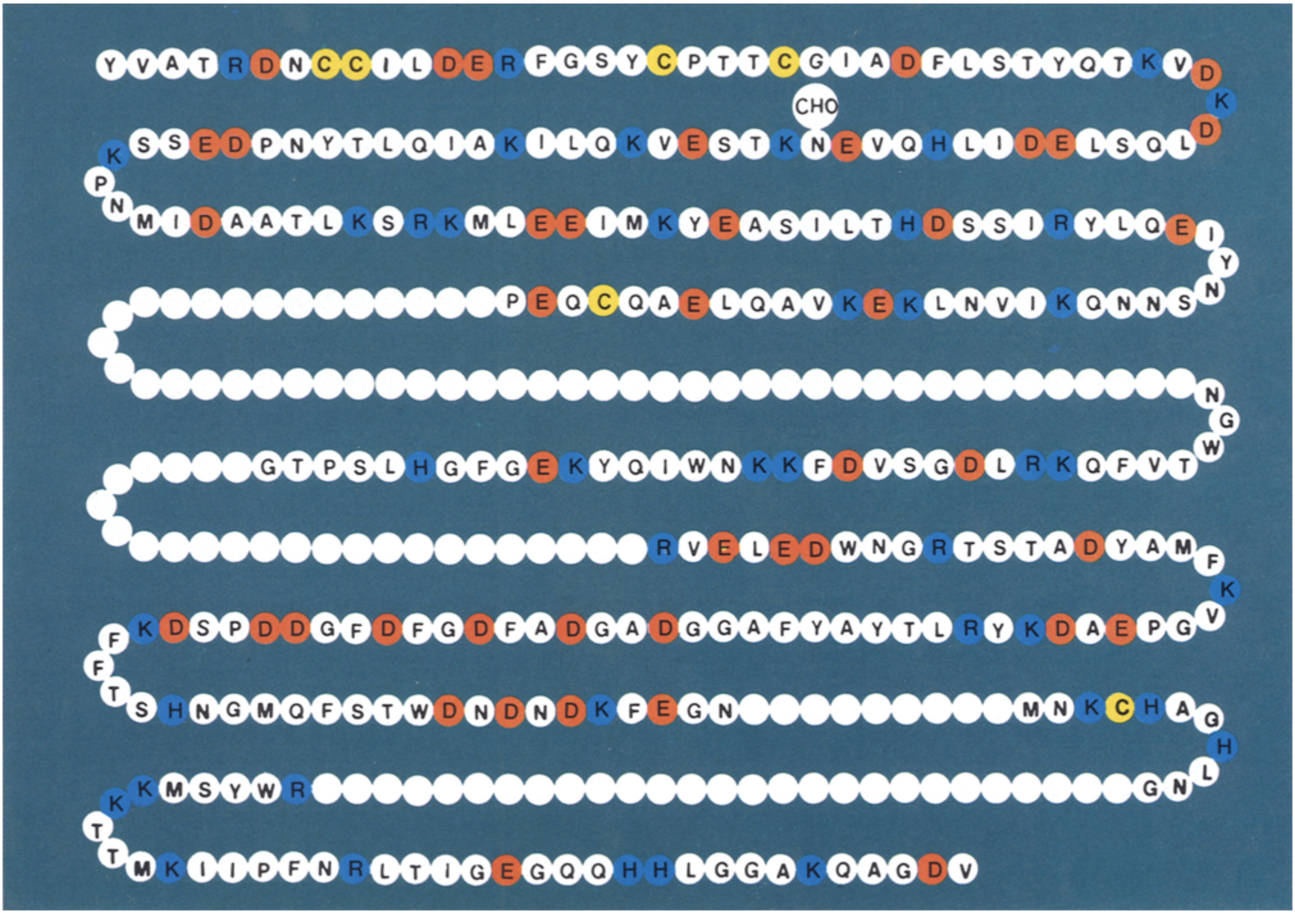

Y V A T R D N C C I L D E R F G S Y C P T T C G I A D F L S T Y Q T K V
D K D
CHO
K S S E D P N Y T L Q I A K I L Q K V E S T K N E V Q H L I D E L S Q L
P N
M I D A A T L K S R K M L E E I M K Y E A S I L T H D S S I R Y L Q E
I Y N
P E Q C Q A E L Q A V K E K L N V I K Q N N S
N G W
G T P S L H G F G E K Y Q I W N K K F D V S G D L R K Q F V T
R V E L E D W N G R T S T A D Y A M
F K V
K D S P D D G F D F G D F A D G A D G G A F Y A Y T L R Y K D A E P G
F F T
S H N G M Q F S T W D N D N D K F E G N
M N K C H A
G H L
K M S Y W R
G N
K T T
M K I I P F N R L T I G E G Q Q H H L G G A K Q A G D V

Figure 9. Partial structure of the fibrin γ-chain.

possible relationship to other structures. This is so because we know that protein structures undergo changes either through evolution or through short term in vivo processes. One may instance the relationship between lysozyme and α-lactalbumin, secretin and glucagon, trypsin and thrombin, or β-liprotropin and the peptides with morphin-like activities in the pituitary. Until now such relationships have not eluded us (or at least we hope so), but will that be so also in the future with the data volume growing at an exponential rate? I have misgivings about that. It is true that the Atlas of Protein Sequencing through its computer service at present performs this task. However, this is done on an insecure,year to year financial basis. A most important function is therefore not sufficiently safeguarded. We may in time expect the unravelling of a new systema naturalis among the biomolecules. It would be tragic if this development would be endagered through our own negligence.

REFERENCES

1. ABDERHALDEN, E. & H. BROCKMANN: Beitrag zur Konstitutionsermittlung von Proteinen bzw. Polypeptiden, Biochem. Z. 225, 386-408 (1930)
2. BRANDT, W.F., P. EDMAN, A. HENSCHEN & C. von HOLT: Abnormal behaviour of proline in the

isothiocyanate degradation, Hoppe-Seyler's Z.f.physiol.Chemie, 357, 1505-1508 (1976)

3. BRANDT, W. F. & C. von HOLT: The determination of the primary structure of histone F3 from chicken erythrocytes by automatic Edman degradation, Eur.J.Biochemistry 46, 419-429 (1974)
4. COLLEN, D., B. KUDRYK, B. HESSEL & B. BLOMBÄCK: Primary structure of human fibrinogen and fibrin. Isolation and partial characterization of chains of fragment D, J.Biol.Chem. 250, 5708-5817 (1975)
5. EDMAN, P.: Method for determination of the amino acid sequence in peptides, Acta Chem. Scand. 4, 283-293 (1950)
6. EDMAN, P.: On the mechanism of the phenylisothiocyanate degradation of peptides, Acta Chem.Scand. 10, 761-768 (1956)
7. EDMAN, P. & G. BEGG: A protein sequenator, Eur.J.Biochemistry 1, 80-91 (1966)
8. EDMAN, P. & A. HENSCHEN: Sequence determination. In: S. B. Needleman, ed., Springer Verlag (Berlin, Heidelberg, New York) Protein Sequence Determination pp. 232-279 (1975)
9. GOLDBERGER, R. F. & C.B. ANFINSEN: The reversible masking of amino groups in ribonuclease and its possible usefulness in the synthesis of the protein, Biochemistry 1, 401-405 (1962)
10. GROSS, E. & B. WITKOP: Selective cleavage of the methionyl peptide bonds in ribonuclease with cyanogen bromide, J. Amer.Chem.Soc. 83, 1510-1511 (1961)
11. HAGEL, P.: N-terminal amino acid determination. Ph.D. thesis, University of Amsterdam (1970)
12. HENSCHEN, A. & P. EDMAN: Large scale preparation of S-carboxymethylated chains of human fibrin and fibrinogen and the occurence of γ-chain variants, Biochim.Biophys.Acta 263, 351-367 (1972)
13. HENSCHEN, A., F. LOTTESPEICH, T. SEKITA & R. WARBINEK: Amino acid sequence of human fibrin. Preliminary note on the order of peptides obtained by cleaving the γ-chain at the methionyl and arginiyl bonds. Hoppe-Seyler's Z.f.physiol. Chemie 357, 605-608 (1976)
14. HERMODSON, M. A., L. H. ERICSSON, K. TITANI, H. NEURATH & K. A. WALSH: Application of sequenator analysis to the study of proteins, Biochemistry 11, 4493-4502 (1972)
15. LINDERSTRØM-LANG, K. U.: Proteins and Enzymes, University Press (Stanford, Calif.) (1952)
16. SANGER, F. & H. TUPPY: The amino acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates, Biochem.J. 49, 463-481 (1951)
17. TAKAGI, T. & R. F. DOOLITTLE: Amino acid sequence studies on plasmin-derived fragments of human fibrinogen: Amino-terminal sequences of intermediate and terminal fragments. Biochemistry 14, 940-946 (1975)