S. A. Bernhard*
D. F. Bradley†
W. L. Duda

Automatic Determination of Amino Acid Sequences

Abstract: A fundamental problem for biochemistry is the determination of the linear sequence of amino acids in proteins. This paper describes a computer-oriented logic for obtaining such determination. The logic applies successively stronger decision rules to extract the required information on the protein sequence.

Introduction

One of the fundamental problems facing biochemists is the determination of the linear sequence of the 20 amino acids that constitute proteins. The purpose of this paper is to present a computer-oriented method whereby a highly specific amino acid sequence is derived from fragmentary information about its parts. The solution of the problem is obtained through programmed algorithms which take the fragmentary data as input and generate the complete sequence as output.

Protein chains generally contain on the order of 100 subunits (amino acids) of 20 distinct types. These subunits, or monomers, are linked in the polymer linearly in a chain of definite polarity, e.g.,

$$A_1 \rightarrow A_5 \rightarrow A_3 \rightarrow \cdots A_{20} \rightarrow A_9 \rightarrow A_2 \rightarrow A_1 \rightarrow A_3.$$

Each type of protein presumably has a unique arrangement of subunits. Deoxyribonucleic acid, DNA, is composed of much longer chains ($\gg$100) of subunits (nucleotides) of only four distinct types, e.g.,

$$N_1 \rightarrow N_4 \rightarrow N_3 \rightarrow \cdots N_1 \rightarrow N_2.$$

Again, there is a definite polarity.

DNA (and ultimately ribonucleic acid, RNA) carries the genetic information for all biological replication, including the replication of proteins which proceeds under genetic control. All biological function is governed by specific proteins (such as enzymes, oxygen carriers and structural tissues), the particular function of each one being defined by its unique arrangement of subunits. Current theory postulates that the arrangement of subunits in nucleic acids linearly determines the arrangement of subunits in proteins. One of the very fundamental problems in this field is the "breaking of the 'genetic code'" by which the "gene" on a DNA sequence, written in a four-letter nucleotide language $N_1 \rightarrow N_2 \rightarrow N_4 \rightarrow N_3$, determines or is translated into the functional protein written in a 20-letter amino acid language $A_1 \rightarrow A_5 \rightarrow A_9 \rightarrow$ et cetera.

There have been three distinct types of approach toward breaking the code:

(1) Investigations into the mathematical nature of coding schemes by which four-letter nucleotide sequences determine 20-letter amino acid sequences.[1-17] General considerations, as well as experimentally determined sequences, rule out such schemes as "overlapping" codes (see for example the discussion in Ref. 1).

(2) Observations of the effect of chemically induced errors in the nucleic acid sequence (mutations) on the specific amino acid sequence of specific proteins.[1,21,22] By genetic techniques it is possible to select, from within a species, variant members differing from the usual member in the composition and/or sequence of amino acids in one functional protein. Analysis of the amino acid sequence of this protein coupled with genetic "mapping"

* Institute of Molecular Biology, University of Oregon, Eugene, Oregon
† National Institute of Mental Health, Bethesda, Maryland. (Now at Chaim Weitzmann Institute of Science, Rehovot, Israel.)
‡ Presented in part at IBM Medical Symposium, Endicott, N. Y., September, 1960. (See Footnote 47.)

of each of the mutants affords a powerful tool for "breaking the code."

(3) Biochemical and biophysical investigations of nucleic-acid-regulated protein synthesis.[18–20,23–27] For example Nirenberg et al.[18,19] have shown that polyuridylic acid controls the incorporation of phenylalanine into insoluble peptides in a complex, *in vitro* enzyme system, suggesting that a sequence of several uridine nucleotide bases is the nucleic acid code symbol for "phenylalanine" in protein language.

Fundamental to some of the efforts to "break the code" is an understanding of the actual sequences of amino acids in specific proteins. Knowledge of the sequences for many different proteins will certainly be required before all the details of biological coding are established.

Complete amino acid sequences have been determined for a small number of proteins,[29–34] following essentially the procedure used by Sanger to determine the sequence in insulin. Work is in progress on a number of others.[35–37] The sequence in insulin (51 residues) was determined in about 10 years. Subsequently, advances in instrumentation and novel reactions have shortened the process considerably. Ribonuclease (124 residues), tobacco mosaic virus protein (158 residues) and myoglobin (153 residues), for example, were worked out in a few years.

We initiated this investigation to see whether machine analysis of sequencing data could reduce the amount of effort involved in working out protein and nucleic acid sequences. In this paper we present preliminary results indicating the feasibility of a new approach to this problem.

Sequence determination by machine

A computer has been programmed to use a small set of strategic rules which enable it to scan compositional and sequence data on peptides to determine the sequence of the protein to the maximum extent consistent with the information content of the inputs. The computer initially examines the fragments, making only *categorical* deductions which are retained in its memory. A specific example of this procedure worked out "by hand" is given below. After all categorical deductions have been made the computer re-examines the inputs, applying successively stronger strategic rules to extract more information on how to find the sequence of the protein. Since the stronger rules force the computer to work harder, the weaker rules are applied first. The formulation of the set of strategic rules—and the translation of these rules of logic into a language with which a digital computer can operate—has been one of the most challenging aspects of this project.

Essentially the program tells the computer to perform the following tasks:

1. Sort the fragments (inputs) according to first (N-terminal) symbol and chain length.
2. Construct a binary matrix to represent inclusion relationships within each set of fragments having the same first symbol.
3. Evaluate and store in memory the maximal sets of fragments with distinct first symbols.
4. Compute partial sequences within fragments from degree of intersection of fragments with identical first symbols.
5. Compute degree of overlapping of all fragments.

Information quality of inputs

The completely programmed computer will accept sets of data of high (e.g., Thr·Pro·Lys·Asp), low (Thr, Pro, Lys, Asp) or intermediate (Thr·Pro·(Lys, Asp)) quality. The number of such sets of data which will be required to solve a given sequence will depend in part upon their quality and in part upon their relationship to sites along the original sequence. It may be impossible to find the sequence of a molecule if only low quality data is used. Thus the sequence ABABAB can only be determined to the extent of knowing AB(AB)AB, i.e, the order of AB within the brackets is not known. This will hold even if we have all possible fragments of the form $L(X_1 \cdots X_n)$ (where L is the leftmost member and $(X_1 \cdots X_n)$ is the composition).

The number of fragments required to obtain the sequence of a molecule "strongly" (i.e., leaving only small pockets of ambiguity) depends in large measure upon the selectivity with which we can obtain the fragments. If the fragments reflect a random selection from the original molecule, then "strong" sequencing of a molecule of length n can generally be obtained with $n \ln n$ fragments. If the set of fragments is randomly distributed and if we can select from this set certain fragments with particular properties, e.g., the set of fragments all of which begin with threonine, then strong sequencing can be accomplished with n fragments (where n is the count on the fragments selected from the random set). Chemically, this process could be exemplified by an indifferent fission followed by an analysis of those fragments possessing a particular chromatographic property.

Finding sequencing with low quality inputs

Of particular interest is the use of the computer to determine sequences in cases in which the *average* quality of the input data is significantly lower than has been used to determine the presently known protein sequences. Low quality data which can be obtained in reasonably large quantity with minimal expenditure of highly skilled human effort is (1) the number of times an element occurs in the protein or fragment and (2) the N-terminal element.

The first question to be asked of the computer is whether the sequence of a protein can in fact be found from a sufficiently complete set of such low quality data. Although low quality data has been utilized to varying degrees in the determination of most protein sequences, no protein sequence has been determined using only such data. To investigate the feasibility of such a technique we have carried out a series of "model experiments" with hypothetical data of this nature. These model experiments can be carried out by hand for a short sequence, using the same set of rules which are programmed into the computer.

Sequence analysis of insulin fragment

We assume a knowledge of the composition of the protein, that is, a knowledge of the kinds and frequencies of acids in the molecule. Further, when the molecule is broken down into fragments, we know the composition of the fragments. Finally, we assume that the N-terminal acid of the molecule and of each fragment is known.

The hypothetical molecule we are to reconstruct, P_{12}, has the composition of a part of the insulin molecule described above, namely, Cys·(Arg, Phe, Phe, Thr, Gly, Gly, Ala, Glu, Pro, Lys, Tyr) further abbreviated by 0·(1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9).

Fragments corresponding to those produced by a hypothetical "indifferent fission" hydrolysis experiment were obtained by using a random number generator to decide at which bond fission would occur. The resulting single amino acids and dipeptides were discarded. The remaining fragments are listed in Table 1. In a typical run during which the probability of bond breakage was 0.3, 28% of the bonds were actually broken and the median peptide length was 4.2, or 35% of the total fragment length.

Table 1 **Fragments remaining in hypothetical molecule after simulated fission occurred and single amino acids and dipeptides were discarded.**

1. 6(1224)	7. 6(14)	13. 4(146)
2. 0(46)	8. 3(578)	14. 4(229)
3. 2(2379)	9. 9(378)	15. 2(3789)
4. 7(58)	10. 1(224)	16. 0(146)
5. 1(223479)	11. 0(1446)	17. 1(2249)
6. 4(22)	12. 2(239)	18. 4(12246)

A typical procedure for determining the sequence of the fragments of Table 1 *by computer methods* is as follows:

A. The fragments are sorted and arranged according to the *known left-hand symbols*:

0(46)	1(223479)	2(2379)	3(578)
0(1446)	1(224)	2(239)	
0(146)	1(2249)	2(3789)	
4(22)	6(1224)	7(58)	9(378)
4(146)	6(14)		
4(229)			
4(12246)			

B. Deductions (→) from the initial information based upon the identity of left-hand symbols are now possible from the composition data, viz., 0(1, 2, 2, 3, 4, 4, 5, 6, 7, 8, 9): let n = total number of occurrences of the particular left-hand symbol in the entire sequence.

○ $n = 1$
0(46)
0(146)
0(46)1
0(1446)
0(46)14

① $n = 1$
1(224)
1(2249)
1(224)9
1(223479)
1(224)9(37)

② $n = 2$
2(239)
2(2379)
2(239)7

2(3789) → **2(3789)**

③ $n = 1$
3(578) → **3(578)**

④ $n = 2$
4(22)
4(229)
4229

4(146)
4(12246)
4(146)22

⑤ $n = 1$
no fragments

⑥ $n = 1$
6(14)
6(1224)
6(14)22

⑦ $n = 1$
7(58) → **7(58)**

⑧ $n = 1$
no fragments

⑨ $n = 1$
9(378) → **9(378)**

C. The total number of deductions in *B* are summarized. This now represents the total information up to this point, viz.,

0(46)14	**4229**
1(224)9(37)	**4(146)22**
2(239)7	**6(14)22**
2(3789)	**7(58)**
3(578)	**9(378)**

D. Left-hand symbols, from *C*, are now systematically identified with internal symbols (bracketed).

Symbol ○ Fragments — No deductions

0(46)14

Symbol ① Fragments containing "1" — Deductions:

0(146)14 — 0(46)14

1(224)9(37) — 1(224)9(37) → 0(46)14229(37)

4(146)22 — 4(146)22 → 04614229(37)

6(14)22 — 6(14)22 → **04614229(37)**

We now have **04614229(37)** plus unused fragments:

2(239)7 4229
2(3789) 7(58)
3(578) 9(378)

Symbol ② Fragments containing "2" — Deductions:

04614229(37)

2(239)7 — 2(239)7 → 0461422937

2(3789) — 2(3789) → 04614229378

4229 — 4229 → **04614229378**

We now have **04614229378** plus unused fragments:

3(578)
7(58)
9(378)

Symbol ③ Fragments containing "3" — Continuing:

3(578) — 04614229378

3(578) → 046142293785

9(378) — 9(378) → **046142293785**

We now have **046142293785** plus unused fragment:

7(58)

Symbol ⑦ Fragment containing "7" — Continuing:

046142293785

7(58) — 7(58) → **046142293785**

All of the information which was supplied in the original set of fragments (Table 1) is now included in the final deduction. Some of the deductions were vacuous. They nevertheless have been carried out to guarantee that the information carried in fragments 7(58), 9(378), 6(14)22 and 4229 is used. These fragments hence confirm the unambiguous sequence **046142293785.**

This example, using a chain length of 12, has been chosen to illustrate the procedure which the computer follows. The particular example could be and was worked out by hand. For longer chains difficulty in sorting by distinct left-hand symbols and the number of deductions to be made increases very rapidly. At relatively short chain lengths these two procedures outstrip human capabilities and can satisfactorily be worked through only by machine methods. Recently a punch card method for facilitating the deduction process by hand has been reported.[40]

The number of peptides required to give sufficient overlap to completely determine the sequence in any particular case is a variable. The entire sequence in the example was determinable with only eight peptides selected from the randomly generated sample of 18. There is some probability that the sequence would not have been determinable even if all 18 were used.

We have carried out similar experiments with a random degradation (p breakage = 0.2) of ribonuclease, containing 124 amino acids. The complete sequence was unambiguously determined using 300 fragments in the first model experiment and 250 fragments in a second experiment. The average number of fragments required (determined from many model experiments) has yet to be ascertained for this protein.

There are a number of points to be established by studies of this type: (1) the average number of fragments required for a given protein chain length, (2) the optimum size distribution of fragments, (3) the relative gain in information by knowing in addition the penultimate left-most element or the carboxyl-terminal amino acid of a particular fragment, (4) the relation between the degree of ambiguity at any time and the number of fragments used up to that point, (5) the relative information values of specific (e.g., enzymatic) degradation of the sequence vs less or nonspecific degradations, (6) the effects of errors in chemical analyses and techniques for their detection and correction.

Residual ambiguity and partial sequences

The question of "residual ambiguity" deserves particular attention. It may not be necessary in a given application to know the *complete* sequence for the entire protein. Perhaps only the sequence around a particular active site is needed. It is definitely advantageous not to require complete sequences since ambiguity decreases rapidly at first but only very slowly near the end, the removal of the last bits of ambiguity requiring rather large numbers of pieces of data (peptide fragments).

An important feature of the proposed method would be that an unambiguously determined sequence for a particular protein programmed into the computer could be used to locate the *position* of the change in sequence in a mutant of the protein.

Application to polynucleotides

Although the method has been described in terms of proteins it is, in principle, equally applicable to the determination of nucleic acid sequences. However, a much greater number of fragments will be required because (1) the nucleic acid alphabet has only four letters and (2) nucleic acid sequences are, in general, much longer. However, this difficulty may be obviated in the future by either:

(1) Selective fragmentation of nucleic acids to smaller units which still contain demonstrable genetic information,[12,41] or

(2) The purification of shorter nucleic acid sequences, some of which are the genetic-biochemical transducers in the synthesis of single proteins.[43,45]

The recent development of computer programs to analyze the composition of oligonucleotides from ultraviolet spectra of their hydrolysates[46] may accelerate progress in this field considerably.

Summary

This paper describes preliminary results indicating the way in which complex mathematical analysis can be utilized to determine the sequences of amino acids in large proteins. A new, logical system for sequencing proteins by digital computers based on the reassembly of nonspecifically fragmented protein subchains (peptides) is described. The method is particularly well adapted to use low quality data such as (1) the overall amino acid composition plus (2) the *N*-terminal element of each peptide. Unique sequence analysis is obtained through the use of a sufficient number of overlapping peptides produced by "indifferent fission." Preliminary studies, using the logical sequence tracing system described, indicate that sequences can be determined with surprisingly high efficiency using *only* such data.

The methods described here have been programmed and tested successfully with randomly generated amino acid sequences to a length of 750 residues. A paper describing these results is in preparation.

References

1. F. H. C. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, "General nature of the genetic code for proteins," *Nature* 192, 1227 (1961).
2. G. Gamow, A. Rich, and M. Ycas, "The problem of information transfer from the nucleic acids to proteins," *Adv. Biol. and Med. Physics IV*, p. 23, Academic Press, 1956, New York.
3. R. G. Hart, "On the distribution of purine and pyrimidine bases in the nucleic acid of tobacco mosaic virus," *Proc. Natl. Acad. Sci. U.S.* **43**, 457 (1957).
4. K. K. Reddi, "The arrangement of purine and pyrimidine nucleotides in TMV RNA," *Proc. Natl. Acad. Sci. U.S.* **45**, 293 (1959).
5. H. Jehle, "Amino acid sequence selection in protein synthesis," *Proc. Natl. Acad. Sci. U.S.* **45**, 1360 (1959).
6. H. S. Shapiro and E. Chargaff, "Studies on the nucleotide arrangement in DNA's IV. Patterns of nucleotide sequence in the DNA of rye germ and its fractions," *Biochim. Biophys. Acta* **39**, 68 (1960).
7. K. Burton and G. B. Petersen, "The frequencies of certain sequences of nucleotides in DNA," *Biochem. J.* **75**, 17 (1960).
8. M. Ycas, "Correlation of viral RNA and protein composition, *Nature* 188, 209 (1960).
9. C. I. Davern, "Bias in base pair orientation in DNA," *Nature* **188**, 209 (1960).
10. F. Lanni, "Analysis of sequence patterns in ribonuclease I. Sequence vectors and vector maps," *Proc. Natl. Acad. Sci. U.S.* **46**, 1563 (1960).
11. R. Simha and J. M. Zimmerman, "Synthesis kinetics and sequence distribution in synthetic polynucleotides," *J. Polymer Sci.* **42**, 309 (1960).
12. N. Sueoka, "Variation and heterogeneity of base composition of DNA's: A compilation of old and new data," *J. Mol. Biol.* **3**, 31 (1961).
13. J. Josse, A. D. Kaiser, and A. Kornberg, "Enzymatic synthesis of DNA VIII. Frequencies of nearest neighbor base sequences in DNA," *J. Biol. Chem.* **236**, 864 (1961).
14. S. B. Weiss and T. Nakamoto, "The enzymatic synthesis of RNA: nearest neighbor base frequencies," *Proc. Natl. Acad. Sci. U.S.* **47**, 1400 (1961).
15. C. R. Woese, "Coding ratio for the RNA viruses," *Nature* **190**, 697 (1961).
16. C. R. Woese, "Non-random occurrence of amino acid replacements," *Nature* **191**, 1196 (1961).
17. I. Leslie, "Biochemistry of heredity: a general hypothesis," *Nature* **189**, 4761 (1961).
18. M. W. Nirenberg, J. H. Matthai, and O. W. Jones, "An intermediate in the biosynthesis of polyphenylalanine directed by synthetic template RNA," *Proc. Natl. Acad. Sci. U.S.* **48**, 104 (1962).
19. J. H. Mattai and M. W. Nirenberg, "Characteristics and stabilization of DNA-ase-sensitive protein synthesis in *E. coli.* extracts" *Proc. Natl. Acad. Sci. U. S.* **47**, 1580 (1961); The dependence of cell-free protein synthesis in *E. coli.* upon naturally occurring or synthetic polyribonucleotides, *ibid.* **47**, 1588 (1961).
20. J. Speyer, P. Lengyel, C. Basilio, and S. Ochoa, "Synthetic polynucleotides and the amino acid code II." *Proc. Natl. Acad. Sci. U. S.* **48**, 63 (1962).
21. S. Brenner, L. Barnett, F. H. C. Crick, and A. Orgel, "The theory of mutagenesis," *J. Mol. Biol.* **3**, 121 (1961).
22. T. Alderson, "Mechanism of formaldehyde-induced mutagenesis. The uniqueness of adenylic acid in the mediation of mutagenic activity of formaldehyde," *Nature* **187**, 485 (1960).

23. H. M. Dintzis, "Assembly of the peptide chains of hemoglobin," *Proc. Natl. Acad. Sci. U. S.* **47**, 247 (1961).
24. D. M. Prescott, and R. F. Kimball, "Relation between RNA, DNA, and protein synthesis in the replicating nucleus of euplotes," *Proc. Natl. Acad. Sci. U. S.* **47**, 686 (1961).
25. A. M. Michelson, "A hypothesis for the biosynthesis of RNA and protein," *Nature* **181**, 375 (1958).
26. E. P. Geiduschek, T. Nakamoto, and S. B. Weiss, "The enzymatic synthesis of RNA: complementary interaction with DNA," *Proc. Natl. Acad. Sci. U. S.* **47**, 1405 (1961).
27. A. Rich, "A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids," *Proc. Natl. Acad Sci. U. S.* **46**, 1044 (1960).
28. J. A. Cohen, *J. Cell. Comp. Physiol.* Supplement 1, Dec. 1959.
29. F. Sanger, "Chemistry of insulin," *Science* **129**, 1340 (1959).
30. A. Tsugita, D. T. Gish, J. Young, H. Fraenkel-Conrat, C. A. Knight, and N. M. Stanley, "The complete amino acid sequence of tobacco mosaic virus," *Proc. Natl. Acad. Sci. U. S.* **46**, 1463 (1960).
31. C. H. W. Hirs, S. Moore, and W. H. Stein, *J. Biol. Chem.* **235**, 633 (1960).
32. D. H. Spackman, W. H. Stein, and S. Moore, *J. Biol. Chem.* **235**, 648 (1960).
33. "The amino acid sequence of sperm whale myoglobin." A. B. Edmundson and C. H. W. Hirs, "Chemical studies," *Nature* **190**, 663 (1961); J. C. Kendrew, H. C. Watson, B. E. Strandberg, R. E. Dickerson, D. C. Phillips, and V. C. Shore, "A partial determination by X-ray methods and its correlation with chemical data," *ibid.* **190**, 666 (1961).
34. D. T. Gish, "Studies on the amino acid sequence of tobacco mosaic virus protein IV. The amino acid sequences of an eicosa peptide and a heptadecapeptide isolated from a tryptic digest of TMV protein," *J. Am. Chem. Soc.* **83**, 3303 (1961).
35. G. Braunitzer, N. Hilschmann, K. Rudloff, B. Hilse, B. Liebold, and R. Muller, "The haemoglobin particles: chemical and genetic aspects of their structure," *Nature* **190**, 480 (1961).
36. F. Sorm, B. Keil, V. Holeysovsky, B. Meloun, O. Mikes, and J. Vanecek, "On proteins XLIX. Comparison of the microstructures of chymotrypsinogen and trypsinogen," *Coll. Czech. Chem. Comm.* **23**, 935 (1958).
37. B. Keil, F. Sorm, V. Holeysovsky, V. Kostka, B. Meloun, O. Mikes, V. Tomasek, and J. Vanecek, "On proteins LVI. On the partial structures of bovine chymotrypsinogen and trypsinogen," *Coll. Czech. Chem. Comm.* **24**, 3491 (1959).
38. P. Edman, *Acta Chem. Scand.* **4**, 277 (1950).
39. H. Fraenkel-Conrat, J. I. Harris, A. C. Levy, in *Methods of Biochemical Analysis*, Vol. II., *Ed.* D. Glick, Interscience Publishers, N. Y., 1954, p. 359.
40. R. V. Eck, "A simplified strategy for sequence analysis of large proteins," *Nature* **193**, 241-243 (1962).
41. G. Bernardi and C. Sadron, "Kinetics of the enzymatic degradation of DNA into subunits," *Nature* **191**, 809 (1961).
42. E. Otaka, Y. Oota, and S. Osawa, "Sub-unit of ribosomal RNA from yeast," *Nature* **191**, 598 (1961).
43. A. Bendich, H. B. Pahl, G. C. Horngold, H. S. Rosencranz, and J. R. Fresco, "Fractionation of DNAs on columns of anion exchangers," *J. Am. Chem. Soc.*, **80**, 3949 (1958).
44. B. P. Doctor, J. Apgar, and R. W. Holley, "Fractionation of yeast amino acid acceptor RNAs by countercurrent distribution," *J. Biol. Chem.* **236**, 1117 (1961).
45. H. von Portatius, P. Doty, and M. L. Stephenson, "Separation of L-valine acceptor 'soluble RNA' by specific reaction with polyacrylic acid hydrazide," *J. Am. Chem. Soc.* **83**, 3351 (1961).
46. J. C. Reid and A. W. Pratt, "Vector analysis of ultraviolet mixture spectra: the composition of RNA," *Biochem. Biophys. Res. Comm.* **3**, 337 (1960).
47. S. A. Bernhard, D. F. Bradley, and W. L. Duda, "The Logical Sequencing of Amino Acids, Parts I and II," *Abstracts of IBM Medical Symposium.* Endicott, N. Y., September, 1960.

Received July 5, 1962