Journal of Macromolecular Science, Part A
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597274

Chemical Evolution of Protein Sequences

Isadore Nicholson[a]
[a] Chemistry Department, C. W. Post College Long Island University Brookville, New York

To cite this Article Nicholson, Isadore(1970) 'Chemical Evolution of Protein Sequences', Journal of Macromolecular Science, Part A, 4: 7, 1619 — 1625
To link to this Article: DOI: 10.1080/00222337008069372
URL: http://dx.doi.org/10.1080/00222337008069372

J. MACROMOL. SCI.–CHEM., A4(7), pp. 1619-1625, November, 1970

Chemical Evolution of Protein Sequences

ISADORE NICHOLSON

Chemistry Department
C. W. Post College
Long Island University
Brookville, New York 11548

SUMMARY

An estimation of the hydrolytic stability of naturally occurring peptides and homologous sequences within protein chains has been made using Newman's "Rule of Six." All possible sequences and their corresponding cumulative steric values, "six numbers," were computer-generated. Those sequences having identical six numbers were grouped as families. The results of this study suggest that, in abiogenesis, peptides and proteins may have most often simply evolved to sequences which are members of the permutationally more populated sequence families.

With a knowledge of which amino acids and how many of each are present in peptide molecules (data readily obtained today using commercial amino acid analyzers), I have found that a simple molecular parameter can lead one to a family of likely sequences for that peptide.

The results have implications with regard to the problem of the chemical evolution of proteins and the origin of life.

The hydrolysis of the backbone peptide bond has been found [1-3] to correlate with steric effects as expressed in Newman's Rule of Six [4] which states "in reactions involving addition to an unsaturated function containing a double bond, the greater the number of atoms in the six position, the greater will be the steric effects." A "six number," termed (sn), is used to denote the number of atoms in the six position, as in Fig. 1. Thus, for dipeptides one may calculate the cumulative sn value,

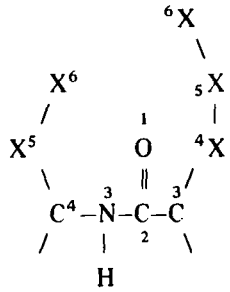

Fig. 1. Atoms denoted by X comprise sidechains of the peptide. Where the X6 atoms are present and multivalent, the sn value will be greater than 1.

i.e., alanyl-glycine has an sn value of 0 whereas glycyl-alanine has a corresponding value of 3.

Obviously, the sn values are noncommutative and vary with the sequence.

The trend of hydrolysis of many dipeptides and also polyaminoacids (i.e., homopolymers) has been found to be in excellent agreement with the sn value of the structures involved.

Whitfield [5] has prepared a matrix of data that contains the sn values for the dipeptide links for all of the commonly occurring amino acids except asparagine and glutamine.

If, in fact, proteins formed from amino acids in the primeaval oceans or on land by the agency of vulcanic heat [6], lightning [7], or ultraviolet light [8] underwent a chemical type of Darwinian selection, then the surviving proteins of today are of the most stable sequences; namely, the least hydrolyzable, considering the available amino acid starting mixture.

I have used a computer (IBM-1130) to calculate all possible sequence permutations along with the cumulative sn value for each sequence.

Since the computer working time increases very rapidly as a function of the number of amino acids programmed (see Table 1), polypeptides composed of over nine amino-acids were not investigated.

Cyclic oligopeptides were not included in this study since steric interactions due to ring structures are not explicable in terms of Newman numbers. Unhappily, this restriction excludes the ocytocin and vasopressin group.

As an example, a biologically active normally biosynthesized (as opposed to an in vitro hydrolytic peptide fragment) peptide hormone angiotensin II was analyzed as to the size of families of sequences having identical cumulative sn values, see Fig. 2.

Calculated sn values		Number of sequences having sn values
44		1,440
47		14,400
48		2,880
49		None
50		4,320
51		17,280
Total number of sequences	=	40,320

Fig. 2. Angiotensin II (horse).
Known sequence [9] Asp-Arg-Val-Tyr-Ileu-His-Pro-Phe
calculated $\Sigma sn = 5 + 6 + 9 + 5 + 9 + 8 + 9 = 51$

Table 1

No. of amino acids in peptide	No. of possible sequences
4	24
5	120
6	720
7	5040
8	40320
9	362280

Thus, angiotensin II falls into the category of the most populated family which is also the hydrolytically most stable family with a $\Sigma sn = 51$.

One may thus suggest that in this case, genetic stability is derived from chemical stability.

Two classes of peptide sequences were investigated: 1) homologous sequences within protein chains, and 2) oligopeptides containing no ring due to disulfide or peptide linking. The results are indicated in Table 2.

In no case tested, did the naturally occurring sequence fall into the least populated sequence family. Almost as rarely did the natural sequence fall into the most populated sequence family. The naturally occurring sequences have been found in the permutationally more populated sequence families.

Table 2

Sequence	Reference	Name	Σsn of the most populated families	Σsn of the family naturally occurring sequence	Sequence having maximum Σsn
arg-pro-pro-gly-phe-ser-pro-phe-arg	10	Bradykinin (kallidin I) (9 amino acids, 10 families)	53, 56	56	61
See Fig. 2	9	Angiotensin (8 amino acids, 5 families)	47, 51	51	51
val-thr-ala-ala-his-cys-gly-val	11	Chymotrysin (8 amino acids 9 families)	37, 31	31	40
val-thr-ala-gly-his	11	α-lytic protease and liver aldolase homolog (5 amino acids, 8 families)	21, 19 = 17 = 15	19	24
asn-gly-thr-ser-met-ala	12	Subtilisin homolog (6-amino acids, 8 families)	22, 21	22	25

asp-gly-ser-asp-pro	13	Fibrinopeptide homolog (5 amino acids, 9 families)	21, 23 = 17	17	25
gly-asp-ser-gly-gly	14	Chymotrypsin (bovine), trypsin (bovine), Trypsin (pig) (5 amino acids, 6 families). Homologous active site	9, 6	9	9
cys-lys-pro-val-asn-thr-phe-val-his-glu	15	Common sequence of rat, bovine, horse ribonuclease (10 amino acids, 10 families)	59, 63, 62	62	65
phe-pro-val-asp-asp-asp-asp-lys	7	Trypsinogen activation peptide, sheep 1 (amino acids, 5 families)	47, 43, 46	46	47
val-thr-pro-gly-his-ala-cys-thr	17	Aldolase (rabbit, muscle), Active site (8 amino acids, 11 families)	44, 38, 40	40	46

One may hypothesize that the original sequences formed were dictated by probability and that chemical evolution involved a reshuffling of amino acids to more and more hydrolytically stable peptides (i.e., peptides whose cumulative Newman number is maximal).

ACKNOWLEDGMENT

I wish to thank Mr. George Lipka of the C. W. Post Computer Center for programming services.

REFERENCES

[1] R. L. M. Synge, *Biochem. J.*, **39**, 351 (1945).
[2] K. Heyns, W. Walter, and H. F. Grützmacher, *J. Polym. Sci.*, **30**, 573 (1958).
[3] R. E. Whitfield, *Science*, **142**, 577 (1963).
[4] M. S. Newman, *Steric Effects in Organic Chemistry*, Wiley, New York, 1956, Chapter 4.
[5] R. E. Whitfield and W. L. Wasley, in *Chemical Reactions of Polymers*, (E. M. Fettes, ed.), Wiley-Interscience, New York, 1964, pp. 425.
[6] S. W. Fox, T. Waehneldt, C. R. Windsor, and J. Ryan, *Second Annual Report*, Institute of Molecular Evolution (Univ. Miami), Coral Gables, Florida, 1966.
[7] C. Ponnamperuma and J. Flores, Abstracts, 152nd National Meeting of the American Chemical Society, New York, Sept. 1966.
[8] C. Ponnamperuma and E. Peterson, *Science*, **147**, 1572 (1965).
[9] L. T. Skeggs, K. Lentz, J. Kahn, and N. P. Shumway, *J. Exp. Med.*, **106**, 439 (1957).
[10] D. F. Elliot, G. P. Lewis, and E. W. Horton, *Biochem. Biophys. Res. Commun.*, **3**, 87 (1960).
[11] D. E. Morse and B. L. Horecker, *Science*, **161**, 813 (1968).
[12] E. L. Smith, F. S. Markland, et al., *J. Biol. Chem.*, **241**, 5974 (1966); H. F. Noller and S. A. Bernhard, *Biochemistry*, **4**, 1118 (1965).
[13] M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Spring, Md., 1969, pp. D-70.
[14] Ibid, pp. 49.
[15] Ibid, pp. D-226.

[16] Ibid, pp. D-116.
[17] C. Y. Lai, P. Hoffee, and B. L. Horecker, *Arch. Biochem. Biophys.*, **112**. 567 (1965).

Accepted by editor January 26, 1970
Received for publication March 16, 1970