VOL4 NO. 1 March 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

IUS Soft Computing Research Group

Conformational Parameters for Amino Acids in Helical, ß-Sheet, and Random Coil Regions Calculated from Proteins: After 40 Years

Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba

Article Info

Article history:
Article received on February 2015
Received in revised form March 2015

Keywords:
Protein structural classes, Secondary structure; Conformation of proteins; Statistical methods

Abstract

Forty years ago, Peter Y. Chou and Gerald D. Fasman (1974a), relying on the information from fifteen proteins calculated α helix, β-sheet, and coil conformational parameters, P_α,P_β,and P_c,for the 20 naturally occurring amino acids from the frequency of occurrence of each amino acid residue in the α, β, and coil conformations. Secondary structure of these 15 proteins had been determined by X-ray crystallography. Although the accuracy could not go over to the level of 60% too much, these values utilized for a long time to provide a simple procedure, devoid of complex computer calculations, to predict the secondary structure of proteins from their known amino acid sequences. In the same article of Peter Y. Chou and Gerald D. Fasman, a detailed analysis of the helix and β-sheet boundary residues in proteins provided amino acid frequencies at the N- and C-terminal ends which were used to delineate helical and β regions. Charged residues are found with the greatest frequency at both helical ends, but they were mostly absent in β-sheet regions. In the same article a mechanism of protein folding was proposed, whereby helix nucleation starts at the centers of the helix where the P_α values are highest, and propagates in both directions, until strong helix breakers where P_α values are lowest, terminate the growth at both ends. Similarly, residues with the highest P_β values will initiate β regions and residues with the lowest P_β values will terminate β regions. The helical region with the largest P_α was proposed as the site of the first fold during protein renaturation. The mechanism whereby proteins fold into their native conformation, capable of biological activity, has been a long sought after goal. With the elucidation of the three-dimensional structure of many proteins through X-ray crystallography, a new momentum has been given to understanding the factors governing this complex assembly of polypeptide chains. In this paper, using similar statistics from 20 347 proteins, the level of reliability of formerly found results is discussed.

1. INTRODUCTION

The impetus for the prediction of protein conformation was initiated with studies on α-amino acids. Helix formation in α-amino acids is characterized by a cooperative process (Zimm and Bragg, 1959; Applequist, 1963) in which the Zimm-Bragg parameters σ and s are defined respectively as the cooperativity factor for helix initiation, and the equilibrium constant for converting a coil residue to a helical state at the end of a long helical sequence. Potentiometric titration data on poly (L-glutamic acid) (Nagasawa and Holtzer, 1964) and poly(L-lysine) (Hermans, 1966a) gave s values of 1.25 and 1.15, respectively, in aqueous solution at 25°C. Since s = 1 corresponds to the critical value at which long chains substantially convert into the helical form (Zimm and Bragg, 1959), the greater than unity s values for poly(G1u) and poly(Lys) (both in the un-ionized form) indicate that these two homopolymers form stable helices; this has been confirmed by optical studies.

An analysis of all 20 amino acids in 15 proteins was presented in Chou and Fasman (1974a), whereby the frequency of their occurrence in various conformational states is compared with the experimental Zimm-Bragg σ and s parameters. Boundary residues of helical and β-sheet sections were analyzed, and yield clues for the termination of these conformational regions. The helix and β-conformational parameters provided a quantitative measure of regions in proteins with the highest helical and β-sheet potential, and may be useful in understanding protein folding mechanisms. These parameters had been used with limited success in predicting protein secondary conformation from known amino acid sequences along years that follow.

2.METHOD

Secondary structures of proteins are obtained in the form of the x-ray analyses in three conformations helix "h", sheet "s" , and others ".". Others are interpreted as coils "c".

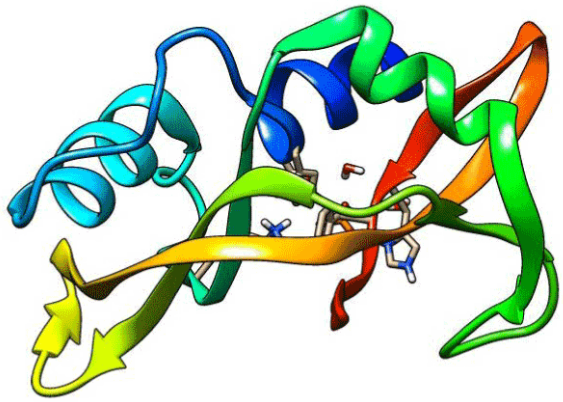

Figure 1. α-helices, β-sheets, and coils on the same picture

Symbols for Amino Acids

Proteins are chains in the three dimensional space built from smaller chemical molecules called amino acids. Thereare 20 different amino acids. Each of them is denoted by a different letter in the Latin alphabet as shown below.

TABLE1. Names and symbols of 20 amino acids

#	Amino acid	Chemical	alphabet
1	Alanine	Ala	A
2	Arginine	Arg	R
3	Asparagine	Asn	N
4	Aspartic acid	Asp	D
5	Cysteine	Cys	C
6	Glutamine	Gln	Q
7	Glutamic acid	Glu	E
8	Glycine	Gly	G
9	Histidine	His	H
10	Isoleucine	Ile	I
11	Leucine	Leu	L
12	Lysine	Lys	K
13	Methionine	Met	M
14	Phenylalanine	Phe	F
15	Proline	Pro	P
16	Serine	Ser	S
17	Threonine	Thr	T
18	Tryptophan	Trp	W
19	Tyrosine	Tyr	Y
20	Valine	Val	V

Based on the protein chain it is easy to create its relevant sequence of amino acids replacing an amino acid in chainby its code in Latin alphabet. As a result a word on the amino acids' alphabet is received. This word can be calleda protein primary structureon the condition that letters in this word are in the same order as amino acids in theprotein chain are.

A secondary structure of a protein is a subsequence of amino acids coming from the relevant protein. Thesesub chains form in the three dimensional space regular structures which are the same in shape for differentproteins. In the analysis, a similar representation for the secondary structures as for the primary oneshas been used. A secondary structure is represented by a word on the relevant alphabet of secondary structures – eachkind of a secondary structure has its own unique letterα-helix, H; β-sheetS, and coil C. An alphabet of secondary structures consisting of threedifferent secondary structures has been considered in the analysis.

Pioneering Research

Chou and Fasman (1974a) made a survey on the following 15 proteins whose amino acid sequences and conformations via X-ray crystallography are known: carboxypeptidase A (Quiocho and Lipscomb, 1971), a-chymotrypsin (Blow, 1969), cytochrome 115 (Mathews er al., 1972), elastase (Shotton and Watson, 1970), ferricytochrome c (Dickerson et al., 1971), a and P-hemoglobin

(Perutz et of., 1968), insulin (Blundell et al., 1972), lysozyme (Blake et al., 1967), myogen (Nockolds et al., 1972), myoglobin (Kendrew et al., 1961), papain (Drenth et al., 1971), ribonuclease S (Wyckoff et al., 1970), staphylococcal nuclease (Arnone et al., 1971), and subtilisin BPN (Wright et al., 1969).

Helix, β-sheet, and coil regions

The amino acid residues in the helix, β-sheet, and coil regions of these 15 proteins were tabulated in Table 2. (TABLE 2 of Chou and Fasman (1974a)). It should be noted that the β-sheet residues differ slightly from those reported earlier (Chou and Fasman, 1973) due to more recent detailed X-ray diffraction analysis references cited above. Where the β regions were not specified explicitly in the original papers, as in the case of chymotrypsin, elastase, and ribonuclease S, the schematic diagrams showing hydrogen bonding in these proteins were used to delineate the β-sheet regions. Residues at β bends which did not show hydrogen bonding were not included in the β-sheet regions. All regions designated as helical (alicyll, 310, distorted helix) by the X-ray crystallographic studies have been included as helical residues in Table 2. Despite minor changes in some helical regions, based on the latest X-ray studies denoted in a following paper, the analysis did not significantly alter the calculated helix parameters so that the helical residues listed in Table 2 are identical with those reported earlier (Chou and Fasman, 1973).

TABLE 2: Amino acid residues in the helix, inner helix, β-sheet, and coil regions of 15 proteins in Chou and Fasman, 1974a.

A. Acid	AA	H	In H	In β	In Coil
Ala	228	119	62	38	71
Arg	78	22	9	12	44
Asn	133	35	12	15	83
ASP	111	39	10	15	57
CYS	54	15	3	12	27
Gin	95	40	16	20	35
Glu	113	62	28	5	46
GlY	232	45	22	32	155
His	74	33	11	9	32
Ile	106	38	22	29	39
Leu	196	94	64	41	61
LYS	175	67	34	22	86
Met	28	12	6	8	8
Phe	82	33	16	18	31
Pro	85	18	0	9	58
Ser	202	57	24	25	120
Thr	156	47	21	32	77
TrP	44	18	10	9	17
Tyr	100	22	10	22	56
Val	181	74	44	51	56
Total	2473	890	424	424	1159

Before the time of Chou and Fasman, some surveys of helical and nonhelical residues had been performed, they were based on even fewer proteins (Cook, 1967; Kotelchuck et al., 1963; Ptitsyn, 1969; Finkelstein and Ptitsyn, 1971; Kabat and Wu, 1973a,b; Wu and Kabat, 1971, 1973). Hence, the analysis using 15 proteins containing 2473 residues should had more statistical reliability than earlier literature data. In Chou and Fasman 1974a in addition, the β-sheet regions of proteins as well as the helical and β-sheet boundary regions had been analyzed in greater detail. In Chou and Fasman 1974b, authors further improved their analysis.

3. NORMALIZATION PROCEDURE

In Chou and Fasman, 1974a, a normalization procedure had also been used to derive helix, β, and coil conformational parameters which can be used in predicting protein conformation. In this article the probability of finding the j amino acid residue in a protein is given by

$$p_j = \frac{n_j}{\sum n_j} \quad (1)$$

wheren_j is the number of j residues in proteins, and $\sum n_j$ is the total number of residues in proteins. The probability of finding the j residue in the conformational state, k, of proteins is

$$p_{j,k} = \frac{n_{j,k}}{\sum n_{j,k}} \quad (2)$$

where$n_{j,k}$ is the number of j residues in the k state and $\sum n_{j,k}$ is the total number of residues in the k state. The probability or frequency of occurrence of the j residue in the k state of proteins is

$$f_{j,k} = \frac{n_{j,k}}{n_j} \quad (3)$$

Therefore, the average frequency of finding the 20 amino acid residues in the k state of proteins, $\langle f_k \rangle$ can be written as

$$\langle f_k \rangle = \frac{\sum f_{j,k}}{\sum j} = \frac{\sum n_{j,k}}{\sum n_j} \quad (4)$$

where $\sum j = 20$. When f, is normalized by ,$\langle f_k \rangle$the conformational parameter $P_{j,k}$of the j amino acid residue is

$$P_{j,k} = \frac{f_{j,k}}{\langle f_k \rangle} \quad (5)$$

Substituting eq 1-4 in (5) results in

$$P_{j,k} = \frac{p_{j,k}}{p_j} \quad (6)$$

Hence the conformational parameter of the j amino acid residue in the k state, $P_{j,k}$is equal to the probability of finding the j residue in the k state divided by the probability of finding the j residue in proteins. The k conformational state in proteins is either the α, β, or

random coil. Throughout the text the subscript j is omitted to aid in clarity.

4. NORMALIZATION RESULTS

To make a comparison with today's available data, we simplified the procedure and limited ourselves only amino acid residues in the helix, β-sheet, and coil regions in Table 2 of 15 proteins. When we apply the normalization in the above, Table 2 columns are replaced by the columns of Table 3.

Table 3: Normalized amino acid residues in the helix, β-sheet, and coil regions of 15 proteins in Chou and Fasman, 1974a.

"o"	"H"	"E"	"C"
"A"	0.52	0.17	0.31
"R"	0.28	0.15	0.56
"N"	0.26	0.11	0.62
"D"	0.35	0.14	0.51
"C"	0.28	0.22	0.50
"Q"	0.42	0.21	0.37
"H"	0.55	0.04	0.41
"G"	0.19	0.14	0.67
"E"	0.45	0.12	0.43
"I"	0.36	0.27	0.37
"L"	0.48	0.21	0.31
"K"	0.38	0.13	0.49
"M"	0.43	0.29	0.29
"F"	0.40	0.22	0.38
"P"	0.21	0.11	0.68
"S"	0.28	0.12	0.59
"T"	0.30	0.21	0.49
"W"	0.41	0.20	0.39
"Y"	0.22	0.22	0.56
"V"	0.41	0.28	0.31

5. RESULTS AND DISCUSSION

In Chou and Fasman, 1974a in addition to the helical regions, the β-sheet regions of proteins as well as the helical and β-sheet boundary regions had been analyzed in greater detail using 15 proteins containing 2473 residues. We wondered how today abundant data has changed these result. For comparison we only choose amino acid residues in the helix, β-sheet, and coil regions.

We visited several databases and collected data for 4 341 108 residues which reside in 20 347 non redundant proteins. Normalized results for amino acid residues in the helix, β-sheet, and coil regions of 20347 proteins are tabulated in Table 4.

Table 4: Normalized amino acid residues in the helix, β-sheet, and coil regions of 20 347 proteins from several non redundant databases.

"o"	"H"	"E"	"C"
"A"	0.51	0.16	0.33
"R"	0.44	0.19	0.37
"N"	0.29	0.12	0.58
"D"	0.33	0.11	0.56
"C"	0.32	0.28	0.40
"Q"	0.47	0.15	0.38
"H"	0.30	0.19	0.50
"G"	0.17	0.13	0.70
"E"	0.50	0.14	0.37
"I"	0.39	0.36	0.25
"L"	0.49	0.23	0.28
"K"	0.42	0.16	0.42
"M"	0.43	0.20	0.37
"F"	0.38	0.30	0.32
"P"	0.19	0.09	0.72
"S"	0.29	0.16	0.54
"T"	0.28	0.25	0.47
"W"	0.41	0.27	0.32
"Y"	0.37	0.29	0.33
"V"	0.33	0.40	0.28

Percentage errors made in probabilities of amino acid residues in the helix, β-sheet, and coil regions of proteins are found as in Table 5.

Table 5: Percentage errors made in probabilities of amino acid residues in the helix, β-sheet, and coil regions.

o	H	E	C	Overall
% Error	0.14	0.25	0.16	0.17

These percentage errors would not make any difference if the most preferred states for amino acids were the same in two investigations. In Table 5, the most preferred states are shown in two investigation: cf is for Chou and Fasman, 1974a, and ta for this article. 1 is for helix, 2 for β-sheet, and 3 for coil.

Table 6: The most preferred states for 20 amino acids are shown in two investigation: cf is for Chou and Fasman, 1974a, and ta for this article. 1 is for helix, 2 for β-sheet, and 3 for coil.

o	cf	ta
A	1	1
R	3	1
N	3	3
D	3	3
C	3	3
Q	1	1
H	1	3
G	3	3
E	1	1
I	3	1
L	1	1
K	3	3
M	1	1
F	1	1
P	3	3
S	3	3
T	3	3
W	1	1
Y	3	1
V	1	2

In this table we see that Arginine, Glutamic acid, Isoleucine, Tyrosine, and Valine changed their most preferable states. If we believe in statistics which rely on two thousand times larger data, wrong most preferable states for five amino acids among twenty is something tolerable.

REFERENCES

Applequist, J. (1963), On the helix-coil equilibrium in polypeptides. J. Chem. Phys. 38,934-941.

Arnone, A,, Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Jr., Richardson, D. C., Richardson, J. S., and, Yonath, A. (1971), A high resolution structure of an inhibitor complex of the extracellular nuclease of staphylococcus aurous, J. Biol. Chem. 246, 2302.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V.R. (1967), On the conformation of the hen egg-white lysizyme, Proc. Roy. Soc., B 167,365-377.

Blow, D. M. (1969), The Study of a-Chymotrypsin by X-Ray Diffraction, Biochem. J. 112, 261-268.

Blundell, T. L., Cutfield, J. F.,Dodson, E. J., Dodson, G. G., Hodgkin, D. C., and Mercola, D. A. (1972), The crystal structure of rhombohedral 2 zinc insulin, Cold Spring Harbor Symp. Quant. Biol. 36, 233-241.

Chou, P. Y., and Fasman, G. D.(1973),Structural and functional role of Leu residues in proteins, J. Mol. Biol. 74, 263-281.

Chou, P. Y., and Fasman, G. D. (1974a), Conformational Parameters for Amino Acids in Helical, β-Sheet, and Random Coil Regions Calculated from Proteins, Biochemistry 13, 211-222.

Chou, P. Y., and Fasman, G. D. (1974b),Prediction of Protein Conformation, Biochemistry 13,222-245.

Cook, D. A. (1967),The relation between amino acic sequence and protein conformation, J. Mol. Bid. 29, 167-171.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A., and Margoliash, E. (1971), General features of horse and bonito proteins at 2.8A^{o} resolution, J. Biol. Chem. 246, 1511-1535.

Drenth, J., Jansonius, J. N., Koekoek, R., and Wolthers, B. G. (1971),The structure of papain. Adv Protein Chem. 1971;25:79–115.

Finkelstein, A. V., and Ptitsyn, 0. B. (1971), Statistical analysis of the correlation among amino acid residues in helical, beta-structural and non-regular regions of globular proteins, J. Mol. Biol. 62, 613-624.

Hermans, J. (1966),Experimental Free Energy and Enthalpy of Formation of the α Helix, J. Phys. Chem. 70, 510-515.

Kabat, E. A., and Wu, T. T. (1973a),The influence of nearest neighbor amino acids residues on aspects of secondary structure proteins: attempts to location of a-helices, and b-sheets, Biopolymers 12, 751-774.

Kabat, E. A., and Wu, T. T. (1973b), The influence of nearest neighbor amino acids on the conformation of the middle amino acids of the proteins. Comparison of the predicted and experimental determination of b-sheets in conconavalin A, Proc. Natl. Acad. Sci. U.S. 70,1473-1477.

Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C., and Shore, V. C. (1961),The amino-acid sequence x-ray methods, and its correlation with chemical data, Nature (London) 190, 666-670.

Kotelchuck, D., Dygert, M., and Scheraga, H. A. (1969), The influence of short-range interactions on protein conformation. III. Dipeptide distributions in proteins of known sequence and structure, Proc. Natl. Acad. Sci. U.S. 63,615-622.

Mathews, F. S., Levine, M., and Argos, P. (1972), Three-dimensional Fouriersynthesis of calfliver cytochrome b5 at 2.8 A^o resolution, J. Mol. Biol. 64,449-464.

Nagasawa, M., and Holtzer, A. (1964), The Use of the Debye-Hckel Approximation in the Analysis of Protein Potentiometric Titration Data, J. Amer. Chem. Soc. 86, 531-538.

Nockolds, C. E., Kretsinger, R. H., Coffee, C. J., and Bradshaw, R. A. (1972), Structure of a calcium-binding carp myogen, Proc. Nut. Acad. Sci. U.S. 69,581-584.

Perutz, M. F., Muirhead, H., Cox, J. M., and Goaman, L. C. G. (1968),Three-dimensional Fourier Synthesis of Horse Oxyhaemoglobin at 2.8 Å Resolution: The Atomic Model,Nature(London) 219,131-139.

Ptitsyn, 0.B. (1969), Statistical analysis of the distributionof amino acid residues among helical and non-helical regions in globular proteins, J. Mol. Bid. 42,501-510.

Quiocho, F. A., and Lipscomb, W. N. (1971), Carboxy-peptidase A: a protein and an enzyme. Adv Protein Chem., 25:1–78.

Shotton, D. M., and Watson, H. C. (1970), Three dimensional structure of tosyl-elastase, Nature (London) 225,811-816.

Wright, C. S., Alden, R. A., and Kraut, J. (1969), Structure of subtilisin BMN at 2.5A^o resolution, Nature (London)221,235-242.

Wu, T.T., and Kabat, E. A.(1971),An Attempt to Locate the Non-helical and Permissively Helical Sequences of Proteins: Application to the Variable Regions of Immunoglobulin Light and Heavy Chains, Proc. Natl. Acad. Sci. U.S. 68(7), 1501-1506.

Wu, T. T., and Kabat, E. A. (1973),An attempt to evaluate the influence of neighboring amino acids (n-1) and (n+1) on the backbone conformation of amino acid (n) in proteins. Use in predicting the three-dimensional structure of the polypeptide backbone of other proteins, J. Mol. Bid. 25;75(1):13-31.

Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B., and Richards, F. M. (1970), The three-dimensional structure of ribonuclease-S, J.Bid. Chem. 245, 305-328.

Zimm, B. H., and Bragg, J. K. (1959), Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains, J. Chem. Phys. 31, 526-531.