A STATISTICAL ANALYSIS OF THE CORRELATION OF AMINO ACID RESIDUES IN β-REGIONS OF GLOBULAR PROTEINS*

JAGDEESH BANDEKAR**

(*Molecular Biophysics Unit, Indian Institute of Science, Bangalore* 560012

Received on June 16, 1976 and in revised form on September 1, 1976

ABSTRACT

Starting out with the null hypothesis that the amino acid residues are distributed without any preference for any parts of β-regions, a statistical analysis is performed on the amino acid residues occurring at the N-terminal, C-terminal, and in the inner β-regions, of 21 globular proteins with a total of 3523 residues. The observed β-regions are taken from x-ray data. Statistically reliable differences in the individual amino acid content of β-regions have revealed the following features: Ala and Trp show an above-average tendency while Cys shows a below-average tendency to be at the N-terminal β-regions; Val, Leu, Ile, and Phe prefer the inner β-regions, while Pro, Asn, and Lys show a below-average preference to be in the inner β-regions. Highly hydrophobic residues are found to prefer the inner β-regions, weakly hydrophobic and charged residues prefer the non-β-regions.

The results of this study only partially support the rules due to Chou and Fasman for prediction of protein conformation and hence indicate the need for their modification.

Keywords: Protein conformation, β-structrue, statistical analysis, secondary structure.

1. INTRODUCTION

Ordered parts of proteins and polypeptides are comprised of helical and β-regions. Whereas a great deal of work has been done (*see*, for example, Burgess *et al.* [1]) to characterize the amino acid residues occurring in helical regions, relatively little is known concerning the β-regions of proteins and polypeptides. However, recently some effort seems to be channelled

---

* Contribution No. 80 from the Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.

** Presently in solid state and Structural Chemistry Unit.

in this course. Finkelstein and Ptitsyn [2] reported statistically reliable differences in amino acid content of β-regions: Ile, Leu, Val had above-average preferences to be in β-regions; highly hydrophobic residues showed a below-average tendency to be in β-regions. However, owing to the limited amount of data, namely, 163 β-residues in 9 globular proteins, the analysis was only qualitative. Nagano [3] made a statistical analysis of the structures of 95 proteins of known sequence belonging to 13 families of crystallographically known conformations. His reports were not very detailed: mostly hydrophobic residues were reported to be characteristic of β-regions. It was suggested by him that β-regions initiate from some arrangement of hydrophobic residues rich in Ile, Val, and that the other residues might stack on the core by H-bond formation. Chou and Fasman [4, 5] carried out a survey of fifteen proteins of known primary and secondary structures. These proteins contained 424 β-residues. It was observed that the charged residues were conspicuously absent at the β-sheet boundary regions, nor were they favoured in the central β-regions. Lim [6, 7] gave a stereochemical theory of secondary structure in globular proteins. Ananthanarayanan and Bandekar [8, 9] filled in an important gap in the theory of order-disorder transitions in globular proteins and polypeptides by showing that one-dimensional near-neighbour Ising model type approach is valid in case of β-region prediction also.

In this work, a statistical analysis is performed on a total of 713 β-residue in 21 globular proteins.

2. METHODS

Two residues at each end of a β-region were included in computing frequencies of residues at the N- and C-terminals of a β-region. The remaining residues in that β-region counted as residues in inner β-regions. Two residues on either side of N- and C-terminals and not in β-regions were grouped into non-β-regions. This arbitrary criterion of including two end-residues will be discussed later. The following proteins were included in this study: Apolactate dehydrogenase, carboxypeptidase A, concanavalin A, α-chymotrypsin, cytochrome b_5, cytochrome c, elastase, ferredoxin, α- and β-hemoglobin, insulin, lysozyme, myogen, myoglobin, papain, ribonuclease S, rubredoxin, staphylococcal nuclease, subtilisin BPN', thermolysin, and trypsin inhibitor. The β-regions were taken as reported by Chou and Fasman [4, 5] and Lim [6, 7]. Fig. 1 gives a diagrammatic representation of the breakdown of the number of β-regions with amino acid residues ranging from 2 to 18. Table I gives the numbers n_{jk} of the amino acid residues of the serial number

i (j= 1, 2,. . .,20) included in the non-β-regions (k=1), amino- (k=2) and carbonyl- (k = 3) ends of β-regions, and inner β-regions (k = 4).

The notation due to Ptitsyn [16] is followed throughout this paper. Based on the over-all distribution of the 20 amino acid residues among the four indicated regions, $n_{0,jk}$ and n_{jk} were computed (*see* Table I). Pearson criterion (*see*, for example, Yeomans [10] (was used to show that the deviations are not accidental. This consists in computing the experimental value of

$$\chi^2 = \sum_{j=1}^{20} \sum_{k=1}^{4} \frac{(\triangle n_{jk})^2}{n_{0,\,jk}} \qquad (1)$$

Substituting $\triangle n_{jk}$ and $n_{o,jk}$ from Table I into equation (1), one gets $\chi^2 = 140$. However, accordiing to the Pearson criterion, if the deviations were accidental, the probability that χ^2 exceeds 96 is only 0·1% (*see*, for example, Lewis [11]). This indicates that some of the deviations are far from being accidental and reflect a definite difference, to a 0·1 % level of significance, in the distribution of various amino acid residues. The permissible deviations under the normal distribution, $\triangle n_p$ are computed as indicated by Ptitsyn [16] and those deviations whose probabilities are less than 5% are underlined in Table I. Table I reveals the following features: Gly shows an above-average tendency to be in non-β-regions; Ala occurs mainly at the N-terminal ends of the β-regions, at the cost of non-β regions; Val prefers inner β-regions at the expense of non-β-regions; Leu prefers inner and C-terminal ends of β-regions, but not the non-β-regions; Ile prefers inner β-regions, and shows a below-average tendency to be in non-β-regions; Ser prefers non-β-regions; Pro shows a below-average tendency to be in inner β-regions and an above-average tendency to be in non-β-regions; Phe prefers the inner β-regions at the cost of non-β regions; Tyr prefers C-terminal β-regions; Asn and His prefer non-β-regions at the expense of inner β-regions; Trp prefers N-terminal ends of β-regions and Cys displays a below average tendency to be at the N-terminal ends of β-regions. This analysis was continued by grouping the amino acid residues into four groups each with some common features: strongly hydrophobic residues (Cys, Ile, Leu, Met, Phe, Trp, Tyr, and Val), weakly hydrophobic residues (Ala, Asn Gln, Gly, Ser, and Thr), acidic residues (Asp and Glu), and basic residues (Arg, His, and Lys). The results are shown in Table II. It is evident from Table II that the strongly hydrophobic residues prefer to be in the inner β-regions at the cost of non-β-regions, weakly hydrophobic residues prefer, along with the acidic and basic residues, to be in the non-β regions at the expense of inner β-regions.

TABLE I

*Distribution of amino acid residues among different regions**

Residue	Non-β-regions		Amino-ends of β-rigions		Carbonyl-ends of β-regions		Inner β-regions	
	n	Δ*n*	*n*	Δ*n*	*n*	Δ*n*	*n*	Δ*n*
Gly	41	10.52	11	— 5.12	16	— 0.30	22	— 5.08
Ala	19	— 8.77	25	10.31	13	—1.85	25	0.33
Val	24	— 9.19	22	4.45	12	—5.75	40	10.51
Leu	7	—15.01	10	— 1.64	20	8.23	28	8.44
Ile	11	—12.37	16	3.64	11	—1.50	31	10.24
Ser	37	10.92	9	— 4.79	10	—3.94	21	— 2.17
Thr	23	— 0.71	12	— 0.54	9	—3.68	26	4.94
Met	4	— 0.40	2	— 0.33	3	0.65	4	0.09
Pro	18	7.50	6	0.45	4	—1.61	3	— 6.33
Phe	6	— 5.52	5	— 1.09	4	—2.16	19	8.77
Tyr	13	— 4.27	7	— 2.13	17	7.76	14	— 1.35
Asp	21	4.41	7	— 1·78	10	1.13	11	— 3.74
Asn	23	8.44	8	0.30	9	1.21	3	— 9.94
Glu	12	2.52	3	— 2.01	8	2.93	5	— 3.43
Gln	12	1.50	6	0.45	6	0.39	7	— 2.33
His	8	1.57	3	— 0.40	2	—1.44	6	0.28
Lys	26	7.04	9	— 1.03	14	3.86	7	— 9.85
Arg	10	— 0.16	6	0.63	5	—0·43	9	— 0.03
Trp	6	— 0.09	7	3.78	1	—2.26	4	— 1.42
Cys	10	2.21	1	— 3.12	3	—1.16	9	2.08

a The numbers underlined are statistically reliable to 5% level of significance.

TABLE II

*Groupwise distribution of amino acid residues among the four regions**

Residues	Non-β regions		Amino-ends of β-regions		Carbonyl-ends of β-regions		inner β-regions	
	n	Δ*n*	*n*	Δ*n*	*n*	Δ*n*	*n*	Δ*n*
Highly hydrophobic (Cys, Ile, Leu, Met, Phe, Trp, Tyr, Val)	81	—44.66	70	3·55	71	3·80	149	37·37
Weakly hydrophobic (Ala, Asn, Gly, Gln, Ser, Thr)	155	21·90	71	0·61	63	—8·17	87	—31·25
Acidic (Asp, Glu)	33	6·92	10	—3·78	18	4·06	16	— 7·17
Basic (His, Lys, Arg)	44	8·43	18	—0·80	21	1·99	22	— 9·60
Chou and Fasman condition B–3 (Glu, Pro)	30	10·02	9	—1·56	12	1·32	8	— 9·76
Nagano hypothesis (Leu, Ile, Val, Phe)	48	—42·10	53	5·36	47	—1·18	118	37·96
High P residues due to Chou and Fasman (Val, Met, Ile, Cys, Tyr, Phe$_5$ Leu, Gln)	87	—43·06	69	0·23	76	6·46	152	36·45

* The numbers underlined are statistically reliable to 5% level of significance.

3. DISCUSSION

The results obtained above have been obtained from the data of 21 proteins with a total of 3,523 residues, of which 713 occur in β-regions. There are a total of 103 β-regions, if two or more consecutive residues falling in β-conformation with respect to their conformational angles ϕ and ψ (*see*, for example, Ramachandran and Sasisekharan [14]) are taken to define a β-region. The criterion used in this study uses a minimum of five consecutive β-residues to define a β-region. From Fig. 1 it is evident that this arbitrary definition neglects the contribution of 19 regions out of a total of 103, or 67 β-residues out of the total of 713. This neglect is assumed to be insignificant.

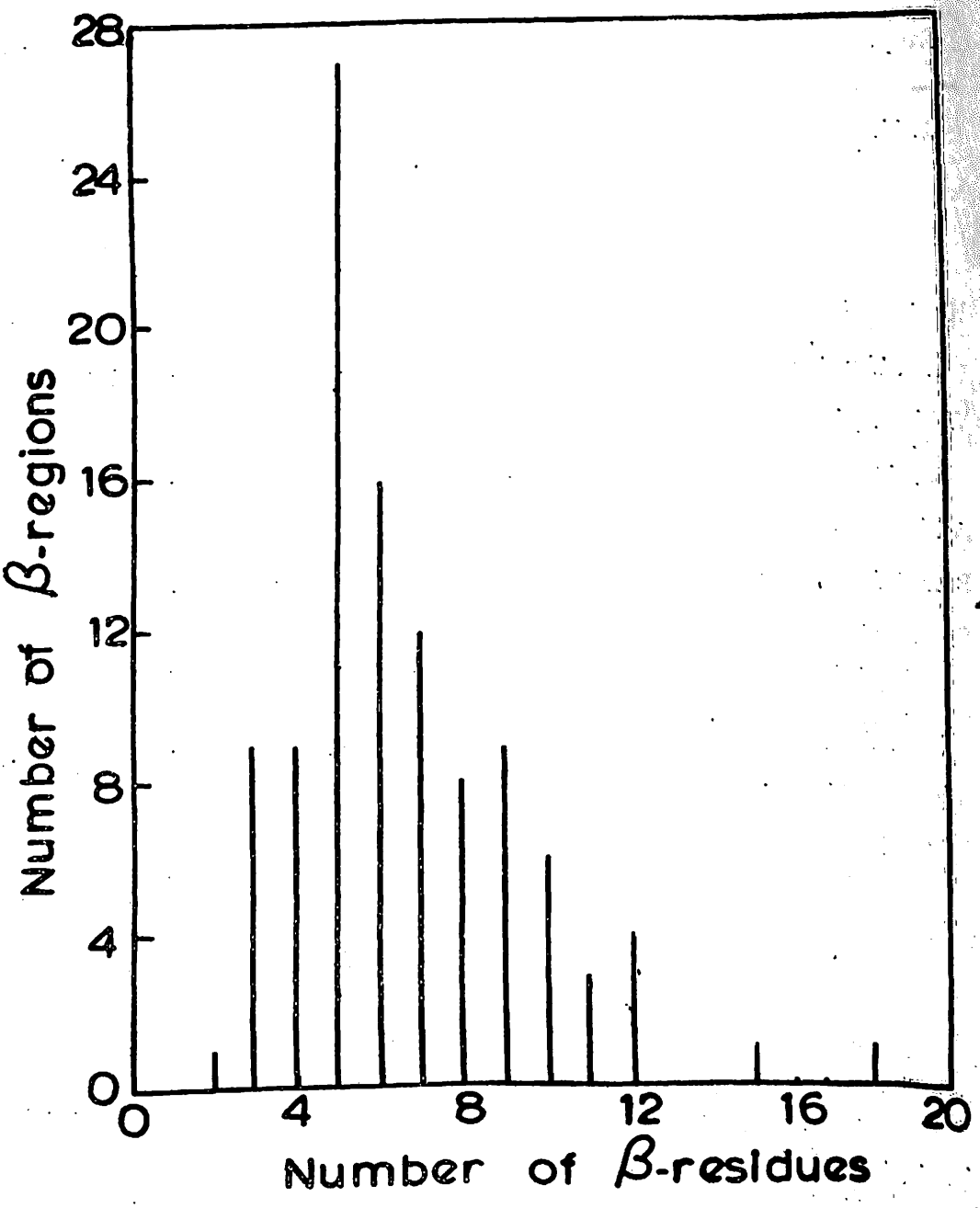

FIG. 1. Distribution of the β-residues per β-region for the 21 proteins as given in the text.

Chou and Fasman [4] take three residues for defining β-end regions and, in the process, neglect 36 regions out of a total of 66, or 168 β-residues out of 424. They neglect 55% of β-regions and 40% of β-residues. While it is

justified to take three residues on both ends of a helical region because of its intra-chain hydrogen-bonding the same cannot be said with regard to a β-region. Aleksanyan and Skvortsov [15] reported recently that β-structure is weaker in cooperativity than the α-structure. Finkelstein and Ptitsyn [2] used data on nine globular proteins in their statistical analysis of the correlation among amino acid residues in helical, β-structural, and non-regular regions. This analysis neglected 27 β-regions out of 43 or 126 β-residues out of 261. This neglect is thought to be considerable.

Based on the above arguments, the results of the present investigation are considered to be statistically very reliable. Because of the limited data, Finkelstein and Ptitsyn [2] studied the behaviour of hydrophobic and hydrophilic residues in β-regions without further classification into N-, C-terminal and inner β-regions. They found that highly hydrophobic residues prefer β-regions and weakly hydrophobic as well as highly hydrophilic residues show below-average preference for β-regions. Table II shows that the strongly hydrophobic residues prefer inner β-regions while weakly hydrophobic and charged residues prefer the non-β regions, at the cost of inner β-regions. The present study only partially supports conditions B–3 and B–4 of Chou and Fasman [5]: a look at Tables I and II shows that Pro and Glu together prefer non-β-regions; charged residues display an above-average tendency to be in non-β-regions and they do this at the cost of inner β-regions; Trp *does not* show any non-average behaviour at the C-terminal ends. Also, Table I shows no non-average behaviour for Arg. Chou and Fasman [4] wrote that β-residues with high P_β values are found with *equal* frequency in β-regions. Table II shows that the residues with high P_β values show an above-average tendency to be in the inner β-regions and a below-average tendency to be in the non-β-regions.

A set of new parameters f_β, $f_{\beta N}$, $f_{\beta I}$, P_β as obtained from the present analysis is given in Table III. This set was tried on thermolysin, a protein not included by Chou and Fasman in their statistical analysis. The new set of parameters gave a 27% β-content as against the 36% β-content given by the parameters due to Chou and Fasman [4]. The observed β-content is 22%.

According to Nagano [3] residues characteristic of β-structure are: Val, Leu, Ile, Phe, Thr, Gln, Met and Cys. Nagano [3] hypothesized that the initiation sites for β-structure are regions rich in Val, Leu, Ile and Phe. Table II shows that these residues prefer to be in inner β-regions, and not

TABLE III

*Frequency of β-residues and conformational parameters**

Residue	f_β	$f_{\beta N}$	$f_{\beta C}$	$f_{\beta I}$	P_β
Gly ..	0·152	0·033	0·048	0·066	0·75
Ala ..	0·202	0·076	0·040	0·076	0·99
Val ..	0·318	0·086	0·047	0·157	1·56
Leu	0·249	0·038	0·077	0·107	1·22
Ile ..	0·355	0·096	0·066	0·187	1·74
Ser ..	0·147	0·030	0·033	0·070	0·72
Thr ..	0·228	0·055	0·041	0·119	1·12
Met ..	0·205	0·045	0·068	0·091	1·00
Pro ..	0·097	0·048	0·032	0·02	0·47
Phe ..	0·310	0·044	0·035	0·168	1·52
Tyr ..	0·295	0·048	0·116	0·096	1·44
Asp ..	0·175	0·038	0·055	0·060	0·86
Asn ..	0·115	0·042	0·047	0·016	0·56
Glu ..	0·112	0·020	0·053	0·033	0·55
Gln ..	0·195	0·047	0·047	0·055	0·96
His ..	0·137	0·032	0·021	0·063	0·67
Lys ..	0·159	0·043	0·067	0·034	0·78
Arg ..	0·192	0·058	0·048	0·087	0·94
Trp ..	0·220	0·119	0·017	0·068	1·08
Cys ..	0·216	0·014	0·041	0·122	1·06

* See text for the definitions of f_β and P_β.

in non-β-regions. They are average-behaved at the N- and C-terminal β-regions. This hypothesis due to Nagano [3] cannot presently be tested since nothing is known concerning the sites of initiation for β-regions. It is, however, possible that the β-initiation takes place at the center and propagates in both directions until terminated at both ends by β-breakers. While acidic residues prefer the N-terminals and basic residues prefer the C-terminals of helical regions [16], charged residues show only average behaviour (*see* Table I) at β-region boundaries. However, unlike Chou and Fasman [4] have concluded, charged residues *are not* conspicuously absent at the β-region boundaries. In this sense, it is hard to see any directionality in β-regions.

Acknowledgements

It is a pleasure to thank Prof. G. N. Ramachandran who has been a source of inspiration to the author for the invaluable help extended in the course of this work. Thanks are also expressed to Dr. V. S. Anantha-narayanan, Dr. M. Vijayan, and Prof. K. B. Athreya for their kind help. Assistance from the SERC (DST) project in the Molecular Biophysics Unit is gratefully acknowledged.

References

[1] Burgess, A. W., Ponnuswamy, P. K. and Scheraga, H. A. *Isr. J. Chem.*, 1974, **12**, 239.

[2] Finkelstein, A. V. and Ptitsyn, O. B. *J. Mol. Biol.*, 1971, **62**, 613.

[3] Nagano, K. *J. Mol. Biol.* 1973 **75**, 401.

[4] Chou, P. Y. and Fasman, G. D. *Biochemistry*, 1974, **13**, 211.

[5] Chou, P. Y. and Fasman, G. D. *Biochemistry*, 1974, **13**, 222.

[6] Lim, V. I. *J. Mol. Biol.*, 1974 *a*, **88**, 857.

[7] Lim, V. I. *J. Mol., Biol.* 1974 *b*, **88**, 873.

[8] Ananthanarayanan, V. S. and Bandekar, J. *Curr. Sci.*, 1975, **44**, 609.

[9] Ananthanarayanan, V. S. and Bandekar, J. *Int. J. Pept. Protein Res.*, 1976, **8**,

[10] Yeomans, K. A. .. *Applied Statistics*, Vols. I and II, London: Allen Lane, The Penguin Press. 1968,

[11] Lewis, T. .. *Biometrika*, 1953, 40, 421.

[12] Bliss, C. I. .. *Statistics in Biology*, Vols. I and II, New York: McGraw-Hill Book Co., 1970.

[13] Owen, D. B. .. *Handbook of Statistical Tables*, Reading: Addison Wesley Publishing Co., 1962.

[14] Ramachandran, G. N. and Sasisekharan, V. *Adv. Protein Chem.*, 1968, 23, 284.

[15] Aleksanyan, V. I. and Skvortsov, A. M. *Mol. Biol.* 1974, 8, 142.

[16] Ptitsyn, O. B. .. *J. Mol. Biol.*, 1969, 42, 501.