THE CHOU-FASMAN SECONDARY STRUCTURE PREDICTION METHOD WITH AN EXTENDED DATA BASE

Patrick ARGOS, Michael HANEI and R. Michael GARAVITO+

Department of Physics, Southern Illinois University, Edwardsville, IL 62026 and *+Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA*

Received 6 June 1978

1. Introduction

In the last 25 years many protein primary sequences and tertiary structures have been determined. This knowledge has prompted the development of several schemes to predict secondary structural regions in proteins (α-helices, β-sheets and loops) [1].

The most widely used technique has been that of Chou and Fasman [2–4]. It consists of predictive rules that can be easily applied by a visual inspection of the primary sequence; the method also requires little mathematical sophistication for its comprehension.

Chou and Fasman determined conformational parameters (normalized frequencies) for each amino acid to predict helical and sheet secondary structural regions in proteins. Their data sample used to calculate these parameters effectively included 1939 residues. Levitt [8] has recently determined the normalized frequencies from a 5523 residue sample. Both sets of parameters were employed in a computerized version of the Chou-Fasman secondary structure prediction procedure [3]. No improvement was observed in the correctness of predictions made with the extended data base. Cross-correlation functions were calculated to measure the degree of linear dependence between the Chou-Fasman or Levitt conformational parameters and various physical properties of amino acids.

Address correspondence to: Dr Patrick Argos, Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

2. Methods

The Chou-Fasman technique [4] essentially relies on the normalized frequency (protein conformational parameters) with which all 20 amino acids appear in β-sheet and α-helical regions. The frequency of an amino acid in a given secondary structure was obtained by dividing its occurrence in each conformational region with its total occurrence, as observed in a 15 protein data sample. The conformational parameters were then calculated by normalizing this frequency through division by the average frequency of a residue in a sheet or helical region. All 20 amino acids can then be listed in a hierarchical order ranging from a strong former to a strong breaker of a particular secondary structure. The residues in a protein whose secondary structure is to be predicted are then assigned a conformational parameter as well as an ability to make or break a sheet or helix. If segments consisting of a certain number of consecutive residues in the primary sequence can be found with an average conformational parameter greater than a given value, a nucleation site for a sheet or helix is predicted. The parameters for residues on the N- and C-terminal sides of the nucleation site are then examined to find breaking clusters which terminate the secondary structural span. A computer program was written [5] that followed the Chou-Fasman rules as nearly as possible.

Levitt and Greer [6] have recently devised an automatic and objective technique to identify regions of secondary structure in globular proteins. Through an examination of the atomic coordinates of a large number of proteins, they ascertained patterns of

Table 1

Helical (P_α) and sheet (P_β) conformational parameters as calculated by Chou and Fasman (CAF) and Levitt (LEV)

P_α

	CAF	LEV
H_α	1.53 GLU	MET 1.47
	1.45 ALA	GLU 1.44
	1.34 LEU	LEU 1.30
		ALA 1.29
		GLN 1.27
h_α	1.24 HIS	LYS 1.23
	1.20 MET	HIS 1.22
	1.17 GLN	
	1.14 VAL	
	1.14 TRP	
	1.12 PHE	CYS 1.11
I_α	1.07 LYS	PHE 1.07
	1.00 ILE	ASP 1.04
i_α	0.98 ASP	TRP 0.99
		ILE 0.97
		ARG 0.96
		VAL 0.91
		ASN 0.90
		SER 0.82
	0.82 THR	THR 0.82
	0.79 ARG	
	0.79 SER	
	0.77 CYS	
b_α	0.73 ASN	TYR 0.72
	0.61 TYR	
B_α	0.59 PRO	GLY 0.56
	0.53 GLY	PRO 0.52

P_β

	CAF	LEV
H_β	1.67 MET	
	1.65 VAL	
	1.60 ILE	
		VAL 1.49
		ILE 1.45
h_β	1.30 CYS	PHE 1.32
	1.29 TYR	
	1.28 PHE	
	1.23 GLN	TYR 1.25
	1.22 LEU	THR 1.21
	1.20 THR	
	1.19 TRP	
		TRP 1.14
		HIS 1.08
I_β		LEU 1.02
		ARG 0.99
	0.97 ALA	MET 0.97
		SER 0.95
i_β		GLY 0.92
	0.90 ARG	ALA 0.90
	0.81 GLY	GLN 0.80
	0.80 ASP	
b_β		LYS 0.77
		ASN 0.76
		GLU 0.75
	0.74 LYS	CYS 0.74
	0.72 SER	ASP 0.72
	0.71 HIS	
	0.65 ASN	PRO 0.64
	0.62 PRO	
B_β	0.26 GLU	

The CAF data base consisted of 1939 residues while the LEV sample included 5523 residues. The CAF qualitative assignment of an amino acid's ability to form secondary structure is as follows: H, strong former; h, former; I, weak former; i, indifferent; b, breaker; B, strong breaker peptide hydrogen bonds, inter-C_α distances and inter-C_α torsion angles that define precisely secondary structural regions. With the use of the Brookhaven protein data bank [7] file, Levitt [8] has recently calculated the Chou-Fasman normalized frequencies from a 5523 residue data base consisting of 31 protein classes. Proteins with a high degree of sequence homology and tertiary structural similarity were assigned a protein class and only allowed to give a weighted contribution to the data sample. Thus the total number of residues examined was 11 569. If the Chou-Fasman 15 protein data base with a total of 2473 residues were similarly weighted, their data sample would effectively consist of 1939 residues. The resulting Levitt and Chou-Fasman conformational parameters are listed in table 1. The qualitative assignment of an amino acid's ability to form or break a secondary structural pattern is also shown in table 1 for the Chou-Fasman (CAF) and Levitt (LEV) parameters.

It is clear that there are significant hierarchical shifts in residues which frequently occur in proteins. For example, valine has moved from a 'former' of helices to an 'indifferent' status. Histidine has changed from a 'breaker' of sheet formation to a 'former' of β-strands.

Cross-correlation functions were calculated between the CAF and LEV normalized frequencies for the helical and sheet cases. The cross-correlation function (CCF) between two series X and Y, each consisting of N elements, is defined as [9]:

$$\mathrm{CCF} = \frac{\sum_{j=1}^{N} (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^{N} (X_j - \bar{X})^2 \sum_{j=1}^{N} (Y_j - \bar{Y})^2}$$

where $\bar{X}$ and $\bar{Y}$ are the mean values of the elements in the respective series. The function measures the degree of linear dependence between elements of the two different series. A value of CCF near +1 indicates that the size of the series elements (large or small) follow each other while a −1 CCF value indicates a large element in one series follows a small element in the other series. A value near zero shows little correlation. For the CAF and LEV helical parameters, CCF_α was determined to be 0.87; for the β-sheet case, CCF_β was 0.67. It is clear that the greatest changes resulting from the near tripling of the data base occurred in the sheet normalized frequencies.

3. Evaluation of predictions

A sensitive parameter used to evaluate the correctness of a prediction is the correlation coefficient, C, proposed by Matthews [1]. The correlation between helical prediction and observation would be calculated by:

$$C_\alpha = \frac{(a/N - \bar{P}\,\bar{O})}{[\bar{P}\,\bar{O}\,(1-\bar{O})(1-\bar{P})]^{1/2}}$$

where N is the total number of residues in the protein and $\bar{O}$ and $\bar{P}$ are, respectively, the fraction of the protein observed and predicted helical. A perfect prediction would be indicated by $C_\alpha = 1.0$, while a random prediction would yield $C_\alpha = 0.0$. Total disagreement in observation and prediction would result in a C_α value of -1.0. A C_β coefficient can be similarly determined. A useful prediction would probably have a correlation greater than 0.4 [5].

4. Results and discussion

Predictions of α-helices and β-sheets for 24 proteins with known structures [7] were calculated from the computerized version of the Chou-Fasman procedures. The amino acids were classified in their breaking and forming categories as shown in table 1. The list of proteins included 11 of those in the original CAF data sample. The mean C_α and C_β were weighted according to the number of residues within a given structure. The correlation coefficient values were: $C_{\alpha,\mathrm{CAF}} = 0.35$, $C_{\alpha,\mathrm{LEV}} = 0.36$, $C_{\beta,\mathrm{CAF}} = 0.36$ and $C_{\beta,\mathrm{LEV}} = 0.36$. Though the predictions were not identical with the different data bases, an increased data base did not generally improve the prediction quality. The LEV parameters increased the mean C_α by only 0.01, while the average C_β did not change.

Table 2
Correlation coefficients for the secondary structure prediction of 4 proteins not included in the Chou-Fasman or Levitt data bases

Protein	Chain length	α-Helix		β-Structure	
		CAF	LEV	CAF	LEV
Dihydrofolate reductase [11] (*Escherichia coli* – MTX)	156	0.42	0.40	0.35	0.20
Rhodanese [12] (bovine)	293	0.39	0.17	0.39	0.11
Neurotoxin [13] (*Laticauda semifasciata*)	62	–	–	0.19	–0.04
Penicillopepsin [14] (*Penicillium janthinellum*)	322	0.06	0.17	0.20	0.08
Average		0.26	0.22	0.29	0.10

CAF indicates the correlation coefficient resulting from predictions using the Chou-Fasman frequencies; LEV indicates the use of the Levitt parameters. The numbers in parentheses refer to the reference describing the tertiary structure as determined by X-ray diffraction techniques

These results are particularly surprising given the considerable shifts of amino acids in the hierarchy of ability to form or break secondary structures.

Correlation coefficients for 4 proteins not included in either the LEV or CAF data bases are given in table 2. The lower mean values probably result from the smaller sample. However, the sheet corelation for the LEV parameters is significantly decreased. This is likely a result of the more uniform distribution of the β conformational parameters within the amino acid hierarchy determined from the LEV extended data base. Plots of the β conformational parameter versus the rank of an amino acid within the β hierarchy are shown in fig.1 for the CAF and LEV cases. It is clear that a nearly 3-fold increase in data results in a more linear relationship. The demarcation between strong and weak β-sheet formers and breakers becomes less clear and points toward necessary alterations in the CAF procedure which relies on such categorizations. The same trend, though not as striking, can be observed in the helical parameters. Correspondingly, the mean C_α for the LEV-based predictions also decreases, though not as dramatically as the mean C_β.

Jones [10] lists various physical and chemical properties of amino acids: bulkiness, polarity, hydrophobicity, R_F rank from paper chromatographic studies, and refractivity. Cross-correlation functions were calculated between the CAF or LEV normalized frequencies and each of the amino acid properties; the resultant CCF values are given in table 3. Bulkiness, hydrophobicity and chromatographic rank show the largest positive cross-correlations with the CAF and LEV sheet parameters. The LEV sheet frequencies are correlated less well with polarity than are the CAF parameters. Polarity and especially refractivity appear better correlated with the LEV rather than the CAF helical conformational values.

It appears that the Chou-Fasman secondary structure prediction method has not improved its prediction ability with an extended data base. Modification of the Chou-Fasman procedure thus seems necessary, especially in the light of the more linear distribution of the sheet conformational parameter with amino acid rank.

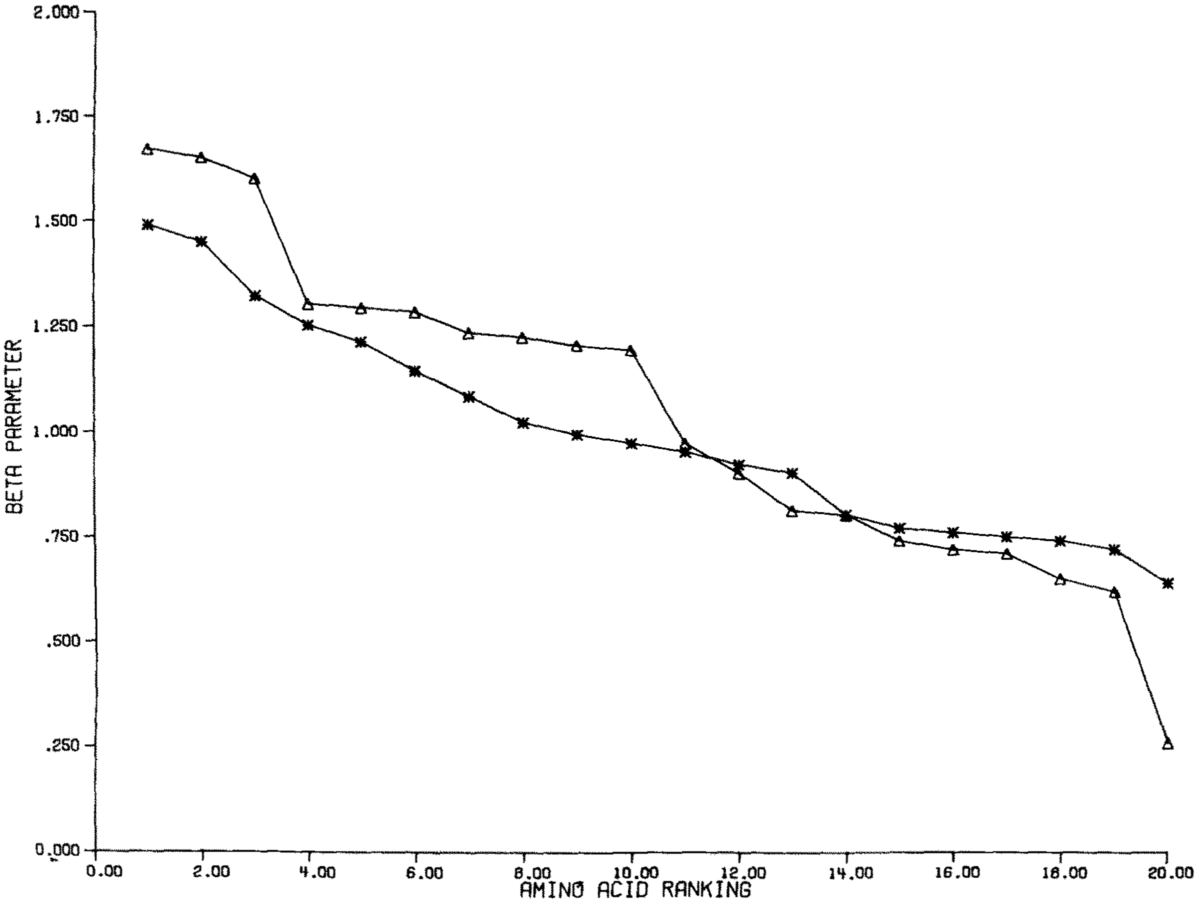

Fig.1. Plot of the β-sheet conformational parameter versus the amino acid rank within the hierarchical ability to form β-structures. An (*) indicates parameters calculated from the Levitt 5523 residue sample while (△) refers to those obtained by Chou and Fasman from 1939 residues.

Table 3
Cross-correlation functions between Chou-Fasman (CAF) or Levitt (LEV) conformational parameters and several amino acid physical properties

Amino acid property	CAF(α)	LEV(α)	CAF(β)	LEV(β)
Polarity	0.24	0.32	–0.56	–0.32
Bulkiness	0.30	0.16	0.56	0.56
Hydrophobicity	0.07	–0.07	0.44	0.44
Chromatographic rank	0.15	–0.08	0.62	0.65
Refractivity	0.09	0.23	0.34	0.22

Cross-correlation function values between the amino acid property given in the first column and the CAF or LEV parameters are respectively given in the columns designated by CAF and LEV. An α or β indicates respectively the use of the helix or sheet normalized frequencies

Acknowledgements

The authors wish to express gratitude to Dr Michael Levitt of the MRC Laboratory of Molecular Biology in Cambridge, England, for providing conformational parameters. Without the outstanding secretarial abilities of Ms Sharon Wilder, this manuscript would never have made the printing stage. This research was supported by a grant from the state of Illinois.

References

[1] Matthews, B. W. (1975) Biochim. Biophys. Acta 405, 442–451.

[2] Chou, P. Y. and Fasman, G. D. (1978) in: Advances in Enzymology (Meister, A. ed) vol. 46, Academic Press, New York, in press.

[3] Chou, P. Y. and Fasman, G. D. (1974) Biochemistry 13, 211–222.

[4] Chou, P. Y. and Fasman, G. D. (1974) Biochemistry 13, 222–245.

[5] Argos, P., Schwarz, J. and Schwarz, J. (1976) Biochim. Biophys. Acta 439, 261–273.

[6] Levitt, M. and Greer, J. (1977) J. Mol. Biol. 114, 181–293.

[7] Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973, USA.

[8] Levitt, M. (1977) unpublished results.

[9] Jenkins, G. M. and Watts, D. G. (1968) Spectral Analysis and Its Applications, Holden-Day, San Francisco.

[10] Jones, D. D. (1975) J. Theor. Biol. 50, 167–183.

[11] Matthews, D. A., Alden, R. A., Bolin, J. T., Freer, S. T., Hamlin, R., Xuong, N., Kraut, J., Poe, M., Williams, M. and Hoogsteen, K. (1977) Science 197, 452–455.

[12] Bergsma, J., Hol, W. G. J., Jansonius, J. N., Kalk, K. H., Ploegman, J. H. and Smit, J. D. G. (1975) J. Mol. Biol. 98, 637–643.

[13] Tsernoglou, D. and Petsko, G. A. (1976) FEBS Lett. 68, 1–4.

[14] Hsu, I. N., Delbaere, L. T. J., James, M. N. G. and Hofmann, T. (1977) Nature 266, 140–145.