Int. J. Peptide Protein Res. **19**, 1982, 394–401

Protein secondary structure

Studies on the limits of prediction accuracy

JAUME PALAU,† PATRICK ARGOS* and PERE PUIGDOMENECH†

†*Unit of Biophysics and Molecular Biology, Institute of Biology of Barcelona, Center for Investigations and Research of Cataluña, C.S.I.C., Barcelona, Spain, and*
**Department of Biological Sciences, Purdue University, West Lafayette, Indiana, USA*

Received 17 July, accepted for publication 2 November 1981

A secondary structure prediction technique is proposed which includes nucleation site determination through multiplication of conformational preference parameters as well as weighting factors to represent structurally stabilizing short range interactions. The prediction accuracy of the method is calculated using data bases categorized according to the four protein structural classes and with differing assignments of secondary structural regions. The results indicate that nearest neighbor prediction techniques (a) are insensitive to various assignment criteria for the secondary structural spans, (b) have nearly achieved their upper limit of prediction accuracy, and (c) can be somewhat improved through the use of stereochemical weighting factors and conformational parameters derived from the four structural groups.

Key words: protein secondary structure; protein structure prediction; structure prediction accuracy.

X-ray diffraction studies of crystalline proteins have to date resulted in nearly 100 known tertiary structures (1, 2). With their advent have come many secondary structure prediction methods which require only a knowledge of the amino acid sequence (cf. 3–5). These techniques generally rely on a statistical or informational analysis of the frequency with which the 20 amino acids appear within the observed secondary structures (α-helix, β-strands and reverse turns). The most popular is that of Chou & Fasman (6, 7) who calculate conformational preference parameters for each of the amino acids in particular secondary structures. The normalized propensity parameters are defined as the ratio of the frequency with which an amino acid appears in a secondary structure to its frequency within the entire sample. If a contiguous segment of four or five amino acids have an average propensity value greater than a threshold assignment, a secondary structure nucleation site is declared; terminal regions are then determined where the average propensity falls below a set value. The present paper will propose a modified version of the Chou and Fasman technique. A nucleation center is determined by multiplying the frequency probabilities for a continuous segment of five residues; this contrasts with the Chou and Fasman additive propensity principle. Furthermore, weighting factors are introduced which account for certain stabilizing interactions observed within secondary structures.

The prediction routines are far from perfect; they are generally about 60% correct depending on the criteria of assessment (3). Obviously

0367–8377/82/040394–08 $02.00/0

several questions have arisen regarding the etiology of their limited success and the possible extent of their improvement. Will an increased data base lead to better and more accurate predictions? Are the methods limited by the lack of agreement in the assignment of secondary structural regions within a protein? Can the prediction techniques be improved by calculating frequency factors from the four protein structural classes proposed by Levitt & Chothia (8)? The present work will address these queries. Conformational propensities are calculated for each amino acid with the use of various data bases, which include those categorized according to the four protein structural classes as well as two samples resulting from different criteria for secondary structural assignments; namely, that proposed by Levitt & Greer (9) and by Chou & Fasman (10). The resulting frequency factors are then used in the prediction scheme proposed here and their effect on the prediction quality is assessed. It appears that the nearest neighbor prediction technique is not sensitive to the two assignment criteria and has nearly achieved its upper limit of prediction success with a mean near 60%. However, utilization of the protein class conformational parameters which differ considerably may provide some improvement in prediction accuracy.

DATA BASES

Calculations of the prediction parameters were made by using two sets of protein samples. One set labeled as "CF" is formed by 33 proteins, which correspond to the protein sample of Chou & Fasman (10) with the following additions: bacteriophage thioredoxin (11), worm myohemerythrin (12–14), bovine superoxide dismutase (14) and chicken muscle triose phosphate isomerase (15). The secondary structures were delineated according to the (ψ, ϕ) dihedral angles of the mainchain peptides (5) as well as certain threshold distances between particular protein atoms (10); for example, after suitable exclusion of helical regions, a tetrapeptide is assigned as a β-turn if the $(C_{\alpha,i} - C_{\alpha,i+3})$ distance is less than 7 Å.

Levitt & Greer (9) have analyzed automatically and objectively the atomic coordinates of 60 known protein structures to identify regions of sheet, helix, and turn conformations. Their criteria for this delineation were based on patterns in peptide hydrogen bonds, inter-C_α distances and inter-C_α torsion angles. The secondary structural segments of the 60 proteins are given in Tables 7, 8, 9 and 10 of their publication (9). Only 44 of the listed proteins were utilized in the present analysis. Proteins without primary structures and one subunit of redundant dimeric pairs were eliminated. Also excluded were repeated protein structures: rubredoxin at 2.0 Å, concanavalin A (Rockefeller), alpha chymotrypsin A (Michigan), ferricytochrome c "inner" and "outer", ribonuclease S, semiquinone flavodoxin, subtilisin *novo*, lactate dehydrogenase-NAD, and D-glyceraldehyde-3-phosphate dehydrogenase "red". This second data set is labeled as "LG".

All proteins in the CF data pool are also part of the LG sample with the exception of superoxide dismutase. The LG assignments were only accepted if helical, sheet and turn regions contained respectively at least five, three and four consecutive amino acids. All designations in the CF data base satisfied these conditions. Each of the data sets were also divided in groups according to their secondary structural character: all-α, all-β, $\alpha + \beta$, and α/β proteins, classifications suggested by Levitt & Chothia (8).

PREDICTION ALGORITHM

The proposed algorithm is based on two factor types: those corresponding to frequency parameters and those derived from short range interactions. The prediction parameter, θ_i^s, which indicates the propensity that the ith amino acid of the protein sequence will be in secondary structural state s, can be expressed as:

$$\theta_i^s = \left[\prod_{j=1}^{n} (W_i^s)_j\right] \left[\prod_{k=i-m}^{i+m} P_k^s\right] \quad (1)$$

where $(W_i^s)_j$ is the jth weighting factor representing structurally stabilizing short range interactions, and P_k^s is the frequency factor for the amino acid in a given position k which varies from $i - m$ to $i + m$. If the amino acid in position k is of type l (Arg, Ala, and so forth) where $l = 1$ to 20, then the frequency factor P_l^s can be expressed as:

$$P_l^s = \frac{N_l^s/N^s}{N_l/N} = \frac{f_l^s}{f_l} \qquad (2)$$

where N_l^s is the number of times within the data base that a particular amino acid l appears within secondary structural type s; N^s is the number of all amino acid types in the data base that are in structural configuration s; N_l is the number of times that the amino acid appears in the data irrespective of structural configuration, and N is the total number of all residues in the data pool. Clearly f_l^s represents the fraction or percentage of the secondary structural s residues that are composed of amino acid type l; similarly f_l is the fraction of the entire sample that is composed of the l amino acid type. Eqn. 2 is essentially the definition given by Chou & Fasman (6, 7) for their conformational parameters. The standard error for the conformational preference parameters (P_l^s) can be estimated as (10, 19):

$$\sigma_P = \frac{1}{f_l}\left[\frac{f_l^s(1-f_l^s)}{N^s}\right]^{1/2} \qquad (3)$$

In the present work the following forms of eqn. 1 have been used:

$$\theta_i^\alpha = W_i^\alpha \left[\prod_{k=i-2}^{i+2} P_k^\alpha\right] \qquad (4)$$

$$\theta_i^\beta = W_i^\beta \left[\prod_{k=i-2}^{i+2} P_k^\beta\right] \qquad (5)$$

and

$$\theta_i^t = \left[\prod_{k=i-2}^{i+2} P_k^t\right] \qquad (6)$$

The symbols α, β and t refer respectively to the helical, sheet and turn configurations. The value of m in eqn. 1 has been set equal to two. The weighting factor W_i^α for the ith position along the protein sequence is a frequency parameter which represents the occurrence of hydrophobic triplets in helical positions 1-2-5 and 1-4-5. The hydrophobic clusters are apparently important as helix stabilizers (17). The term W_i^α can be expressed as:

$$W_i^\alpha = \left[\frac{f_{\phi t}^\alpha}{f_\phi^\alpha}\,\frac{f_\phi^c}{f_{\phi t}^c}\right]^{v_i} \qquad (7)$$

where $f_{\phi t}^\alpha$ is the fraction of all possible pentapeptides in the helical regions of the data sample that contain hydrophobic residues in the 1-2-5 and 1-4-5 positions; f_ϕ^α is the fraction of all possible pentapeptides that are in helices irrespective of their hydrophobic nature; $f_{\phi t}^c$ is the fraction of helical residues that are hydrophobic and f_ϕ^c is the fraction of amino acids in the data sample that are hydrophobic. The second ratio serves to normalize with respect to hydrophobicity. The amino acids considered hydrophobic were Leu, Ile, Val, Met, Phe, Tyr, Trp and Ala. The fractional ratios are raised to the power v_i which is the number of times the residue in the ith position of the amino acid sequence to be predicted appears in hydrophobic triplets at positions 1-2-5 and 1-4-5. All possible pentapeptides in which the ith amino acid can participate are searched to determine v_i. A similar definition is given for W_i^β where hydrophobic doublets in the 1-3 positions of β strands are considered. Since the weighting factor increases the secondary structural prediction parameter of the ith amino acid by a power law, it reflects the stabilizing local hydrophobic interactions. Other weighting factors can be introduced to express further systematic interactions in secondary structures.

ASSESSMENT OF PREDICTION ACCURACY

Several quality indices have been utilized to assess the correctness of secondary structure predictions. Schulz & Schirmer (3) have critically reviewed the present evaluation chaos and suggest the use of a composite index Q_{pos} which is expressed as:

$$Q_{pos} = A_\alpha + A_\beta + A_t + A_{coil} \qquad (8)$$

where A_s is the percentage of all the protein residues that are predicted and observed in secondary structural state s. The Q_{pos} value is termed a positive correct prediction index. In the present work, the prediction accuracy for several proteins was determined through a weighted composite index (Q_{pos}^w); that is,

$$Q_{pos}^w = \frac{\sum_{i=1}^{l} N_i Q_{pos,i}}{\sum_{i=1}^{l} N_i} \qquad (9)$$

where l is the number of proteins predicted, N_i is the number of amino acids in the ith molecule, and Q_{pos} is the composite index for the ith protein.

Correlation coefficients were calculated between various parameters. If $[X_j]$ and $[Y_j]$ represent data sets each of n members, then the coefficient (CCF) correlating the two parameter series is given (18) by

$$CCF = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(Y_j - \bar{Y})}{\left[\sum_{j=1}^{n} (X_j - \bar{X})^2 \sum_{j=1}^{n} (Y_j - \bar{Y})^2\right]^{1/2}} \quad (10)$$

where

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

and

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

RESULTS AND DISCUSSION

Frequency factors from the CF and LG data sets

Table 1 lists the conformational preference parameters (eqn. 2) for the 20 amino acids in the helical, sheet, and turn structural states utilizing both the LG and CF data samples. The LG values are in essential agreement with the frequency factors calculated by Levitt (19). The CF values also correlate well with the LG results despite the differing criteria for secondary structural assignment. The CCF values between the P^{α}, P^{β}, and P^{t} paired lists were respectively 0.89, 0.91, and 0.81. Furthermore, the prediction accuracy of the algorithm discussed here was little affected

TABLE 1

Conformational preference parameters (P^{α}, P^{β}, and P^{t}) for the 20 amino acids as calculated from the LG and CF data samples. The standard errors (σ_p) are given in parentheses

Amino acid	P^{α} (LG)	P^{α} (CF)	P^{β} (LG)	P^{β} (CF)	P^{t} (LG)	P^{t} (CF)
Leu	1.30 (0.06)	1.22 (0.08)	1.03 (0.06)	1.24 (0.12)	0.49 (0.08)	0.56 (0.07)
Ile	0.87 (0.07)	1.01 (0.09)	1.47 (0.10)	1.59 (0.17)	0.55 (0.11)	0.57 (0.08)
Val	0.95 (0.05)	1.05 (0.08)	1.44 (0.07)	1.73 (0.14)	0.51 (0.08)	0.55 (0.07)
Met	1.32 (0.17)	1.47 (0.21)	0.96 (0.15)	0.94 (0.25)	0.52 (0.21)	0.71 (0.18)
Phe	1.09 (0.09)	1.10 (0.11)	1.13 (0.10)	1.41 (0.19)	0.88 (0.17)	0.72 (0.11)
Tyr	0.71 (0.07)	0.72 (0.09)	1.35 (0.10)	1.45 (0.19)	1.28 (0.20)	1.12 (0.14)
Trp	1.03 (0.13)	1.02 (0.17)	1.24 (0.15)	1.28 (0.28)	0.88 (0.25)	0.90 (0.20)
Ala	1.30 (0.06)	1.32 (0.07)	0.81 (0.05)	0.90 (0.09)	0.84 (0.10)	0.65 (0.06)
Thr	0.80 (0.06)	0.86 (0.08)	1.19 (0.07)	1.20 (0.14)	1.05 (0.13)	0.96 (0.10)
Ser	0.78 (0.05)	0.77 (0.07)	1.02 (0.06)	0.70 (0.09)	1.29 (0.12)	1.46 (0.11)
Cys	0.92 (0.11)	0.70 (0.13)	1.12 (0.13)	1.12 (0.24)	0.69 (0.19)	1.43 (0.22)
Asn	0.90 (0.07)	0.74 (0.08)	0.81 (0.08)	0.82 (0.13)	1.48 (0.19)	1.45 (0.14)
Gln	1.04 (0.09)	1.25 (0.13)	1.03 (0.10)	0.95 (0.17)	1.00 (0.19)	0.94 (0.14)
Asp	1.02 (0.07)	0.97 (0.08)	0.71 (0.06)	0.75 (0.10)	1.28 (0.15)	1.47 (0.12)
Glu	1.43 (0.09)	1.48 (0.11)	0.59 (0.06)	0.44 (0.09)	0.78 (0.13)	0.75 (0.09)
His	1.33 (0.12)	1.06 (0.13)	0.85 (0.10)	0.86 (0.17)	0.53 (0.15)	0.96 (0.15)
Lys	1.23 (0.06)	1.13 (0.08)	0.77 (0.05)	0.75 (0.09)	0.95 (0.12)	0.95 (0.08)
Arg	0.93 (0.10)	1.04 (0.13)	1.03 (0.11)	0.75 (0.16)	0.91 (0.19)	0.93 (0.15)
Gly	0.63 (0.04)	0.59 (0.05)	0.94 (0.05)	0.83 (0.09)	1.76 (0.14)	1.53 (0.10)
Pro	0.63 (0.07)	0.57 (0.08)	0.75 (0.07)	0.46 (0.11)	1.47 (0.20)	1.51 (0.16)

by the use of either data base in determining the preference parameters (*vide infra*).

Preference parameters for structural classes

The LG protein sample was categorized according to the four structural classes: all-α, all-β, $\alpha + \beta$, and α/β proteins (8). The all-α and all-β groups consist primarily of α-helices and β-strands respectively. The α/β proteins generally possess alternate helical and strand configurations while the $\alpha + \beta$ set tends to successive helices followed by successive strands in the peptide backbone. Levitt & Greer (9) list their data sample in four tables according to the structural divisions. Table 2 shows the preference parameters and standard errors as calculated for each class from the LG data base. The results suggest distinctive utilization of certain amino acids depending on the amount and topology of the different secondary structures. Leu, Gln and Glu are preferred in helices of structures that also contain β-sheets. Val is more frequently used in helices of $\alpha + \beta$ structures while Thr is utilized in all-α proteins. There are also distinct preferences for certain β-strand amino acids as they appear in the structural classes. Val and Ile dominate the sheets of α/β proteins while Ile and Phe are preferred in $\alpha + \beta$ structures which also uniquely use Asn. The all-β proteins are less selective and utilize Val, Met, Tyr, and Phe. Though the propensity parameters for β-bends show large standard errors, turn residues consistently appearing in all structural classes include Gly, Asn, Pro and Ser. The all-α proteins are unique in their turn usage of Asp, Gln and Phe and strong non-usage of Thr. The all-β proteins prefer Arg and avoid Tyr.

Structural constraints would be expected to dictate the dominance of particular amino acids. Janin & Chothia (20) have observed that α/β proteins prefer Val and Ile as constituent residues in their β-strands in order to form a smooth sheet surface against which helices pack. The strand preference parameters show that Ile and Val are the preferred amino acids in α/β strands. Lifson & Sander (21, 22) have statistically determined that Val and Ile make up 32.5% of the residues in parallel β-strands and only 22.4% for the anti-parallel case. Since α/β structures are largely composed of parallel strands, it is consistent that the Ile and Val P^{β} values are the largest (Table 2). The amino acid preferences noted here are thus likely to have structural explanations. For example, Thr would aid helix initiation in all-α proteins through hydrogen bonding between its γ-oxygen and the mainchain (cf. 23). Val and Leu may be preferred helical residues in proteins with β-structures to facilitate helix-sheet packing.

Secondary structure predictions

The prediction algorithm proposed here was applied to the proteins of the LG and CF data samples. Frequency parameters calculated for the entire data base and for each of the structural classes were used to predict respectively the secondary structure for all the molecules and for those proteins in each of the structural groups. Furthermore, predictions were attempted for all proteins with conformational propensity parameters derived from a given class. The $f^{s}_{\phi t}$ values for triplets and doublets (eqn. 7) are given in Table 3 for the LG proteins. The CF results were similar. The percentage of all possible helical pentapeptides that contain hydrophobic triplets in the 1-2-5 and 1-4-5 positions varies amongst the structural classes: 19.0% (all-α), 12.4% (all-β), 12.9% ($\alpha + \beta$) and 14.8% (α/β). Apparently packing helices in all-α structures requires the hydrophobic-hydrophilic helical sideness. The parallel β-strands of α/β proteins also appear more demanding in doublet hydrophobicity.

The weighted composite indices (eqn. 9) resulting from the application of the proposed prediction method using various frequency parameters are given in Table 4. It is clear that prediction quality is only somewhat improved for a particular structural family using conformational propensity values calculated from the same protein class. The LG and CF data samples are predicted with nearly the same accuracy despite the different criteria delineating secondary structural regions and the different size of the data bases. These results would suggest that the limit of prediction accuracy from singlet amino acid methodologies has nearly been reached. This threshold is near 0.56 which is the Q_{pos} index utilizing

TABLE 2

Conformational preference parameters (P^{α}, P^{β}, and P^{t}) as calculated from the LG data sets based on each of the four structural classes (all-α, all-β, α + β, and α/β). A (—) indicates that no residues of the given type were found in the particular secondary structural state. N^{s} is the number of residues observed in a particular secondary structural state while N^{sc} is the total number of residues observed in a particular structural class. The standard errors (σ_p) are given in parentheses

Amino acid	P^{α} all-α	P^{α} α + β	P^{α} α/β	P^{β} all-β	P^{β} α + β	P^{β} α/β	P^{t} all-α	P^{t} all-β	P^{t} α + β	P^{t} α/β
Leu	1.04 (0.09)	1.39 (0.19)	1.22 (0.11)	1.02 (0.11)	0.94 (0.18)	1.05 (0.07)	—	0.67 (0.21)	0.77 (0.31)	0.50 (0.14)
Ile	0.98 (0.16)	0.84 (0.18)	0.99 (0.12)	1.14 (0.14)	1.67 (0.27)	1.38 (0.08)	0.43 (0.39)	0.27 (0.16)	0.36 (0.19)	0.79 (0.20)
Val	1.03 (0.10)	1.18 (0.17)	0.82 (0.09)	1.66 (0.10)	1.22 (0.20)	1.62 (0.06)	0.14 (0.16)	0.69 (0.16)	0.52 (0.22)	0.50 (0.13)
Met	1.11 (0.26)	0.90 (0.32)	1.45 (0.25)	1.41 (0.40)	1.30 (0.49)	0.82 (0.16)	0.88 (0.93)	—	0.76 (0.64)	0.50 (0.29)
Phe	0.96 (0.12)	1.02 (0.24)	0.92 (0.14)	1.32 (0.19)	1.56 (0.33)	1.23 (0.10)	2.20 (0.65)	0.47 (0.27)	0.37 (0.30)	0.96 (0.28)
Tyr	0.80 (0.17)	0.73 (0.14)	0.72 (0.13)	1.35 (0.15)	1.26 (0.21)	1.23 (0.11)	1.53 (0.95)	0.54 (0.23)	1.24 (0.36)	1.78 (0.40)
Trp	1.17 (0.27)	0.64 (0.29)	1.11 (0.24)	1.06 (0.21)	1.25 (0.44)	1.26 (0.15)	—	1.24 (0.53)	1.10 (0.74)	0.57 (0.32)
Ala	1.08 (0.08)	1.34 (0.14)	1.15 (0.11)	0.89 (0.09)	0.82 (0.13)	0.98 (0.07)	0.69 (0.28)	0.87 (0.22)	0.91 (0.22)	0.92 (0.18)
Thr	1.15 (0.15)	0.74 (0.14)	0.97 (0.11)	1.15 (0.09)	0.98 (0.19)	1.20 (0.08)	0.28 (0.34)	1.12 (0.22)	0.87 (0.28)	1.11 (0.24)
Ser	0.95 (0.11)	1.05 (0.16)	0.84 (0.09)	0.86 (0.07)	0.65 (0.15)	0.98 (0.07)	1.43 (0.53)	1.08 (0.18)	1.34 (0.32)	1.40 (0.22)
Cys	1.22 (0.38)	1.27 (0.24)	1.03 (0.24)	1.04 (0.16)	0.71 (0.22)	1.01 (0.17)	—	0.83 (0.37)	0.50 (0.37)	0.62 (0.35)
Asn	1.05 (0.17)	0.83 (0.14)	0.87 (0.14)	0.67 (0.10)	1.27 (0.20)	0.66 (0.11)	1.52 (0.85)	1.36 (0.32)	1.32 (0.35)	1.57 (0.34)
Gln	0.95 (0.21)	1.13 (0.21)	1.43 (0.20)	1.06 (0.13)	1.01 (0.23)	0.63 (0.14)	1.44 (0.90)	1.06 (0.33)	1.06 (0.34)	0.66 (0.26)
Asp	0.86 (0.10)	1.06 (0.17)	1.00 (0.12)	0.71 (0.10)	0.98 (0.19)	0.74 (0.09)	2.42 (0.60)	1.24 (0.34)	0.90 (0.34)	1.22 (0.25)
Glu	1.09 (0.12)	1.69 (0.22)	1.37 (0.14)	0.72 (0.12)	0.54 (0.15)	0.59 (0.09)	0.63 (0.22)	0.91 (0.32)	0.53 (0.27)	0.92 (0.22)
His	1.02 (0.13)	1.11 (0.31)	0.95 (0.18)	1.04 (0.20)	1.26 (0.35)	1.17 (0.13)	0.22 (0.25)	0.91 (0.50)	1.08 (0.53)	0.39 (0.22)
Lys	1.01 (0.09)	1.08 (0.14)	1.20 (0.11)	1.00 (0.12)	0.73 (0.14)	0.83 (0.08)	1.18 (0.38)	0.66 (0.25)	1.27 (0.28)	0.86 (0.18)
Arg	0.93 (0.19)	0.91 (0.20)	1.06 (0.17)	1.06 (0.17)	0.99 (0.24)	1.03 (0.12)	—	1.30 (0.45)	0.77 (0.42)	0.90 (0.31)
Gly	0.85 (0.09)	0.47 (0.09)	0.64 (0.07)	0.87 (0.07)	0.94 (0.14)	0.90 (0.07)	2.64 (0.59)	1.69 (0.24)	1.61 (0.29)	1.61 (0.22)
Pro	0.91 (0.15)	0.48 (0.15)	0.72 (0.12)	0.69 (0.10)	0.69 (0.21)	0.73 (0.11)	1.34 (0.79)	1.54 (0.35)	1.62 (0.49)	1.30 (0.30)
N^{s}	1350	636	1110	1414	439	1071	85	249	187	309
N^{sc}							1648	2425	1709	3116

TABLE 3

Percentage of hydrophobic triplets (1-2-5 and 1-4-5) in α- and non-α regions and hydrophobic doublets (1-3) in β- and non-β regions. The percentages are relative to all possible pentapeptides or doublets in the respective spans. Values in parentheses correspond to the total number of hydrophobic triplets and doublets found in the various structural classes. The data base was that of Levitt & Greer (9)

	Whole sample	all-α	all-β	α + β	α/β
		Triplets 1-2-5			
α	7.9 (257)	10.1 (136)	6.2 (9)	6.1 (39)	6.6 (73)
non-α	2.8 (157)	1.0 (3)	2.3 (52)	1.8 (19)	4.2 (84)
		Triplets 1-4-5			
α	8.1 (263)	8.9 (120)	6.2 (9)	6.8 (43)	8.2 (91)
non-α	2.4 (138)	0.3 (1)	2.1 (47)	1.7 (18)	3.6 (73)
		Doublets 1-3			
β	16.8 (491)	0.0 (0)	15.1 (214)	13.4 (59)	20.4 (218)
non-β	6.0 (358)	1.3 (22)	3.7 (37)	9.7 (123)	8.8 (181)

frequency parameters and predictions for the entire LG protein sample. Short range interactions are apparently responsible for only about 60% of the secondary structure in a protein. Long range interactions and more weighting factors should improve the proposed algorithm; however, the pool of known structures is presently inadequate for the necessary statistical significance.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Dr. Carl Moser who arranged financial and data processing support to perform the research at the CECAM Center, University of Paris at Orsay, France. P.A. wishes to acknowledge travel support from the American National Science Foundation and grant support from the United States Public Health Service (no. GM27682) and the American Cancer Society

TABLE 4

Mean composite prediction indices (Q_{pos}) of the proposed method using frequency factors determined from the various data sets. LG and CF refer respectively to the Levitt & Greer (9) and Chou & Fasman (10) protein data samples

Protein class predicted	Frequency factor class	Q_{pos} (LG)	Q_{pos} (CF)
all-α	all-α	0.74	0.66
all-α	all proteins	0.73	0.65
all-β	all-β	0.47	0.53
all-β	all proteins	0.49	0.55
α + β	α + β	0.55	0.56
α + β	all proteins	0.51	0.61
α/β	α/β	0.55	0.58
α/β	all proteins	0.55	0.56

(Faculty Research Award No. FRA173). J.P. and P.P. acknowledge support from the Comisión Asesora de Investigación Cientifica y Técnica of Spain.

REFERENCES

1. Matthews, B.W. (1976) *Ann. Rev. Phys. Chem.* **27**, 493–523
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 532–542
3. Schulz, G.E. & Schirmer, R.H. (1979) *Principles of Protein Structure*, pp. 108–130, Springer-Verlag, New York
4. Sternberg, M.J.E. & Thronton, J.M. (1978) *Nature* **271**, 15–20
5. Chou, P.Y. & Fasman, G.D. (1978) in *Advances in Enzymology* (Meister, A., ed.), vol. 47, pp. 45–148, John Wiley and Sons, New York
6. Chou, P.Y. & Fasman, G.D. (1974) *Biochemistry* **13**, 211–221
7. Chou, P.Y. & Fasman, G.D. (1974) *Biochemistry* **13**, 222–245
8. Levitt, M. & Chothia, C. (1976) *Nature* **261**, 552–558
9. Levitt, M. & Greer, J. (1977) *J. Mol. Biol.* **114**, 181–293
10. Chou, P.Y. & Fasman, G.D. (1977) *J. Mol. Biol.* **115**, 135–175
11. Soderberg, B.O., Sjoberg, B.M., Sonnerstam, V. & Branden, C.I. (1978) *Proc. Natl. Acad. Sci. US* **75**, 5827–5830
12. Hendrickson, W.A., Klippenstein, G.L. & Ward, K.B. (1975) *Proc. Natl. Acad. Sci. US* **72**, 2160–2164
13. Hendrickson, W.A. & Ward, K.B. (1975) *Biochem. Biophys. Res. Commun.* **66**, 1349–1356
14. Klippenstein, G.L., Cote, J.L. & Ludlam, S.E. (1976) *Biochemistry* **15**, 1128–1136
15. Richardson, J.S., Thomas, K.A. & Richardson, D.C. (1975) *Biochem. Biophys. Res. Commun.* **63**, 986–992
16. Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. & Waley, S.G. (1975) *Nature* **255**, 609–614
17. Palau, J. & Puígdoménech, P. (1974) *J. Mol. Biol.* **88**, 457–469
18. Jenkins, G.M. & Watts, D.G. (1968) *Spectral Analysis and Its Applications*, Holden-Day, San Francisco
19. Levitt, M. (1978) *Biochemistry* **17**, 4277–4285
20 Janin, J. & Chothia, C. (1980) *J. Mol. Biol.* **143**, 95–129
21. Lifson, S. & Sander, C. (1979) *Nature* **282**, 109–111
22. Lifson, S. & Sander, C. (1980) in *Protein Folding* (Jaenicke, R., ed.), pp. 289–316, Elsevier, Amsterdam
23. Kendrew, J.C., Watson, H.C., Strandberg, B.E. & Dickerson, R.E. (1961) *Nature* **190**, 666–670

Address:

Patrick Argos
Department of Biological Sciences
Purdue University
West Lafayette, Indiana 47907
USA