Quiocho, F. A., and Lipscomb, W. N. (1971), *Advan. Protein Chem. 25*, 1.
Rifkind, J., and Applequist, J. (1964), *J. Amer. Chem. Soc. 86*, 4207.
Robson, B., and Pain, R. H. (1971), *J. Mol. Biol. 58*, 237.
Robson, B., and Pain, R. H. (1972), *Nature (London), New Biol. 238*, 107.
Sage, H. J., and Fasman, G. D. (1966), *Biochemistry 5*, 286.
Schiffer, M., and Edmundson, A. B. (1967), *Biophys. J. 7*, 121.
Shotton, D. M., and Watson, H. C. (1970), *Nature (London) 225*, 811.
Snell, C. R., and Fasman, G. D. (1972), *Biopolymers 11*, 1723.
Snell, C. R., and Fasman, G. D. (1973), *Biochemistry 12*, 1017.
Sugiyama, H., and Noda, H. (1970), *Biopolymers 9*, 459–469.
Terbojevich, M., Cosani, A., Peggion, E., Quadrifoglio, F., and Crescenzi, V. (1972), *Macromolecules 5*, 622.
Tooney, N. M., and Fasman, G. D. (1968), *J. Mol. Biol. 36*, 355.
Van Wart, H. E., Taylor, G. T., and Scheraga, H. A. (1973), *Macromolecules 6*, 266.
Warashina, A., and Ikegami, A. (1972), *Biopolymers 11*, 529.
Wetlaufer, D. B. (1973), *Proc. Nat. Acad. Sci. U. S. 70*, 697.
Wright, C. S., Alden, R. A., and Kraut, J. (1969), *Nature (London) 221*, 235.
Wu, T. T., and Kabat, E. A. (1971), *Proc. Nat. Acad. Sci. U. S. 68*, 1501.
Wu, T. T., and Kabat, E. A. (1973), *J. Mol. Biol. 75*, 13.
Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B., and Richards, F. M. (1970), *J. Biol. Chem. 245*, 305.
Zimm, B. H., and Bragg, J. K. (1959), *J. Chem. Phys. 31*, 526.
Zimm, B. H., and Rice, S. A. (1960), *Mol. Phys. 3*, 391.

Prediction of Protein Conformation†

Peter Y. Chou and Gerald D. Fasman*

ABSTRACT: A new predictive model for the secondary structure of globular proteins (α helix, β sheet, and β turns) is described utilizing the helix and β-sheet conformational parameters, P_α and P_β, of the 20 amino acids computed in the preceding paper (Chou and Fasman, 1974). This simple and direct method, devoid of complex computer calculations, utilizes empirical rules for predicting the initiation and termination of helical and β regions in proteins. Briefly stated: when four helix formers out of six residues or three β formers out of five residues are found clustered together in any native protein segment, the nucleation of these secondary structures begins and propagates in *both* directions until terminated by a sequence of tetrapeptides, designated as breakers. These rules were successful in locating 88% of helical and 95% of β regions, as well as correctly predicting 80% of the helical and 86% of the β-sheet residues in the 19 proteins evaluated. The accuracy of predicting the three conformational states for all residues, helix, β, and coil, is 77% and shows great improvement over earlier prediction methods which considered only the helix and coil states. The β-turn conformational parameters, P_t, for all 20 amino acids are computed. Their use enables the prediction of chain reversal and tertiary folding in proteins. A procedure for predicting conformational changes in specific regions is also outlined. Despite some evidence of long-range interactions in stabilizing protein folding, the present predictive model illustrates that short-range interactions (*i.e.*, single residue information as represented by P_α and P_β) and medium-range interactions (*i.e.*, neighboring residue information as represented by $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$) play the predominant role in determining protein secondary structure. Although the three-dimensional structures of only 19 proteins have been elucidated to date *via* X-ray studies, the amino acid sequences of hundreds of proteins have already been determined. Since the present predictive model is capable of delineating the helix, β, and coil regions of proteins of known sequence with 80% accuracy, application of this method will be of assistance to all those interested in studying the correlation between protein conformation and biological activity as well as an aid to crystallographers in interpreting X-ray data.

Since experimental evidence has shown that the conformation of proteins is determined predominantly by their amino acid sequence (Anfinsen *et al.*, 1961), many attempts have been made to predict protein structures from their primary sequence. The earlier prediction models classified amino acids[1] qualitatively as helix breakers (Guzzo, 1965), helix formers (Prothero, 1966), as well as helical and antihelical pairs (Periti *et al.*, 1967). Other methods have used matching helical fragments of known conformation (Low *et al.*, 1968) and "helical wheels" to locate hydrophobic arcs in assigning helical regions (Schiffer and Edmundson, 1967). While the above predictive models were based on a few proteins with known conformation, determined by X-ray crystallography, the 60–70% accuracy obtained was encouraging. From conformational energy calculations, Kotelchuck and Scheraga (1969) designated residues as helix making or helix breaking and correctly predicted 61% of helical and 78% of the total residues in four proteins. Using a slightly modified model, Leberman (1971) obtained slightly better results; however, many helical regions were still left unpredicted.

Utilizing the Zimm–Bragg (1959) helix initiation and growth

† Contribution No. 923 from the Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02154. *Received July 2, 1973.* This research was generously supported in part by grants from the U. S. Public Health (GM 17533), National Science Foundation (GB 29204X), American Heart Association (71-1111), and the American Cancer Society (P-577).

[1] The abbreviations for amino acids and polymers conform to the tentative rules of the IUPAC–IUB Commission on Biochemical Nomenclature, as published in *J. Biol. Chem. 247*, 323 (1972).

parameters, σ and s, based on experimental studies of random copolyamino acids, Lewis *et al.* (1970) assigned amino acids as helix breakers (s = 0.385), formers (s = 1.05), and indifferent (s = 1.00) with $\sigma = 5 \times 10^{-4}$ for all residues. The helix probability profiles calculated from these parameters correctly predicted 64% of helical and 68% of the total residues in 11 proteins. Essentially the same accuracy of prediction was obtained when several residues were revised from helix indifferent to formers (Lewis and Scheraga, 1971), and it was suggested that a more satisfactory predictive method may emerge when the σ and s parameters for all 20 amino acids have been determined experimentally.

Using the (ϕ, ψ) angles of the middle amino acid for various tripeptide sequences in proteins of known X-ray structure, Wu and Kabat (1971, 1973) constructed a 20 × 20 table for all tripeptides showing their frequency of occurrence in helical and nonhelical regions. This table was then used in conjunction with the helical wheel method to predict helical regions in cytochrome *c* and various immunoglobulins. Applying informational theory, Robson and Pain (1971) used the distribution of pairs of amino acids separated by 0, 1 . . . . 4 residues, and showed slightly improved computational predictions for five proteins as compared to prediction rules made from single-residue information. Finkelstein and Ptitsyn (1971) reported no correlation due to interactions between adjacent pair residues in helical and nonhelical regions and concluded that the secondary structure of a polypeptide chain depends mainly on side-chain interactions with the backbone rather than side chain–side chain interactions.

Despite the many predictive methods cited above for protein secondary structure, the β-pleated sheet regions have usually been totally neglected, and were included as nonhelical regions along with the irregular random coil segments of proteins. The first attempt at predicting β-sheet regions of proteins was made by Ptitsyn and Finkelstein (1970), but their qualitative rules predicted only one of the six β regions of ribonuclease and two of the six β regions of papain. More recently Kabat and Wu (1973a,b), as well as Nagano (1973), have developed methods for predicting β-sheet regions. However, as discussed in the previous paper (Chou and Fasman, 1974), their results were not as successful as our present predictive model.

In this paper a new predictive model is outlined utilizing the helix and β-sheet potentials of all 20 amino acids that have been quantitatively determined from a statistical analysis of 15 proteins with known conformation (Chou and Fasman, 1973, 1974). These protein conformational parameters P_α and P_β,[2] computed in the previous paper, can be used in predicting the exact sequences of both the helical and β regions of proteins. While the P_α and P_β values can be used to compute the average values, $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$, for any polypeptide fragment, their use in predicting helical and β regions necessitates empirical rules for the initiation and termination of these secondary structural regions. It is the purpose of this paper to formulate a set of predictive rules so that the helix, β-sheet, and coil regions of proteins can be located in a simple direct manner without recourse to complex computer analysis. The overall accuracy of prediction of the secondary structure of globular proteins employing this new procedure is 80%.

Methods

Analysis of helix–coil boundary residues and central helix residues of proteins (Chou and Fasman, 1974) showed that residues with the highest helical potential, P_α, reside mostly at the helix center while strong helix breakers (low P_α values) cluster just beyond the helix ends. Hence helix nucleation could start at the center and propagate in both directions until strong helix breakers terminate helix growth. Furthermore, the frequent occurrence of negatively charged residues, in addition to proline, at the N-terminal helix and positively charged residues at the C-terminal helix (Cook, 1967; Ptitsyn, 1969; Chou and Fasman, 1973, 1974) are utilized in locating the helix boundaries. These facts serve as useful starting points in predicting helical regions of proteins.

Before calculating the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for any protein segment, it is advantageous to first assign all residues, in any amino acid sequence, as formers, breakers, and indifferent to helix and β regions. In this way clusters of formers and breakers for both these secondary structures can be quickly located. In Table I is listed the helix and β-sheet potentials P_α and P_β of 20 amino acids in their hierarchical order with the following classifications: H_α (strong helix former), h_α (helix former), I_α (weak helix former), i_α (helix indifferent), b_α (helix breaker), B_α (strong helix breaker); H_β (strong β former), h_β (β former), I_β (weak β former), i_β (β indifferent), b_β (β breaker), and B_β (strong β breaker). The symbols H and h can be thought of as strong and moderate hydrogen bonding, respectively, with subscripts α and β denoting helical or β conformation. When

TABLE I: Assignment of Amino Acids as Formers, Breakers, and Indifferent for Helical and β-Sheet Regions in Proteins Based on P_α and P_β Values.[a]

Helical Residues[b]	P_α		β-Sheet Residues[c]	P_β	
$Glu^{(-)}$	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
$His^{(+)}$	1.24		Cys	1.30	
Met	1.20		Tyr	1.29	
Gln	1.17	h_α	Phe	1.28	
Trp	1.14		Gln	1.23	h_β
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
$Lys^{(+)}$	1.07	I_α	Trp	1.19	
Ile	1.00		Ala	0.97	I_β
$Asp^{(-)}$	0.98		$Arg^{(+)}$	0.90	
Thr	0.82		Gly	0.81	i_β
Ser	0.79	i_α	$Asp^{(-)}$	0.80	
$Arg^{(+)}$	0.79		$Lys^{(+)}$	0.74	
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	$His^{(+)}$	0.71	b_β
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		$Glu^{(-)}$	0.26	B_β

[a] Chou and Fasman (1974). [b] Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker. I_α assignments are also given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix). [c] β-sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker. b_β assignment is also given to Trp (near the C-terminal β region).

[2] α refers to all helical conformations (α_I, α_{II}, 3_{10}, and distorted helices). β refers to the β-sheet conformation. c refers to the coil conformation or irregular regions of proteins which are not α or β.

these assignments are compared to the classification of Lewis and Scheraga (1971), it is seen that they assigned Ser as a helix breaker (Ser is helix indifferent, Table I), Phe and Tyr as helix indifferent (Phe is a helix former, Tyr is a helix breaker, Table I), with the other residues in agreement with the present findings. While Ptitsyn and Finkelstein (1970) made no attempts in identifying the β-sheet potential of individual amino acids, their general classification that hydrophobic residues are β formers and that charged residues, as well as Pro, are β breakers appears in reasonable agreement with the present assignments for β-sheet residues in Table I. As the data sampling herein of helical residues (890) is more than double that of the β residues (424) in the 15 proteins, it is not expected that the P_α values will change greatly with the possible exceptions of Met, Trp, and Cys since these residues do not occur frequently in proteins. However, future values of P_β based on large numbers of proteins, as their conformation becomes known, could alter some of the present assignments for β-sheet residues.

With these limitations in mind, one can proceed with the prediction of secondary structures in proteins using Table I and the following predictive rules.

A. Search for Helical Regions. 1. Helix Nucleation. Locate clusters of four helical residues (h_α or H_α) out of six residues along the polypeptide chain. Weak helical residues (I_α) count as 0.5 h_α (*i.e.*, three h_α and two I_α residues out of six could also nucleate a helix). Helix formation is unfavorable if the segment contains $^1/_3$ or more helix breakers (b_α or B_α), or less than $^1/_2$ helix formers.

2. *Helix Termination.* Extend the helical segment in *both* directions until terminated by tetrapeptides with $\langle P_\alpha \rangle < 1.00$. The following helix breakers can stop helix propagation: b_4, b_3i, b_3h, b_2i_2, b_2ih, b_2h_2, bi_3, bi_2h, bih_2, and i_4. Once the helix is defined, some of the residues (especially h or i) in the above tetrapeptides may be incorporated at the helical ends. The notations i, b, h in the tetrapeptide breakers also include I, B, and H, respectively. Adjacent β regions can also terminate α regions.

3. Pro cannot occur in the inner helix or at the C-terminal helical end.

4. *Helix Boundaries.* Pro, $Asp^{(-)}$, $Glu^{(-)}$ prefer the N-terminal helical end. $His^{(+)}$, $Lys^{(+)}$, $Arg^{(+)}$ prefer the C-terminal helical end. I_α assignments are given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix) if necessary to satisfy condition A-1.

Rule 1: any segment of six residues or longer in a native protein with $\langle P_\alpha \rangle \geq 1.03$ as well as $\langle P_\alpha \rangle > \langle P_\beta \rangle$, and satisfying conditions A-1 through A-4, is predicted as helical.

B. Search for β-Sheet Regions. 1. β-Sheet Nucleation. Locate clusters of three β residues (h_β or H_β) out of five residues along the polypeptide chain. β-sheet formation is unfavorable if the segment contains $^1/_3$ or more β-sheet breakers (b_β or B_β), or less than $^1/_2$ β-sheet formers.

2. *β-Sheet Termination.* Apply condition A-2 outlined for helix termination in stopping β-sheet propagation, substituting α by β and *vice versa*.

3. $Glu^{(-)}$ occurs rarely in the β region. Pro occurs rarely in the inner β region.

4. *β-Sheet Boundaries.* Charged residues occur rarely at the N-terminal β-sheet end, and infrequently at the inner β region and C-terminal β end. Trp occurs mostly at the N-terminal β-sheet end and rarely at the C-terminal β-end.

Rule 2: any segment of five residues or longer in a native protein with $\langle P_\beta \rangle \geq 1.05$ as well as $\langle P_\beta \rangle > \langle P_\alpha \rangle$, and satisfying conditions B-1 through B-4, is predicted as β sheet.

Thus there are only two basic rules for protein prediction. While the predictive search conditions elaborated above may appear to be extensive, they are given so that incorrected predictions of α and β regions may be minimized. Many of the search conditions overlap or reinforce each other so that with some practice in prediction, only conditions A-1, A-2, and B-1 need be applied to correctly identify the helices and β sheets of proteins. Only when there is overlap of α and β residues is there a need to calculate $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values as prescribed in rules 1 and 2. Even the computation of $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ can be bypassed if the assignments of α and β potentials are grouped in their hierarchical order. Thus a region of six residues with $(H_2h_2ib)_\alpha$ and $(Hh_3iB)_\beta$ assignments can be predicted to be α since there are two strong α formers (H_α) and one α breaker (b_α) compared to one strong β former (H_β) and one strong β breaker (B_β). Hence the discrete P_α and P_β values for 20 amino acids in Table I serve as refinement techniques in identifying the α- and β-sheet potential of a particular fragment in a quantitative manner. However, in most cases it will be found adequate to use only the former, breaker, indifferent assignments, and the termination tetrapeptides (condition A-2) to locate the secondary structural regions of proteins. Furthermore, the utilization of Tables IV and VI on the frequency of helix and β-sheet boundary and central residues from the preceding paper (Chou and Fasman, 1974) will aid in the delineation of secondary structural boundary regions, and will prove useful in predicting regions with similar α and β potentials.

Results

Applying the above conditions and rules, the helical and β-sheet regions in 15 proteins are predicted in Table II, and compared to the latest conformation determined by X-ray crystallography (see Chou and Fasman, 1974). The $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for the predicted α and β regions were also computed and listed in Table II so that the α and β-sheet potentials of the various fragments can be compared relatively. It is seen that 70 of the 81 α regions (86%) as well as 64 of the 66 β regions (97%) are correctly localized. Of the 70 helices predicted, 112 of the 140 helical end boundaries are defined with an accuracy of ± 2 residues (*e.g.*, when the α region 3–18 of myoglobin is predicted as 4–22, the accuracy at the N- and C-helix ends is -1 and $+4$ residues, respectively), while 10 helical ends are predicted incorrectly by more than ± 4 residues. Of the 64 β regions identified, 107 of the 128 β-sheet boundaries are defined with an accuracy of ± 2 residues while 8 β ends are incorrect by more than ± 4 residues. At the same time, 9 α regions and 23 β-sheet regions are predicted at sites in disagreement with X-ray findings. These overpredictions, as well as those secondary structures missed, using the above prediction criteria will be analyzed in the Discussion. The percentage of residues correctly predicted in the conformational state k can be expressed as

$$\%_{ok} = \frac{100(n_k - \text{number incorrect})}{n_k} \quad (1)$$

where k represents the α, β, or coil regions in the native structure of proteins as determined by X-ray analysis. Lewis and Scheraga (1971) have used eq 1 for prediction of helices where $k = \alpha$. The per cent of total residues ($n_\alpha + n_\beta + n_c$) in the proteins predicted correctly is then $\%_{oN} = (\%_{o\alpha} + \%_{o\beta} + \%_{oc})/3$ or more simply

$$\%_{oN} = \frac{100(N - \text{total incorrect})}{N} \quad (2)$$

TABLE II: Comparison of Experimental[a] and Predicted Helical and β-Sheet Regions in the 15 Proteins Included in Computing the Conformational Parameters P_α and P_β.[b]

	Helical Regions[c]				β-Sheet Regions[c]			
	X-ray	Predicted	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	X-ray	Predicted	$\langle P_\beta \rangle$	$\langle P_\alpha \rangle$
Carboxypeptidase A	14–28	13–29	1.15	1.10	32–36	32–38	1.20	1.12
	72–88	72–88	1.11	1.05	49–53	47–52	1.30	1.05
	94–103	98–102	1.11	1.06	60–67	61–68	1.22	1.06
	112–122	116–122	1.15	0.92	104–109	103–111	1.25	1.13
	173–187	173–184	1.15	1.09	–	137–141[f]	1.33	0.98
	215–231	215–233	1.17	0.95	190–196	191–195	1.16	1.14
	254–262[g]	–	(0.86)	(1.00)	200–204	200–204	1.24	1.16
	288–306	289–305	1.13	1.11	–	206–211	1.17	0.76
					–	234–238[f]	1.06	0.67
					239–241	–	(0.95)	(0.74)
					–	243–249	1.39	0.92
					265–271	261–269	1.13	0.91
					–	277–281	1.16	0.99
α-Chymotrypsin	–	55–60[g]	1.10	1.07	29–35	29–34	1.21	1.12
	–	78–84	1.18	0.93	39–47	39–47	1.08	0.81
	–	111–116	1.13	0.98	50–54	50–54	1.27	0.99
	164–173[g]	–	(0.83)	(0.93)	65–68	61–68	1.23	1.04
	234–245	233–245	1.16	1.14	86–91	85–89	1.25	1.16
					103–108	103–108	1.20	1.15
					119–122	117–123	1.24	1.06
					134–140	134–146	1.17	0.84
					155–163	155–163[d]	1.07	1.17
					179–184	179–184	1.22	1.07
					199–203	197–201	1.12	0.87
					206–214	206–214	1.22	1.04
					226–230	226–232	1.20	0.93
Cytochrome b_5	8–15	9–15	1.27	0.86	4–6	4–8	1.23	0.85
	33–38	34–39	1.31	0.75	21–25	21–25	1.29	1.13
	42–49	43–50	1.31	0.84	28–32	29–33	1.23	0.98
	55–62	54–61	1.19	0.90	50–54	–	(0.87)	(0.99)
	64–74	65–74	1.07	0.93	75–79	75–79	1.10	1.08
	80–86[g]	–	(0.92)	(0.76)				
Cytochrome *c*	9–13[e]	2–13	1.07	1.02	None			
	14–18	14–21	1.11	1.08				
	49–54[g]	–	(0.93)	(0.86)	–	46–50	1.15	0.87
	62–70	59–69	1.24	0.87	–	80–85	1.32	1.05
	71–75[g]	–	(0.87)	(1.00)				
	91–101	88–101	1.15	0.92				
Elastase	–	55–63	1.19	0.98	29–36A	30–36	1.17	0.98
	164–170[g]	–	(0.91)	(1.06)	37–47	38–54	1.15	0.98
	237–245	–[f]	(0.97)	(1.08)	50–56			
					65–69	64–69	1.31	0.95
					80–91	79–90	1.25	1.01
					102–110	103–114	1.18	1.16
					–	117–123	1.30	1.04
					133–144	136–144	1.18	0.88
					149–152	150–163	1.18	1.11
					155–163			
					–	165–169	1.18	0.93
					179–188	180–185	1.20	0.94
					192–203	199–216	1.15	1.01
					206–216			
					221–231	220–223	1.14	0.93
						226–235	1.25	0.97
					–	237–243	1.19	1.03

TABLE II (*Continued*)

	Helical Regions[c]				β-Sheet Regions[c]			
	X-ray	Predicted	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	X-ray	Predicted	$\langle P_\beta \rangle$	$\langle P_\alpha \rangle$
α-Hemoglobin	3–18	4–17	1.13	1.00	None			
	20–35	20–36	1.14	0.92				
	36–42[q]	–	(0.84)	(1.08)	–	38–43	1.15	0.88
	52–71	53–73	1.12	1.04				
	80–89	79–84	1.11	0.94				
		86–93	1.20	1.02				
	94–112	96–113	1.12	1.09				
	118–138	120–138	1.07	1.06				
β-Hemoglobin	4–18	6–23	1.28	0.87	None			
	19–34	26–34	1.18	1.10				
	35–41[q]	–	(0.89)	(1.10)	–	35–42	1.12	0.92
	50–56	51–55[q]	1.07	1.14				
	57–76	59–71	1.10	0.98				
		73–78	1.17	0.89				
	85–94	85–98	1.20	1.06				
	99–117	101–118	1.13	1.07				
	123–143	122–135	1.10	1.07				
		137–143	1.26	1.02				
Insulin	A2–8	A2–7[h]	1.06	1.22	B2–7	B1–7	1.15	1.07
	A13–19	A13–18	1.12	0.98	B24–29	B24–28	1.13	0.85
	B9–19	B10–19	1.19	1.15				
Lysozyme	5–15	7–15	1.28	0.93	1–3	2–6	1.19	0.87
	25–35	27–35	1.16	1.01	38–46	38–43	1.05	1.00
	79–84	79–84	1.05	1.01	50–54	50–58	1.16	0.91
	88–99	88–99[i]	1.04	1.15	57–60			
	108–115	107–114	1.08	1.05				
	119–124	119–125[j]	1.10	1.19				
Myogen (Carp)	–[k]	1–6	1.17	1.15	None	None		
	7–15	8–21	1.27	0.97				
	26–33	26–33	1.21	1.04				
	40–51	40–52	1.14	1.09				
	67–71[k]	57–77	1.21	0.95				
	78–89	81–88	1.15	0.95				
	102–107[k]	99–108	1.11	1.07				
Myoglobin	3–18	4–22	1.23	1.02	None	None		
	20–35	24–36	1.05	1.04				
	36–42	38–43	1.24	0.83				
	51–57	48–57	1.27	0.87				
	58–77	58–77	1.06	0.96				
	–[l]	81–85	1.40	0.58				
	86–95	86–97	1.13	0.93				
	100–118	101–119	1.11	1.04				
	124–149	123–128	1.12	1.01				
		130–149	1.18	0.96				
Papain	24–41	26–35	1.16	1.15	5–7	4–9	1.18	0.97
	50–57	50–57	1.20	0.89	–	37–42	1.27	0.95
	67–78	68–77	1.16	1.10	–	91–95	1.27	0.90
	117–126	120–126[m]	1.14	1.14	111–112	110–114	1.33	1.08
	138–143	136–143	1.14	1.00	130–131	130–135	1.35	1.12
					162–167	161–167	1.16	0.99
					169–175	170–174	1.29	1.00
					185–191	185–189	1.24	0.79
					206–208	199–208	1.15	0.85
Ribonuclease S	3–13	2–13	1.24	0.93	41–48	43–48	1.19	1.03
	24–35	28–35[n]	1.04	1.10	60–65	60–65	1.09	1.06
	50–59	49–59	1.14	1.05	69–76	69–76	1.11	0.83
					79–87	79–85	1.17	0.91
					96–110	95–102	1.11	1.02
						105–110	1.31	1.10
					116–124	115–124	1.13	1.02

TABLE II (*Continued*)

	Helical Regions[c]				β-Sheet Regions[c]			
	X-ray	Predicted	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	X-ray	Predicted	$\langle P_\beta \rangle$	$\langle P_\alpha \rangle$
Staphylococcal nuclease	–	5–10	1.22	0.74	12–19	12–18	1.19	1.16
	54–67	56–67	1.18	0.92	21–27	22–27	1.30	1.03
	–	69–76	1.24	0.94	30–36[o]	32–41	1.24	1.09
	99–106	98–110	1.14	1.07	–[o]	88–94	1.16	1.00
	122–134	121–142	1.19	0.84	–[o]	111–115	1.37	1.09
Subtilisin BPN′	5–10[q]	–	(0.81)	(1.05)	–	4–11	1.20	0.87
	14–20	13–19	1.15	0.92	28–32	28–32	1.33	1.14
	64–73	64–75	1.10	1.07	45–50	44–51	1.16	1.03
	103–117	111–116	1.26	1.10	–	79–84	1.29	0.95
	132–145	132–145	1.22	1.00	89–94	89–96	1.21	1.19
	–	195–200	1.27	1.10	–	103–108	1.27	0.95
	223–238	222–238	1.15	1.01	120–124	119–124	1.34	1.04
	242–252[q]	–	(0.94)	(1.03)	148–152	147–152	1.42	1.24
	269–275	267–275	1.21	1.17	–[p]	174–180	1.33	1.12
					–[p]	205–209	1.19	1.00
					–	241–246	1.19	0.97
					–	250–255	1.12	0.95

[a] References to the X-ray data are given in preceding paper (Chou and Fasman, 1974). Slight revisions of α and β regions based on latest X-ray studies are included here for cytochrome b_5 (Mathews *et al.*, 1972a), insulin (Blundell *et al.*, 1972), lysozyme (Imoto *et al.*, 1972), and ribonuclease S (Richards and Wyckoff, 1971). [b] Chou and Fasman (1974). [c] The computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values refer to the predicted regions. The values cited in parentheses refer to α and β regions determined from X-ray but omitted in predictions, also denoted by a dash under the "predicted" column. Overpredictions of α and β regions are denoted by a dash under the "X-ray" column. [d] Pro-161 at C-terminal prohibits α formation (A-3), despite $\langle P_\alpha \rangle > \langle P_\beta \rangle$. [e] 1–11 is α in tuna cytochrome *c* (Takano *et al.*, 1972). [f] 237–243 has α potential ($\langle P_\alpha \rangle = 1.03$), but is predicted as β since $\langle P_\beta \rangle = 1.19$. [g] Helical formation for 51–55 is favored by Pro-51 and Asp-52 at the N-terminal (A-4). [h] Glu-4 prevents β formation (B-3). [i] 88–99 has four β breakers preventing β formation (A-1). [j] Trp-123 at the C-terminal prevents β formation (B-4). [k] Interpretation of the electron density map at the amino terminal is still uncertain (Nockolds *et al.*, 1972). 15 additional α residues (58–62, 65–69, 77, and 98–101) are listed by Kretsinger *et al.* (1972). [l] 82–85 has dihedral angles φ and ψ of the α conformation, and appears to form part of the F helix, 86–95 (Watson, 1969). [m] 120–126 has only one H_β but four H_α residues, so it is predicted as α. [n] 28–35 contains three β-breaking residues (A-1), so it is predicted as α. In ribonuclease A (Kartha *et al.*, 1967), 28–35 is α instead of 24–35. [o] Residues 37–41, 88–92, and 110–114 appear to be in β regions according to Figures 5 and 6 of Cotton *et al.* (1972). [p] The φ, ψ angles for 174–180 and 205–209 are in the β conformation (Alden *et al.*, 1971). [q] Helix appears distorted or 3_{10} type. See Table X for details. [r] φ, ψ angles of Val-139, Val-141, Thr-236, and Tyr-238 (all β formers) are found in the β conformation (Quiocho and Lipscomb, 1971). [s] Recent X-ray refinements of chymotrypsin show 55–59 to be in a 3_{10} conformation (Birktoft and Blow, 1972).

where N is the total number of residues in the protein. Hence, when 27 of 31 α residues and 49 of 55 β residues in ribonuclease S were predicted correctly, the $\%_\alpha$ and $\%_\beta$ obtained were 87 and 89%, respectively, using eq 1. Since 2 α and 2 β residues were overpredicted in this enzyme, the total number of incorrect residues is 14 so that $\%_N$ for ribonuclease ($N = 124$) is 89%. Thus α and β residues missed in the prediction will be reflected in the $\%_\alpha$ and $\%_\beta$ values, while the total residues predicted incorrectly, including overpredictions of secondary structures, will be reflected in the $\%_N$. A summary of the correctness of the predictions as well as overpredictions in the proteins evaluated herein is presented in Table III. When all residues were considered, the predictive accuracy is 81% for α residues, 85% for β-sheet residues, and 77% for total residues (α, β, and coil residues). Since there are three possible conformations, random guesses will yield only 33% for $\%_N$ whereas the above predictive criteria exceed the statistical average by 44%. Most of the predictions on protein secondary structure in the literature involve only identification of helical and nonhelical regions, and as summarized by Kotelchuck and Scheraga (1969) gives 67–78% correctness in $\%_{\alpha+n\alpha}$ ($\%_{\alpha+n\alpha}$ is defined here as % of residues correct when only the helix and nonhelix conformations are considered) and 20–72% correctness in $\%_\alpha$. In order to compare the prediction accuracy of the procedure developed in this paper with literature values, $\%_{\alpha+n\alpha}$ values are listed in the last column of Table III. The 87% correctness obtained in $\%_{\alpha+n\alpha}$ is better than all previous prediction criteria of helical and nonhelical regions as well as 37% higher than the statistical average of 50%.

Another measure in the quality of prediction of a given type of secondary structure may be characterized by (Ptitsyn and Finkelstein, 1970)

$$Q_k = \frac{\%_k + \%_{nk}}{2} \tag{3}$$

where $\%_k$ is defined by eq 1, and $\%_{nk}$ represents the percentage of correctly predicted residues not incorporated in the conformational state k

$$\%_{nk} = \frac{100(n_{nk} - \text{number incorrect})}{n_{nk}} \tag{4}$$

TABLE III: Summary of Correct and Incorrect Predictions of Helical and β-Sheet Regions in the 15 Proteins Included in Computing the Conformational Parameters P_α and P_β.[a]

Protein	$(n_\alpha/n_\beta/N)$[b]	(α_m/α_o)[c]	(β_m/β_o)[d]	Total Incorrect[e]	$\%_{o\alpha}$[f]	Q_α[g]	$\%_{o\beta}$[f]	Q_β[h]	$\%_N$[i]	$\%_{o\alpha+n\alpha}$[j]
Carboxypeptidase A	(113/46/307)	23/4	9/40	71	80	89	81	83	76	91
α-Chymotrypsin	(22/82/241)	10/20	5/18	53	55	73	94	92	78	88
Cytochrome b_5	(48/23/93)	12/3	6/3	20	75	84	74	85	78	84
Cytochrome c	(41/0/104)	12/16	−/11	37	71	73	-	89	64	73
Elastase	(16/124/240)	16/9	24/36	71	0	48	81	75	70	90
α-Hemoglobin	(109/0/141)	14/8	−/6	23	87	81	-	96	84	84
β-Hemoglobin	(115/0/146)	19/8	−/8	28	83	79	-	95	81	82
Insulin	(25/12/51)	3/0	1/1	5	88	94	92	95	90	94
Lysozyme	(54/21/129)	5/2	6/5	16	91	94	71	83	88	95
Myogen	(52/0/108)	5/33	−/0	38	90	66	-	100	65	65
Myoglobin	(121/0/153)	5/11	−/0	19	96	81	-	100	87	87
Papain	(54/30/212)	13/2	4/29	42	76	88	87	86	80	93
Ribonuclease S	(31/55/124)	4/2	6/2	14	87	93	89	93	89	95
Staphylococcal nuclease	(35/22/149)	2/28	4/17	51	94	85	82	85	66	80
Subtilisin BPN′	(86/27/275)	28/13	0/48	72	67	80	100	91	74	85
Total[k]	(922/442/2473)	171/159	65/224	560	81	86	85	87	77	87

[a] Chou and Fasman (1974). [b] The three numbers listed in parentheses represent the number of helical residues (n_α), β-sheet residues (n_β), and total residues (N) in the protein. [c] α_m and α_o are respectively the number of helical residues missed in prediction and overpredicted. [d] β_m and β_o are respectively the number of β-sheet residues missed in prediction and overpredicted. [e] The total residues predicted incorrectly = $(\alpha_m + \alpha_o + \beta_m + \beta_o)$ − (number of incorrect residues counted twice in α_m and β_o or β_m and α_o). [f] $\%_{o\alpha} = 100(n_\alpha - \alpha_m)/n_\alpha$ and $\%_{o\beta} = 100(n_\beta - \beta_m)/n_\beta$, are respectively the % of helical and β residues predicted correctly. [g] $Q_\alpha = (\%_{o\alpha} + \%_{on\alpha})/2$ is the average of the percentages of helical and nonhelical residues predicted correctly in the protein from eq 3, where $\%_{on\alpha} = 100 \times (N - n_\alpha - \alpha_o)/(N - n_\alpha)$. [h] $Q_\beta = (\%_{o\beta} + \%_{on\beta})/2$ is the average of the percentages of β and non-β residues predicted correctly from eq 3, where $\%_{on\beta} = 100(N - n_\beta - \beta_o)/(N - n_\beta)$. [i] $\%_N = 100(N$ − total incorrect$)/N$ is the per cent of total residues in the protein predicted correctly from eq 2. [j] $\%_{o\alpha+n\alpha} = 100(N - \alpha_m - \alpha_o)/N$ represents the per cent of total residues predicted correctly when the β conformation is neglected in the prediction. [k] The figures in the first four columns of the bottom row are summations of the 15 proteins listed above, and are used to compute the percentages in the bottom row.

where $n_{nk} = N - n_k$. It should be noticed that the number incorrect in eq 4 is the number of k residues overpredicted, whereas the number incorrect in eq 1 is the number of k residues missed in prediction. Using ribonuclease S again as an illustrative example where $n_\alpha = 31$, $n_\beta = 55$, $N = 124$, $\alpha_0 = 2$, $\beta_0 = 2$ (see Table III), with $n_{n\alpha} = 93$ and $n_{n\beta} = 69$, then $\%_{on\alpha} = 98\%$ and $\%_{on\beta} = 97\%$ using eq 4. Since $\%_{o\alpha} = 87\%$ and $\%_{o\beta} = 89\%$, as noted earlier, Q_α and Q_β are both 93% from eq 3 in the present case. The qualities of predictions (Q_α and Q_β) for the other proteins are listed in Table III. In general when predicting proteins with low α content, $Q_\alpha > \%_{o\alpha}$ as in the case of chymotrypsin, elastase, papain, and subtilisin (see Table III). Furthermore, predicting heme proteins with little or no β regions tends to yield high Q_β. Hence, Q_α and Q_β tend to give a higher % correct evaluation than does $\%_{o\alpha}$ and $\%_{o\beta}$, and thus could be quite misleading in representing the prediction accuracy of helical and β regions of proteins. Therefore, we have chosen to use instead $\%_{o\alpha}$, $\%_{o\beta}$, and $\%_N$ to measure the prediction accuracy for helical, β, and total residues in a protein.

The 15 proteins in Table II were used in computing the P_α and P_β values of the 20 amino acids in Table I, and these conformational parameters were then utilized in the evaluation of secondary structures in these same proteins. Hence the 80% correctness in correlation (Table III) does not really measure the success of our predictive model of protein conformation unless the same accuracy can be obtained by predicting proteins *not* included in the original set of 15 proteins from which the P_α and P_β values were derived.

Toward this end, four other proteins, whose X-ray structure was known, were used to evaluate the prediction accuracy. These include lamprey hemoglobin (Hendrickson *et al.*, 1973) with high α and no β content, concanavalin A (Edelman *et al.*, 1972; Hardman and Ainsworth, 1972) with high β and no α content, thermolysin (Colman *et al.*, 1972) with moderate α and β content, and pancreatic trypsin inhibitor (Huber *et al.*, 1972) containing both α and β regions. Since trypsin inhibitor has only 58 residues, a complete predictive analysis is given in Figure 1 to illustrate that our predictive model applies to small proteins as well as to large globular ones. Helical (h_α and H_α) and β-forming (h_β and H_β) residues are first enclosed in parentheses and underlined respectively in Figure 1. This gives a quick glimpse as to where α and β regions will nucleate. The assignments of helical (α) and β potential (β) according to Table I are also given under each residue. Hence residues 4, 6, 7 are h_α (Table I) and 2, 3 are I_α (A-4) so that $n_\alpha = 4$ in the regions 2–7 and helix initiation is possible (A-1). Arg-1 is not included as α because it prefers the C-terminal helix rather than the N-terminal (A-4). Pro-8 is not α because it cannot occur at the C-terminal helix (A-3). Furthermore the tetrapeptide 8–11 $(BBbi)_\alpha$ is a helix breaker (A-2) so that the α region is well confined to residues 2–7. Since there are also three β residues in this fragment, the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values are computed, yielding $\langle P_\alpha \rangle = 1.06$ and $\langle P_\beta \rangle = 0.91$. As $\langle P_\alpha \rangle >$

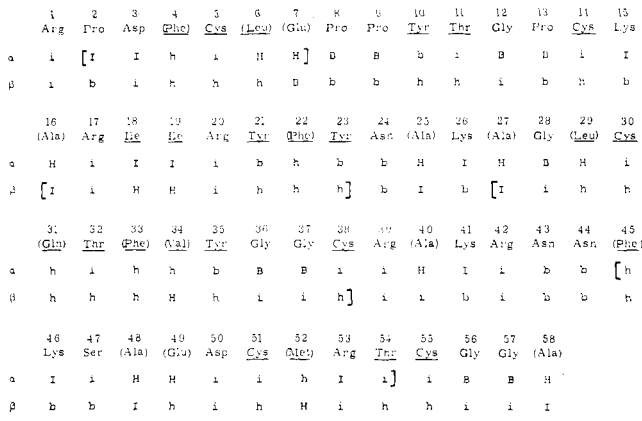

FIGURE 1: Predictive analysis of helical and β-sheet regions in pancreatic trypsin inhibitor. Assignments in the first row under each residue refer to helical potential (α), and in the second row to β potential (β) as defined in Table I. Helical and β residues are also enclosed in parentheses and underlined, respectively. Amino acid sequence as determined by Kassell and Laskowski (1965).

1.03 and $\langle P_\alpha \rangle > \langle P_\beta \rangle$ for residues 2–7, this region is predicted α (rule 1). The α and β potential of residues 2–7 can also be compared by grouping the α and β assignments in this region as $(H_2hI_2i)_\alpha$ and $(h_3ibB)_\beta$. Since there are two strong α formers (H_α) and no α breakers (b_α), and there are no strong β formers (H_β) and two β breakers ($bB)_\beta$, region 2–7 is predicted as α. Similarly the helix breakers $(Iibb)_\alpha$ at 41–44 and $(iBBH)_\alpha$ at 55–58 help to define α region 45–54, where $\langle P_\alpha \rangle = 1.05$ for $(H_2h_2I_2i_4)_\alpha$ is greater than $\langle P_\beta \rangle = 0.99$ for $(Hh_4Ii_2b_2)_\beta$.

The clustering of β residues (underlined, and also denoted by h or H in the second column β) between 18 and 23 plus β breakers $(ibhb)_\beta$ at 12–15 and $(bIbI)_\beta$ at 24–27 set the β region at 16–23, where $\langle P_\beta \rangle = 1.23$ for $(H_2h_3Ii_2)_\beta$ is greater than $\langle P_\alpha \rangle = 0.92$ for $(HhI_2i_2b_2)_\alpha$. Another clustering of β residues between 29 and 38 plus the β breakers $(hbIb)_\beta$ at 23–26 and $(iibi)_\beta$ at 39–42 define the second β region at 27–38, where $\langle P_\beta \rangle = 1.16$ for $(Hh_7Ii_3)_\beta$ is greater than $\langle P_\alpha \rangle = 0.90$ for $(H_2h_3i_3bB_3)_\alpha$. It can be readily seen that this type of hierarchical analysis yields the correct prediction of α and β conformation and hence can be used with confidence in most cases, but computation of $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ is advisable. It is interesting that the β-breaking regions 24–26 $(bIb)_\beta$ serves as a bend forming the antiparallel β-sheet regions 16–23 and 27–38. Regions which were not predicted to be either α or β are classified as irregular or in the coil conformation, and should have both $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values less than unity. That this is so can be seen in Table IV where the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values are listed for all the secondary structural regions, α, β, and coil. Although the β-breaking region 24–26 has a $\langle P_\alpha \rangle = 1.08$, there are not enough helical residues near this tripeptide to nucleate a helical region (A-1). Corroboration that these residues are nonhelical is reflected in the $\langle P_\alpha \rangle = 0.93$ for the region 16–38. It can also be seen in Table IV the α residues 55 and 56 and β residue 24 were unpredicted, while three α (2, 7, 45) and two β (37, 38) residues were overpredicted. Hence 87% of helical and 95% of β residues are predicted correctly (eq 1) and 86% of total residues correctly identified as either α, β, or coil (eq 2).

The secondary structures of concanavalin A, lamprey hemoglobin, and thermolysin were similarly analyzed and the comparison of the predicted regions with X-ray results are given in Table V, along with trypsin inhibitor. As in Table II, the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for the predicted α and β regions are listed

TABLE IV: Conformational Prediction for Pancreatic Trypsin Inhibitor. $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ Values Computed for Helical, β-Sheet, and Coil Regions.

Predicted[a]	X-Ray[b]	$\langle P_\alpha \rangle$[c]	$\langle P_\beta \rangle$[c]
1 c		0.79	0.90
2–7 α	3–6 α	1.06[d]	0.91
8–15 c		0.70	0.90
16–23 β	16–24 β	0.92	1.23
24–26 c		1.08[e]	0.79
27–38 β	27–36 β	0.90	1.16
39–44 c		0.93	0.80
45–54 α	46–56 α	1.05	0.99
55–58 c		0.82	0.98

[a] Based on predictive analysis of Figure 1; α = helical; β = β sheet; c = coil. [b] Huber *et al.* (1972). [c] The computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values refer to the predicted regions. [d] Condition A-4 allows I_α assignments to Pro and Asp at the N-terminal helix end so that condition (A-1) requiring four h_α residues out of six to nucleate a helix is sastisfied. [e] Residues 24–26 are not predicted helical despite $\langle P_\alpha \rangle = 1.08$ since condition A-1 is not satisfied. Furthermore $\langle P_\alpha \rangle = 0.93$ for residues 16–38, indicating that this entire region to be nonhelical.

for comparison. It is seen that 17 of the 18 α regions (94%), as well as 20 of the 22 β regions (91%), are correctly localized. Of the 17 helices predicted, 25 of the 34 helical ends are defined with an accuracy of ± 2 residues, while five helical ends are incorrect by more than ± 4 residues. Of the 20 β regions predicted, 32 of the 40 β-sheet boundaries are defined with an accuracy of ± 2 residues, while only one is incorrect by more than ± 4 residues. At the same time five α- and ten β-sheet regions are predicted at sites in disagreement with X-ray data. The predictive correctness in these four proteins is summarized in Table VI and compares favorably with Table III. Hence the overall accuracy is encouraging in the prediction of proteins whose conformation were not included in the original set used to compute the P_α and P_β values. This shows that the essential information relating to α, β, and coil conformation of the four proteins in Table V does not reside exclusively in the proteins themselves but rather in the intrinsic conformational potentials of individual amino acid residues as calculated from 15 other proteins. Hence the P_α and P_β values of Table I and the predictive model outlined in this paper provide a useful and simple method in correctly identifying the secondary structural regions of any globular protein once its amino acid sequence is known. Thus for the 19 proteins evaluated herein, 88% of helical and 95% of β regions were correctly located, while 80% of the helical and 86% of the β-sheet residues were accurately predicted.

Since the P_α and P_β conformational parameters of Table I were derived from 15 globular proteins, one should not indiscriminately apply the present predictive model to fibrous proteins which have many repetitive amino acid sequences not found in globular proteins. For example, the six different collagen chains listed by Dayhoff (1972) have an average of 51% Pro and Gly residues. Since Pro and Gly are strong helix breakers (Table I) in globular proteins, one may erroneously predict that collagen has a random conformation instead of a triple-helical structure (Pauling and Corey, 1951). In the case of silk fibroin, which has a β-pleated structure composed of

TABLE V: Comparison of Experimental and Predicted Helical and β-Sheet Regions in Four Proteins Not Included in Computing the Conformational Parameters P_α and P_β.[a]

	Helical Regions[b]				β-Sheet Regions[b]			
	X-ray	Predicted	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	X-ray	Predicted	$\langle P_\beta \rangle$	$\langle P_\alpha \rangle$
Concanavalin A[c]	–	38–43	1.13	1.08	4–9	3–12	1.18	1.08
	81–85[g]	81–86	1.13	1.08	25–29	25–29	1.28	0.90
	–	155–160	1.16	1.08	48–55	47–55	1.16	0.95
	–	180–189	1.17	1.00	59–66	60–67	1.14	1.01
					73–78	73–80	1.13	0.97
					92–97	88–96	1.15	1.05
					106–116	106–113	1.14	0.98
					125–132	124–134	1.11	1.09
					140–144	140–144	1.21	1.17
					173–177	173–177	1.13	1.06
					190–199	190–200	1.18	1.12
					209–215	209–215	1.19	1.03
					–	229–234	1.11	1.08
Lamprey hemoglobin[d]	12–29	8–24	1.14	0.96	None			
	30–44	–	(0.99)	(1.10)	–	2–6	1.21	0.89
	45–52	45–52	1.16	0.98	–	35–43	1.27	1.08
	62–66	61–78	1.13	1.00				
	67–88	80–88	1.15	1.12				
	92–106	92–100	1.08	0.93				
	–	104–110	1.14	1.04				
	111–127	115–128	1.21	1.19				
	132–148	133–147	1.13	1.12				
Thermolysin[e]	–	53–58	1.19	0.94	4–13	4–17	1.10	0.88
	65–88	67–74	1.16	0.94	15–32	20–32	1.21	0.84
	137–152	137–150	1.16	1.06	37–46	37–50	1.09	0.94
	159–180	158–180	1.06	0.99	52–58	–	(0.92)	(1.10)
	235–246	238–246	1.07	1.04	60–63	61–66	1.13	1.04
	259–274	261–271	1.03	1.00	–	75–84	1.20	0.80
	280–296	281–295	1.15	1.07	97–106	98–110	1.07	0.95
	302–313	301–313	1.17	1.14	112–116	–	(0.95)	(1.03)
					119–123	120–124	1.21	0.89
					–	127–131	1.27	0.93
					–	151–157	1.21	0.92
					–	192–197	1.26	0.78
					–	221–225	1.24	0.79
					–	251–260	1.21	0.82
					–	272–276	1.31	0.95
Trypsin inhibitor (pancreatic)[f]	3–6	2–7	1.06	0.91	16–24	16–23	1.23	0.92
	45–56	45–54	1.05	0.99	27–36	27–38	1.16	0.90

[a] Chou and Fasman (1974). [b] The computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values refer to the predicted regions. The values cited in parentheses refer to α and β regions determined from X-ray but omitted in predictions, also denoted by a dash under the "predicted" column. Overpredictions of α and β regions are denoted by a dash under the "X-ray" column. [c] The 2.0-Å resolution, as determined by Edelman *et al.* (1972), was used instead of the 2.4-Å resolution of Hardman and Ainsworth (1972). [d] Hendrickson *et al.* (1973). [e] Colman *et al.* (1972). [f] Huber *et al.* (1972). [g] The position of the single helical turn is not stated explicitly but can be traced to residues 81–85 from the stereodrawings (Edelman *et al.*, 1972). However, Hardman and Ainsworth (1972) do state explicitly that 81–85 is helical.

58% Ala and Gly residues (Lucas *et al.*, 1958), application of $(P_\beta)_{Ala} = 0.97$ and $(P_\beta)_{Gly} = 0.81$ would underpredict the β regions. Furthermore, one should approach the conformational prediction of isolated protein fragments with caution. It is well known that the helicity of isolated myoglobin fragments in water is much lower than it is in the native protein, although these peptides do assume greater helical conformation in 95% aqueous methanol (Epand and Scheraga, 1968b). This illustrates that tertiary folding in proteins can provide a nonaqueous type of environment which aids the stability of secondary structures. While utilization of P_α and P_β may lead to erroneous results for isolated protein fragments in water, it does give accurate results for small proteins in their native state (*e.g.*, insulin and pancreatic trypsin inhibitor). Application of the present prediction model in locating the secondary structures of glucagon, lac repressor, and various hormones is

TABLE VI: Summary of Correct and Incorrect Predictions of Helical and β-Sheet Regions in Four Proteins Not Included in Computing the Conformational Parameters P_α and P_β.[a]

Protein	$(n_\alpha/n_\beta/N)$	α_m/a_o	β_m/β_o	Total Incorrect	$\%_{o\alpha}$	Q_α	$\%_{o\beta}$	Q_β	$\%_N$	$\%_{o\alpha+n\alpha}$
Concanavalin A	(5/85/238)	0/23	5/22	50	100	95	94	90	79	90
Lamprey hemoglobin	(117/0/148)	30/10	−/14	45	74	71	-	91	70	73
Thermolysin	(109/69/316)	28/8	17/61	89	74	85	75	75	72	89
Trypsin inhibitor (pancreatic)	(15/19/58)	2/3	1/2	8	87	90	95	95	86	91
Total	(246/173/760)	60/44	23/99	192	76	84	89	86	75	86

[a] Refer to footnotes of Table III.

in progress. Furthermore, the use of P_α and P_β may elucidate the trend of conformational changes of abnormal hemoglobins due to mutations, and of histones due to changes in environmental conditions. These results will be published elsewhere.

Discussion

Utilization of X-Ray Data. Since all predictive models of protein conformation must be compared with the experimental results of X-ray crystallography, it is important to assess the accuracy of the X-ray data. Hence if the secondary structures of proteins can only be delineated with 90% precision by X-ray methods, it is unlikely that predictions of residues in the helical, β, and coil regions of proteins can exceed this number. This would then place an upper limit on predictive accuracy. The uncertainty in identifying residues belonging to a given conformation from X-ray studies has been pointed out by several groups recently: "the fraction of helical structure is hard to assess" (Birktoft and Blow, 1972), and "the counting of amino acid residues in various secondary structures is not clear cut" (Chen *et al.*, 1972). Saxena and Wetlaufer (1971) have raised the point that X-ray data interpretation is limited both by experimental and conceptual resolving power. Since there are different interpretations of secondary structures, especially near helical boundaries and frequently for β regions, any reported X-ray structure should not be considered as denfinitive and final. Hence one should bear these factors in mind when utilizing X-ray data as a basis for predictive models, as well as when comparing predictions with X-ray structure results.

Reliability of P_α and P_β Values. Since the P_α and P_β conformational parameters provide a quantitative measure of the helical and β-sheet potential in any native protein segment (Chou and Fasman, 1974) and can also be used in assigning amino acids as formers, breakers, and indifferent to the α and β conformations (Table I), it is instructive to examine the reliability of P_α and P_β in the present predictive model. Although uncertainty in the interpretation of X-ray data could lead to erroneous assignments of residues in the α and β conformation, thereby resulting in erroneous P_α and P_β values, such errors would likely cancel out if a large number of observations were utilized. That is, random errors become less significant as the number of experimental data increases. Thus when the X-ray data on 15 proteins investigated by different laboratories are used (Chou and Fasman, 1973, 1974), the computed P_α and P_β values based on this large statistical sampling should be more accurate than those based on any individual protein. It is interesting to note that regions 4–8 and 9–15 of cytochrome b_5 were correctly identified as β [$\langle P_\beta \rangle$ = 1.23] and α [$\langle P_\alpha \rangle$ = 1.27], respectively, agreeing with the 2.0-Å resolution data, in contrast to the 2.8-Å X-ray data which showed that these regions were in the coil conformation (Mathews *et al.*, 1972a,b). Hence correct conformational predictions can be made even on regions which were incorrectly identified by X-ray and used in the computation of P_α and P_β. Several β regions, not used in computing the P_β values, were correctly predicted: these include lysozyme 1–3, 38–40; ribonuclease S 60–65, 116–124; and insulin B 2–7, which were later identified by refined X-ray data to be in the β conformation.

Of the 15 proteins used to compute the P_α and P_β values, five are heme proteins whose 637 residues represent 26% of the total residues in the sample, while the rest consists of predominantly proteolytic enzymes. Hence the present statistical analysis appears more balanced in representing globular proteins than the earlier predictive models which were based for the most part on heme proteins. It is encouraging that our predictive model gave the same accuracy ($\%_N$ = 77%) for heme proteins as for nonheme proteins.

In order to assess the constancy of the P_α and P_β values, the conformational data of thermolysin and concanavalin A, with 554 residues (21% α and 29% β), were added to the original sample of 2473 residues. The newly computed P_α and P_β values changed on the average by ±4% and ±8%, respectively, for the 20 amino acids. Residues which showed the greatest changes in P_α were Tyr (0.67), Cys (0.84), Ser (0.73), Pro (0.55), Trp (1.07), and Met (1.13). The greatest changes in P_β were for Pro (0.45), Glu (0.33), Phe (1.53), Asn (0.75), Ser (0.80), Cys (1.16), and Gln (1.10). However, it should be noticed that the hierarchical order of the helical and β potentials in Table I remained essentially the same. With the exceptions of Trp which may be now assigned as a weak α former (I_α) and Phe which may be classified as a strong β former (H_β), there is no need to change the α or β assignments for the other amino acids, since the P_α and P_β values of Table I appear quite constant.

Nucleation of Secondary Structures. Since protein chains are synthesized sequentially in the cell from the N-terminal to the C-terminal end (Dintzis, 1961), most protein predictive models have assumed that helix initiation must also start from the amino end. However, in the previous paper (Chou and Fasman, 1974) it was shown that the strongest helical forming residues, which also have the highest experimental σ values, are found at the central helical cores, so that helix nucleation could start at the center and propagate in both directions until strong helix breakers terminate the growth on both ends. While this mechanism is in contrast to the usual helix folding proposals starting from the N-terminal, it is in agreement with recent experimental evidence that sequential folding probably does

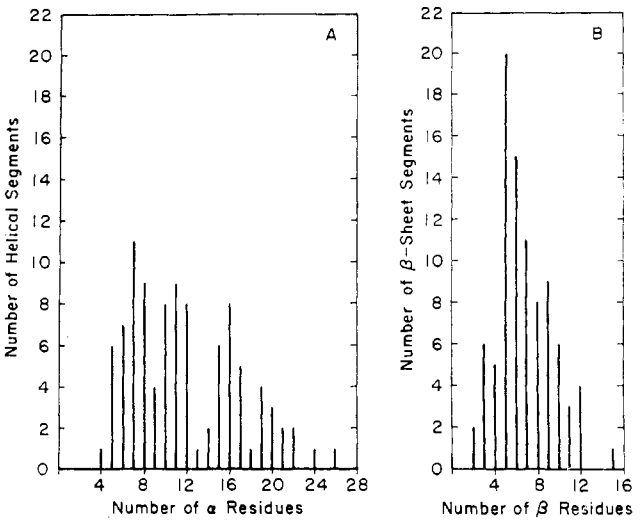

FIGURE 2: Distribution of the number of residues per helical segment (A) and per β-sheet segment (B) for the 19 proteins listed in Tables II and V as determined from X-ray crystallography.

not occur from the amino end. Quiocho and Lipscomb (1971) have pointed out that the tertiary structure of carboxypeptidase A makes it infeasible for orderly folding from the N- to the C-terminal. Similarly, Anfinsen (1972) has suggested, on the basis of experiments with staphylococcal nuclease fragments, that secondary structural folding occurs *after* complete amino acid assembly in proteins. The present predictive model provides search methods for helix and β-sheet initiation as well as propagation by locating clusters of helical and β residues as nucleation centers. This process was demonstrated in correctly locating the α and β regions of pancreatic trypsin inhibitor (Figure 1).

Length of Helical Regions. The search condition for locating four helical residues out of six in determining a helical region (A-1) is based on the fact that 3.6 consecutive residues are required in forming a hydrogen bond in a single turn of the α helix (Pauling *et al.*, 1951). In selecting a criterion for identifying helical regions, Phillips (1967) proposed a minimum of four residues within 40° of the ϕ, ψ angles required for helices. Theoretical results show that the energy of the α-helical conformation becomes very low (*i.e.*, favoring α formation) when the oligopeptide has grown to about six residues (Brant, 1968; De Coen, 1970). In addition, Goodman *et al.* (1969) have shown experimentally, from circular dichroism (CD) and nuclear magnetic resonance (nmr) studies, the critical size for helix formation in oligopeptides to be seven residues. Furthermore, the examination herein of the 99 α regions, determined from X-ray, in the 19 proteins (Figure 2A) shows that the onset of helicity occurs at five or six residues, which agrees well with the theoretical results cited earlier.

From the above knowledge on α-helix formation, it was decided to set the minimum number of residues at six for helix prediction (A-1). In this way four consecutive helical residues will *not* be predicted to be α if their neighboring residues are found to be α breakers. Hence the four α formers (Glu-Glu-Ala-Val) in chymotrypsin 20–23 were correctly predicted as nonhelical since neighboring residues 18–19 (Asn-Gly) and 24–25 (Pro-Gly) are all α breakers (A-2). A hierarchical analysis of the region 18–25 in chymotrypsin shows $(H_3hbB_3)_\alpha$ with a $\langle P_\alpha \rangle = 1.00$. Hence this region is predicted to be nonhelical because of $n_b > n/3$ (A-1), Pro-24 near the C-terminal of segment 18–25 (A-3), and $\langle P_\alpha \rangle < 1.03$ (rule 1), even though there are four α residues out of six (A-1). In the above example, it is seen how the various conditions for helix formation overlap, and application of any one of these would be sufficient for the correct prediction.

Although the length for optimum helical prediction was found to be four residues by Robson and Pain (1971) and to be five residues by Kotelchuck and Scheraga (1969), our present model of at least six residues for α prediction appears in closer agreement to the critical helix size as determined by theoretical, experimental, and X-ray results. Also by not requiring four α residues in a row to initiate a helix as in the earlier models (Kotelchuck and Scheraga, 1969; Leberman, 1971), there is an improvement in α prediction (Table III) using the present helical search method (A-1). Finally, by considering the tetrapeptide α breakers on both sides of a six-residue segment with α forming potential, a total of 14 residues are surveyed even though the computed $\langle P_\alpha \rangle$ for the predicted segments does not include the P_α values of the α-breaking residues. Recently, Ponnuswamy *et al.* (1973) have found that medium-range interactions up to nine residues appear to play a role in conformational stability in proteins in addition to short-range interactions. It can be seen that one of the main features of our present model is the inclusion of both short-range interactions (*i.e.*, single residue information as represented by P_α) as well as medium-range interactions (*i.e.*, neighboring residue information as represented by $\langle P_\alpha \rangle$ for *at least* six residues) in the prediction of protein conformation.

Length of β Regions. The fact that the length of β regions in proteins is shorter than the helical segments can be readily seen in Figure 2. Possible reasons for this have been discussed in the previous paper (Chou and Fasman, 1974). Since, as yet, there is no theoretical treatment as to the minimum length of β segments, the experimental results from X-ray studies (Figure 2B) have been utilized. While there are two 2-residue β segments (papain 111–112, 130–131), six 3-residue β segments, and five 4-residue β-segments, this number increases dramatically to 20 and 15 for the 5-residue and 6-residue β segments, respectively. It should also be noticed (Figure 2B) that there are no β segments longer than 12 residues with the one exception of the 15-residue β segment of ribonuclease S 96–110. In contrast, Figure 2A shows 36 α segments with greater than 12 residues. Hence it was decided to set the search condition for β regions (B-1) as localizing at least three β residues out of five. Of the 113 β segments predicted in the 19 proteins, none had less than five residues [*i.e.*, no violation of condition (B-1)]. While there are 13 β segments with less than five residues as found by X-ray studies, all but one of these were correctly predicted (with the exception of carboxypeptidase 239–241 whose $\langle P_\beta \rangle = 0.95$) since their neighboring residues also favor β formation. Finally, it should be pointed out that single β residues have been found on occasion by X-ray crystallography but were not included in Figure 2B. These include Phe-149 and Glu-183 in papain (Drenth *et al.*, 1971) as well as Thr-11 and Phe-45 in trypsin inhibitor (Huber *et al.*, 1972) which appear hydrogen bonded to other β regions. None of these four single β residues was predicted to be in a β region, even though three of these are β formers, because neighboring residues are not β promoting. It is extremely difficult to predict the location of single β residues, but since they occur rarely in the 19 proteins surveyed here, they have little effect on the overall accuracy of β-sheet prediction.

α- and β-Breaking Tetrapeptides. In earlier predictive models, Kotelchuck and Scheraga (1969), as well as Leberman (1971), proposed that two helix-breaking residues in a row were necessary to terminate helix growth which was propagating from the N- to the C-terminal end. Based on a statistical

analysis of seven proteins, Kotelchuck *et al.* (1969) also noted that helix-breaking residues occurred more frequently at the C-terminal than at the N-terminal nonhelical region. However, the analysis of 15 proteins by Chou and Fasman (1974) (see preceding paper, Table IV, and Figure 1) showed that the helix breakers (Gly, Pro, Tyr, Asn) occurred with about equal frequency at both nonhelical ends (57 and 59 breakers bordering the N- and C-helix ends, respectively), and that the $\langle P_\alpha \rangle$ values for all residues at both nonhelical ends were also similar ($\langle P_\alpha \rangle_{nN} = 0.96$ and $\langle P_\alpha \rangle_{nC} = 0.94$ for the three residues bordering the N- and C-helix ends). Since 3.6 residues are required to form a single turn of the α helix (Pauling *et al.*, 1951), it was decided, in the present predictive model, that any tetrapeptide with 50% or more helix breaking or indifferent residues will be energetically unfavorable for helix propagation. This forms the basis of the helix-breaking tetrapeptides defined in condition A-2 which provide helix termination at both ends. Furthermore any tetrapeptide with 50% or more β breaking of indifferent residues will be assumed to hinder the propagation of β-sheet regions (B-2). In addition, Glu, Trp, Pro, as well as charged residues are helpful in locating the β-sheet terminals as given in conditions B-3 and B-4. Similarly, Pro and charged residues are important aids in locating the helical ends as given in conditions A-3 and A-4.

Tabulation of helix-breaking tetrapeptides showed that they appear approximately equally distributed between the N- and C-terminal helical ends. The same may be said for the β-breaking tetrapeptides. There are approximately 4% of tetrapeptides found at the predicted α and β terminals which are not breaking tetrapeptides as defined in condition A-2 but these regions were terminated for other reasons such as the location of Pro residues (A-3, B-3) and the location preferences of charged residues (A-4, B-4). When there are cases of α- and β-region overlaps, it is essential to compute the respective $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ in these regions as well as utilizing all the conditions and rules so that an accurate prediction is possible. Figure 2 may also be used in support of the prediction of longer stretches of helix than β sheet, if a region has many residues with both α- and β-forming potentials. Overall, it is seen that α- and β-breaking tetrapeptides perform the important function of terminating secondary structures at both ends thereby delineating the length of predicted α and β regions of proteins.

$\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ for Predicted Segments. In order to decide on the magnitude of the cutoff point for $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ in rules 1 and 2 for predicting α and β regions, respectively, the $\langle P_\alpha \rangle > 1$ and $\langle P_\beta \rangle > 1$ values for all the predicted regions (α, β, and coil) with at least five residues were examined. There were only two cases where $1.00 < \langle P_\alpha \rangle < 1.03$, and eight cases where $1.00 < \langle P_\beta \rangle < 1.05$. These regions failed to satisfy the nucleation conditions for helix (A-1) and β sheet (B-1) and hence were predicted to be in the coil conformation, in agreement with X-ray results. It was also found that 54 of the 106 regions with $\langle P_\alpha \rangle > 1$ also have β-forming potential (*i.e.*, $\langle P_\beta \rangle > 1$). Without a predictive model for β sheets, these regions could have been easily interpreted as α since $\langle P_\alpha \rangle > 1$. However, since $\langle P_\beta \rangle > \langle P_\alpha \rangle$ for these identical 54 regions, they were predicted as β, so that excessive overpredictions of α regions were avoided. Likewise, there are 55 of 114 regions with $\langle P_\beta \rangle > 1$ which also have α-forming potential (*i.e.*, $\langle P_\alpha \rangle > 1$) and were predicted as α since $\langle P_\alpha \rangle > \langle P_\beta \rangle$, thus eliminating redundant β predictions.

The cutoff points for helices $\langle P_\alpha \rangle \geq 1.03$ (rule 1) and β sheets $\langle P_\beta \rangle \geq 1.05$ (rule 2) were not made arbitrarily, but in adherence to the search conditions for locating these secondary structural regions. That is, regions predicted above the cutoff points should also satisfy the search requirements A-1 through A-4 for helices and B-1 through B-4 for β sheets. For example, chymotrypsin 1–13 $(H_3h_4I_{i2}b_3)_\beta$ with $\langle P_\beta \rangle = 1.11$ was predicted as coil instead of β because Pro-4 and Pro-8 cannot be in the center of a β region as set by condition B-3. Similarly insulin A8-12 $(Hh_2b_2)_\beta$ with $\langle P_\beta \rangle = 1.11$ was predicted as coil since $(n_b)_\beta > n/3$ prohibits β formation (B-1). There are also five regions with $\langle P_\beta \rangle > \langle P_\alpha \rangle > 1$ but were predicted as α in agreement with X-ray studies because various conditions (B-1–B-4) were not satisfied whereas the A-1–A-4 conditions were met. One such example is ribonuclease 28–35 $(Hh_3I_{i2}b)_\alpha$ with $\langle P_\alpha \rangle = 1.04$ and $(H_2h_2ib_3)_\beta$ with $\langle P_\beta \rangle = 1.10$ [(B-1) with $(n_b)_\beta > n/3$]. The above example indicates that one should be cautious in applying $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values alone in making predictions, and that they should be used in conjunction with the search conditions provided herein.

β Turns in Proteins. Recently β bends (Lewis *et al.*, 1971) and hairpin turns (Kuntz 1972) have been proposed as a mechanism for tertiary folding of globular proteins. These bends or turns consist of only four amino acids, thus enabling a polypeptide chain to reverse itself by nearly 180°, with hydrogen bonding between the CO group of residue i and the NH group of residue $i + 3$ (Venkatachalam, 1968). The prediction of β bends in the light and heavy chains of several immunoglobulins (Bunting *et al.*, 1972) showed that these turns may hold the hypervariable regions together and are strongly conserved. More recently, Crawford *et al.* (1973) characterized the β turns in terms of dihedral angles found from the atomic coordinates of seven proteins, and tabulated the percentage distribution of amino acids in these reverse turns. These data showed that β turns can occur between helices, β sheets, or in the coil region. These authors found that Pro, Asn, Gly, and Tyr occurred with the greatest frequency (ranging from 49 to 45%) in the reverse turns. It is interesting that these four residues are also the strongest helix breakers (Chou and Fasman, 1973, 1974) as shown in Table I.

Of the 125 β turns, in seven proteins, compiled by Crawford *et al.* (1973), 28 were actually part of 3_{10} and distorted helices. These 28 regions which do not show chain reversal were excluded in a new compilation (Table VII herein) which also includes the β turns from five additional proteins: cytochrome *c* (Takano *et al.*, 1972), staphylococcal nuclease (Cotton *et al.*, 1972), elastase (Shotton *et al.*, 1972), papain (Drenth *et al.*, 1971), and thermolysin (Colman *et al.*, 1972). Since these β turns were obtained from stereodiagrams, Table VII should be regarded as tentative, and may be subject to revision when atomic coordinates become available. Nevertheless, it is more comprehensive than similar tables previously published (Lewis *et al.*, 1971; Bunting *et al.*, 1972) Also the exclusion of 3_{10} helices from the data of Crawford *et al.* (1973) may lead to better correlation of the chain-reversing role of β turns. A glance at Table VII shows that Gly and Ser are above average in their frequency of occurrence at all four positions of the β turn, while Asn, Cys, and Tyr are above average at three of the four positions. The amino acids with the greatest frequency of occurrence at the 1st position of the β turn are Asp and Tyr (each with 14%), at the 2nd position is Pro (27%), at the 3rd position is Asn (22%) and in the 4th position is Trp (21%).

In order to compare the β-turn potentials for the 20 amino acids with their α and β potentials, the conformational parameter for the β turn, $P_t = f_t/\langle f_t \rangle$, was obtained[3] in the same man-

[3] $f_t = n_t/n$, where n_t and n are respectively the total occurrences of each residue in the β turns and in all the regions of the 12 proteins. $\langle f_t \rangle = N_t/N$ is the average frequency of all residues in the β-turn regions.

TABLE VII: Frequency of Occurrence of Amino Acids in the β Turns of 12 Proteins.[a]

Amino Acid	n[b]	(i)[c]	f_i[d]	$(i+1)$[c]	f_{i+1}[d]	$(i+2)$[c]	f_{i+2}[d]	$(i+3)$[c]	f_{i+3}[d]
Ala	204	10	0.049	10	0.049	7	0.034	6	0.029
Arg	79	4	0.051	10	0.127[f]	2	0.025	8	0.101
Asn	139	14	0.101	12	0.086	30	0.216	9	0.065
Asp	102	14	0.137	9	0.088	7	0.069	6	0.059
Cys	45	4	0.089	1	0.022	5	0.111	4	0.089
Gln	101	5	0.050	9	0.089	3	0.030	9	0.089
Glu	94	1	0.011	3	0.032	5	0.053	2	0.021
Gly	222	23	0.104	20	0.090	35	0.158	25	0.113
His	60	5	0.083	3	0.050	2	0.033	2	0.033
Ile	118	8	0.068	4	0.034	2	0.017	6	0.051
Leu	156	6	0.038	3	0.019	5	0.032	8	0.051
Lys	150	9	0.060	12	0.080	10	0.067	11	0.073
Met	28	2	0.070	2	0.070	1	0.036	2	0.070
Phe	64	2	0.031	3	0.047	4	0.063	4	0.063
Pro	81	6	0.074	22	0.272	1	0.012	5	0.062
Ser	201	20	0.100	19	0.095	19	0.095	21	0.104
Thr	162	10	0.062	15	0.093	9	0.056	11	0.068
Trp	44	2	0.045	0	0.000	2	0.045	9	0.205
Tyr	118	16	0.136	3	0.025	13	0.110	12	0.102
Val	175	4	0.023	5	0.029	2	0.011	5	0.029
Total	2343	165	$\langle f_i \rangle$[e] = 0.070	165	$\langle f_{i+1} \rangle$[e] = 0.070	165	$\langle f_{i+2} \rangle$[e] = 0.070	165	$\langle f_{i+3} \rangle$[e] = 0.070

[a] The data from seven proteins (carboxypeptidase A, α-chymotrypsin, ribonuclease S, myoglobin, lysozyme, subtilisin BPN', and cytochrome b_5) compiled by Crawford *et al.* (1973) were revised by excluding the 28 helices of the 3_{10} type which were included along with β turns. The data from five proteins (cytochrome *c*, nuclease, elastase, papain, and thermolysin) were obtained from stereodiagrams. [b] n = total occurrence of each residue in the 12 proteins. [c] i, $i+1$, $i+2$, $i+3$ represent the total occurrence of each residue in the 1st, 2nd, 3rd, and 4th position of the β turn. [d] The frequency of occurrence is given by $f_i = i/n$, $f_{i+1} = (i+1)/n$, $f_{i+2} = (i+2)/n$, $f_{i+3} = (i+3)/n$. [e] The average frequency of occurrence of β turns in the 12 proteins is $\langle f_j \rangle = \Sigma j/N = 165/2343 = 0.07$, where $j = i$, $i+1$, $i+2$, or $i+3$. [f] All f_i, f_{i+1}, f_{i+2}, f_{i+3} greater than $\langle f_j \rangle = 0.07$ are underlined, indicating that residues at these positions have above average probability to occur in β turns.

ner as P_α and P_β (Chou and Fasman, 1974) by the process of normalization. In Table VIII it can be seen that there is almost an inverse relationship between the β-turn potential and α potential for the 20 amino acids. That is, the strongest β-turn formers are also the strongest helix breakers. Aside from Glu, the weak β-turn formers (those with low P_t values) are hydrophobic in character, and are strong helix and β formers (as denoted by their high P_α and P_β values). Three of the four highest potential β-turn formers (Asn, Ser, Pro) are also β-sheet breakers. Since the tetrapeptide breakers for α and β regions in our predictive model are now found to be composed of strong β-turn formers, it is possible that once the β-turn conformation is identified it could also serve as termination points at the boundaries of helices and β sheets. An examination of the β turns of lysozyme in Table IX shows this indeed to be true, as there are 10 β turns within ± 2 residues of the helix and β-sheet boundaries. Only the N-terminal β-region 1–3 of lysozyme lacks an adjacent β turn; otherwise chain reversal appears to occur near practically all the terminals of α and β regions. Hence the β turn provides a plausible mechanism for α termination since chain reversals of 180° will be energetically unfavorable for helix propagation. As can be seen in Table IX, the α region 25–35 of lysozyme is preceded and followed by β turns 20–23 and 36–39. Similarly, β turns 104–107 and 115–118 form the boundaries of helix 108–115. The β turn also interestingly plays a dual role in the termination and formation of β sheets. While chain reversal results in β termination (β-turn corners at positions $i+1$ and $i+2$ are not counted as β-sheet residues), it also assists the formation of hydrogen bonding between adjacent sections of antiparallel β sheets. This can be illustrated by the β turn 54–57 of lysozyme which serves as a hinge between the β regions 50–54 and 57–60 as shown in Table IX. Finally, β turns tend to bring distant parts of the polypeptide chain together, so that additional interactions such as that between parallel β sheets, between helices, and between helix and β sheet may lend greater stability to the globular protein.

Prediction of β Turns. The average frequency of β turns in the 12 proteins surveyed is 27% (Table VIII) which is intermediate in frequency to the helices (36%) and the β sheets (17%) found in 15 proteins (Chou and Fasman, 1973, 1974). Together, these three conformations make up 80% of these globular proteins, the remainder being the random conformation. Hence reliable predictive models of protein structure should include these three conformations. The search conditions for α- and β-sheet regions have already been outlined in Methods. Some preliminary guidelines for locating β turns in proteins are given below. The relative probability that a tetrapeptide will form a β turn is (Lewis *et al.*, 1971)

$$p_t = f_i f_{i-1} f_{i-2} f_{i+3} \quad (5)$$

where f_i, f_{i+1}, f_{i-2}, and f_{i+3} are respectively the frequency of occurrence for a certain residue at the 1st, 2nd, 3rd, and 4th

TABLE VIII: Frequency of β Turn Residues in 12 Proteins, and Comparison of β Turn Conformational Parameters, P_t, for 20 Amino Acids with Their P_α and P_β Values.

Amino Acid	n^a	$n_t{}^b$	$f_t{}^c$	$P_t{}^d$	$P_\alpha{}^e$	$P_\beta{}^e$
Gly	222	99	0.446	1.68	0.53	0.81
Asn	139	62	0.446	1.68	0.73	0.65
Ser	201	83	0.413	1.56	0.79	0.72
Pro	81	33	0.407	1.54	0.59	0.62
$Asp^{(-)}$	102	34	0.333	1.26	0.98	0.80
Tyr	118	39	0.331	1.25	0.61	1.29
Cys	45	14	0.311	1.17	0.77	1.30
Trp	44	13	0.295	1.11	1.14	1.19
$Lys^{(+)}$	150	40	0.267	1.01	1.07	0.74
$Arg^{(+)}$	79	21	0.266	1.00	0.79	0.90
Thr	162	43	0.265	1.00	0.82	1.20
Phe	64	12	0.188	0.71	1.12	1.28
$His^{(+)}$	60	11	0.183	0.69	1.24	0.71
Met	28	5	0.179	0.67	1.20	1.67
Ile	118	18	0.153	0.58	1.00	1.60
Ala	204	31	0.152	0.57	1.45	0.97
Gln	101	15	0.149	0.56	1.17	1.23
Leu	156	22	0.141	0.53	1.34	1.22
$Glu^{(-)}$	94	11	0.117	0.44	1.53	0.26
Val	175	14	0.080	0.30	1.14	1.65
Total	$N = 2343$	$N_t = 620$	$\langle f_t \rangle^f = 0.265$	$\langle P_t \rangle^g = 1.00$	$\langle P_\alpha \rangle^g = 1.00$	$\langle P_\beta \rangle^g = 1.00$

[a] n = total occurrence of each residue in the 12 proteins. [b] n_t = total occurrence of each residue in the β turns, where β turn overlap residues were *not* counted twice. [c] $f_t = n_t/n$ is the frequency of residues in β-turn regions. [d] $P_t = f_t/\langle f_t \rangle$ is the conformational parameter for the β turn. [e] As defined in Table I. [f] $\langle f_t \rangle = N_t/N$ is the average frequency of all residues in the β-turn regions. [g] $\langle P_t \rangle$, $\langle P_\alpha \rangle$, and $\langle P_\beta \rangle$ are respectively the average conformational parameter for the β turn, helix, and β regions.

position of a β turn. For example, using eq 5 and Table VII, it is found that lysozyme 17–20 (Leu-Asp-Asn-Tyr) has a p_t = (0.038)(0.088)(0.216)(0.102) = 0.74 $\times$ 10^{-4}. Although the β turn involves only four residues, it is likely that neighboring residues before and after the β turn may be important in stabilizing this conformation. A survey similar to that carried out for adjacent helix and β-sheet regions (Chou and Fasman, 1974) is in progress, so that more than four residues may be utilized in predicting the β turn if warranted. The averaged probability for any given tetrapeptide to be in the β turn is $p_t = \langle f_j \rangle^4 = (0.07)^4 = 0.24 \times 10^{-4}$, where $\langle f_j \rangle$ is the average frequency of occurrence of β turns, for the set of 12 proteins sampled in Table VII. Preliminary investigation shows p_t = 0.5×10^{-4} to be a reasonable cut-off value in predicting the β turns of the 12 proteins studied herein. In particular, all seven β turns of staphylococcal nuclease and five out of six β turns of ribonuclease S were correctly predicted to within ± 1 residue. At the same time there were two and seven β-turn overpredictions, respectively, in the nuclease and ribonuclease, which is a slight improvement over previous predictions (Lewis *et al.*, 1971). Of the 14 β turns of lysozyme, nine were correctly predicted to within ± 1 residue using $p_t = 0.5 \times 10^{-4}$ as the cut-off point. From Table IX, it can be seen that seven of the ten β turns of lysozyme have $\langle P_\alpha \rangle < 1$, $\langle P_t \rangle > 1.10$, and $\langle P_t \rangle > \langle P_\beta \rangle$. For the 108 β turns in carboxypeptidase, chymotrypsin, ribonuclease, lysozyme and subtilisin listed by Crawford *et al.* (1973), the average $\langle P_\alpha \rangle$ was approximately 0.90. Hence, any tetrapeptide with $\langle P_\alpha \rangle < 0.90$ and $\langle P_t \rangle > \langle P_\beta \rangle$ will have a high probability of occurrence in the β turn. This can be confirmed by using eq 5 to check whether the tetrapeptide has a $\langle p_t \rangle > 0.5 \times 10^{-4}$. Since there is almost an inverse relationship between P_t and P_α as shown in Table VIII, the greater the difference between $\langle P_t \rangle > \langle P_\alpha \rangle$ for any tetrapeptide along the protein chain, the more likely this tetrapeptide will be in a β turn.

Using the above rules, the 55 possible tetrapeptide combinations of pancreatic trypsin inhibitor (see Figure 1 for sequence) were examined for β-turn occurrences. These p_t values, obtained from eq 5 and Table VII, are plotted in Figure 3 showing ten sites with β-turn probability greater than 0.5×10^{-4}. (Note residues 36 and 42 are also above the cut-off.) When these ten probable β turns are examined, it is seen that all have $\langle P_\alpha \rangle < 0.90$, and with the exception of 19–22, all have $\langle P_t \rangle > \langle P_\beta \rangle$. The tetrapeptide 19–22 with $\langle P_\beta \rangle = 1.27 > \langle P_t \rangle = 0.89$ was predicted to be within the β region 16–23 in agreement with the X-ray results (Table IV). The prediction of 45–54 as α precludes the inclusion of 53–56 as a β turn even though this tetrapeptide may be part of a 3_{10} helix at the C-terminal. Furthermore, 1–4 is not predicted as a β turn since it is at the N-terminal. The above examples illustrate the importance of analyzing the probable β turns above the p_t = 0.5×10^4 cut-off point, so that overpredictions can be avoided. A rough method of identifying β turns without resorting to p_t computation is the location of three out of four residues in any tetrapeptide with $f_j > 0.07$ through the use of Table VII. When $p_t > 0.5 \times 10^{-4}$ for tetrapeptides at both sites j and $j + 1$, only the site with the higher p_t value is predicted as a β turn. Hence the prediction of 35–38 and 41–44 as β turns did not necessitate the inclusion of 36–39 and 42–45 since the latter tetrapeptides have lower p_t values (Figure 3). Three other tetrapeptides which were predicted as β turns, because of $p_t > 0.5 \times 10^{-4}$, are 8–11, 10–13, and 12–15 which are part of

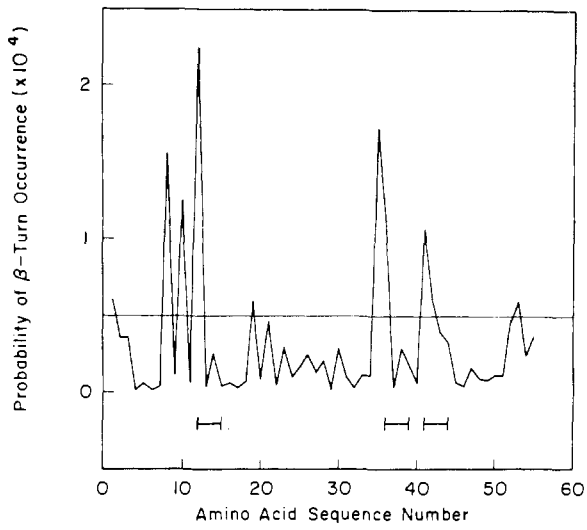

FIGURE 3: Probability that a tetrapeptide β turn begins at site j of pancreatic trypsin inhibitor. The horizontal line corresponds to an arbitrary cut-off value of 0.5×10^{-4}. The horizontal bars indicate the positions of the observed reverse turns starting at j = 12, 36, and 41 (Huber *et al.*, 1972). The sharp peaks at j = 8 and j = 10 are due to Pro-8, Pro-9, and Pro-13 which may assume the poly(Pro) conformation.

the predicted coil region 8–15 (Table IV). It can be seen that the five predicted β turns correspond to the five peaks greater than $p_t = 1 \times 10^{-4}$ in Figure 3. Three of these were found as reverse turns by X-ray (Huber *et al.*, 1972). The two overpredictions involved Pro-8–Pro-9 in tetrapeptide 8–11 and Gly-12–Pro-13 in tetrapeptide 10–13. These regions may assume the poly(Pro) conformation. From the predicted helical, β-sheet, and β-turn regions of pancreatic trypsin inhibitor, a schematic diagram of the secondary structural folding in this protein can be drawn, and is presented in Figure 4. Because of the closely predicted β turns at 8–11, 10–13, and 12–15, the first chain reversal is not evident until residue 12. Although the tetrapeptide 23–26 (Tyr-Asn-Ala-Lys) has $p_t = 0.3 \times 10^{-4}$, which is below the cut-off point (Figure 3), the possibility of a β turn is suggested by three of the four residues having $f_j > 0.07$. Furthermore $\langle P_\alpha \rangle = 0.97$ and $\langle P_t \rangle = 1.11 > \langle P_\beta \rangle = 0.91$ for 23–26, and 24–26 was predicted as a coil region between two β regions 16–23 and 27–38 (Table IV). Hence, chain reversal is represented at 24–26 in Figure 4 to allow hydrogen bonding between the antiparallel β sheets. The predicted β region 27–38 was shortened to 27–34 so that the third chain reversal predicted at 35–38 is accommodated. The β turn could also have been initiated at 36–39 since $p_t = 1.05 \times 10^{-4}$ and $\langle P_t \rangle = 1.38 > \langle P_\beta \rangle = 0.96$, but 35–38 was chosen because of its higher values of $p_t = 1.72 \times 10^{-4}$ and $\langle P_t \rangle = 1.45 > \langle P_\beta \rangle = 1.05$. The fourth and last chain reversal is predicted at 41–44 just before the predicted helical segment 45–54 in agreement with X-ray. Since trypsin inhibitor was not included in the sample of 12 proteins used to derive the f_j values of Table VII as well as the P_t values of Table VIII, the correct identification of the β turns in this protein shows the reliability of the f_j and P_t values used here for β-turn predictions. As can be seen in Figure 4, the β turns in trypsin inhibitor appear as terminators of both α- and β-sheet regions, similar to the case in lysozyme (Table IX). The schematic diagram also helps in visualizing that single β sheet residues Thr-11 and Phe-45 which were found by X-ray studies (Huber *et al.*, 1972) but not predicted (Table IV) can form possible hydrogen bonds with the nearby β region. With the knowledge

TABLE IX: β Turns as Terminators of Helix and β-Sheet Boundaries in Lysozyme from X-Ray Data with Their $\langle P_\alpha \rangle$, $\langle P_\beta \rangle$, and $\langle P_t \rangle$ Values.[a]

Helical Regions[b]	β-Turns[c]	$\langle P_\alpha \rangle$	$\langle P_\beta \rangle$	$\langle P_t \rangle$
5–15	17–20	0.92	0.99	1.18
25–35	20–23	0.64	1.07	1.30
	36–39	0.84	0.83	1.41
79–84	74–77	0.89	0.96	1.27
88–99	98–101	0.97	1.19	0.93
108–115	104–107	0.98	1.03	1.15
	115–118	0.80	1.01	1.22
119–124	124–127	0.77	1.15	1.11
β-Sheet Regions[b]	**β Turns[c]**	**$\langle P_\alpha \rangle$**	**$\langle P_\beta \rangle$**	**$\langle P_t \rangle$**
1–3	–	–	–	–
38–46	36–39	0.84	0.83	1.41
50–54	54–57	1.01	1.22	0.84
57–60	60–63	0.97	1.00	1.17

[a] $\langle P_\alpha \rangle$, $\langle P_\beta \rangle$, and $\langle P_t \rangle$ are respectively the average P_α, P_β, and P_t values (see Table VIII) for the β-turn tetrapeptides. [b] Imoto *et al.* (1972). [c] Crawford *et al.* (1973). β turns listed here are all within ± 2 residues to the helix and β-sheet boundaries.

of β-turn locations, as well as the predicted helices and β sheets in a protein, precise model building may even lead to possible identification of the elusive single β-sheet residues. Similarly, the predicted conformational folding of a protein will also aid in identifying the correct topological linkage of disulfide bridges which can be verified chemically. A horizontal fold along the β-sheet axis of trypsin inhibitor (Figure 4) could bring Cys-14 and Cys-38 together. A pivot of 90° around the β-turn region 41–44 could link Cys-5 with Cys-55, as well as Cys-30 with Cys-51, in agreement with the chemical (Kassell and Laskowski, 1965) and X-ray data (Huber *et al.*, 1972).

Utilizing the β-turn potentials, it is now possible to remove a large number of residues (average 27%), previously termed random, and assign to them a definitive three-dimensional structure. Prediction of the α, β, and β-turn regions allows 80% of the residues in proteins to be conformationally assigned, thus giving a much more satisfactory status to the prediction of protein conformation. Hence the correct prediction of β turns in proteins will lead to greater understanding of protein structure and function.

Unpredicted Helical Regions. While it is satisfying to cite prediction agreement with X-ray results, it is equally instructive to examine secondary regions which were omitted in the predictions. In this way, one can assess whether the unpredicted regions were due to the consequence of neglecting tertiary structure in the model or to an incomplete set of rules for determining the various secondary structures. Since the present predictive model offers distinct α and β potentials for all 20 amino acids based on 15 proteins, improvements over earlier prediction methods were expected and were found as can be seen in Table III. Eleven of the 16 unpredicted α regions of Leberman (1971) were also correctly predicted to be helical (Table II).

Although 80% of all α residues and 88% of all α regions in the 19 proteins were correctly identified by our predictive

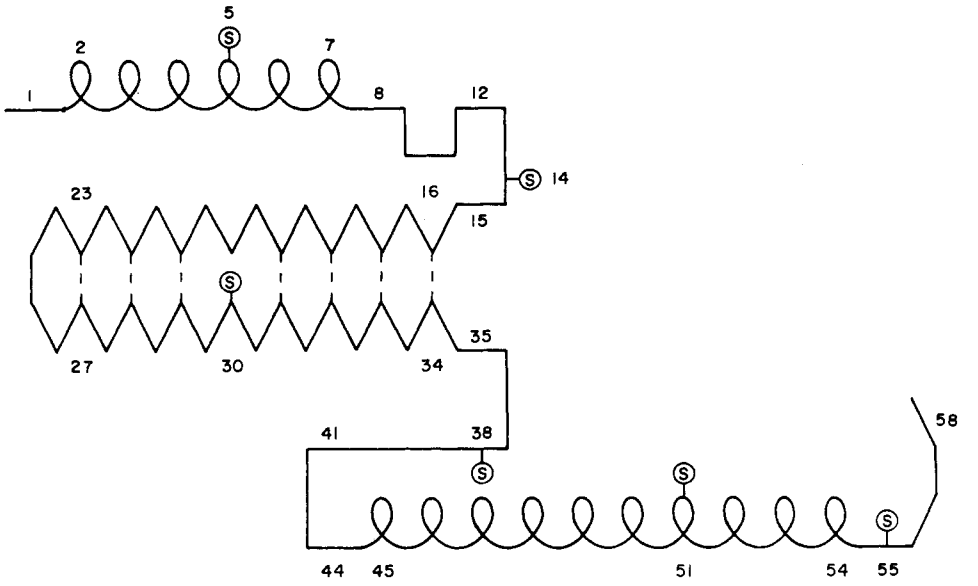

FIGURE 4: Schematic diagram of helical, β-sheet, and reverse β-turn regions *predicted* in pancreatic trypsin inhibitor, based on Table IV and Figure 3. Residues are represented in their respective conformational state: helical (ℓ), β-sheet (Λ), coil (−). Chain reversals denote β-turn tetrapeptides. Hydrogen bonding between the antiparallel β-sheets are represented by dashed lines. Conformational boundary residues are numbered as well as the six Cys residues indicated by Ⓢ. It should be noted that in the present scale each helical loop (ℓ) represents a single helical residue and *not* a single turn consisting of 3.6 residues.

method, 12 helical regions were not predicted since $\langle P_\alpha \rangle < 1$ for these regions, even when neighboring residues were considered (Table X). Six of these regions were predicted to be β sheets since $\langle P_\beta \rangle > 1.05$. It is noteworthy that all of the 12 helices omitted in the predictions were actually distorted α helices or 3_{10} helices: α-hemoglobin 36–42 and β-hemoglobin 35–41 are 3_{10} helices (Perutz *et al.*, 1968); sequences 49–54 and 71–75 of cytochrome *c* are distorted helices (Dickerson *et al.*, 1971); residues 80–86 of cytochrome b_5 (Mathews *et al.*, 1972a) and 259–262 of carboxypeptidase (Crawford *et al.*, 1973) are also reported as 3_{10} helices; Birktoft and Blow (1972) describe residues 164–173 of α-chymotrypsin as "a mixture of α and 3_{10} helix, forming irregular hydrogen bonds." Regions 164–170 and 237–245 of elastase may be distorted helices as in chymotrypsin due to their structural similarity (Shotton and Watson, 1970). Although elastase 237–243 does have helical potential with $\langle P_\alpha \rangle = 1.03$, it was predicted as β-sheet since $\langle P_\beta \rangle = 1.19 > \langle P_\alpha \rangle$ (Table II). Finally, subtilisin 242–252 was reported as a distorted helix by Wright *et al.* (1969), and subtilisin 5–8, 103–106 and 244–247 were analyzed by Crawford *et al.* (1973) to be 3_{10} helices.

From Table X it can be seen that there are many Pro residues as well as Asn and Tyr within and near distorted helices. It appears that these few α-breaking residues can hinder the formation of a regular α helix. Furthermore it is known that poly(Pro) forms a left-handed helix in water and organic solvents (Cowan and McGavin, 1955). Although theoretical calculations show that poly(Tyr) is more stable as a right-handed helix (Ooi *et al.*, 1967; Chen and Woody, 1971), both right-handed helices (Fasman *et al.*, 1964) and left-handed helices (Applequist and Mahr, 1966) have been reported experimentally for poly(Tyr). Although no experimental data are available on poly(Asn), the ϕ, ψ angles of many Asn residues in proteins are in the left-handed helical conformation. Examples include Asn-19, -37, and -77 of lysozyme (Imoto *et al.*, 1972) and Asn-18, -101, and -204 of chymotrypsin (Birktoft and Blow, 1972). Hence these residues (Pro, Asn, and Tyr) in proteins may oppose the regular right-banded twist, thus resulting in helix distortion. In addition, these distorted helices may also be influenced by the steric effects imposed by another adjacent helical or β segment, as can be seen from Table X.

It is worthy of note that Tyr-48 and Pro-71 which interact with the heme of cytochrome *c* (Dickerson *et al.*, 1971) may lend helix stability to regions 49–54 and 71–75 (Table X). Similarly, Thr-39 and Tyr-42 of α-hemoglobin, as well as Thr-38 and Phe-41 of β-hemoglobin, are in contact with the heme (Perutz *et al.*, 1968) so that these α regions form despite $\langle P_\alpha \rangle < 1$. Heme proteins generally have much higher helicity than nonheme proteins, and removal of the heme group in myoglobin causes a 20% decrease in the helical content (Harrison and Blout, 1965). Hence it may be necessary to consider prosthetic group–polypeptide interactions in future predictive models. Since the unpredicted helices (Table X) are not regular α helices, a clearer delineation of the different types of helices such as α_I, α_{II}, and 3_{10} (Némethy *et al.*, 1967), when the X-ray data are more refined, may lead to further improvements in predicting the helical regions of proteins.

Overpredicted helical regions for the 19 proteins analyzed using our predictive model, but not found by X-ray studies, are given in Table XI. An examination of these 12 segments shows that $\langle P_\alpha \rangle > 1.03$ and $\langle P_\alpha \rangle > \langle P_\beta \rangle$; hence they were predicted as α in adherence to rule 1. When four adjacent residues on both sides of the predicted helical segments were included in the computation of $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$, the α potentials, $\langle P_\alpha \rangle$, for 11 of the 12 segments were still larger than unity. Subtilisin 191–204, with $\langle P_\alpha \rangle < 0.98$, was the only exception in Table XI with $\langle P_\alpha \rangle < 1$ when adjacent residues were considered. Chymotrypsin 55–60 was predicted as helical with $\langle P_\alpha \rangle = 1.10$ in disagreement with X-ray data (Blow, 1969). However, more recently Birktoft and Blow (1972) have obtained refined atomic coordinates for tosyl-α-chymotrypsin showing that residues 55–59 are in a 3_{10} conformation, making two hydrogen bonds, although a full turn of a 3_{10} helix was not made. Since elastase 55–63 is homologous to chymotrypsin 55–60, higher resolution studies of elastase may reveal this region to be helical. It is surprising that the myogen sequences 1–6, 16–21, 57–66, and 72–77 were not found to be helical by X-

TABLE X: Helical Regions Found by X-Ray Studies but Omitted in Predictions.

Protein Segment	Unpredicted Helical Regions[a]	$\langle P_\alpha \rangle$[b]	$\langle P_\beta \rangle$[b]
(1) α-Hemoglobin 36–42	*Met-Phe-Leu-Gly*-[*Phe*(*Pro*)*Thr-Thr-Lys-Thr-Tyr**]-Phe(Pro)His-Phe 36 42	0.84 (0.94)	1.08 (1.10)
(2) β-Hemoglobin 35–41	*Leu-Leu-Val-Val*-[*Tyr**(*Pro*)*Trp-Thr-Gln-Arg-Phe*]-Phe-Asp*-Ser-Phe 35 41	0.89 (1.01)[c]	1.10 (1.17)
(3) Cytochrome *c* 49–54	(Pro)Gly-Phe-Thr-Tyr*-[*Thr-Asp-Ala-Asn**-*Lys-Gly*]-Lys-Gly-Ile-Thr 49 54	0.93 (0.86)	0.86 (1.01)
(4) Cytochrome *c* 71–75	*Glu-Tyr**-*Leu-Glu-Asn**-[(*Pro*)*Lys-Lys-Tyr**-*Ile*]-(Pro)Gly-Thr-Lys 71 75	0.87 (0.89)	1.00 (0.91)
(5) Cytochrome b_5 80–86	ILE-ILE-GLY-GLU-LEU-[*His*(*Pro*)*Asp-Asp-Arg-Ser-Lys*]-Ile-Thr-Lys(Pro) 80 86	0.92 (0.95)	0.76 (0.89)
(6) Carboxypeptidase A 254–262	[*Ser-Ile-Asp-Trp-Ser-Tyr**-*Asn**-*Gln-Gly*]-Ile-Lys-TYR*-SER-PHE-THR 254 262	0.86 (0.85)	1.00 (0.98)
(7) α-Chymotrypsin 164–173	Leu(Pro)Leu-Leu-[*Ser-Asn**-*Thr-Asn**-*Cys-Lys-Lys-Tyr**-*Trp-Gly*]-Thr-Lys-Ile-Lys 164 173	0.83 (0.94)	0.93 (0.99)
(8) Elastase 164–170	Tyr*-Leu(Pro)-Thr-Val-[*Asp-Tyr**-*Ala-Ile-Cys-Ser-Ser*]-Ser-Ser-Tyr*-Trp 164 170	0.91 (0.91)	1.06 (1.07)
(9) Elastase 237–245	THR-ARG-VAL-Ser-Ala-Tyr*-Ile-Ser-[*Trp-Ile-Asn**-*Asn**-*Val-Ile-Ala-Ser-Asn**] 237 245	0.97 (0.97)	1.08 (1.10)
(10) Subtilisin BPN′ 5–10	Ala-Gln-Ser-Val-[(*Pro*)*Tyr**-*Gly-Val-Ser-Gln*]-Ile-Lys-Ala(*Pro*)*Ala-Leu* 5 10	0.81 (0.97)	1.05 (1.06)
(11) Subtilisin BPN′ 103–110	[*Gln-Tyr**-*Ser-Trp-Ile-Ile-Asn**-*Gly*]-*Ile-Glu-Trp-Ala-Ile-Ala-Asn** 103 110	0.87 (0.96)	1.14 (1.03)
(12) Subtilisin BPN′ 242–252	*Leu-Ser-Lys-His*(Pro)-Asn*-Trp-[*Thr-Asn**-*Thr-Gln-Val-Arg-Ser-Ser-Leu-Gln-Asn**] 242 252	0.94 (0.92)	1.03 (0.99)

[a] Unpredicted helical regions are denoted by brackets. Helical and β-sheet residues from X-ray studies are respectively italicized and capitalized. Pro is denoted by ( ), Asn and Tyr are denoted by asterisks (*) to indicate their frequent occurrences within and near the unpredicted helices. [b] The computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values refer to the unpredicted helical regions in the brackets. The values in parentheses refer to $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ computed for the region in brackets *plus* four adjacent residues on both sides (*i.e.*, eight additional P_α and P_β values have been used in the averaging), hence taking into account near-neighbor interactions. [c] Though this is the only fragment in Table X whose $\langle P_\alpha \rangle > 1$ when adjacent residues are considered, it is not predicted as helical because rule 1 stating $\langle P_\alpha \rangle > 1.03$ is not satisfied.

TABLE XI: Helical Regions Predicted but Not Found By X-Ray Studies.

Protein Segment	Overpredicted Helical Regions[a]	$\langle P_\alpha \rangle$[b]	$\langle P_\beta \rangle$[b]
(1) α-Chymotrypsin 55–60	TRP-VAL-VAL-THR-[Ala-Ala-His-Cys-Gly-Val]-Thr-Thr-Ser-Asp (+) (−) 55 60	1.10 (1.02)	1.07 (1.15)[c]
(2) α-Chymotrypsin 78–84	Gly-Ser-Ser-Ser-[Glu-Lys-Ile-Gln-Leu-Lys]-Ile-ALA-LYS-PHE-LYS (−) (+) (+) (+) (+) 78 84	1.18 (1.05)	0.93 (0.96)
(3) α-Chymotrypsin 111–116	LYS-LEU-Ser-Thr-[Ala-Ala-Ser-Phe-Ser-Gln]-Thr-Val-SER-ALA (+) 111 116	1.13 (1.07)	0.98 (1.02)
(4) Elastase 55–63	TRP-VAL-MET-THR-[ALA-ALA-His-Cys-Val-Asp-Arg-Glu-Leu]-Thr-PHE-ARG-VAL (+) (−) (+) (−) (+) 55 63	1.19 (1.11)	0.98 (1.15)[c]
(5) Myogen 1–6	[Ala-Phe-Ala-Gly-Val-Leu]-*Asn-Asp-Ala-Asp* (−) (−) 1 6	1.17 (1.12)	1.15 (0.96)
(6) Myogen 16–21	*Ala-Ala-Ala-Leu*-[Glu-Ala-Cys-Lys-Ala-Ala]-Asp-Ser-Phe-Asn (−) (+) (−) 16 21	1.29 (1.22)	0.87 (0.91)
(7) Myogen 57–66	Asp-Lys-Ser-Gly-[Phe-Ile-Glu-Glu-Asp-Glu-Leu-Lys-Leu-Phe]-*Leu-Gln-Asn-Phe* (−) (+) (−) (−) (−) (−) (+) 57 66	1.26 (1.13)	0.89 (0.91)
(8) Myogen 72–77	*Gln-Asn-Phe-Lys*-[Ala-Asp-Ala-Arg-Ala-Leu]-*Thr-Asp-Gly-Glu* (+) (−) (+) (−) (−) 72 77	1.24 (1.11)	0.97 (0.91)
(9) Myoglobin 81–85	*Lys*-Lys-Lys-Gly-[His-His-Glu-Ala-Glu]-*Leu-Lys-Pro-Leu* (+) (+) (+) (+) (+) (−) (−) (+) 81 85	1.40 (1.16)	0.58 (0.75)
(10) Staphylococcal 5–10 nuclease	Ala-Thr-Ser-Thr-[Lys-Lys-Leu-His-Lys-Glu]-Pro-ALA-THR-LEU (+) (+) (+) (+) (−) 5 10	1.22 (1.10)	0.74 (0.90)
(11) Staphylococcal 69–76 nuclease	*Met-Val-Glu*-Asn-[Ala-Lys-Lys-Ile-Glu-Val-Glu-Phe]-Asn-Lys-Gly-Gln (−) (+) (+) (−) (−) (+) 69 76	1.24 (1.13)	0.94 (0.95)
(12) Subtilisin BPN′ 195–200	Ser-Val-Gly-Pro-[Glu-Leu-Asp-Val-Met-Ala]-Pro-Gly-Val-Ser (−) (−) 195 200	1.27 (0.98)[d]	1.10 (1.01)

[a] Overpredicted helical fragments are denoted by brackets. Helical and β-sheet residues from X-ray studies are respectively italicized and capitalized. Charged residues are denoted by (+) or (−) to indicate frequent occurrences in and near overpredicted helical regions. [b] The computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values refer to the overpredicted helical regions in the brackets. The values in parentheses refer to $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ computed for the region in the bracket *plus* four adjacent residues on both sides (*i.e.*, eight additional P_α and P_β values have been used in the averaging), hence taking into account near-neighbor interactions. [c] Though $\langle P_\beta \rangle > \langle P_\alpha \rangle$ when adjacent residues (belonging to the β region) are considered, chymotrypsin 55–60 (containing four h_α and two h_β) and elastase 55–63 (six h_α and three h_β) are still predicted to be helical since there are twice as many h_α residues as h_β residues in these fragments. [d] This is the only fragment in Table XI where $\langle P_\alpha \rangle < 1$ where adjacent residues are considered.

TABLE XII: Overpredicted β-Sheet Regions Containing β Turns.

Protein	Overpredicted β Region[a]	Observed β Turn[b]	Predicted β Turn[c]
Carboxypeptidase[d]	206–213	206–209, 213–216	206–209, 208–211
	234–238	232–235	234–237, 235–238
	243–249	242–245, 244–247	
	277–281	275–278, 277–280	275–278
Elastase[e]	117–123	115–118	
Subtilisin[f]	4–11	5–8	4–7
	79–84	83–86	
	103–108	103–106	102–105, 107–110
	241–246	238–241, 244–247	238–241
Thermolysin	127–131	125–128	122–125, 125–128
	151–157	151–154	
	192–197	194–197	194–197
	221–225	225–228	221–224, 225–228
	251–260	249–252, 259–262	257–260, 259–262
	272–276	276–279	276–279

[a] See Tables II and V for $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for these regions. [b] Observed β turns for carboxypeptidase and subtilisin are from atomic coordinates reported by Crawford *et al.* (1973). Observed β turns for elastase (Shotton *et al.*, 1972) and thermolysin (Colman *et al.*, 1972) are from stereodrawings. [c] Predicted β turns have $p_t > 0.5 \times 10^{-4}$ as calculated from eq 5 and Table VII. [d] The ϕ, ψ angles of residues 209, 236, 238, 248, and 281 appear to be in the β conformation (Quiocho and Lipscomb, 1971). [e] The homologous sequence 118–123 in α-chymotrypsin is β (Birktoft and Blow, 1972). [f] The ϕ, ψ angles of residues 79, 82, 103, and 242 appear to be in the β conformation (Alden *et al.*, 1971).

ray as these regions have high $\langle P_\alpha \rangle$ values due to the presence of many Ala and Glu residues in these segments (see Table XI). However, it should be pointed out that the X-ray analysis for myogen is still ambiguous at the amino terminal as well as in the region 55–62 (Nockolds *et al.*, 1972).

A glance at Table XI shows that many of the overpredicted helices have many charged residues. Hence the charge repulsion of four $Lys^{(+)}$ residues in and near chymotrypsin 78–84 may prevent α formation, as was found by X-ray (Birktoft and Blow, 1972), even though this segment has $\langle P_\alpha \rangle = 1.18$. It is well known that poly(Lys) and poly(Glu) in aqueous solutions are nonhelical when charged (see review, Fasman, 1967). Clustering of charged residues causing side-chain repulsion may also hinder α formation in myogen 57–66, 72–77, myoglobin 81–85, and staphylococcal nuclease 5–10, 69–76 (Table XI). Hence these regions were overpredicted as α since our predictive model places no limit on the number of charged residues within a helical region. Nevetheless, there is some indication that the overpredicted helix in myoglobin 81–85 may be actually helical. Watson (1969) has indicated that the ϕ, ψ angles of myoglobin 82–85 are in the helical conformation but a regular hydrogen bond is prevented by Pro-88 of the F helix 86–95. Although some of the overpredicted helices in Table XI are incorrect due to neglect of excess charge hindering α formation, some of these regions may prove to be correct on refinement of X-ray data. Further rationalization may be necessary due to tertiary structure restrictions.

Unpredicted β-Sheet Regions. Of the 615 β-sheet residues in the 19 proteins surveyed herein, 527 of these were correctly predicted yielding 86% accuracy in β prediction. At the same time, 84 of the 88 β regions (95%) in these proteins were correctly localized, thus showing the validity of the β conformational parameter assignments in Table I, as well as the search method for β sheets. The four β segments which were omitted in prediction but found by X-rays studies are carboxypeptidase 239–241, cytochrome b_5 50–54, thermolysin 52–58 and 112–116. These segments have $\langle P_\beta \rangle < 1$, and lack three β residues out of five residues to nucleate a β-sheet region (B-1). It is interesting that there is not a single β-forming residue with $P_\beta > 1$ in cytochrome b_5 50–54 (Ala-Gly-Asp-Gly-Ala), yet this region forms a parallel β sheet with β region 27–32 (Mathews *et al.*, 1972a). An examination of neighboring residues of 50–54 in cytochrome b_5 shows both regions 42–49 and 55–62 to be α (Table II). Hence it is quite likely that the β-sheet structural integrity of 50–54 is due to the adjacent helices rather than the residues themselves. The above example illustrates that there may be certain cases where even near neighbor residue information (*i.e.*, medium-range interactions) is insufficient to determine the precise conformation, and that long-range interactions (*i.e.*, neighboring structural regions) may sometimes play a more important role. However, the present predictive model shows that for most cases short-range and medium-range interactions are the predominant factor in determining the secondary conformation of a polypeptide chain in proteins.

Overpredicted β-Sheet Regions. Of the 88 β regions found by X-ray studies in the 19 proteins, only four were omitted in prediction. However, there were 32 overpredictions in the 114 β regions predicted as can be seen from Tables II and V. There are 29 Tyr and 27 Thr residues in these 32 regions of overpredicted β sheets. Since Tyr ($P_\beta = 1.29$, $P_t = 1.25$) and Thr ($P_\beta = 1.20$, $P_t = 1.00$) have both β sheet and β turn forming potentials (Table VIII), it is possible that some of the overpredicted β regions are actually involved in β turns. That this is indeed the case can be seen in Table XII where 20 β turns were observed in 15 of the overpredicted β regions. In 16 of these observed β turns $\langle P_t \rangle > \langle P_\beta \rangle$, indicating β-turn preference. Furthermore, 17 tetrapeptides in these overpredicted β regions are predicted as β turns on the basis of $p_t > 0.5 \times 10^{-4}$ (Table XII). This indicates that many of the overpredicted β regions can be eliminated if β-turn prediction is made in conjunction with β-sheet prediction. It can be seen from Table VIII that there is a considerable number of overlaps for the β-turn and β-sheet forming potentials. The P_t and P_β values are quite similar for Tyr, Cys, Trp, Arg, Thr, and His. Hence, when these residues are clustered together, it may be difficult to assess whether the region is β sheet or β turn on the basis of $\langle P_\beta \rangle$ and $\langle P_t \rangle$ alone. Calculations of p_t in thse regions will provide a better indication of β-turn occurrence. A more comprehensive analysis of β-turns and β-sheet regions for all 19 proteins is in progress so that this dilemma may be resolved.

In Table XIII are listed 15 overpredicted β regions with $\langle P_\beta \rangle > 1.05$ and $\langle P_\beta \rangle > \langle P_\alpha \rangle$ which are not apparently involved in β turns. It is noteworthy that the irregular C-helical region of both α- and β-hemoglobin (Perutz *et al.*, 1969) as well as a homologous sequence in the regular B helix of lamprey hemoglobin (Hendrickson *et al.*, 1973) are predicted as β sheet (Table XIII), although the β conformation has not been observed in the hemoglobins. Cytochrome *c* residues 46–50 and 80–85 were also overpredicted as β regions as these sequences were not found by X-ray studies (Dickerson *et al.*, 1971). However, an examination of the five overpredicted β regions in the four heme proteins (Table XIII) reveals that 17 residues

TABLE XIII: β-Sheet Regions Predicted but Not Found by X-Ray Studies.

Protein	Overpredicted β-Sheet Regions[a]	$\langle P_\beta \rangle$[b]	$\langle P_\alpha \rangle$[b]
(1) α-Hemoglobin 38–43	*Leu-Gly-Phe-Pro*-[*Thr*(Thr*)Lys-Thr*(Tyr*)*(Phe)]-Pro(His)(Phe)-Asp 38 43	1.15 (1.01)[c]	0.88 (0.88)
(2) β-Hemoglobin 35–42	(*Leu*)*Leu-Val-Val*-[*Tyr*-Pro-Trp(Thr*)Gln-Arg(Phe)*(Phe)]-Asp(Ser)(Phe)Gly 35 42	1.12 (1.15)	0.92 (0.95)
(3) Lamprey hemoglobin 35–43	*Glu-Thr*-Ser-Gly*-[*Val-Asp-Ile-Leu-Val-Lys(Phe)Phe-Thr**]-*Ser-Thr*-Pro-Ala* 35 43	1.27 (1.14)	1.08 (0.97)
(4) Cytochrome *c* 46–50	Gln-Ala-Pro-Gly-[(Phe)Thr*(Tyr*)*Thr*-Asp*]-*Ala-Asn-Lys-Asn* 46 50	1.15 (0.98)[c]	0.87 (0.85)
(5) Cytochrome *c* 80–85	Pro-Gly-Thr*-Lys-[(Met)(Ile)Phe-Ala-Gly(Ile)]-Lys-Lys-Lys-Thr* 80 85	1.32 (1.13)	1.05 (1.03)
(6) Carboxypeptidase 137–141	Thr*-Ser-Ser-Ser-[Leu-Cys-Val-Gly-Val]-Asp-Ala-Asn-Arg 137 141	1.33 (1.10)	0.98 (0.99)
(7) Concanavalin A 229–234	Gly-Ser-Thr*-Gly-[Arg-Leu-Leu-Gly-Leu-Phe]-Pro-Asp-Ala-Asn 229 234	1.11 (1.01)[c]	1.08 (0.94)
(8) Papain 37–42	*Thr*-Ile-Glu-Gly*-[*Ile-Ile-Lys-Ile-Arg*-Thr*]-Gly-Asn-Leu-Asn 37 42	1.27 (1.02)[c]	0.95 (0.90)
(9) Papain 91–95	Tyr*-Tyr*-Glu-Gly-[Val-Gln-Arg-Tyr*-Cys]-Arg-Ser-Arg-Glu 91 95	1.27 (1.00)[c]	0.90 (0.90)
(10) Staphylococcal 88–94 nuclease	Lys-Tyr*-Gly-Arg-[Gly-Leu-Ala-Tyr*-Ile-Tyr*-Ala]-Asp-Gly-Lys-Met 88 94	1.16 (1.04)[c]	1.00 (0.89)
(11) Staphylococcal 111–115 nuclease	Gly-Leu-Ala-Lys-[Val-Ala-Tyr*-Val-Tyr*]-Lys-Pro-Asn-Asn 111 115	1.37 (1.10)	1.09 (1.07)
(12 Subtilisin BPN′ 174–180	Lys-Tyr*-Pro-Ser-[Val-Ile-Ala-Val-Gly-Ala-Val]-Asp-Ser-Ser-Asn 174 180	1.33 (1.11)	1.12 (1.00)
(13) Subtilisin BPN′ 205–209	Pro-Gly-Val-Ser-[Ile-Gln-Ser-Thr*-Leu]-Pro-Gly-Asn-Lys 205 209	1.19 (1.08)	1.00 (0.89)
(14) Subtilisin BPN′ 250–255	*Val-Arg-Ser-Ser*-[*Leu-Gln-Asn*-Thr*-Thr*-Thr*]-Lys-Leu-Gln-Asp 250 255	1.12 (1.01)[c]	0.95 (0.97)
(15) Thermolysin 75–84	*Val-Asp-Ala-His*-[*Tyr*-Tyr*-Ala-Gly-Val-Thr*-Tyr*-Asp-Tyr*-Tyr**]-*Lys-Asn-Val-His* 75 84	1.20 (1.08)	0.80 (0.89)

[a] Overpredicted β regions are denoted by brackets. Residues which are in the helical regions from X-ray studies are italicized. Tyr and Thr residues are denoted by asterisks (*) to show their frequent occurrences within and near the overpredicted β-sheet regions. The 17 residues which are denoted by ( ) are in contact with the heme. [b] The computed $\langle P_\beta \rangle$ and $\langle P_\alpha \rangle$ values refer to the overpredicted β regions in the brackets. The values in parentheses refer to $\langle P_\beta \rangle$ and $\langle P_\alpha \rangle$ computed for the regions in brackets *plus* two adjacent residues on both sides (*i.e.*, four additional P_β and P_α values have been used in the averaging), hence taking into account near-neighbor interactions. [c] These seven fragments in Table XIII have $\langle P_\beta \rangle <$ 1.05 when adjacent residues are considered, indicating that the overpredicted β regions in brackets are somewhat weak.

within and near these regions are involved in heme–polypeptide interactions. Hence the influence of the porphyrin ring on polypeptide regions in contact with it may stabilize helical formation even if $\langle P_\beta \rangle > \langle P_\alpha \rangle$. This again illustrates that under certain circumstances long-range interactions can play a primary role in the folding of secondary structures.

Overpredicted β regions were also found in nonheme proteins (Table XIII). Carboxypeptidase 137–141 with $\langle P_\beta \rangle = 1.33$ does appear to have Val-139 and Val-141 in the β conformation according to the ϕ, ψ angles reported by Quiocho and Lipscomb (1971). Papain 37–42 was predicted as β since $\langle P_\beta \rangle = 1.27 > \langle P_\alpha \rangle = 0.95$ whereas X-ray studies showed that this region was at the end of helix 24–41 (Drenth *et al.*, 1971). The presence of $Lys^{(+)}$-39 and $Arg^{(+)}$-41 may in fact hinder β formation (B-4), but they do fit the C-terminal helix end proposal (A-4) in agreement with X-ray findings. This shows that overpredicted β regions as well as α regions (Table XI) may be avoided by careful analysis of charge localizations. Papain 91–95, with four out of five β-forming residues, was predicted as β since $\langle P_\beta \rangle = 1.27$, although analysis of adjacent residues shows that this region is weakly β (Table XIII). Staphylococcal nuclease 88–94 and 111–115 were predicted to be β in disagreement with X-ray findings (Arnone *et al.*, 1971), but more recent X-ray analysis showed that 88–92 and 110–114 are in the β conformation (Cotton *et al.*, 1972). Subtilisin 174–180 and 205–209 were predicted to be β although not found by X-ray studies (Wright *et al.*, 1969). However, later refinements of the ϕ, ψ angles of subtilisin do indeed show the β conformation for sequences 174–180 and 205–209 (Alden *et al.*, 1971), in agreement with the predictions herein.

In summary, many of the overpredicted β sheets are due to the occurrence of β turns and heme–protein interactions within these regions. Future predictive models must take these facts into account. Another reason for the apparent greater number of β than α overpredictions is that residues in the β conformation are more difficult to delineate than α regions from X-ray analysis. This is because β regions are generally shorter than helices (Figure 2) and they also involve long-range interactions in the case of parallel β sheets. Hence, future X-ray refinements may show some of the overpredicted β regions (Table XIII) to be actually β. There is some indication that this is true for nuclease 88–94 and 111–115 and subtilisin 174–180 and 205–209. It will be instructive to analyze the ϕ, ψ angles for all the overpredicted β regions when such X-ray data become available. Only then will the shortcomings of the present model be seen clearly, allowing future improvements on predicting β sheets.

Comparison of Estimates of Helix and β-Sheet Content from X-Ray, CD, and Prediction. The use of circular dichroism (CD) in the study of protein conformation has been reviewed by Beychok (1968) and more recently by Adler *et al.* (1973). The CD estimates of the fraction of helix, β, and random coil content in a protein can be obtained by using poly(Lys) as a model compound for the three different conformations (Greenfield and Fasman, 1969). Recently, the CD spectra of proteins with known structure from X-ray studies have been utilized in determining a new set of reference values for the helix, β, and coil conformations (Saxena and Wetlaufer, 1971; Chen *et al.*, 1972). Since the fraction of helicity (f_α) and fraction of β sheet (f_β) in proteins can also be obtained from the predictive model outlined in this paper, they can be compared with the results obtained by CD as well as X-ray studies. Such a comparison is presented in Table XIV where the latest CD results are cited for the 19 proteins studied herein. When there are conformational differences between CD estimates from different laboratories for a given protein, the CD result which was closer to the X-ray estimates was chosen for inclusion in Table XIV. In spite of this, the CD estimates for the percentage of helicity appear on the average about 10% lower than the X-ray results. The main reason for this underprediction is probably that the helical segments in globular proteins are short (see Figure 2A) whereas long polypeptide helical segments were mostly used for the CD reference spectra. Even when the average helix size was considered to be 11 residues, f_α (obtained from CD) was still underestimated for insulin and cytochrome *c*. This is probably due to the fact that these proteins have several helical segments even shorter than 11 (Chen *et al.*, 1972). The same can be said for concanavalin A, cytochrome b_5, and elastase which have several short helices (Table II) resulting in lower f_α estimates from CD. Estimates of f_β from CD are in closer agreement, differing from X-ray results by approximately $\pm 5\%$. The fraction helicity (f_α) estimated from the prediction herein is 6% lower than that obtained from X-ray studies for 12 proteins and 10% higher for seven proteins, with an average of a $\pm 7\%$ difference from X-ray for the 19 proteins. Estimations of f_β from prediction is approximately 6% higher than X-ray findings, reflecting the overpredicted β regions discussed in the previous sections (also Table XII and XIII). It should be noted that the f_α and f_β determined from X-ray are not absolute, as higher resolution studies often lead to slight revisions of the results reported. Several revisions in X-ray determination (see footnote of Table XIV) show slight improvement in agreement with the predicted values. There is recent CD evidence that mammalian cytochrome c_1 has 25% β structure (Yu *et al.*, 1971), so that the 10% β for cytochrome *c* found by the present prediction method is not unreasonable.

These results show that our predictive model for protein conformation yields slightly better accuracy than CD studies when compared with X-ray analysis, especially for estimating the helical content of proteins. In addition, the prediction model has an advantage over CD studies in that it *can* locate where the secondary structural regions are (*e.g.*, Tables II and V). Furthermore, predictions of β turns in proteins, through use of Tables VII and VIII, provide valuable information on protein tertiary folding whereas characteristic CD spectra of the β-turn conformation has yet to be determined. Despite the advantages of protein predictions over CD studies, circular dichroism is still one of the most powerful methods for studying conformational changes of polypeptides and proteins in solution.

One of the main reasons for citing the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for all predicted regions (Tables II and V), as well as discussing these values throughout the text, is to demonstrate that there are many regions in proteins with *both* helical and β-forming potentials, and one conformation is usually preferred depending on the environmental conditions. Hence, although there is only 2% helix in the native structure of concanavalin A, as found by X-ray (Edelman *et al.*, 1972; Hardman and Ainsworth, 1972), McCubbin *et al.* (1971) showed that higher helicity can be induced in the native protein through the use of 2-chloroethanol. Their CD spectra for concanavalin A shows 0% helicity in water and 55% helicity in 70% chloroethanol. These results are not necessarily in conflict with the conformational prediction for concanavalin A herein. As can be seen in Table V, four helical regions are predicted whose 28 residues account for the 12% helicity predicted for concanavalin A (Table XIV). There are also nine regions which were predicted as β but have α potential, $\langle P_\alpha \rangle$, above unity (Table V). It has been demonstrated that organic solvents can induce a conformational change from β sheets to helices for both polypep-

TABLE XIV: Determination of the Fraction Helicity (f_α) and Fraction β Sheet (f_β) of 19 Proteins from X-Ray Studies, CD Spectra, and Prediction from P_α and P_β Values.

Protein	Conformation	X-Ray[a]	CD[b]	Predicted[c]	Protein	Conformation	X-Ray[a]	CD[b]	Predicted[c]
Carboxypeptidase A	α	0.37	0.26[d]	0.31	Lamprey hemoglobin	α	0.79	0.50[t]	0.66
	β	0.15	0.18[d]	0.25		β	0	0[t]	0.09
α-Chymotrypsin	α	0.09	0.08	0.13	Lysozyme	α	0.42	0.29	0.40
	β	0.34	0.10	0.39		β	0.16	0.16	0.16
Concanavalin A	α	0.02	0[e]	0.12	Myogen (Carp)	α	0.48[n]	0.41[j]	0.74
	β	0.38	0.30[e]	0.47		β	0	0[j]	0
Cytochrome b_5	α	0.52	0.19[f]	0.42	Myoglobin	α	0.79	0.77	0.83
	β	0.25	0.36[f]	0.22		β	0	0.02	0
Cytochrome c	α	0.39	0.27	0.43	Papain	α	0.25	0.21	0.20
	β	0	0.06	0.10		β	0.14	0.10	0.26
Elastase	α	0.07	0[g]	0.04	Ribonuclease	α	0.25	0.18	0.23
	β	0.52	0.50[g]	0.57		β	0.44	0.44	0.41
α-Hemoglobin	α	0.77	0.72[h]	0.73	Staphylococcal nuclease	α	0.23[o]	0.27	0.41
	β	0	0[h]	0.04		β	0.15[o]	0.10	0.23
β-Hemoglobin	α	0.79	0.67[h]	0.71	Subtilisin BPN'	α	0.31	0.21[k]	0.26
	β	0	0[h]	0.05		β	0.10[p]	0.31[k]	0.27
Insulin	α	0.49	0.31	0.43	Thermolysin	α	0.34	0.20[l]	0.28
	β	0.24	0.18	0.24		β	0.22	–	0.36
					Trypsin inhibitor (pancreatic)	α	0.26	0.25[m]	0.29
						β	0.33	–	0.33

[a] $f_\alpha = n_\alpha/N$ and $f_\beta = n_\beta/N$ from X-ray studies are computed from data in Tables III and VI under the column $(n_\alpha/n_\beta/N)$ for the respective proteins, where n_α, n_β, and N represent the helical, β, and total residues in a protein. [b] All CD estimates of f_α and f_β are from Chen *et al.* (1972) unless otherwise noted. When f_α and f_β are not explicitly given, they were calculated from the CD data using eq 1 and Table V of Chen *et al.* (1972). All CD studies were done at 25° under neutral pH in aqueous solution. [c] Predicted f_α and f_β are calculated from the α and β regions predicted in Tables II and V. [d] Saxena and Wetlaufer (1971). [e] McCubbin *et al.* (1971). [f] Huntley and Strittmatter (1972). [g] Visser and Blout (1971). [h] Beychok *et al.* (1967). [i] Calculated from $[\theta]_{210} = -17{,}500$ and $[\theta]_{222} = -17{,}000$ for deoxylamprey hemoglobin (Sugita *et al.*, 1968). [j] Based on $[\theta]_{222} = -13{,}870$ for hake parvalbumin (Parello and Péchère, 1971), which is homologous to carp myogen (Nockolds *et al.*, 1972). Another related protein, troponin, was found to have $f_\alpha = 0.42$ from CD (Staprans and Watanabe, 1970). [k] Calculated from $[\theta]_{216} = -10{,}000$ and $[\theta]_{225} = -7520$ for subtilisin Novo (Myers and Glazer, 1971). [l] Estimated from optical rotatory dispersion by Drucker and Yang (1969). [m] Estimated from optical rotatory dispersion by Pospišilova *et al.* (1967). [n] 15 additional α residues are listed by Kretsinger *et al.* (1972) increasing f_α to 0.62 (see footnote *k* of Table III). [o] 17 additional β residues were found by Cotton *et al.* (1972) increasing f_β to 0.26; additional hydrogen-bonded residues may also indicate an increase in f_α (see footnote *o* of Table III). [p] The additional β regions found by Alden *et al.* (1971) in subtilisin give $f_\beta = 0.17$.

tides and proteins. Epand and Scheraga (1968a) have shown that poly(Val) changes from the β conformation to a helix upon addition of methanol to an aqueous solution of the polymer. Timasheff *et al.* (1966) has demonstrated that in 30% methanol, a sharp transition of β-structured regions to helices occurs in β-lactoglobulins. Thus, the nine predicted β regions (with 83 residues) in concanavalin A with $\langle P_\alpha \rangle > 1$ can assume the α conformation in organic solvents. Together with the other 28 helical residues predicted for the native structure, the sum of the total α residues becomes, at minimum, 111 or 47% helicity for concanavalin A in organic solvents, in general agreement with CD studies (McCubbin *et al.*, 1971). Likewise, there are six predicted β regions (with 70 residues) in elastase with $\langle P_\alpha \rangle > 1$, in addition to the nine residue helical fragment with $\langle P_\alpha \rangle = 1.19$ predicted for the native protein (Table II). Hence there are 79 potential helical residues in elastase which accounts for 33% helicity. Although only 7% of elastase is found in the helical conformation by X-ray (Shotton and Watson, 1970), CD studies show that elastase is approximately 35% helical in sodium dodecyl sulfate (Visser and Blout, 1971). It is plausible that these α regions are those with $\langle P_\alpha \rangle > 1$ predicted for elastase in Table II. Thus the $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values computed for the various α and β segments in proteins can help in elucidating the regions potentially capable of undergoing conformational change. This approach may lead to greater understanding of protein folding mechanisms.

Addendum

Recently, the X-ray structure of carp myogen has been refined to 1.85 Å (Kretsinger and Nockolds, 1973). The helical regions have been redefined as segments 8–18, 26–33, 40–51, 60–70, 79–88, and 99–107. The predictions herein for myogen (Table III) are considerably improved when compared to these latest results; thus the accuracy of $\%_{o\alpha}$ increased from 90 to 97%, and that of $\%_{oN}$ increased from 65 to 84%. A recent 2-Å resolution analysis of tuna cytochrome *c* [T. Takano, personal communication] showed helical regions at residues 1–12, 50–54, 61–69, and 89–101, with region 70–75 assuming a helical-like structure with no hydrogen bonds. Since these data are of higher resolution than the 2.8-Å study for horse cytochrome *c*, the predictions herein based on these results showed that $\%_{o\alpha}$

increased from 71 to 85%, and $\%_N$ increased from 64 to 74% (Table III). Furthermore, refined X-ray analysis of staphylococcal nuclease (E. E. Hazen, personal communication) showed helical regions at 54–67, 98–108, and 121–142, while β regions were located at 12–19, 22–27, 30–41, 71–75, 87–94, and 110–112, increasing our prediction accuracy from $\%_N$ = 66 to 84% for nuclease. The above X-ray data corrections thus improve our predictions for myogen, cytochrome *c*, and nuclease, and are encouraging, as they indicate that the conformational parameters P_α and P_β can accurately locate the secondary structural regions of proteins.

References

Adler, A. J., Greenfield, N., and Fasman, G. D. (1973), *Methods Enzymol. 27D*, 675.

Alden, R. A., Birktoft, J. J., Kraut, J., Robertus, J. D., and Wright, C. S. (1971), *Biochem. Biophys. Res. Commun. 45*, 337.

Anfinsen, C. B. (1972). *Biochem. J. 128*, 737.

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H., Jr. (1961), *Proc. Nat. Acad. Sci. U. S. 47*, 1309.

Applequist, J., and Mahr, T. G. (1966), *J. Amer. Chem. Soc. 88*, 5419.

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Jr., Richardson, D. C., Richardson, J. S., and, in part, Yonath, A. (1971), *J. Biol. Chem. 246*, 2302.

Beychok, S. (1968), *in* Poly-α-Amino Acids, Fasman, G. D., Ed., New York, N. Y., Marcel Dekker, p 293.

Beychok, S., Tyuma, I., Benesch, R. E., and Benesch, R. (1967), *J. Biol. Chem. 242*. 2460.

Birktoft, J. J., and Blow, D. M. (1972), *J. Mol. Biol. 68*, 187.

Blow, D. M. (1969), *Biochem. J. 112*, 261.

Blundell, T., Dodson, G. Hodgkin, D., and Mercola, D. (1972), *Advan. Protein Chem. 26*, 279.

Brant, D. A. (1968), *Macromolecules 1*, 291.

Bunting, J. R., Athey, T. W., and Cathou, R. E. (1972), *Biochim. Biophys. Acta 285*, 60.

Chen, A. K., and Woody, R. W. (1971), *J. Amer. Chem. Soc. 93*, 29.

Chen, Y. H., Yang, J. T., and Martinez, H. M. (1972), *Biochemistry 11*, 4120.

Chou, P. Y., and Fasman, G. D. (1973), *J. Mol. Biol. 74*, 263.

Chou, P. Y., and Fasman, G. D. (1974), *Biochemistry 13*, 211.

Colman, P. M., Jansonius, J. N., and Matthews, B. W. (1972), *J. Mol. Biol. 70*, 701.

Cook, D. A. (1967), *J. Mol. Biol. 29*, 167.

Cotton, F. A., Bier, C. J., Day, V. W., Hazen, E. E., Jr., and Larsen, S. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 243.

Cowan, P. M., and McGavin, S. (1955), *Nature (London) 176*, 501.

Crawford, J. L., Lipscomb, W. N., and Schellman, C. G. (1973), *Proc. Nat. Acad. Sci. U. S. 70*, 538.

Dayhoff, M. O. (1972), Atlas of Protein Sequence and Structure, Vol. 5, Silver Spring Md., National Biomedical Research Foundation, p D-291.

De Coen, J. L. (1970), *J. Mol. Biol. 49*, 405.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A., and Margoliash, E. (1971), *J. Biol. Chem. 246*, 1511.

Dintzis, H. M. (1961), *Proc. Nat. Acad. Sci. U. S. 47*, 247.

Drenth, J., Jansonius, J. N., Koekoek, R., and Wolthers, B. G. (1971), *Advan. Protein Chem. 25*, 79.

Drucker, H., and Yang, J. T. (1969), *Fed. Proc., Fed. Amer. Soc. Exp. Biol. 28*, 3306.

Edelman, G. M., Cunningham, B. A., Reeke, G. N., Jr., Becker, J. W., Waxdal, M. J., and Wang, J. L. (1972), *Proc. Nat. Acad. Sci. U. S. 69*, 2580.

Epand, R. F., and Scheraga, H. A. (1968a), *Biopolymers 6*, 1551.

Epand, R. M., and Scheraga, H. A. (1968b), *Biochemistry 7*, 2864.

Fasman, G. D. (1967), *in* Poly-α-Amino Acids, Fasman, G. D., Ed., New York, N. Y., Marcel Dekker, p 499.

Fasman, G. D., Bodenheimer, E., and Lindblow, C. (1964), *Biochemistry 3*, 1665.

Finkelstein, A. V., and Ptitsyn, O. B. (1971), *J. Mol. Biol. 62*, 613.

Goodman, M., Verdini, A. S., Toniolo, C., Phillips, W. D., and Bovey, F. A. (1969), *Proc. Nat. Acad. Sci. U. S. 64*, 444.

Greenfield, N., and Fasman, G. D. (1969), *Biochemistry 8*, 4108.

Guzzo, A. V. (1965), *Biophys. J. 5*, 809.

Hardman, K. D., and Ainsworth, C. F. (1972), *Biochemistry 11*, 4910.

Harrison, S. C., and Blout, E. R. (1965), *J. Biol. Chem. 240*, 299.

Hendrickson, W. A., Love, W. E., and Karle, J. (1973), *J. Mol. Biol. 74*, 331.

Huber, R., Kukla, D., Rühlmann, A., and Steigemann, W. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 141.

Huntley, T. E., and Strittmatter, P. (1972), *J. Biol. Chem. 247*, 4641.

Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C., and Rupley, J. A. (1972), *Enzymes, 3rd Ed. 7*, 666.

Kabat, E. A., and Wu, T. T. (1973a), *Biopolymers 12*, 751.

Kabat, E. A., and Wu, T. T. (1973b), *Proc. Nat. Acad. Sci. U. S. 70*, 1473.

Kartha, G., Bello, J., and Harker, D. (1967), *Nature (London) 213*, 862.

Kassell, B., and Laskowski, M., Sr. (1965), *Biochem. Biophys. Res. Commun. 20*, 463.

Kotelchuck, D., Dygert, M., and Scheraga, H. A. (1969), *Proc. Nat. Acad. Sci. U. S. 63*, 615.

Kotelchuck, D., and Scheraga, H. A. (1969), *Proc. Nat. Acad. Sci. U. S. 62*, 14.

Kretsinger, R. H., and Nockolds, C. E. (1973), *J. Biol. Chem. 248*, 3313.

Kretsinger, R. H., Nockolds, C. E., Coffee. C. J., and Bradshaw, R. A. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 217.

Kuntz, I. D. (1972), *J. Amer. Chem. Soc. 94*, 4009.

Leberman, R. (1971), *J. Mol. Biol. 55*, 23.

Lewis, P. N., Go, N., Gō, M., Kotelchuck, D., and Scheraga, H. A. (1970), *Proc. Nat. Acad. Sci. U. S. 65*, 810.

Lewis, P. N., Momany, F. A., and Scheraga, H. A. (1971), *Proc. Nat. Acad. Sci. U. S. 68*, 2293.

Lewis, P. N., and Scheraga, H. A. (1971), *Arch. Biochem. Biophys. 144*, 576.

Low, B. W., Lovell, F. M., and Rudko, A. D. (1968), *Proc. Nat. Acad. Sci. U. S. 60*, 1519.

Lucas, F., Shaw, J. T. B., and Smith, S. G. (1958), *Advan. Protein Chem. 13*, 107.

McCubbin, W. D., Oikawa, K., and Kay, C. M. (1971), *Biochem. Biophys. Res. Commun. 43*, 666.

Mathews, F. S., Argos, P., and Levine, M. (1972a), *Cold Spring Harbor Symp. Quant. Biol. 36*, 387.

Mathews, F. S., Levine, M., and Argos, P. (1972b). *J. Mol. Biol. 64*, 449.

Myers, B., II, and Glazer, A. N. (1971), *J. Biol. Chem. 246*, 412.
Nagano, K. (1973), *J. Mol. Biol. 75*, 401.
Némethy, G., Phillips, D. C., Leach, S. J., and Scheraga, H. A. (1967), *Nature (London) 214*, 363.
Nockolds, C. E., Kretsinger, R. H., Coffee, C. J., and Bradshaw, R. A. (1972), *Proc. Nat. Acad. Sci. U. S. 69*, 581.
Ooi, T., Scott, R. A., Vanderkooi, G., and Scheraga, H. A. (1967), *J. Chem. Phys. 46*, 4410.
Parello, J., and Péchère, J. F. (1971), *Biochimie 53*, 1079.
Pauling, L., and Corey, R. B. (1951), *Proc. Nat. Acad. Sci. U. S. 37*, 272.
Pauling, L., Corey, R. B., and Branson, H. R. (1951), *Proc. Nat. Acad. Sci. U. S. 37*, 205.
Periti, P. F., Quagliarotti, G., and Liquori, A. M. (1967), *J. Mol. Biol. 24*, 313.
Perutz, M. F., Muirhead, H., Cox, J. M., and Goaman, L. C. G. (1968), *Nature (London) 219*, 131.
Phillips, D. C. (1967), *Proc. Nat. Acad. Sci. U. S. 57*, 484.
Ponnuswamy, P. K., Warme, P. K., and Scheraga, H. A. (1973), *Proc. Nat. Acad. Sci. U. S. 70*, 830.
Pospišilová, D., Meloun, B., Fric̆, I., and Šorm, F. (1967), *Collect. Czech. Chem. Commun. 32*, 4108.
Prothero, J. W. (1966), *Biophys. J. 6*, 367.
Ptitsyn, O. B. (1969), *J. Mol. Biol. 42*, 501.
Ptitsyn, O. B., and Finkelstein, A. V. (1970), *Biofizika 15*, 757.
Quiocho, F. A., and Lipscomb, W. N. (1971), *Advan. Protein Chem. 25*, 1.
Richards, F. M., and Wyckoff, H. W. (1971), *Enzymes, 3rd Ed. 4*, 647.
Robson, B., and Pain, R. H. (1971), *J. Mol. Biol. 58*, 237.
Saxena, V. P., and Wetlaufer, D. B. (1971), *Proc. Nat. Acad. Sci. U. S. 68*, 969.
Schiffer, M., and Edmundson, A. B. (1967), *Biophys. J. 7*, 121.
Shotton, D. M., and Watson, H. C. (1970), *Nature (London) 225*, 811.
Shotton, D. M., White, N. J., and Watson, H. C. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 91.
Staprans, I., and Watanabe, S. (1970), *J. Biol. Chem. 245*, 5962.
Sugita, Y., Dohi, Y., and Yoneyama, Y. (1968), *Biochem. Biophys. Res. Commun. 31*, 447.
Takano, T., Swanson, R., Kallai, O. B., and Dickerson, R. E. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 397.
Timasheff, S. N., Townend, R., and Mescanti, L. (1966), *J. Biol. Chem. 241*, 1863.
Venkatachalam, C. M. (1968), *Biopolymers 6*, 1425.
Visser, L., and Blout, E. R. (1971), *Biochemistry 10*, 743.
Watson, H. C. (1969), *Progr. Stereochem. 4*, 299.
Wright, C. S., Alden, R. A., and Kraut, J. (1969), *Nature (London) 221*, 235.
Wu, T. T., and Kabat, E. A. (1971), *Proc. Nat. Acad. Sci. U. S. 68*, 1501.
Wu, T. T., and Kabat, E. A. (1973), *J. Mol. Biol. 75*, 13.
Yu, C. A., Yong, F. C., Yu, L., and King, T. E. (1971), *Biochem. Biophys. Res. Commun. 45*, 508.
Zimm, B. H., and Bragg, J. K. (1959), *J. Chem. Phys. 31*, 526.

Influence of Temperature on the Intrinsic Viscosities of Proteins in Random Coil Conformation†

Faizan Ahmad and Ahmad Salahuddin*

ABSTRACT: Measurements have been made of the intrinsic viscosities of proteins consisting of one polypeptide chain, in 6 M guanidine hydrochloride plus β-mercaptoethanol at different temperatures in the range 25–55°. The molecular weight dependence and the values of Huggins constant and end-to-end distance were determined at 25° for the reduced denatured proteins including papain and were consistent with their linear random coil behavior. The behavior of reduced ovalbumin in 9 M urea was found to be similar. The dimensions of randomly coiled proteins generally decreased with increasing temperature. The unperturbed dimensions calculated from viscosity data also showed variation with temperature. The intrinsic viscosity–temperature profiles of random coil proteins were characterized with a minimum at 35° and a hump at 40°. Similar observations were made on randomly coiled ovalbumin in 9 M urea. However, such features were not noticed in the curves for cross-linked randomly coiled ovalbumin and lysozyme. Our results, therefore, suggest that proteins which are well-behaved, linear random coils in denaturing solvents at 25° show conformational anomalies at higher temperatures that are independent of amino acid composition, chain length, and the nature of the denaturing solvent.

Intrinsic viscosity measurements have been successfully used in the detection of conformational changes in proteins (Tanford, 1968). For native globular proteins intrinsic viscosity, $[\eta]$, is low, about 3–4 cm^3/g, regardless of their molecular weights. Moreover, it is independent of temperature and of the nature of the solvent, as long as the protein molecule retains its native conformation and the extent of its solvation is not markedly changed (Tanford, 1961). On the other hand, intrinsic viscosity for linear random coil proteins shows strong molecular weight dependence and this was successfully used by Tanford (1968) in the demonstration of the linear

† From the Department of Biochemistry, Jawaharlal Nehru Medical College, Aligarh Muslim University, Aligarh 202001, India. *Received August 8, 1973.* A part of this work was submitted by F. A. in partial fulfilment of the requirements of the degree of Master of Philosophy in Chemistry, Aligarh Muslim University, Aligarh. One of us (F. A.) thanks the Council of Scientific and Industrial Research, Government of India, for the award of a fellowship.