Biochemistry

Volume 13, Number 2 January 15, 1974

Conformational Parameters for Amino Acids in Helical, β-Sheet, and Random Coil Regions Calculated from Proteins†

Peter Y. Chou and Gerald D. Fasman*

ABSTRACT: The helix, β-sheet, and coil conformational parameters, P_α, P_β, and P_c, for the 20 naturally occurring amino acids have been computed from the frequency of occurrence of each amino acid residue in the α, β, and coil conformations in 15 proteins, whose structure has been determined by X-ray crystallography. These values have been utilized to provide a simple procedure, devoid of complex computer calculations, to predict the secondary structure of proteins from their known amino acid sequences. The computed P_α values are within 10% of the experimental Zimm–Bragg helix growth parameters, s, evaluated from poly(α-amino acids). The environmental effects on the s values of polypeptides and proteins are discussed, showing that P_α values may be more reliable in predicting protein conformation. A detailed analysis of the helix and β-sheet boundary residues in proteins provide amino acid frequencies at the N- and C-terminal ends which are used to delineate helical and β regions. Charged residues are found with the greatest frequency at both helical ends, but are mostly absent in β-sheet regions. The frequencies at the helical ends may also be correlated to the experimental Zimm–Bragg helix initiation parameter, σ, evaluated from poly(α-amino acids). A mechanism of protein folding is proposed, whereby helix nucleation starts at the centers of the helix (where the P_α values are highest) and propagates in both directions, until strong helix breakers (where P_α values are lowest) terminate the growth at both ends. Similarly, residues with the highest P_β values will initiate β regions and residues with the lowest P_β values will terminate β regions. The helical region with the highest α potential (*i.e.*, largest $\langle P_\alpha \rangle$) is proposed as the site of the first fold during protein renaturation. The mechanism of folding of myoglobin is discussed. Thus, the protein conformational parameters and the conformational boundary frequencies determined for the first time in their hierarchical order in this paper will enable accurate prediction of protein secondary structure as well as providing insights into tertiary folding.

The mechanism whereby proteins fold into their native conformation, capable of biological activity, has been a long sought after goal. With the elucidation of the three-dimensional structure of many proteins through X-ray crystallography, a new momentum has been given to understanding the factors governing this complex assembly of polypeptide chains.

The impetus for the prediction of protein conformation was initiated with studies on poly(α-amino acids). Helix formation in poly(α-amino acids) is characterized by a cooperative process (Zimm and Bragg, 1959; Applequist, 1963) in which the Zimm–Bragg parameters σ and s are defined respectively as the cooperativity factor for helix initiation, and the equilibrium constant for converting a coil residue to a helical state at the end of a long helical sequence. Potentiometric titration data on poly(L-glutamic acid) (Nagasawa and Holtzer, 1964) and poly(L-lysine) (Hermans, 1966a) gave s values of 1.25 and 1.15, respectively, in aqueous solution at 25°. Since $s = 1$ corresponds to the critical value at which long chains substantially convert into the helical form (Zimm and Bragg, 1959), the greater than unity s values for poly(Glu)[1] and poly(Lys) (both in the un-ionized form) indicate that these two homopolymers form stable helices; this has been confirmed by optical studies (Fasman, 1967). Temperature studies on the conformational stability of block and random copolypeptides in water have provided additional s values: $s_{Gly} = 0.57$ (Ananthanarayanan *et al.*, 1971), $s_{Ser} = 0.78$ (Hughes *et al.*, 1972), $s_{Ala} = 1.07$ (Platzer *et al.*, 1972), and $s_{Leu} = 1.30$ (Ostroy *et al.*, 1970). The low s values for Gly

† Publication No. 922 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); *Biochemistry 11*, 942 (1972). All amino acids are of the L configuration unless otherwise noted.

and Ser confirm earlier qualitative studies which showed that Gly incorporation in a copolymer with γ-ethyl-L-glutamate led to helix destabilization (Block and Kay, 1967), and that poly(Ser), of low molecular weight, in water assumed a random chain conformation, while high molecular weight samples and block copolymers assumed random or β structures (Tooney and Fasman, 1968). Likewise the $s > 1$ values for Ala and Leu reflect the strong helix-forming power of these two residues, as shown by the extreme stability of poly(Ala) in water at 95° (Doty and Gratzer, 1962), and the increased helical stability of copolymers of Glu and Leu with increasing incorporation of leucine (Fasman *et al.*, 1964).

Recent circular dichroism studies in H_2O on random copolymers of Leu with hydroxypropylglutamine (Chou *et al.*, 1972) showed that increased incorporation of Leu in these copolymers resulted in greater helical stability. Furthermore, thermodynamic calculations based on circular dichroism studies showed that s_{Leu} increased from 1.20 at 0° to 1.28 at 70°. The comparable s_{Ala} values (Platzer *et al.*, 1972) at these temperatures are 1.08 and 1.00. The higher s values for Leu than for Ala may be attributed to its larger nonpolar side chain which facilitates hydrophobic stabilization of the helix. The importance of hydrophobic interactions in the Leu helix is also reflected in the larger s_{Leu} values at higher temperatures which agrees well with theoretical findings (Némethy and Scheraga, 1962). In addition, utilizing potentiometric titrations of (Lys, Leu) random copolymers (Snell and Fasman, 1972) yielded $s = 1.13$ for poly(Lys) and an extrapolated value of $s = 1.92$ for poly(Leu) at 25°. An earlier study of Leu block polymers with poly(D,L-Lys) as the end blocks gave $s_{Leu} = 1.30$ at 25° (Ostroy *et al.*, 1970). Hence under three different copolymer environments, the average s value, $\langle s_{Leu}^{25°} \rangle = 1.47$, suggesting that Leu may be the strongest helical-forming residue, since existing s values for other amino acids are all lower. This led to an investigation of Leu residues in various regions of 15 proteins with known conformation, and it was shown that while Glu, Ala, and Leu were found most frequently in helical regions, Leu was clearly the most prominent residue in the inner helical cores of proteins (Chou and Fasman, 1973).

An analysis of all 20 amino acids in 15 proteins is presented in this paper whereby the frequency of their occurrence in various conformational states is compared with the experimental Zimm–Bragg σ and s parameters. Boundary residues of helical and β-sheet sections are analyzed, and yield clues for the termination of these conformational regions. The helix and β-conformational parameters provide a quantitative measure of regions in proteins with the highest helical and β-sheet potential, and may be useful in understanding protein folding mechanisms. These parameters have been used successfully in predicting protein secondary conformation from known amino acid sequences as shown in the following paper (Chou and Fasman, 1974).

Methods

A survey was made on the following 15 proteins whose amino acid sequence and conformation *via* X-ray crystallography are known: carboxypeptidase A (Quiocho and Lipscomb, 1971), α-chymotrypsin (Blow, 1969), cytochrome b_5 (Mathews *et al.*, 1972), elastase (Shotton and Watson, 1970), ferricytochrome *c* (Dickerson *et al.*, 1971), α and β-hemoglobin (Perutz *et al.*, 1968), insulin (Blundell *et al.*, 1972), lysozyme (Blake *et al.*, 1967), myogen (Nockolds *et al.*, 1972), myoglobin (Kendrew *et al.*, 1961), papain (Drenth *et al.*, 1971), ribonuclease S (Wyckoff *et al.*, 1970), staphylococcal nuclease (Arnone *et al.*, 1971), and subtilisin BPN′ (Wright *et al.*, 1969). The amino acid residues in the helix, β-sheet, and coil regions of these proteins are tabulated in Table I. It should be noted that the β-sheet residues differ slightly from those reported earlier (Chou and Fasman, 1973) due to more recent detailed X-ray diffraction analysis (references cited above). Where the β regions were not specified explicitly in the original papers, as in the case of chymotrypsin, elastase, and ribonuclease S, the schematic diagrams showing hydrogen bonding in these proteins were used to delineate the β-sheet regions. Residues at β bends which did not show hydrogen bonding were not included in the β-sheet regions. All regions designated as helical (α_I, α_{II}, 3_{10}, distorted helix) by the X-ray crystallographic studies have been included as helical residues in Table I. Despite minor changes in some helical regions, based on the latest X-ray studies (denoted in the following paper), the analysis did not significantly alter the calculated helix parameters so that the helical residues listed in Table I are identical with those reported earlier (Chou and Fasman, 1973). Although previous surveys of helical and nonhelical residues have been made, they were based on fewer proteins (Cook, 1967; Kotelchuck *et al.*, 1969; Ptitsyn, 1969; Finkelstein and Ptitsyn, 1971; Kabat and Wu, 1973a,b; Wu and Kabat, 1971, 1973). Hence, the present analysis using 15 proteins containing 2473 residues should have more statistical reliability than earlier literature data. In addition, the β-sheet regions of proteins as well as the helical and β-sheet boundary regions have been analyzed in greater detail. Finally, a new normalization procedure has been used to derive helix, β, and

TABLE I: Amino Acid Residues in the Helix, Inner Helix,[a] β-Sheet, and Coil Regions of 15 Proteins.

Amino Acid	No. of Residues	Residues in Helix	Residues in Inner Helix	Residues in β Region	Residues in Coil Region
Ala	228	119	62	38	71
Arg	78	22	9	12	44
Asn	133	35	12	15	83
Asp	111	39	10	15	57
Cys	54	15	3	12	27
Gln	95	40	16	20	35
Glu	113	62	28	5	46
Gly	232	45	22	32	155
His	74	33	11	9	32
Ile	106	38	22	29	39
Leu	196	94	64	41	61
Lys	175	67	34	22	86
Met	28	12	6	8	8
Phe	82	33	16	18	31
Pro	85	18	0	9	58
Ser	202	57	24	25	120
Thr	156	47	21	32	77
Trp	44	18	10	9	17
Tyr	100	22	10	22	56
Val	181	74	44	51	56
Total	2473	890	424	424	1159

[a] The three helical end residues on both N- and C-terminals of a helical region are omitted.

TABLE II: Frequency of Helical, Inner Helical,[a] β, and Coil Residues in 15 Proteins with Their Conformational Parameters P_α, $P_{\alpha i}$, P_β, and P_c.

Amino Acid	f_α[b]	P_α[c]	$f_{\alpha i}$[b]	$P_{\alpha i}$[c]	f_β[b]	P_β[c]	f_c[b]	P_c[c]
Ala	0.522	1.45	0.272	1.59	0.167	0.97	0.311	0.66
Arg	0.282	0.79	0.115	0.67	0.154	0.90	0.564	1.20
Asn	0.263	0.73	0.090	0.53	0.113	0.65	0.624	1.33
Asp	0.351	0.98	0.090	0.53	0.137	0.80	0.514	1.09
Cys	0.278	0.77	0.056	0.33	0.222	1.30	0.500	1.07
Gln	0.421	1.17	0.168	0.98	0.211	1.23	0.368	0.79
Glu	0.549	1.53	0.248	1.45	0.044	0.26	0.407	0.87
Gly	0.190	0.53	0.091	0.53	0.138	0.81	0.668	1.42
His	0.446	1.24	0.149	0.87	0.122	0.71	0.432	0.92
Ile	0.358	1.00	0.208	1.22	0.274	1.60	0.368	0.78
Leu	0.480	1.34	0.327	1.91	0.209	1.22	0.311	0.66
Lys	0.383	1.07	0.194	1.13	0.126	0.74	0.491	1.05
Met	0.429	1.20	0.214	1.25	0.286	1.67	0.286	0.61
Phe	0.402	1.12	0.195	1.14	0.219	1.28	0.378	0.81
Pro	0.212	0.59	0	0	0.106	0.62	0.682	1.45
Ser	0.282	0.79	0.119	0.70	0.124	0.72	0.594	1.27
Thr	0.295	0.82	0.128	0.75	0.205	1.20	0.494	1.05
Trp	0.409	1.14	0.227	1.33	0.203	1.19	0.386	0.82
Tyr	0.220	0.61	0.100	0.58	0.220	1.29	0.560	1.19
Val	0.409	1.14	0.243	1.42	0.282	1.65	0.309	0.66

$\langle f_\alpha \rangle^e = 0.359$ $\langle P_\alpha \rangle^f = 1.00$ $\langle f_{\alpha i} \rangle^e = 0.171$ $\langle P_{\alpha i} \rangle^f = 1.00$ $\langle f_\beta \rangle^d = 0.171$ $\langle P_\beta \rangle^e = 1.00$ $\langle f_c \rangle^d = 0.469$ $\langle P_c \rangle^e = 1.00$

[a] The three helical end residues on both N- and C-terminals of a helical region are omitted. [b] f_α, $f_{\alpha i}$, f_β, and f_c are respectively the frequency of residues in the helical, inner helical, β, and coil regions. [c] P_α, $P_{\alpha i}$, P_β, and P_c are respectively the conformational parameters for the helix ($f_\alpha/\langle f_\alpha \rangle$), the inner helix ($f_{\alpha i}/\langle f_{\alpha i} \rangle$), the β region ($f_\beta/\langle f_\beta \rangle$), and the coil region ($f_c/\langle f_c \rangle$). [d] $\langle f_\alpha \rangle$, $\langle f_{\alpha i} \rangle$, $\langle f_\beta \rangle$, $\langle f_c \rangle$ are respectively the average frequency of residues in helical, inner helical, β, and coil regions. [e] $\langle P_\alpha \rangle$, $\langle P_{\alpha i} \rangle$, $\langle P_\beta \rangle$, and $\langle P_c \rangle$ are respectively the average conformational parameter for the helix, inner helix, β, and coil regions.

coil conformational parameters which can be used in predicting protein conformation.

Results and Discussion

Protein Conformational Parameters. The frequency of all 20 amino acids in the helical, inner helical,[2] β-sheet, and coil regions can be obtained when their occurrence in each conformational region (columns 2–5 of Table I) is divided by their total occurrence (column 1 of Table I) in the 15 proteins. That is, $f_{j,k} = n_{j,k}/n_j$ (see Appendix). The results are tabulated in Table II, showing that Glu, Ala, Leu, and His are found most frequently in helical regions, while Leu, Ala, Glu, and Val occur most often in the inner helical cores of proteins. Met, Val, Ile, and Cys appear to be the strongest β-sheet formers, and Pro, Gly, Asn, and Ser are the most frequent coil residues in proteins. That these findings agree well with conformational studies of poly(α-amino acids) in solution has already been discussed (Chou and Fasman, 1973). The percentages of residues in the 15 proteins found in the helical, β, and coil regions are respectively 36, 17, and 47%. These are represented as average fractions $\langle f_\alpha \rangle = 0.360$, $\langle f_\beta \rangle = 0.171$, and $\langle f_c \rangle = 0.469$ in Table II. The average fraction of residues found in the inner helical regions, $\langle f_{\alpha i} \rangle = 0.171$, is the same as that found in the β regions.

When the frequency of residues in the helical, inner helical, β, and coil regions for the 20 amino acids are divided by the average frequency of residues in these respective regions, their protein conformational parameters[3] are obtained: $P_\alpha = f_\alpha/\langle f_\alpha \rangle$, $P_{\alpha i} = f_{\alpha i}/\langle f_{\alpha i} \rangle$, $P_\beta = f_\beta/\langle f_\beta \rangle$, $P_c = f_c/\langle f_c \rangle$, and these values are tabulated in Table II. This normalization procedure gave P_α and $P_{\alpha i}$ values resembling the Zimm–Bragg helix growth parameter s. The known average experimental s values, $\langle s \rangle$, for seven amino acid residues determined from homopolymer and copolymer data are compared, in Table III, to the helix conformational parameter P_α calculated from the frequency of helical residues in proteins. The good agreement between the experimental s values from poly(α-amino acids) and the calculated P_α values from proteins, while somewhat surprising, is not totally fortuitous. Therefore, P_α could be a parameter which measures the propensity of a residue to be in the helical conformation, as does s. While $P_\alpha > 1$ corresponds to the case where the fraction helicity of a residue is greater than the average fraction helicity of the protein (36% in the 15 proteins analyzed), $s > 1$ implies that more than 50% of the residues in an *infinitely* long polypeptide chain are helical. However, it should be noted that for *finite* polymer chains (which are more appropriate models for proteins), the fraction helicity becomes less than 50% at $s = 1$ (Zimm and Bragg, 1959). That is, at the transition temperature, T_c, of the infinite polymer chain, where $f_\infty = 0.5$ and $s = 1$, smaller values of fraction helicity, f_N, are found as the chain length (N) decreases, *i.e.*, longer polypeptide chains

[2] Inner helix: the three helical end residues on both N- and C-terminals of a helical region are omitted.

[3] The s_h', s_{hi}', s_β', and s_c' notations used in our previous paper (Chou and Fasman, 1973) have been changed respectively to P_α, $P_{\alpha i}$, P_β, and P_c. This was done to avoid confusion with the s' parameter of Zimm and Rice (1960). Furthermore, the s parameter of Zimm and Bragg (1959) refers to the one-dimensional polypeptide model, whereas the P values are derived from the native three-dimensional protein structure.

TABLE III: Comparison of Experimental Zimm–Bragg Helix Growth Parameters, s, and the Average s Value, $\langle s \rangle$, from Poly(α-amino acids) in Aqueous Solutions[a] at 25°, with Helix Conformational Parameters, P_α, Calculated from 15 Proteins.

Amino Acid	$\langle s \rangle$	P_α
Glu	1.26[b–h]	1.53
Ala	1.11[h–k]	1.45
Leu	1.40[l–o]	1.34
His	1.36[p]	1.24
Lys	1.15[d,e,n,q]	1.07
Ser	0.78[r]	0.79
Gly	0.57[s]	0.53

[a] Solvent conditions are denoted after references. [b] Nagasawa and Holtzer (1965); 0.2 M NaCl. [c] Miller and Nylund (1965); 0.2 M NaCl. [d] Hermans (1966a); 0.1 M KCl. [e] Ciferri *et al.* (1968); 0.1 M KCl. [f] Olander and Holtzer (1968); 0.2 M NaCl. [g] Bychkova *et al.* (1971); 0.1 M NaCl. [h] Warashina and Ikegami (1972); 0.1 M NaCl. [i] Ingwall *et al.* (1968); H_2O. [j] Sugiyama and Noda (1970); 0.06 M NaBr. [k] Platzer *et al.* (1972); H_2O. [l] Ostroy *et al.* (1970); H_2O. [m] Chou *et al.* (1972); H_2O. [n] Snell and Fasman (1972); 0.05 M KF. [o] Alter *et al.* (1972); H_2O. [p] Terbojevich *et al.* (1972); 0.02 M KCl. [q] Barskaya and Ptitsyn (1971); 0.02 M NaCl. [r] Hughes *et al.* (1972); H_2O. [s] Ananthanarayanan *et al.* (1971); H_2O.

melt out at higher temperatures. Using $\sigma = 2 \times 10^{-4}$, for the cooperativity factor for helix initiation Chou *et al.* (1972) plotted the theoretical curves of f_N *vs.* s for homopolymers with different chain lengths (N = 150–250), and showed that at $s = 1$, $f_{150} = 0.26$, $f_{200} = 0.32$, and $f_{250} = 0.36$. Further calculations using $\sigma = 1 \times 10^{-3}$ showed that at $s = 1$, $f_{100} = 0.34$, $f_{150} = 0.39$, $f_{200} = 0.42$, and $f_{250} = 0.43$. Using an intermediate value of $\sigma = 5 \times 10^{-3}$, the following fraction helicities were obtained at $s = 1$: $f_{100} = 0.27$, $f_{150} = 0.35$, $f_{200} = 0.39$, $f_{250} = 0.41$. It is interesting to note that all the proteins studied herein fall within the chain-length region of 100–300 (except for insulin with $N = 51$), and that their average fraction helicity $\langle f_\alpha \rangle = 0.36$ used to define $P_\alpha = 1$ is quantitatively similar to the fraction helicity of polymers of similar chain length at $s = 1$. From the definition of $\langle f_\alpha \rangle$ (eq 4 Appendix), the individual characteristics of all the amino acid residues are averaged out over all residues in the polymeric chains, and then $\langle f_\alpha \rangle$ is exactly the same as the fraction of helix in homopolymers. Another formulation of P_α is to equate it to $p_{j,\alpha}/p_j$, where $p_{j,\alpha}$ is the probability of finding the jth residue in the α conformation, and p_j is the probability of finding the jth residue in proteins. Likewise, the P_β and P_c parameters listed in Table II can be expressed as ratio of probabilities (see Appendix).

Helix Boundary Residues in Proteins. While the conformational parameters P_α and P_β in Table II reveal the hierarchical strength of helical and β-sheet residues in proteins, these parameters do not indicate which residues are mainly responsible for the initiation and termination of helical and β regions. The latter information can be obtained by an analysis of the conformational boundary residues in proteins. The frequency of helical boundary residues, which includes the three helical residues on both ends of a helical region (f_{hN}, N-terminal and f_{hC}, C-terminal) and the three nonhelical residues adjacent to the helical end residues (f_{nhN} and f_{nhC}), are shown in Table IV. The subscript hN refers to the three residues at the beginning of a helical region closer to the N-terminal of the protein, while the subscript hC refers to the three residues at the end of a helical region closer to the C-terminal. The preference of negatively charged residues (Asp and Glu) to occur at the N-terminal helical ends and of positively charged residues (His, Lys, and Arg) to occur at the C-terminal helical ends has been noted earlier (Cook, 1967; Ptitsyn, 1969). In addition, it was found that these charged residues also tend to cluster in the coil region neighboring the helical ends (f_{nhN} and f_{nhC} in Table IV), especially near the C-terminal helices. The greater frequency of His residues at the C-terminal helix as well as at both N-terminal and C-terminal nonhelix regions may be correlated to its greater participation at the active sites of enzymes.

Recently, Robson and Pain (1972), using the data from 11 proteins, made a correlation on the directional effect for the 20 amino acids in their transfer of helix-forming information. Their findings are in general agreement with the data in Table IV; however, there are several major differences: (1) Lys was found to have a marked directional effect, but His and Arg had marginal influence at the helical ends, whereas Table IV shows that Arg, His, and Lys occur with the least frequency at the N-terminal helix, and with the most frequency at the C-terminal helix as well as the C-terminal nonhelix region. (2) Gly and Ser behave similarly to Glu in their directionality, whereas Table IV shows that $(f_{hN})_{Glu} = 0.195$ and $(f_{hC})_{Glu} = 0.124$ are much greater than $(f_{hN})_{Ser} = 0.079$, $(f_{hN})_{Gly} = 0.060$ and $(f_{hC})_{Ser} = 0.084$, $(f_{hC})_{Gly} = 0.039$. (3) Thr has similar characteristics to Lys, whereas Table IV shows that these two residues have apparently opposite characteristics with $(f_{hN})_{Lys} = 0.057$, $(f_{hC})_{Lys} = 0.160$, $(f_{hN})_{Thr} = 0.122$, and $(f_{hC})_{Thr} = 0.045$. (4) Trp was impartial to directionality, whereas Table IV shows $(f_{hN})_{Trp} = 0.136$ and $(f_{hC})_{Trp} = 0.045$, *i.e.*, Trp prefers to occupy the N-terminal rather than the C-terminal helical ends.

In the ten nonheme proteins with known X-ray structure, Chou and Fasman (1973) found that of the 58 out of 1836 residues identified as active site or substrate binding residues, Asp (11), His (10), Tyr (6), Arg (5), Glu (4), Lys (4), Ser (4), and Trp (4) occurred with the greatest frequency. It is interesting that Asp, His, Tyr, and Arg, which contribute more than half of the active site residues (32 out of 58) in the proteins analyzed, all have inner helix conformational parameters, $P_{\alpha i}$, less than unity (Table II). The recent findings of Nakanishi *et al.* (1972) that the central core of the α helix has greater stability than the terminal portions led them to conclude that a rapid fluctuation exists at the helix–coil borderland region which diminishes sharply as the inner helical core is approached. We may speculate that the propensity of polar residues Asp, His, Tyr, and Arg at the helix–coil boundaries of enzymes is due to the fact that these regions are more flexible than the rigid inner helix core, thus facilitating substrate binding as well as enzymatic catalysis. Examples of active site residues found in these boundary regions include Arg-10 and His-12 (C-terminal helix 3–13) and His-48 (N-terminal helix 50–59) in ribonuclease; Glu-35 (C-terminal helix 25–35) and Trp-108 (N-terminal helix 108–115) in lysozyme; His-64 (N-terminal helix 64–73) and Ser-221 (N-terminal helix 223–238) in subtilisin BPN′; His-69, Arg-71, and Glu-72 (N-terminal helix 72–88), Tyr-265 (C-terminal helix 254–262) of carboxypeptidase A; and Asp-90, -92, -94 (C-terminal helix 78–89) of myogen. It is quite possible that these polar residues found at conformational junctions serve as signals for protein folding, thus bringing essential distant residues in juxtaposition necessary to perform their enzymatic function.

TABLE IV: Frequency of Helical Boundary and Central Residues[a] in 15 Proteins.

	f_{hN}[b]		f_{hC}[c]		f_{nhN}[d]		f_{nhC}[e]		f_{hI}[f]
Pro	0.212	$His^{(+)}$	0.216	$His^{(+)}$	0.162	$His^{(+)}$	0.149	Ala	0.184
$Asp^{(-)}$	0.207	$Lys^{(+)}$	0.160	Pro	0.141	$Asp^{(-)}$	0.135	Phe	0.183
$Glu^{(-)}$	0.195	Gln	0.158	Ser	0.104	$Lys^{(+)}$	0.120	Leu	0.179
Ala	0.140	$Arg^{(+)}$	0.154	Gly	0.103	Asn	0.120	$Glu^{(-)}$	0.177
Trp	0.136	Cys	0.148	Asn	0.098	$Arg^{(+)}$	0.115	Val	0.166
Thr	0.122	Met	0.143	Ile	0.085	Gly	0.112	Gln	0.158
Gln	0.116	$Glu^{(-)}$	0.124	Phe	0.085	Ile	0.094	Met	0.143
Phe	0.098	Ala	0.118	Gln	0.084	Pro	0.094	$Lys^{(+)}$	0.126
Asn	0.090	Val	0.116	Leu	0.082	Cys	0.093	Trp	0.114
Ser	0.079	Phe	0.110	$Asp^{(-)}$	0.081	Thr	0.090	$Asp^{(-)}$	0.099
Cys	0.074	Leu	0.102	$Glu^{(-)}$	0.080	Tyr	0.090	Cys	0.093
Met	0.071	Asn	0.090	Tyr	0.080	Phe	0.073	$Arg^{(+)}$	0.090
Tyr	0.070	Ser	0.084	$Lys^{(+)}$	0.074	Met	0.071	Ser	0.084
Ile	0.066	Ile	0.075	Val	0.072	Leu	0.061	Thr	0.083
Val	0.061	$Asp^{(-)}$	0.054	Met	0.071	$Glu^{(-)}$	0.053	Asn	0.075
Gly	0.060	Tyr	0.050	Trp	0.068	Ala	0.044	Ile	0.075
$Lys^{(+)}$	0.057	Thr	0.045	Thr	0.064	Gln	0.042	$His^{(+)}$	0.068
Leu	0.056	Trp	0.045	Cys	0.056	Ser	0.040	Tyr	0.050
$His^{(+)}$	0.054	Gly	0.039	Ala	0.044	Trp	0.023	Gly	0.034
$Arg^{(+)}$	0.038	Pro	0	$Arg^{(+)}$	0.038	Val	0.022	Pro	0
$\langle f \rangle$[g]	0.097		0.097		0.082		0.080		0.112

[a] Helix boundary residues include the three helical residues on both ends of a helical region and the three nonhelical residues adjacent to the helical end residues, a total of six residues on each end of the helix. Helix central residues include the three helical residues at the middle of a helix. [b] f_{hN}, frequency of residues in the N-terminal helix region. [c] f_{hC}, frequency of residues in the C-terminal helix region. [d] f_{nhN}, frequency of residues in the N-terminal nonhelical region. [e] f_{nhC}, frequency of residues in the C-terminal nonhelical region. [f] f_{hI}, frequency of residues in the center of a helix region, comprised of three residues only. [g] $\langle f \rangle$, average frequency of residues in the region defined at the top of each column.

Estimation of σ from Helical Boundary Residues. The Zimm–Bragg σ parameter is an entropic factor for the initiation of a helix, and represents the statistical weight assigned to the first helical residue of a helical sequence bordering a coil region (Zimm and Bragg, 1959). In the Lifson–Roig (1961) theory of helix–coil transitions, the initiation factor v (corresponding to $\sigma^{1/2}$) is used as the statistical weight for both the beginning and the end of a helical sequence. Zimm and Bragg (1959) have discussed the rationale for σ to be of the order of 10^{-2} or less and attribute this to the difficulty of forming the first turn of the helix. Poland and Scheraga (1970) have also explained the smaller than unity values of σ based on statistical weights since the ends of a helical sequence have a relative low probability of occurrence. Hence one may compare the frequency of N-terminal and C-terminal helical residues in proteins (Table IV), which also occur with low probabilities ($\langle f \rangle < 0.1$), with the σ values of their repective amino acids.

While σ is independent of neighboring residues in homopolymers, the same is not the case in proteins. That is, a helical residue may have a smaller helix initiation factor, σ, if its neighboring residues are helix breakers. Similarly a coil residue may have a higher σ value if it is surrounded by residues with high helix potential. This can be observed by comparing the fraction helicity of Ala in proteins (52%, Table II) with its higher helicity (>90%) in poly(Ala) in block copolymers (Doty and Gratzer, 1962; Ingwall *et al.*, 1968). Likewise, while poly(Gly) has not been observed to form an α helix (Ananthanarayanan *et al.*, 1971), there are 19% of Gly residues in helical regions of proteins. These examples clearly demonstrate how different neighboring groups, especially in proteins, can affect the helix–coil equilibria of amino acid residues.

The frequency of N-terminal (f_{hN}) and C-terminal helical residues (f_{hC}) are based on the *three* helical residues at both ends of a helical region. Since $v = \sigma^{1/2}$ is assigned only to the first and last residues of a helical sequence (Lifson and Roig, 1961), the statistical weights of the first and last helical residues can be correlated by $v_N = \sigma^{1/2} \propto f_{hN}/3$ and $v_C = \sigma^{1/2} \propto f_{hC}/3$, respectively. It is expected that v_{Nj} and v_{Cj} would be proportional to the frequency with which the j amino acid occurs at the respective amino (f_{hNj}) and carboxyl (f_{hCj}) helical ends. Hence if the statistical weight of finding a particular amino acid at either helical end is twice that of a second amino acid, one would expect to find the first amino acid twice as often as the second amino acid at that end. Therefore one can propose a helix boundary parameter, $P_{\sigma NC} = (f_{hN}/3) \cdot (f_{hC}/3)$, so that $\sigma = v_N v_C \propto P_{\sigma NC}$, since f_{hN} and f_{hC} are simply the frequency or probability that amino acids occur at the helical ends. The $P_{\sigma NC}$ value for each amino acid can then be obtained by multiplying its respective f_{hN} and f_{hC} values in Table IV, and dividing by 9. To verify that three helical residues was the proper number to use at each end of the helix, a calculation using four amino acids at both helical ends, $f_{hN}/4$ and $f_{hC}/4$, was made.[4] The results obtained for $P_{\sigma NC} = (f_{hN}/4)(f_{hC}/4)$ were similar to that for $P_{\sigma NC} = (f_{hN}/3) \cdot (f_{hC}/3)$, with the exception of $(P_{\sigma NC})_{Leu}$ whose value more than doubled. This was expected since Leu residues occur twice as frequently as do the other amino acids in the inner helical

[4] The $f_{hN}/4$ values were based on *four* residues at both helical ends, whereas $f_{hN}/3$ and $f_{hC}/3$ utilized *three* residues at both helical ends.

TABLE V: Comparison of Experimental Zimm–Bragg Helix Initiation Parameters, σ (average σ literature values $\langle\sigma\rangle$), from Poly(α-amino acids) in Aqueous Solutions at 25°, with Helix Boundary Parameters, $P_{\sigma NC}$,[a] and Helix Nucleation Parameters, $P_{\sigma I}$,[b] Calculated from 15 Proteins.

Amino Acid	$\langle\sigma\rangle$	$P_{\sigma NC}$	$P_{\sigma I}$[r]
Ala	4.5×10^{-3} [c-e]	1.8×10^{-3}	3.8×10^{-3}
Leu	3.6×10^{-3} [f-i]	6.3×10^{-4}	3.6×10^{-3}
Glu	3.4×10^{-3} [f,j-m]	2.7×10^{-3}	3.5×10^{-3}
Lys	2.1×10^{-3} [d,n,g]	1.0×10^{-3}	1.8×10^{-3}
Ser	0.9×10^{-4} [o]	7.3×10^{-4}	7.8×10^{-4}
Gly	0.5×10^{-4} [p]	2.6×10^{-4}	1.3×10^{-4}
Pro	5.0×10^{-6} [q]	0	0

[a] $P_{\sigma NC} = (f_{hN}/3)(f_{hC}/3)$, where f_{hN} and h_{hC} are defined and listed in Table IV. [b] $P_{\sigma I} = (f_{hI}/3)^2$ where f_{hI} is defined in Table IV. [c] Ingwall *et al.* (1968). [d] Sugiyama and Noda (1970). [e] Platzer *et al.* (1972). [f] Miller and Nylund (1965). [g] Snell and Fasman (1972). [h] Ostroy *et al.* (1970). [i] Alter *et al.* (1972). [j] Rifkind and Applequist (1964). [k] Bychkova *et al.* (1971). [l] Warashina and Ikegami (1972). [m] Kubota *et al.* (1972). [n] Barskaya and Ptitsyn (1971). [o] Hughes *et al.* (1972). [p] Ananthanarayanan *et al.* (1971). [q] Ganser *et al.* (1970) (in trifluoroethanol–*n*-butyl alcohol). [r] The $P_{\sigma I}$ values for the other 13 amino acids with unknown σ's are: Phe, 3.7×10^{-3}; Val, 3.1×10^{-3}; Gln, 2.8×10^{-3}; Met, 2.3×10^{-3}; Trp, 1.4×10^{-3}; Asp, 1.1×10^{-3}; Cys, 9.6×10^{-4}; Arg, 9.0×10^{-4}; Thr, 7.7×10^{-4}; Asn, 6.3×10^{-4}; Ile, 6.3×10^{-4}; His, 5.1×10^{-4}; Tyr, 2.8×10^{-4}.

regions, *i.e.*, $(f_{\alpha i})_{Leu} = 0.327$ as compared to $\langle f_{\alpha i}\rangle = 0.171$ (Table II). The selection of three helical end residues used to calculate $P_{\sigma NC}$ was not arbitrary, since these residues have *only* one hydrogen bond to the fourth nearest neighboring residue. Inner helical residues have two hydrogen bonds whereas coil residues have no hydrogen bonds. Because the hydrogen-bonding scheme is identical in the three helical end residues, it was decided that $P_{\sigma NC} = (f_{hN}/3)(f_{hC}/3)$ would yield a better correlation to σ for helix initiation. Since the helix–coil boundary is not exactly clear from X-ray crystallographic analysis (Kuntz, 1972), thus making it difficult to pinpoint the helical end residue, it is probably more accurate to take the first and last turn of each helical region containing three residues at both helical ends, and divide by 3. The average experimental σ values, $\langle\sigma\rangle$, for seven amino acids derived from poly(α-amino acids) are compared to the calculated helix boundary parameters $P_{\sigma NC}$ from proteins in Table V. It is interesting that the strongest helical promoting residues in proteins, Glu and Ala, with the highest P_α values (Table II) also have the highest $P_{\sigma NC}$ values, while the weakest helical forming residues, Gly and Pro, with the lowest P_α values have the lowest $P_{\sigma NC}$ values. This trend is also paralleled by the experimental $\langle\sigma\rangle$ values.

Despite the overall agreement in magnitude between the $\langle\sigma\rangle$ and $P_{\sigma NC}$ values, although experimental σ values often vary by a factor of 10, there are certain discrepancies to be pointed out. Though the $P_{\sigma NC}$ of Leu is 1/5 of $\langle\sigma_{Leu}\rangle$, it is still smaller than the $P_{\sigma NC}$ value of 14 amino acids. This may be surprising considering that Leu is a strong helix residue in both polypeptides and proteins. However, since most of the Leu residues are found in the inner helical cores of proteins, it has less probability of occurring at the helical ends, resulting in the small calculated value of $P_{\sigma NC}$. At present no distinction has been made in the literature for the statistical weights assigned to the first and last helical residues, *i.e.*, v_N and v_C (or $\sigma_N^{1/2}$ and $\sigma_C^{1/2}$). The computed $P_{\sigma NC}$ herein utilizes both f_{hN} and f_{hC} values, although it can be seen from Table IV that these values are not necessarily in agreement for a particular amino acid. In fact, the positively charged residues His, Lys, and Arg which occur most frequently at the C-terminal helix are found least frequently at the N-terminal helix. The reverse is true for Pro residues, which is the strongest initiator at the N-terminal helix but is the weakest at the C-terminal helix. Because of the asymmetry at helical ends it may be more accurate to define $P_{\sigma N} = f_{hN}/3 \propto v_N = \sigma_N^{1/2}$ and $P_{\sigma C} = f_{hC}/3 \propto v_C = \sigma_C^{1/2}$, instead of using a single value of $P_{\sigma NC}$ for initiating both helical ends. M. Froimowitz and G. D. Fasman (to be published) utilized $P_{\sigma N}$ and $P_{\sigma C}$ in the Lifson–Roig theory and have shown its applicability in the prediction of protein conformation. Recently Gō *et al.* (1971) have proposed a model for the helix–coil transition in copolymers where the three helical residues on both ends of a helical sequence are assigned statistical weights of $\sigma^{1/4}$, but no distinction was made between σ_N and σ_C for the helical ends.

β-Sheet Boundary Residues in Proteins. The frequency of β-sheet boundary residues, which include the three β-sheet residues on both ends of a β region ($f_{\beta N}$, N-terminal and $f_{\beta C}$, C-terminal) and the three non-β residues adjacent to the β-end residues ($f_{n\beta N}$ and $f_{n\beta C}$), is shown in Table VI. While charged residues are frequently found at the helix boundaries of proteins (Table IV), they are conspicuously absent at the β-sheet boundary regions. The values of $f_{\beta N}$ and $f_{\beta C}$ in Table VI show that the charged residues occur rarely at the N-terminal β region, and below average at the C-terminal region (except for Arg with $f_{\beta C} = 0.09$ which is greater than $\langle f_{\beta C}\rangle = 0.074$). In addition, the charged residues are not favored in the central β region (composed of the three residues at the middle of a β region) since their frequencies in the central β region, $f_{\beta I}$, are less than the average frequency $\langle f_{\beta I}\rangle$ of this region. While the polarity of helical ends as well as the asymmetric distribution of charged residues at the helix junctions (Table IV) may serve some directional purpose in protein folding, there is no apparent function discernible in the distribution of charged residues at the β-boundary regions (Table VI). However, it can be seen that Trp and Gln prefer the N-terminal rather than the C-terminal β-region, whereas Arg and Cys prefer the C-terminal rather than the N-terminal β-region in proteins. In addition, Cys residues tend to occur more frequently at the central β region than at the boundaries. The strong β residues with high P_β values (Table II) are found with equal frequency at the N-terminal, C-terminal, as well as in the central β-sheet region. Contrast this with the strong helical residue, Leu, which occurs rarely at the N-terminal, moderately at the C-terminal, and predominantly in the inner helical region (Tables II and IV). There is more of an even distribution of residues in the various moieties of a β region when compared to a helix region, which probably arises from the basic difference in these conformations. The hydrogen-bonding scheme is mainly interchain in β sheets and intrachain in helices; hence long-range interactions should play a more important role in the case of β regions. Helix formation depends mainly on its nearest neighbors and is a rapid cooperative phenomenon (Engel and Schwarz, 1970) whereas β-sheet formation has slower kinetics as distant residues are brought into juxtaposition. Conio *et al.* (1971) have shown that the coil → β transition is slower than coil → helix for poly(Tyr) in water.

TABLE VI: Frequency of β-Sheet Boundary and Central Residues[a] in 15 Proteins.

	$f_{\beta N}$[b]		$f_{\beta C}$[c]		$f_{n\beta N}$[d]		$f_{n\beta C}$[e]		$f_{\beta I}$[f]
Val	0.160	Val	0.127	Pro	0.165	Trp	0.136	Cys	0.167
Trp	0.159	Leu	0.117	Asn	0.150	Gly	0.116	Val	0.166
Met	0.143	Ile	0.113	Ser	0.094	Ser	0.114	Ile	0.142
Gln	0.137	Tyr	0.110	Cys	0.093	Thr	0.103	Phe	0.134
Tyr	0.130	Met	0.107	Tyr	0.090	Pro	0.094	Leu	0.112
Ile	0.123	Cys	0.093	$Arg^{(+)}$	0.090	$Arg^{(+)}$	0.090	Met	0.107
Phe	0.110	$Arg^{(+)}$	0.090	Thr	0.090	$His^{(+)}$	0.081	Trp	0.091
Ala	0.079	Phe	0.085	Gly	0.086	Tyr	0.080	Thr	0.090
Thr	0.071	Ala	0.075	$Lys^{(+)}$	0.074	Asn	0.068	Tyr	0.090
$Asp^{(-)}$	0.063	$Lys^{(+)}$	0.069	Gln	0.074	Gln	0.063	Gly	0.078
Leu	0.061	$Asp^{(-)}$	0.063	$Asp^{(-)}$	0.072	Val	0.061	Ala	0.066
Pro	0.059	Asn	0.060	$Glu^{(-)}$	0.071	Ile	0.057	$Lys^{(+)}$	0.063
Ser	0.054	Thr	0.058	Met	0.071	Leu	0.056	$Asp^{(-)}$	0.063
Asn	0.053	Ser	0.054	Ile	0.047	$Asp^{(-)}$	0.045	$Arg^{(+)}$	0.051
Gly	0.047	Gly	0.052	Val	0.039	$Glu^{(-)}$	0.044	Asn	0.045
$Lys^{(+)}$	0.046	Gln	0.042	Leu	0.031	Ala	0.035	Glu	0.042
Cys	0.037	$His^{(+)}$	0.041	$His^{(+)}$	0.027	$Lys^{(+)}$	0.034	Ser	0.040
$His^{(+)}$	0.014	Pro	0.035	Ala	0.026	Phe	0.024	$His^{(+)}$	0.027
$Arg^{(+)}$	0.013	$Glu^{(-)}$	0.035	Phe	0.024	Cys	0.019	Pro	0.024
$Glu^{(-)}$	0.001	Trp	0.023	Trp	0.023	Met	0	$Glu^{(-)}$	0.001
$\langle f \rangle$[g]	0.074		0.074		0.071		0.070		0.079

[a] β-sheet boundary residues include the three β residues on both ends of a β region and the three non-β residues adjacent to the β-sheet end residues, a total of six residues on each end of the β region. β-sheet central residues include the three β residues at the middle of a β region. [b] $f_{\beta N}$, frequency of residues in the N-terminal β region. [c] $f_{\beta C}$, frequency of residues in the C-terminal β region. [d] $f_{n\beta N}$, frequency of residues in the N-terminal non-β region. [e] $f_{n\beta C}$, frequency of residues in the C-terminal non-β region. [f] $f_{\beta I}$, frequency of residues in the center of a β region comprised of three residues only. [g] $\langle f \rangle$, average frequency of residues in the region defined at the top of each column.

Comparison of $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ Values for Helix and β Regions. In the 15 proteins analyzed there are 2473 residues of which 890 residues are in 80 helical segments and 424 residues in 64 β-sheet segments. From these data the following *average* values can be calculated: amino acids/protein = 165, helical residues/protein = 59.2, helical segments/protein = 5.33, residues/helical segment = 11.1, β residues/protein = 28.3, β segments/protein = 4.3, residues/β segment = 6.6. As relatively few proteins have been studied, these averages have limited significance. It is possible that helices are longer than β regions because once the first turn of a helix is formed the rest of the helix propagates easily and quickly in a zipper-like mechanism. It is interesting that the average number of residues per helical segment in the 15 proteins is 11.1, corresponding to three turns of the α helix which contains 3.6 residues per turn. When the average $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values, $\langle \tilde{P}_\alpha \rangle$ and $\langle \tilde{P}_\beta \rangle$, are computed for the 80 helical and 64 β regions in the 15 proteins, it is seen that $\langle \tilde{P}_\beta \rangle = 1.11$ for all β segments and $\langle \tilde{P}_\alpha \rangle = 1.08$ for all helical segments. These data suggest that β regions may be more stable than helix regions in proteins. The β sheet has already been shown to be more conformationally stable than the α helix in the case of poly(α-amino acids) (Davidson and Fasman, 1967). The fact that proteins in the β conformation, *e.g.*, silk, form extremely thermodynamically stable structures is also a well-established fact (Lucas *et al.*, 1958).

Since the P_α and P_β values of Table II provide a relative measurement of the helix and β forming potential of each amino acid residue, it is interesting to calculate the average P_α and P_β values at the N-terminal, C-terminal, and internal regions of both helix and β segments. Intuitively it is difficult to say whether the strongest helical-forming residues (those with highest P_α values) occur at the N-terminal helix, $\langle P_\alpha \rangle_N$ (where they assist helix initiation), or at the C-terminal helix, $\langle P_\alpha \rangle_C$ (where they assist in preventing the helix from converting to the coil state), or at the internal helix, $\langle P_\alpha \rangle_I$ (where they assist in stabilizing the central helix core). A similar statement can be made concerning the β region. When the average P_α and P_β values are computed, it is seen (Figure 1) that $\langle P_\alpha \rangle_N = 1.04$, $\langle P_\alpha \rangle_C = 1.09$, and $\langle P_\alpha \rangle_I = 1.13$ for the helix regions, and $\langle P_\beta \rangle_N = 1.17$, $\langle P_\beta \rangle_C = 1.11$, and $\langle P_\beta \rangle_I = 1.17$ for the β regions. It is clear that residues with the highest helical potential occupy the central helix core, with P_α values tapering off as both helical ends are approached, with a further drop in P_α at the helix boundaries, N-terminal nonhelical region, $\langle P_\alpha \rangle_{nN} = 0.96$ and C-terminal nonhelical region, $\langle P_\alpha \rangle_{nC} = 0.94$. While residues with the highest β potential are found to occupy the central β-sheet core, there is not much of a lowering of P_β values as the β region ends are approached. In fact, as noted in Table VI, there appears to be a scattering of strong β formers in the N-terminal and C-terminal as well as the central portion of β regions. Only at the β-sheet boundaries is there a sharp drop in the P_β values with $\langle P_\beta \rangle_{nN} = 0.92$ (non-β N-terminal) and $\langle P_\beta \rangle_{nC} = 0.97$ (non-β C-terminal). The average P_α and P_β values for the three central coil residues are $\langle P_\alpha \rangle_{cI} = 0.93$ and $\langle P_\beta \rangle_{cI} = 0.97$, which are similar to the P_α and P_β values for the coil residues at the boundary of the helix and β regions. Thus, both the helix and β-sheet regions exhibit a potential wall at their conformational boundaries, indicating that certain helix and β-sheet breaking residues play important roles in conformational termination.

Estimation of σ from Helical Central Residues. It is noted

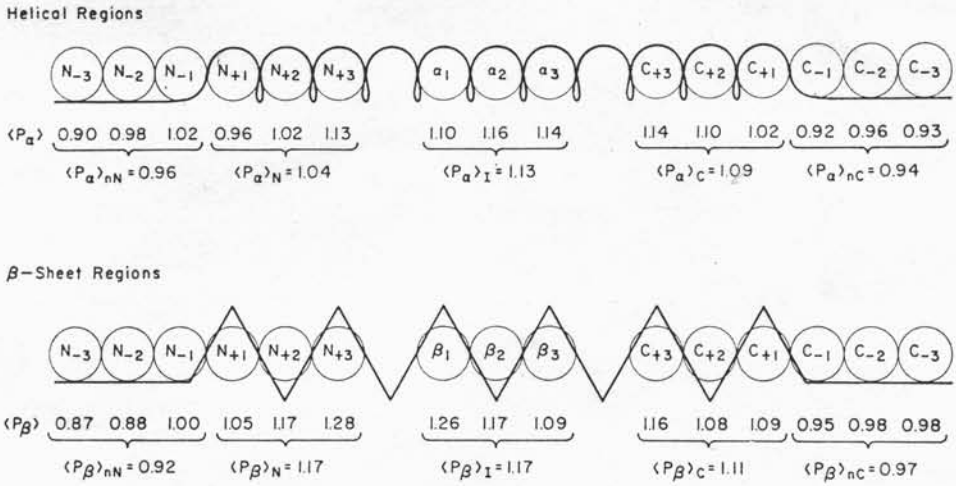

FIGURE 1: Comparison of $\langle P_\alpha \rangle$ values for helical boundary and central regions and $\langle P_\beta \rangle$ values for β-sheet boundary and central regions in 15 proteins. $\langle P_\alpha \rangle_{nN}$, $\langle P_\alpha \rangle_N$, $\langle P_\alpha \rangle_I$, and $\langle P_\alpha \rangle_{nC}$ are respectively the average P_α values (see Table II and eq 5 of Appendix) for the N-terminal nonhelical, N-terminal helical, central helical, C-terminal helical, and C-terminal nonhelical regions, with each region comprised of three residues. $\langle P_\beta \rangle_{nN}$, $\langle P_\beta \rangle_N$, $\langle P_\beta \rangle_I$, $\langle P_\beta \rangle_C$, and $\langle P_\beta \rangle_{nC}$ are respectively the average P_β values (see Table II and eq 8 of Appendix) for the N-terminal non-β, N-terminal β, central β, C-terminal β, and C-terminal non-β regions, with each region comprised of three residues.

that the strongest helix breakers, Gly, Pro, Tyr, and Asn (Table II), are also frequently found at the N-terminal and C-terminal nonhelical regions (Table IV). That the values of (f_{nhN}) and (f_{nhC}) are about 0.80–0.14 for these residues suggests that these helix breakers are equally effective in terminating a helical section growing at either the N- or C-terminal. Since the strongest helical residues are found at the exact center of helical cores (Figure 1) a postulate may be made for the mechanism whereby helix initiation could actually start at the center, with the strongest helical residues responsible for nucleation, and propagates in *both* directions, until strong helix breakers terminate the helix growth at both the N- and C-terminals of helical regions. The f_{hI} values in Table IV are based on three residues at the middle of a helical region.[5] If it is assumed that one of these center residues serves as a nucleation point, then a helix nucleation parameter may be redefined as $P_{\sigma I} = (f_{hI}/3)^2 \propto \sigma$. When initiation starts at the helix center, the central nucleating residue must have both amino and carbonyl groups hydrogen bonded, whereas at the helical ends only one hydrogen bond is involved. Hence squaring $(f_{hI}/3)$ in computing $P_{\sigma I}$, $[P_{\sigma I} = (f_{hI}/3)^2]$, is similar to $P_{\sigma NC} = (f_{hN}/3)(f_{hC}/3)$ as in both cases two hydrogen bonds are frozen in the helix nucleation process. The newly evaluated $P_{\sigma I}$ values based on helix initiation at the center of a helix instead of at the ends are listed in the last column of Table V and appear in closer agreement with the experimental $\langle \sigma \rangle$ values. In particular there is better agreement between $(P_{\sigma I})_{Leu}$ and $\langle \sigma_{Leu} \rangle$ than between $(P\sigma_{NC})_{Leu}$ and $\langle \sigma_{Leu} \rangle$, as well as the general order of $\langle \sigma \rangle$ and $P_{\sigma I}$. The earlier discrepancy, which indicated that Leu was a strong helix former but a weak helix initiator, is now resolved when it is seen that Leu is also a strong helix initiator when initiation is considered at the helix center. Since experimental σ values are difficult to evaluate in some cases, the $P_{\sigma I}$ values listed in Table V may be used as approximate estimates for unknown σ parameters.

Prediction of Protein Conformation. In order to accurately predict the complete secondary structure of proteins, it is essential to obtain the precise conformational potentials of the 20 amino acids in helical, β-sheet, and random coil conformations as well as a set of rules governing the folding of these regions. In this paper the manner in which the conformational parameters, P_α, P_β, and P_c, have been derived is outlined, and the advantages of their use in evaluating protein conformation will be discussed. In the following paper (Chou and Fasman, 1974), the application of these conformational parameters is elaborated on in conjunction with some simple rules for initiating and terminating helical and β-sheet regions. As the complete hierarchical order of both helix and β-sheet forming potentials has been established for the first time in the present paper, a comparison of the utilization of these conformational potentials *vs.* other parameters in the literature will be instructive and will demonstrate the simplicity and advantages of this new approach to protein conformational prediction.

Applying information theory techniques, Robson and Pain (1971) assigned different helix-forming parameters for the 20 amino acids, and were quite successful in predicting the helical regions in five globular proteins. Their helix-forming information parameters, I, obtained by more sophisticated computer methods, are in essential agreement with the more easily calculated P_α parameters obtained here. However, their finding that Gly ($I = +0.18$) is more favorable in helix formation than Phe ($I = +0.11$) disagrees with the finding herein that Gly ($P_\alpha = 0.53$) is the strongest helix breaker and that Phe ($P_\alpha = 1.12$) is a moderate helix former. In support of our assignments it will be noted that Poly(Glu[76]Phe[24]) was shown to be more helical (48%) than poly(Glu) (30%) (at pH 5.2, where poly(Glu) is 60% charged, $\mu = 0.2$, and 20°), indicating that Phe is indeed a helix promoter (Sage and Fasman, 1966), agreeing with the prediction herein (see also Addendum).

Since only a few experimental s values are available, Lewis *et al.* (1970) divided the 20 amino acids into three groups: helix formers ($s = 1.05$), helix indifferent ($s = 1.00$), and helix breakers ($s = 0.385$), and were able to correctly predict the conformation of 68% of all residues as well as 64% of the helical residues in 11 proteins. These predictions were obtained even though Tyr was incorrectly assigned as a helix former, and Ala, Gln, and His incorrectly assigned as helix indifferent. In a later paper, Lewis and Scheraga (1971) assigned Tyr as helix

[5] In cases where there are four inner residues, both middle residues were tabulated as belonging to the exact helical center. When helical length is shorter than nine residues, there are overlaps in the counting of end and central residues. The same case holds for β-sheets.

indifferent and Ala, Glu, and His as helix formers, but obtained essentially the same results since their s values for helix former (1.05) and helix indifferent (1.00) are very similar. If Tyr is assigned as a helix breaker, according to the present survey in Table II, using $P_\alpha = 0.61$, as well as using the different P_α values for each amino acid, more accurate prediction of helical residues in native proteins is possible by simply averaging the P_α values in any segment considered, and any sequence with $\langle P_\alpha \rangle > 1$ is considered helical, and $\langle P_\alpha \rangle < 1$ considered nonhelical. A more complete set of rules is provided in the following paper (Chou and Fasman, 1974).

Lewis and Scheraga (1971), using the Zimm–Bragg matrix formulation, incorrectly predicted residues 91–94 (Tyr-Ile-Tyr-Ala) to be helical in staphylococcal nuclease. Using $(P_\alpha)_{\text{Tyr}} = 0.61$, $(P_\alpha)_{\text{Ile}} = 1.00$, and $(P_\alpha)_{\text{Ala}} = 1.45$ from Table II leads to $\langle P_\alpha \rangle = 0.92$, *i.e.*, a correct prediction that this segment is in a nonhelical conformation. Likewise the helix-breaking nature of Tyr-92, -97, and -115 in ribonuclease leads to $\langle P_\alpha \rangle = 0.95$ for residues 92–115, suggesting that this segment contains no helices, which is verified by X-ray studies (Kartha *et al.*, 1967; Wyckoff *et al.*, 1970). This is a marked improvement over earlier predictions which stated that there are helices in this region, *e.g.*, residues 105–113 (Prothero, 1966), 104–112 (Schiffer and Edmundson, 1967), and 96–110 (Lewis and Scheraga, 1971). Utilizing both P_α and P_β values from Table II shows in fact that region 96–110 of ribonuclease is β sheet ($\langle P_\beta \rangle = 1.12 > \langle P_\alpha \rangle = 1.05$) in agreement with X-ray analysis. Furthermore the other five β regions (identified with $\langle P_\beta \rangle > \langle P_\alpha \rangle$) and all the six random coil regions (identified with $\langle P_\alpha \rangle < 1, \langle P_\beta \rangle < 1$) of ribonuclease were correctly predicted by our present method (Chou and Fasman, 1974). Similarly, the $\langle P_\beta \rangle$ values correctly identified the three β regions of staphylococcal nuclease, whereas previous prediction attempts (Lewis and Scheraga, 1971; Pain and Robson, 1970) assigned these regions as helical.

Recently, Kabat and Wu (1973a,b) have used a 20 × 20 table of tripeptides based on the ϕ, ψ dihedral angles of 12 proteins to locate helix breaking and β-sheet breaking residues in proteins. Because their frequency data on β sheets are quite sparse, they only attempted predictions of β conformation in two proteins. In the case of papain, which has 30 β residues (Drenth *et al.*, 1971), Kabat and Wu (1973a) missed 10 β residues ($\%_\beta$ residues correctly predicted, $\%_\beta = 67\%$) and overpredicted 65 β residues as compared to our underprediction of 4 β residues ($\%_\beta = 87\%$) and overprediction of 29 β residues (Chou and Fasman, 1974). In concanavalin A, which has 12 β regions (Edelman *et al.*, 1972), Kabat and Wu (1973b) located seven β regions correctly, whereas our predictive method correctly identified all 12 β regions with only one overpredicted β region (Chou and Fasman, 1974). Since there are three times more β-sheet residues than β-sheet tripeptides in proteins, our P_β parameters provide more information on the β-forming potential of the 20 amino acids than the Kabat and Wu method (1973a,b). Utilizing pairwise interactions, Nagano (1973) has developed a computer method for predicting helices, β-sheets, and loops in proteins. However, his prediction for carboxypeptidase A gave $\%_\alpha = 78\%$ and $\%_\beta = 46\%$ as compared to $\%_\alpha = 80\%$ and $\%_\beta = 81\%$ found by our present predictive criteria. These results strongly indicate that P_α and P_β values can be used in a quantitative manner for evaluating protein conformation.

Protein Chain Folding. Lewis *et al.* (1971) have proposed that regions of high helical probability aid in directing the folding of polypeptide chains to yield the final native protein conformation. They also indicated that the β bends could play an important role in the folding mechanism of β structures. However, no suggestion was offered as to which section of the protein would fold first. With the computed conformational parameters P_α and P_β in Table II, it is now possible to compare the various helical and β segments in each protein by means of their $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values. Since there are indications that helix initiation proceeds faster than β formation (Snell and Fasman, 1973) we postulate that the first fold in a protein will occur near the helix region with the highest helical potential, *i.e.*, largest $\langle P_\alpha \rangle$ value. The next fold will occur near the helix region with the second largest $\langle P_\alpha \rangle$ value, and so forth. Likewise the β-sheet regions in proteins will fold in the order of their $\langle P_\beta \rangle$ values. There may be cases where β regions with higher $\langle P_\beta \rangle$ values than corresponding $\langle P_\alpha \rangle$ values of helices will fold first.

The P_α values listed in Table II may be used to propose the folding mechanisms of myoglobin. Recently Puett (1972) chose arbitrary s values for various fragments of apomyoglobin which best described his experimental data. Realizing the limitations of his model, which treats only Pro as a helix breaker, he nevertheless concluded that fragments 37–87 and 100–119 have the lowest ($s = 1.06$) and the highest ($s = 1.30$) helix-forming potential, respectively, and speculated that fragment 100–119 represents the nucleus for chain folding. These conclusions disagree with the computed $\langle P_\alpha \rangle$ values of 1.15 and 1.09 for fragments 37–87 and 100–119, respectively, indicating that the former fragment is a more likely candidate as a nucleation center for chain folding in myoglobin. Three-dimensional diagrams of the myoglobin molecule (Dickerson and Geis, 1969) show that helices C, D, and E of fragment 37–87 are packed tightly around the heme and might therefore have higher stability, whereas fragment 100–119 (G-helix) lies on the periphery of the molecule, and therefore would be less stable and an unlikely nucleation region. The recent proposal of McLachlan (1972) based on repeating sequences and gene duplication in proteins suggests that an invariable core near the active center acts as a nucleation center for chain folding with peripheral sections to be evolved in later stages of growth. This gives further support that fragment 37–87, containing the E helix and caging the heme, is more directly involved than the peripheral G-helix in nucleating the folding of the myoglobin chain. The nucleation of folding of globular proteins has also been discussed recently by Wetlaufer (1973).

Role of Solvent on s Values. The experimental s values for Glu and Ala appear lower than their corresponding P_α values found in proteins (Table III), indicating that these residues may be in a different environment in proteins than in polypeptides. That is, an isolated polypeptide chain predominantly assumes a linear character, with little or no tertiary structure, while a globular protein will have some of its chains buried in the interior of the molecule simulating a nonaqueous environment. It is interesting that Harrison and Blout (1965) in their treatment of the rotatory dispersion of apomyoglobin found that the optical parameters which best described this protein in aqueous solution to be similar to those of polypeptides dissolved in organic solvents. Hermans (1966b) has shown that $s_{\text{Glu}} = 1.19$ in aqueous solutions changed to $s_{\text{Glu}} = 1.43$ in 20% ethanol, while Conio and Patrone (1969) obtained $s_{\text{Glu}} = 1.50$ in 27% ethanol. These higher s values at 25° for poly(Glu) in a less aqueous environment are in closer agreement to the $P_\alpha = 1.53$ value for Glu residues found in proteins. The better agreement in the s and P_α values for Leu (Table III) is probably due to the strong hydrophobic character of the leucine side chains which keep the polypeptide backbone in a

nonaqueous medium. An indication that the leucine residues in proteins occupy a similar nonaqueous environment is reflected in the high P_α and $P_{\alpha i}$ values (Table II) and the fact that leucine is most frequently found in the inner helical core of proteins. Glu and Ala occur second and third in frequency among inner helical residues in proteins (Table II), indicating that these residues are also essentially in a nonaqueous environment, especially if the helices are in the interior of globular proteins. However, they are quite exposed to water when present in isolated polypeptides. This may account for the discrepancies between the s and P_α values of Table III, although the average difference among the seven residues is only 10%.

The close similarity of s and P_α for Lys in Table III suggests that unlike Glu and Ala, Lys residues in polypeptides and proteins have approximately similar environments. The contribution of the nonpolar side chain of Lys, with its four methylene groups, to hydrophobic stabilization of the helix has been demonstrated experimentally (Hatano and Yoneyama, 1970; Grourke and Gibbs, 1971). Theoretically, Némethy and Scheraga (1962) have shown that the standard free energy and enthalpy of formation of hydrophobic bonds between Leu and Glu as well as Leu and Ala are respectively $\Delta G_{H\phi}^\circ = -0.5$ kcal/mol and $\Delta H_{H\phi}^\circ = 0.7$ kcal/mol, whereas for Leu and Lys interaction, $\Delta G_{H\phi}^\circ = -1.0$ kcal/mol and $\Delta H_{H\phi}^\circ = 1.4$ kcal/mol. Since the hydrophobic stabilization of Lys is double that of Glu or Ala, one can expect that the *nonpolar moiety* of poly(Lys) produces a more nonaqueous environment than poly(Glu) and poly(Ala) whose side chains are shorter. On the other hand, Lys residues in proteins occur less frequently than Glu and Ala in the inner helix as can be seen from $(P_{\alpha i})_{Lys} = 1.13$, $(P_{\alpha i})_{Glu} = 1.45$, and $(P_{\alpha i})_{Ala} = 1.59$, and more frequently than Glu and Ala in the coil regions since $(P_c)_{Lys} = 1.05$, $(P_c)_{Glu} = 0.87$, and $(P_c)_{Ala} = 0.66$ (Table II). This means that in proteins, Lys is more solvated or in a more aqueous environment than Glu and Ala.[6] Hence one can compare the nonpolar moiety of poly(Lys) to the inner helix Lys residues of proteins with low dielectric media, and the hydrophilic portion of poly(Lys) to the Lys residues in the coil regions of proteins with high dielectric media. Likewise, the infrequency of Ser and Gly in the inner helix of proteins $(P_{\alpha i})_{Ser} = 0.70$ and $(P_{\alpha i})_{Gly} = 0.53$, as well as their high frequency in the coil regions of proteins, $(P_c)_{Ser} = 1.27$ and $(P_c)_{Gly} = 1.42$, are similar to the polypeptide case where Ser (Hughes *et al.*, 1972) and Gly (Ananthanarayanan *et al.*, 1971) were shown to be helix breakers and prefer the random conformation. Thus amino acid residues whose environments in polypeptides are similar to those in proteins will have good agreement between s and P_α values. Applequist (1963) had predicted on theoretical grounds that the side chain and polymer–solvent interactions would be expressed in the s value. The experimental data (Hermans, 1966b; Conio and Patrone, 1969) on polypeptides in different solvents have shown this to be true. In addition, the P_α values found in this paper for amino acids in proteins reflect the intrinsic helix-forming potential of the various residues as well as environmental factors on this helix growth parameter. Since experimental s values for all the amino acids are not available (Lewis and Scheraga, 1971) the P_α and P_β values published herein may be used for conformational prediction in proteins, as well as for correlation of experimental data in the case of protein fragments, *e.g.*, apomyoglobin (Puett, 1972). However, one should also bear in mind near-neighbor interactions as well as the temperature and solvent effects on the helix growth parameters, s, as discussed above.

Thus by utilizing the P_α and P_β values, for all the amino acids found in proteins, determined herein, as well as the nucleation and termination rules in the following paper (Chou and Fasman, 1974), reasonable estimates of protein conformation can now be made with confidence in a direct, simple manner which surpasses previous attempts in its reliability.

Acknowledgment

We thank Dr. S. Kubota for many stimulating discussions.

Addendum

Recently, Van Wart *et al.* (1973) obtained thermodynamic parameters on random copolymers of Phe and hydroxypropylglutamine in water. Their experimental values of $s_{Phe} = 1.08$ and $\sigma_{Phe} = 1.8 \times 10^{-3}$ at 25° agree well with the values herein of $(P_\alpha)_{Phe} = 1.12$ (Table II) and $(P_{\sigma I})_{Phe} = 3.7 \times 10^{-3}$ (Table V) computed from the frequencies of Phe residues in the helical regions of proteins. This illustrates that Phe residues in poly(Phe) as well as in proteins have a similar conformational preference.

Appendix

The probability of finding the j amino acid residue in protein is

$$p_j = n_j/\Sigma n_j \tag{1}$$

where n_j is the number of j residues in proteins, and Σn_j is the total number of residues in proteins.

The probability of finding the j residue in the conformational state, k, of proteins is

$$p_{j,k} = n_{j,k}/\Sigma n_{j,k} \tag{2}$$

where $n_{j,k}$ is the number of j residues in the k state and $\Sigma n_{j,k}$ is the total number of residues in the k state.

The probability or frequency of occurrence of the j residue in the k state of proteins is

$$f_{j,k} = n_{j,k}/n_j \tag{3}$$

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

$$\langle f_k \rangle = \Sigma f_{j,k}/\Sigma j = \Sigma n_{j,k}/\Sigma n_j \tag{4}$$

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

$$P_{j,k} = f_{j,k}/\langle f_k \rangle \tag{5}$$

Substituting eq 1–4 in eq 5 results in

$$P_{j,k} = p_{j,k}/p_j \tag{6}$$

[6] Recent X-ray studies show that all 14 Lys residues in chymotrypsin are on the external surface of the molecule while 4 out of 5 Glu and 13 out of 22 Ala residues are external (Birktoft and Blow, 1972). In carp myogen (Kretsinger and Nockolds, 1973) all 13 Lys are external while 1 Glu out of 6 and 2 Ala out of 18 are found to be internal.

Hence the conformational parameter of the j amino acid residue in the k state, $P_{j,k}$, is equal to the probability of finding the j residue in the k state divided by the probability of finding the j residue in proteins. The k conformational state in proteins is either the α, β, or random coil. Throughout the text the subscript j is omitted to aid in clarity.

References

Alter, J. E., Taylor, G. T., and Scheraga, H. A. (1972), *Macromolecules 5*, 739.

Ananthanarayanan, V. S., Andreatta, R. H., Poland, D., and Scheraga, H. A. (1971), *Macromolecules 4*, 417.

Applequist, J. (1963), *J. Chem. Phys. 38*, 934.

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.

Barskaya, T. V., and Ptitsyn, O. B. (1971), *Biopolymers 10*, 2181.

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

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1967), *Proc. Roy. Soc., B 167*, 365.

Block, H., and Kay, J. A. (1967), *Biopolymers 5*, 243.

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

Blundell, T. L., Cutfield, J. F., Dodson, E. J., Dodson, G. G., Hodgkin, D. C., and Mercola, D. A. (1972), *Cold Spring Harbor Symp. Quant. Biol. 36*, 233.

Bychkova, V. E., Ptitsyn, O. B., and Barskaya, T. V. (1971), *Biopolymers 10*, 2161.

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

Chou, P. Y., and Fasman, G. D. (1974), *Biochemistry 13*, 222.

Chou, P. Y., Wells, M., and Fasman, G. D. (1972), *Biochemistry 11*, 3028.

Ciferri, A., Puett, D., Rajagh, L., and Hermans, J. (1968), *Biopolymers 6*, 1019.

Conio, G., and Patrone, E. (1969), *Biopolymers 8*, 57.

Conio, G., Patrone, E., and Salaris, F. (1971), *Macromolecules 4*, 283.

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

Davidson, B., and Fasman, G. D. (1967), *Biochemistry 6*, 1616.

Dickerson, R. E., and Geis, I. (1969), The Structure and Action of Proteins, New York, N. Y., Harper and Row, p 48.

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

Doty, P., and Gratzer, W. B. (1962), *in* Polyamino Acids, Polypeptides and Proteins, Stahmann, M. A., Ed., Madison Wis., University of Wisconsin Press, p 111.

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

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.

Engel, J., and Schwarz, G. (1970), *Angew. Chem., Int. Ed. Engl. 9*, 389.

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

Fasman, G. D., Lindblow, C., and Bodenheimer, E. (1964), *Biochemistry 3*, 155.

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

Ganser, V., Engel, J., Winklmair, D., and Krause, G. (1970), *Biopolymers 9*, 329.

Gō, N., Lewis, P. N., Gō, M., and Scheraga, H. A. (1971), *Macromolecules 4*, 692.

Grourke, M. J., and Gibbs, J. H. (1971), *Biopolymers 10*, 795.

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

Hatano, M., and Yoneyama, M. (1970), *J. Amer. Chem. Soc. 92*, 1937.

Hermans, J. (1966a), *J. Phys. Chem. 70*, 510.

Hermans, J. (1966b), *J. Amer. Chem. Soc. 88*, 2418.

Hughes, L. J., Andreatta, R. H., and Scheraga, H. A. (1972), *Macromolecules 5*, 187.

Ingwall, R. T., Scheraga, H. A., Lotan, N., Berger, A., and Katchalski, E. (1968), *Biopolymers 6*, 331.

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.

Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C., and Shore, V. C. (1961), *Nature (London) 190*, 666.

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

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

Kubota, S., Gaskin, F., and Yang, J. T. (1972), *J. Amer. Chem. Soc. 94*, 4328.

Kuntz, I. D. (1972), *J. Amer. Chem. Soc. 94*, 8568.

Lewis, P. N., Gō, 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.

Lifson, S., and Roig, A. (1961), *J. Chem. Phys. 34*, 1963.

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

McLachlan, A. D. (1972), *J. Mol. Biol. 64*, 417.

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

Miller, W. G., and Nylund, R. E. (1965), *J. Amer. Chem. Soc. 87*, 3542.

Nagano, K. (1973), *J. Mol. Biol. 75*, 401.

Nagasawa, M., and Holtzer, A. (1964), *J. Amer. Chem. Soc. 86*, 538.

Nakanishi, M., Tsuboi, M., Ikegami, A., and Kanehisa, M. (1972), *J. Mol. Biol. 64*, 417.

Némethy, G., and Scheraga, H. A. (1962), *J. Phys. Chem. 66*, 1773.

Nockolds, C. E., Kretsinger, R. H., Coffee, C. J., and Bradshaw, R. A. (1972), *Proc. Nat. Acad. Sci. U. S. 69*, 581.

Olander, D. S., and Holtzer, A. (1968), *J. Amer. Chem. Soc. 90*, 4549.

Ostroy, S. E., Lotan, N., Ingwall, R. T., and Scheraga, H. A. (1970), *Biopolymers 9*, 749.

Pain, R. H., and Robson, B. (1970), *Nature (London) 227*, 62.

Perutz, M. F., Muirhead, H., Cox, J. M., and Goaman, L. C. G. (1968), *Nature (London) 219*, 131.

Platzer, K. E. B., Ananthanarayanan, V. S., Andreatta, R. H., and Scheraga, H. A. (1972), *Macromolecules 5*, 177.

Poland, D., and Scheraga, H. A. (1970), Theory of Helix-Coil Transitions in Biopolymers, New York, N. Y., Academic Press, p 10.

Prothero, J. W. (1966), *Biophys. J. 6*, 367.

Ptitsyn, O. B. (1969), *J. Mol. Biol. 42*, 501.

Puett, D. (1972), *Biochim. Biophys. Acta 257*, 537.

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).