# Conformational Preferences of amino acids in globular proteins

**ARTICLE** *in* BIOCHEMISTRY · NOVEMBER 1978

Impact Factor: 3.02 · DOI: 10.1021/bi00613a026 · Source: PubMed



**1 AUTHOR:**

Michael Levitt
Stanford University
**231** PUBLICATIONS **25,688** CITATIONS

[Reprinted from Biochemistry, (1978) 17, 4277.]


# Conformational Preferences of Amino Acids in Globular Proteins?

Michael Levitt‡

**ABSTRACT:** In a previous paper [Levitt, M.. and Greer. J. (1977), J. Mol. *Biol. 11* 4. 18 l-2391, an objective compilation of the secondary-structure regions in more than 50 different globular proteins was produced automatically. In the present paper, these assignments of secondary structure are analyzed to give the frequency of occurrence of the 20 naturally occurring amino acids in $\alpha$ helix, $\beta$ sheet, and reverse-turn secondary structure. Nineteen of these amino acids have a weak but statistically significant preference for onlv one type of secondary structure. These preferences correlate well with the chemical structure of the particular amino acids giving a more objective classification of the conformational properties of amino acids than available before.

Interest in the conformational preferences of the amino acids began soon after the three-dimensional structure of myoglobin, the first protein structure to be solved by X-ray crystallography, was published in 1960 (Kendrew et al.,1960). This protein structure showed that certain regions of the polypeptide chain have a well-defined local or secondarv structure. namely, they were $\alpha$ helices, while other regions were more irregular. Attention was addressed to whether certain amino acids occurred in either the $\alpha$-helices or the irregular regions more often than expected by chance. The early studies (Guzzo. 1965: Prothero. 1966: Cook, 1967) showed that some amino acids did have preferences for $\alpha$ helix. but the number of accurate assignments that could be made was limited by the small sample of protein structures available at that time. As more structures of proteins were solved, attempts were made to determine preferences of amino acids for the $\beta$ sheet (Ptitsyn and Finkelstein. 1970) and reverse-turn (Lewis et al., 197 1; Crawford et al.. 1973) secondary structures and also to refine the preferences for $\alpha$ helix (Kotelchuck and Scheraga, 1969; Ptitsyn, 1969: Pain and Robson, 1970; Lewis and Scheraga, 197 1).

More recent studies have analyzed much larger numbers of protein structures (between 15 and 29 structures). leading to assignments of conformational preferences for all the amino acids and for all three types of secondary structure (Nagano. 1973: Chou and Fasman, 1974a.b: Tanaka and Scheraga. 1976: Maxfield and Scheraga. 1976; Robson and Suzuki, 1976; Chou and Fasman. 1977). These studies have generallv used the assignments of secondary structure given by the particular crystallographic group (Nagano. 1973: Chou and Fasman. 1974, 1977: Robson and Suzuki. 1976) or have used the values of $(\phi, \psi)$ angles to assign the "secondary structure" (Tanaka and Scheraga, 1976; Maxfield and Scheraga, 1976). Assignments made by the former method are subjective and do not obey consistent criteria, and different groups use assignments that sometimes differ appreciably (see Levitt and Greer. 1977). Assignments made by the latter method are more objective and consistent but are sensitive to small errors in the $(\phi, \psi)$ angles and do not always correspond to the usual assignments of secondary structure made on the basis of both the local residue conformation and the pattern of hydrogen bonds with other residues.

The uncertainties concerning the experimental secondary structure were largely overcome in a recent paper by Levitt and Greer (1977). In their study an automated procedure, which was, therefore, objective and consistent, was used to assign the secondary structure of more than 60 protein structures on the basis of both the local conformation and the pattern of hydrogen bonds.

In the present studv these assignments are analyzed to produce frequencies of occurrence of the 20 naturallv occurring amino acids in $\alpha$-helix, $\beta$-sheet, and reverse-turn secondary structure. With such a large number of protein structures, the occurrences in different structures can be weighted to eliminate redundancies and still give accurately determined frequencies. Statistical tests, which involve integrating the probability distribution function of the observed frequencies, are used to set 80% confidence limits on the frequencies and also to give the probabilities that a particular amino acid favors. is indifferent to. or breaks a particular secondarv structure.

The results obtained here show that 19 of the 20 naturallv occurring amino acids do have a statistically significant preference for $\alpha$-helix, P-sheet, or reverse-turn secondary structure. These preferences correlate with the chemical structure and stereochemistry of the amino acids as follows: ( 1) amino acids with a bulky side chain (those branched at the B-carbon atom like Val. Ile, and Thr or those with a large aromatic ring like Phe, Tvr, and Trp) favor $\beta$ sheet: (2) amino acids with short polar side chains (like Ser. Asp, Asn) or with a special side chain (like Gly, which has no side chain. and Pro, which has a cyclized side chain) prefer reverse turns; (3) all other amino acids prefer $\alpha$ helix with the exception of Arg, which has no preference.

An important difference between the preferences calculated here and those calculated bv others is that here each amino acid prefers onlv one type of secondary structure, while in other studies several amino acids prefer two types of secondary structure. The present study, therefore, gives a more clear-cut classification of the conformational properties of amino acids in globular proteins. which can be used in future analysis of proteins.

## Experimental Procedures

The basic method used here is simple and well known: the frequency with which a particular amino acid occurs in a particular ttrpe of secondarv structure is determined by counting the number of times it occurs in known protein structures. The present studv does have some special features:

† From the MRC Laboratory of Molecular Biology. Cambridge, England, and The Salk Institute of Biological Studies. San Diego. California 920 12. *Received* **April** 17, 1978.

‡ Present address: The Salk Institute of Biological Studies. San Diego. Calif. 920 12.

(1) The assignments of secondary structure to regions of the polypeptide chain have been done automatically from the X-ray coordinates and for very many more protein structures (Levitt and Greer, 1977). (2) The data are weighted to allow for the many related protein structures solved to date by X-ray crystallography. (3) The counting statistics are analyzed carefully and used to define the statistically significant preferences of amino acids f'or different types of secondary structure.

*Protein Data and Weighting Schemes.* The proteins used here are those that have been' analyzed in a previous paper (Levitt and Greer, 1977). The secondary-structure assignments presented in that work form the basic data sample used here.

The entire data sample of 66 protein structures (1 1 569 amino acids) is highly redundant.There are many cases where (a) two proteins are the independentlv solved halves of a dimer. (b) the same protein was solved by hifferent groups, and (c) two or more proteins are closely related and have homologous sequences. Rather than choose one representative protein from each redundant class, all the available data was used after weighting it with a weight dependent on the protein concerned. The weight, $w$, was taken as $1/M$ when $M$ proteins were judged to be in the same redundant class. Nagano (1973) included the sequences of homologous proteins in his sample even when the X-ray structure of these proteins had not been solved. He also used a weight of $1/\sqrt{M}$ which gives more weight to redundant data than in the present work.

Another problem with the basic sample of proteins is that not all the secondary-structure assignments are equally reliable. As noted previously (Levitt and Greer, 1977). $\alpha$ helices found bv the pattern of hydrogen bonds (H-bond method) are more reliable than those found by the local residue conformation (a-angle method), and $\beta$ sheets found by the H-bond method and the separation of a-carbons. $C^{\alpha}$-$C^{\alpha}$ method, together are more reliable than those found by either method alone. In some studies only these more reliable assignments were used: the residues in less reliable $\alpha$-helix and $\beta$-sheet assignments were counted as unassigned. The same original assignments of reverse-turn secondary structure were used in both cases.

*Analysis of Counting Statistics.* The number of times a particular amino acid is counted in a particular type of secondary structure reflects the intrinsic preference of that amino acid for the type of secondary structure. Here we assume that this preference depends only on the type of amino acid and type of secondary structure but not on the position of the amino acid in the sequence nor on the positions of any other amino acids (Ptitsyn and Finkelstein. 1970). In this way, the number of times amino acid type $i$ occurs in secondary structure type $j$, $n_{ij}$, depends on the total number of that amino acid, $N_i$, and the probability or frequency of occurrence,& for that secondary structure. The protein data can be considered as a sequence of amino acids of type $i$ that can occur at random in secondary structure of type $j$, so that the probability of' counting $N_{ij}$ occurrences is given by the binomial probability distribution (for simplicity, the subscripts on $n_{ij}$, $N_i$, and $f_{ij}$ are dropped).

$$P(n) = [N!/n!\,(N-n)!]f^{n}(1-f)^{N-n} \qquad (1)$$

In the present study, values of $n$ and N are obtained by counting. and we wish to determine the most likely value for the intrinsic probability, $f$, and the distribution about this value. At fixed values of $n$ and N. eq 1 can be regarded as the probability distribution for $f$, but it must be renormalized to give $\int_0^1 P(f)df = 1$. Assuming that all values of $f$ as equally likely a *priori,* gives:

$$P(f) = [(\mathrm{N}+1)!/(n)!(N-n)!]f^{n}(1-f)^{N-n} \qquad (2)$$

From this distribution function, the most likely value of $f'$ is f= n/N, the mean value is $\bar{f} = (n+1)/(N+2)$, and the variance $\sigma^2 = \bar{f^2} - \bar{f}^2 = [(n+1)/(N+2)][(N-n+1)/(N+2)(\mathrm{N}+3)]$. For values of $n$ and N larger than about 10, $\bar{f} \simeq f' = n/N$, $\sigma^2 = f'(1-f')/N$, and the distribution $P(f)$ tends to a Gaussian function. This approximation has been used before (Ptitsyn. 1969; Chou and Fasman, 1977) to define confidence limits for $f$ as $f^{\min} = f' - \sigma$ and $f^{\max} = f' + \sigma$ such that it was 67.5% certain that $f$ fell between them. A better approximation was derived by Lindley (1964) who showed that the above probability distribution (eq 2) is very nearly Gaussian for a new variable $x = \ln[f/(1-f)]$, with a mean $x = \ln[n/(N-n)]$ (i.e., $f$ = n/N, the most likely value) and a variance $\sigma_x^2 = 1/(n+1) + 1/(N-n+1)$. Although this approximation works well (Maxfield and Scheraga, 1976), we prefer to calculate more exact confidence limits by numerical integration of eq 2. In order to be $C$% certain that $f$ falls between $f^{\min}$ and $f^{\max}$ requires that $\int_0^{f^{\max}} P(f)\,df = 1 - C/200 = \int_{f^{\max}}^{\infty} P(f)\,df$.

The calculated frequences, $f'$, are also used to define the preference of a particular amino acid for a particular secondary structure. If there are $N_{ij}$ amino acids in secondary structure j out of N-r amino acids in total, then ail those amino acids which have no preference for the secondary structure would occur with the same random frequence $N_j/N_{\mathrm{T}} = f_j$. Here the preferences of any amino acid are classified as follows: if $f'_{ij}$ is 10% greater than the random frequency, $f_j$, the amino acid is taken as favoring secondary structure type j (denoted by *h):* if $f'_{ij}$ is 10% less than this random frequency, the amino acid is taken as breaking secondary structure type j (denoted by *b*). if $f'_{ij}$ is within 1 0% of $f_j$, the amino acid is taken as indifferent to secondary structure by j (denoted by i).

The certaintv with which this classification is made can be calculated by integrating the probability distribution for $f$, $P(f)$. over three regions: from $f_{ij} = 0$ to $f_{ij} = 0.9f_j$; from $f_{ij} = 0.9f_j$ to 1. $1f_j$; and from $f_{ij} = 1.1f_j$ to $f_{ij} = \infty$ to give $I_0^{0.9}$, the probabiiitv of $f_{ij} < 0.9f_j$; $I_{0.9}^{1.1}$, the probabilitv of $f_{ij} > 0.9f_j$ and $f_{ij} < 1.1f_j$; and $I_{1.1}^{\infty}$, the probability of $f_{ij} > 1.1f_j$, respectively. These three probabilities are simply those that the amino acid breaks. is indifferent to, or favors secondary structure of type j.

As the frequencies $f_{ij}$ depend on the amounts of each type of secondary structure-present in the sample. it is convenient to normalize them to give tendencies $P_{ij} = f_{ij}/f_j$ (Chou and Fasman. 1974a). The classification scheme given above corresponds to $P_{ij} > 1.1$ for an amino acid that favors the secondarv structure, $P_{ij} < 0.9$ for one that breaks the secondary structure, and $0.9 < P_{ij} < 1.1$ for one that is indifferent.

## Results and Discussion

*The Protein Data and Weighting Scheme.* The 66 protein structures used here are listed in Table I. These proteins fall into 3 1 families of structures having up to nine members in a family (the hemoglobin family). When there are A4 members in a family, occurrences of amino acids in secondary structure are counted with weight $w = 1/M$. In this way, the 1309 residues in the hemoglobin family are equivalent to only 145 effective residues ($M = 9$). Because each member of a familv is an independent X-ray determination and sometimes has a somewhat different sequence, the present weighting scheme may be too strict: the 1309 residues in the hemoglobin family are probably equivalent to more than 145 effective residues

TABLE I: Classification of Protein Data to Eliminate Redundancies.

| proteins in class | no. in class ($M$) | wt ($W$) | no. of res | effect. no. of res |
|---|---|---|---|---|
| Ca-bind. protein B (carp) | 1 | 1 | 106 | 106 |
| azomyohemerythrin, hemerythrin | 2 | 0.5 | 231 | 115 |
| carboxy-hb, cyanmet-hb. metmyoglobin, α- & β-aquomet-hb, α- & β-deoxy-hb (horse & human) | 9 | 0.111 | 1309 | 145 |
| rubredoxin (1.5 Å) | 1 | 1 | 54 | 54 |
| var pt of Bence-Jones REI (dimer), IgG Fab' (dimer). Bence-Jones McG (dimer) | 6 | 0.167 | 1070 | 179 |
| prealbumin (dimer) | 2 | 0.5 | 246 | 123 |
| superoxide dismutase | 1 | 1 | 151 | 151 |
| Con A (argonne & rockefeller) | 2 | 0.5 | 474 | 237 |
| alkaline Ser protease | 1 | 1 | 185 | 185 |
| trypsin, a-Chy (MRC & Mich.), Chy[a], elastase | 6 | 0.333 | 1205 | 238 |
| insulin (dimer) | 2 | 0.5 | 102 | 51 |
| trypsin inhibitor | 1 | 1 | 58 | 58 |
| ferredoxin | 1 | 1 | 54 | 54 |
| ferricyt $b_5$ | 1 | 1 | 87 | 87 |
| Ox. high-potent. Fe protein | 1 | 1 | 85 | 85 |
| ferricyt c (tuna) "outer" & "inner". ferricyt c (tuna & Bonito), ferricyt $c_2$, cyt $c_{550}$ | 6 | 0.167 | 658 | 110 |
| ribonucl A & S | 2 | 0.5 | 248 | 124 |
| I ysozyme | 1 | 1 | 129 | 129 |
| nuclease (Staph. aureus) | 1 | 1 | 142 | 142 |
| papain | 1 | 1 | 212 | 212 |
| t hermolysin | 1 | 1 | 316 | 316 |
| thioredoxin | 1 | 1 | 108 | 108 |
| ox. & semiquinone flavodoxins | 2 | 0.5 | 276 | 138 |
| adenylate kinase | 1 | 1 | 194 | 194 |
| triose phosphate isomerase (dimer) | 2 | 0.5 | 494 | 247 |
| carbonic anhydrase B & C | 2 | 0.5 | 512 | 256 |
| subtilisin BPN' & Novo | 2 | 0.5 | 550 | 275 |
| carboxypeptidases A & B[b] | 2 | 0.6 | 613 | 368 |
| lactactate dehydrogenase (apo & NAD) | 2 | 0.5 | 658 | 329 |
| D-glyceraldehyde-3-P dehydrogenase (green & red) | 3 | 0.5 | 666 | 333 |
| alcohol dehydrogenase | 1 | 1 | 374 | 374 |
| totals | 66 | 0.477[c] | 11569 | 5523 |

[a] The three chymotrypsin structures were given weights of 1/9 each. [b] As carboxypeptidase B is a preliminary structure that differs very much from carboxypeptidase A (Levitt and Greer, 1977). it was given a weight of 0.2, while carboxypeptidase A was given a weight of 1. [c] This is the mean weight calculated as the ratio of the total effective number of residues to the total number of residues.

Nevertheless, even with this weighting scheme. there are 5523 effective residues ( 1 1 569 actual residues) in the data base used here, which is substantially more than used in previous studies.

Counts of occurrences of the 20 amino acids in the three types of secondary structure were made in three ways (Table II): (I) using the most reliable secondary structure assignments of Levitt and Greer (1977) together with the weights in Table I; (II) using the same reliable assignment but without weights: and, (III) using all secondary structure assignments together with the weights. The counts were then converted to normalized secondary structure frequency values, $P_{ij}$, defined (see Methods) so that $P_{ij} = 1$ for the frequency of occurrence expected by chance. In general. $P_{ij}$ values derived with weights

TABLE II: Normalized Frequencies Calculated Using Differently Weighted Protein Data.

| amino acid | α helix I[a] | α helix II | α helix III | β sheet I | β sheet II | β sheet III | reverse turn I | reverse turn II | reverse turn III |
|---|---|---|---|---|---|---|---|---|---|
| Ala | 1.29 | 1.32 | 1.25 | 0.90 | 0.86 | 0.89 | 0.78 | 0.79 | 0.78 |
| cys | 1.11 | 0.92 | 1.12 | 0.74 | 1.04 | 0.85 | 0.80 | 0.79 | 0.80 |
| Leu | 1.30 | 1.31 | 1.32 | 1.02 | 1.04 | 1.03 | 0.59 | 0.57 | 0.59 |
| Met | 1.47 | 1.39 | 1.43 | 0.97 | 0.93 | 0.99 | 0.39 | 0.51 | 0.39 |
| Glu | 1.44 | 1.44 | 1.45 | 0.75 | 0.66 | 0.65 | 1.00 | 1.02 | 1.00 |
| Gln | 1.27 | 1.10 | 1.24 | 0.80 | 1.00 | 0.82 | 0.97 | 0.92 | 0.97 |
| His | 1.22 | 1.31 | 1.25 | 1.08 | 0.85 | 1.04 | 0.69 | 0.81 | 0.69 |
| Lys | 1.23 | 1.25 | 1.24 | 0.77 | 0.77 | 0.81 | 0.96 | 0.99 | 0.96 |
| Val | 0.91 | 0.93 | 0.88 | 1.49 | 1.43 | 1.48 | 0.47 | 0.46 | 0.47 |
| Ile | 0.97 | 0.93 | 0.94 | 1.45 | 1.47 | 1.41 | 0.51 | 0.50 | 0.51 |
| Phe | 1.07 | 1.02 | 1.08 | 1.32 | 1.21 | 1.22 | 0.58 | 0.77 | 0.58 |
| Tyr | 0.72 | 0.73 | 0.75 | 1.25 | 1.31 | 1.25 | 1.05 | 0.93 | 1.05 |
| Trp | 0.99 | 0.97 | 1.03 | 1.14 | 1.26 | 1.15 | 0.75 | 0.79 | 0.75 |
| Thr | 0.82 | 0.79 | 0.81 | 1.21 | 1.27 | 1.13 | 1.03 | 0.97 | 1.03 |
| Gly | 0.56 | 0.61 | 0.57 | 0.92 | 0.89 | 0.93 | 1.64 | 1.67 | 1.64 |
| Ser | 0.82 | 0.76 | 0.82 | 0.95 | 1.02 | 0.96 | 1.33 | 1.30 | 1.33 |
| Asp | 1.04 | 1.03 | 1.03 | 0.72 | 0.69 | 0.74 | 1.41 | 1.47 | 1.41 |
| Asn | 0.90 | 0.95 | 0.87 | 0.76 | 0.73 | 0.86 | 1.28 | 1.25 | 1.28 |
| Pro | 0.52 | 0.58 | 0.60 | 0.64 | 0.68 | 0.71 | 1.91 | 1.78 | 1.91 |
| Arg | 0.96 | 0.98 | 0.99 | 0.99 | 0.97 | 1.02 | 0.88 | 0.90 | 0.88 |

[a] I, most reliable secondary structure assignment only, with weight:;; II, most reliable assignments. unweighted; III, all assignments, with weights. In each of these sets of protein data there are the following effective number of residues in α helices, β sheets, reverse turns, undefined regions, and in total: set I 17 15, 1555, 112 1, I 116, and 5507; set II 3804. 3276. 2386. 2033, and 11499; set III 1790, 1957, 1121, 639. and 5507.

(sets I and III) are more similar to each another than they are to $P_{ij}$ values derived without weights (set II). The effect of using the most reliable secondary structure assignment is smaller (compare $P_{ij}$ values in sets I and II); 75 of the 1790 a-helical residues and 402 of the 1957 P-sheet residues are classified as less reliable by the criteria given under the Experimental Section. In spite of the small differences in the three sets of $P_{ij}$ values. a clear pattern emerges from Table II: $P_{ij}$ values significantly greater than 1 occur in the α-helix column for the first 8 amino acids, in the P-sheet column for the next 6 amino acids. and in the reverse-turn column for the next 5 amino acids. In only one case (Cys in α helix) does a $P_{ij}$ value in one of these three groups fall below 1 .O (set II). In the remainder of this work, attention will be focused on the set I values as they have been derived using the reliable secondary structure and the rather conservative weighting scheme.

***Statistics of Occurrence in α Helix, β Sheet, and Reverse Turns.*** The statistics of the occurrence of amino acids in α helix, β sheet. and reverse turns in the reliable assignment., weighted data base (set I) are analyzed in Tables III-V. The number of occurrences, $n_{ij}$, is not necessarily integral with this data base, as each occurrence is counted with a weight $W = 1/M$ (these nonintegral values have been used to calculate the exact $P_{ij}$ values). The confidence limits, $P^{min}$ and $P^{max}$, have been chosen such that $P_{ij}$ lies between these values with an 80% chance. The three integrals, $I_0^{0.9}$, $I_{0.9}^{1.1}$, and $I_{1.1}^{\infty}$, give the probabilitv that the $P_{ij}$ value of each amino acid for a particular secondary structure is less than 0.9 ($I_0^{0.9}$), is between 0.9 and 1 .1 ($I_{0.9}^{1.1}$), and is greater than 1.1 ($I_{1.1}^{\infty}$). These integrals are also the probability that the particular amino acid dislikes the secondary structure (structure code b). is indifferent to it (structure code i), and favors the secondary structure (structure code h). In Tables III-V the preference for α helix, β sheet.

TABLE III: Statistical Analysis[a] of Amino Acid Preferences for $\alpha$ Helix.

| amino acid | $n_{ij}$ | $P_{ij}$ | $P^{min}$ | $p^{max}$ | $I_0^{0.9}$ | $I_{0.9}^{1.1}$ | $I_{1.1}^{\infty}$ | struct code |
|---|---|---|---|---|---|---|---|---|
| Ala | 186 | 1.29 | 1.20 | 1.38 | 0.1 E-7[b] | 0.004 | 0.996 | h |
| cys | 42 | 1.11 | 0.95 | 1.30 | 0.049 | 0.389 | 0.562 | h |
| Leu | 152 | 1.30 | 1.19 | 1.40 | 0.1 E-6 | *0.007* | 0.993 | h |
| Met | 39 | 1.47 | 1.26 | 1.71 | 0.001 | 0.009 | 0.990 | h |
| Giu | 126 | 1.44 | 1.31 | 1.56 | 0.1 E-8 | 0.001 | 0.999 | h |
| Gin | *67* | 1.27 | 1.11 | 1.41 | 0.00 1 | 0.082 | 0.917 | h |
| His | *50* | 1.22 | 1.06 | 1.40 | 0.005 | 0.157 | *0.837* | h |
| Lys | 147 | 1.23 | 1.12 | 1.33 | 0.7 E-5 | 0.049 | 0.95 1 | h |
| Vai | 124 | 0.91 | 0.82 | 0.99 | 0.454 | 0.542 | 0.004 | i,b |
| lie | 90 | *0.97* | 0.87 | 1.09 | 0.171 | 0.742 | 0.087 | i |
| Phe | 65 | 1.07 | 0.95 | 1.23 | 0.039 | 0.509 | 0.452 | i,h |
| Tyr | 48 | 0.72 | 0.62 | 0.86 | 0.954 | 0.045 | 0.00 1 | b |
| Trp | 26 | 0.99 | 0.80 | 1.21 | 0.253 | 0.465 | 0.28 1 | i |
| Thr | 90 | 0.82 | 0.73 | 0.92 | 0.83 1 | 0.169 | 0.3 E-3 | b |
| Gly | 90 | 0.56 | 0.49 | 0.63 | 0.999 | 0.9 E-8 | 0.7 E-17 | b |
| Ser | 112 | 0.82 | *0.74* | 0.91 | 0.873 | 0.127 | 0.5 E-4 | b |
| Asp | 107 | 1.04 | 0.94 | 1.15 | 0.036 | 0.708 | 0.256 | i |
| Asn | 68 | 0.90 | *0.79* | 1.03 | 0.454 | 0.52 1 | 0.025 | i,b |
| Pro | 36 | 0.52 | 0.43 | 0.64 | 0.999 | 0.4 E-4 | 0.3 E-8 | b |
| Arg | 50 | 0.96 | 0.82 | 1.11 | 0.290 | 0.592 | 0.1 18 | i |

[a] $P_{ij} = (n_{ij}/N_i)/(N_j/N)$, where the values of $N_i$, the number of times amino acid $i$ occurs in the whole sample, are as follows: Ala = 464, cys = 12 1, Leu = 378, Met = 84, Gin = 382, Glu = 1'7 1. His = 13 1, Lys = 385. Val = 440. Iie = 296. Phe = 193, Tyr = 211, Trp = 84. Thr = 351, Gly = 519, Ser = 439. Asp = 330. Asn = 241. Pro = 219, Arg = 168. The number of times secondary structure j occurs in the whole sample ($N_j$) are as follows: $\alpha$ helix = *17* 15. $\beta$ sheet = 1555, reverse-turn = 1 121. undefined = 1 116. The total number of residues, N, is 5507. Note that because of the weights used none of the numbers is exactly an integer: the accurate *nonintegral* values have been used to calculate $P_{ij}$ values. [b] The symbol E denotes "10 to the power of'. i.e., 0.1 E-7 = 0.1 X $10^{-7}$.

TABLE IV: Statistical Analysis[a] of Amino Acid Preferences for $\beta$ Sheet.

| amino acid | $n_{ij}$ | $P_{ij}$ | $P^{min}$ | $p^{max}$ | $I_0^{0.9}$ | $I_{0.9}^{1.1}$ | $I_{1.1}^{\infty}$ | struct code |
|---|---|---|---|---|---|---|---|---|
| Ala | 118 | 0.90 | 0.81 | 0.99 | 0.480 | 0.5 15 | 0.005 | i.b |
| Cys | 25 | 0.74 | 0.58 | 0.92 | 0.874 | 0.119 | 0.007 | b |
| Leu | 109 | 1.02 | 0.92 | 1.13 | 0.06 1 | 0.751 | 0.188 | i |
| Met | 23 | 0.97 | 0.77 | 1.21 | 0.3 12 | 0.434 | 0.254 | i |
| Giu | 59 | 0.75 | 0.64 | 0.86 | 0.957 | 0.043 | 0.7 E-5 | b |
| Gin | 39 | 0.80 | *0.67* | 0.96 | 0.776 | 0.224 | 0.010 | b |
| His | 40 | 1.08 | 0.91 | 1.27 | 0.084 | 0.443 | 0.473 | i,h |
| Lys | 84 | *0.77* | 0.68 | 0.87 | 0.945 | 0.055 | 0.3 E-4 | b |
| Val | 185 | 1.49 | 1.38 | 1.59 | 0.1 E- 13 | 0.5 E-6 | 0.999 | h |
| Ile | 122 | 1.45 | 1.33 | 1.59 | 0.1 E-8 | 0.1 E-4 | 0.999 | h |
| Phe | 72 | 1.32 | 1.17 | 1.48 | 0.1 E-3 | 0.029 | 0.971 | h |
| Tyr | 75 | 1.25 | 1.11 | 1.41 | 0.5 E-5 | *0.074* | 0.925 | h |
| Trp | 27 | 1.14 | 0.92 | 1.38 | 0.073 | 0.315 | 0.612 | h |
| Thr | 120 | 1.21 | 1.10 | 1.33 | 0.00 1 | 0.100 | 0.899 | h |
| Gly | 135 | 0.92 | 0.83 | 1.01 | 0.362 | 0.63 1 | *0.007* | i.b |
| Ser | 118 | 0.95 | 0.86 | 1.05 | 0.230 | *0.739* | 0.03 1 | i |
| Asp | 67 | 0.72 | 0.62 | 0.82 | 0.983 | 0.016 | 0.7 E-C | b |
| Asn | 51 | 0.76 | 0.64 | 0.88 | 0.932 | 0.068 | 0.4 E-3 | b |
| Pro | 39 | 0.64 | 0.52 | 0.76 | 0.996 | 0.004 | 0.6 E-5 | b |
| Arg | 47 | 0.99 | 0.84 | 0.16 | 0.208 | 0.583 | 0.209 | i |

[a] See footnote to Table III

or reverse turn have been assigned on the basis of whether $P_{ij} < 0.90$ (b), $P_{ij} > 1.10$ (h), or $0.9 < P_{ij} \leq 1.10$ (i), rather than on the basis of the values of the probability integrals.

Seven of the eight amino acids (Table III) defined as $\alpha$-helix favoring (h) are given this structure code with more than 80% confidence: for Cys the confidence of the assignment is only 56%. All five of the amino acids defined as $\alpha$-helix breaking (b) are given this structure code with more than 80% confidence. The seven amino acids that are indifferent to $\alpha$ helix are given this structure code with lower confidence. Three of the seven amino acids defined as $\alpha$-helix indifferent have a significant chance of having a different structure code: for Val the probabilities are i (54%), b (45%); for Phe they are i (51%), h (45%); and for Asn they are i (52%), b (45%).

Five of the six amino acids (Table IV) defined as P-sheet favoring (h) have been given this structure code with at least 90% confidence; for Trp the value of $P_{ij}$ (1.14) is significantly greater than 1.1, but the probability distribution of $P_{ij}$ is par-

TABLE V: Statistical Analysis[a] of Amino Acid Preferences for Reverse Turns.

| amino acid | $n_{ij}$ | $P_{ij}$ | $P^{min}$ | $p_{max}$ | $I_0^{0.9}$ | $I_{0.9}^{1.1}$ | $I_{1.1}^{\infty}$ | struct code |
|---|---|---|---|---|---|---|---|---|
| Ala | 73 | 0.77 | 0.67 | 0.88 | 0.920 | 0.080 | 0.2 E-3 | b |
| Cys | 20 | 0.8 1 | 0.63 | 1.05 | 0.658 | 0.277 | 0.065 | b |
| Leu | 45 | 0.58 | 0.49 | 0.70 | 0.999 | 0.00 1 | 0.1 E-7 | b |
| Met | | 0.41 | 0.27 | 0.66 | 0.992 | 0.007 | 0.00 1 | b |
| Glu | 57 | 0.99 | 0.85 | 1.15 | 0.189 | 0.610 | 0.20 1 | i |
| Gln | 34 | 0.98 | 0.80 | 1.19 | 0.27 1 | 0.496 | 0.233 | i |
| His | 18 | 0.68 | 0.5 1 | 0.89 | 0.902 | 0.090 | 0.008 | b |
| Lys | 75 | 0.96 | 0.83 | 1.09 | 0.258 | 0.653 | 0.089 | |
| Val | 42 | 0.47 | 0.39 | 0.56 | 0.999 | 0.2 E-6 | 0.1 E-11 | b |
| Ile | 31 | 0.51 | 0.42 | 0.64 | 0.999 | 0.00 1 | 0.6 E-7 | b |
| Phe | 23 | 0.59 | 0.45 | 0.75 | 0.989 | 0.010 | 0.00 1 | b |
| Tyr | 45 | 1.05 | 0.88 | 1.24 | 0.118 | 0.505 | 0.377 | i,h |
| Trp | 13 | 0.76 | 0.55 | 1.05 | 0.7 1 1 | 0.219 | 0.070 | b |
| Thr | 74 | 1.04 | 0.90 | 1.18 | 0.084 | 0.624 | 0.292 | |
| Gly | 173 | 1.64 | 1.51 | 1.77 | O.lE-15 | 0.5 E-8 | 1 .000 | h |
| Ser | 118 | 1.32 | 1.19 | 1.46 | 0.4 E-5 | 0.013 | 0.987 | h |
| Asp | 95 | 1.41 | 1.26 | 1.57 | 0.1 E-5 | 0.003 | 0.997 | h |
| Asn | 63 | 1.28 | 1.11 | 1.47 | 0.00 1 | 0.077 | 0.922 | h |
| Pro | 85 | 1.91 | 1.70 | 2.1 1 | 0.5 E- 12 | 0.2 E-7 | 1 .000 | h |
| Arg | 30 | 0.88 | 0.71 | 1.08 | 0.525 | 0.390 | 0.08 5 | b,i |

[a] See footnote to Table II I.

TABLE VI: Conformational Preferences[a] of Amino Acids for $\alpha$ Helix, $\beta$ Sheet and Reverse Turns.

| type of secondary struct | favoring (h) | indifferent (i) | breaking (b) |
|---|---|---|---|
| $\alpha$ helix | Ala, Leu, Met, His. Glu, Gln. Lys, (Cys) | Val. Ile. Phe, Trp. Asp, Asn, Arg | Tyr. Thr. Gly, Ser. Pro |
| $\beta$ sheet | Val, Ile, Phe, (Trp), Tyr, Thr | Ala, Leu. Met, His. Gly, Ser. Arg | Glu. Gln. Lys, Asp, Asn, Pro. cys |
| reverse turn | Gly. Ser. Asp. Asn. Pro | Gly. Gln, Lys, Tyr, Thr. (Arg) | Ala, Leu, Met. His, Val, Ile. Phe, (Trp). (Cys), (Arg) |

[a] These preferences are assigned with at least 75% confidence for the h and b classes (24 out of 40 are with at least 95% confidence). unless the amino acid is enclosed in parentheses when the confidence is as low as 56%. The confidence with which the i structure code is assigned is generally lower than for the h and b structure codes.

ticularly broad with only 27 counted occurrences, giving a 61% chance of the h assignment. All six of the ,&sheet breaking amino acids (b) have been given this preference with more than 75% confidence. Three of the seven P-sheet indifferent (i) amino acids have a significant chance of a different structure code: for Ala the probabilities are i (52%), b (48%); for His they are i (44%), h (47%): and for Gly they are i (63%). b (36%). Note that as the His $P_{it}$ value of 1.08 is less than 1.10. His is given the i structure code even though the h structure code would have been slightly more probable.

All five amino acids (Table V) defined as reverse-turn favoring (h) have been given this structure code with at least 92% confidence. Seven of the nine amino acids defined as reverse-turn breaking (b) have been given this structure code with more than 90% confidence; the other two have the b assignment with 66 and 7 1% confidence (for Cys and Trp). Two of the six amino acids defined as reverse-turn indifferent (i) have a significant chance of an alternative structure code: for Tyr the probabilities are i (50%). h (38%); for Arg they are i (39%), b (53%).

*Conformational Preferences and Classification.* Table VI summarizes the conformational preferences of the amino acids derived from their $P_{ij}$ values. The most remarkable feature of this classification is that 19 of the 20 amino acids favor only one type of secondary structure: no amino acid favors two types of secondary structure and only Arg favors none. Eighteen of the 20 amino acids dislike only one type of secondary structure: Pro dislikes both $\alpha$ helix and $\beta$ sheet; Cys dislikes both $\beta$ sheet and reverse turns (weakly).

Certain correlations between the chemical structure of the amino acid and the conformational preference are clear in Table VI. All the amino acids whose side chain branches at the $C^{\beta}$ atom (Val, Ile, and Thr) and three of the four amino acids with aromatic side chains (Phe. Tyr, Trp) favor $\beta$ sheet. His. the other aromatic amino acid. also favors $\beta$ sheet but too weakly to be included in the present classification (see Table IV). All the amino acids with a short polar side chain (Ser. Asp, and Asn) and the two *special* amino acids with no side chain (Gly) and a cyclized side chain (Pro) favor reverse turns. All the other amino acids (except for Arg). which do not fall into the above classes. favor $\alpha$ helix.

It is possible to understand these preferences on the basis of amino acid stereochemistry. The P-sheet favoring amino acids all have restricted conformational freedom as a result of the branching at the $C^{\beta}$ or the large aromatic side chains, indicating that "bulkiness" favors $\beta$ sheet. The reverse-turn favoring amino acids all have a tendency to change the direction of the chain: the polar side-chain groups of Ser, Asp, and Asn can hydrogen bond back to the main chain and stabilize turns: Gly has great backbone conformational freedom without any side-chain steric hinderance, and Pro is almost locked into a turn by virtue of side-chain cyclization. The $\alpha$-helix favoring residues from the most diverse class with nonpolar (Ala. Cys. Leu, and Met), polar (Gln and His), and charged (Glu and Lys) amino acid side chains. It is almost as if amino acids favor $\alpha$ helix by default, with those that are neither bulky nor short and polar falling into this class. On the basis of the chemical

structure. Arg would be expected to fall into the helix-favoring classification.

The dislikes of the amino acids can also be correlated with their chemical structures. All three amino acids with hydroxyl groups as part of the side chain (Thr, Thr, and Ser) break $\alpha$ helix, as do the special amino acids Gly, which has no side-chain, and Pro, which cannot form hydrogen bonds inside $\alpha$ helices. All five amino acids with charged side chains (Glu. Lys, and Asp) or with amide group side chains (Gln and Asn) break $\beta$ sheet. Proline, which cannot form two hydrogen bonds. also breaks sheets. All the nonpolar amino acids (Ala. Leu. Met, Phe, Val, Ile, Phe, and Trp) dislike reverse turns. His, which is both aromatic (nonpolar) and polar, and Cys, which is sometimes nonpolar, also break reverse turns.

It is also possible to understand these dislikes on the basis of stereochemistry. The hydroxyl group must interfere with the backbone hydrogen bonds that stabilize $\alpha$ helix; such interference is easily conceivable for Ser and Thr, where the -OH is close to the backbone, but not for Tyr, where the -OH is attached to distant side of an aromatic ring. The charged (Glu. Lys, and Asp) and amide group side chains (Gln and Asn) must interfere with the hydrogen bonds that stabilize $\beta$ sheets. Note that, although polar side chains seem to disrupt the peptide hydrogen bonds of both $\alpha$ helix and $\beta$ sheet, thev act differently: Tyr and Thr which disrupt $\alpha$ helix actually favor $\beta$ sheet; Glu. Gln, and Lys. which disrupt $\beta$ sheet. actually favor $\alpha$ helix. The nonpolar amino acids dislike reverse turns as their hydrophobic side chains prefer to be in the protein interior, while the reverse turns are usually on the protein surface.

Seventeen of the 20 amino acids can be classified into six classes on the basis of their conformational preferences (see Table VII). Class H 1 consists of the nonpolar $\alpha$-helix favoring amino acids (Ala, Leu, Met, and His); class H2 consists of the charged or polar $\alpha$-helix favoring amino acids (Gln, Glu, and Lys). It is not clear why His behaves more like a nonpolar amino acid, although its bulky aromatic side chain may increase its preference for $\beta$ sheet, so that it is dissimilar to Gln, Glu. and Lys which disrupt $\beta$ sheet. Class B1 consists of nonpolar P-sheet favoring residues (Val, Ile, Phe, and Trp); class B2 consists of polar P-sheet favoring residues (Tyr and Thr). Class T1 consists of the turn-favoring residues with very short side chains (Gly and Ser); class T2 consists of the turn-favoring residues with longer side chains (Asp and Asn). Three amino acids fall into special classes as follows: Cys in class H3, Pro in class T3. and Arg in class N 1.

*Comparison with Previously Published Preferences.* Many workers have assigned conformational preferences to amino acids using frequencies of occurrence in protein structures, helixjcoil transition parameters. and energy calculations (see Table VIII). Most of the early studies (Guzzo, 1965; Prothero, 1966; Cook, 1967: Ptitsyn and Finkelstein, 1970; Crawford et al., 1973) assigned the preferences to only a few of the 20 naturally occurring amino acids; about one-third of these assignments still agree with the present assignments made in this work (TW). Some early studies did assign $\alpha$-helix preferences to all 20 amino acids using energy calculations (Kotelchuck and Scheraga, 1969) and analysis of protein structures (Lewis and Scheraga, 197 1; Pain and Robson, 1970); only one-half of these assignments agree with this work. Four recent studies have analyzed more than 15 protein structures (Nagano, 1973; Chou and Fasman, 1974a,b; Robson and Suzuki, 1976; Chou and Fasman. 1977. Chou et al., 1977); overall there is good agreement between these assignments and those made in the present study.

In some cases the assignment made previouslv ,disagrees with the present assignment but would not be too improbable in the present studv. For $\alpha$ helix, Cvs could be i with 43% chance, Phe could be h with 45% chance, and Asn could be b with 45% chance. For $\beta$ sheet. Met could be h with 25% chance, Gly could be b with 35% chance, and Ser could be b with 23% chance. For reverse turns, His could be i with 21% chance and Trp could be i with 28% chance. For other cases, the disagreements are more significant. For $\alpha$ helix, His could be i with 16% chance, Lys could be i with 5% chance. and Val could be h with 0.4% chance. For $\beta$ sheet, Cys could be h with 0.7% chance, Gln could be b with 1% chance, His could be b with 8.4% chance, and Asp could be i with 1.6% chance. For reverse turns, Cys could be h with 7.2% chance, Glu could be b with 6.9% chance, Phe could be i with 1.5% chance.

TABLE VII: Classification of Amino Acids by Conformational Preferences.

| class | structure code for $\alpha$ helix | $\beta$ sheet | reverse turn |
|---|---|---|---|
| helix favoring | | | |
| (H 1) Ala, Leu, Met. His | h | i | b |
| (H2) Gln, Glu, Lys | h | b | i |
| (H3) Cys | h | b | b |
| sheet favoring | | | |
| (B 1) Val. Ile, Phe, Trp | i | h | b |
| (B2) Tyr, Thr | b | h | i |
| reverse-turn favoring | | | |
| (T1) Ser, Gly | b | i | h |
| (T2) Asp, Asn | i | b | h |
| (T3) Pro | b | b | h |
| neutral | | | |
| (N1) Arg | i | i | b,i |

There is a measure of agreement in all the previously published conformational preferences with the following 19 out of 60 assignments remaining unchanged: $\alpha$-helix favoring Ala. Leu. and Met; $\alpha$-helix breaking Ser and Pro; $\beta$-sheet favoring Val, Ile, Phe, and Thr; P-sheet breaking Glu, Asn, and Pro: reverse-turn favoring Gly, Ser, Asp, Asn, and Pro; reverse-turn breaking Ala, Leu, Val, and Ile.

The studies by Robson and Suzuki (1976) on 25 proteins and by Chou and Fasman (1977) on 29 proteins are in closest agreement with the present study, although there are still 18 and 16 disagreements, respectively. One main difference between the present assignments and the others is that here no amino acid favors more than one type of secondary structure. In the assignments made by Robson and Suzuki (1976), Cys favors both $\beta$ sheet and reverse turns, while Leu, Met. Val. and Phe favor both $\alpha$ helix and $\beta$ sheet. In the assignments made by Chou et al. (1977), Cys favors both $\beta$ sheet and reverse turns, while Leu, Gln, and Phe favor both $\alpha$ helix and $\beta$ sheet. Such a tendency to favor more than one type of secondary structure makes it difficult to classify the amino acids on the basis of their conformational preferences. The classification made by Robson and Suzuki (1976) is in fact quite different from the present one (see Table VII), with the following six classes of amino acids: Glu, Ala: Pro, Gly; Val, Leu, lle, Met, Phe; Ser, Thr, Asn. Gln, His: Asp. Lys, Arg, Trp: Cys. Tyr.

*Predicting Secondary Structures Using Conformational Preferences.* One of the major aims of previous studies of the conformational preferences of amino acids has been the prediction of the secondary structure from the amino acid sequence (Prothero, 1966; Kotelchuck and Scheraga, 1969: Robson and Pain. 1970: Chou and Fasman, 1974b: Maxfield and Scheraga, 1976). The biggest obstacle to a reliable prediction is that amino acids do not show strong likes or dislikes

TABLE VIII: Comparison[c] of the α-Helix, &Sheet. and Reverse-Turn Preferences of This Work with Published Preferences.[a]

| amino acid | α helix; no. of structures: TW 66 | G 3 | P 4 | SE 4 | C 4 | KS 4 | LS | PR 11 | PF 9 | CLS 8 | N 17 | CF 15 | RS 25 | CF' 29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | h | | h | h | h | h | h | h | h | h | h | h | h | h |
| Cys | h | | | h | | h | | | b | | | b | | b |
| Leu | h | | h | h | h | h | h | h | h | h | h | h | h | h |
| Met | h | | | h | | h | h | h | | h | h | h | h | h |
| Glu | h | b | h | | | h | h | h | h | h | h | h | h | h |
| Gln | h | | h | | | h | h | h | | h | h | h | | h |
| His | h | b | | | | b | h | h | h | h | h | h | | |
| Lys | h | | | | | b | | b | | | h | | h | h |
| Val | i | | h | h | h | h | h | | | h | | h | h | |
| Ile | | | h | h | | h | h | h | | h | i | ı | i | i |
| Phe | | | | h | b | h | i | i | | h | h | h | h | h |
| Tyr | b | | | h | | b | i | b | b | | b | b | b | b |
| Trp | | | | h | | h | h | h | | h | i | h | i | i |
| Thr | b | | h | | | b | | b | | | b | b | b | b |
| Gly | b | | | | | | b | b | b | | b | b | b | b |
| Ser | b | | | | | b | b | b | b | | b | b | b | b |
| Asp | | b | | | b | b | | b | | | | | | |
| Asn | | | | | b | b | b | b | b | | b | b | b | b |
| Pro | b | b | b | | b | b | b | b | b | | b | b | b | b |
| Arg | | | | | | h | | | | | b | b | | |
| score[b] | | 3/4 | 3/8 | 4/9 | 4/7 | 10/20 | 8/20 | 6/20 | 7/10 | 4/10 | 4/10 | 7/20 | 4/20 | 4/20 |

| β sheet: no. of structures: TW 66 | PF 9 | CLS 8 | N 17 | CF 15 | RS 25 | CF' 29 |
|---|---|---|---|---|---|---|
| | | | | | | b |
| b | | | h | h | h | h |
| i | h | h | h | h | h | h |
| | | h | h | h | h | |
| b | | | b | b | b | b |
| b | | | h | h | | h |
| | | | b | b | b | b |
| b | | h | b | b | b | b |
| h | h | h | h | h | h | h |
| h | h | h | h | h | h | h |
| h | | h | h | h | h | h |
| h | | h | | h | h | h |
| h | | | h | h | | h |
| h | | h | h | h | h | |
| | | | b | b | b | b |
| | | | b | b | | b |
| b | | | | b | b | b |
| b | | | b | b | b | b |
| b | | | b | b | b | b |
| | | | b | | | |
| | 1/3 | 3/8 | 10/20 | 8/20 | 10/20 | 7/20 |

| reverse turn: no. of structures: TW 66 | LMS 3 | PF 9 | CLS 8 | CF 15 | RS 25 | CF' 29 |
|---|---|---|---|---|---|---|
| b | b | b | | b | b | b |
| b | b | | h | h | h | h |
| b | b | b | | b | b | b |
| b | | | | b | b | b |
| i | b | b | | b | b | b |
| | b | | | b | | |
| b | b | | | b | | |
| | b | | h | i | | |
| b | b | b | | b | b | b |
| b | b | | | b | b | b |
| b | b | | h | b | b | b |
| | h | | h | h | h | h |
| b | h | | h | h | h | i |
| | h | | h | | | |
| h | h | h | h | h | h | h |
| h | h | h | h | h | h | h |
| h | h | h | h | h | h | h |
| h | h | | h | h | h | h |
| h | | | h | h | h | i |
| h | h | | h | | | |
| h | h | h | h | h | h | h |
| h | h | h | h | h | h | h |
| h | h | | h | h | h | h |
| h | h | h | h | h | h | h |
| h | h | h | h | h | h | h |
| | h | h | | | | |
| | 8/10 | 7/9 | 6/11 | 5/20 | 5/20 | 5/20 |

[a] The sources of the published amino acid preferences are indicated by the following abbreviations: TW, this work; G, Guzzo (1965); P, Prothero (1966); C, Cook (1967); SE, Schiffer and Edmunson (1967); KS, Kotelchuck and Scheraga (1969); LS, Lewis and Scheraga (1971); PR, Pain and Robson (1970); PF, Ptitsyn (1969) and Finkeistein and Ptitsyn (1971); CLS, Crawford et al. (1973); N, Nagano (1973); CF, Chou & Fasman (1974); RS, Robson and Suzuki (1976); CF', Chou and Fasman (1977) and Chou et al. (1977); LMS, Lewis et al. (1971). The number below each abbreviation gives the number of protein structures analyzed in that study (no allowance is made for redundant structures in a family). [b] Score gives the number of disagreements with the preferences of this work (TW) divided by the number of preferences defined. [c] In all cases the preferences have been assigned as in this work: $P > 1.1$ (h), $P < 0.9$ (b), otherwise (i).

for secondary structure: for α-helix, Met (the strongest former) occurs in α-helix only 1.5 times more often than expected by chance, while Pro (the strongest breaker) occurs only 0.5 times less often. Because only about one-third of the residues in the sample are α helical, only 39 out of 84 Met residues are in α helix (46%), while as many as 36 out of 219 Pro residues are in α helix (16%). The frequencies of occurrence in β sheet and reverse turns show similar trends, and any rule which automatically assigned residues by their individual preferences would fail badly.

The situation is really a little more promising, as α helices and β sheets consist of several adjacent residues so that runs

TABLE IX: Averaged[a] $P^{\alpha}$ and $P^{\beta}$ Vaiues of Residues That Are Really in $\alpha$ Helix and $\beta$ Sheet.

| range of $\langle P^{\alpha} \rangle$ or $\langle P^{\beta} \rangle$ | no. of res with $\langle P^{\alpha}_i \rangle$ in | | | no. of res with $\langle P^{\beta}_i \rangle$ in | | |
|---|---|---|---|---|---|---|
| | $\alpha$ helix | $\beta$ sheet | other | $\alpha$ helix | $\beta$ sheet | other |
| 0.6–0.7 | 12 | 23 | 46 | 2 | 2 | 6 |
| 0.7-0.8 | 43 | 171 | 197 | 878 | 37 | 115 |
| 0.8-0.9 | 171 | 434 | 454 | 414 | 247 | 508 |
| 0.9–1.o | 440 | 605 | 492 | 678 | 579 | 590 |
| 1.0–1.1 | 554 | 491 | 340 | 450 | 605 | 360 |
| 1.1-1.2 | 416 | 175 | 114 | 143 | 369 | 75 |
| 1.2–1.3 | 137 | 35 | 22 | 20 | 87 | 9 |
| 1.3-1.4 | 12 | 1 | 0 | 0 | 9 | 2 |
| total[b] | 1785 | 1935 | 1665 | 1785 | 1935 | 1665 |

[a] The $\alpha$-helix and $\beta$-sheet tendencies $P^{\alpha}$ and $P^{\beta}$ are averaged over five adjacent residues so that $\langle P^{\alpha} \rangle$ of the ith residue is $\langle P^{\alpha}_i \rangle = 0.2 \sum_{k=-2}^{k=2} P^{\alpha}_{i+k}$. [b] No averaged tendency is calculated for residues less than two away from the chain ends, explaining why there is a total of $1785 + 1935 + 1665 = 5385$ effective residues here. All assignments of secondary structure made in Levitt and Greer (1977) are included here.

of $\alpha$-helix favoring or P-sheet favoring residues indicate the particular type of secondary structure more clearly. This feature has been incorporated into many of the rules used to find $\alpha$ helices, with $\alpha$ helix predicted only where three out of five (or four out of six) successive residues favor $\alpha$ helix (Prothero. 1966; Kotelchuck and Scheraga, 1969; Chou and Fasman. 1974b). Unfortunately, considering adjacent residues in this way only helps to a limited degree. Table IX shows that the central residue of a stretch of five residues with a high average helix preference, $\langle P^{\alpha} \rangle$, is often not an $\alpha$-helix residue at all. Only 62% of the 9 12 residues at the center of a stretch of five residues with an average $\alpha$-helix tendency greater than 1.1 actually occur in $\alpha$ helix, while only 65% of the 7 14 residues at the center of a stretch of five residues with an average $\beta$-sheet tendency greater than 1.1 actually occur in $\beta$ sheet. Even those residues with the highest 4% of the $\langle P^{\alpha} \rangle$ values have a 28% chance of not being in $\alpha$ helix, and those with the highest 2% of the $\langle P^{\beta} \rangle$ values have a *24%* chance of not being in $\beta$ sheet. Clearly, no straightforward algorithm that considered local sequence alone could predict which of the residues in a stretch with high average $P^{\alpha}$ value should in fact be in $\beta$ sheet, reverse turns. or irregular structure. The occurrence of many residues in $\alpha$ helix or $\beta$ sheet depends. to a large extent, on interactions occurring in the overall three-dimensional structure of the protein. Nevertheless. the clear correlation found here between the preference of an amino acid for a particular type of secondary structure and the sterochemistry of the side chain does indicate that local interactions do have ***some*** influence on the chain conformation.

## Conclusion

Although the present analysis of secondary structure in globular proteins has used much more data and more consistent assignments of secondary structure than before, the present results are similar to previous results in that the preferences of amino acids for secondary structure are all very weak. The present study differs from previous studies in that here each amino acid prefers only one type of structure and most amino acids dislike only one type of structure, leading to a very clear classification of amino acid by their preferences and dislikes for secondary structure. This classification has shown that the chemical structure and sterochemistry of the amino acid plays a major part in determining its preference and dislike for secondary structure. The rule that has emerged from this study can be summarized as follows: Bulky amino acids. namely, those that are branched at the $\beta$ carbon or have a large aromatic side chain, prefer $\beta$ sheet. The shorter polar side chains prefer reverse turns, as do Gly and Pro, the special side chains. All other side chains prefer $\alpha$ helix, except Arg which has no preference. The polar side chains with hydroxyl groups disrupt $\alpha$ helix. the other polar side chain disrupt $\beta$ sheet, and the hydrophobic side chains disrupt reverse turns.

## Acknowledgments

I am extremely grateful to the numerous X-rav crystallographers who have kindly provided me with such a wealth of data.

## References

Chou. P. Y., and Fasman, *G.* D. (1974a). *Biochemistry 13,* 21 l-221.

Chou, P. Y., and Fasman, G. D. (1974b), *Biochemistry 13, 222-245.*

Chou, P. Y.. and Fasman, G. D. (1977), *J. Mol. Biol. 115,* 135–175.

Chou, P. Y., Fasman. G. D., and Adler, A. J. (1977), in The Molecular Biologv of the Mammalian Genetic Apparatus, Ts'o, P. 0. P., Ed., Amsterdam, Elsevier/North Holland Biomedical Press, pp 1–52.

Cook, D.A. (1967). *J. Mol. Biol. 29,* 167-171.

Crawford, J. L., Lipscomb. W. N.. and Schellman. C. G. *(1973). Proc. Natl. Acad. Sci. U.S.A. 70, 538-542.*

Finkelstein. A. V.. and Ptitsyn. 0. B. (*197* 1), *J. Mol. Biol. 62,* 613-624.

Guzzo, A. V. (1965), *Biophvs J. 5,* 809-821.

Kendrew, J. C., Dickerson, R. E.. Strandberg, B. E.. Hart, R. G., Davies. D. R., Phillips, D. C.. and Shore, V. C. (1960), ***Nature (London)*** *185,* ***422-427.***

Kotelchuck, D., and Scheraga. H. A. (1969), *Proc. Natl. Acad. Sci. U.S.A. 62,* 14-2 1.

Lindley, D. V. (1964), *Ann. Math. Stat. 3.5,* 1622.

Levitt, M., and Greer, J. (1977), *J.* ***Mol.*** *Biol. 114,* 181–*239.*

Lewis, P. N., and Scheraga, H. A. (1971), *Arch. Biochem. Biophvs. 144, 576-583.*

Lewis, P. N.. Momany, F. A., and Scheraga, H. A. (197 1), *Proc. Natl. Acad. Sci. U.S.A. 68, 2293-2297.*

Maxfield, F. R., and Scheraga, H. A. (1976), *Biochemistry* 15, 5138-5153.

Nagano. K. (1973), *J. Mol. Biol. 75, 401-420.*

Pain. R. H.. and Robson, B. (1970), *Nature* ***(London) 227,*** *62-63.*

Prothero, J. W. (1966), *Biophvs. J. 6,* 367–370.

Ptitsyn, 0. B. (1969). *J.* ***Mol.*** *Biol. 42,* 501–5 10.

Ptitsyn, 0. B., and Finkelstein, A. V. (1970), *Biofizika* 1.5, *757.*

Robson, B., and Suzuki, E. (1976), *J. Mol. Biol.* 107, 327-356.

Schiffer, M., and Edmundson, A. B. (1967), 121-135.

Tanaka, S., and Scheraga. H. A. (1976), ***Macromolecules 9,*** 142-159.