# Conformational Preference Functions for Predicting Helices in Membrane Proteins

DAVOR JURETIĆ,[1,*] BYUNGKOOK LEE,[2] NENAD TRINAJSTIĆ,[3] and ROBERT W. WILLIAMS[4]

[1]Natural Sciences and Arts Department, University of Split, N. Tesle 12, 58000 Split, Croatia; [2]NCI, Building 37, Room 4B15, National Institutes of Health, Bethesda, Maryland 20892, USA; [3]Rudjer Bošković Institute, Bijenička 54, 41000 Zagreb, Croatia; and [4]Department of Biochemistry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, Maryland 20814-4799, USA

## SYNOPSIS

A suite of FORTRAN programs, PREF, is described for calculating preference functions from the data base of known protein structures and for comparing smoothed profiles of sequence-dependent preferences in proteins of unknown structure. Amino acid preferences for a secondary structure are considered as functions of a sequence environment. Sequence environment of amino acid residue in a protein is defined as an average over some physical, chemical, or statistical property of its primary structure neighbors. The frequency distribution of sequence environments in the data base of soluble protein structures is approximately normal for each amino acid type of known secondary conformation. An analytical expression for the dependence of preferences on sequence environment is obtained after each frequency distribution is replaced by corresponding Gaussian function. The preference for the $\alpha$-helical conformation increases for each amino acid type with the increase of sequence environment of buried solvent-accessible surface areas. We show that a set of preference functions based on buried surface area is useful for predicting folding motifs in $\alpha$-class proteins and in integral membrane proteins. The prediction accuracy for helical residues is 79% for 5 integral membrane proteins and 74% for 11 $\alpha$-class soluble proteins. Most residues found in transmembrane segments of membrane proteins with known $\alpha$-helical structure are predicted to be indeed in the helical conformation because of very high middle helix preferences. Both extramembrane and transmembrane helices in the photosynthetic reaction center M and L subunits are correctly predicted. We point out in the discussion that our method of conformational preference functions can identify what physical properties of the amino acids are important in the formation of particular secondary structure elements. © 1993 John Wiley & Sons, Inc.

## INTRODUCTION

Accurate prediction of membrane protein folding is one of the most urgent tasks in life sciences. For many such proteins location of transmembrane and extramembrane $\alpha$-helices must be determined before better understanding can be gained on how they function. Because only a small number of membrane proteins has been examined by X-ray crystallography,[1] statistical analysis of solved structures has limited utility as a help for predicting secondary structure segments in such proteins. In contrast, a structure of almost 100 different globular soluble proteins is known with good atomic resolution.[2] All of the best current algorithms for secondary structure prediction from amino acid sequence data[3-9] have been developed and tested using the data base of soluble protein structures. They also suffer from the same set of serious shortcomings. The prediction accuracy from 60 to 70% in the three-state model (helix, sheet, and coil) is not enough to serve as a solid foundation for a tertiary structure prediction. For integral membrane proteins the use of these algorithms is questionable since their training was performed on soluble proteins. Indeed, the most widely used Chou–Fasman's[3] and Garnier–Osgu-

Biopolymers, Vol. 33, 255–273 (1993)
© 1993 John Wiley & Sons, Inc. CCC 0006-3525/93/020255-19

[*] To whom correspondence should be addressed.

thorpe–Robson (GOR) algorithms[4] are poor predictors of membrane protein folding motifs.[10] In the rare cases when we are satisfied with the prediction accuracy, it is not possible to tell what features of the amino acids were recognized by used statistical algorithms for the prediction of a particular secondary structure.

The alternative approach to the statistical one is to guess from the start about physical or chemical properties of amino acids that should be important for the creation of, for example, transmembrane segments in membrane proteins. The Kyte and Doolittle method[11] and similar ones[12,13] plot smoothed hydrophobicity values along protein sequence and identify transmembrane segments as the most hydrophobic segments. The membrane-associated helices can be identified with the aid of the hydrophobic moment plot.[14]

In this report we develop a method for finding the dependence of conformational preference of an amino acid on local average hydrophobicities of its sequence neighbors. Assuming that soluble protein structures contain statistical information about folding of hydrophobic domains, we sought to extract from the data base of such proteins a set of simple preference functions that are useful for predicting folding motifs in membrane environment as well. The major part of this study uses the buried surface area hydrophobicity scale of Rose et al.[15] to calculate conformational preferences. That scale is based on partitioning of amino acids in protein interior and on the surface. The conservation of the side-chain hydrophobicity over residues from four protein families is excellent when that scale is used to measure hydrophobicity.[16] Furthermore, when Rose's scale is used to calculate preferences, it is the best predictor of the $\alpha$-helices in the photosynthetic reaction center M subunit.[17] Just by comparing simple preference functions based on buried surface area in consecutive primary structure segments and by assigning the conformation to the highest value, the secondary structure of 5 membrane proteins is predicted with an accuracy of 66% and their $\alpha$-helical conformation with an accuracy of 79% (this work and recently published preliminary account[18]).

The method of conformational preference functions[19] is not restricted to buried surface area property or to hydrophobicity parameters derived from solution measurements. It can be used with any set of 20 conformational parameters. In the case of buried surface area parameters we shall show that our method is better than using these parameters directly to predict secondary structure such as an $\alpha$-helix. In general, preference functions can identify those physical-chemical properties of the amino acids that are important in the secondary structure formation.

## METHODS

### Structural Data Base

The data sample consisted of monomers of 90 different proteins (a total of 16109 residues) known with resolution equal or better than 0.3 nm (Table I). Proteins were selected from the Brookhaven Protein Data Bank (PDB).[20] Secondary structures ($\alpha$-helix, $\beta$-sheet, turn, and undefined) were assigned to all residues using the program DSSP by Kabsch and Sander.[21] An undefined structure was defined as a piece of low curvature not in H-bonded structure.[21] The $\alpha$-helix included the 3, 4, and 5 helix, the $\beta$-sheet included the $\beta$-bridge, and the turn included the bend. In the three-state prediction model, turn and undefined conformation were lumped together in the coil conformation. For secondary structure prediction $\alpha$-helix conformation is further subdivided into middle helix, N-terminal helix, C-terminal helix, and short helix. A short helix has 5 or less residues. Longer helices have 3 (if length is less than 8 residues) to 4 N-terminal and C-terminal residues, while all remaining residues are considered to be in the middle helix conformation.

Using the sliding window method,[22] local environment $X$ is assigned to each residue. The environment $X$ of the residue $n$ is defined to be the average of a selected property over 8 residues from $n - 4$ to $n + 4$, *excluding* residue $n$. The first 4 and last 4 residues in each sequence do not have assigned environment $X$. Unless specified otherwise, the property examined is the average buried surface area of amino acids in soluble proteins given by Rose et al.[15]

### Performance Measures

Four parameters are used for expressing performance and two for reporting the secondary structure prediction accuracy. In predicting any type of secondary structure for $N$ residues, we distinguish the numbers of residues that are associated with positive correct prediction $w$, negative correct prediction $x$, underpredictions $y$, and overprediction $z$. The correlation coefficient[23]

$$C_\alpha = \frac{w_\alpha x_\alpha - y_\alpha z_\alpha}{((x_\alpha + y_\alpha)(x_\alpha + z_\alpha)(w_\alpha + y_\alpha) \times (w_\alpha + z_\alpha))^{1/2}} \tag{1}$$

estimates how well the predicted secondary structure conformation is correlated with the observed one for each secondary structure type $\alpha$. It ranges from $-1.0$ (perfectly anticorrelated) to 1.0 (perfectly correlated). The success rate (or percentage of correctly predicted residues when multiplied by 100) $Q_\alpha = w_\alpha / N_\alpha$ estimates the prediction accuracy for a particular conformation $\alpha$ in one protein. A composite quality index for the three state model, $Q_3 = (w_\alpha + w_\beta + w_c)/N$ is the sum of conformational $Q$ indexes multiplied by a fraction of residues in each conformation. An overall $Q_\alpha$ and $Q_3$ index for a list of proteins is calculated as weighted average of all indexes for individual proteins so that longer proteins give correspondingly larger contribution to the overall prediction accuracy.

Two additional parameters in Tables V and VI, *h* and *hp*, report the percentage of residues, which are *not* in middle-helical conformation, among residues with local environment higher than 1.37 nm$^2$ (*h*) and percentage of residues, which are *not* in middle-helical conformation, among residues having middle-helix preference higher than 1.4 (*hp*). These numerical values are related to Rose's buried surface area scale only[15] and are chosen so that roughly 10% of the residues from the protein data base have higher buried surface environment and helical preference. A jackknife statistical procedure was used with protein data base when performance parameters were calculated for 11 $\alpha$-class proteins from that data base (Table V). Single $\alpha$-class proteins were removed from the data set of 90 proteins during the training procedure.

Abbreviations for the amino acid residues are given below together with the values for buried surface (in parentheses, the values are expressed in nm$^2$) from Rose et al.[15]: G, Gly(0.63); S, Ser(0.86); A, Ala(0.87); P, Pro(0.93); D, Asp(0.98); N, Asn(1.03); T, Thr(1.07); E, Glu(1.14); K, Lys(1.16); Q, Gln(1.19); C, Cys(1.32); V, Val(1.41); H, His(1.56); I, Ile(1.58); R, Arg(1.62); L, Leu(1.64); M, Met(1.73); Y, Tyr(1.78); F, Phe(1.94); and W, Trp(2.25).

## Preference Functions from Normal Approximation for Frequency Distribution of Amino Acids over Environment

Frequency distribution for lysine in the $\alpha$-helix conformation is shown in Figure 1A. Sequence environment on the *x* axis is considered as a continuous variable $X$ for a Gaussian curve chosen as a close fit for that distribution. Since $X$ values were obtained by the averaging procedure, normal function was expected from the Central Limit Theorem[24] to be a good fit for frequency distribution. The $\chi^2$ test of goodness of fit was used on all 20 amino acids in the protein data base. The tested null hypothesis was that the frequency distribution is approximately normal. As an example, the results for the choice of buried solvent-accessible surface scale (Rose's scale[15]) are presented in Table II. Reported numbers are probabilities $p$ that higher $\chi^2$ values can be found. The $p$ value between 0.10 and 0.90 is usually taken as evidence that there is no reason to reject the null hypothesis tested.[24] The software developed by SAS Institute (SAS Institute, Inc., Box 8000, Cary, North Carolina 27511) was used for the Kolmogorov–Smirnov normality tests (not reported), while the program developed in our laboratory was used for the $\chi^2$ tests. The tests were performed for all hydrophobicity scales used in this paper.

We define preference function $P_{ij}(X)$ as

$$P_{ij}(X) = p_{ij}(X)N/N_j \tag{2}$$

where $p_{ij}(X)$ is the frequency with which amino acid of type $i$ found in local environment $X$ occurs in a particular type $j$ of secondary structure. Using normal approximation mentioned above, $p_{ij}(X)$ is expressed as a ratio of one Gaussian function of $X$ (for conformation $j$) to the sum of all Gaussian functions of $X$ (for all conformations):

$$p_{ij}(X) = (N_{ij}/\sigma_{ij})\exp(-(X-\mu_{ij})^2/2\sigma_{ij}^2)/ \left(\sum_j (N_{ij}/\sigma_{ij})\exp(-(X-\mu_{ij})^2/2\sigma_{ij}^2)\right) \tag{3}$$

The average $\mu_{ij}$ and sample standard deviation $\sigma_{ij}$ of parameters $X$ are listed in Table III. The number of amino acids found in each conformation ($N_{ij}$) and fraction of conformation $j$ in the protein data set ($N_j/N$) are also listed in Table III.

Frequency distributions with $p$ values outside safe range were occasionally found with a seemingly random distribution among hydrophobicity scales, amino acid types, and secondary structure conformations. However, preference functions based on normal approximation for frequency distributions could still be used instead of preference points. For instance, the $p$ value of less than 0.01 for the frequency distribution of lysine in the undefined conformation (Table II) did not prevent close fit of preference function for lysine in the $\alpha$-helix conformation to preference points (Figure 1B).

**Table I Data Set of Protein Structures Used to Derive Preference Functions**[a]

| No. | PDB | Resol. | Protein | Class[b] | Unit | No. aa |
|---|---|---|---|---|---|---|
| 1 | 156B | 2.5 | Cytochrome B 562 *Escherichia coli* | a | | 103 |
| 2 | 155C | 2.5 | Cytochrome C 550 *Paracoccus denitrificans* | a | | 134 |
| 3 | 451C | 1.6 | Cytochrome C551 *Pseudomonas aeruginosa* | a | | 82 |
| 4 | 1ABP | 2.4 | L-Arabinose binding protein *Escherichia coli* | a/b | | 306 |
| 5 | 1ACX | 2.0 | Actinoxanthin *Actinomyces globisporus* | b | | 108 |
| 6 | 1BP2 | 1.7 | Phospholipase A2 Bovine pancreas | a + b | | 123 |
| 7 | 1CAC | 2.0 | Carbonic anhydrase form C Human | a + b | | 256 |
| 8 | 1CC5 | 1.5 | Cytochrome C5 *Azotobacter vinelandii* | a | | 83 |
| 9 | 1CCR | 1.5 | Cytochrome C Rice embryos | a | | 111 |
| 10 | 1CTF | 1.7 | L7-L12 50s ribosomal protein *Escherichia coli* | a/b | | 68 |
| 11 | 1CTX | 2.8 | $\alpha$-Cobratoxin Cobra | a/b | | 71 |
| 12 | 1ECO | 1.4 | Hemoglobin *Chironomus thummi* | a | | 136 |
| 13 | 1FBJ | 2.6 | Ig*a Fab fragment (J539) galactan-binding Mouse | b | L | 213 |
| 14 | 1FC2 | 2.8 | Immunoglobin Fc and fragment B Human | a | C | 44 |
| 15 | 1FX1 | 2.0 | Flavodoxin *Desulfovibrio vulgaris* | a/b | | 147 |
| 16 | 1GCN | 3.0 | Glucagon Porcine pancreas | a | | 29 |
| 17 | 1GCR | 1.6 | $\gamma$-II Crystallin Calf eye lens | b | | 174 |
| 18 | 1GP1 | 2.0 | Glutathione peroxidase Bovine erythrocyte | a/b | A | 184 |
| 19 | 1GPD | 2.9 | D-Glyceraldehyde-3-phosphate dehydrogenase Lobster | a/b | G | 333 |
| 20 | 1HHO | 2.1 | Hemoglobin A Human | a | A | 141 |
| 21 | 1HIP | 2.0 | High potential iron protein *Chromatium vinosum* | | | 85 |
| 22 | 1HMG | 3.0 | Hemagglutinin Influenza virus | b | A | 328 |
| 23 | 1HMZ | 2.0 | Hemerythrin Sipunculid worm | a | | 113 |
| 24 | 1IG2 | 3.0 | Immunoglobulin G1 Human | b | L | 216 |
| 25 | 1INS | 1.5 | Insulin Pig | a | B | 30 |
| 26 | 1LDX | 2.9 | Lactate dehydrogenase Mouse testicles | a/b | | 329 |
| 27 | 1LZ1 | 1.5 | Lysozyme Human | a/b | | 130 |
| 28 | 1MBD | 1.4 | Myoglobin Sperm whale | a | | 153 |
| 29 | 1MLT | 2.0 | Melittin Honey bee | a | A | 26 |
| 30 | IPP2 | 2.5 | Phospholipase A2 Western diamondback rattlesnake | a | R | 122 |
| 31 | 1PPT | 1.37 | Avian pancreatic polypeptide Turkey | a | | 36 |
| 32 | 1PYP | 3.0 | Inorganic pyrophosphatase Baker's yeast | a/b | | 281 |
| 33 | 1RHD | 2.5 | Rhodanase Bovine liver | a/b | | 293 |
| 34 | 1RN3 | 1.45 | Ribonuclease A Bovine pancreas | a + b | | 124 |
| 35 | 1SN3 | 1.8 | Scorpion neurotoxin Scorpion | a + b | | 65 |
| 36 | 1UBQ | 1.8 | Ubiquitin Human erythrocytes | | | 76 |
| 37 | 2ABX | 2.5 | Alpha-bungarotoxin Branded krait | | | 74 |

**Table I** (*Continued*)

| No. | PDB | Resol. | Protein | Class[b] | Unit | No. aa |
|---|---|---|---|---|---|---|
| 38 | 2ACT | 1.7 | Actinidin Kivi fruit | a + b | | 218 |
| 39 | 2ADK | 3.0 | Adenylate kinase Porcine muscle | a/b | | 194 |
| 40 | 2ALP | 1.7 | α-lytic protease *Lysobacter enzymogenes* | b | | 198 |
| 41 | 2APR | 1.8 | Acid proteinase Bread mold | b | | 325 |
| 42 | 2ATC | 3.0 | Aspartate carbamoyl transferase *Escherichia coli* | a/b | A | 305 |
| 43 | 2AZA | 1.8 | Azurin *Alcaligenes denitrificans* | a/b | A | 129 |
| 44 | 2B5C | 2.0 | Cytochrome B5 Bovine liver | a | | 85 |
| 45 | 2CDV | 1.8 | Cytochrome C3 *Desulfovibrio vulgaris* | a | | 107 |
| 46 | 2CGA | 1.8 | Chymotrypsinogen A Bovine pancreas | b | A | 245 |
| 47 | 2CPP | 1.63 | Cytochrome P450cam *Pseudomonas putida* | a | | 405 |
| 48 | 2CYP | 1.7 | Cytochrome C peroxidase Baker's yeast | a/b | | 293 |
| 49 | 2EBX | 1.4 | Erabutoxin Sea snake | b | | 62 |
| 50 | 2EST | 2.5 | Elastase Porcine pancreas | b | E | 242 |
| 51 | 2FD1 | 2.0 | Ferredoxin *Azotobacter vinelandii* | a + b | | 106 |
| 52 | 2GN5 | 2.3 | Gene 5 DNA binding protein Filamentous bacteriophage fd m13 | | | 87 |
| 53 | 2GRS | 2.0 | Glutathionine reductase Human erythrocyte | a/b | | 461 |
| 54 | 2LH7 | 2.0 | Leghemoglobin Yellow lupus root nodules | a | | 153 |
| 55 | 2LHB | 2.0 | Hemoglobin V Sea lamprey | a | | 149 |
| 56 | 2MDH | 2.5 | Cytoplasmic malate dehydrogenase Pig heart | a/b | A | 324 |
| 57 | 2MT2 | 2.3 | Cd, Zn, metallothionein Rat liver | | | 61 |
| 58 | 2PAB | 1.8 | Prealbumin Human plasma | a/b | A | 114 |
| 59 | 2PKA | 2.1 | Kallikrein A Pig pancreas | a/b | B | 152 |
| 60 | 2RHE | 1.6 | Immunoglobulin Bence-Jones λ variable domain Human | b | | 114 |
| 61 | 2RHV | 3.0 | Rhinovirus 14 Human virus | b | 1 | 273 |
| 62 | 2SBT | 2.8 | Subtilisin novo *Bacillus amyloliquefaciens* | a/b | | 275 |
| 63 | 2SGA | 1.5 | Proteinase A *Streptomyces griseus* | b | | 181 |
| 64 | 2SNS | 1.5 | Staphylococcal nuclease *Staphylococcus aureus* | a/b | | 141 |
| 65 | 2SOD | 2.0 | Cu, Zn, Superoxide dismutase Bovine erythrocyte | b | O | 151 |
| 66 | 2STV | 2.5 | Satellite tobacco necrosis virus coat protein Tobacco | b | | 184 |
| 67 | 2TBV | 2.9 | Tomato bushy stunt virus Tomato | b | A | 284 |
| 68 | 2TGT | 1.7 | Trypsinogen Bovine pancreas | b | | 233 |
| 69 | 3C2C | 1.68 | Cytochrome C2 *Rhodospirilum rubrum* | a | | 112 |
| 70 | 3CNA | 2.4 | Concanavalin A Jack bean | a/b | | 237 |
| 71 | 3CPV | 1.85 | Calcium-binding parvalbumin B Carp | a | | 108 |
| 72 | 3CYT | 1.8 | Cytochrome C Albacore tuna | a | O | 103 |
| 73 | 3FXC | 2.5 | Ferredoxin *Spirulina platensis* | a + b | | 98 |

**Table I** (*Continued*)

| No. | PDB | Resol. | Protein | Class[b] | Unit | No. aa |
|---|---|---|---|---|---|---|
| 74 | 3GAP | 2.5 | Catabolite gene activator protein *Escherichia coli* | a + b | A | 208 |
| 75 | 3ICB | 2.3 | Calcium-binding protein Bovine intestine | a | | 75 |
| 76 | 3LDH | 3.0 | Lactate dehydrogenase Dogfish muscle | a/b | | 329 |
| 77 | 3PGK | 2.5 | Phosphoglycerate kinase Baker's yeast | a/b | | 415 |
| 78 | 3PGM | 2.8 | Phosphoglycerate mutase Baker's yeast | a/b | | 230 |
| 79 | 3RP2 | 1.9 | Rat mast cell protease Rat | b | A | 224 |
| 80 | 4ADH | 2.4 | Apo-Liver alcohol dehydrogenase Horse liver | a/b | | 374 |
| 81 | 4DFR | 1.7 | Dihydrofolate reductase *Escherichia coli* | a/b | A | 159 |
| 82 | 4FXN | 1.8 | Flavodoxin (semiquinone form) Clostridium MP | a/b | | 138 |
| 83 | 4SBV | 2.8 | Southern bean mosaic virus coat protein | b | A | 199 |
| 84 | 5CPA | 1.54 | Carboxypeptidase A Bovine pancreas | a/b | | 307 |
| 85 | 5PTI | 1.8 | Trypsin inhibitor Bovine pancreas | a + b | | 58 |
| 86 | 5RXN | 1.2 | Rubredoxin *Clostridium pasteurianum* | a + b | | 54 |
| 87 | 6PAD | 2.8 | Papain Papaya | a + b | | 213 |
| 88 | 6PCY | 1.9 | Plastocyanin Poplar leaves | b | | 99 |
| 89 | 7TLN | 2.3 | Thermolysin *Bacillus thermoproteolyticus* | a + b | | 316 |
| 90 | 8CAT | 2.5 | Catalase Beef liver | a/b | A | 498 |

[a] Each protein is listed with its Brookhaven Protein Data Bank indentification code (PDB), crystallographic resolution in Ångstroms (Resol.), common name and source, the folding type (Class), the name of subunit used (Unit), and the number of residues in that subunit (No. aa).

[b] All a's and b's stand for $\alpha$'s and $\beta$'s, respectively.

## A Flow Diagram of PREF Algorithms

A flow diagram of the method employed by the suite of FORTRAN programs, PREF, is shown in Figure 2. The 1st set of programs DSSP,[21] SS4 and SS7 serves to determine secondary structure from the x-ray data (DSSP), to convert Kabsch–Sander files into our four-structure format (SS4) and to classify $\alpha$-helix segments into short helix, N-terminal helix, middle helix, and C-terminal helix (SS7). A frequency distribution over local environments $X$ for all amino acids in all conformations is calculated by PR if SS4 files are used, or by FREQ if SS7 files are used. Both programs also use selected hydrophobicity scale and appropriate class limits for environments. Files created by PR and FREQ are further analyzed by NORM and GAUS respectively to determine Gaussian parameters needed to construct preference functions. All the results from Table III except Table III(A1) data are one example for the GAUS.DAT file. It is the end result of the training procedure with PREF. For the testing procedure, sequence-dependent preferences are calculated by SP or SEDP using Eqs. (2) and (3) and the same scale of 20 conformational parameters. These programs produce a profile of smoothed preferences. The smoothing consists of calculating arithmetic mean of 7, 5, or 3 preferences for helical, sheet, and coil (turn and undefined) conformations respectively, and of associating the result with the central residue. The primary structure of a polypeptide with unknown secondary structure serves as an input file for SP or SEDP, which assigns conformation to each residue by comparing the value of smoothed preference for the various possible secondary structure types and by choosing the conformation with highest preference. A list of proteins of known secondary structure can also be used as an input file for SP or

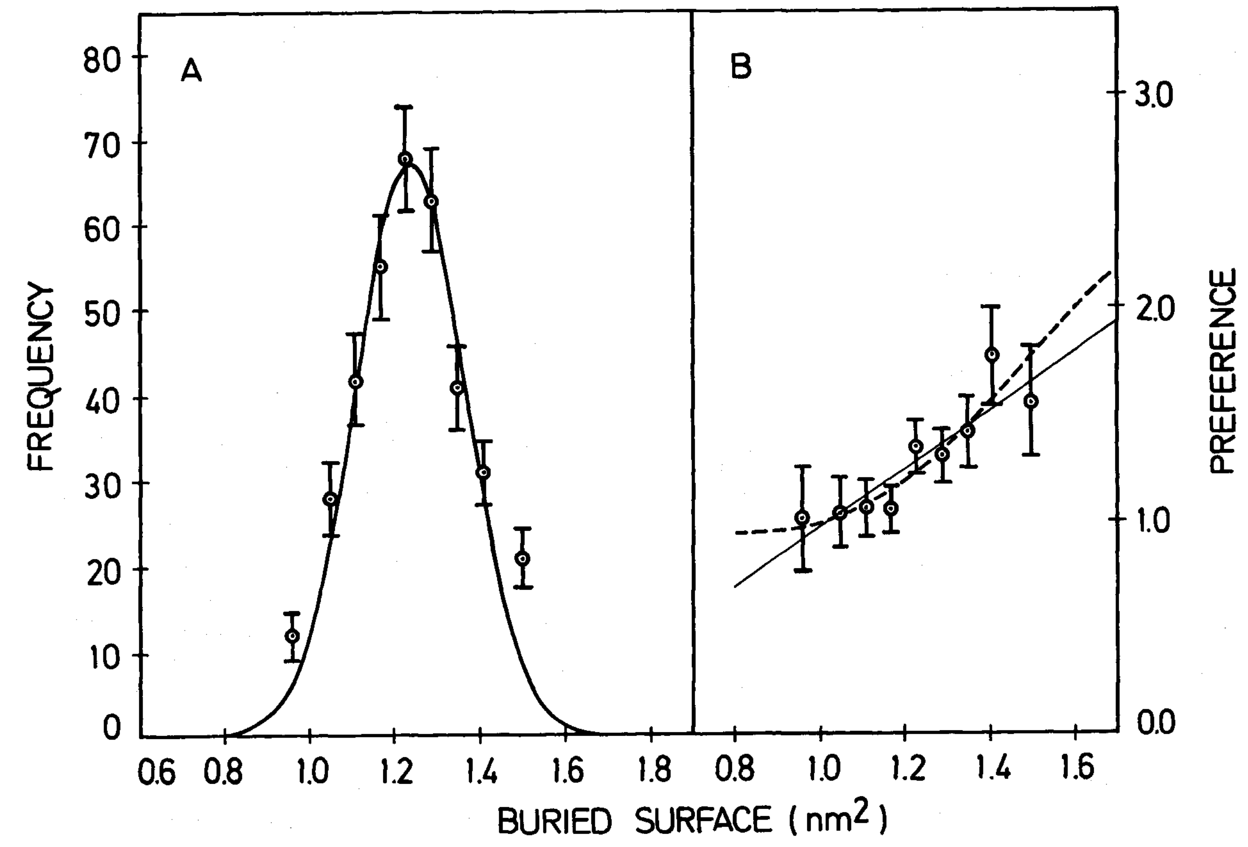

**Figure 1.** Lysine in the $\alpha$-helix conformation. (A) Frequency distributions for environments (see Methods) of lysine in the $\alpha$-helix. (B) Dependence of preference for lysine in the $\alpha$-helix on sequence environment. Sequence environment on the $x$ axis is obtained from Rose's scale of buried surface areas.[15] The data base of 90 soluble proteins is used to calculate frequency and preference points on the $y$ axis. The four-state model ($\alpha$-helix, $\beta$-sheet, turn, and undefined conformation) is used. Nine frequency points are obtained by grouping the environments in total of nine classes and counting the number of occurrence of lysine in the $\alpha$-helix conformation in each class.[44] Chosen class separation is 0.06 nm$^2$. In (A) vertical bars are interval estimation (one standard deviation) based on the assumption that frequency values are approximately normally distributed in the $y$ direction.[25] The full line for frequency distribution is calculated as numerator of Eq. (3), multiplied by a constant factor: (0.06 nm$^2$) / $\sqrt{2\pi}$ to find the area of the histogram. It is normal frequency distribution for 361 lysine environments determined by their mean: 1.2368 nm$^2$ and standard deviation 0.1272 nm$^2$. In (B) the line with short dashes is derived from Eqs. (2) and (3). The straight line is linear regression line through 9 preference points. Vertical bars are one standard error above and below preference points.

SEDP. In that case the accuracy of secondary structure prediction is reported. Predicted helical conformations (middle helix, N-terminal helix, C-terminal helix, and short helix) are lumped together again into the $\alpha$-helix conformation by the SEDP program when performance statistics is reported for the three-state model ($\alpha$-helix, $\beta$-sheet, and coil). The first four residues at the N-terminal and last four residues at the C-terminal protein end are automatically assigned the undefined conformation. If middle helix preferences are not desired, only the left branch of the program flow is used (PR, NORM and SP). A suite of programs, PREF, is available on the collaborative basis.

## RESULTS

### The Dependence of Amino Acid Preferences on Local Sequence Environment

Frequency distributions for local sequence environments can be approximated with normal distribu-

**Table II Normality Tests for the Frequency Distribution of Environments Calculated from Buried Surface Amino Acids Scale[15,a]**

| Conformation | | | | |
|---|---|---|---|---|
| | Ala | Arg | Asn | Asp |
| α-Helix | < 0.01 (517) | 0.90 (181) | 0.12 (153) | 0.25 (245) |
| β-Sheet | 0.80 (211) | 0.60 (114) | 0.60 (105) | 0.08 (90) |
| Turn | 0.20 (264) | 0.85 (139) | 0.30 (257) | 0.45 (296) |
| Undefined | 0.50 (277) | 0.75 (114) | 0.97 (210) | 0.05 (268) |
| | Cys | Gln | Glu | Gly |
| α-Helix | 0.87 (81) | 0.85 (188) | 0.40 (352) | 0.20 (189) |
| β-Sheet | 0.55 (95) | 0.50 (108) | 0.88 (118) | 0.02 (191) |
| Turn | 0.25 (69) | 0.25 (143) | 0.90 (205) | 0.96 (191) |
| Undefined | 0.20 (93) | 0.35 (119) | 0.15 (138) | 0.70 (329) |
| | His | Ile | Leu | Lys |
| α-Helix | 0.40 (106) | 0.45 (231) | 0.12 (439) | 0.75 (361) |
| β-Sheet | 0.50 (79) | 0.05 (302) | 0.50 (330) | 0.17 (162) |
| Turn | 0.20 (78) | 0.97 (102) | 0.50 (194) | 0.60 (283) |
| Undefined | 0.50 (96) | 0.45 (156) | 0.80 (233) | < 0.01 (204) |
| | Met | Phe | Pro | Ser |
| α-Helix | 0.18 (98) | 0.50 (189) | 0.45 (109) | 0.35 (228) |
| β-sheet | 0.70 (73) | 0.15 (172) | 0.20 (59) | 0.60 (244) |
| Turn | 0.55 (38) | 0.01 (102) | 0.07 (254) | 0.45 (340) |
| Undefined | 0.45 (52) | 0.88 (108) | 0.70 (283) | 0.40 (323) |
| | Thr | Trp | Tyr | Val |
| α-Helix | 0.50 (224) | 0.50 (70) | 0.17 (134) | 0.80 (296) |
| β-sheet | 0.35 (268) | 0.45 (75) | 0.98 (176) | 0.75 (437) |
| Turn | 0.12 (330) | 0.70 (44) | 0.17 (118) | 0.25 (141) |
| Undefined | 0.96 (250) | 0.80 (43) | 0.60 (112) | 0.35 (250) |

[a] Reported numbers are probabilities $p$ that higher $\chi^2$ values can be found. Total number of environments for each case is given in parentheses. Class limits were (in nm$^2$): 1.02–1.44 in the steps of 0.06. The $\chi^2$ values were calculated by our FORTRAN program, while probabilities were extracted from the standard table (Appendix I in Ref. 68) for six degrees of freedom.

tions (Figure 1A and Table II). The distribution parameters for buried surface sequence environment $X$ (Rose's scale[15]) are collected in the Table III for each amino acid type $i$ in each secondary conformation $j$. The distribution averages $\mu_{ij}$ are higher in the $\alpha$-helix conformation than in the other three secondary conformations in most cases. If the means of the distributions in numerator and denominator of Eq. (3) are shifted, their ratio (which is proportional to preference Eq. (2)) will show the dependence on $X$. Increased preference for helix with increased $X$ is indeed observed for all amino acid types. The slope ($b$) of linear regression line is positive in each case and it is higher for corresponding amino acid types in the middle helix conformation (Methods) than in the $\alpha$-helix conformation (Table IV A, B).

Lysine has been selected to show that the preference function model based on the buried surface area is a good model even for an amino acid that is almost invariably located at the solvent-accessible surface of proteins. The preference function for lysine in the $\alpha$-helix (Figure 1B) was calculated from Eqs. (2) and (3) in the four conformations model. There are many better examples than lysine, such as valine, isoleucine, phenylalanine, leucine, glycine, and serine that have even more significant ($p \ll 0.01$ on Student's $t$ test in all cases; see Table IV) dependence of their $\alpha$-helix preferences on sequence environment.

The case of leucine illustrate the advantage of calculating leucine preference functions instead of assuming that leucine preference is independent of the nature of its neighbors in the sequence.[25] Figure 3 shows the dependence of the $\alpha$-helix conformation preference of leucine on local environment of (A) Rose's buried surface areas[15] and (B) Fauchère and Pliška hydrophobicities.[26] Figure 3A predicts a 2.6-fold increase in the probability of leucine assuming helix conformation, when its sequence neighbors have high ($X$ = 1.5 nm$^2$) rather than low ($X$ = 1.0 nm$^2$) potential to bury their surface area during protein folding. For the choice of Fauchère and Pliška hydrophobicity scale,[26] which is based on partitioning of amino acids between polar and nonpolar solvents, the $\alpha$-helix preference for leucine does not show any dependence on the local hydrophobic environment (Figure 3B).

Also for leucine, Figure 3 shows the dependence of $\beta$-sheet conformational preference on the local environment of (C) buried surface areas[15] and (D) hydrophobicities.[26] The probability of leucine assuming $\beta$-sheet conformation does not depend on the potential of its sequence neighbors to bury their surface area during protein folding (Figure 3C). Figure 3D predicts an almost 3-fold *increase* in the probability of leucine assuming $\beta$-sheet conformation, when its sequence neighbors have a high ($X$ = 2.5) rather than low ($X$ = 0.5) hydrophobic sequence environment. Linear regression analysis of data points (Table IV) confirmed the impression from Figure 3D about the significant positive dependence of leucine preference for $\beta$-sheet on the sequence hydrophobic environment and from Figure 3A about the significant positive dependence of leucine preference for $\alpha$-helix on buried surface areas of its neighbors ($p \ll 0.01$ on Student's $t$ test in both cases).

Other 19 amino acids were also analyzed as in Figure 3A. The increase in helix probability with an increase in buried surface environment is even higher for some other amino acids. Keeping the same definition of high and low buried sequence environment, the probability ratio is 3.0 for phenylalanine and 4.2 for tryptophan (Table IV). For the middle helix conformation (Methods), probability increase is still steeper. The probability ratio is 10.4, 11.4, and 21.4 for leucine, phenylalanine, and tryptophan, respectively (Table IV). In the leucine example, linear regression analysis gave 5.3 for the positive slope of middle helix preference function, which can be compared to the slope around 2.5, as seen in Figure 3A. Due to the smaller number of residues in each of the four new helical conformations, the errors in slope determination are also higher, but the dependence of preferences on sequence environment of buried surface areas remained very significant ($p \ll 0.01$ on Student's $t$ test in the case of leucine in middle helix conformation). Helix preference dependence on buried surface environment is significant ($p < 0.01$) for 12 amino acid types, while middle helix preference dependence on buried surface environment is significant for 13 amino acid types.

For low buried surface environment (1 nm$^2$), the probability of the middle helix conformation becomes negative for serine and proline if the linear fit for preference points is used (Table IV). When preference functions are used, such breakdown of the model (negative preferences) cannot occur.

These findings can be summarized as follows: All 20 natural amino acids show positive correlation between their (a) $\alpha$-helix preference and buried accessible surface area of their local primary structure neighbors, (b) middle-helix preference and buried accessible surface area of their local primary structure neighbors, and (c) $\beta$-sheet preference and hydrophobic environment that was calculated from hydrophobicity scale.[26] The dependence of $\alpha$-helix

**Table III Parameters[a] for the Construction of Gaussian Curves**

(A) Four Secondary State Conformations

| AA | $N_{ij}$ | $\mu_{ij}$ | $\alpha_{ij}$ | AA | $N_{ij}$ | $\mu_{ij}$ | $\alpha_{ij}$ |
|---|---|---|---|---|---|---|---|
| | (1) $\alpha$-Helix ($N/N_j$ = 3.4966) | | | | (2) $\beta$-Sheet ($N/N_j$ = 4.5037) | | |
| Ala | 517 | 1.2184 | 0.1254 | Ala | 211 | 1.2300 | 1.1303 |
| Cys | 81 | 1.2011 | 0.1143 | Cys | 95 | 1.1969 | 0.1173 |
| Leu | 439 | 1.2428 | 0.1229 | Leu | 330 | 1.2218 | 0.1202 |
| Met | 98 | 1.2386 | 0.1089 | Met | 73 | 1.2170 | 0.1300 |
| Glu | 352 | 1.2395 | 0.1229 | Glu | 118 | 1.2335 | 0.1169 |
| Gln | 188 | 1.2538 | 0.1155 | Gln | 108 | 1.2469 | 0.1216 |
| His | 106 | 1.2380 | 0.1184 | His | 79 | 1.2529 | 0.1274 |
| Lys | 361 | 1.2368 | 0.1272 | Lys | 162 | 1.2457 | 0.1171 |
| Val | 296 | 1.2301 | 0.1287 | Val | 437 | 1.2034 | 0.1243 |
| Ile | 231 | 1.2422 | 0.1196 | Ile | 302 | 1.1994 | 0.1213 |
| Phe | 189 | 1.2465 | 0.1243 | Phe | 172 | 1.1940 | 0.1279 |
| Tyr | 134 | 1.2314 | 0.1217 | Tyr | 176 | 1.2128 | 0.1246 |
| Trp | 70 | 1.2261 | 0.1194 | Trp | 75 | 1.2033 | 0.1232 |
| Thr | 224 | 1.2486 | 0.1243 | Thr | 268 | 1.2146 | 0.1230 |
| Gly | 189 | 1.2597 | 0.1218 | Gly | 191 | 1.2475 | 0.1352 |
| Ser | 228 | 1.2548 | 0.1129 | Ser | 244 | 1.2201 | 0.1317 |
| Asp | 245 | 1.2430 | 0.1179 | Asp | 90 | 1.2567 | 0.1256 |
| Asn | 153 | 1.2726 | 0.1305 | Asn | 105 | 1.2279 | 0.1186 |
| Pro | 109 | 1.2532 | 0.1232 | Pro | 59 | 1.2264 | 0.1429 |
| Arg | 181 | 1.2476 | 0.1247 | Arg | 114 | 1.2282 | 0.1318 |
| | (3) Turn ($N/N_j$ = 3.9212) | | | | (4) Undefined ($N/N_j$ = 4.1997) | | |
| Ala | 264 | 1.1697 | 0.1251 | Ala | 277 | 1.1813 | 0.1211 |
| Cys | 69 | 1.1652 | 0.1187 | Cys | 93 | 1.1546 | 0.1312 |
| Leu | 194 | 1.1779 | 0.1185 | Leu | 233 | 1.1801 | 0.1219 |
| Met | 38 | 1.1750 | 0.1195 | Met | 52 | 1.1817 | 0.1119 |
| Glu | 205 | 1.2337 | 0.1194 | Glu | 138 | 1.2070 | 0.1261 |
| Gln | 143 | 1.2007 | 0.1255 | Gln | 119 | 1.1653 | 0.1157 |
| His | 78 | 1.2164 | 0.1217 | His | 96 | 1.2017 | 0.1187 |
| Lys | 283 | 1.2060 | 0.1172 | Lys | 204 | 1.1791 | 0.1252 |
| Val | 141 | 1.1779 | 0.1178 | Val | 250 | 1.1731 | 0.1235 |
| Ile | 102 | 1.1831 | 0.1333 | Ile | 156 | 1.1790 | 0.1233 |
| Phe | 102 | 1.2075 | 0.1229 | Phe | 108 | 1.1757 | 0.1169 |
| Tyr | 118 | 1.1908 | 0.1144 | Tyr | 112 | 1.1663 | 0.1212 |
| Trp | 44 | 1.1436 | 0.1185 | Trp | 43 | 1.1565 | 0.1056 |
| Thr | 230 | 1.1760 | 0.1181 | Thr | 250 | 1.2025 | 0.1279 |
| Gly | 616 | 1.2084 | 0.1284 | Gly | 329 | 1.2028 | 0.1315 |
| Ser | 340 | 1.1912 | 0.1345 | Ser | 323 | 1.1895 | 0.1180 |
| Asp | 296 | 1.2115 | 0.1272 | Asp | 268 | 1.2171 | 0.1292 |
| Asn | 257 | 1.2111 | 0.1213 | Asn | 210 | 1.2200 | 0.1264 |
| Pro | 254 | 1.2141 | 0.1233 | Pro | 283 | 1.2071 | 0.1202 |
| Arg | 139 | 1.2042 | 0.1382 | Arg | 114 | 1.2096 | 0.1262 |

(B) Helical Conformations

| AA | # | $\mu_{ij}$ | $\alpha_{ij}$ | AA | # | $\mu_{ij}$ | $\alpha_{ij}$ |
|---|---|---|---|---|---|---|---|
| | (1) Middle helix ($N/N_j$ = 11.0172) | | | | (2) N-Helix ($N/N_j$ = 11.9105) | | |
| Ala | 175 | 1.2509 | 0.1165 | Ala | 128 | 1.1977 | 0.1311 |
| Cys | 20 | 1.2355 | 0.1185 | Cys | 19 | 1.1874 | 0.1022 |

**Table III** (*Continued*)

(B) Helical Conformations

| AA | # | $\mu_{ij}$ | $\alpha_{ij}$ | AA | # | $\mu_{ij}$ | $\alpha_{ij}$ |
|---|---|---|---|---|---|---|---|
| | (1) Middle helix ($N/N_j$ = 11.0172) | | | | (2) N-Helix ($N/N_j$ = 11.9105) | | |
| Leu | 136 | 1.2768 | 0.1115 | Leu | 95 | 1.2243 | 0.1226 |
| Met | 37 | 1.2419 | 0.1074 | Met | 19 | 1.2189 | 0.0941 |
| Glu | 81 | 1.2662 | 0.1223 | Glu | 126 | 1.2225 | 0.1177 |
| Gln | 53 | 1.2466 | 0.1050 | Gln | 56 | 1.2504 | 0.1207 |
| His | 25 | 1.2632 | 0.1124 | His | 20 | 1.1985 | 0.1110 |
| Lys | 110 | 1.2600 | 0.1139 | Lys | 60 | 1.1773 | 0.1368 |
| Val | 86 | 1.2530 | 0.1240 | Val | 83 | 1.2230 | 0.1316 |
| Ile | 84 | 1.2739 | 0.1100 | Ile | 56 | 1.2262 | 0.1380 |
| Phe | 57 | 1.2877 | 0.1022 | Phe | 51 | 1.2253 | 0.1303 |
| Tyr | 39 | 1.2721 | 0.1109 | Tyr | 33 | 1.2094 | 0.1294 |
| Trp | 26 | 1.2492 | 0.1291 | Trp | 17 | 1.1947 | 0.1250 |
| Thr | 59 | 1.2780 | 0.1071 | Thr | 67 | 1.2278 | 0.1344 |
| Gly | 49 | 1.2845 | 0.1205 | Gly | 64 | 1.2280 | 0.1233 |
| Ser | 55 | 1.2916 | 0.1084 | Ser | 53 | 1.2496 | 0.1024 |
| Asp | 61 | 1.2839 | 0.1163 | Asp | 98 | 1.2164 | 0.1093 |
| Asn | 48 | 1.2956 | 0.1127 | Asn | 38 | 1.2600 | 0.1412 |
| Pro | 17 | 1.3035 | 0.1245 | Pro | 54 | 1.2406 | 0.1287 |
| Arg | 61 | 1.2852 | 0.1183 | Arg | 45 | 1.2256 | 0.1315 |
| | (3) Short helix ($N/N_j$ = 18.2196) | | | | (4) C-Helix ($N/N_j$ = 12.1779) | | |
| Ala | 81 | 1.1957 | 0.1338 | Ala | 133 | 1.2093 | 0.1182 |
| Cys | 27 | 1.1715 | 0.1207 | Cys | 15 | 1.2260 | 0.1040 |
| Leu | 66 | 1.2139 | 0.1174 | Leu | 142 | 1.2358 | 0.1299 |
| Met | 8 | 1.2062 | 0.1047 | Met | 34 | 1.2535 | 0.1201 |
| Glu | 60 | 1.2215 | 0.1242 | Glu | 85 | 1.2521 | 0.1260 |
| Gln | 32 | 1.2375 | 0.1347 | Gln | 47 | 1.2770 | 0.1063 |
| His | 23 | 1.2052 | 0.1212 | His | 38 | 1.2621 | 0.1182 |
| Lys | 57 | 1.2404 | 0.1326 | Lys | 134 | 1.2428 | 0.1241 |
| Val | 46 | 1.2293 | 0.1161 | Val | 81 | 1.2133 | 0.1359 |
| Ile | 27 | 1.1948 | 0.1065 | Ile | 64 | 1.2345 | 0.1111 |
| Phe | 28 | 1.1807 | 0.1146 | Phe | 53 | 1.2574 | 0.1294 |
| Tyr | 28 | 1.2168 | 0.1179 | Tyr | 34 | 1.2182 | 0.1230 |
| Trp | 12 | 1.1925 | 0.0753 | Trp | 15 | 1.2487 | 0.1205 |
| Thr | 42 | 1.2536 | 0.1294 | Thr | 56 | 1.2389 | 0.1217 |
| Gly | 38 | 1.2713 | 0.1144 | Gly | 38 | 1.2697 | 0.1217 |
| Ser | 56 | 1.2175 | 0.0977 | Ser | 64 | 1.2602 | 0.1282 |
| Asp | 50 | 1.2368 | 0.1173 | Asp | 36 | 1.2547 | 0.1277 |
| Asn | 28 | 1.2504 | 0.1457 | Asn | 39 | 1.2726 | 0.1291 |
| Pro | 33 | 1.2561 | 0.1165 | Pro | 5 | 1.2000 | 0.0557 |
| Arg | 30 | 1.1903 | 0.0984 | Arg | 45 | 1.2567 | 0.1266 |

[a] The average $\mu_{ij}$ and sample standard deviation $\alpha_{ij}$ of sequence environments $X$ (see Methods) are given together with the total number of environments $N_{ij}$ in the protein data set for amino acid type $i$ in the secondary conformation $j$. For each conformation $j$ the fraction of that conformation in the protein data set is given as the inverse value $N/N_j$.

preference on hydrophobic environment and of $\beta$-sheet preference on buried surface environment is such that about 50% of amino acids have positive correlation with higher environment $X$, but that dependence is often not significant. Middle helix preferences have stronger dependence on buried surface environment than $\alpha$-helix preferences (Table IV) and because of that might be more useful in predicting helical structures when buried sequence environment is used to improve a prediction.

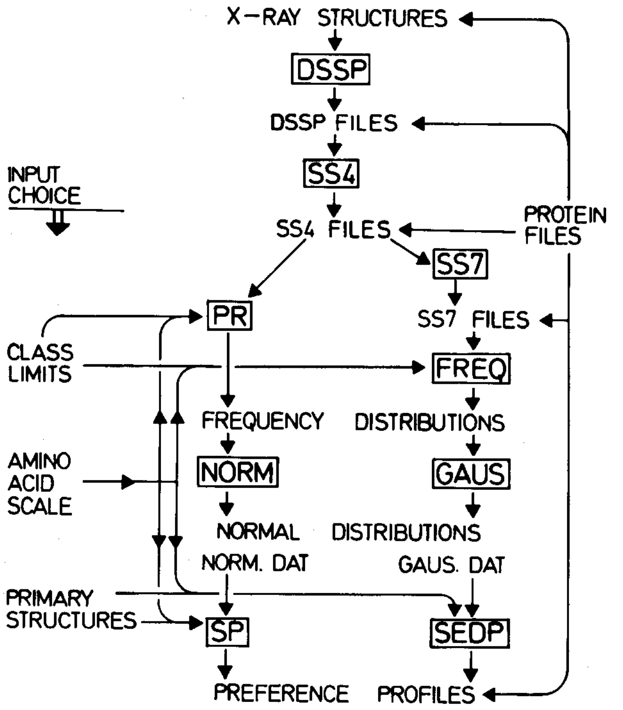

**Figure 2.** Flowchart of method employed by PREF. The DSSP is Kabsch–Sander's program.[21] All other programs (also in squares) are written by us in the FORTRAN language. The training procedure starts with the choice of the data base of crystallographically solved protein structures (x-ray structures) and an appropriate branch of the program flow. The middle-helix preferences are used only in the right branch (SS7, FREQ, GAUS and SEDP programs). A large number of hydrophobicity scales, physical property scales, and statistical scales are given in the programs and one of them must be also chosen at the start of each training or testing procedure. Class limits for environments (8 numbers) are needed only for the training procedure. The essential program for users who want to avoid the training procedure is SEDP (or SP) with corresponding data file GAUS.DAT (or NORM.DAT). The primary structure input for SP (or SEDP) contains the secondary structure, too, in the form of all U residues (U stands for undefined conformation) when secondary structure is not known, or in the form of H (for an $\alpha$-helix), B (for an $\beta$-sheet), T (for turn), and U residues when secondary structure is known.

## The Secondary Structure Prediction

To make optimal use of preference functions, one would have to find optimal environmental variable $X$ for each structural type and to combine the prediction results in a novel secondary structure prediction algorithm. We shall explore in this section only the utility of using preference functions based on Rose's hydrophobicity scale[15] for average buried surface areas. Preliminary tests with membrane

**Table IV Linear Regression Analysis[a] of Data Points for the Dependence of Preferences on Buried Surface Environment Calculated from the Rose's Hydrophobicity Scale[15]**

| AA | $a$ | $b$ | $s_a$ | $s_b$ | $F$ |
|---|---|---|---|---|---|
| | | (A) $\alpha$-Helix | | | |
| Ala | −0.322 | 1.349 | 0.372 | 0.299 | 20.30 |
| Cys | 0.244 | 0.475 | 0.692 | 0.558 | 0.72 |
| Leu | −1.696 | 2.457 | 0.358 | 0.288 | 72.81 |
| Met | −2.314 | 2.925 | 0.904 | 0.728 | 16.14 |
| Glu | 0.704 | 0.664 | 0.340 | 0.274 | 5.89 |
| Gln | −0.994 | 1.726 | 0.998 | 0.804 | 4.61 |
| His | −0.359 | 1.138 | 0.805 | 0.649 | 3.08 |
| Lys | −0.397 | 1.370 | 0.323 | 0.260 | 27.67 |
| Val | −1.312 | 1.877 | 0.319 | 0.257 | 53.22 |
| Ile | −1.648 | 2.212 | 0.308 | 0.248 | 79.41 |
| Phe | −1.895 | 2.543 | 0.580 | 0.467 | 29.62 |
| Tyr | −0.915 | 1.468 | 0.715 | 0.576 | 6.49 |
| Trp | −2.718 | 3.214 | 0.930 | 0.749 | 18.40 |
| Thr | −1.086 | 1.571 | 0.461 | 0.372 | 17.85 |
| Gly | −0.941 | 1.178 | 0.169 | 0.137 | 74.43 |
| Ser | −1.667 | 1.931 | 0.387 | 0.312 | 38.39 |
| Asp | −0.071 | 0.805 | 0.440 | 0.354 | 5.17 |
| Asn | −1.317 | 1.706 | 0.621 | 0.501 | 11.60 |
| Pro | −1.270 | 1.506 | 0.520 | 0.419 | 12.93 |
| Arg | −0.707 | 1.464 | 0.533 | 0.430 | 11.61 |
| | | (B) Middle helix | | | |
| Ala | −3.018 | 3.847 | 1.103 | 0.889 | 18.74 |
| Cys | −3.062 | 3.231 | 1.123 | 0.905 | 12.75 |
| Leu | −5.041 | 5.325 | 0.617 | 0.497 | 114.70 |
| Met | −3.485 | 4.174 | 1.351 | 1.089 | 14.70 |
| Glu | −2.819 | 3.296 | 0.839 | 0.676 | 23.74 |
| Gln | −1.266 | 1.888 | 0.975 | 0.785 | 5.78 |
| His | −1.452 | 1.821 | 0.862 | 0.695 | 6.88 |
| Lys | −2.088 | 2.754 | 0.588 | 0.474 | 33.77 |
| Val | −2.474 | 2.813 | 1.041 | 0.839 | 11.24 |
| Ile | −5.020 | 5.245 | 0.712 | 0.574 | 83.54 |
| Phe | −4.282 | 4.499 | 1.356 | 1.092 | 16.97 |
| Tyr | −1.978 | 2.302 | 1.325 | 1.068 | 4.65 |
| Trp | −7.313 | 7.496 | 4.242 | 3.418 | 4.81 |
| Thr | −2.324 | 2.506 | 0.907 | 0.731 | 11.75 |
| Gly | −2.057 | 2.078 | 0.599 | 0.483 | 18.55 |
| Ser | −3.002 | 2.981 | 0.380 | 0.306 | 94.93 |
| Asp | −2.572 | 2.791 | 0.722 | 0.582 | 22.97 |
| Asn | −3.038 | 3.127 | 0.819 | 0.660 | 22.44 |
| Pro | −1.522 | 1.515 | 0.472 | 0.380 | 15.90 |
| Arg | −2.948 | 3.434 | 1.723 | 1.389 | 6.12 |

[a] Our FORTRAN program was used to find the intercept ($a$), slope ($b$), standard error of intercept ($s_a$), standard error of slope ($s_b$), and $F$ value—$F = (b/s_b)^2 = t^2$—for the linear regression line drawn through preference points in each case. The $t$ test serves to discover whether or not an observed correlation coefficient $r$ is significantly greater than zero. For $N = 9$ and $F = (N - 2)/(-1 + 1/r^2) > 12.24$ (Appendix F in Ref. 68) the probability $p$ to have such a correlation coefficient in a sample drawn from population with zero correlation is less than 0.01.

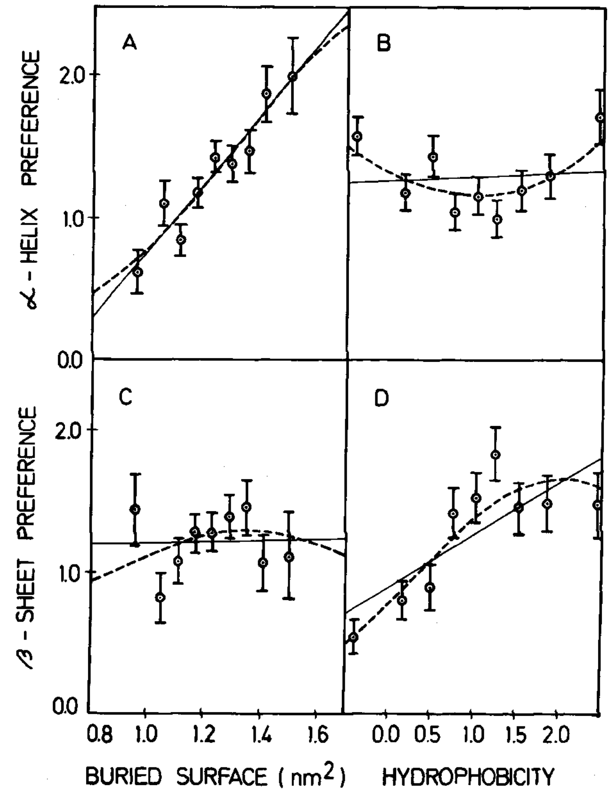

**Figure 3.** Leucine preferences for $\alpha$-helix structure (A and B) and for $\beta$-sheet structure (C and D). Rose's scale[15] is used to calculate environment $X$ of buried solvent accessible surfaces on the $x$ axis and preference functions (see Methods) in (A) and (C). The Fauchère–Pliška hydrophobicity scale[26] is used to calculate environment $X$ and preference functions in (B) and (D). The same notation is used as in Figure 1B.

proteins of known structure indicated that introduction of middle helical conformation increases the accuracy of a prediction.[17] Therefore, for all reported predictions, preference functions for seven different secondary structure conformations were used including four helical conformations: middle helix, N-terminal helix, C-terminal helix, and short helix (right branch of programs flow in the Figure 2). With this procedure, adopted middle helical preferences in some membrane proteins are very high (Figure 4).

For the training set of 90 different water-soluble proteins the overall percentage of correctly predicted residues (the success rate, or $Q_3$ index multiplied by 100) is 53% for the three-state model and 42% for the four-state model. The three-state prediction for helical residues is slightly better $Q_\alpha = 62\%$ of helical residues. The corresponding results for $\beta$-sheet and coil (turn and undefined residues) are $Q_\beta = 25\%$ and $Q_c = 57\%$, respectively. The correlation coefficients of Matthews[23] for helix, sheet, and coil residues are $C_\alpha = 0.22$, $C_\beta = 0.19$, and $C_c = 0.31$, respectively. Better overall results than that obtained from Chou–Fasman's predictive schemes[3,27] were not expected since decision constants and Chou–Fasman rules such as helix propagation rules and helix stop signals were not incorporated in the algorithm. With an optimal choice of decision constants,[27] the GOR program (algorithm from 1978[4]) results in 55%, 0.22, 0.23, 0.28 for $Q_3$, $C_\alpha$, $C_\beta$, and $C_c$, respectively, for the same set of 90 soluble proteins.

For the first testing set of proteins, a subset of 11 $\alpha$-class proteins was chosen from the 90-protein list (Table V). Each of 11 proteins was first excluded from the training set of proteins when preference functions were extracted from the training set. We have reported the $y$ and $z$ values (see Methods) for $\alpha$-helices to show that $Q_\alpha = 74\%$ (in average) for that group of proteins is not due to excessive overestimation of helical conformation. Overall success rate in the three-state model $Q_3 = 66\%$, and correlation coefficient for helical residues $C_\alpha = 0.38$ are considerably better performance parameters than for the whole protein data set. Corresponding performance parameters for the case when $\alpha$-helix conformation is not divided into middle helix, N-terminal helix, and C-terminal helix (left branch of program flow in Figure 2) are $Q_3 = 59\%$ and $C_\alpha = 0.35$. Two parameters in Table V—*h* and *hp*—report the percentage of residues among residues with local environment higher than 1.37 nm$^2$ (*h*) that are *not* in middle-helical conformation and percentage of residues among residues having middle-helix preference higher than 1.4 (*hp*) that are *not* in middle-helical conformation.

For the second testing set of 5 membrane proteins the overall accuracy is also 66% (Table VI). None of these proteins were included in the training set of 90 soluble proteins. Overall helix prediction accuracy and correlation coefficient are 79% and 0.34, respectively for this set of proteins. Due to incompletely known structure of three proteins from that list, we have assigned all extramembrane residues in "known" structures of rhodopsin, bacteriorhodopsin, and lactose permease to the undefined conformation. This will tend to decrease the apparent accuracy. For instance, if lactose permease, whose structure is known in less details than for other 4 proteins,[28] is omitted from the membrane protein list, overall prediction accuracy $Q_3$, helix prediction accuracy $Q_\alpha$, and helix correlation coefficient $C_\alpha$ increase to 67%, 79%, and 0.40, respectively. When rhodopsin[29] is also omitted, and only 3 proteins with

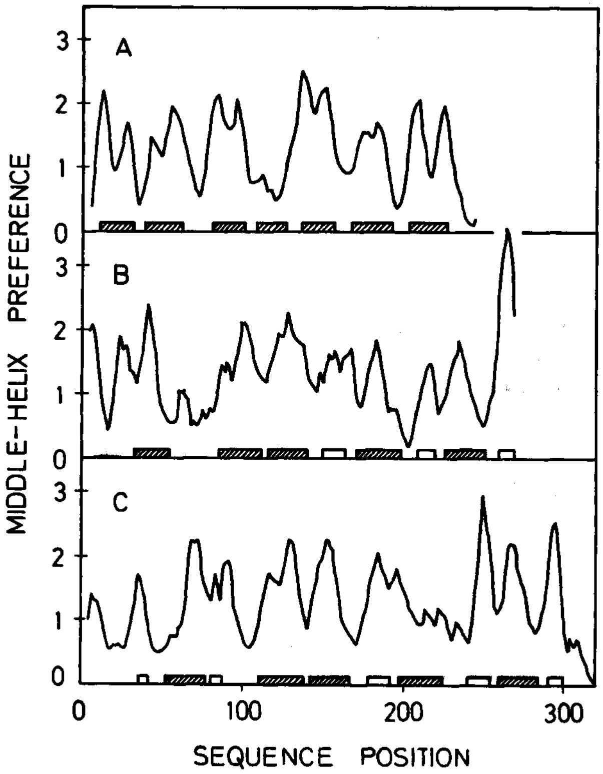

**Figure 4.** A middle helix preference profile for (A) bacteriorhodopsin,[29] (B) photosynthetic reaction center L subunit,[31] and (C) photosynthetic reaction center M subunit.[31] Experimental data for the $\alpha$-helical segments[30,31] are shown on the $x$ axis in the form of empty boxes for segments found outside membrane or shaded boxes for the transmembrane segments. The preferences are smoothed by computer (see text) and resulting points are connected by hand.

best known structure are left in the list, i.e., bacteriorhodopsin[30] and two subunits of photosynthetic reaction center,[31,32] then $h$ and $hp$ parameters from Table VI decrease to 16 and 13 respectively (on average). It is also worth noticing that residues with buried surface environment higher than 1.37 nm$^2$ and with middle helix preference higher than 1.4 are no longer just 10% of the total number of residues, but form 37% and 43% residues, respectively, of all residues in these 5 membrane proteins. Although sheet conformation of residues was also predicted by our program, quality indexes for predicting $\beta$-residues were not included in Table VI, since from all 5 proteins only several residues from the amino termini of the photosynthetic reaction center L and M subunits are known to be in the $\beta$-sheet conformation.[31,32] The success rate and correlation coefficient for predicting turn (or undefined) residues in 5 membrane proteins are 48% and 0.42, respectively. The reason for low turn prediction accuracy in the testing list of membrane proteins is quite clear. The assignment of undefined conformation to all extramembrane segments of rhodopsin, bacteriorhodopsin, and lactose permease increases the number of $N_c$ residues and decreases $Q_c$ index that contains $N_c$ in the denominator.

The GOR program,[4] with a choice of decision constants (DC): $DC_{helix} = -100$, $DC_{turn} = DC_{coil} = 0$, $DC_{beta} = 50$, appropriate for proteins having more than 50% of helical residues, results in considerably lower values of 60% and 0.27 for $Q_3$ and $C_\alpha$, respectively, for the same set of 5 membrane proteins. For photosynthetic reaction center M and L subunits only, the corresponding (averaged) values are $Q_3 = 60\%$ and $C_\alpha = 0.38$ with the same choice of decision constants. For the testing set of 11 $\alpha$-class proteins, the corresponding (averaged) values are $Q_3 = 66\%$ and $C_\alpha = 0.37$, which is comparable but not better than results presented in Table V.

## Strength and Limitations of This Prediction Method

Limitations of our method are best seen by predicting the secondary structure of specific proteins. With the choice of Rose buried surface area scale[15] predictions of $\beta$-class proteins, either soluble or membrane bound, are poor. For instance, if the proposed $\beta$-barrel structure for Omp A outer membrane protein of *Escherichia coli*[33] is assumed correct, then our program results are 44%, 17%, and 0.13, for the three-state prediction accuracy, $\beta$-sheet prediction accuracy, and sheet correlation coefficient, respectively. Only one of assumed 8 $\beta$-strands are predicted (the C-terminal one), while other 7 strands are mostly predicted as the $\alpha$-helices. With outer membrane porin of *E. coli*,[34] performance parameters are similar: 38% for the three-state prediction accuracy, 18% for the $\beta$-sheet prediction accuracy, and 0.14 for the $\beta$-sheet correlation coefficient.

For the class of all-$\alpha$ proteins, that basically have only 2 conformational states—helix and coil—predictions are generally of good accuracy. One example is hemerythrin (1hmz) predicted with an overall accuracy of 80%, $\alpha$-helix accuracy of 91%, and correlation coefficients for helix and coil of 0.49 and 0.58 respectively (Table V). From the plot of middle helix preferences along the sequence (not shown), one can see that all helices in this protein are clearly separated and identified.

The $\alpha$-helix preference profiles for two membrane proteins—bacteriorhodopsin[30] and photosynthetic reaction center[31,32]—are shown in Figure 4. Only

Table V Prediction Results[a] on Testing a Set of $\alpha$-Class Soluble Proteins

| Protein (PDB Code) | $C_\alpha$ | $Q_3$ | $Q_\alpha$ | $w_\alpha$ | $x_\alpha$ | $y_\alpha$ | $z_\alpha$ | $h$ | $hp$ |
|---|---|---|---|---|---|---|---|---|---|
| 1. Cytochrome c550 [155c] | 0.33 | 59 | 77 | 27 | 51 | 8 | 35 | 50 | 0 |
| 2. Cytochrome b562 [156b] | 0.35 | 70 | 73 | 51 | 21 | 19 | 12 | 0 | 0 |
| 3. Cytochrome c [1ccr] | 0.48 | 69 | 64 | 30 | 53 | 17 | 11 | 44 | 40 |
| 4. Hemoglobin [1eco] | 0.27 | 63 | 63 | 64 | 23 | 38 | 11 | 9 | 11 |
| 5. Cytochome p450 [2cpp] | 0.36 | 60 | 82 | 172 | 103 | 38 | 92 | 38 | 12 |
| 6. Leghemoglobin [2lh7] | 0.40 | 69 | 73 | 87 | 25 | 32 | 9 | 7 | 0 |
| 7. Hemoglobin V [2lhb] | 0.41 | 69 | 69 | 77 | 29 | 35 | 8 | 18 | 0 |
| 8. Calcium-binding parvalbumin b [3cpv] | 0.28 | 58 | 69 | 36 | 33 | 16 | 23 | 56 | 50 |
| 9. Calcium-binding protein [3icb] | 0.62 | 80 | 88 | 38 | 23 | 5 | 9 | 0 | 0 |
| 10. Cytochrome c551 [451c] | 0.37 | 63 | 49 | 20 | 35 | 21 | 6 | 0 | 0 |
| 11. Hemerythrin [1hmz] | 0.49 | 80 | 91 | 72 | 18 | 7 | 16 | 32 | 24 |
| Weighted average | 0.38 | 66 | 74 | | | | | 26 | 12 |

[a] The performance parameters for predicting helix conformation ($\alpha$) are given as the correlation coefficient $C$,[23] the prediction accuracy $Q$, number of residues associated with positive correct prediction $w$, negative correct prediction $x$, underpredictions $y$, and overprediction $z$. Overall prediction accuracy $Q_3$ in the three state model is also given. Two additional parameters, $h$ and $hp$, report the percentage of residues, which are *not* in middle helical conformation, among residues with local environment higher than 1.37 nm$^2$ ($h$) and percentage of residues, which are *not* in middle helical conformation, among residues having middle helix preference higher than 1.4 ($hp$). Class limits used were 1.050, 1.112, 1.152, 1.190, 1.226, 1.262, 1.303, 1.366 (in nm$^2$). The average is determined by weighting each protein's statistical parameter with the number of residues in the protein.

smoothed middle helix preferences are shown. In almost all cases the peaks with high middle helix preference correspond to the sequence segments that are helical and have environment $X$ higher than average for the protein data base. The valleys in the middle helix preference profile are generally the segments with high turn preference (not shown).

Six out of seven helices in bacteriorhodopsin are recognized (Figure 4A). One missed helix in bacteriorhodopsin is helix D. There are both experimental[35] and theoretical[36] indications that only helix D is able to unfold and become partially disordered and susceptible to proteases. This may happen because helix D is glycine-rich helix. Glycine presence can lead to helix underprediction by our method. Glycine has the smallest value for its average buried area (Methods). During the application of the PREF suite of programs, it decreases helical preference of all nearby amino acids in two stages. First, helical preferences for all other nearby amino acids are decreased, because glycine decreases the buried surface area environment of these amino acids, and because, as mentioned earlier, a positive correlation exists for all 20 natural amino acids found in proteins between environment value and helical or middle helical propensity. Second, in the

Table VI Prediction Results[a] on Testing Set of Membrane Proteins

| Protein [Reference] | $C_\alpha$ | $Q_3$ | $Q_\alpha$ | $w_\alpha$ | $x_\alpha$ | $y_\alpha$ | $z_\alpha$ | $h$ | $hp$ |
|---|---|---|---|---|---|---|---|---|---|
| 1. Rhodopsin (human) [29] | 0.44 | 67 | 85 | 161 | 90 | 29 | 68 | 21 | 16 |
| 2. Bacteriorhodopsin [30] | 0.36 | 69 | 79 | 130 | 47 | 34 | 37 | 15 | 9 |
| 3. Lactose permease (*E. coli*) [28] | 0.17 | 60 | 77 | 199 | 61 | 58 | 98 | 23 | 28 |
| 4. Photosynthetic reaction center L subunit [31, 32] | 0.33 | 67 | 75 | 132 | 56 | 44 | 41 | 20 | 18 |
| 5. Photosynthetic reaction center M subunit [31, 32] | 0.45 | 67 | 76 | 148 | 89 | 48 | 38 | 14 | 11 |
| Weighted average | 0.34 | 66 | 79 | | | | | 19 | 18 |

[a] The performance parameters are defined in the footnote of the Table V.

smoothing process, residue helical preference can be further reduced by including a low value caused by glycine in the average of preferences.

For the photosynthetic reaction center, subunits L and M, all helices, both in and out of the membrane, are clearly recognized by our program (Figure 4B, C). However, some transmembrane helices have wrongly predicted start or end, and some segments are overpredicted as helical segments. Several $\beta$-sheet residues and 2 antiparallel $\beta$-sheets found at the amino terminal of L and M respectively are not predicted.

## DISCUSSION

The basic result of this paper is that the introduction of relatively simple preference functions alone is enough to predict helices in membrane proteins. Importantly, preference functions are constructed using the data base of soluble protein structures and a hydrophobicity scale also derived from the analysis of such structures.[15] The location of helices can be predicted by other secondary structure prediction procedures, such as the GOR information theory procedure[4,5] or neural network procedures.[6-9,37,38] All seven transmembrane helices in bacteriorhodopsin have been predicted by neural network procedure,[37] but their location in the sequence is considerably different from the location expected on the basis of experimental[30] and theoretical[36] investigations. We have also used an improved version of neural network program,[9] trained on the class of all-$\alpha$ proteins, to predict the $\alpha$-helix conformation in membrane proteins. The performance parameters are in several cases better than those listed in Table VI. For other secondary structure prediction procedures, "trained" on water-soluble proteins, correlation between predicted structure of membrane proteins and structure measured in experiments is so low that such procedures are considered inappropriate for membrane proteins.[10] The most important deficiency of such methods is that they do not identify what physical properties of the polypeptide segments are important for helix formation.

Because of its simplicity, the Chou–Fasman prediction scheme[3] is still widely used. The success of Chou–Fasman's prediction scheme may be in part due to the fact that steric effects are predominant in their conformational parameters for $\alpha$-helix.[39] Conformational preference functions, introduced in this work, can take such effects explicitly into account through the combination of statistics and physical-chemical considerations. The protein-folding process reduces the accessible surface area by a factor of about 3–4 depending on protein molecular weight.[40] During the initial stages of folding, the formation of autonomous folding units ($\alpha$-helices) is probably the most efficient mechanism for water exclusion from polypeptide surface. It is clear that high environment (Rose's scale[15]) can be considered as a high potential for the exclusion of side-chain surfaces from the contact with water and for $\alpha$-helix formation. The data presented indicate that $\alpha$-helix formation may require steric protection offered by primary structure neighbors of residues with helix propensity. Our results add to the recently reported procedures for identifying potential folding initiation sites[41-43] since they suggest that an $\alpha$-helix can nucleate more easily in a local primary structure environment of higher initial solvent-accessible surface area that can become buried in protein interior.

That the $\alpha$-helix preference of a given amino acid can have a very different value depending on its sequence neighbors has some implications on the secondary structure prediction algorithms. A given amino acid can influence helix-forming potential either because of its inherent preference for a helix or because it influences its neighbors. For example, Figure 3A illustrates that higher environment (Rose's scale[15]) of leucine neighbors fosters $\alpha$-helix conformation of that amino acid. The sequence neighbors of leucine are then more bulky with higher propensity to become buried during folding process. Such correlation between environment of buried areas and $\alpha$-helix conformation is observed for all amino acid types (Table IV). An bulky amino acid, such as arginine, tends to help helix formation of its sequence neighbors because of its large size. However, its own preference for helix is either better or poorer than that of other amino acids depending on its environment. A secondary structure prediction algorithm that does not recognize this dual role of an amino acid is likely to perform less successfully than those that do.

To calculate preference functions and to use them in predicting profiles of secondary structure preferences, a suite of FORTRAN programs, PREF, has been created (Figure 2). Preference functions based on Gaussian curves for a frequency distribution of an amino acid over environments are not only a close fit for preference points (Figure 1 and 3), but can be easily incorporated into any secondary structure prediction scheme that uses conformational preferences. Figure 4 illustrates that most segments of transmembrane proteins, predicted by us to be in the middle helix conformation, are indeed in the $\alpha$-

helix conformation. Same segments have higher than average environment of buried areas.

Sequence-dependent preferences are better predictors of helices in the class of all-$\alpha$ proteins than the environmental property used to derive these preferences (compare parameters $h$ and $hp$ in the Table V). The same conclusion cannot be derived for membrane proteins until a larger number of such proteins of known structure can be analyzed. However, plots of middle helix preferences along the sequence of integral membrane proteins (Figure 4) give a clearer picture of where helices are located than hydrophobicity plots that can be found in the literature[13] for such proteins. Of course, hydrophobicity plots are just that—i.e., such plots can be used to locate highly hydrophobic sequence segments rather than segments that have high probability to fold into particular secondary conformation.

Before using preference functions, a particular scale of physical, chemical, or statistical parameters for 20 natural amino acids must be chosen from many proposed in the literature.[45,46] The PREF prediction depends on the choice of that scale, so that computer experiments with PREF can be used to *select* the hydrophobicity scale that gives the best secondary structure prediction. Identification of an optimal scale of physical parameters for each secondary structure conformation can suggest what features of amino acids are important during formation of that conformation.

It is indeed unlikely that one scale will be best for all protein classes, for all conformations, and for all applications.[47] In our preliminary investigations[17] we have found that Rose's scale[15] is the best predictor for localizing helices in the photosynthetic reaction center M subunit.[48,49] Since Rose's scale is among the best conserved scales for all examined $\alpha$-class protein families,[16] it is not surprising that it is also a good predictor of protein-folding pattern in such proteins (Table V). That the same scale works well for globular membrane proteins with transmembrane helices (Ref. 17 and Table VI) indicates that similar principles operate during folding of such proteins and of hydrophobic cores of $\alpha$-class soluble proteins.

For the data base of soluble globular proteins containing roughly equal amounts of $\alpha$-helix, $\beta$-sheet, and coil residues, PREF results with Rose's scale[15] are similar to that obtained with older version of the GOR algorithm.[4,27] In the case of $\alpha$-class proteins (Table V) and for globular membrane proteins with transmembrane helices (Table VI), secondary structure predictions using PREF are comparable or better than that obtained with GOR algorithm. Although the GOR algorithm takes implicitly into account all physical properties of neighboring amino acids in the sequence (in the information about amino acid type), it is conceivable that for specific protein classes the prediction accuracy may depend, in addition, on the explicit choice of an input coding scheme based on the physical properties of amino acids that are crucial for the folding process into dominant secondary conformation. Gibrat et al. have recently estimated[50] that 10% of the residues in the data base of soluble globular proteins of known structure have their conformation almost exclusively determined by the local sequence. These residues are speculated to act as seeds for the nucleation sites during the folding. The fourfold increase in the percentage of residues (from 10 to 40%) having high local buried surface environment and high middle helix preference as well in the data base of integral membrane proteins is consistent with the hypothesis that in such proteins considerably more than 10% of the residue conformations is determined by the local sequence. If true, this hypothesis would help explain accurate prediction of membrane protein structures with PREF algorithms that take into account only local sequence information. However, what seems to be the case for the data set of integral membrane proteins, as those used in this study, must be tested for other membrane protein classes.

A set of PREF algorithms leaves many possibilities to gain additional insight into protein-folding problems when using these algorithms. We did only preliminary work in exploring some of these possibilities such as taking less or more than 8 nearest neighbors in the definition of sequence environment (8 or more neighbors must be averaged for optimal results) or in using a different scale of physical-chemical properties in that definition. The scales considered very similar, such as Chothia's scale[40] of solvent-accessible surfaces,[51] Rose's scale[15] of buried surface areas, and Fauchère-Pliška's scale[26] of solution hydrophobicities, can produce completely different dependencies of preference functions on sequence hydrophobic environment (Figure 3 and unpublished results). Statistical scales of Chothia and Rose are well correlated (the correlation coefficient: 0.84), so that it is not surprising that either higher initial solvent-accessible surface area of neighboring residues, or their higher potential to bury such area, increase the preference of the central residue (irrespective of its type) for the $\alpha$-helical conformation (Ref. 44 and Table IV of this work). However, in soluble proteins, more hydrophobic sequence neighbors *do not* increase the preference of the central residue for the $\alpha$-helical conformation (Figure 3B and similar results for other amino acids that are not shown). The difference observed in the behavior of the preference function as a function of accessible surface area vs hydrophobicity is consistent with an earlier observation[52] that the local clustering pattern of accessible surface area is different from that of the hydrophobicity.

A scale different from Rose's[15] may be better in predicting sheet conformation, for which formation intermolecular forces are predominant over steric effects.[39] As a rough guide of how useful some amino acid scale will be in predicting secondary structure, one can examine how strongly preference function for that conformation depends on the sequence environment associated with that scale. For instance, the results presented in Figure 3 for leucine, when repeated for all other amino acid types, would indicate that the Fauchère-Pliška scale[26] is a good predictor of $\beta$-sheet conformation, while Rose's scale[15] is a good predictor of $\alpha$-helix conformation. Indeed, the test with the Fauchère-Pliška scale[26] showed that $\beta$-sheet conformation of $\beta$-class proteins is much better predicted with this hydrophobicity scale than with Rose's scale[15] of buried surfaces (not shown). Systematic evaluation of some 55 different sets of scales of conformational parameters with a set of PREF algorithms, trained on soluble and tested on membrane proteins, gave advantage to those parameters that are specific to protein molecules (to be published). In particular, interresidue contact energies derived by Miyazawa and Jernigan[53] (from crystal structures of globular soluble proteins), helix propagation parameters,[54] and a combination of Chou–Fasman's conformational parameters for $\alpha$-helix and $\beta$-sheet[55] are also good predictors of membrane proteins folding motifs (not shown). When predicting membrane protein secondary structure, an improvement in performance parameter $C_\alpha$ can be achieved for some scales when middle helix conformation is defined. The example is Rose's scale of buried surfaces[15] that we used in this paper. Nevertheless, that is not a general rule.

Helix prediction can be obviously improved by introducing in the algorithm the cooperativity rules, designed to eliminate lone helical residues, and helical end rules designed to locate segments where helical growth should stop.[56,57] Overall prediction in the three-state model should increase in accuracy when protein class prediction from sequence information[58–60] is performed first and decision constants introduced in the algorithm. Also, the information from the arrangement of hydrophobic residues, and of supersecondary structures, can be assessed following the work of Lim,[61] Eisenberg,[14] and others.[62–66] As is the case with other statistical methods, the choice of protein data base for the training procedure is the most important initial step. The statistics is better for larger number of proteins in the training data set of proteins, but structures not known at high enough resolution must be avoided.[67] Accordingly, better results are expected when a larger data base of high-resolution structures is used or when preference functions are trained on a specific class of proteins (unpublished results).

One of the authors (DJ) expresses his thanks to Dr. R. Venable for help in programming on NIH Apollo system and to Dr. D. G. Kneller for the permission to use his version of neural network program for the prediction of secondary structure. This work was supported by National Science Foundation Grant GM7160, Uniformed Services University Grant GM7610, and Croatian Ministry of Science Grant 1-03-171.

## REFERENCES

1. Kühlbrandt, W. (1988) *Quart. Rev. Biophys.* **21,** 429–477.
2. Gregoret, L. M. & Cohen, F. E. (1990) *J. Mol. Biol.* **211,** 959–974.
3. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13,** 211–222.
4. Garnier, J., Osguthorpe, J. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
5. Gibrat, J.-F., Garnier, J. & Robson, B. (1987) *J. Mol. Biol.* **198,** 425–443.
6. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202,** 865–884.
7. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
8. McGregor, M. J., Flores, T. P. & Steinberg, M. J. E. (1989) *Protein Eng.* **2,** 521–526.
9. Kneller, D. G., Cohen, F. E. & Langridge, R. (1990) *J. Mol. Biol.* **214,** 171–182.
10. Wallace, B. A., Cascio, M. & Mielke, L. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 9423–9427.
11. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
12. Kuhn, L. A. & Leigh, Jr., J. S. (1985) *Biochim. Biophys. Acta* **828,** 351–361.
13. Engelman, D. M., Steitz, T. A. & Goldman, A. (1986) *Ann. Rev. Biophys. Biophys. Chem.* **15,** 321–353.
14. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984) *J. Mol. Biol.* **179,** 125–142.
15. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H. & Zehfus, M. H. (1985) *Science* **229,** 834–838.
16. Kelly, L. & Holladay, L. A. (1987) *Protein Eng.* **1,** 137–140.

17. Williams, R. W. & Loughran, S. (1987) *Biophys. J.* **51**, 234a.
18. Juretić, D. (1991) *Periodicum Biologorum* **93**, 279–280.
19. Juretić, D. & Lee, B. (1989) *Biophys. J.* **55**(2/2), 354a.
20. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Rodgers, J. R., Kennard, O., Shimanouchi, T. & M. Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
21. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577–2637.
22. Rose, G. D. (1978) *Nature* **272**, 586–590.
23. Matthews, B. W. (1975) *Biochim. Biophys. Acta* **405**, 442–451.
24. Fisher, R. A. (1973) *Statistical Methods for Research Workers*, Hafner Publishing Company, New York.
25. Levitt, M. (1978) *Biochemistry* **17**, 4277–4285.
26. Fauchère, J.-L. & Pliška, V. (1983) *Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.
27. Williams, R. W., Chang, A., Juretić, D. & Loughran, S. (1987) *Biochim. Biophys. Acta* **916**, 200–204.
28. Roepe, P. D. & Kaback, H. R. (1989) *Biochemistry* **28**, 6127–6132.
29. Birge, R. P. (1990) *Biochim. Biophys. Acta* **1016**, 293–327.
30. Stern, L. J., Ahl, P. L., Mart, T., Mogi, T., Dunach, M., Berkowitz, S., Rothschild, K. J. & Khorana, H. G. (1989) *Biochemistry* **28**, 10035–10042.
31. Deisenhofer, J., Epp, O., Mikki, K., Huber, R. & Michel, H. (1985) *Nature* **318**, 618–624.
32. Michel, H., Weyer, K. A., Gruenberg, I., Dunger, I., Oesterhelt, D. & Lottspeich, F. (1986) *EMBO J.* **5**, 1149–1158.
33. Vogel, H. & Jähnig, F. (1986) *J. Mol. Biol.* **190**, 191–199.
34. Schiltz, E., Kreusch, A., Nestel, U. & Schulz, G. E. (1991) *Eur. J. Biochem.* **199**, 587–594.
35. Fimmel, S., Choli, T., Dencher, N. A., Buldt, G. & Wittmann-Liebold, B. (1989) *Biochim. Biophys. Acta* **978**, 231–240.
36. Jähnig, F. & Edholm, O. (1990) *Z. Phys. B Cond. Matter* **78**, 137–143.
37. Bohr, H., Bohr, J., Brunak, S., Cotterill, M. J., Lautrup, B., Norskov, L., Olsen, O. H. & Petersen, S. B. (1988) *FEBS Lett.* **241**, 223–228.
38. Kim, J. R. & Lee, B. (1991) private communication.
39. Charton, M. & Charton, B. I. (1983) *J. Theor. Biol.* **102**, 121–134.
40. Chothia, C. (1976) *J. Mol. Biol.* **105**, 1–14.
41. Moult, J. & Unger, R. (1991) *Biochemistry* **30**, 3816–3824.
42. Hugson, F. M., Wright, P. E. & Baldwin, R. L. (1990) *Science* **259**, 1544–1548.
43. Jeng, M. F., Englander, S. W., Elove, G. A., Wand, A. J. & Roder, H. (1990) *Biochemistry* **29**, 10433–10437.
44. Juretić, D. & Williams, R. W. (1991) *J. Math. Chem.* **8**, 229–242.
45. Kubota, Y., Nishikawa, K., Takahashi, S. & Ooi, T. (1982) *Biochim. Biophys. Acta* **701**, 242–252.
46. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. & DeLisi, C. (1987) *J. Mol. Biol.* **195**, 659–685.
47. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171**, 479–488.
48. Yeates, T. O., Komiya, H., Rees, D. C., Allen, J. P. & Feher, G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 6438–6442.
49. Allen, J. P., Feher, G., Yeates, T. O., Komiya, H. & Rees, D. C. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 6162–6166.
50. Gibrat, J.-F., Robson, B. & Garnier, J. (1991) *Biochemistry* **30**, 1578–1586.
51. Lee, B. & Richards, F. M. (1971) *J. Mol. Biol.* **55**, 379–400.
52. Lipman, D. J., Pastor, R. W. & Lee, B. (1987) *Biopolymers* **26**, 17–26.
53. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.
54. Wojcik, J., Altman, K.-H. & Scheraga, H. A. (1990) *Biopolymers* **30**, 121–134.
55. Chou, P. Y. & Fasman, G. D. (1978) *Ann. Rev. Biochem.* **47**, 251–276.
56. Richardson, J. S. & Richardson D. C. (1988) *Science* **240**, 1648–1652.
57. Presta, L. G. & Rose, G. D. (1988) *Science* **240**, 1632–1641.
58. Deleage, G. & Roux, B. (1987) *Protein Eng.* **1**, 289–294.
59. Klein, P., Kanehisa, K. & DeLisi, C. (1985) *Biochim. Biophys. Acta* **815**, 468–476.
60. Klein, P. & DeLisi, C. (1986) *Biopolymers* **25**, 1659–1672.
61. Lim, V. I. (1974) *J. Mol. Biol.* **88**, 873–894.
62. Busetta, B. & Hospital, M. (1982) *Biochim. Biophys. Acta* **701**, 111–118.
63. Ptitsyn, O. B. & Finkelstein, A. V. (1983) *Biopolymers* **22**, 15–25.
64. Taylor, W. R. & Thornton, J. M. (1984) *J. Mol. Biol.* **173**, 487–514.
65. Rees, D. C., DeAntonio, L. & Eisenberg, D. (1989) *Science* **245**, 510–513.
66. Šali, A. & Blundell, T. L. (1990) *J. Mol. Biol.* **212**, 403–428.
67. Norris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thorton, J. M. (1992) *Proteins* **12**, 345–364.
68. Croxton, F. E. & Cowden, D. J. (1948) *Applied General Statistics*, Pretince-Hall, New York.

*Received October 29, 1991*
*Accepted June 24, 1992*