Environmental features are important in determining protein secondary structure

J. RANDY MACDONALD AND W. CURTIS JOHNSON, JR.

Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331, USA

(RECEIVED January 26, 2001; FINAL REVISION March 15, 2001; ACCEPTED March 15, 2001)

Abstract

We have investigated amino acid features that determine secondary structure: (1) the solvent accessibility of each side chain, and (2) the interaction of each side chain with others one to four residues apart. Solvent accessibility is a simple model that distinguishes residue environment. The pairwise interactions represent a simple model of local side chain to side chain interactions. To test the importance of these features we developed an algorithm to separate α -helices, β -strands, and "other" structure. Single residue and pairwise probabilities were determined for 25,141 samples from proteins with <30% homology. Combining the features of solvent accessibility with pairwise probabilities allows us to distinguish the three structures after cross validation at the 82.0% level. We gain 1.4% to 2.0% accuracy by optimizing the propensities, demonstrating that probabilities do not necessarily reflect propensities. Optimization of residue exposures, weights of all probabilities, and propensities increased accuracy to 84.0%.

Keywords: Protein folding; prediction secondary structure

Supplemental material: See www.proteinscience.org.

It is generally believed that the information necessary to predict secondary structure lies in the amino acid sequence of a protein. Thus many laboratories interested in the protein-folding problem have developed algorithms to predict protein secondary structure from sequence. The traditional methods assume that the propensity of an amino acid to be in a particular secondary structure is proportional to the probability that the residue is found in that secondary structure. Probabilities can be found from the samples in the protein data bank of x-ray crystallographic structures (PDB). The Chou-Fasman method (Chou and Fasman 1978) and the Garnier-Osguthorpe-Robson (GOR) method (Garnier et al. 1978, 1994) are the two most popular algorithms to take this approach. The Chou-Fasman method uses as its features the probability that each single amino acid is in an α -helix or a β -strand. The method reveals that certain amino

acids have a higher probability to be in a given secondary structure than other amino acids, but all residues are found in every secondary structure. This simple algorithm with 40 features achieves about 57% accuracy (Williams et al. 1987). The GOR method incorporates the additional features of considering pairwise interactions with the nearest eight amino acids on either side of the target amino acid. However, although the target amino acid is identified by its α -helical, β -strand, or other secondary structure (called coil), the algorithm does not consider the secondary structure of the amino acid with which it is interacting. The GOR IV method, with its 19,260 features, achieves an accuracy of ~65% (Garnier et al. 1996).

More recently, workers have used neural networks to predict secondary structure from sequence (Qian and Sejnowski 1988; Holley and Karplus 1989; Stolorz et al. 1992; Rost and Sander 1993). Secondary structure information in the PDB is used to teach the neural network how to make the most accurate predictions. These nonlinear algorithms have achieved accuracies >70%, but the features in the PDB that lead to these predictions tend to be obscured.

The highest prediction accuracy is achieved by multiple alignment of protein sequences (Pongor and Szalay 1985;

Reprint requests to: Dr. W. Curtis Johnson, Jr., Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331, USA; e-mail: johnsowc@ucs.orst.edu; fax: (541) 737-0481.

Article and publication are at www.proteinscience.org/cgi/doi/10.1110/ps.420101.

Levin et al. 1986; Nishikawa and Ooi 1986; Sweet 1986; Zvelebil et al. 1987; Cuff and Barton 1999). This gets the most out of the PDB by assuming that homologous sequences will have the same secondary structure. The algorithms based on this principle are extremely valuable, because they lead to high prediction accuracy, but again the features in the PDB that lead to this accuracy are obscured.

Our laboratory has been interested in amino acid environment as a set of features that determines secondary structure (Zhong and Johnson 1992; Waterhous and Johnson 1994). We used various solvents to change the environment of selected peptides, and showed experimentally that almost any peptide sequence can be made to form an α -helix, a β -strand, or a random coil by the proper choice of solvent. In general, alcohols support helices, hydrophobic solvents support β -strands, and aqueous solution gives mostly random coil. Thus the secondary structure formed by a peptide depends on environment as well as sequence. It appears that to be successful, schemes that predict secondary structure should take effective solvent (environment) into account.

We also showed experimentally that the relative propensity of the common amino acids to form a helix changes with environment (Krittanaï and Johnson 2000). We measured profiles of free energy for α -helix propagation as a function of the percentage methanol in a mixed aqueous/methanol solvent system. Crossovers among the profiles demonstrate changes in the relative order of helical propensity as the solvent environment changes. Again, this is a clear experimental demonstration that effective solvent should be taken into account when predicting secondary structure.

Finally, it is clear that the rank order of amino acids for helix propagation differs markedly for different peptides. In particular, we find that the amino acids W and F are the best helix formers in 88% methanol when substituted at the X position in the peptide acetyl-(VAXAK)₃-NH₂. In contrast, these two amino acids are nearly the worst helix formers in 90% 2,2,2-trifluoroethanol when substituted at the X position in the peptide acetyl-(VAEAK)(TSXSR)(VAEAK)-NH₂. There are numerous examples of this context dependence for the helical propensity of amino acids (Myers et al. 1997; Krittanaï and Johnson 2000; J.R. Lawrence and W.C. Johnson, in prep.), so it does not appear that helical propensities for single amino acids can be used to successfully predict protein structure. This is another aspect of the effect of environment.

We investigate environment by returning to the traditional methods for predicting secondary structure from sequence. We utilize samples in the PDB with low homology to determine the probability that each amino acid is an α -helix, β -strand, or other structure as the first set of features. Because we are using samples with low structural homology, it is not fair to compare this work with multiple sequence alignment, which makes use of high homology. We

also consider pairwise interactions among amino acids as a second set of features related to environment that will give context dependence to the formation of secondary structure. In our computer experiments we separate three types of samples: known α -helices, known β -strands, and known other structures. This simplifies the pairwise interactions, because the target amino acid in an α -helix will interact only with other α -helical amino acids, and the same for β -strands and other structures. In addition, we consider the exposure of each amino acid side chain to the solvent as a third set of features related to environment. α -Helices and β -strands have differing solvent environments in proteins, and their amino acid pairs should have different patterns of exposure depending on their secondary structure. Solvent exposure can be traced back to the sequence that directs the folding of the protein, so it is not unreasonable to believe that these patterns of solvent exposure help determine secondary structure. We determine solvent exposure explicitly from the x-ray crystallographic structures in the PDB. Of course this adds the complication that the tertiary structure must be known to determine the exposures used to predict secondary structure. Conversely, it is just an explicit acknowledgment that the relationship between protein sequence and secondary structure is complex. Presumably, an algorithm could be developed to predict secondary structure from sequence that bootstraps these exposures and iterates the final secondary and tertiary structures.

We find that adding the features of exposure increases the accuracy of separating the three secondary structures, as does adding the features of pairwise interactions. Combining exposure with pairwise interactions, and optimizing these features and the propensities in the algorithm leads to 84.0% accuracy in separation.

Results and Discussion

Experiment 1 begins the main thrust of this work, which is to investigate how the environmental features of protein structure can be best exploited for prediction of protein secondary structure. This experiment was done to give a baseline from which to measure the effect of adding in the features. What we see in Table 1 is that simple single residue probabilities distinguish α -helices (H), β -strands (E), and other structure (O) at the 74.4% level or about twice as well as a random guess.

Experiment 2 shows what single residue probabilities gain by including the set of features defining side chain exposure to solvent as buried (b) or exposed (x). We optimize the exposure cutoffs for each amino acid, but see little improvement over single residue probabilities without exposures. Merely adding features does not increase accuracy. In this case we double the number of features (20 amino acids $\times$ 2 exposures per amino acid for each secondary structure) improving the accuracy from 74.4% to 75.3%.

Experiment 3 was done to investigate what context as near neighbor interaction can contribute to secondary structure prediction. Here we simplify this set of features as the interaction of amino acid pairs positioned 1 through 4 residues apart (indexed as $k = 1-4$). We optimized the weights of the pairwise parameters at each of the four positions for each secondary structure. When probabilities for single amino acids ($k = 0$) were included, they did not improve accuracy as their weighting factor was 0.00. Weighting factors for the pairs are given in Table 2 (see below). We see that $k = 1$ contributes very little to identifying α -helices, and that $k = 2-4$ contribute about equally. This goes along with our intuition. Amino acids at $k = 3$ and 4 are close to the target amino acid, and their side chains interact directly. Amino acids at $k = 2$ are on opposite sides of the helix, so when exposure is considered we expect them to be related by opposite interaction with water. Only $k = 2$ makes a significant contribution to identifying β -strands. Again, this goes along with our intuition. Amino acids at $k = 2$ are on the same side of a β -strand and interact directly. All the k 's contribute about equally to identifying other structure. This is not surprising, because other has all types of relative orientations of the amino acid pairs that are independent of k .

With pairwise interactions we have increased the number of features to 1600 for each structure (4 k positions $\times$ 400 pairs) versus only 40 features for each structure in Experiment 2. We see an increase in accuracy over our single residue baseline experiment to 78.3%.

Experiment 4 optimizes the propensities for pairwise interactions beyond simple probabilities as described in Materials and Methods. Samples that are poorly predicted are poorly predicted for a reason. Straight probabilities lose this information and overweight the correct predictions. We optimized the propensities to uncover information in the poorly predicted samples. Here our optimization procedure increases the accuracy to 79.7% or a gain of 5.3% over our single residue baseline experiment. The gain of 1.4% over Experiment 3 demonstrates that probability does not necessarily reflect propensity.

Experiment 5 was done to show how exposure information combines with pairwise information to improve the accuracy of distinguishing the three structures. We opti-

Table 1. Separating α -helices, β -strands, and other structure

Expt.	Amino Acids	Exposures	Propensities	Accuracy
1	singles	no	probabilities	74.4%
2	singles	optimized	probabilities	75.3%
3	pairs	no	probabilities	78.3%
4	pairs	no	optimized	79.7%
5	pairs	optimized	probabilities	82.0%
6	pairs	optimized	optimized	84.0%

Table 2. Optimized weighting factors that adjust the importance of $\ln P$ for each k value

k	HH	EE	OO
1	0.05	0.00	0.65
2	0.65	1.00	0.50
3	0.55	0.10	0.50
4	0.55	0.20	0.40

mized the exposure cutoff for each amino acid side chain as described in Materials and Methods. These optimized fractions for pairs that define a side chain to be exposed are given in Table 3. We see that most cutoff fractions are around 0.5. Exceptions are (1) P, which is always considered buried; (2) C, which is considered buried unless >80% of its side chain is exposed to solvent; and (3) I, K, W, and Y, which are considered exposed when only 20% of their side chain is exposed. Including optimized exposures improves the accuracy 3.7% to 82.0%. When compared with Experiment 2, where the exposures were similarly optimized, it is clear that including pairwise interactions improves accuracy by 6.7%. With pairwise probabilities there are twice as many exposure combinations as for single residue probabilities giving 6400 features for each secondary structure.

Finally we carried out Experiment 6 to show what could be gained by using optimized exposures and pairwise parameters, followed by an optimization of the resulting propensities. In the end we are able to distinguish the three structures with an accuracy of 84.0% or 9.6% better than our baseline experiment. These results show that even with

Table 3. Total areas of the amino acid sidechains and the optimized exposure cutoff for amino acid pairs

Amino acid	Area ($\text{\AA}^2$)	Exposure cutoff
A Ala	35.0	0.50
C Cys	36.0	0.80
D Asp	64.0	0.50
E Glu	83.0	0.40
F Phe	104.0	0.40
G Gly	15.0	0.30
H His	93.0	0.40
I Ile	92.0	0.20
K Lys	107.0	0.10
L Leu	85.0	0.40
M Met	98.0	0.50
N Asn	67.5	0.50
P Pro	70.0	1.00
Q Gln	86.0	0.60
R Arg	122.5	0.50
S Ser	42.0	0.50
T Thr	64.0	0.40
V Val	75.0	0.30
W Trp	157.5	0.20
Y Tyr	117.5	0.20

a very simple model of solvent exposure and local sequence context, we are able to exploit these two sets of features of protein structure to resolve the three structures with an absolute improvement of >50%, or at a level >2.5 times better than would be expected by chance alone. Clearly, proteins contain more features than amino acid identity that must also be considered when predicting secondary structure from sequence.

Table 4 shows the five amino acid pairs with the highest propensities from Experiment 6 for each secondary structure, k , and exposure. The fifth pair in each series of five has from 40% to 80% of the propensity of the pair with the highest propensity in the series. We see that the pairs for each secondary structure and exposure are similar for each of the k 's. As we might expect, most buried amino acids are hydrophobic whereas most exposed amino acids are hydrophilic and charged. The exposed–buried pairs tend to be the same as the buried–exposed pairs, but with the order reversed. One disappointing aspect of the table is that some pairs show up with a high propensity for both α -helix and β -strand. This emphasizes the difficulty in separating these two secondary structures. For instance, for both $k = 3$ and 4 LL has the highest probability for both α -helix and β -strand as a buried–buried pair. The hallmark of other structure is the amino acid G, which has a high propensity when it is buried. Some doubly exposed pairs show up in all three structures. For instance, KK has a high propensity when doubly exposed for all three structures at $k = 2$.

Materials and methods

Features

We are interested in finding the features related to protein sequence that determine secondary structure. Experimental evidence shows that the local environment exerts a profound influence on the propensity of an amino acid to be found in a particular secondary structure (Zhong and Johnson 1992; Waterhouse and Johnson 1994; Myers et al. 1997; Krittana and Johnson 2000; J.R. Lawrence and W.C. Johnson, in prep.). This is not surprising.

Table 4. The five amino acid pairs with the highest propensities for each secondary structure, k , and exposure from Experiment 6

		HH	EE	OO
$k = 1$	bb	AL, LA, AA, LL, AV	LL, LV, VL, AV, AL	GL, GA, LP, LG, PG
	bx	LK, AK, LE, VK, LR	LK, VK, VE, LE, LT	GK, GE, PE, GT, LK
	xb	KL, EL, KA, EA, AA	KV, KL, TL, KI, EV	KG, EG, SG, DG, AG
	xx	EK, KE, EE, KK, KA	IK, KI, KV, IV, IE	KK, EK, KD, KE, DK
$k = 2$	bb	AA, GL, LA, AL, GA	LL, VL, LV, LA, IL	GG, GL, LG, AG, PG
	bx	LK, LE, AK, LA, LV	LK, VY, LY, GV, VV	GK, GE, GD, GS, GT
	xb	EL, KL, DL, EA, KA	KL, KB, KI, VG, YL	KG, EG, AG, KL, EL
	xx	KK, KL, LK, EK, LE	ET, TT, KK, KT, EE	KK, KE, EK, KD, DK
$k = 3$	bb	LL, AL, LA, AA, VA	LL, LA, LG, VL, LV	GG, GL, LG, GA, GP
	bx	LK, AK, GK, LE, LY	VK, LE, LK, VE, IE	GK, GE, PK, GT, GD
	xb	KL, KA, EL, KV, IL	KL, EL, KV, KL, VV	KG, EG, AG, SG, DG
	xx	EK, KE, AK, EE, DK	KV, KK, KI, TK, VI	KK, DK, EK, KE, EE
$k = 4$	bb	LL, AA, AL, VL, LA	LL, LV, VG, LG, LA	GG, PG, GL, LG, AG
	bx	GK, PA, PK, AK, LK	VK, AK, GV, LK, LD	GE, GK, GD, GS, PK
	xb	KA, KL, AA, EL, EG	KL, KV, KA, KI, VG	KG, EG, DG, AG, SG
	xx	EK, KE, AA, EA, KD	KK, TT, TK, VV, KV	KK, KE, EK, DK, EE

α -Helices and β -strands have differing solvent environments in proteins. For instance, when every other residue is considered, α -helices have their side chains on opposite sides of the secondary structure whereas β -strands have the side chains on the same side. Thus we investigated the features of amino acids in proteins that reflect the influence of local environment. One such feature is the solvent exposure of an amino acid side chain within a protein structure. We chose to use a simple model of side chain exposure that designates residues as either solvent x or b . Another feature that reflects the local environment of an amino acid side chain is its interaction with its near neighbors. These near neighbor interactions were simplified by considering pairs of amino acid residues at sequence positions i and $i + k$ for k equal to 1–4. We considered only pairs within a contiguous sequence of residues in the same secondary structure.

Samples

To investigate the features listed above we selected protein chains with low homology from the PDB. These chains were chosen from a subset of the PDB called pdb-select (Hobohm and Sander 1994), using the following criteria: (1) All chains selected have <30% structural homology; (2) all structures were solved by X-ray crystallography; (3) all structures were solved to a resolution of $\geq 2.5\text{\AA}$ with an R-factor of ≥ 0.2 . Fourteen-hundred-sixty-four protein chains met these criteria.

The secondary structure of all residues in our sample set was determined from the PDB structure files by using the Xtlstr program (King and Johnson 1999). To minimize end effects in the secondary structures, two residues were cut from the ends of each α -helix, and one from the ends of each β -strand. The minimum length for each sample was set to five residues. This gives us 25,141 samples of α -helices, β -strands, and other structures.

Exposure

The Connolly method (Connolly 1983) was used to determine the solvent exposure of the amino acid side chains from PDB structure files. Prior to the surface calculations, hydrogen atoms were added to the PDB structure files (PDBH) using the program MOLMOL (Koradi et al. 1996). Input molecular surface files for each PDBH structure file were prepared using the program Pdb2ms. The output of the Connolly molecular surface program gives the solvent-exposed area of each atom in the PDBH file. These atoms are recombined into the side chains of each residue. The total surface area for each of the 20 amino acid side chains was computed as the average of several isolated and randomly chosen amino acids from our PDBH sample set. In each case only the surface area of side chain atoms, not including the α -carbons, is used to determine the areas for each of the 20 amino acids. The fraction of exposure for each side chain is then the exposed area of the side chain divided by its total area. Initially a solvent exposure cutoff of 0.50 of total possible side chain surface area was used to determine whether a residue is buried or solvent exposed, but later we optimized the cutoff for each side chain. Table 3 lists the total surface area of each side chain and the optimized fraction for pairs that defines it to be exposed. That is, A must have an exposure >0.50 to be exposed. Interestingly, P is always considered buried.

We realize that partitioning the side chains into two groups, buried or exposed, is a first estimate of how exposure determines secondary structure. We realize that solvent environment of the buried side chains may differ, because the inside of a protein is really a microsolvent. We also realize that we are not taking into consideration intermediates in the folding pathway, but are only considering the final structure. Finally, we understand that using

this scheme one must know the tertiary structure to predict secondary structure. However, one must start somewhere and we believe this paper demonstrates that the knowledge of solvent exposure is important to predicting secondary structure. Presumably, an algorithm could be developed to predict secondary structure from sequence that bootstraps the necessary exposures and iterates the final secondary and tertiary structures.

Secondary structure probabilities

Amino acid pairs at sequence positions i and $i + k$ can be indexed by their k value. We only considered the statistics of pairs with k values 1 through 4. In addition, the statistics of single amino acids were considered and indexed as $k = 0$. Because the equations for single amino acids and amino acid pairs differ in their indices, we treated these two possibilities separately. For single amino acids ($k = 0$) we calculated the occurrence of the 20 common amino acids, three secondary structures, and two solvent exposures giving 120 combinations. For pairs of amino acids ($k = 1$ through 4) we calculated the occurrence of 400 common amino acid pairs, three secondary structures, and pairs of solvent exposures. Because the pairs occur only in a single secondary structure for our simple system, this results in 3200 combinations for each k .

Probabilities were calculated separately for each of the three secondary structures. For single amino acids ($k = 0$), the probabilities (P) were calculated as the occurrence of amino acid i with exposure e in the given secondary structure m , $O(i,m,e,0)$, divided by the sum of these occurrences over i and e . Thus these are true probabilities that sum to 1.0 for $k = 0$ and a given secondary structure.

$$P(i,m,e,0) = O(i,m,e,0) / \sum_i \sum_e O(i,m,e,0)$$

For instance, the amino acid A (alanine) occurs 6301 times in our samples as an α -helix with ≥ 0.50 of its side chain buried, classifying it as a buried residue. In all there are 92,526 occurrences of amino acids, both buried and exposed, as an α -helix. Thus the probability that a helical amino acid will be a buried A is 6301 over 92,526, or 6.81×10^{-2} .

For pairwise interactions ($k = 1$ through 4), probabilities can be calculated for amino acid i with exposure e in secondary structure m , and amino acid j with exposure f in secondary structure n . However, here we simplified the problem by considering samples that are all α -helix, all β -strand, or all other structure. Thus both amino acids in the pair i,j will have the same secondary structure (m,m), and for each k there will only be three tables of probabilities corresponding to the three secondary structures. These probabilities for each k were calculated as the occurrence of pair i,j with exposures e,f in secondary structure m,m , $O(i,j,m,m,e,f,k)$, divided by the sum of these occurrences over all pairs and exposures. Again, these are true probabilities that sum to 1.0 for a given k and secondary structure.

$$P(i,j,m,m,e,f,k) = O(i,j,m,m,e,f,k) / \sum_{i,j} \sum_{e,f} O(i,j,m,m,e,f,k)$$

For instance, amino acid pair AE occurs 466 times in our samples as an α -helix at $k = 2$ with A buried and E exposed. In all there are 76,294 occurrences as an α -helix of all 400 amino acid pairs with their side chains in the four possibilities of exposure. Thus the probability that a helical amino acid pair will be an AE (buried – exposed) is 466 over 76,294, or 6.11×10^{-3} .

Many pairwise combinations have few or no occurrences. For instance, for the helical occurrences at $k = 2$ we have 960 entries of the 1600 that are $< 5\%$ of the largest entry. We consider this fortunate. First, it means that we need fewer samples than might be

expected to generate reliable probability tables. Second, it means there is the potential for clear separation among the amino acid pairs that code for secondary structure. For instance, if for $k = 2$ the pair AE with A buried and E exposed occurred only in α -helices, leaving no occurrences for this pair in the corresponding tables for β -strand and other, then we would know with certainty that AE (buried – exposed) means α -helix. If there were this type of clear separation for all pairs, then the prediction of secondary structure from sequence would be solved. In practice we see fair separation, but not perfect separation. Our largest number of helical occurrences at $k = 4$ is 749 for the pair LL with both chains buried. The corresponding number of occurrences is 95 for β -strand and 114 for other. Clearly there is not much difference between three occurrences and no occurrences. We generate nearly the same probability for three occurrences and no occurrences while not affecting the probability when there is a large number of occurrences by adding a floor value of 10^{-4} to all probabilities in every table. For instance, the floor value changes the probability that a helical pair will be AE (buried – exposed) to 6.21×10^{-3} .

Scoring the sequences

We set up an algorithm to score each sample. The propensities of a single amino acid or amino acid pair for a particular secondary structure is assumed to be proportional to the probability that the single or pair is found in that structure. Rather than multiplying probabilities, the scoring algorithm uses the natural log of the probabilities so they can be added. For a given secondary structure, the scores for a given k value will be the sum of the logs of the probabilities of that secondary structure for all the singles or pairs in the sample. The score is normalized for the length (l) of the sample by dividing by the length. Expressing this idea as an equation using the indices in the equations for probabilities, we have for $k = 0$

$$s_o(m) = \left[\sum_{i=1}^l \ln P(i,m,e,0) \right] / l$$

and for $k = 1$ through 4

$$s_k(m) = \sum_{k=1}^4 \left\{ w(m,k) \left[\sum_{i=1}^{l-k} \ln P(i,i+k,m,m,e,f,k) \right] / (l-k) \right\} / \sum_{k=1}^4 w(m,k)$$

The factors $w(m,k)$ are an adjustable weight for each $\ln P(i,j,m,m,e,f,k)$. For each sample, scores are calculated for each possible secondary structure, and the predicted structure is the one with the highest score.

Optimization

For each of the 20 amino acids we allowed the solvent exposure cutoff to vary over the range of 0.0 to 1.1, and used feedback from the scoring algorithm to find the combination of exposure cutoffs that gave the maximum number of correctly scored samples. The optimized exposure cut offs are given in Table 3. Optimization of the local sequence context was carried out by weighting the probabilities depending on their k -value and secondary structure. Probabilities for the single amino acids ($k = 0$) did not improve the scores when included with the pairwise probabilities. That is, the weighting factor for $k = 0$ was identically zero. The combination of twelve weights $w(m,k)$ for the pairwise probabilities that gave the maximum number of correctly scored samples was found by

varying the individual weights between 0–2.0. The optimized weights are given in Table 2. Both the exposure cutoffs and k-value weights were systematically varied until a combination that gave the maximum number of correctly scored secondary structure files was achieved.

Optimization of the propensities was also carried out. We chose a subset of the samples that were most poorly predicted, which are a fraction (f) of the original set [with probabilities P(o)] that are ordered by their scores. Probabilities P(f) are calculated for this sample subset in the same way we calculated them for the original set. These probabilities will be different and will improve the predictions within the subset. We then abandon the assumption that propensities are equivalent to probabilities, and calculated new propensities (Q) by modifying $\ln P(o)$ with a fraction g of $\ln P(f)$ so that

$$\ln Q = (\ln P(o) + g \ln P(f)) / (1 + g)$$

We optimized the parameters f and g by systematically varying their values until we arrived at propensities that gave the highest number of correctly scored samples. The resulting values were $f = 0.29$ and $g = 0.63$. As an example of how the propensities are optimized, let us consider the amino acid pair AE (buried – exposed) as an α -helix at $k = 2$, which has a probability of 6.21×10^{-3} as calculated with the floor value in the example above. We order all the samples according to their scores for their *known* secondary structures. We choose the fraction $f = 0.29$ of these samples that comprise the lowest scores. This subset of samples will include all the losers and winners that were closest to losing. We now recalculate the probabilities for this subset, and find that the probability that pair AE (buried – exposed) at $k = 2$ will be in an α -helix is 1.14×10^{-3} .

The new propensity for this particular pair will lie between the probability for the entire set of samples (6.21×10^{-3}), and the probability for the subset (1.14×10^{-3}). We add a fraction $g = 0.63$ of the log of the probability for the subset to the log of the probability for the entire sample set, and normalize by dividing by 1.63. The resulting propensity for pair AE (buried – exposed) at $k = 2$ to be in an α -helix is then 3.23×10^{-3} .

Validation

The prediction accuracy presented here for the various computer experiments are the result of cross validation, where 20 samples at a time are removed from the sample set and predicted using the remaining sample set to compute the probabilities.

Software

Unless otherwise noted, all software was developed by ourselves on personal computers running the Linux operating system (RedHat) along with the associated software tools. Computer programs and subroutines used were written in awk, in fortran, as shell scripts, or as a combination of all three languages.

Supplemental material

Tables of the logs of the optimized propensities that correspond to Experiments 4 and 6 are included as an electronic appendix. These tables were generated using all the samples, but the results in Table 1 were cross validated, which means that the sample being considered was not included in the samples that generated the tables of optimized propensities used to calculate the score for that sample.

Acknowledgments

We thank Dr. P. Shing Ho and Dr. P. Andrew Karplus of Oregon State University for many helpful conversations. We thank Dr.

Beth Basham from Dr. Ho's laboratory, who wrote the programs for converting the Molecular Surface Area output into solvent accessible surface area of all amino acids in the pdb structure files. We thank Dr. Reto Koradi, the principal author of the Molmol program that was used to calculate the hydrogen atom positions in the pdb structure files, and Dr. Michael Connolly, the author of the Molecular Surface Area program that was used to calculate the solvent accessible surface areas, for putting their programs in the public domain. This work was supported by PHS grant GM-21479 from the National Institutes of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Chou, P.Y. and Fasman, G.D. 1978. Empirical predictions of protein conformation. *Ann. Rev. Biochem.* **47**: 251–276.
- Connolly, M.L. 1983. Analytical molecular surface calculation. *J. Appl. Cryst.* **16**: 548–558.
- Cuff, J.A. and Barton, G.J. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **34**: 508–519.
- Garnier, J., Osguthorpe, D.J., and Robson, B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**: 97–120.
- Garnier, J., Levin, J.M., Gibrat, J.F., and Biou, V. 1994. Secondary structure prediction and protein design. *Biochem. Soc. Symp.* **57**: 11–24.
- Garnier, J., Gibrat, J.F., and Robson, B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Meth. Enzymol.* **266**: 540–553.
- Hobohm, U. and Sander, C. 1994. Enlarged representative set of protein structures. *Protein Sci.* **3**: 522–524.
- Holley, L.H. and Karplus, M. 1989. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci.* **86**: 152–156.
- King, S.M. and Johnson, W.C. 1999. Assigning secondary structure from protein coordinate data. *Proteins Struct. Funct. Genet.* **35**: 313–320.
- Koradi, R., Billeter, M., and Wüthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graphics* **14**: 51–55.
- Krittana, C. and Johnson, W.C., Jr. 2000. The relative order of helical propensity of amino acids changes with solvent environment. *Proteins Struct. Funct. Genet.* **39**: 132–141.
- Levin, J.M., Robson, B., and Garnier, J. 1986. An algorithm for secondary structure determination in proteins based on sequence simulation. *FASEB Lett.* **205**: 303–308.
- Myers, J.K., Pace, C.N., and Scholtz, J.M. 1997. A direct comparison of helix propensity in proteins and peptides. *Proc. Natl. Acad. Sci.* **94**: 2833–2837.
- Nishikawa, K. and Ooi, T. 1986. Amino acid sequence homolog applied to the prediction of protein secondary structure, and joint prediction with existing methods. *Biochim. Biophys. Acta* **871**: 45–54.
- Pongor, S. and Szalay, A.A. 1985. Prediction of homology and divergence in the secondary structure of polypeptides. *Proc. Natl. Acad. Sci.* **82**: 366–370.
- Qian, N. and Sejnowski, T.J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**: 865–884.
- Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**: 584–599.
- Stolorz, P., Lapedes, A., and Xia, Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* **225**: 363–377.
- Sweet, R.M. 1986. Evolutionary similarity among peptide segments is a basis for prediction of protein folding. *Biopolymers* **25**: 1565–1577.
- Waterhous, D.V. and Johnson, W.C., Jr. 1994. Importance of environment in determining secondary structure in proteins. *Biochemistry* **33**: 2121–2128.
- Williams, R.W., Chang, A., Juretic, D., and Loughran, S. 1987. Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta* **916**: 200–204.
- Zhong, L. and Johnson, W.C., Jr. 1992. Environment affects amino acid preference for secondary structure. *Proc. Natl. Acad. Sci.* **89**: 4462–4465.
- Zvelebil, M.J., Barton, G.J., Taylor, W.R., and Sternberg, M.J.E. 1987. Prediction of protein secondary structure and active sites using alignment of homologous sequence. *J. Mol. Biol.* **194**: 957–961.