Protein Engineering vol.3 no.8 pp.659–665, 1990

Predicting surface exposure of amino acids from protein sequence

Stephen R.Holbrook[1,2], Steven M.Muskal[3] and Sung-Hou Kim[1,3]

[1]Chemical Biodynamics Division, Lawrence Berkeley Laboratory and [3]Department of Chemistry, University of California, Berkeley, CA 94720, USA

[2]To whom correspondence should be addressed

The amino acid residues on a protein surface play a key role in interaction with other molecules, determine many physical properties, and constrain the structure of the folded protein. A database of monomeric protein crystal structures was used to teach computer-simulated neural networks rules for predicting surface exposure from local sequence. These trained networks are able to correctly predict surface exposure for 72% of residues in a testing set using a binary model (buried/exposed) and for 54% of residues using a ternary model (buried/intermediate/exposed). In the ternary model, only 11% of the exposed residues are predicted as buried and only 5% of the buried residues are predicted as exposed. Also, since the networks are able to predict exposure with a quantitative confidence estimate, it is possible to assign exposure for over half of the residues in a binary model with >80% accuracy. Even more accurate predictions are obtained by making a consensus prediction of exposure for a homologous family. The effect of the local environment of an amino acid on its accessibility, though smaller than expected, is significant and accounts for the higher success rate of prediction than obtained with previously used criteria. In the absence of a three-dimensional structure, the ability to predict surface accessibility of amino acids directly from the sequence is a valuable tool in choosing sites of chemical modification or specific mutations and in studies of molecular interaction.

Key words: hydrophobicity/neural network/buried–exposed amino acids/protein surface/solvent accessibility

Introduction

The concept of solvent accessibility as a measure of surface exposure of protein atoms and residues was pioneered by Lee and Richards (1971) and applied by others (Alden and Kim, 1979; Connolly, 1983; Kabsch and Sander, 1983; Richmond, 1984) in formulating hypotheses of antigenic determinants (i.e. Hopp and Woods, 1981), chemical reactivity (Holbrook and Kim, 1983), subunit binding (Argos, 1988), protein folding (Ponder and Richards, 1987), molecular docking and site-directed mutagenesis. These applications of the concept of surface accessibility have generally required explicit knowledge of the experimentally determined three-dimensional structure of the protein of interest.

While early studies showed an approximately equal distribution of polar and non-polar atoms on the protein surface (Lee and Richards, 1971), later analyses of the distribution of amino acid residues have indicated the expected preferences of residues with charged side chains for the molecular surface and residues with large aliphatic side chains for the molecular interior (Shrake and Rupley, 1973). Subsequent studies have concentrated on explicitly defining and quantitating these preferences in order to relate amino acid sequence to surface exposure.

To date, attempts to predict surface exposure of protein residues from amino acid sequence have focused on deriving hydrophobicity scales [see Eisenberg (1984) for a compilation of some of these scales]. Some attempt to incorporate information from the local environment has been made by averaging hydrophobicities over a sliding window (for example, see Hopp and Woods, 1981). These scales, however, differ not only in relative magnitudes of amino acid hydrophobicity, but also in their order, especially for small and uncharged polar residues.

In order to predict surface exposure of protein residues, it is first necessary to define categories for the buried and exposed residues. Clearly, a complete description would require calculation of the exact accessible surface area presented by each atom of the residue. A more reasonable expectation would be an estimate of the total area exposed by each residue. Janin (1979) has given a binary definition of buried versus accessible residues as those of less than or greater than 20 $Å^2$ exposure respectively. Lawrence *et al.* (1987) have shown that a binary description extracts only about half the available information from the observed distributions of amino acid accessibilities. More recent definitions (Rose *et al.*, 1985) use the fractional exposure of residues in folded proteins compared with a standard, fully exposed state such as found in extended tripeptides.

Computer-simulated neural networks have recently been applied to a variety of problems in biological structure and function including identification of splice sites in mRNA and translational initiation sites in DNA (Stormo *et al.*, 1982; Nakata *et al.*, 1985), protein secondary structure prediction (Qian and Sejnowski, 1988; Holley and Karplus, 1989), protein beta-turn classification (McGregor *et al.*, 1989) and the predilection for disulfide-bond formation by cysteines (Muskal *et al.*, 1990). We have applied this computational technique to extract information about surface accessibility of protein residues from a database of crystallographically determined, high-resolution protein structures. Neural networks trained on this database can then be used to predict the accessibility of residues of proteins of known sequence, but unknown tertiary structure. Neural networks have the advantages of being able to incorporate both positive and negative information, not needing a preconceived model, and being able to incorporate higher-order correlations in their patterns. Such advantages make neural networks a convenient and powerful tool in the study of biological structure and function in general and for the prediction of surface exposure of protein residues in particular.

Methods

Database

The atomic coordinates of the proteins used in this study were from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

Twenty non-homologous crystallographically refined proteins of higher than 2.0 Å resolution containing a total of 3581 amino acid residues were selected for the neural network training set and five non-homologous proteins of equally high resolution totalling 963 residues were used for testing the trained networks. We considered it important to limit our database to structures known to high resolution so that the surface side chains are ordered and their areas well defined. The database was restricted to monomeric, single-domain, globular proteins since it has been shown that the residues on the surface of protein subunits or domains have a different amino acid distribution than residues exposed to solvent in monomeric proteins (Argos, 1988). Finally, experiments with permuted versions of the training and testing sets, have shown that this testing set is representative. The proteins included in the testing and training databases are listed in Table I.

Solvent accessibilities were calculated with the DSSP program of Kabsch and Sander (1983) and compared with standard, fully exposed values given by Rose *et al.* (1985) to give fractional accessibilities. These fractional accessibilities were then grouped and used as data to train and test the neural networks. Two definitions of surface accessible residues were used: (i) a binary model in which buried residues are defined as those with <20% of the standard state exposure and accessible residues as those >20% fully exposed; and (ii) a ternary model in which a residue is either fully buried (0–5% exposure), intermediate (5–40%) exposure, or fully accessible (>40% exposure). The choice of a 20% cutoff for buried residues in the binary model corresponds to a natural break in the observed distribution of residue accessibility in proteins given by Rose *et al.* (1985). The limits of 5% for fully buried and 40% for fully exposed bracket the mean fractional area lost on folding of the most buried and most accessible (excluding lysine) amino acids in proteins (Rose *et al.*, 1985). Miller *et al.* (1987) have shown that surface–interior partition coefficients and transfer free energies calculated from them are not sensitive to the precise definition of amino acid burial/exposure. Amino acid distributions of our training and testing sets according to both these models are shown in Table II.

Table I.

PDB code	Protein	Residues	Resolution	*R* factor
(A) Neural network protein training set				
1BP2	Pancreatic phospholipase A_2	123	1.7	0.17
1CPV	Parvalbumin B	108	1.85	0.11
1CTF	L7/L12 ribosomal protein	68	1.7	0.17
1GCR	γ-Crystallin	174	1.6	0.23
1LZ1	Lysozyme	130	1.5	0.18
1MBD	Myoglobin	153	1.4	–
1PCY	Plastocyanin	99	1.6	0.17
1RN3	Ribonuclease A	124	1.45	0.24
1TPP	β-Trypsin	223	1.4	0.19
2ACT	Actinidin	218	1.7	0.17
2ALP	α-Lytic protease	198	1.7	0.13
2APR	Acid proteinase	325	1.8	0.14
2SGA	Proteinase A	181	1.5	0.13
3DFR	Dihydrofolate reductase	162	1.7	0.15
3TLN	Thermolysin	316	1.6	0.21
4FXN	Flavodoxin	138	1.8	0.20
451C	Cytochrome C 551	82	1.6	0.19
5CPA	Carboxypeptidase	307	1.54	–
9PAP	Papain	212	1.65	0.16
(B) Neural network protein testing set				
1NXB	Neurotoxin B	62	1.38	0.24
1UBQ	Ubiquitin	76	1.8	0.18
2CPP	Cytochrome P450	405	1.63	0.19
2PRK	Proteinase K	279	1.5	0.17
2SNS	Staphylococcal nuclease	141	1.5	–

Neural networks

The neural networks used in this study were of the feedforward type with either zero (perceptron) or one hidden layer and weights adjusted by backpropagation as described in the preceding paper (Muskal *et al.*, 1990). The protein sequences were presented to

Table II. Amino acid distribution in the training–testing database

(A) Bindary distribution

Amino acid residue	Class I: 0–20% exposure	Class II: 20–100% exposure	% Exposed residues
Phe	132	25	15.9
Ile	205	44	17.6
Cys	91	24	20.9
Leu	239	67	21.9
Val	233	82	26.0
Trp	56	14	20.0
Met	49	14	22.2
Ala	218	166	43.2
His	46	45	49.5
Tyr	101	108	51.7
Gly	202	252	55.5
Thr	116	178	60.5
Ser	136	233	63.1
Pro	55	112	67.1
Asp	67	172	72.0
Glu	49	159	76.4
Asn	61	184	75.1
Gln	48	146	75.3
Arg	42	143	77.3
Lys	29	201	87.4

(B) Ternary distribution

Amino acid residue	Class I: 0–5% exposure	Class II: 5–40% exposure	Class III: 40–100% exposure
Phe	80	69	8
Ile	142	96	11
Cys	58	51	6
Leu	158	128	20
Val	161	122	32
Trp	30	36	4
Met	38	20	5
Ala	130	160	94
His	19	52	20
Tyr	38	123	48
Gly	136	159	159
Thr	62	130	102
Ser	83	106	180
Pro	27	66	74
Asp	29	96	114
Glu	19	75	114
Asn	30	82	133
Gln	19	69	106
Arg	5	102	78
Lys	5	94	131

the neural networks as windows, or subsequences, of 1−13 residues centered around and usually including the amino acid of interest, which slide along the entire sequence. For experiments involving only the flanking residues, the central residue was omitted from the window. The total number of windows or patterns for a particular protein is therefore equal to the number of residues in the protein. When the window extends beyond the N- or C-terminus, a special null indicator is given for each overlapping residue. Each amino acid in the window was represented by activating one of 21 input nodes, one for each of the 20 possible amino acids and one node for residues exceeding the N- or C-terminus. The input nodes contained a zero except for the node corresponding to the amino acid in the sequence, which contained a one. While the patterns of these subsequences formed the input to the networks, the output consisted of either two or three nodes, corresponding to either a binary (buried/exposed) or ternary (buried/intermediate/exposed) definition of accessibility.

Results

Perceptron networks, with no hidden layers, have been trained and tested for both binary and ternary models of surface exposure. The windowsizes (number of flanking residues and the central amino acid) of the patterns presented to the perceptrons were varied so as to test the effect of local sequence on surface exposure. Because of computational limitations, only the optimal windowsizes were then tested in networks containing hidden layers in order to extract any higher order correlations.

Binary models

Windowsize was varied between 1 (no neighbors) and 13 (six amino acids on either side of the central) residues for both training and testing networks containing two outputs, one for buried (<20% exposure) and one for exposed residues (>20% exposure). Table III shows the results of these experiments. The correct overall prediction for the training set is seen to reach a maximum of ~74% at windowsize 11 with a correlation coefficient of 0.48. The highest percentage of correct prediction, 72%, and correlation coefficient, 0.44, for the testing set was obtained with a windowsize of nine residues. While this is highly accurate when compared with the 52% expected from a residue-independent prediction based on the overall distribution frequency alone, it is only a 2% increase over the 70% obtained with networks trained on patterns of only single amino acids (windowsize 1). To investigate the significance of this difference and the influence of flanking residues on exposure or burial of the central residue we trained a network using examples consisting of only the flanking residues and excluding the central residue which we were trying to predict. Using four flanking residues on each side of the residue of interest (windowsize 8), we were able to predict exposure of the central residue in 55.3% of the cases with a correlation coefficient of 0.10 for both the buried and exposed nodes. This increase in prediction accuracy and correlation coefficient over that expected for a random prediction indicates that the sequence of the flanking residues has a small, but significant effect on exposure of the central residue.

For a windowsize of one, corresponding to a single amino acid, the weights determining the contribution of each amino acid to activation of the buried or exposed output are essentially a measure of hydrophobicity/hydrophilicity. They are listed in Table IV and compared with other hydrophobicity scales. Not surprisingly, these are similar to the fractional buried percentages

Table III. Perceptron prediction of solvent accessibility (binary model)

Window size	% Correctly predicted buried	% Correctly predicted exposed	% Correctly predicted overall	Correlation coefficient
Training set				
1	63.3	74.4	69.1	0 38
3	69.9	70.3	70.1	0.40
5	72.6	69.6	71.0	0.42
7	73.7	70.3	71.9	0.44
9	75.0	70.3	72.5	0.45
11	75.4	72.6	73.9	0.48
13	73.6	73.3	73.4	0.47
Testing set				
1	62.0	78.0	70.0	0.40
3	65.5	72.4	69.5	0.39
5	69.0	72.6	70.8	0.42
7	68.8	74.7	71.8	0.44
9	67.4	76.6	72.0	0.44
11	67.4	76.1	71.8	0.44
13	66.1	75.3	70.7	0.42

Table IV. Hydrophobicity scales

Amino acid[a]	Neural net[b] window size 1	Eisenberg[c] consensus	Janin[d] free energy	Rose[e] $A^o - <A>/A^o$
Phe	60	61	50	88
Ile	58	73	70	88
Cys	51	04	90	91
Leu	45	53	50	85
Val	42	54	60	86
Trp	41	37	30	85
Met	14	26	40	85
Ala	6	25	30	74
His	3	−40	−10	78
Tyr	−13	2	−40	76
Gly	−16	16	30	72
Thr	−29	−18	−20	70
Ser	−33	−26	−10	66
Pro	−33	−7	−30	64
Asp	−45	−72	−60	62
Glu	−47	−62	−70	62
Asn	−56	−64	−50	63
Gln	−56	−69	−70	62
Arg	−64	−180	−140	64
Lys	−79	−110	−180	52
Correlation with neural net weights	−	0.75	0.82	0.94

[a]Amino acids may be classified as hydrophobic (Phe−Trp), neutral or ambiphilic (Met−Thr) and hydrophilic (Ser−Lys) for the neural network weights.
[b]Normalized differences between weights to buried and exposed nodes for perceptron network using a binary model with windowsize 1.
[c]From Eisenberg (1984). The ΔG values have been muliplied by 100. In comparison to the neural network scale the greatest differences are for: Cys, which may reflect mixture of S−S and S−H bound cysteines in the database; His, which may also have different populations depending on ionization state; and Pro which may serve special functions within proteins.
[d]From Janin (1979). These ΔG values have been muliplied by 100. Again Cys has an anomalous value. In this scale it tends to have a very high free energy for burial.
[e]From Rose *et al.* (1985). These numbers are the mean fractional area loss of the amino acid surface area on protein folding times 100.

of Rose *et al.* (1985), since these are also based on surface accessibility in protein structures.

As a control experiment, accessibilities were predicted based on hydrophobicity values of either a single residue or an average hydrophobicity of consecutive residues. When the Eisenberg (1984) consensus, and Janin (1979) hydrophobicity scales were applied for a single residue, the exposure state was correctly predicted for 67 and 68% of the residues in the testing set respectively, significantly lower than the accuracy of the neural network predictions (72%). The scale of Rose *et al.* (1985), which is based on fractional residue accessibility in proteins, as were the neural networks, correctly predicted the exposure state of 70% of the residues. Using an average hydrophobicity over a run of residues actually decreased the accuracy of prediction, i.e. using the scale of Rose *et al.* (1985) averaged over a window nine residues wide predicted only 64% of the residue exposures correctly. This implies that the commonly used technique of averaging hydrophobicities over a sliding window is not appropriate for obtaining the most accurate prediction of residue surface exposure.

Differences between weights linking input nodes to the buried and exposed output nodes obtained for a windowsize 9 perceptron network are displayed for representative amino acids in Figure 1. In this figure, valine, tryptophan and methionine are examples of commonly buried (nonpolar) residues and glycine, proline and glutamic acid represent neutral and polar (exposed) residues. The influence of these residues in flanking positions is shown in addition to their preference as the central residue.

Ternary models

In experiments involving three-state prediction (buried, partially exposed and fully exposed), we varied windowsize from one to nine residues, at which point prediction of the testing set began to decrease. Table V gives the results of these experiments for both the training and testing datasets. These may be compared with the 34% accuracy expected from a 'random' prediction based on distribution frequency alone and disregarding residue type. For both datasets, the fully buried and exposed residues are predicted with greater accuracy than the partially exposed residues. As in the experiments with a binary representation, the exposed residues in the testing set are consistently predicted 10% more accurately than the buried. The overall peak in prediction with the ternary model occurs for the testing set at windowsize 7 after which a decline occurs. The weights for the links connecting the central amino acid input nodes with the buried, intermediate and exposed output nodes are compared in Figure 2. No weights due to flanking residues are shown.

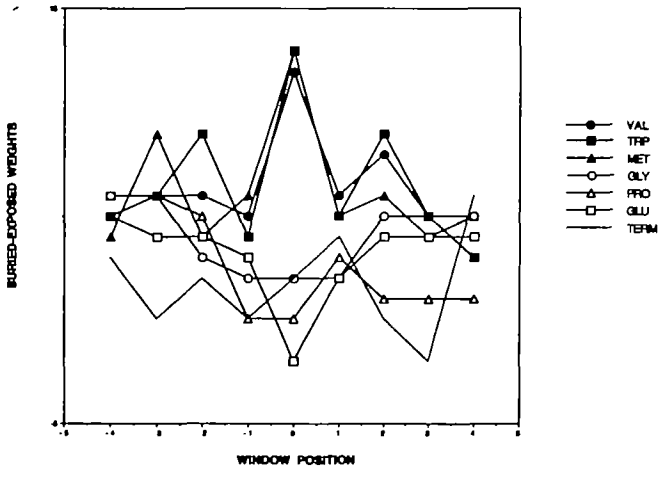

Fig. 1. A graphical representation of neural network weights from the binary distribution model of amino acid exposure (buried/exposed). Differences between weights to the buried and exposed output nodes are plotted versus window position. The amino acid position corresponds to the window of flanking residues (window size 9) from −4 to +4 (N- to C-terminal direction) around the central or zero position. Representative hydrophobic, neutral and hydrophilic residue weights are shown along with weights from the amino acid terminus. Weights for residues not shown in this figure generally follow the same trends as those presented. Prominent peaks are seen for tryptophan, valine and methionine in flanking positions two or three residues before or after the central amino acid. These may be due to their presence in alpha helices and beta strands. Glycine and proline show a broad distribution of weights favorable to exposure of the central residue, while glutamic acid has a more pronounced central peak, but little contribution when present in flanking sequences. The chain terminus in window positions −1 to −4 flanks the N-terminal residue, while in positions 1−4 it flanks the C-terminal residue.

Experiments with networks containing a hidden layer of computational nodes between the input and output layers resulted in an improvement in prediction for windowsize 7 and three output states. The maximal improvement was observed when using 10 hidden nodes, which predicted the testing set with 54.2% overall accuracy, compared with the best prediction of 52.0% with a perceptron network (Table V).

Using this three-state network with hidden nodes, a residue which is predicted to be fully exposed is actually found to be fully or partially exposed >89% (307/344) of the time, while a residue predicted to be buried is found fully or partially buried in 95% (235/248) of the cases. The difference in prediction percentage for buried and exposed is in large part due to overprediction of the fully exposed state and under prediction of the fully buried state by the network. If only fully exposed or fully buried residues are considered (cases observed or predicted to be partially exposed are ignored) we are able to predict the correct states for 87% of the residues. As shown in Table VI, the hydrophobic residues Phe, Ile, Cys, Leu, Val and Trp are predicted with very high accuracy (86−100%), as are the hydrophilic residues Lys, Arg, Gln, Asn, Asp, Pro and Ser (75−100%). The ambiphilic residues Gly and Thr are, as expected, predicted with less accuracy (68 and 60% respectively), but the amibiphilic residues Met, Ala and His are predicted with 90−100% accuracy. Even the hydrophobic residue Val is correctly predicted to be exposed in one case and the hydrophilic residue Pro is predicted correctly to be buried in one case. Clearly, including the flanking amino acid sequence in predicting residue exposure is a strength of our approach when compared with methods which assign a specific hydrophobicity index to each residue type regardless of its neighbors.

Table V. Perceptron prediction of solvent accessibility (ternary model)

Windowsize	% Correct fully buried	% Correct intermediate	% Correct fully exposed	% Correct overall
Training set				
1	50.4	45.2	53.1	49.1
3	52.4	45.5	61.3	52.4
5	54.7	49.3	59.5	54.1
7	57.2	51.9	59.7	55.9
9	61.0	52.4	60.8	57.5
Testing set				
1	49.4	44.8	57.9	50.2
3	48.1	40.8	67.6	51.1
5	50.2	40.6	62.4	50.1
7	51.5	44.8	62.1	52.0
9	50.9	41.1	60.3	49.8

Confidence of predictions

An advantage of neural network analysis is that a prediction of surface exposure is based on quantitative activity values at each of the output nodes. Therefore a confidence level may be assigned to each prediction based on the strength of the output activities. While the accuracy of prediction increases with the minimum activity accepted, a corresponding decrease is seen in the percent of the total residues whose accessibility is predicted. For example, using the binary model of accessibility, while 100% of tested residues are predicted with an accuracy of 72%, over half of the residues with the strongest activities are predicted with >80% accuracy.

Prediction refinement

The overall number of buried residues in a globular protein has been shown to approximately obey the equation:

$$N_b^{1/3} = N_t^{1/3} - b \quad (1)$$

where N_b is the total number of buried residues in a protein and N_t the total number of amino acids. The parameter b has been determined by Miller *et al.* (1987) to equal 2.0 for a definition of buried residues as those with <5% fractional accessibility. This equation can be used to adjust our predictions to an overall expected value for a given protein and thereby improve our prediction accuracy. For example, proteinase K (2PRK), a member of our test set, is predicted by this formula to have 93 buried residues, while our ternary neural network predicts only 63 as buried (the actual number of buried residues is 106). If an empirical bias term of 0.1 is added to the network activities for the buried node the number predicted as buried increases to 94, the prediction accuracy for this node increases from 42 to 56% and the overall prediction for this protein from 50.9 to 51.6%. Because of the relatively large error in this formula (~15% for our testing set) it is only reasonable to apply to cases where the number of buried residues predicted by the neural network differs from that calculated by equation (1) by a large percentage. Besides 2PRK, only 1NXB and 1UBQ of our test set show large differences. Application of the formula to 1NXB improves the accuracy of prediction from 54.8 to 56.5%. The overall accuracy of exposure prediction for 1UBQ actually decreases from 65.8 to 60.5% when an attempt is made to correct for overprediction of the buried node; however, the number of fully exposed residues predicted as buried decreases by 1/3 due to the improved distribution. Thus, it is clear that for cases where N_b is significantly different from the number predicted by the network and the protein is globular, application of an empirical bias term is appropriate.

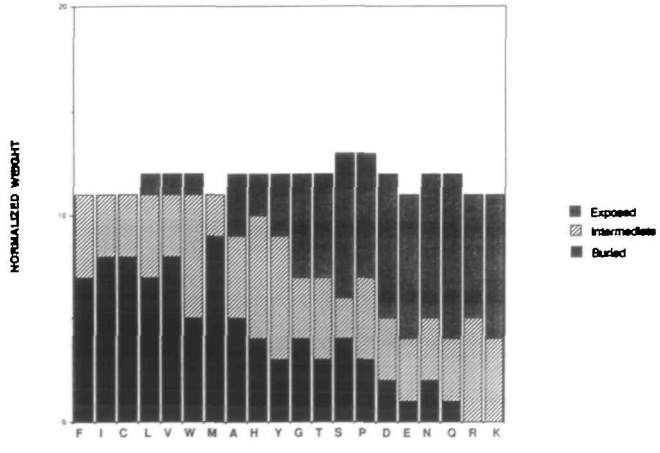

Fig. 2. *Histogram of the neural network weights of the central amino acids in the ternary distribution model of amino acid exposure (buried/intermediate/exposed). Amino acid types are listed across the histogram, while relative magnitudes of the buried, intermediate and exposed weights are stacked vertically. The residue types are ordered from hydrophobic to hydrophilic as given in Table IV.*

Consensus prediction for homologous protein families

The crystal structures of homologous proteins have shown that not only is the overall chain fold conserved, but also the surface exposure of various residue side chains. For example, the structures of hen egg-white lysozyme and human lysozyme are both known to high resolution (2.0 and 1.5 Å respectively). The surface exposure of these two proteins as calculated from their crystal structures agree for over 91% of the residues according to a binary model (buried/exposed) even though only 60% of the residues are identical, thus validating the concept that homologous proteins have very similar exposure patterns. This concept can be used to increase exposure prediction accuracy in the following manner. The amino acid sequences of members of a homologous family are aligned using standard methods with gaps inserted where necessary. Next, the exposure for each member of the family is calculated using a trained neural network. The prediction scores are then matched according to the sequence alignment and a minimum, maximum and average prediction score calculated over all members of the family for each residue position. At the location of gaps neutral prediction scores are assigned, i.e. 0.5 where predictions range from 0.0 to 1.0. We have tested this procedure using six members of the lysozyme family: human, hen egg white, bovine, rat, California quail and Hanuman Langur lysozymes. The last four were chosen from the protein Identification Resource, PIR (George *et al.*, 1986), having codes of LZBO, LZRT, LZQJEC and A29736

Table VI. Amino acid exposure prediction distribution

Amino acid[a]	Exposed-predicted exposed	Buried-predicted buried	Exposed-predicted buried	Buried-predicted exposed	% Correct
Phe	0	12	2	0	86
Ile	0	28	1	0	97
Cys	0	7	0	0	100
Leu	0	41	2	0	95
Val	1	29	0	0	100
Trp	0	2	0	0	100
Met	0	5	0	0	100
Ala	6	13	0	2	90
His	2	0	0	0	100
Tyr	0	0	0	0	*na*
Gly	12	9	6	4	68
Thr	6	0	0	4	60
Ser	30	3	0	11	75
Pro	11	1	1	1	86
Asp	12	0	0	0	100
Glu	25	0	1	2	89
Asn	22	0	0	7	76
Gln	24	0	0	5	83
Arg	9	0	0	1	90
Lys	31	0	0	0	100
Totals	191	150	13	37	87.2

[a]Hydrophobic residues (Phe–Trp), ambiphilic residues (Met–Thr) and hydrophilic residues (Ser–Lys) are separated.

respectively. These proteins are between 58% (LZQJEC) and 87% (A29736) homologous with human lysozyme. Using the higher resolution structure of human lysozyme as the standard for the actual exposure state, we employed the following algorithm for prediction. The predicted exposure state of a residue was chosen as the greater of (i) the maximum prediction of the exposed state and (ii) the maximum prediction of the buried state over all members of the protein family. This procedure allowed us to increase our prediction accuracy from 76.9 to 80% for all residues, and from 87 to 95% for half the residues which are most strongly predicted. Other algorithms employing the averages, or minimum prediction scores gave similar results. We suspect that increasing the number of proteins used as members of the homologous family and thereby increasing variability of amino acid type at each position could further improve the results.

Discussion

The surface exposure of amino acid residues in proteins obviously depends on primary, secondary and tertiary structure. Since for most proteins, only sequence data are available, we have used only the amino acid identity and that of its sequential neighbors in predicting surface exposure. In this article we show that predictions of surface exposure made by neural networks trained on amino acid and flanking residue sequence are more accurate than those made with standard amino acid hydrophobicity scales alone or averaged over a sliding window. Lipman *et al.* (1987) have shown that there is a significant tendency for buried and accessible residues to run in clusters along the sequence, but that distributions of hydrophobicities or hydrophilicites do not show this tendency. This observation implies that use of standard hydrophobicities alone for prediction of residue exposure is limited in power.

Somewhat surprisingly, we observe only a 2% overall increase in accessibility prediction for our testing set by including neighboring residues in both the binary and ternary models (using hidden nodes a 4% increase was obtained for the ternary model). This corresponds to a 7% decrease in error for the binary model (30 to 28%) and an 8% decrease for the ternary model (49.8 to 45.8%). Because of the bias toward prediction of residues as exposed in the testing sets we feel that the correlation coefficient is a better measure of prediction accuracy for these networks. Thus, the increase from 0.40 to 0.44 (10%) by including flanking residues in the binary model may be considered a good estimate for the influence of neighboring residues on amino acid exposure in proteins. When only the flanking residues were used for prediction with the binary model, it was still possible to correctly predict over 55% of the cases (correlation coefficient 0.10), ~10% over random. These results strongly support the notion that the flanking sequence exerts a small, but significant contribution toward predicting surface exposure.

An advantage of using neural networks for prediction of residue accessibility is that analysis of the network weights allows us to make a physical interpretation of the major factors influencing exposure. From the plot of network weights in the binary model shown in Figure 1, it is apparent that the primary factor governing exposure of the strongly hydrophobic (Val, Trp, Met) and hydrophilic (Glu) residues is the identity of the central amino acid itself; however, for neutral or ambiphilic residues, e.g. proline and glycine, the flanking sequence is more influential. Nevertheless, the weights show that hydrophobic residues two and three amino acids before or after the central amino acid favor its burial. This is likely due to the preponderance of buried residues in β-strand and to a lesser degree α-helical structures and the periodicity of these structures. Since exposed residues are favored over buried in turn and coil regions, exposure of the central residue is favorably influenced by neighboring residues, e.g. proline and glycine, which preferentially are found in these regions (high turn propensities). As turns and coils are not periodic structures, less positional specificity is observed for the exposed residues than for buried residues which prefer regular secondary structure.

The use of a three-state exposure model offers several advantages over the two-state model. First, the definition of buried and exposed residues is clarified since intermediate cases are classified as a third category. Second, it is possible to reproduce the observed distribution more closely by allowing more classes. Finally, if it is not necessary to distinguish between fully and partially exposed residues, it is possible to predict exposure with very high accuracy.

Analysis of weights to the output nodes of the three-state model shows a greater contribution of neighboring residues to the exposure of the central residue, especially for the intermediate (partially exposed) node, which is not strongly determined by the central residue alone (not shown). The weights of Figure 2 suggest that larger residues (i.e. W, H, Y and R) tend toward intermediate exposure (correlation coefficient 0.35) regardless of their hydrophobicity. Since size of the residue is more important in the ternary model, the order of the weights of the central residues of Figure 2 no longer follow the hydrophobicity scale of Table IV. Generally, high weights for neighboring hydrophobic residues tend to favor burial of the central residue and high weights for neighboring hydrophilic residues favor exposure of the central residue (not shown). These patterns, therefore, appear to reflect the tendency of buried or exposed residues to cluster in runs or patches (Lippman, 1987).

In this paper, we have presented both a two-state and three-state model for surface exposure of protein residues and made highly accurate predictions of accessibility based solely on the identity of the amino acid of interest and its flanking sequence. We believe that this ability to predict surface exposure of protein residues can be a valuable guide to the protein engineer in locating sites for site-specific mutations, to immunologists in identifying antigenic determinants, and to theorists by placing a strong constraint on protein folding.

Acknowledgements

The authors are grateful to Marge Hutchinson for her assistance and advice. We acknowledge the support of the Health Effects Research Division, Health and Environmental Research, Office of Energy Research of the US Department of Energy. One of the authors (S.M.M.) is supported by a University of California Regents Fellowship.

References

Alden,C.J. and Kim,S.-H. (1979) *J. Mol. Biol.*, **132**, 411–434.
Argos,P. (1988) *Protein Engineering*, **2**, 101–113.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Connolly,M.L. (1983) *Science*, **221**, 709–713.
Eisenberg,D. (1984) *Annu. Rev. Biochem.*, **53**, 595–623.
George,D.G., Barker,W.C. and Hunt,L.T. (1986) *Nucleic Acids Res.*, **14**, 11–15.
Holbrook,S.R. and Kim,S.-H. (1983) *Biopolymers*, **22**, 1145–1166.
Holley,L.H. and Karplus,M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 152–156.
Hopp,T. and Woods,K. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.
Janin,J. (1979) *Nature*, **277**, 491–492.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Lawrence,C., Auger,I. and Mannella,C. (1987) *Proteins*, **2**, 153–161.
Lee,B.K. and Richards,F.M. (1971) *J. Mol. Biol.*, **55**, 379–400.
Lipman,D.J., Pastor,R.W. and Lee,B. (1987) *Biopolymers*, **26**, 17–26.

Lippmann,R. (1987) IEEE ASSP, 4–22.
McGregor,M.J., Flores,P.T. and Sternberg,M.J. (1989) *Protein Engineering*, **2**, 521–526.
Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) *J. Mol. Biol.*, **196**, 641–656.
Muskal,S.M., Holbrook,S R. and Kim,S.-H. (1990) *Protein Engineering*, **3**, 667–672.
Nakata,K., Kanehisa,M. and DeLisi,C. (1985) *Nucleic Acids Res.*, **13**, 5327–5340.
Ponder,J.W. and Richards,F.M. (1987) *J. Mol. Biol.*, **193**, 775–791.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Richmond,T.J. (1984) *J. Mol. Biol.*, **178**, 63–89.
Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) *Science*, **229**, 834–838.
Rumelhart,D.E., McClelland,J.L. and the PDP research group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1, MIT Press, Cambridge, MA.
Shrake,A. and Rupley,J.A. (1973) *J. Mol. Biol.*, **79**, 351–371.
Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) *Nucleic Acids Res.*, **10**, 2997–3011

Received on October 17, 1989; accepted on March 23, 1990