Protein Engineering vol.11 no.1 pp.11–19, 1998

Prediction of protein hydration sites from sequence by modular neural networks

L.Ehrlich, M.Reczko[1,2], H.Bohr[3] and R.C.Wade[4]

European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, [1]German Cancer Research Centre, 69120 Heidelberg, Germany, [3]Center for Biological Sequence Analysis, The Technical University of Denmark, Building 206, DK-2800 Lyngby, Denmark

[2]Present address: Synaptic Ltd, Aristotelous 313, Acharnai, Greece

[4]To whom correspondence should be addressed

The hydration properties of a protein are important determinants of its structure and function. Here, modular neural networks are employed to predict ordered hydration sites using protein sequence information. First, secondary structure and solvent accessibility are predicted from sequence with two separate neural networks. These predictions are used as input together with protein sequences for networks predicting hydration of residues, backbone atoms and sidechains. These networks are trained with protein crystal structures. The prediction of hydration is improved by adding information on secondary structure and solvent accessibility and, using actual values of these properties, residue hydration can be predicted to 77% accuracy with a Matthews coefficient of 0.43. However, predicted property data with an accuracy of 60–70% result in less than half the improvement in predictive performance observed using the actual values. The inclusion of property information allows a smaller sequence window to be used in the networks to predict hydration. It has a greater impact on the accuracy of hydration site prediction for backbone atoms than for sidechains and for non-polar than polar residues. The networks provide insight into the mutual interdependencies between the location of ordered water sites and the structural and chemical characteristics of the protein residues.

Keywords: hydration/ligand binding site/neural network/secondary structure/solvent accessible surface

Introduction

Proteins act in aqueous solution and water plays a fundamental role in their structure, function and dynamics. Ordered water molecules are observed in and around proteins by X-ray crystallography and nuclear magnetic resonance (Levitt and Park, 1993). Some of these can be considered as an integral part of the protein structure (Loris *et al.*, 1994). For example, the positions of 21 buried water molecules were found to be conserved in a study of homologous serine proteases (Finer-Moore *et al.*, 1992; Sreenivasan and Axelsen, 1992) and preferred consensus hydration sites were found in a set of FKBP12–drug complexes (Faerman and Karplus, 1995). Solvent molecules may be important for protein folding and stability (Clore and Gronenborn, 1992; Pedersen *et al.*, 1994), e.g. they are often found at turns in proteins (Thanki *et al.*, 1991; Loris *et al.*, 1994). They also serve to fill and stabilize cavities (Wade, 1990; Hubbard and Argos, 1994; Hubbard *et al.*, 1994; Williams *et al.*, 1994) where they can have long residence times exceeding a nanosecond (Otting *et al.*, 1991). Water molecules are important in the interaction of proteins with their ligands. Ordered water molecules are generally displaced (Rand and Fuller, 1992) from the protein surface on ligand binding providing an entropic contribution to the ligand binding free energy. They may however be trapped at the ligand–protein interface and such water molecules can bridge hydrogen bonds, fill space and modulate protein–ligand specificity, see e.g. Quiocho *et al.* (1989). They may also play a functional role (Affleck *et al.*, 1992; Meyer, 1992).

Knowledge of hydration sites is thus necessary for the full characterization of protein structure and function. Fast, accurate and efficient methods to predict hydration sites would be useful for the assignment of water sites during crystallographic and NMR refinement, for the setting up of molecular dynamics simulations, for the prediction of the effects of protein mutations and the locations of ligand binding sites, and for deducing protein functional mechanisms that involve individual water molecules.

A number of methods to locate the coordinates of hydration sites around the three-dimensional structures of proteins have been developed. Energy-based methods range from the calculation of the interaction energy of a water molecule with a protein e.g. in the GRID method (Goodford, 1985; Boobbyer *et al.*, 1989; Wade, 1990; Wade and Goodford, 1993) and CARTE program (Goodfellow and Vovelle, 1989) to molecular dynamics (Wade *et al.*, 1990; Helms and Wade, 1995; Roux *et al.*, 1996; Zhang and Hermans, 1996) calculations to compute free energies of hydrating cavities in proteins. Other approaches include rule-based approaches, such as the algorithm based on the directionality of hydrogen bonds developed by Vedani and Huhta (1991) to determine protein solvation sites; knowledge-based approaches, such as the AQUARIUS programs of Pitt and Goodfellow (1991), Pitt *et al.* (1995) and the work of Roe and Teeter (1993); and the particle-correlation-based approach of Hummer *et al.* (1995) to compute water density distributions.

Artificial neural networks provide a powerful tool for pattern recognition and have proven useful for a range of chemical and biological applications (Burns and Whitesides, 1993). Here, neural networks are used to predict protein hydration sites from protein sequence information using crystal structures in the Brookhaven protein data bank (Bernstein *et al.*, 1977) to train the networks. A preliminary account of networks to predict hydration was given by Wade *et al.* (1992). The new networks described here give improved predictions. This is partially due to the inclusion of both solvent accessibility and secondary structure information. The location of water binding sites on proteins is related to both these features (Thanki *et al.*, 1990, 1991; Kuhn *et al.*, 1992; Morris *et al.*, 1992) and the networks provide insights into the interdependencies of these quantities. As both quantities may be predicted using neural networks with only amino acid sequence as input (Bohr *et al.*, 1988; Qian and Sejnowski, 1988; Holley and Karplus, 1989;

Holbrook *et al.*, 1990; Rost and Sander, 1994a,b), the networks described here to predict water binding sites do not require the tertiary structure of the test proteins to be known and are thus widely applicable. That is, networks are used in a tiered, modular fashion: in the first tier, separate recurrent networks are used to predict secondary structure and solvent accessibility. In the second tier, a network predicts hydration sites using output from the networks of the first tier as input.

Materials and methods

Data sets

Protein structures from the Brookhaven Protein Databank (Bernstein *et al.*, 1977) were used for training and testing the networks (see Appendix 1). Forty proteins (6913 residues) were used for training [as in our previous study (Wade *et al.*, 1992)]. Seventy-seven proteins (18698 residues) were used for testing the final networks. All the protein structures chosen contain more than 32 crystallographic water sites, were solved to better than 2.0 Å resolution, and have *R*-factors less than 22%. They cover a wide range of protein tertiary structure types. The 77 protein test set contains only single-chain proteins with less than 25% homology to each other and to proteins in the training set (Hobohm and Sander, 1994).

For the water predictions, all symmetry-related water sites within 5 Å of any non-hydrogen atom in the proteins were generated using QUANTA (Molecular Simulations Inc., SD, USA), INSIGHT (Molecular Simulations Inc., San Diego, USA) and WHATIF (Vriend, 1990) software packages. If two water sites were positioned closer than 1.2 Å to each other, this was taken to indicate the presence of one partially disordered water molecule rather than two water molecules and one of these water sites (the symmetry generated one where possible) was deleted.

Definition and representation of the properties predicted

The following definitions of the properties predicted were used:

Secondary structure. The definitions of Kabsch and Sander (1983) based on hydrogen bond formation as implemented in the DSSP program (Kabsch and Sander, 1983) were used. Three classes were considered: α-helix, β-sheet and coil (i.e. all non-α and non-β structures). For the networks, secondary structure was represented by three binary neurons corresponding to the three classes.

Solvent accessibility. Solvent accessibility was calculated for a probe sphere of radius 1.4 Å as implemented in the DSSP program (Kabsch and Sander, 1983) in a manner similar to the algorithm of Shrake and Rupley (1973). Solvent accessibilities were then normalized following the method of Holbrook *et al.* (1990) to the values given by Rose *et al.* (1985). The resultant fractional accessibilities were then binarized using a threshold of 0.2. For the networks, solvent accessibility was represented by two (complementary) binary neurons, one signifying surface and the other buried residues.

Hydration

(i) Residue hydration. A residue was defined as hydrated when a crystallographically observed water molecule lay within 3.5 Å of any of its atoms.

(ii) Backbone oxygen and nitrogen hydration. One of these atoms was defined as hydrated when a water molecule lay within 3.5 Å of it and closer to it than any other atom in the same residue. Hydration of backbone carbon atoms was not considered.

(iii) Sidechain hydration. A sidechain was defined as hydrated when a water molecule lay within 3.5 Å of any of its non-carbon sidechain atoms. No predictions were made for residues without sidechains or with sidechains consisting only of carbon and hydrogen atoms.

The cut-off distance of 3.5 Å was chosen to include all water molecules making hydrogen bond interactions with polar protein atoms and the stronger interactions with non-polar atoms. The same distance was used by Thanki *et al.* (1990, 1991) in their analysis of the solvation of protein structures. However, it has since been observed (Walshaw and Goodfellow, 1993) that solvent sites around apolar protein carbon atoms peak at about 4 Å. Most of these solvent molecules are involved in hydrogen bond interactions and therefore are included in our analysis but a minority are not. These were only statistically significant around Ala and Phe residues (Walshaw and Goodfellow, 1993).

For the networks, hydration was given by two (complementary) binary neurons, one signifying hydrated by one or more ordered water sites, the other, unhydrated. Note that two or more residues may be hydrated by a single water molecule or that two or more water molecules may hydrate a single residue. Thus the number of hydrated residues does not correspond to the number of crystallographically observed water sites. Analogous statements are also true for the hydration of backbone atoms and sidechains.

Representation of the input protein sequence

The following representations of the amino acid sequence were used:

RA. Each residue was specified by 20 binary (0/1) neurons corresponding to the 20 standard amino acids. Unusual or unknown residues or the absence of a residue at a position in the input window was indicated by zero values for all 20 neurons.

RB. Each residue was represented by seven real-valued neurons encoding its physical properties, six corresponding to hydrophobicity (Eisenberg *et al.*, 1982), size, polarity, charge, aromaticity, aliphaticity and the seventh indicating whether the residue was proline (Bohr *et al.*, 1992). These seven neurons provide a unique description of each amino acid.

Tests were also made using just one neuron coding for hydrophobicity data but this resulted in considerably reduced performance, indicating that other factors are important in determining hydration.

Network architecture

For each type of hydration prediction, three networks were used. The results of networks to predict secondary structure and solvent accessibility were fed into the network to predict hydration sites as shown in Figure 1. The networks were built, trained and tested with the SNNS program (Stuttgart Neural Network Simulator, version 3.2, University of Stuttgart, Germany).

For the solvent accessibility predictions, an Elman type recurrent neural network (Elman, 1990) was used. This is essentially a multi-layer feed-forward network with extra feedback connections between the hidden neurons. The network consisted of 17 × 20 (3400) input neurons, three hidden neurons and two output neurons. The input window represents a continuous sequence of 17 amino acid residues centred on the residue for which a prediction is being made. Each residue is coded by 20 binary neurons.

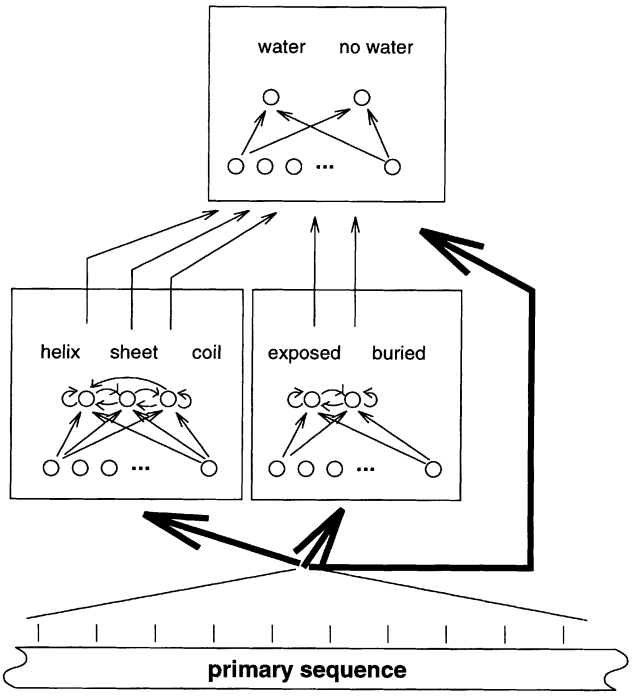

Fig. 1. Schematic diagram of how the modular networks are used to predict hydration sites.

For the secondary structure predictions, a recurrent neural network similar to the Jordan type (Jordan, 1986) was used. This is basically a perceptron (without hidden neurons) and with recurrent connections between the output neurons. The network consisted of 13 × 20 (2600) input neurons and three output neurons.

For the hydration predictions, perceptrons and feed-forward networks with one hidden layer of three neurons were used. The networks had two binary output neurons. Networks with input neurons representing sequence windows of 1, 3, 9 and 17 residues were tested. Input neurons representing secondary structure (3) and solvent accessibility (2) were also used. Each residue was represented either by 20 binary neurons corresponding to each amino acid type or by seven real-valued neurons describing its physical properties. For single amino acid residue windows, the results were mostly similar for the two representations of amino acid type (differences within error bars). For large (9 and 17 residue) windows, the 20 neuron representation of amino acids was found to be impractical for this property because the increased number of neurons and connections resulted in much slower and less reliable training of the networks.

Training protocols

Training was performed using the Quickprop algorithm (Fahlman, 1989). For each network type, training was performed for one network until there was negligible improvement in performance, as measured by percentage correct and Matthews coefficient (Matthews, 1975) on testing. The same number of steps was then used for the remaining equivalent networks.

The network architecture was optimized for prediction by cross-validation using a leave-some-out procedure within the 40 protein training set. The training set was divided into five groups of eight proteins. Five networks were trained on 32 proteins (different combinations of four of these five groups) and tested on the remaining group of eight proteins. The five trained networks were used to make predictions for the separate test sets. These five predictions were combined to produce a single prediction by making a 'winner takes all' decision for the output of each network and then determining the majority decision. The full training set of 40 proteins was also used to train networks whose resultant predictive performance was very similar to that derived from the corresponding five networks.

As discussed later and shown in Figure 4a, considerably more of the backbone atoms were unhydrated than hydrated. This imbalance made training difficult due to the insensitivity of the error function, the sum of the squared errors for all data points, to learning. To reduce the imbalance, networks were trained with all atoms with one or more water ligands input three times for the nitrogen predictions and twice for the oxygen predictions for every input of an atom without a water ligand. This resulted in improved learning for hydrated atoms and improved predictive ability even though it made the overall hydration profile of the training sets very different from that of the test sets.

For networks for hydration site prediction using predicted secondary structure and solvent accessibility, tests were made for training with actual or predicted values of these properties and for presenting the predicted values in real or binarized form.

Performance evaluation

Performance was evaluated in terms of the following quantities after making all values of output neurons integer quantities:

Q. the % of patterns correctly predicted; and

C. the Matthews correlation coefficient (Matthews, 1975) defined as follows. If the two possible output values are denoted by 0 and 1, and if p is the number of correctly predicted examples of 1s, $\bar{p}$ is the number of correctly predicted examples of 0s, q is the number of examples of 1s incorrectly predicted and $\bar{q}$ is the number of examples of 0s incorrectly predicted, then the Matthews coefficient C_M is defined by:

$$C_M = \frac{p\bar{p} - q\bar{q}}{\sqrt{(p + q)\,(p + \bar{q})\,(\bar{p} + q)\,(\bar{p} + \bar{q})}} \qquad (1)$$

When there is total agreement between predicted and target values (ideal performance), C_M is 1. For complete disagreement, C_M is −1. When $C_M = 0$, predictions are equivalent to those obtained based on the proportion of 0s and 1s in the dataset. Thus values of C_M greater than zero indicate that the network has learned correlated features in the data and has predictive ability.

Results and discussion

Predictions of secondary structure and solvent accessibility

Performance data are given in Table I. The predictive ability achieved for the three secondary structure categories (Q = 62% and C ~0.36) is similar to that achieved in other neural network predictions of secondary structure using sequence input only (Rost and Sander, 1993; Chandonia and Karplus, 1995). The combination of the results of five networks provides a slight improvement in the results compared with using one network. Considerable improvements could be made by increasing the training set size (Chandonia and Karplus, 1996) and by using evolutionary information (Rost and Sander, 1994a; Chandonia and Karplus, 1996; Frishman and Argos,

Table I. Performance of neural networks for predicting secondary structure and solvent accessibility[a]

Crossvalidation: train on 32 proteins, test on eight proteins						
Network	Secondary structure				Solvent accessibility	
	Q_{ss} (%)	C_{α}	C_{β}	C_{coil}	Q_{sa} (%)	C_{sa}
A	63.97	0.39	0.39	0.40	73.80	0.47
B	63.14	0.41	0.39	0.44	68.71	0.37
C	62.88	0.34	0.37	0.41	69.16	0.38
D	58.43	0.27	0.29	0.39	67.61	0.35
E	59.51	0.22	0.34	0.34	67.75	0.35
Average (sd)	61.59 (2.45)	0.33 (0.08)	0.36 (0.04)	0.40 (0.04)	69.41 (2.54)	0.38 (0.05)
Testing set (77 proteins)						
jury (sd)	62.01 (7.05)	0.34 (0.19)	0.33 (0.15)	0.40 (0.11)	69.77 (5.34)	0.39 (0.11)

[a]Prediction quality was assessed by two quantities: percentage of residues for which predictions were correct, *Q*, and Matthews coefficient, *C*, as defined in the Methods section. The subscripts *ss* and *sa* refer to secondary structure and solvent accessibility respectively. As the Matthew's coefficient is defined for two states, three coefficients are given for the secondary structure predictions. C_{α} is the coefficient for a residue being in an α helix versus in any other structure; C_{β} and C_{coil} are defined analogously.

Table II. Performance of neural networks for predicting protein hydration for the 77 protein test set[a,b]

Protein atoms	Sequence only		Sequence + actual properties		Sequence + predicted properties	
	Q (%)	*C*	*Q* (%)	*C*	*Q* (%)	*C*
Residue	70.39 (5.84)	0.31 (0.10)	76.83 (7.07)	0.43 (0.12)	72.68 (6.54)	0.32 (0.11)
N	69.07 (4.78)	0.10 (0.08)	67.94 (5.52)	0.20 (0.11)	62.61 (4.92)	0.14 (0.07)
O	54.82 (3.82)	0.09 (0.08)	59.83 (5.57)	0.25 (0.09)	52.69 (5.00)	0.14 (0.07)
Sidechain	65.23 (6.67)	0.25 (0.12)	65.04 (6.35)	0.26 (0.11)	64.69 (5.98)	0.26 (0.11)

[a]Predictions are given for the combined decision of five networks with standard deviations in parentheses.
[b]Predictions were made with the network architectures giving the best performance for each quantity as follows. For residue, N and O atom predictions: for input of sequence only, there were 63 input neurons representing 9 residues by 7 properties, and a hidden layer of 3 neurons. When properties were also input, there were 12 input neurons, 7 representing the residue by its properties, 2 representing solvent accessibility and 3 representing secondary structure. There was also a hidden layer of 3 neurons. For side-chain predictions: for input of sequence only, a perceptron with 20 input neurons representing 1 residue by its identity was used. When properties were also input, there were 25 neurons, 20 representing the residue, 2 representing solvent accessibility and 3 representing secondary structure.

1996; Riis and Krogh, 1996) in the form of adding a profile of aligned homologous sequences to each sequence considered. With a certain amount of sequence similarity (usually around 20%) it is possible to achieve an overall accuracy of about 72% (Rost and Sander, 1994a; Chandonia and Karplus, 1996; Riis and Krogh, 1996). It seems, however, that neural networks may not be able to achieve greater than about 70% accuracy in predicting secondary structures of non-homologous proteins when no sequence profile is available. Prediction of secondary structure based on sequence windows can be related to formation of structures early in the protein folding process and, since these structures are approximately 70% identical to the ones in the native state (Goldberg *et al.*, 1990), it is expected that the performance of secondary structure predictions will not exceed this value.

For solvent accessibility predictions, the overall accuracy is about 70% with a Matthews coefficient of 0.39 which is similar performance to that obtained by others (Holbrook *et al.*, 1990) without use of homology information. Improvements can also be achieved for solvent accessibility predictions by incorporating evolutionary information (Holbrook *et al.*, 1990; Rost and Sander, 1994b).

Prediction of hydration sites

Performance data for predictions for the test proteins are given in Table II for the best network architectures according to cross-validation within the 40 protein training set. Very similar performance values were obtained during cross-validation and for the test set of 77 proteins. Residue hydration is predicted from sequence alone with about 71% accuracy and a Matthews coefficient of about 0.31. The Matthews coefficient indicates that residue hydration is predicted less well than secondary structure and solvent accessibility when sequence alone is used as input. When secondary structure and solvent accessibility are used as inputs, predictive ability is improved and with the actual values of the properties, is about 77% with a Matthews coefficient of about 0.43—i.e. at least as good as the predictions for secondary structure and solvent accessibility. With predicted values of secondary structure and solvent accessibility, the predictive performance for residue hydration drops to a level intermediate between that with sequence only and with actual values of secondary structure and solvent accessibility.

The Matthews coefficient indicates that hydration of backbone atoms is poorly predicted from sequence alone and is improved by the addition of information on secondary structure and solvent accessibility, whether actual or predicted, although this might result in a reduction of the overall percentage correct as the Matthews coefficient improves. This effect is due to the large discrepancy between the number of hydrated and the number of unhydrated backbone atoms in the data sets. Hydration of side-chains is better predicted from sequence

alone and is little improved when information on secondary structure and solvent accessibility is added.

Networks were trained with both actual and predicted secondary structure and solvent accessibility. The use of predicted values was tested because it was anticipated that if there were systematic errors in the predictions, these would be learnt but would be the same for the test set as for the training set and thus could even aid prediction of hydration. In fact, there is little difference in predictive performance when the networks are trained with actual or predicted secondary structure and solvent accessibility. This means that actual values can be used for the training set. For a test protein, secondary structure and solvent accessibility values predicted by any method can be used. This increases the flexibility and applicability of the method. There was also little difference between using real-valued or integer input neurons to describe secondary structure and solvent accessibility.

The present neural networks show considerably better predictive ability for the same protein dataset (the eight protein test set) than those in our previous study in which information on sequence and secondary structure was used (Wade *et al.*, 1992). One reason is the inclusion of solvent accessibility information in the current study. However, improved predictive ability is obtained even for networks with only sequence information as input. The main reason for the improvement in prediction is the use of a better network architecture and training algorithm. In the present study, overtraining was avoided by using the Quickprop algorithm and crossvalidation for predictive performance. Another difference is the inclusion of symmetry-related water molecules in the present datasets. However, this has little positive effect on predictive ability because many of the additional water molecules make bridging interactions between protein molecules in the crystal and thus their positions are influenced by interactions with residues not known to the neural network. Analyses of the structures of proteins solved from different crystals show that these water sites are often not conserved (Loris *et al.*, 1994; Zhang and Matthews, 1994).

Dependence on adjacent sequence. The results are sensitive to the input sequence window size and the optimal window size is dependent on whether secondary structure and solvent accessibility are used as input for the networks predicting hydration sites. In all cases tested, the results with the maximum window size tested of 17 amino acid residues were not better than with a window size of nine residues. This is probably because of the lower signal-to-noise ratio.

For predictions of sidechain hydration, results were always best with a single residue window regardless of whether secondary structure and solvent accessibility information was used. This indicates that sidechain hydration is mostly dependent on the type of the amino acid and the adjacent residues are of little importance. For sidechain hydration predictions, the weights of connections from methionine, and to a lesser extent cysteine, tend to be larger than other weights. The connection from methionine has the largest magnitude weight in four out of five networks for both 'sequence only' and 'sequence + properties' networks. Cysteine has the largest magnitude weight in the remaining networks. The sidechains of these residues are rarely hydrated and the networks predict that they are all unhydrated thus achieving an accuracy of 92–94% but a Matthews coefficient of zero for these residue types.

For networks for predicting the hydration of backbone nitrogens, analysis of the weights for the connections from the input to the hidden layer neurons showed that the most important input information was whether the residue for which predictions were made was proline or not. The connection from this input neuron had the highest magnitude weight in three out of five cross-validation networks for both 'sequence only' and 'sequence + properties' networks. In all cases, the weight of this connection was among the 10 highest. No such prominence was observable for networks predicting residue or backbone oxygen hydration. The importance of proline for backbone nitrogen predictions has a clear physical basis: proline is the only residue whose backbone nitrogen lacks the ability to donate a hydrogen bond. Almost all proline nitrogens in the test set are unhydrated and the networks predict that they will all be unhydrated thus achieving high accuracy of 91% for this residue but a Matthews coefficient of zero.

However, despite this important attribute of the residue predicted for backbone nitrogen hydration, the effect of adjacent residues is apparent for the predictions for backbone atoms and whole residues. Predictions based on sequence only are best for a window size of nine residues; predictions based on sequence, secondary structure and solvent accessibility are better for a single aminoacid window, with window size making a difference of up to 0.1 in the Matthews coefficient obtained on testing by cross-validation. Thus secondary structure and solvent accessibility data can largely substitute for extended amino-acid sequence data. Their inclusion results in much smaller networks that are much simpler and faster to train than those with long sequence windows. The networks can more easily decipher the explicit property information rather than the information implicitly given in a sequence window by residue type.

Analysis of the weights of connections in the 'sequence only' networks shows that connections from neurons representing whether flanking residues are proline or not tend to have high weights. For residue hydration predictions, the neurons with information about whether the residues immediately before and after the predicted residue are proline tend to have larger mean weight magnitudes, different by a factor of about three from the mean weight magnitudes of other neurons, and are always amongst the 10 connections with highest weights. For backbone oxygen and nitrogen predictions, it is the proline-likeness of the $i - 2$ and $i + 2$ residues that is most important on average. Each of these connections is seen in three out of five of the networks amongst the 10 connections with highest weights. Thus it appears that the most important descriptor of the structure of the sequence flanking a residue for which predictions are made is whether the $i \pm 1$ and $i \pm 2$ residues are proline or not. Prolines disrupt or kink alpha helices. They tend to occur in turns where they also tend to be solvent exposed. Thus they are good descriptors of secondary structure and solvent accessibility. Their dominance in the sequence only networks probably explains why the adjacent sequence becomes rather unimportant when information about secondary structure and solvent accessibility is explicitly given to the networks.

To assess the validity of considering a limited section of the amino acid sequence for determining water sites, the separation in terms of residues along the sequence between atoms that were ligands to each of the water molecules in the training set protein structures was calculated. This is shown in Figure 2. (Note that for this calculation, residues were renumbered consecutively with all subunits following in

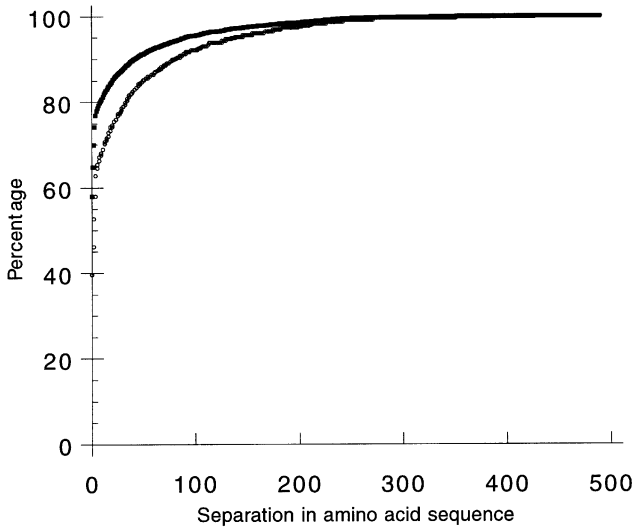

Fig. 2. Plot showing the separation in amino acid sequence of the residues liganded to crystallographically observed water sites. The unfilled circles show the maximum separation in sequence between a pair of protein atom ligands for each water site. The filled circles show the separation in sequence for all pairs of protein atoms liganding to water sites. Both curves rise steeply up to a separation of four residues and then start to flatten off.

sequence and so the separation in sequence may be a little underestimated). From these histograms, it can be seen that about 40% of the water molecules ligand to only one residue. As separation increases from 0–4 residues, there is a rapid increase (to 63%) in the proportion of water molecules accounted for, and that as the separation increases further, the proportion of water molecules increases more gradually. This indicates that a window size of nine residues may be appropriate for optimizing the signal to noise ratio and this is born out by the results. Accuracy should increase with larger windows but at a slower rate per residue added to the window than for windows of less than nine amino acids. Windows of 17 amino acid residues were tested as a window of this size was used for our secondary structure predictions. Other studies, see e.g. Chandonia and Karplus (1996), have shown that a window size of 13 to 19 amino acids is suitable for the prediction of secondary structure as interactions up to nine amino acids apart in sequence are most important for secondary structure. However, in the hydration predictions, this large window size (17) results in too much noise in the data.

Relationship between secondary structure, solvent accessibility and hydration sites. Figure 3 shows the relationship between residue solvent accessible area and hydration for the 40 protein training set. As expected, the least accessible residues (solvent accessible area less than 25 Å^2) are more often without water ligands while residues with solvent accessible areas greater than 25 Å^2 are more likely to be hydrated than not. The threshold in fractional solvent accessible area of 0.2 corresponds to 25 Å^2 in the calculated solvent accessibility.

Analysis of the 40 protein training set shows that 55% of residues in helices and 54% of residues in extended β-structures are hydrated by ordered water molecules. Thus, on average, these secondary structure types are hydrated to similar extents. On the other hand, coil regions are much more hydrated than regions with defined secondary structure with 73% of residues in coil regions hydrated by ordered water molecules.

Thus both solvent accessibility and secondary structure provide information about hydration. However, beneath these

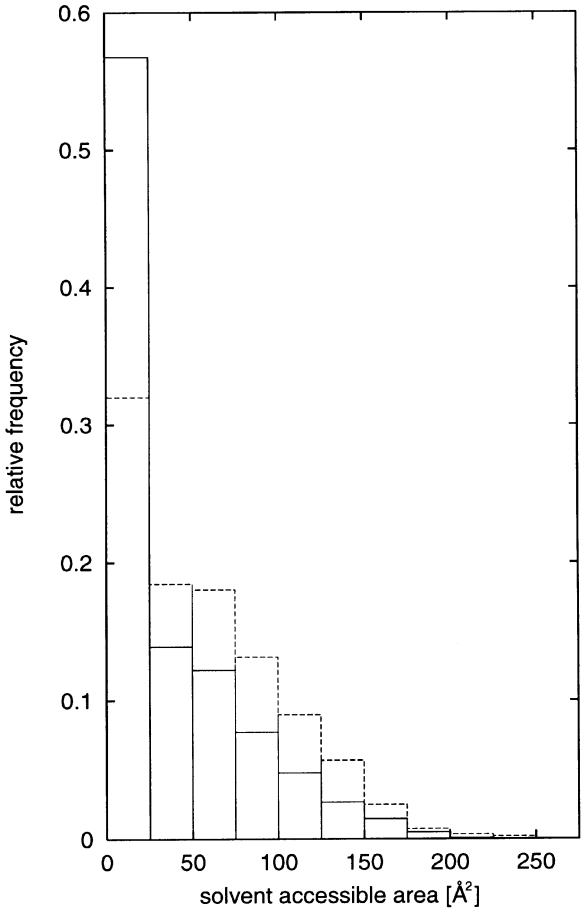

Fig. 3. Histogram of residue hydration versus residue solvent accessible area. Dashed lines show residues hydrated by one of more ordered water sites. Continuous lines show residues that are not hydrated by ordered water sites.

statistics, there are other relationships that are less obvious. Very exposed residues can be very mobile and thus in less well defined regions of a crystal structure lacking secondary structure. The solvent around such residues is also likely to be very mobile and therefore hard to locate in an electron density map. Thus such residues are unlikely to have water ligands and yet will have a high solvent accessibility. On the other hand, Kuhn *et al.* (1992) have shown that grooves on protein surfaces bind three times more ordered water molecules than non-groove surfaces, and that, it is only in grooves that bound water molecules discriminate between different side-chains. Thus it should be remembered that the neural networks are trained to predict the locations of ordered water molecules; this is different from the prediction of regions that are solvent exposed.

The predictive performance of the networks also indicates factors important in determining hydration. Predictions of side-chain hydration shows little improvement with the inclusion of surface area and secondary structure information while those for back-bone oxygen and nitrogen atoms show a significant improvement. In addition, when sequence is the only input, predictive ability is improved by including information about adjacent residues for the backbone atoms but not for the sidechains. This indicates that hydration of a residue's sidechain is predominantly a function of residue type, whereas hydration of a residue's main-chain atoms is more dependent on the nature of its surroundings. Note though that many of

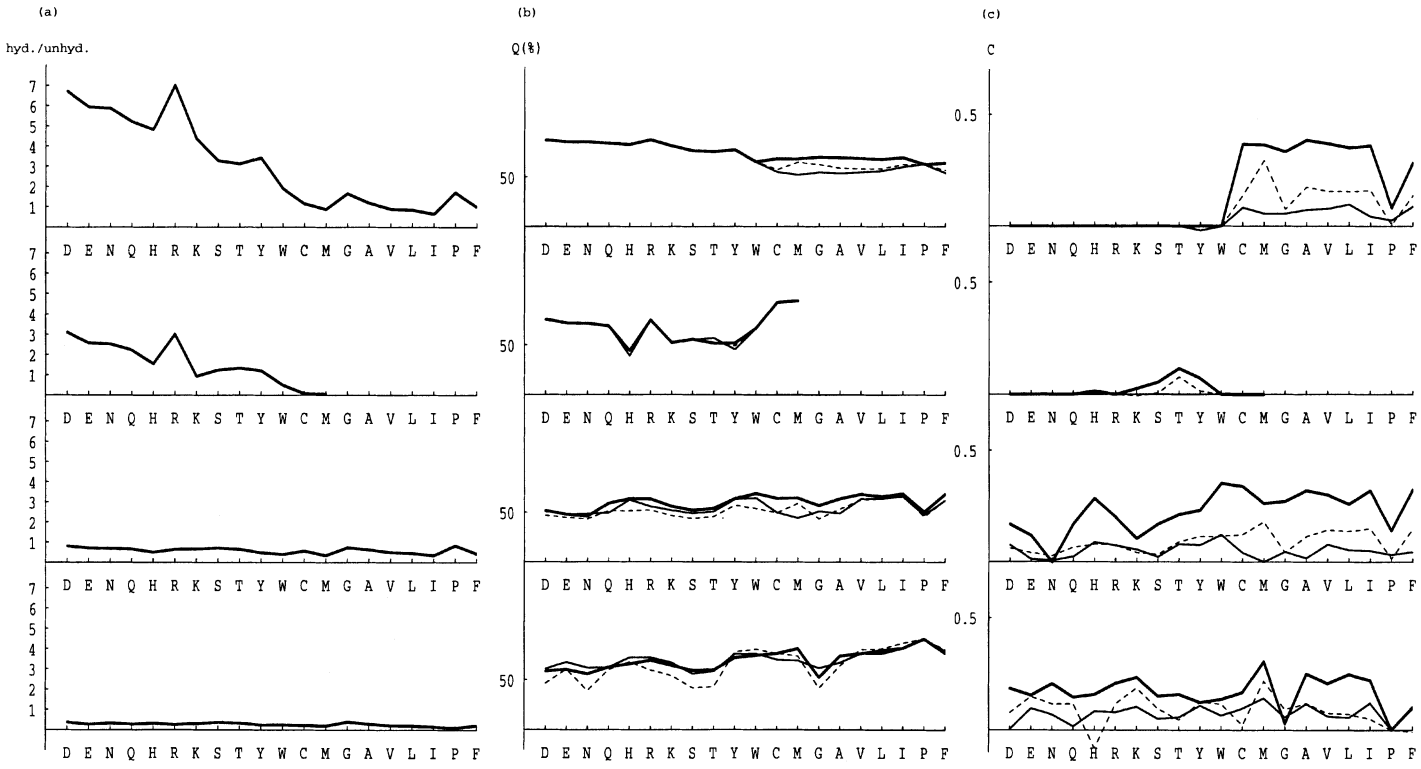

Fig. 4. Relation between hydration by ordered water sites and amino-acid type. (**a**) Ratio of hydrated to unhydrated residues; (**b**) percentage Q of predictions correct: (**c**) Matthews coefficient C for predictions. In (b) and (c), the results are shown with different input data: sequence only (plain lines), sequence + actual properties (bold lines), sequence + predicted properties (dashed lines). Plots are for ordered water sites around, from top to bottom, whole residues, sidechains, backbone oxygen, and backbone nitrogen atoms.

the waters defined as hydrating a main-chain atom will also be interacting with sidechain atoms e.g. through longer hydrogen bonds.

Consistent with previous observations, main-chain amide groups are hydrated much less often than the carbonyl groups (see Figure 4a). If the hydrogen-bond capacity of these groups is satisfied, they are very unlikely to be hydrated (Thanki *et al.*, 1991). Thus main-chain hydration is very dependent on secondary structure with hydration being more likely in β-sheets than α-helices. The free carbonyl groups at the C termini and the NH groups at the N termini of helices and turn/loop regions provide favourable hydration sites (Madhusudan *et al.*, 1993). As shown in Figure 4, oxygen hydration can be predicted better than nitrogen hydration. An increase in predictive ability on going from polar to non-polar residues is observed. This increase is more dominant in the oxygen (as compared with the nitrogen) predictions. A contributor to these differences between oxygen and nitrogen hydration predictions is that the ratio of hydrated to unhydrated oxygen atoms is closer to 0.5 than that for nitrogens so the dataset is more balanced. This is important even though it was taken into account during training by presenting hydrated atoms to the network more often than unhydrated atoms.

In an analysis of the hydration of different residue types (Thanki *et al.*, 1990), it was found that hydration of serine and threonine was dependent on secondary structure while that of tyrosine was not, probably due to its longer side chain. Length of sidechain may also be the reason why hydration of other large sidechains is little influenced by secondary structure. From the network performance shown in Figure 4, it can be seen that the addition of secondary structure and solvent accessibility does not improve predictions for titratable residues and the networks do not detect correlations between the hydration of these sidechains and the sequence. Many more of them are hydrated than not hydrated (see Figure 4a) and it is difficult for the networks to distinguish the unhydrated minority. However addition of secondary structure and solvent accessibility information does improve predictions slightly for serine, threonine and tyrosine, with the greatest impact being from threonine. This trend is consistent with that found in the statistical analysis (Thanki *et al.*, 1990) but the influence of secondary structure is small, as can also be seen in the data presented by Thanki *et al.* (1990).

Considering prediction of residue hydration as a function of amino acid type, it is striking (see Figure 4) that the predictive ability of the networks is almost completely confined to the non-polar residues. It is their hydration that is dependent on the adjacent sequence and its structural and chemical attributes. While this may be in part due to the variation of the ratio of hydrated to unhydrated residues, and if the information was provided, the networks might be able to distinguish the number of water molecules hydrating each residue, this probably points to a real effect rather than a deficiency in the learning of the networks.

Limitations on predictive ability

There is likely to be considerable 'noise' in the datasets on which the networks were trained and tested. Both surface properties and hydration are influenced by crystallization e.g. presence of high concentrations of salt, crystal contacts (Loris *et al.*, 1994; Zhang and Matthews, 1994). Loops, in particular may adopt different conformations in a crystal and in solution, resulting in different solvent exposure; e.g. water may mediate an interaction between two protein molecules in adjacent units

of the crystal. However, there is evidence that important solvent sites are conserved in different crystal structures of the same protein (Loris *et al.*, 1994; Zhang and Matthews, 1994) and homologous proteins (Sreenivasan and Axelsen, 1992; Poornima and Dean, 1995). Uncertainties in the positions of water sites in protein structures determined by X-ray crystallography may also arise because different criteria may be used to assign water sites in different structures (Zhang and Matthews, 1994). The number of crystallographically assigned water sites per residue varies considerably from protein to protein. For the 40 protein training set, the number of water sites per residue ranged from 0.2 to 3.4 (mean 0.9 sd 0.6). The extent of solvation will depend on the protein structure and the crystallization conditions as well as the manner in which water molecules are assigned.

Neural networks learn both positive and negative information: in this case, where there are ordered water sites and where there are not ordered water sites. Because the assignment of water sites may be more conservative in some protein structures than others, it is possible that the experimental data on the presence of water sites is more important than that on the absence of water sites. (A site without a water molecule in one protein may be very similar to one with a water molecule in another structure). Moreover, all these sites are likely to have varying degrees of water occupancy, as water molecules are mobile and move between water sites. It would therefore be more realistic if the output neurons could take intermediate values between 0 and 1 in order to reflect partial occupancies. However, we have chosen to assign only values of 0 and 1 to the output neurons, as occupancies are not usually assigned to water sites in protein structures.

Many water sites in protein structures are assigned temperature factors and these often provide an indication of how strongly bound water is at the site. Exploratory networks were tested in which training input was weighted according to temperature factors but this did not improve predictions. One of the problems with using temperature factors is that they have not been assigned to all water molecules in the data set.

For comparison with previous work, only 40 proteins were included in the training set. Results might be improved by increasing the size of the training set. However, the similarity in the predictive performance on cross-validation for these 40 proteins and on testing with the 77 protein set indicates that the training set is large enough to contain the features relevant for predictions in the larger and non-homologous test set.

Conclusions

While predictive networks were obtained, predictive ability is mostly confined to the backbone atoms of non-polar residues. Few correlations were detected for polar sidechains indicating a weak relationship between ordered water sites around sidechains and the properties of the adjacent sequence. For the networks, information about secondary structure and solvent accessibility of the predicted residue could largely substitute for information about the identity of the residues near in sequence. The networks described here may be generalizable to the prediction of binding sites for other ligands such as cofactors. They require sufficient experimental data for training, data that is most abundant for hydration sites. The amount of data necessary depends on the 'noise' in the data and the specificity of binding. Thus, its worth noting that Hirst and Sternberg (1991) were able to use neural networks to predict ATP/GTP binding motifs. In themselves, hydration site predictions should be of value for determining the interaction properties of proteins and could even provide input for a third layer of networks for tertiary structure prediction or the prediction of the binding of other ligands. However, it should be noted that predictive ability is significantly lower when predicted solvent accessibility and secondary structure rather than the actual values are used as inputs. This indicates that, not only is the secondary structure important for hydration but also that hydration is important for secondary structure and its absence is one of the reasons for the imperfect secondary structure predictions. Thus it is to be expected that better predictions will be made when atomic coordinates are used for predictions when the three-dimensional structure of the protein is known, and we are currently working on networks to do this. On the other hand, for proteins for which only sequence information is available, the present networks provide a useful tool for computing their hydration properties.

Acknowledgements

We thank Prof. Peter Wolynes for many stimulating discussions and Dr Sandor Suhai for his support. This work was funded in part by NATO Collaborative Research Grant No. 930301.

References

Affleck,R., Xu,Z-F., Suzawa,V., Focht,K., Clark,D.S. and Dordick,J.S. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 1100–1104.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennedy,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bohr,H., Bohr,J., Brunak,S., Cotterill,R.M.J., Lautrup,B., Norskov,L., Olsen,O.H. and Petersen,S.B. (1988) *FEBS Lett.*, **241**, 223–228.

Bohr,H., Goldstein,R.A. and Wolynes,P.G. (1992) *AMSE Periodicals, Modelling, Measurement and Control*, **C31**, 55.

Boobbyer,D.N.A., Goodford,P.J., McWhinnie,P.M. and Wade,R.C. (1989) *J. Med. Chem.*, **32**, 1083–1094.

Burns,J.A. and Whitesides,G.M. (1993) *Chem. Rev.*, **93**, 2583–2601.

Chandonia,J.M. and Karplus,M. (1995) *Protein Sci.*, **4**, 275–285.

Chandonia,J.M. and Karplus,M. (1996) *Protein Sci.*, **5**, 768–774.

Clore,G.M. and Gronenborn,A.M. (1992) *J. Mol. Biol.*, **223**, 853–856.

Eisenberg,D., Weiss,R.M., Terwilliger,T.C. and Wilcox,W. (1982) *Farad. Symp. Chem. Soc.*, **17**, 109–120.

Elman,J.L. (1990) *Cognitive Sci.*, **14**, 179–211.

Faerman,C.H. and Karplus,P.A. (1995) *Proteins*, **23**, 1–11.

Fahlman,S.E. (1989) An empirical study of learning speed in back-propagation networks. In *Proc. of the 1988 Connectionist Models Summer School* Carnegie Mellon University.

Finer-Moore,J.S., Kossiakoff,A.A., Hurley,J.H., Earnest,T. and Stroud,R.M. (1992) *Proteins*, **12**, 203–222.

Frishman,D. and Argos,P. (1996) *Protein Engng.*, **9**, 133–142.

Goldberg,M.E., Semisotov,G., Friguet,B., Kuwajima,K., Ptitsyn,O. and Sugai,S. (1990) *FEBS Lett.*, **263**, 51–56.

Goodfellow,J.M. and Vovelle,F. (1989) *Eur. Biophys. J.*, **17**, 167–172.

Goodford,P.J. (1985) *J. Med. Chem.*, **28**, 849–857.

Helms,V. and Wade,R.C. (1995) *Biophys. J.*, **69**, 810–824.

Hirst,J.D. and Sternberg,M.J. (1991) *Protein Engng.*, **4**, 615–623.

Hobohm,U. and Sander,C. (1994) *Protein Sci.*, **3**, 522–524.

Holbrook,S.R., Muskal,S.M. and Kim,S.-H. (1990) *Protein Engng.*, **3**, 659–665.

Holley,L.H. and Karplus,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 152–156.

Hubbard,S.J. and Argos,P. (1994) *Protein Sci.*, **3**, 2194–2206.

Hubbard,S.J., Gross,K.H. and Argos,P. (1994) *Protein Engng.*, **7**, 613–626.

Hummer,G., Garcia,A.E. and Soumpasis,D.M. (1995) *Biophys. J.*, **68**, 1639–1652.

Jordan,M.I. (1986) Attractor dynamics and parallelism in a connectionist sequential machine. In *Proc. of the Eighth Annual Conference of the Cognitive Society*, pp. 531–546. Erlbaum, Hillsdale, N.J.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Kuhn,L.A., Siani,M.A., Pique,M.E., Fisher,C.L., Getzoff,E.D. and Tainer,J.A. (1992) *J. Mol. Biol.*, **228**, 13–22.

Levitt,M. and Park,B.H. (1993) *Structure*, **1**, 223–226.

Loris,R., Stas,P.P. and Wyns,L. (1994) *J. Biol. Chem.*, **269**, 26722–26733.

Madhusudan, Kodandapani,M.R. and Vijayan,M. (1993) *Acta Crystallographica, Section D (Biological Crystallography)*, **D49**, 234–245.
Matthews,B.W. (1975) *Biochem. Biophys. Acta*, **405**, 442–451.
Meyer,E. (1992) *Protein Sci.*, **1**, 1443–1462.
Morris,A.S., Thanki,N. and Goodfellow,J.M. (1992) *Protein Engng.*, **5**, 717–728.
Otting,G., Liepinsh,E. and Wuthrich,K. (1991) *Science*, **254**, 974–979.
Pedersen,J.T., Olsen,O.H., Betzel,C., Eschenburg,S., Branner,S. and Hastrup,S. (1994) *J. Mol. Biol.*, **242**, 193–202.
Pitts,W.R. and Goodfellow,J.M. (1991) *Protein Engng.*, **4**, 531–537.
Pitts,W.R., Murray-Rust,J. and Goodfellow,J.M. (1993) *J. Comp. Chem.*, **14**, 1007–1018.
Poornima,C.S. and Dean,P.M. (1995) *J. Comput. Aided Mol. Des.*, **9**, 521–531.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.
Quiocho,F.A., Wilson,D.K. and Vyas,N.K. (1989) *Nature*, **340**, 404–407.
Rand,R.P. and Fuller,N.L. (1992) *Biophys. J.*, **61**, A345.
Riis,S.K. and Krogh,A. (1996) *Comp. Biol.*, **3**, 163–183.
Roe,S.M. and Teeter,M.M. (1993) *J. Mol. Biol.*, **229**, 419–427.
Rose,G.D., Geselowitz,A.R., Lesser,G.J., Lee,R.H. and Zehfus,M.H. (1985) *Science*, **229**, 834–838.
Rost,B. and Sander,C. (1993) *J. Mol. Biol.*, **232**, 584–599.
Rost,B. and Sander,C. (1994a) *Proteins*, **19**, 55–72.
Rost,B. and Sander,C. (1994b) *Proteins*, **20**, 216–226.
Roux,B., Nina,M., Pomes,R. and Smith,J.C. (1996) *Biophys. J.*, **71**, 670–681.
Shrake,A. and Rupley,J.A. (1973) *J. Mol. Biol.*, **79**, 351–371.
Sreenivasan,U. and Axelsen,P.H. (1992) *Biochemistry*, **31**, 12785–12791.
Thanki,N., Thornton,J.M. and Goodfellow,J.M. (1990) *Protein Engng.*, **3**, 495–508.
Thanki,N., Umrania,Y., Thornton,J.M. and Goodfellow,J.M. (1991) *J. Mol. Biol.*, **221**, 669–691.
Vedani,A. and Huhta,D.W. (1991) *J. Am. Chem. Soc.*, **113**, 5860–5862.
Vriend,G. (1990) *J. Mol. Graph.*, **8**, 52–56.
Wade,R.C. (1990) *J. Comp. Aided Mol. Des.*, **4**, 199–204.
Wade,R.C., Bohr,H. and Wolynes,P.G. (1992) *J. Am. Chem. Soc.*, **114**, 8284–8285.
Wade,R.C. and Goodford,P.J. (1993) *J. Med. Chem.*, **36**, 148–156.
Wade,R.C., Mazor,M.H., McCammon,J.A. and Quiocho,F.A. (1990) *J. Am. Chem. Soc.*, **112**, 7057–7059.
Walshaw,J. and Goodfellow,J.M. (1993) *J. Mol. Biol.*, **231**, 392–414.
Williams,M.A., Goodfellow,J.M. and Thornton,J.M. (1994) *Protein Sci.*, **3**, 1224–1235.
Zhang,L. and Hermans,J. (1996) *Proteins*, **24**, 433–438.
Zhang,X.-J. and Matthews,B.W. (1994) *Protein Sci.*, **3**, 1031–1039.

Received July 9, 1997; revised October 2, 1997; accepted October 9, 1997

Appendix 1.

The following protein datasets were used (Brookhaven protein databank codes are given).

Training set: 40 proteins were divided into the following 5 groups for cross-validation.

Group 1: 1alc, 1cho, 1ctf, 1gcr, 1mba, 1paz, 1rdg, 1sn3
Group 2: 1snc, 1srn, 1ubq, 1utg, 1ypi, 256b, 2act, 9pap
Group 3: 2cdv, 2cga, 2fbj, 2gbp, 2hhb, 2ltn, 2ovo, 2sec
Group 4: 2wrp, 3app, 3c2c, 8dfr, 3rnt, 4dfr, 4ins, 4pep
Group 5: 4pti, 5cpa, 5cpv, 5cyt, 5pcy, 5rxn, 3dfr, 2aza

The sequence identity between proteins in different groups is less than 33% except for two proteins, 2act and 9pap, which are both in group 2 and are 50% sequence identical.

Testing set: 121p, 131l, 153l, 193l, 1amp, 1arb, 1asu, 1bam, 1bp2, 1chd, 1csh, 1cus, 1cyo, 1dts, 1dyr, 1eca, 1ede, 1fkj, 1fnc, 1gof, 1gpb, 1gpr, 1hmt, 1iae, 1ilk, 1knb, 1lis, 1mrj, 1nar, 1nfp, 1nif, 1onc, 1osa, 1pbe, 1pbp, 1pda, 1pgs, 1phg, 1ppn, 1ptx, 1rcf, 1rec, 1rsy, 1sat, 1sbp, 1scs, 1tca, 1thv, 1thx, 1tml, 1tph, 1vhh, 1xnb, 2abk, 2acq, 2ayh, 2cba, 2cpl, 2ctc, 2dri, 2end, 2gdm, 2hbg, 2hts, 2mnr, 2phy, 2por, 2prk, 3chy, 3cla, 3cox, 3grs, 3pte, 3tgl, 4enl, 7rsa, 8abp (77 proteins)