PROTEINS: Structure, Function, and Genetics 50:629–635 (2003)

Real Value Prediction of Solvent Accessibility From Amino Acid Sequence

Shandar Ahmad,[1*] M. Michael Gromiha,[1,2] and Akinori Sarai[1]
[1]*RIKEN Tsukuba Institute, Ibaraki, Japan*
[2]*Computational Biology Research Center (CBRC), AIST, Tokyo, Japan*

ABSTRACT **The solvent accessibility of amino acid residues has been predicted in the past by classifying them into exposure states with varying thresholds. This classification provides a wide range of values for the accessible surface area (ASA) within which a residue may fall. Thus far, no attempt has been made to predict real values of ASA from the sequence information without a priori classification into exposure states. Here, we present a new method with which to predict real value ASAs for residues, based on neighborhood information. Our real value prediction neural network could estimate the ASA for four different nonhomologous, nonredundant data sets of varying size, with 18.0–19.5% mean absolute error, defined as per residue absolute difference between the predicted and experimental values of relative ASA. Correlation between the predicted and experimental values ranged from 0.47 to 0.50. It was observed that the ASA of a residue could be predicted within a 23.7% mean absolute error, even when no information about its neighbors is included. Prediction of real values answers the issue of arbitrary choice of ASA state thresholds, and carries more information than category prediction. Prediction error for each residue type strongly correlates with the variability in its experimental ASA values. Proteins 2003;50:629–635.**

Key words: structure prediction; accessible surface area; neural network

INTRODUCTION

Solvent accessibility is a key property of amino acid residues, important for both the structure and function of proteins. From the point of view of protein structure, accurate prediction of solvent accessibility would be expected to improve the otherwise slow progress being made in accurately predicting the three-dimensional structures of proteins with known sequences, especially those lacking homology to other proteins with solved structures. Consequently, predictions of secondary structure and solvent accessibility are being actively pursued in an effort to bridge the gap between sequence and structure.[1–7] Thus far, however, all attempts to predict solvent accessibility of amino acid residues have been focussed on predicting locations of accessibility "states" (buried or exposed) with somewhat arbitrary choices of state "thresholds." The most successful methods for predicting accessible surface areas (ASAs) have been able to achieve 70–75% best classification capability, but it is not uncommon for mutually contradictory predictions to be made when the same method and model are applied to more than one threshold e.g., whereas a residue may be predicted to be buried based on a 5% threshold, the same residue may be predicted to be exposed based on a 25% threshold. Furthermore, the very act of state classification of accessible areas means that a subsequently developed model relies upon less information than is actually available from the structural data. We, therefore, feel that the use of state thresholds may be abandoned; instead, real value predictions should be made, utilizing all the ASA information available from the structural data. In our previous work, we showed that a simple linear neural network could be used to predict ASA categories with comparable efficiency, and developed an online server to make use of our predictions.[8] Here, we present a mechanism by which real value ASA predictions can be made. This is done by using a similar multilayer neural network and retailoring it to encode for real value outputs and defining an error function that could be trained for development of the resultant model. We also determined a base line for such predictions by developing what may be called an average value assignment method. Finally, we evaluated the variation in our prediction error, taking into consideration window size, residue position, residue type, and data size.

METHODS

Data Selection

Four different sets of protein structures have been used. This included the frequently referred to 126 protein data set of Rost and Sander,[1] a data set of 215 proteins used in our previous study[8] and by Manesh et al.,[5] a data set of 338 monomeric proteins used by Carugo in a related study,[9] and 502 proteins from the Cuff and Barton[4] data set of 513 proteins, selected by removing some sequences, e.g., those with less than 30 residues. Each of these data sets

Abbreviations: ASA, accessible surface area; MAE, mean absolute error.

*Correspondence to: Shandar Ahmad, RIKEN Tsukuba Institute, 3-1-1, Koyadai, Tsukuba, Ibaraki 305 0074, Japan. E-mail: shandar@rtc.riken.go.jp

Received 25 March 2002; Accepted 10 October 2002

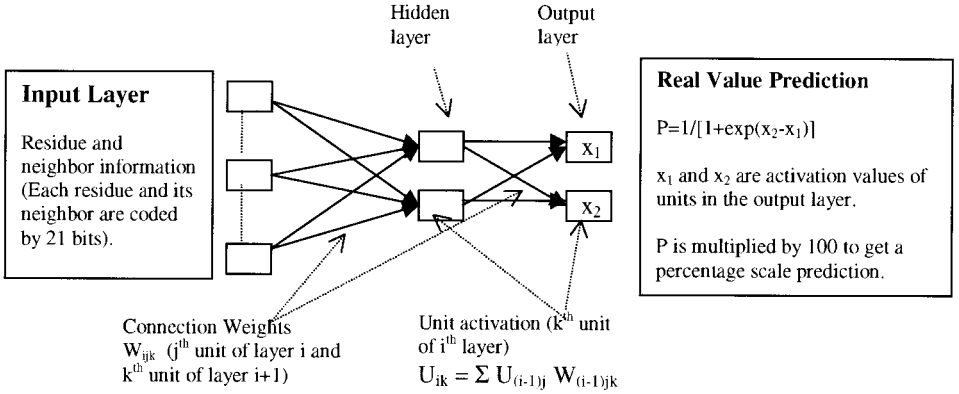

Fig. 1. Real value prediction neural network to map a binary space from input layer to a percentage scale finite space.

consisted of protein sequences with less than 25% homology. The number of residues in these data sets ranges from about 23,000 in Rost and Sander's (126 proteins) data, to more than 83,000 residues in Cuff and Barton's (502 proteins) data. These data sets will be referred to as RS-126, Manesh-215, Carugo-338, and Barton-502, respectively, in the subsequent discussion. Training and validation cycles were carried out on each of these data sets separately, to examine any possible effect of data size on the prediction quality, such that no two sequences from validation, training and test data sets have more than 25% homology. Most of the discussion in this paper refers to Manesh-215 data, unless otherwise mentioned.

Computation of Solvent Accessibility

Solvent accessibility (%) was defined as the ratio between the solvent ASA of a residue within a three-dimensional structure and that in an extended tripeptide (Ala-X-Ala) conformation. Absolute values of ASA in all the data sets except manesh-215 were obtained using the standard DSSP program.[10] The solvent ASA's for Manesh-215 data sets were calculated previously[8] as described elsewhere[11] using the ASC program[12] with the van der Waals radii given by Ooi et al.[13] These values of ASA were retained to examine any variation due to a change in ASA calculation algorithm. The extended state coordinates in ASC are computed using the ECEPP/2 algorithm[14] with the dihedral angles of Oobatake and Ooi.[15] These values of extended state ASA (in Å^2) are 110.2 (Ala), 144.1 (Asp), 140.4 (Cys), 174.7 (Glu), 200.7 (Phe), 78.7 (Gly), 181.9 (His), 185.0 (Ile), 205.7 (Lys), 183.1 (Leu), 200.1 (Met), 146.4 (Asn), 141.9 (Pro), 178.6 (Gln), 229.0 (Arg), 117.2 (Ser), 138.7 (Thr), 153.7 (Val), 240.5 (Trp), and 213.7 (Tyr), respectively.

Design of a Real Value Neural Network

Several investigators have developed multilayer feed-forward type neural networks for the prediction of secondary structure and solvent accessibility states.[16–18] Our final neural network consists of one input layer with 147 units (21 for each residue position, accommodating 3 neighbors on either side), one hidden and one output layer, each with two units, as shown in Fig. 1. Similar to our previously used linear neural network,[8] activation of a unit in any layer is simply a weighted linear sum of the activation signals of the previous layer. Thus, the input layer consists of binary codes of sequence information, and this information is fed-forward to the subsequent layers, according to the expression:

$$U_{ik} = \sum U_{(i-1)j} W_{(i-1)jk}$$

Where U_{ik} is the activation of unit k in the i^{th} layer, W_{ijk} denotes the connection weight linking j^{th} unit of layer "i" to k^{th} unit of layer "$i + 1$," and summation is carried out for all units in the previous layer.

In this way, if x_1 and x_2 are the activation of the two units in the final layer, the real value prediction of ASA is obtained by a transforming the difference $(x_2 - x_1)$, via a sigmoidal function as follows.

$$(ASA)_{pred} = P * 100\ \%, \qquad \text{where } P = 1/[1 + \exp(x_2 - x_1)] \tag{1}$$

The activation signals at the two units in the prediction layer may be interpreted as competing exposure and burial potentials, with the difference in their strengths ultimately determining the ASA of the residue being predicted. Training was carried out by presenting one weight at a time, selected randomly. Unit biases in the whole network were kept zero and no training of biases was carried out. Effect of temperature in the training process was incorporated by reinitializing a small number of weights after definite intervals. We are calling this design a real value prediction neural network (RVP-Net). To summarize, this network differs from what we used previously for category prediction as follows.

The number of units in the output layer of the binary prediction network was the same as the number of exposed states defined, e.g., three units for a three-state prediction.[8] To obtain an ASA state prediction, the total signals

TABLE I. Summary of the Prediction Error (MAE) and Correlation Obtained Using the RVP-Net and Average Assignment Methods for the Manesh-215 Data Set

	Training		Test		Validation	
Method	MAE (%)	Correlation	MAE (%)	Correlation	MAE (%)	Correlation
Real value prediction (RVP) neural network	17.6	0.5179	18.0	0.5011	18.0	0.5006
Average assignment	23.7	0.4554	23.7	0.4543	23.7	0.4543

received by the output layer units were transformed by the expression:

Prediction state = i, such that

$$x_i > x_j \qquad \forall\ 1 \le j \le n;\ i \ne j$$

Where x_i is the activation received by the i'th unit of the output layer and n is the total number of such units. In the present network, there are always only two units in the output layer and their activation signals x_1 and x_2, are transformed according to equation (1) to produce a real value output. Further, the error function, to be minimized in the category prediction work, is based on the single residue accuracy of prediction[8] (i.e., relative number of correct predictions in any ASA state). In the present network, this error function is actually the mean absolute error (MAE) of prediction defined as the absolute difference between the predicted and experimental (desired) values of relative ASA, per residue.

$$\text{MAE} = \sum |(\text{ASA})_{\text{pred}} - (\text{ASA})_{\text{exp}}|/N \qquad (2)$$

where summation is carried out for all residues and N is the total number of predictions.

Three approximately equal sets of non-overlapping data were created from each group of proteins (RS-126, Manesh-215, Carugo-338, Barton-502) and, respectively, designated as training, test, and validation data. Training of the first data set was carried out for a large number of training steps, even though there could be overlearning of the training data. While the training data is allowed to learn, test data error is continuously monitored. Weights are saved for every decrease in the test data prediction error. Thus, finally trained weights represent a stage in the learning process of the training data where test data experienced the minimum mean absolute error (MAE). A certain amount of training is, therefore, present in the test data. Keeping a third data set (validation data) out of the training process ensures that the prediction accuracy for this data represents the predictability of the network and is free from the biases caused by training/test data sets. All six possible combinations from these three data sets were used for training/test/validation. The final results are the averages of the accuracies obtained from the six cycles of training and prediction. Making all possible six combinations from the three sets will put each protein twice in any of the training/test or validation data. Finally reported values of accuracy/error are simply the arithmetic averages for the corresponding data sets, thus formed.

Prediction Error and Correlation

The MAE of percent relative accessibility as defined in equation (2) above has been used as the main indicator of prediction quality in the present work. Pearson's "r" has been reported at some places, and it is calculated as the ratio of the covariance between the predicted and experimental ASA values to the product of the standard deviations in the two. Single residue accuracy and correlation for the state predictions have been defined in related works.[8]

RESULTS AND DISCUSSION

Classifying the 215 high-resolution structures of nonhomologous proteins (Manesh-215) into training, test, and validation sets yielded predictions with MAE values of 17.6, 18.0, and 18.0%, respectively (Table I). Correlation of the experimental and predicted values yielded Pearson's r = 0.5179, 0.5011, and 0.5006, respectively.

As an example, Figure 2 shows the experimental and predicted values for each residue in thioredoxin (PDB code: 1ABA). Note the good linear relationship between the experimental and predicted values. For this protein, Pearson's r = 0.52, the MAE per residue is 22%, and several residues (C14, P16, N35, M37, L51, G78, and L82) are predicted with <2% prediction error. Interestingly, those residues fall within different ranges of ASA values, which is indicative of the high degree of accuracy of this prediction across a wide range of ASAs and amino acid residues. In this protein, about 70% of the residues are predicted with less than a 25% prediction error, despite the fact that these prediction values were obtained when the protein was not part of either the training or test data. Likewise, most of the discussion in this work relates to predictions in the so-called validation data.

Average Assignment Method and the Role of Neighbors

One of the simplest ways to estimate the ASA of individual residues in a protein may be to just look for its average ASA in other known proteins. We can evaluate an average ASA in an experimental data set and assign this average for a new residue as a "prediction," irrespective of its neighboring environment. For example, the average ASA for Asp (aspartic acid) residues in the data set was 40.9%, which could then be assigned to all Asp residues in the prediction data. This so-called "average assignment method" should serve as a quick benchmark of the predictive quality of other methods. In addition, the MAE of predictions made using average assignment has two impor-

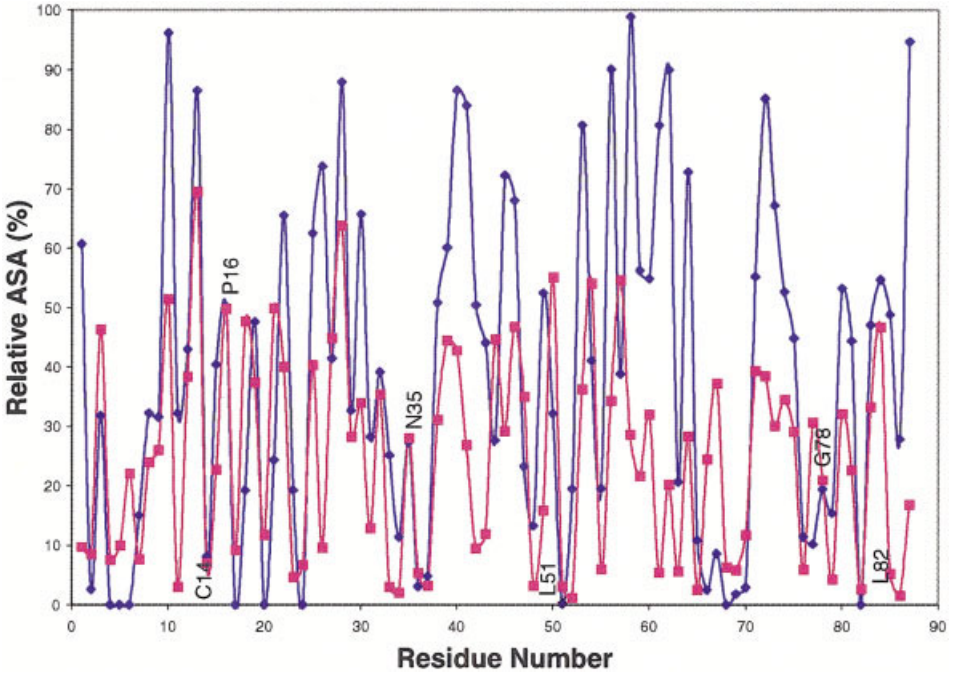

Fig. 2. Example of predicted (red) and experimental (blue) ASA values for a Protein (PDB code 1ABA). Some of the best predictions are labeled with their residue name and number.

tant properties. First, it is the maximum MAE allowed for any real value ASA prediction to be useful. Second, it does not involve information from neighbors; therefore, prediction error can be attributed to the residue itself, which is helpful for calculating the contribution made by neighbors to the final error reduction.

Using the above-mentioned protocol, we calculated the average ASA values in the training set proteins for all 20 residue types, and assigned these values to the corresponding residues in the test and validation data sets. Mean absolute error of prediction in such an assignment was found to be 23.7% for all three data sets (differing only in subsequent decimal places), with Pearson's $r = 0.4554$, 0.4543, and 0.4543, respectively (Table I). Thus, even assigning average ASA values to specific residues in the training data yields a fairly good estimate for the test data. By comparing this value of MAE with that obtained for a three-neighbor network, we could conclude that the neighbors contain information for about 5.7% ASA variation. By training the network using different window sizes, it was found that, of the 5.7% improvement in prediction contributed by neighbors, the largest share of 4% comes from the immediate neighbors, and subsequent neighbors contribute only 1.7% in the reduction of error. No significant improvement was achieved by including more than four neighbors. In that regard, it has been reported that the information from the immediate neighbor is usually sufficient to predict protein structural class with reasonable accuracy.[19] The final value of 18% MAE suggests that this information about solvent accessibility is determined by long-range contacts between the residues in a protein.

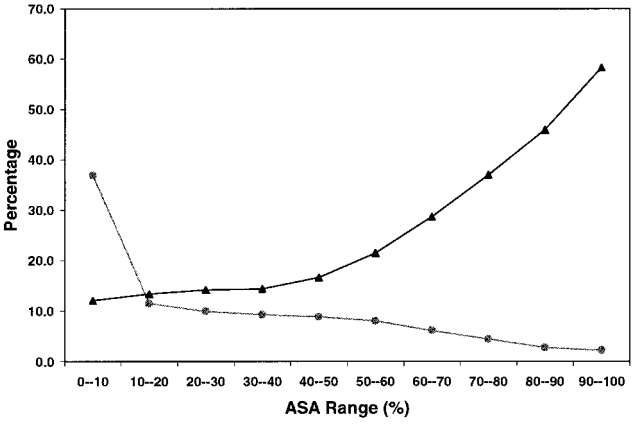

Fig. 3. Prediction error in various ASA ranges. As ASA values increase, so does prediction error (black) due to a corresponding fall in the relative abundance of data (gray).

Variation of Prediction Error With ASA Value

Figure 3 shows the variation in prediction error for different ranges of ASA. We found that the residues, with less than 10% exposure, where 37% of the overall data are located, were predicted within a 12% error. Similar errors were obtained for partially buried residues ($10\% < \text{ASA} < 50\%$), indicating that this method successfully predicts ASA values for 77% of residues within 14% error. Only 15% of residues were observed to be on the surface, and the prediction error was higher for them. This is likely attributable to two factors: (1) There are a very small number of residues in the most exposed ASA range; consequently,

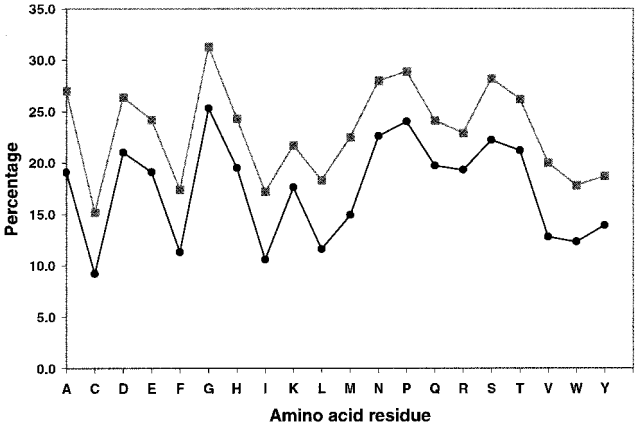

Fig. 4. Residue-specific prediction error and ASA variability. Dark circles represent the prediction error, and gray squares show the corresponding standard deviation in the experimental ASA for that residue type. A very high correlation (r = 0.97) is observed between the prediction error and standard deviation in the original data.

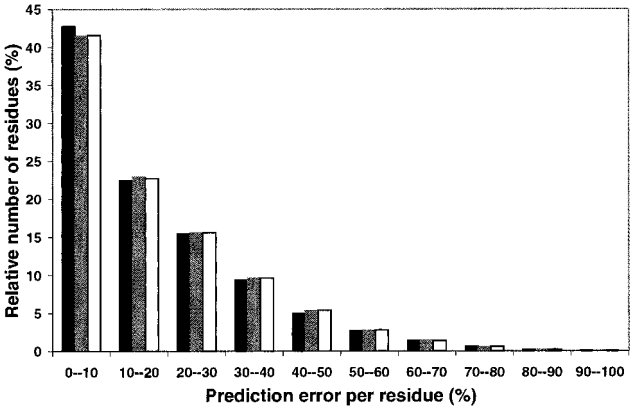

Fig. 5. Prediction error bar diagram showing the relative number of residues predicted within a given range of MAE in the training (black), test (gray) and validation (white) data sets.

their effect on the overall prediction accuracy is overshadowed by the accuracy in lower exposure values (Fig. 3), and (2) exposed residues are not as strictly conserved as the hydrophobic ones; consequently, information about the surrounding residues is necessary for better predictions. This idea is consistent with earlier analyses of protein stability that indicate structural information to be very important for predicting the stability of exposed mutations, while residue information is sufficient for mutations that are buried.[11]

Residue-Specific Variation in Prediction Error

Figure 4 shows the prediction errors we obtained for each of the 20 amino acid residues. Figure 4 also shows the variability of ASA in the overall data, represented by the standard deviation of the ASA values. The prediction curve and the deviations show an excellent correlation (r = 0.97), indicating ASA training to be uniform for all residues. The actual prediction quality varies due to differences in the ASA variability of different residues, however. We found that prediction of the ASA for Cys residues is within a 10% mean prediction error, which may be due to the fact that Cys residues are usually present in the interior of a protein (mean ASA for Cys in the overall data is nearly 10.2%). Interestingly, all of the aromatic residues were predicted with less than a 15% mean prediction error, perhaps because these residues have a relatively strong affinity for other residues, thereby increasing protein stability,[20] and most are located in buried and partially buried regions.[21] Prediction errors, as low as 13%, were obtained for the hydrophobic residues Leu, Val, and Ile. As expected, because most reside on the surface, the deviation of the predicted values from the experimental ASAs was higher for charged residues, falling within a range of 15–25%. The most difficult predictions were for Gly, which, naturally, is attributable to its conformational flexibility and variability (Fig. 4). Thus, predictions for hydrophobic, aromatic, and sulfur-containing residues are more accurate than those for charged residues.

Distribution of Prediction Error

We next analyzed the percentage of residues that could be predicted within a specific range of error. The frequency bar diagram in Figure 5 shows the relative number of residues in each of the ASA ranges, and that the proportions were very similar in the training, test, and validation data sets. Interestingly, about half of the residues were predicted with less than 10% error, and about two thirds were predicted with better than average values of MAE (18.0%); only a few residues were found to have a significantly high prediction error. Taken together, these findings demonstrate the reliability of our method and its utility for estimating solvent accessibility of proteins with unknown structures.

Validation of Results With Different Data Sets

To examine the effect of data size and more general applicability of this method, we have repeated the whole process of training, test, and validation for four independent data sets, as mentioned in Methods. Results for all these data sets are presented in Table II. It is observed that the overall variation in MAE of prediction for the validation data sets varies from 18.0% for the Manesh-215 data set to 19.5% for the RS-126. A very small difference of 1.5%, and strikingly similar values for the other two data sets, indicate the general convergence of prediction MAE to these values. Likewise, the correlation varies from 0.472 to 0.501, which is again a rather small range. Furthermore, as MAE for Carugo-338 is more than Barton-502, there does not seem to be a direct correspondence between the data size and MAE, in this range. Also, we used program ASC[12] to calculate ASA for Manesh-215 and DSSP[10] for all other data sets and found that the MAE and correlation are similar in all these cases. Detailed results of all these predictions for any further analysis can be downloaded from http://www.rtc.riken.go.jp/~shandar/rvp-net/.

Comparison With Other Methods

We could not locate any other method of real value prediction of ASA from sequence, and therefore a direct

TABLE II. Real Value Prediction Results for Different Data Sets

	Training		Test		Validation	
Data set	MAE (%)	Correlation	MAE (%)	Correlation	MAE (%)	Correlation
RS-126	18.8	0.5024	19.4	0.4769	19.5	0.4718
Carugo-338	18.9	0.4953	19.0	0.4901	19.0	0.4870
Barton-502	18.7	0.4876	18.8	0.4822	18.8	0.4800

comparison with any of the existing methods was not possible. The most popular prediction servers such as Jnet and PHD predict ASA states rather than ASA value.[1,4,22] As an approximate comparison, however, we used the values of our real ASA predictions to classify residues into two ASA states in the traditional sense at a 16% threshold. The two-state ASA prediction accuracy and correlation for such a classification were calculated for all our predictions. We observed that the prediction accuracy of (71%) this classification solely with sequence information is very similar to values previously reported by other authors for the category prediction.[18] It may, however, be pointed out that the final accuracy values of PHD or Jnet in terms of binary classification remain better than what we obtain by postprediction classification. This is due to the fact that our predictions are based on sequence information only, whereas these servers have included alignment and evolutionary information into their predictions. For example, RS-126 with PHD shows nearly 75% accuracy,[18] as against 71% we obtained here. In the same work, however, prediction accuracy from sequence information only was reported to be 70%, which is similar to what we get in the present method. Furthermore, Rost and Sander[18] reported that the best correlation (0.432) from single sequence information was obtained for a 10-state network. Our real value prediction correlation (0.472) for the same data set is significantly better than this value. Hence, we suggest that our real value predictions contain more than all the necessary information needed to classify residues in the ASA states.

It will also be interesting to compare the MAE scale to binary prediction accuracy scores. For example, consider the extreme case when a binary prediction provides 100% correct classification, for a state threshold t. Predicting the residue to be buried or exposed now means that the ASA of the residue is either 0% to t (buried) or t to 100% (exposed). We assume there are an equal number of residues in buried and exposed states. In terms of real value of ASA, this implies an average 25% uncertainty for all residues, which is the equivalent of the MAE scale we are using in the present work. Thus, from this perspective, our real value predictions with 19% MAE may be interpreted to have more information (about actual ASA value) than even a 100% correct binary classification.

CONCLUSIONS

We have developed a new method for predicting the exact values of solvent accessibility using a real value neural network. The main aspect of this work is that we do not classify residues into buried and exposed states with arbitrary ASA thresholds at the prediction end. However, if one does wish to have a broad idea of accessibility in terms of categories, our method still allows a post-prediction classification with accuracy similar to pre-prediction classification methods. In addition, it would allow setting thresholds at will and even choosing different thresholds for different residues within the same prediction, which may be desirable because all residues do not have the same average value of ASA and a common threshold for all residues may be difficult to justify. Our predictions on the best predicted data set showed 18% MAE on cross validated data sets, with residues in buried and partially buried regions being better predicted than those on the surface. In addition, the variation in prediction error with respect to residue type strongly correlates with variability in the experimental data. Thus, the real value prediction method presented here should be very useful for predicting ASAs of protein residues in any protein, and may, in turn, be used to improve algorithms for predicting protein structure.

REFERENCES

1. Rost B, Sander C. Improved prediction of protein secondary structure by using sequence profiles and neural networks. Proc Natl Acad Sci 1993;90:7558–7562.
2. Benner SA, Geroff DL, Rozzel JD. Protein structure prediction. Science 1996;274:1448–1449.
3. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
4. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
5. Manesh HN, Sadeghi M, Arab S, Movahedi AM. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
6. Chang I, Cieplak M, Dima RI, Maritan A, Banavar JR. Protein threading by learning. Proc Natl Acad Sci 2001;98:14350–14355.
7. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
8. Ahmad S, Gromiha MM, NETASA: Neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
9. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607–609.
10. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bond and geometrical features. Biopolymer 1983;22:2577–2637.
11. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information for predicting protein stability changes: comparison between buried and partially buried mutations. Protein Eng 1999;12:549–555.
12. Eisenhaber F, Argos P. Improved strategy in analytical surface calculation for molecular system- handling of singularities and computational efficiency. J Comp Chem 1993;14:1272–1280.
13. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface area as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci 1987;84:3086–3090.
14. Momany FA, McGuire RF, Burgess AW, Scheraga HA. Energy parameters in polypeptides 7. Geometric parameters, partial

atomic charges, non bonded interactions, hydrogen bond interactions and intrinsic torsional potentials for naturally occurring amino acids. J Phys Chem 1975;79:2361–2381.
15. Oobatake M, Ooi T. Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol 1993;59:237–284.
16. Holey HL, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci 1989;86:152–156.
17. Chandonia J, Karplus M. Neural network for secondary structure and structure class prediction. Protein Science 1995;4:275–285.
18. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
19. Kumarevel TS, Gromiha MM, Ponnuswamy MN. Structure class prediction: an application of residue distribution along the sequence. Biophys Chem 2000;88:81–101.
20. Kannan N, Vishveshwara S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng 2000;13: 753–761.
21. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A. ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. Nucl Acid Res 2002;30:301–302.
22. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:197–205.