Real Value Prediction of Solvent Accessibility in Proteins Using Multiple Sequence Alignment and Secondary Structure

Aarti Garg, Harpreet Kaur, and G.P.S. Raghava*
Institute of Microbial Technology, Sector-39A, Chandigarh, India

ABSTRACT **The present study is an attempt to develop a neural network-based method for predicting the real value of solvent accessibility from the sequence using evolutionary information in the form of multiple sequence alignment. In this method, two feed-forward networks with a single hidden layer have been trained with standard back-propagation as a learning algorithm. The Pearson's correlation coefficient increases from 0.53 to 0.63, and mean absolute error decreases from 18.2 to 16% when multiple-sequence alignment obtained from PSI-BLAST is used as input instead of a single sequence. The performance of the method further improves from a correlation coefficient of 0.63 to 0.67 when secondary structure information predicted by PSIPRED is incorporated in the prediction. The final network yields a mean absolute error value of 15.2% between the experimental and predicted values, when tested on two different nonhomologous and nonredundant datasets of varying sizes. The method consists of two steps: (1) in the first step, a sequence-to-structure network is trained with the multiple alignment profiles in the form of PSI-BLAST-generated position-specific scoring matrices, and (2) in the second step, the output obtained from the first network and PSIPRED-predicted secondary structure information is used as an input to the second structure-to-structure network. Based on the present study, a server SARpred (http://www.imtech.res.in/raghava/sarpred/) has been developed that predicts the real value of solvent accessibility of residues for a given protein sequence. We have also evaluated the performance of SARpred on 47 proteins used in CASP6 and achieved a correlation coefficient of 0.68 and a MAE of 15.9% between predicted and observed values. Proteins 2005; 61:318–324.** **© 2005 Wiley-Liss, Inc.**

Key words: solvent accessibility; prediction; real value; neural network; multiple alignment; secondary structure

INTRODUCTION

Protein secondary structure prediction is an intermediate step in tertiary structure prediction. In addition to secondary structure, prediction of solvent-accessible surface area (ASA) of residues also helps to understand the complete three-dimensional structure of a protein. Further, ASA is also considered as a key factor in protein folding, because the burial of core residues (hydrophobic residues) is a major driving force for folding.[1] It has been shown that in proteins, the hydrophobic free energies are directly related to ASA of both polar and nonpolar groups.[2] In addition to these, ASA information also resulted in the improvement of prediction of protein subcellular localization, as distribution of surface residues of a protein is correlated with its subcellular environments.[3] Because active sites and hydrations sites of a protein are located on its surface, accurate prediction of the surface residues may also be considered as an important step toward determining its functions and conformational changes.[4] Thus, there is a need to develop an accurate and better method for predicting the solvent accessibility of residues from a protein sequence.

To date, a number of methods have been developed for predicting the surface accessibility of residues based on different databases and computational techniques such as neural networks,[5–7] Bayesian statistics,[8] a multiple linear regression method,[9] a knowledge-based prediction model,[10] information theory,[11] and support vector machines.[12] Most of these methods have been concentrated on two states (buried or exposed) prediction, that is, whether a residue in a given sequence is exposed or buried. The accuracy reported by these methods for two-state prediction is between 72 and 88%.[5–12] Because the predicted state of a residue depends highly on the selected cutoff/threshold, these methods are likely to be less accurate because of the noise introduced due to arbitrary choice of selected thresholds.

To overcome this problem, Ahmad et al.[13,14] developed a method, RVP-net, for predicting the real value of solvent accessibility. The method of RVP-net is a neural network-based method, which predicts the real value of relative solvent accessibility (RSA) with mean absolute error (MAE) of 18–19.5% and correlation coefficient of 0.47–0.50. Recently, a neural network-based regression method, that is, SABLE,[15] has been published, which has a MAE of

Grant sponsor: Council of Scientific and Industrial Research (CSIR); Grant sponsor: Department of Biotechnology (DBT), Govt. of India.

*Correspondence to: G. Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India. E-mail: Raghava@imtech.res.in

Received 20 October 2004; Revised 25 February 2005; Accepted 28 February 2005

Published online 16 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20630

© 2005 WILEY-LISS, INC.

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15.3–15.8% and a correlation coefficient of 0.64–0.67. In addition, the Support Vector Regression-based method, described by Yuan and Huang, achieved a correlation coefficient of 0.66 between the predicted and observed ASA.[16]

In the past, it has been demonstrated that the accuracy of a secondary structure (both regular and irregular) prediction can be improved if we use multiple-sequence alignment (MSA) information. It has also been shown that the prediction accuracy of tight turns can be improved using predicted secondary structure information.[17–20] Based on these observations, a systematic attempt has been made in the present study to improve the prediction accuracy of a real value of solvent accessibility.

MATERIALS AND METHODS

The Datasets

Two nonhomologous datasets of proteins with sequence homology less than 25% have been used. One dataset is comprised of 215 proteins, which was also used earlier by Manesh et al.[11] for ASA prediction. The other dataset is comprised of 502 proteins, obtained from a dataset of 513 proteins[6] by removing those sequences, which have less than 29 residues. These two datasets have also been used earlier by Ahmad et al.[13] for the real value prediction of solvent accessibility. Throughout the study, these datasets have been referred to as Manesh-215 and CB-502, respectively.

Independent Dataset

To evaluate the performance of existing methods and the method developed in the present study, we created an independent dataset of CASP6 proteins, which were neither used during training nor during testing of any method. Originally, it contained 63 proteins, but we have removed 16 structures (containing chains), and used the remaining 47 proteins (http://www.imtech.res.in/raghava/sarpred/supp.html) for the analysis of the present method and evaluation of existing methods such as ACCpro,[21] Jnet,[22] PHDacc,[23] RVPnet,[13] and SABLE.[15]

Calculation of RSA

RSA is a normalized value between 0 and 1. It is calculated as the ratio between the solvent ASA of a residue within a three-dimensional structure and that of an extended tripeptide (Ala-X-Ala) conformation as shown in Equation (1).

$$\text{Relative solvent accessibility} = \frac{\text{ASA in a three-dimensional structure}(\text{Å}^2)}{\text{ASA in an extended tripepetide}(\text{Å}^2)} \quad (1)$$

The absolute values of solvent ASA have been obtained from the standard DSSP program[24] for CB-502 and independent datasets and Ahmad et al.,[13] for Manesh-215 datasets, respectively. The extended state ASA values (in Å^2) were the same as mentioned by Ahmad et al.[13]

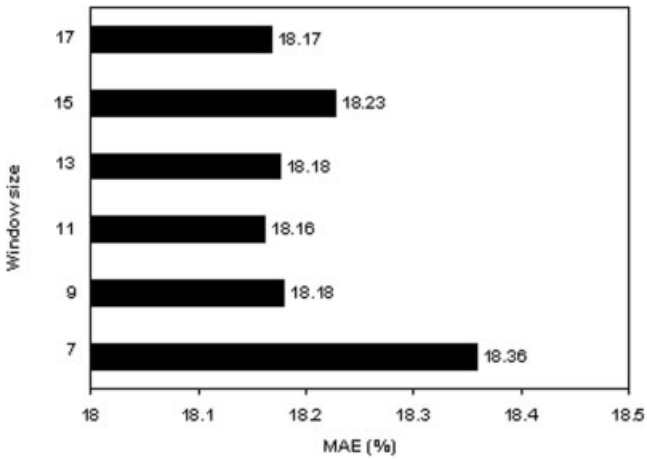

Fig. 1. MAE for different window sizes using a single sequence as an input to the neural network.

Crossvalidation

The evaluation of the prediction method is often carried out by a jackknife or crossvalidation technique.[25] In present study, due to the size of the datasets, the jackknife method (individual testing of each protein in the datasets) was not possible. So, a more limited crossvalidation technique was used, in which the dataset was randomly divided into subsets each containing an equal number of proteins. These subsets were further divided into training, validation, and testing sets.

Training, validation, and testing have been carried out on each of the two datasets (Manesh-215 and CB-502) separately. For the Manesh-215 dataset, fivefold crossvalidation was carried out by dividing the dataset into five equal sets. Three sets have been used for training, one set for validation, and one for testing. The process is repeated five times using a distinct test set each time. Similarly, for the CB-502 dataset, sevenfold crossvalidation has been carried out by dividing the dataset into seven equal sets. The final prediction results have been averaged over the number of subsets.

Neural Network Architecture

Two feed-forward neural networks with a single hidden layer have been used. The performance of the network has been assessed using different window sizes of 7 to 17 residues (Fig. 1). It has been found that an 11-residue window showed a minimum MAE (18.16%) in comparison to other window sizes after a fivefold crossvalidation. Thus, throughout this study, a window of 11 residues, that is, 21 × 11 input vectors, and 10 hidden nodes in a single hidden layer has been used for training and testing the network. ASA has been encoded as relative solvent accessibility values in the range between 0 and 1. The target output consists of a single output value (between 0 and 1) of the central residue in the input pattern.

For the neural network implementation and to generate the neural network architecture, the publicly available free simulation package SNNS, version 4.2, from Stuttgart

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

University has been used.[26] It allows incorporation of the resulting networks into an ANSI C function for use in a stand-alone code. A logistic activation function is used. At the start of each simulation, the weights are initialized with the random values. The training has been carried out using error back-propagation with a sum of square error functions as well as a mean square error function.[27] The learning parameter has been set to 0.001. The magnitude of the error sum in the test and training set is monitored in each cycle of the training. The ultimate numbers of cycles are determined where the network for the single sequences as well as for the multiple alignment profiles during training converges. The overall architecture of the two networks is given below:

First Network: Sequence-to-Structure Net

The input to the first sequence-to-structure network is either a single sequence or multiple alignment profiles. A window of 11 residues has been used, in which a prediction is made for the central residue. Single sequences are encoded as binary bits (0 or 1), and multiple alignment profiles are a PSI-BLAST–obtained position-specific scoring matrix (PSSM).

Second Network: Structure-to-Structure Net

The input to the second structure-to-structure network is predictions obtained from the first net and the predicted secondary structure. Four units encode each residue, where one unit codes for predicted output (between 0 and 1) from first network and the remaining three units code for three secondary structure states (helix, strand, and coil). Secondary structure information is encoded by the actual probabilities of three states provided in the output of the PSIPRED prediction. The probabilities are just the strengths of the prediction for each of the three target states (helix, strand, and coil), and are represented by a real number in the range between 0 and 1.

Multiple Alignment or Position-Specific Scoring Matrices

PSIPRED uses PSI-BLAST to detect homologs of a query sequence against a nonredundant (NR) dataset available at NCBI. After three iterations, it generates a position-specific scoring matrix having the highest score as a part of the prediction process, and here, we have used these intermediate PSI-BLAST–generated position-specific scoring matrices as a direct input to the first level network. The matrix has $21 \times M$ elements, where M is the length of the target sequence, and each element represents the frequency of occurrence of each of the 21 amino acids at one position in the alignment.[28]

Performance Measures

The performance of the method has been assessed for the real value prediction of solvent accessibility as well as for two states (exposed and burried) prediction.

For real value prediction of solvent accessibility, two parameters have been used:

1. MAE is defined as the absolute difference between the predicted and experimental (observed) values of relative solvent accessibility, per residue.

$$\mathrm{MAE} = \frac{\Sigma\,|(\mathrm{RSA})\mathrm{pred} - (\mathrm{RSA})\mathrm{exp}\,|}{N} \qquad (2)$$

where N is the total number of predictions.

2. Pearson's correlation coefficient (r) is the ratio of the covariance between the predicted and experimental values of relative solvent accessibility to the product of the standard deviations in the 2.

$$\mathrm{Pearson's}(r) = \frac{\Sigma XY - \dfrac{\Sigma X \Sigma Y}{N}}{\sqrt{\left(\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}\right)\left(\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}\right)}} \qquad (3)$$

where, X and Y are experimental and predicted values of relative solvent accessibility, respectively.

For assessing the quality of two state predictions, threshold-dependent measures have been used. These measures have been derived from the four scalar quantities: p (the number of correctly predicted exposed residues), n (the number of correctly predicted burried residues), o (the number of burried residues incorrectly predicted as exposed), and u (the number of exposed residues incorrectly classified as burried).

The following measures have been calculated:

$$\mathrm{Accuracy} = \frac{p+n}{t} \qquad (4)$$

$$\text{Matthews's correlation coefficient (MCC)} = \frac{pn - uo}{\sqrt{(p+u)(p+o)(n+u)(n+o)}} \qquad (5)$$

where $t = p + n + o + u$ is the total number of residues. Accuracy has been expressed in percentages.

RESULTS AND DISCUSSION

Prediction Using a Single Sequence

A neural network has been trained and tested on both of the datasets where amino acids in binaries (0 or 1) have been used as input. The performance of the network on both datasets is shown in Table I. It is clear from the results that the performance on the Manesh-215 dataset is marginally better then that of the CB-502 dataset. The averaged Pearson's correlation and MAE achieved on the Manesh-215 dataset using a single sequence was 0.53 and 18.2%, respectively. For the CB-502 dataset, a MAE between the experimental and predicted value is ~1% higher than the Manesh-215 dataset.

Prediction Using Multiple Alignments

To further enhance the prediction performance, we have employed the evolutionary information (in the form of multiple sequence alignment) for prediction. In this case, input to the network is position-specific scoring matrices

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE I. Performance of SNNS Using Single Sequence with and without Secondary Structure Information on Manesh-215 and CB-502 Data Sets

	Manesh-215		*CB-502*	
	First network	Second network	First network	Second network
Pearson's (r)	0.53 ± 0.01	0.61 ± 0.01 (0.61 ± 0.02)	0.52 ± 0.01	0.60 ± 0.01 (0.60 ± 0.01)
MAE (%)	18.2 ± 0.18	16.7 ± 0.20 (16.6 ± 0.16)	18.8 ± 0.19	17.4 ± 0.28 (17.4 ± 0.23)

Values in parentheses correspond to the prediction results obtained by excluding the proteins that were used to develop PSIPRED.

TABLE II. Performance of SNNS Using Multiple Sequence with and without Secondary Structure Information on Manesh-215 and CB-502 Data Sets

	Manesh-215		*CB-502*	
	First network	Second network	First network	Second network
Pearson's (r)	0.63 ± 0.01	0.67 ± 0.01 (0.68 ± 0.02)	0.62 ± 0.02	0.65 ± 0.01 (0.66 ± 0.01)
MAE (%)	16.0 ± 0.20	15.2 ± 0.22 (14.9 ± 0.47)	16.7 ± 0.42	15.9 ± 0.39 (15.9 ± 0.38)

Values in parentheses correspond to the prediction results obtained by excluding the proteins that were used to develop PSIPRED

obtained from PSI-BLAST instead of binaries. As shown in Table II, the correlation coefficient increases from 0.53 to 0.63 and 0.52 to 0.62 for Manesh-215 and CB-502 datasets, respectively, when MSA is used. The MAE decreases by approximately ~2% for both datasets. Thus, the improvement in prediction performance can be attributed to the use of MSA in the form of PSI-BLAST–generated profiles instead of single sequences.

Prediction Using Multiple Alignment and Secondary Structure Information

It is well established that there is a strong relationship between the secondary structure of a residue and its environment because proper consideration of solvent accessibility makes the prediction of secondary states more effective.[29] Hence, secondary structures and ASAs are related and are important features of proteins. If the use of solvent accessibility results in the high accuracy of secondary state prediction, then it may be possible that use of secondary states may result in the improvement of the performance of solvent accessibility prediction. Because PSIPRED is a highly accurate method for predicting the secondary states of proteins, in the present study an attempt has been made to utilize PSIPRED predicted secondary structure information as an input to the second structure-to-structure network along with the output of the first sequence-to-structure network. It has been found that incorporation of secondary structure information further improves the correlation coefficient from 0.63 (with MSA alone) to 0.67 and 0.62 (with MSA alone) to 0.65 for Manesh-215 and CB-502 datasets, respectively (Table II). Also, a corresponding decrease in MAE can be noticed for both datasets. We have also conducted training and testing of the second layer network with predictions obtained from the first layer network (without secondary structure) for the Manesh-215 dataset. After fivefold crossvalidation, the second network without secondary structure information was able to obtain a MAE of 15.9% and a correlation coefficient of 0.63, whereas a second network with secondary structure information achieved a MAE of 15.2% and a correlation coefficient of 0.67. Hence, we can say that use of predicted values from the first network along with secondary structure as an input to the second layer network has improved the performance. Moreover, it is evident from the comparison of Tables I and II that the prediction performance using both MSA and a secondary structure is better than using single sequence and secondary structure information. In addition, we have calculated MAE for helical, strand, and coil regions separately and achieved a MAE of 12.5, 10.9, and 19.1%, respectively. This demonstrates that prediction of solvent accessibility is more reliable in well-defined regular states (as obtained in less MAE) in comparison to the coil region.

Performance of the Method by Excluding Proteins Used to Develop PSIPRED

The results have been further crossvalidated by filtering out the proteins that were used to develop the PSIPRED method. The difference in the results is negligibly small (values in parentheses in Table I and II), which clearly indicates that the results are not biased by PSIPRED performance.

Two States Prediction

In the past, a number of methods have been developed for predicting the states (exposed or buried) of residues. Thus, we have also examined the performance of our method in the terms of two states predictions. We have assigned the state of a residue based on its predicted RSA values (%) and a chosen threshold. For instance, a threshold of 5% means the number of residues having an RSA value (%) greater than and equal to 5 is considered as exposed, and below 5, buried. The value of the threshold has been varied from 5 to 80%. It has been found that prediction accuracy varies from 74.9% at a threshold of 5% to 95.1% at a threshold of 80% for Manesh-215. However, for the CB-502 dataset, accuracy varies from 73.9 to 94.2%, respectively (Table III).

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE III. Results of Two-States Predictions, Obtained Using Multiple Sequence Alignment and Secondary Structure Information

	Manesh-215		CB-502		Independent data set	
Threshold (%)	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC
5	74.9	0.31	73.9	0.27	80.2	0.38
10	77.2	0.50	75.3	0.46	79.9	0.51
20	77.7	0.56	76.4	0.53	77.8	0.54
30	77.8	0.55	76.6	0.52	77.3	0.54
40	78.1	0.49	77.7	0.48	77.1	0.50
50	80.5	0.41	80.4	0.41	78.6	0.43
60	85.3	0.28	84.9	0.29	83.0	0.34
70	90.7	0.15	90.0	0.15	88.4	0.18
80	95.1	0.09	94.2	0.05	92.8	0.06

Acc, accuracy.

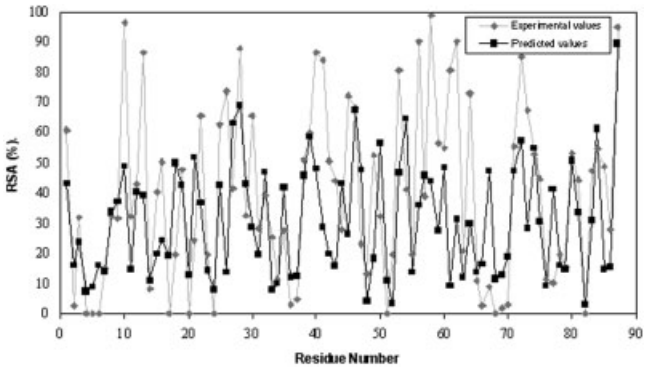

Fig. 2. Predicted and experimental values of (%) RSA of each residue for protein thioredoxin (1ABA).

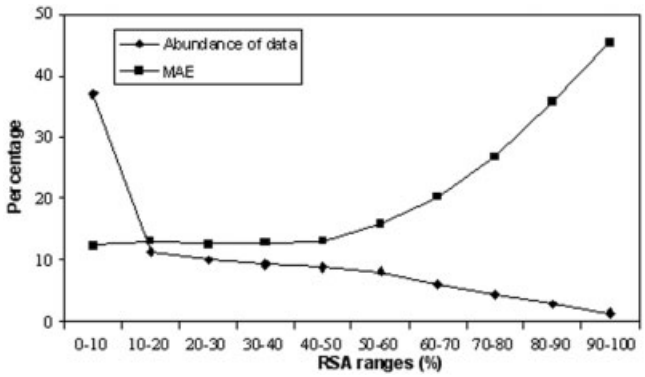

Fig. 3. MAE in various RSA(%) ranges for the Manesh-215 dataset.

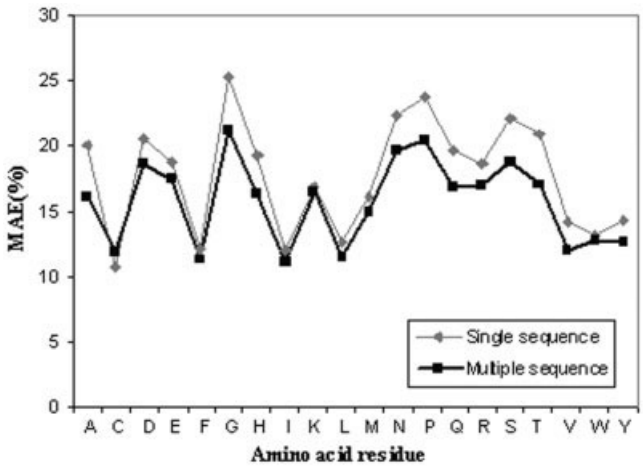

Fig. 4. Residue-specific prediction error obtained using single-sequence and multiple-sequence alignment for the Manesh-215 dataset.

An Example of Prediction of Protein Thioredoxin (1ABA)

To have an objective comparison of the method with Ahmad's RVP-net,[13,14] performance of the method has also been tested on the protein thioredoxin (1ABA) (Fig. 2). This is the same test protein as used by Ahmad et al.[13,14] for testing the RVP-net. For this protein, our method gives the correlation value $r = 0.57$ and a MAE of 19.2%, which is better then $r = 0.52$ and a MAE of 22% as reported by Ahmad et al.[13]

Variation of Prediction Error with a (%)RSA Range

We have also checked the variation in prediction error with respect to the different ranges of RSA values (%) as shown in Figure 3. It has been found that approximately 40% of the residues in the Manesh-215 dataset have an RSA range between 0 and 10%, and for this RSA range, the MAE is 12%. For the same RSA range (0–10%), Ahmad et al.[13] have also reported a MAE of 12%. As one moves from a lower to higher RSA range (60–100%), the number of residues decrease, with an increase in prediction error from 12 to ~45%. However, for the same RSA range, Ahmad et al. had reported a maximum error of 60%.[13] Hence, our present method can predict all the RSA ranges (0–100%) with less of a MAE compared to the RVP-net.[13,14]

Residue-Specific Prediction Error

The prediction errors obtained for each 20 types of amino acids using single and multiple sequence alignments is shown in Figure 4. Using a single sequence as input to the neural network, relative solvent accessibility of all the 20 residues was predicted with an error between 11 and 25%. The minimum error has been obtained for cysteine, while glycine shows a maximum error of 25%.

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE IV. Detailed Comparison of Our Present with Existing Methods for Two-States Classifications at Threshold of 25% on an Independent Data Set (47 CASP6 Proteins)

Methods	Accuracy (%)	MCC
SARpred	77.3	0.55
SABLE	78.0	0.56
ACCpro	73.9	0.48
Jnet	71.6	0.47
RVPnet	71.3	0.43

However, multiple sequence alignment profiles have reduced MAE further between 11 and 21%. Using MSA, a minimum prediction error of 11% has been obtained for hydrophobic amino acids such as leucine, valine, and isoleucine. A reduction of 2% in MAE has also been observed for most of the aromatic amino acids (such as phenylalanine, tyrosine, and tryptophan). Thus, multiple sequence alignment has made the residue-specific predictions more accurate compared to single sequence. However, cysteine residues appear to be better predicted with a single sequence-based approach, suggesting that short, disulphide-rich proteins do not provide enough information of multiple alignments.

Comparison with Existing Methods on an Independent Dataset

There have been several attempts in the past to predict accessibility of amino acids from a sequence with an objective to reduce the gap between the number of known sequences and known three-dimensional structures. Most of these methods are based on two "states" (buried or exposed) prediction such as ACCpro, Jnet, and NETASA. The methods of RVP-net and SABLE predict real value of relative solvent accessibility of amino acids from protein sequence. We evaluated the performance of all existing methods including our present method on 47 CASP6 proteins (independent dataset). First we predicted the RSA values of CASP6 proteins using SARpred, RVPnet, and SABLE Web servers and then compared the observed and predicted RSA values of residues. The correlation coefficient of 0.68, 0.52, and 0.67 and MAE of 15.9, 19.4, and 16.4% between experimental and predicted RSA values has been achieved for SARpred, RVPnet, and SABLE methods, respectively. Further, to compare our present method with existing methods for two states classifications, a threshold of 25% RSA has been chosen, which define an approximately balanced division into the two classes.[15] As shown in Table IV, our method is able to achieve an accuracy of 77.3% with an MCC value of 0.55. It has been observed that the SABLE method has achieved an accuracy of 78.0%, which is ~1% higher than that obtained by our present method. Further, the ACCpro method has achieved an accuracy of 73.9%, whereas, the Jnet method has been able to achieve an accuracy of 71.6% which, is similar to the RVPnet method (71.3%). In summary, comparison using two states classifications at a threshold of 25% for CASP6 proteins showed that performance of the present method is comparable to the SABLE method. However, SARpred is able to achieve an accuracy that is 3 and 5% better than the ACCpro and Jnet methods, respectively.

Availability of a SARpred Server

The program is implemented as a Web server SARpred, available at http://www.imtech.res.in/raghava/sarpred/, using CGI/Perl script. The SNNS-generated network (trained using CB-5012 and Manesh-215 dataset proteins collectively) is converted into the C-program and is used as an interface. Users can enter a primary amino acid sequence in a free format. The prediction results consist of RSA, RSA (%), and ASA (Å^2,) corresponding to the query sequence. Prediction results can also be e-mailed back after a short period of time, depending on the server load.

ACKNOWLEDGMENTS

We are thankful to the developers of SNNS package.

REFERENCES

1. Chan HS, Dill KA. Origins of structures in globular proteins. Proc Natl Acad Sci USA 1990;87:6388–6392.
2. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamics parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
3. Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517–525.
4. Ehrlich L, Reczko M, Bohr H, Wade RC. Prediction of protein hydration sites from sequence by modular neural networks. Protein Eng 1998;11:11–19.
5. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
6. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
7. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
8. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
9. Li X, Pan X-M. New methods for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
10. Mucchielli-Giorgi MH, Hazout S, Tuffery P. PredAcc: prediction of solvent accessibility. Bioinformatics 1999;15:176–172.
11. Manesh HN, Sadeghi M, Arab S, Movahedi AM. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
12. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566–570.
13. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
14. Ahmad S, Gromiha MM and Sarai A. RVP-net: online prediction of real valued accessible surface area of proteins from single sequences. Bioinformatics 2003;19:1849–1851.
15. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–767.
16. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;57:558–564.
17. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
18. Kaur H, Raghava GPS. A Neural network based method for prediction of gamma-turns in proteins from multiple sequence alignment and secondary structure information. Protein Sci 2003; 12:923–929.

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19. Kaur H, Raghava GPS. Prediction of β-turns in proteins from multiple alignment using neural network. Protein Sci 2003;12:627–634.
20. Kaur H, Raghava GPS. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. Proteins 2004;55:83–90.
21. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
22. Cuff JA, Barton GJ. Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 1994;40:502–511.
23. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
24. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bond and geometrical features. Biopolymers 1983;22:2577–2637.
25. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
26. Zell A, Mamier G. Stuttgart neural network simulator, version 4.2. Stuttgart, Germany: University of Stuttgart.
27. Romelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagation errors. Nature 1986;323:533–536.
28. Altschul SF, Madden TL, Alejandro AS, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein databases and search programs. Nucleic Acids Res 1997;25:3389–3402.
29. Maconald JR Jr, Johnson WC. Environmental features are important in determining protein secondary structure. Protein Sci 2001;10:1172–1177.

10970134, 2005, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/prot.20630 by University Of Southampton, Wiley Online Library on [07/11/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License