

Real-SPINE: An Integrated System of Neural Networks for Real-Value Prediction of Protein Structural Properties

Ofer Dor¹ and Yaoqi Zhou^{1,2*}

¹Department of Physiology and Biophysics, Howard Hughes Medical Institute Center for Single Molecule Biophysics, State University of New York at Buffalo, Buffalo, New York 14214

²Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis 46202

ABSTRACT Proteins can move freely in three-dimensional space. As a result, their structural properties, such as solvent accessible surface area, backbone dihedral angles, and atomic distances, are continuous variables. However, these properties are often arbitrarily divided into a few classes to facilitate prediction by statistical learning techniques. In this work, we establish an integrated system of neural networks (called Real-SPINE) for real-value prediction and apply the method to predict residue-solvent accessibility and backbone ψ dihedral angles of proteins based on information derived from sequences only. Real-SPINE is trained with a large data set of 2640 protein chains, sequence profiles generated from multiple sequence alignment, representative amino-acid properties, a slow learning rate, overfitting protection, and predicted secondary structures. The method optimizes more than 200,000 weights and yields a 10-fold cross-validated Pearson's correlation coefficient (PCC) of 0.74 between predicted and actual solvent accessible surface areas and 0.62 between predicted and actual ψ angles. In particular, 90% of 2640 proteins have a PCC value greater than 0.6 between predicted and actual solvent-accessible surface areas. The results of Real-SPINE can be compared with the best reported correlation coefficients of 0.64–0.67 for solvent-accessible surface areas and 0.47 for ψ angles. The real-SPINE server, executable programs, and datasets are freely available on <http://sparks.informatics.iupui.edu>. Proteins 2007;68:76–81. © 2007 Wiley-Liss, Inc.

Key words: residue solvent accessibility; solvent accessible surface area; psi backbone dihedral angles; neural networks; 10-fold cross validation

INTRODUCTION

Proteins are made of a linear chain of various combinations of 20 amino-acid residues (sequence). To perform their biological functions, proteins often have to fold into unique three-dimensional structures. Predicting protein structures from their corresponding sequences is one of the most challenging problems in computational biology.

An integral part of protein structure prediction is to predict structural properties of proteins, such as solvent accessible surface area, backbone dihedral angles, and atomic distances. One popular approach is to simplify structural properties into a few arbitrarily-defined structural classes, such as buried or exposed in residue solvent accessibility and contact or not-in-contact in atomic distances. However, because proteins move freely in three-dimensional space, their associated structural properties are continuously-varying variables. The goal of this article is to develop an integrated system of neural networks that accurately predict the real values of protein structural properties. We demonstrate this system with application to solvent accessibility and backbone ψ dihedral angles.

Several methods for real-value prediction of solvent accessibilities have already been developed.^{1–6} The approaches range from neural networks,^{1,3,4} support vector machines,² multiple linear regression,⁵ and a constrained energy optimization.⁶ Prediction accuracy is often measured by the correlation coefficient between predicted and actual solvent accessible surface areas. There is a steady improvement in correlation coefficient from 0.50 to around 0.65.^{3–5} A method for real-value prediction of backbone ψ dihedral angle is also reported, recently.⁷ The correlation coefficient between predicted and actual dihedral angles is only 0.47.

In this article, we propose a neural-network-based system called Real-SPINE for Real-value prediction of Structural Properties of proteins using Integrated NEural networks. To have a reliable estimate on the accuracy of Real-SPINE, we use a 10-fold cross validation on a large non-redundant dataset available from protein data bank (2640 proteins with less than 25% similarity). In this work, Real-SPINE is applied to predict solvent accessible surface

Grant sponsor: NIH; Grant numbers: R01 GM 966049, R01 GM 068530; Grant sponsor: HHMI, Center for Computational Research, Keck Center for Computational Biology; Grant sponsor: National science foundation of China; Grant number: 20340420391.

*Correspondence to: Yaoqi Zhou, Indiana University School of Informatics, Indiana University-Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis 46202. E-mail: yqzhou@iupui.edu

Received 19 October 2006; Revised 20 December 2006; Accepted 8 January 2007

Published online 30 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21408

areas and backbone ψ dihedral angles. We focus on these two properties in this work because early studies are available for comparison. Real-SPINE yields a 10-fold cross-validated correlation coefficient of 0.74 for solvent accessible surface areas and 0.62 for ψ dihedral angles. In particular, 90% of proteins have a correlation coefficient greater than 0.6 between predicted and actual accessible surface areas.

METHODS

Dataset for Training and Testing

A list of chains (3920 proteins), with sequence identity less than 25% and X-ray resolution lower than 3 Å, was taken from the protein sequence culling server PISCES.⁸ The chains with unknown structure regions were removed. The final dataset contains 2640 protein chains with a total of 591,797 residues.

For a given protein structure, residue solvent-accessible surface areas and ψ dihedral angles are calculated by the DSSP program.⁹ Residue solvent accessibility (RSA) is the solvent-accessible surface areas of a residue in a protein normalized by the accessible surface area of the residue in its “unfolded” state.¹ The ψ dihedral angles (from -180° to 180°) are converted to angles between 0° and 360° by keeping the angles between 0° and 180° unchanged, and adding 360 for angles between -180° and 0° . The angles are further normalized by 360 to make them range between 0 and 1.

Input

Each amino acid is described by a vector of many parameters. These parameters include 20 values from the Position Specific Scoring Matrix (PSSM). The PSSM is obtained from PSI-BLAST¹⁰ with three iterations of searching against nonredundant (NR) sequence database (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>). (i.e., it was produced with command “blastpgp -d filtnr -j 3 -i psitmp.fasta -Q psitmp.prf”). As in PSIPRED,¹¹ the database was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments. We also use seven representative amino-acid properties identified by Meiler et al¹²: a steric parameter (graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability. In addition, one parameter is used for describing the nonexistence of amino acids for some positions of the sliding window centered near the terminus of the chain. A sliding window centered on a residue is employed to make use of information of its neighboring residues. Thus, we have a vector of 28 parameters for each amino acid residue at a given sequence position.

Each network unit is responsible for predicting the real value of a property of one residue such as RSA or ψ . The network unit for a residue is trained and tested with a sliding window of sequential residues centered around that residue. We use a window size of 21 because it is the optimal window size for secondary structure prediction.¹³ The total number of the attributes for learning in Real-

SPINE is $28 \times 21 + 1$. Here, one additional attribute is the bias used for refining the network.

We test the effect of secondary structures. In this test, we employ three raw values for helix, sheet, and coil predicted by SPINE.¹³ The method is trained and tested on the same dataset. In this case, the total number of the attributes for learning in Real-SPINE is $28 \times 21 + 1 + 3$. We have also examined the usefulness of secondary structures with a sliding window. No further improvement was observed.

Algorithm

The overall architecture of neural networks in Real-SPINE, shown in Figure 1, follows the method SPINE for three-state (secondary structure and solvent-accessible surface area) prediction.¹³ The system makes a consensus prediction from two separate predictors (A and B) with one hidden layer. The neural networks employed here are back propagation neural networks with a sigmoid activation function. We set the learning rate and momentum at 0.001 and 0.4, respectively. The momentum and learning rate were chosen for slow learning and minimizing the possibility of missing local optimization. We use 200 hidden units for each predictor. The total number of weights for each predictor is 236,800 for 200 units when sequence profiles, amino-acid properties, and predicted secondary structures are used in input. Each predictor yields one output (either RSA or ψ angle).

All input values and output values for each networks are normalized to be within the range from 0 to 1 with a linear normalization: $X = (X - X_{\min}) / (X_{\max} - X_{\min})$. For each networks, weights are initially generated by random number generators and iteratively optimized to maximize Pearson’s correlation coefficient (PCC) between predicted and actual values of a given structural property. In each training process, a random selection of 5% of the data is set aside for independent test and error estimation to avoid possible overfitting. The weights that produce the best PCC value for the 5% of the data are saved during iterations. Iterations for learning stop if there is more than 400 continuous iterations (epochs) that decrease (or do not increase) the prediction accuracy. If the stopping criterion is not fulfilled, the maximum number of iterations is 4000. The earlier-described parameters were chosen so that the computing time required for learning is

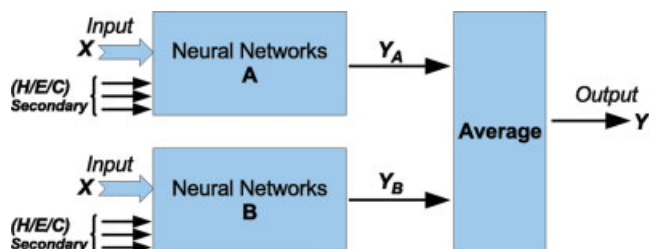


Fig. 1. Block diagram of Real-SPINE for the real-value prediction of structural properties of proteins.

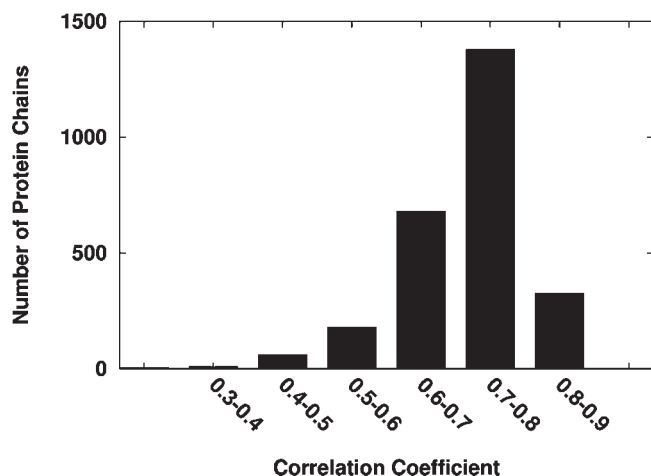


Fig. 2. Number of chains as a function of Pearson's correlation coefficient (PCC) for solvent accessible surface areas.

affordable for the computing resource available to us. There is no attempt to optimize them.

The training is performed for the networks *A* and *B* independently. Because two predictors (*A* and *B*) are started from different initial random weights, different suboptimal weights, and predictions are obtained. The two predictions are averaged to produce the final result.

In this work, we do not test various possible designs of neural networks because it is computationally impossible for a large training data set employed here. We use the design from SPINE because it provides one of the most accurate prediction of secondary structures (ten-fold cross validated accuracy of 80% in Q_3 score).¹³

Cross Validation and Assessment

We randomly divide the training set into 10 parts, nine of which is for training and the rest for testing. The process is repeated 10 times. The final result is assessed by Pearson's correlation coefficient (PCC) between predicted and actual values of a structural property in all 10 tests. We also report the mean absolute error (MAE) and accuracy according to Q-Score. MAE is the absolute difference between predicted and actual values of a normalized structural property that is averaged over all predicted residues. Q-Score is used when structural properties are simplified into several classes. Q-Score is the number of correctly classified residues in total number of residues in that class of structural properties.

RESULTS

Solvent Accessibility

Real-SPINE is first trained for predicting RSA. Real-SPINE with predicted secondary-structure information achieves the correlation coefficient of 0.738 and the MAE of 0.142. This is only a minor improvement from 0.736 and 0.143, respectively, in the absence of secondary structures. Incorporation of sliding windows to secondary structures does not help.

TABLE I. Mean Absolute Errors Between Actual and Predicted Solvent Accessibility for 20 Amino Acid Residues

Type ^a	$\langle RSA \rangle^b$	MAE ^c
A	0.21	0.1262
C	0.09	0.0773
D	0.44	0.1845
E	0.48	0.1723
F	0.12	0.1038
G	0.30	0.1795
H	0.29	0.1533
I	0.11	0.0860
K	0.47	0.1559
L	0.13	0.0944
M	0.15	0.1074
N	0.41	0.1882
P	0.33	0.1724
Q	0.40	0.1712
R	0.38	0.1622
S	0.32	0.1698
T	0.29	0.1543
V	0.13	0.0962
W	0.15	0.1119
Y	0.19	0.1279

^aAmino-acid types.

^bThe average RSA values from 2640 proteins.

^cMean absolute error between predicted and actual RSA values.

The correlation coefficients between predicted and actual solvent accessible surface areas are evaluated for each protein. The distribution of the correlation coefficients for 2640 proteins is shown in Figure 2. More than 60% of proteins have a correlation coefficient of 0.7 or above and 90% of proteins have a correlation coefficient 0.6 or above.

MAE is analyzed according to secondary-structure states of residues. Error is the smallest for sheet residues (MAE = 0.106), followed by helix residues (MAE = 0.125), and coil residues (MAE = 0.179). That is, RSA values are the most predictable for sheets and the least predictable for coils. This can be somewhat expected from the geometry of secondary-structure elements because RSAs of sheet and helical residues likely have more recognizable patterns than those of coil residues.

The mean absolute errors of RSA for 20 residues are shown in Table I. They correlate significantly with their average RSA values with a correlation coefficient of 0.91 (Fig. 3). Thus, RSA values of highly buried residues (mostly hydrophobic residues) are more accurately predicted, at least in average.

For an individual residue, however, the prediction error does not always increase monotonically with the actual exposure level of the residue. Figure 4 shows MAE of RSA prediction as a function of actual RSA for Leu and Lys. While the most accurate prediction is for the most buried Leu, the most accurate prediction for Lys is for 50% buried Lys. We found that Leu and Lys have the highest population at RSA = 0 and RSA = 0.5, respectively. That is, the lowest MAE is associated with the highest population of residues for training. When averaging over all residues,

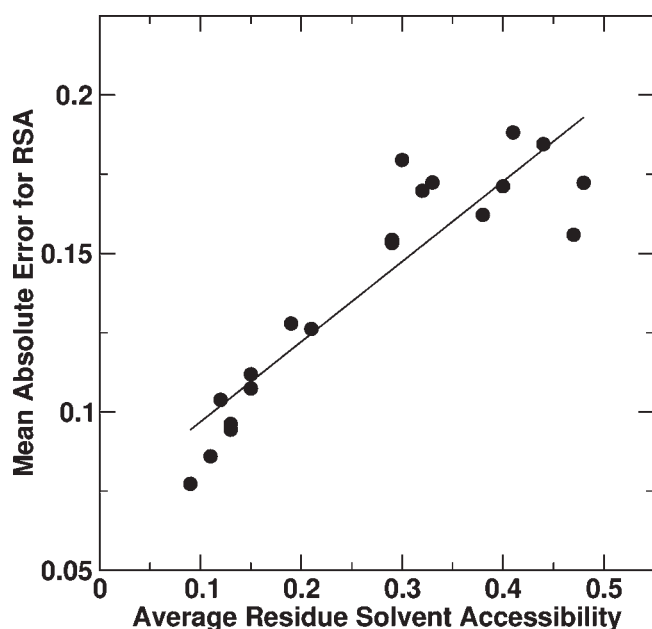


Fig. 3. Mean absolute errors between predicted and actual residue solvent accessibility for 20 amino acid residues as a function of their RSA values averaged over 2640 proteins. Solid line is from linear regression with a correlation coefficient of 0.91.

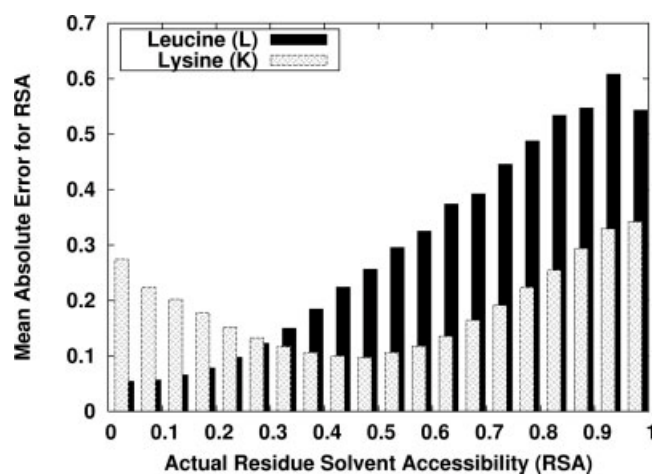


Fig. 4. Mean absolute errors between predicted and actual solvent accessibility for Lys and Leu residues, as a function of actual real solvent accessibility.

Figure 5 shows that there is a monotonic increase in MAE as the actual exposure level of residues increases. This corresponds to the monotonic decrease in number of residues as their exposure level increases (not shown). Thus, the lowest MAE is due to the highest number of residues used in training.

Another way to assess the accuracy of Real-SPINE is to evaluate Q-Score by dividing RSA values into 10 bins (0–0.1, 0.1–0.2, ... and 0.9–1.0). Results for Lys and Leu are shown in Figure 6. The Q-Score negatively correlates with MAE as expected. Q-Score is close to 80% for fully buried Leu (RSA < 0.1) and close to 0 for fully exposed one. The highest Q-Score for Lys is about 30% for RSA between

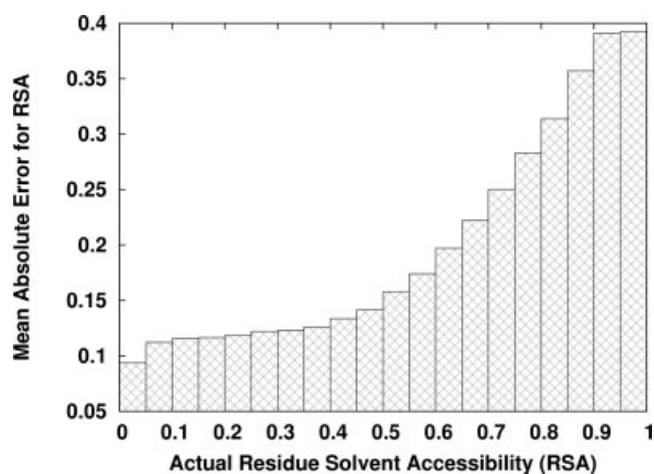


Fig. 5. As in Fig. 4 but for all residues.

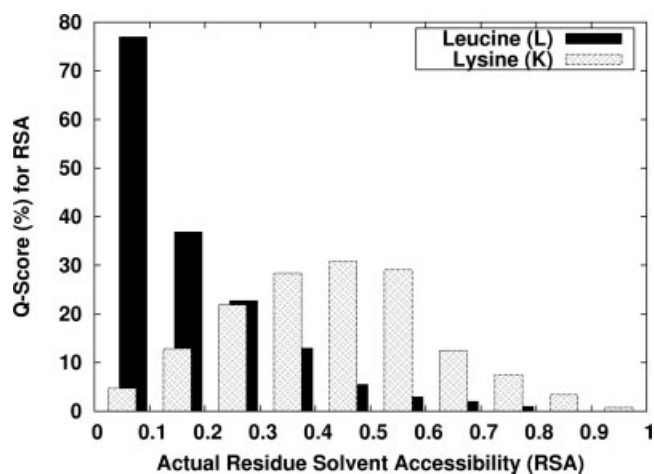


Fig. 6. Q-Score for RSA prediction as a function of actual residue solvent accessibility for Leu and Lys.

0.4 and 0.5. Considerably lower Q-Score is observed for both highly buried and highly exposed Lys. For all residues, Figure 7 indicates that Q-Score is the highest for RSA < 0.1 (about 60%), decreases to between 20% and 30% for $0.1 < \text{RSA} < 0.6$, and becomes lower than 10% after $\text{RSA} > 0.6$. Q-Scores from Real-SPINE are consistently higher than the accuracy from random prediction except for fully exposed residues with $\text{RSA} > 0.8$. Here, the accuracy from a random prediction is fraction of residues in a given RSA bin.

ψ Backbone Angle

Real-SPINE is trained independently for predicting ψ backbone angles. The accuracy measured by PCC and MAE is improved noticeably with the use of secondary structures. Real-SPINE inputted with predicted secondary structures achieves a PCC value of 0.619 and MAE of 0.150, compared with 0.611 and 0.175, respectively, in the absence of secondary structures.

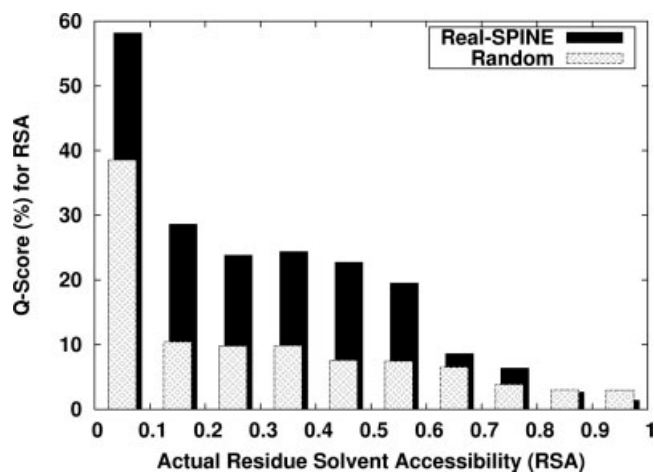


Fig. 7. Q-Score for RSA prediction as a function of actual residue solvent accessibility for all residues. Corresponding accuracy from random prediction is also shown.

TABLE II. Mean Absolute Errors Between Actual and Predicted ψ Backbone Angles for 20 Residues

Type ^a	MAE ^b
A	0.1232
C	0.1507
D	0.1908
E	0.1356
F	0.1386
G	0.2278
H	0.1766
I	0.1029
K	0.1484
L	0.1241
M	0.1269
N	0.2025
P	0.1681
Q	0.1450
R	0.1399
S	0.1694
T	0.1602
V	0.1056
W	0.1433
Y	0.1439

^aAmino acid types.

^bMean absolute error.

ψ backbone angles for helix and sheet residues are more accurately predicted than those of coil residues. MAE values of ψ backbone angles are 0.097 for sheet residues, 0.092 for helical residues, and 0.23 for coil residues, respectively. This is very similar to RSA prediction and indicates that both RSA and ψ backbone angles of helical and sheet residues are easier to predict than those of coil residues.

The MAE between actual and predicted ψ backbone angles for 20 amino acid residues are shown in Table II. We found that it correlates somewhat with the average RSA values (with a correlation coefficient of 0.50). That is, more exposed residues have larger errors in predicted ψ angles.

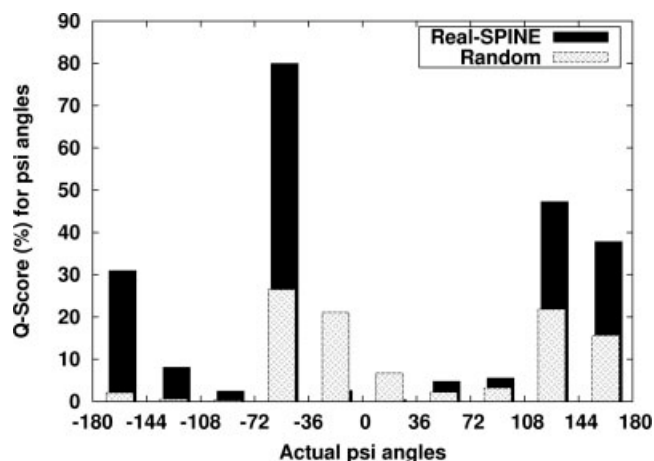


Fig. 8. Q-Score of ψ backbone-angle prediction as a function of ψ angles. The accuracy from random prediction is also shown.

Q-Scores of ψ backbone angles are obtained by dividing -180° to 180° into 10 bins (with 36° per bin). Q-Score is plotted as a function of real ψ angle in Figure 8. The highest Q-Score is 80% for angles between -72° and -36° and followed by about 30–50% for angles between 108° and 180° and between -144° and -180° . Q-Scores from Real SPINE are consistently higher than the accuracy from random prediction except between -36° and 36° . MAE values of ψ backbone angles are also small at the regions where Q-Scores are high (not shown).

Exact Secondary Structures

If known secondary structures (from the DSSP program) rather than predicted secondary structures are used as input, we found that the PCC values can be further increased to 0.747 from 0.738 for solvent-accessible surface areas and 0.735 from 0.619 for ψ angles. In the mean time, MAE can be further decreased to 0.140 from 0.142 for solvent-accessible surface areas and to 0.123 from 0.150 for ψ angles. The effect on ψ angles is more dramatic than on RSA prediction. This is due the direct connection between ψ angles and secondary structures.

DISCUSSION

In this article, we have developed a neural-network method called Real-SPINE to predict the real values of protein structural properties. Application of this method to solvent accessibility and ψ dihedral angles indicates a significant improvement over existing methods in prediction accuracy. For example, the PCC value between predicted and actual RSA values increases to 0.74 from 0.64–0.67, the best results from existing methods.^{3–5} This occurs along with the reduction of mean absolute errors from 0.152–0.162^{2–5} to 0.14. By comparison, the PCC value between RSA predicted by average accessible surface areas and actual RSA values is only around 0.4.⁶ If a two-state definition for RSA is used, the 10-fold cross-validated accuracy with a 25% cutoff is 79.2%. The best reported accuracy for two-state RSA prediction with

the same 25% cutoff is 77–78%.^{3,4,14–16} This two-state accuracy, however, is same as that when the three-state prediction of solvent accessibility predicted from SPINE is reduced to two states.¹³

Improvement in real-value prediction by Real-SPINE can be attributed to a combination of multiple factors. This includes the use of sequence profiles from multiple alignment,⁴ amino-acid properties,³ and predicted secondary structures,⁴ consensus prediction³ made from two separate predictors, and a large-scale learning of more than 200,000 weights with a slow learning rate and overfit protection. In addition, all weights in Real-SPINE are saved according to values of correlation coefficients, rather than the final weights from neural networks.

Predicted RSA^{17–21} have been demonstrating its usefulness in fold recognition. Improved accuracy in both RSA and ψ angle prediction would likely become additional tools for attacking the grand challenge of ab initio protein-structure prediction.

The real-SPINE server and datasets are freely available on <http://sparks.informatics.iupui.edu>. For a given sequence, it produces the real values of solvent accessibility and ψ dihedral angles. It also reports a reliability score (*RS*) that is a measure of the consistency between two predictors (*A*, *B*). $RS = 1 - |A - B|$ where *A* and *B* are the results from two predictors on solvent accessibility or normalized dihedral angles.

ACKNOWLEDGMENT

We thank the authors who made their programs and databases available for comparison.

REFERENCES

- Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* 2003;50:629–635.
- Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins* 2004;57:558–564.
- Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004;56:753–767.
- Garg A, Kaur H, Raghava G. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* 2005;61:318–324.
- Wang J, Lee H, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins* 2005;61:481–491.
- Xu Z, Zhang C, Liu S, Zhou Y. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins* 2006;63:961–966.
- Wood MJ, Hirst JD. Protein secondary structure prediction with dihedral angles. *Proteins* 2005;59:476–481.
- Wang G, Dunbrack RLJ. PISCES: a protein sequence culling server. *Bioinformatics* 2003;19:1589–1591.
- Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res* 1997; 25:3389–3402.
- Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
- Meiler J, Muller M, Zeidler A, Schmaschke F. Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. *J Mol Model* 2001;7:360–369.
- Dor O, Zhou Y. Achieving 80% accuracy for secondary-structure prediction by large-scale neural-network training. *Proteins* 2007;66:838–845.
- Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* 2004;54:557–562.
- Nguyen M, Rajapakse J. Prediction of protein relative solvent accessibility with a two-stage svm approach. *Proteins* 2005;59:30–37.
- Nguyen M, Rajapakse J. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins* 2006;63:542–550.
- Cheng J, Baldi P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 2006; in press.
- Rost B. TOPITS: threading one-dimensional predictions into three-dimensional structures. Third international conference on intelligent systems for molecular biology 1995; AAAI Press Menlo Park, CA USA. pp. 314–321.
- Rost B, Sander C. Protein fold recognition by prediction-based threading. *J Mol Biol* 1997;270:471–480.
- Przybylski D, Rost B. Improving fold recognition without folds. *J Mol Biol* 2004;341:255–269.
- Qiu J, Elber R. SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs. *Proteins* 2006;62:881–891.