

# Achieving 80% Ten-fold Cross-validated Accuracy for Secondary Structure Prediction by Large-scale Training

Ofer Dor<sup>1</sup> and Yaoqi Zhou<sup>1,2,3\*</sup>

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

<sup>2</sup>Indiana University School of Informatics, Indiana University Purdue University, Indianapolis, Indiana 46202

<sup>3</sup>Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana 46202

**ABSTRACT** An integrated system of neural networks, called SPINE, is established and optimized for predicting structural properties of proteins. SPINE is applied to three-state secondary-structure and residue-solvent-accessibility (RSA) prediction in this paper. The integrated neural networks are carefully trained with a large dataset of 2640 chains, sequence profiles generated from multiple sequence alignment, representative amino acid properties, a slow learning rate, overfitting protection, and an optimized sliding-window size. More than 200,000 weights in SPINE are optimized by maximizing the accuracy measured by  $Q_3$  (the percentage of correctly classified residues). SPINE yields a 10-fold cross-validated accuracy of 79.5% (80.0% for chains of length between 50 and 300) in secondary-structure prediction after one-month (CPU time) training on 22 processors. An accuracy of 87.5% is achieved for exposed residues (RSA >95%). The latter approaches the theoretical upper limit of 88–90% accuracy in assigning secondary structures. An accuracy of 73% for three-state solvent-accessibility prediction (25%/75% cutoff) and 79.3% for two-state prediction (25% cutoff) is also obtained. *Proteins* 2007;66:838–845. © 2006 Wiley-Liss, Inc.

**Key words:** solvent accessibility; solvent accessible surface area; neural network

## INTRODUCTION

How to make an accurate prediction of protein secondary structure and residue solvent accessibility (RSA) is a long-standing unsolved problem in structural bioinformatics. One common way to predict secondary structure and solvent accessibility is to classify them into a few states. A three-state classification (helix, sheet, coil) is often used for secondary structure while two-state (buried and exposed) and three-state (buried, exposed, intermediate) assignments are most common for solvent accessibility.

Early methods for secondary-structure prediction are built on statistical analysis of single residue<sup>1–3</sup> and its neighboring residues.<sup>4–7</sup> Current state-of-the-art techniques,<sup>8,9</sup> on the other hand, employ machine-learning models to “learn” from sequence profiles generated from

multiple sequence alignment as well as other sequence-derived information. The most commonly used machine-learning models are neural networks.<sup>10–18</sup> Other methods such as multiple linear regression,<sup>19,20</sup> k-nearest neighborhood,<sup>21</sup> and support vector machines<sup>22–24</sup> have also been used. Some methods<sup>14,15,25–30</sup> are based on consensus prediction from multiple methods or multiple neural networks. In a recent study, a separate neural network for predicting the ends of secondary-structure segments yields an improved prediction of secondary structure.<sup>31</sup> Many approaches have also been developed for predicting RSA.<sup>32–43,35</sup> Methods for a combined prediction of secondary and solvent accessibility<sup>44</sup> or  $\psi$  dihedral angles<sup>45</sup> are also developed. The accuracy of secondary-structure prediction is stagnated around 77%. Most reported accuracies, however, are not multiply cross validated and/or are obtained from several small datasets.

The goal of this paper is to develop an integrated system of neural networks that is suitable for predicting structural properties of proteins. We design a consensus predictor based on a parallelized two-level neural network. The method is called SPINE (prediction of Structural Properties of proteins by Integrated NEural networks). Here, we apply SPINE to secondary-structure and solvent-accessibility prediction. To have a reliable estimate on the three-state accuracy of SPINE, we use a ten-fold cross-validation on a large nonredundant dataset available from protein data bank (2640 proteins with less than 25% similarity). The  $Q_3$  score (the percentage of correctly classified residues) is 79.5% for secondary-structure prediction and 73% for RSA (three-state definition based on 25% and 75% cutoffs). We further examine

The Supplementary Material referred to in this article can be found at <http://www.interscience.wiley.com/jpages/0887-3585/suppmat/>

Grant sponsor: NIH; Grant numbers: R01 GM 966049 and R01 GM 068530; Grant sponsor: HHMI; Grant sponsors: The Center for Computational Research and the Keck Center for Computational Biology at SUNY, Buffalo; Grant sponsor: National Science Foundation of China; Grant number: 20340420391.

\*Correspondence to: Yaoqi Zhou, Indiana University School of Informatics, Indiana University Purdue University, Indianapolis, IN 46202. E-mail: yqzhou@iupui.edu

Received 15 August 2006; Revised 18 September 2006; Accepted 17 October 2006

Published online 18 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21298

the effect of window sizes, number of hidden layers, X-ray resolution, chain sizes, and other attributes on the accuracy of secondary-structure prediction.

## METHOD

### 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.<sup>46</sup> The chains with unknown structure regions were removed. The final dataset contains 2640 protein chains with a total of 591,797 residues. The dataset has a composition of 37.6% helix, 23.0% sheet, and 39.4% coil. In this set, there are 1952 chains whose chain length is between 50 and 300. The composition of 1952 chains is 37.5% helix, 24.3% sheet, and 38.2% coil. To test the dependence on the size of dataset, another group of chains is obtained by randomly removing 50% of the dataset with 2640 chains. This small-size dataset has 1373 chains.

For a given protein structure, the secondary structure of a residue is defined by the program DSSP.<sup>47</sup> Because the DSSP program assigns eight states for secondary structures, we group them into the three states by converting (G, H, I) to H, (B, E) to E, and (T, S, other) to C. Residue solvent accessible surface area is also calculated by the DSSP program. Residue solvent accessibility (RSA) is the solvent accessible surface area of a residue in a protein normalized by the solvent accessible surface area of the residue in its “unfolded” state.<sup>48</sup> We define a residue as buried if  $RSA \leq 25\%$ , somewhat exposed if  $25\% < RSA \leq 75\%$ , and fully exposed if  $RSA > 75\%$ . The compositions for buried, partially exposed, and fully exposed are 55.4%, 37.1%, and 7.5%, respectively, for the 2640-protein set.

To further test the method and facilitate comparison with early studies, other small datasets of protein structures have been used. These datasets include 215 high-resolution structures of proteins (Manesh-215) collected by Naderi-Manesh et al.,<sup>49</sup> 338 monomeric protein dataset (Carugo-338) used by Carugo,<sup>50</sup> and 513 protein dataset (CB-513) developed by Cuff and Barton.<sup>51</sup> These datasets are also made of protein sequences with less than 25% homology.

### SPINE 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-BLAST<sup>52</sup> with three iterations of searching against nonredundant sequence database (<ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz>) (i.e. it was produced with command “blastpgp -d fltnr -j 3 -i psitmp.fasta -Q psitmp.pr”). As in PSIPRED,<sup>12</sup> the database was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments. We also use seven representative amino acid properties

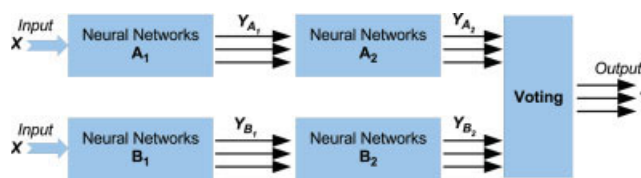


Fig. 1. Block diagram of three-state prediction for secondary structure and residue solvent accessibility.

identified by Meiler et al.<sup>53</sup> 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 non-existence of amino acids for some positions of the sliding window centered at the edge of the chain. A sliding window centered on a residue is employed to make use of information of its neighboring residues. (Also see later.) Thus, we have a vector of 21 or 28 parameters for each amino acid residue at a given sequence position depending on if seven properties of amino acid residues are used.

As with previous studies,<sup>26</sup> each network unit is responsible for predicting the secondary structure (or RSA) of one residue. The network unit for a residue is trained and tested with a sliding window of sequential residues centered around that residue. The total number of the attributes for learning in the SPINE method is either  $21 \times n + 1$  for PSSM only or  $28 \times n + 1$  for PSSM plus properties (PROP). Here, one additional attribute is the bias used for refining the network and  $n$  is the size of sliding window.

### SPINE Algorithm

As shown in Figure 1, the overall architecture of neural networks in SPINE follows the protocol established by Rost and Sander<sup>26</sup> and adapted by many others (e.g. Refs. 12 and 54). The system makes a consensus prediction from two separate predictors (A and B) consisting of two-level neural networks (i.e. a total of four neural networks). The first-level ( $A_1$  or  $B_1$ ) network is a three-state classifier using all the input attributes described earlier. The second-level ( $A_2$  or  $B_2$ ) network is a filter that refines the predicted results from  $A_1$  or  $B_1$  with or without a sliding window.

The neural networks employed here are back propagation neural networks with a sigmoid activation function. The parameters for neural networks are shown in Table I. Specifically, the learning rate and momentum are set 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 either 100 or 200 number of hidden units for the first-level network in order to test the effect of the number of hidden units. The number of hidden units for the second-level network is fixed at 10.

TABLE I. Neural Networks Settings

Networks	Momentum	Learning rate	Maximum epochs	Stopping epochs	Units
A <sub>1</sub> and B <sub>1</sub>	0.4	0.0001	4,000	400	100 or 200
A <sub>2</sub> and B <sub>2</sub>	0.4	0.0001	20,000	400	10

All initial input and output values for each network are generated by random number generators and were normalized to be within the range from 0 to 1 with a linear normalization:  $X = (X - X_{\min}) / (X_{\max} - X_{\min})$ . For each network, weights are iteratively optimized to maximize the number of correctly classified residues divided by total number of residues, the  $Q_3$  value. In each training process, a random selection of 5% of the data is set aside for independent test and error estimation in order to avoid possible overfitting. The weights that produce the more accurate prediction of the 5% of the data are saved for further use. 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 for A<sub>1</sub> and B<sub>1</sub> networks and 20,000 for A<sub>2</sub> and B<sub>2</sub> networks. In every learning the aforementioned parameters were chosen so that the computing time required for learning is affordable for the computing resource available to us. There is no attempt to optimize them.

The training is first performed for the first-level networks (A<sub>1</sub> and B<sub>1</sub>) and followed by the second-level networks (A<sub>2</sub> and B<sub>2</sub>). Because two predictors (A and B) are started from different initial random weights, different suboptimal weights and predictions are obtained. The two predictions are combined by voting. If both predictions favor one state, the state is predicted. If the two predictions are different states, the one with the highest output value is chosen. The output could be either secondary structure as H (helix), E (sheet), or C (coil), or three-state solvent accessibility for fully exposed, partially exposed, and fully buried, as defined in the training dataset.

### Cross validation

We randomly divide the training set into 10 parts, nine of which are for training and the rest for testing. The process is repeated 10 times. The  $Q_3$  score is the total number of correctly predicted residue states (in all 10 tests) divided by the size of training set (total number of residues). The accuracies for helices ( $Q_H$ ), sheets ( $Q_E$ ), and coils ( $Q_C$ ) are also reported in terms of number of correctly predicted residues in total number of residues in a given class (state).

## RESULTS

### The Window Size

Because of the uncertainty about which window size for the first-level network to use, we train and test neu-

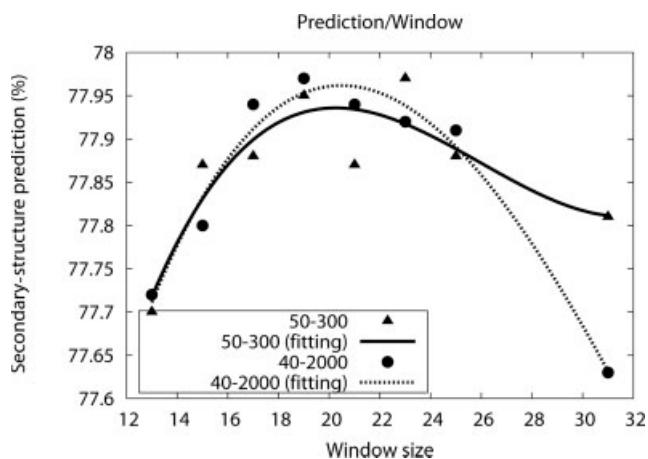


Fig. 2. The accuracy of secondary-structure prediction as a function of window size for A<sub>1</sub> and B<sub>1</sub>. Two sets of results are for 170,093 residues and 1146 chains of lengths between 50 and 300 (triangles) and 248,299 residues and 1373 chains of lengths between 40 and 2000 (solid circles), respectively. Lines are from curve fitting to the two sets of data. This test is conducted with PSSM as the only input.

ral networks on various window sizes by using two datasets of different chain lengths (50–300 and 40–2000) for 1146 and 1373 chains, respectively. (1146 chains are obtained from 1373 chains after proteins with chain lengths less than 50 or greater than 300 are removed.) The resulting  $Q_3$  scores are shown in Figure 2. We found that window sizes between 15 and 25 make a negligible effect on the accuracy of prediction (0.1–0.15%). Nevertheless, there is an optimal window size. Without curve fitting, the optimal window size with the highest  $Q_3$  score is 23 and 19, for the first and second datasets, respectively. From the curve fitting, the results for both datasets suggest an optimal window size of 21. Thus, hereafter, we will use a window size of 21 for all other training and testing.

### Secondary-Structure Prediction

Six experiments with a sliding window size of 21 residues for the first-level network are conducted. The results of ten-fold cross validations are shown in Table II. These experiments are designed to test the effect of employing amino acid properties, the size of databases, and number of hidden layers, chain lengths, and the size of window for the second-level network on the accuracy of secondary-structure prediction. The overall  $Q_3$  score increases 0.5% by employing seven amino acid properties as additional attributes to PSSM, 0.5% by increasing the size of datasets from 1373 to 2640 chains, 0.2% by using sliding

**TABLE II. Six Experiments in Secondary-Structure Prediction**

Experiment	No. of chains	No. of residues	Chain size	Filter window	Input profile	NN units	$Q_3$ Score (%)
1	1373	248,299	40–2000	1	PSSM	100	77.9
2	1373	248,299	40–2000	1	PSSM+PROP	100	78.6
3	2640	591,797	40–2000	1	PSSM+PROP	100	79.1
4	2640	591,797	40–2000	1	PSSM+PROP	200	79.3 <sup>a</sup>
5	2640	591,797	40–2000	11...21 <sup>b</sup>	PSSM+PROP	200	79.5 <sup>c</sup>
6	1952	313,006	50–300	11...21 <sup>b</sup>	PSSM+PROP	200	80.0 <sup>d</sup>

<sup>a</sup>The partial accuracies are  $Q_H = 83.35\%$ ,  $Q_E = 69.95\%$ , and  $Q_C = 80.55\%$ . The compositions of helices, sheets, and coils are 37.6%, 23.0%, and 39.4%, respectively.

<sup>b</sup>The sizes of filter window ( $A_2, B_2$ ) tested are 11, 13, 15, 17, 19, and 21. They yielded the same performance.

<sup>c</sup>The partial accuracies are  $Q_H = 83.72\%$ ,  $Q_E = 71.07\%$ , and  $Q_C = 80.48\%$ . (This setting was used for SPINE server.)

<sup>d</sup>The partial accuracies are  $Q_H = 84.44\%$ ,  $Q_E = 72.23\%$ , and  $Q_C = 80.46\%$ . The compositions of helices, sheets, and coils are 37.5%, 24.3%, and 38.2%, respectively.

**TABLE III. The Accuracy of Secondary-Structure Prediction for Small Datasets<sup>a</sup>**

Dataset	Method	$Q_h$	$Q_e$	$Q_c$	$Q_3$ Score (%)
Carugo-338 <sup>b</sup>	Tenfold <sup>c</sup>	79.76	67.63	80.13	77.07
	Direct <sup>d</sup>	83.02	73.23	81.86	80.14
CB-513 <sup>e</sup>	Tenfold <sup>c</sup>	79.04	64.85	80.98	76.77
	Direct <sup>d</sup>	83.84	72.35	79.91	79.64

<sup>a</sup>Window sizes are 21 and 15 for the first and second level neural networks, respectively. Both PSSM and PROP profiles are used. The number of neural network units is 200 for the first level neural networks.

<sup>b</sup>338 monomeric protein dataset.<sup>50</sup>

<sup>c</sup>Tenfold cross-validated results.

<sup>d</sup>Direct application of SPINE trained by 2640 proteins.

<sup>e</sup>513-protein dataset (CB-513) developed by Cuff and Barton.<sup>51</sup>

window of 11 to 21 in filter neural networks ( $A_2, B_2$ ), and 0.2% by increasing the number of hidden layers from 100 to 200. The final  $Q_3$  score for the ten-fold cross validation of the dataset of 2640 chains is 79.5%. The accuracies of helix, sheet, and coil are 83.35%, 69.95%, and 80.55%, respectively. As with previous studies,<sup>8,9</sup> helical prediction has the highest success rate, followed by coil. Interestingly, the  $Q_3$  score for proteins whose chain lengths between 50 and 300, derived from 40–2000 test results, is even higher by an additional 0.5% (80.0%).

SPINE is further tested on several smaller datasets. Ten-fold cross validations are separately performed on two datasets with 338 (Carugo-338<sup>50</sup>) and 513 (CB-513<sup>51</sup>) proteins. These datasets are also made of protein sequences with less than 25% homology. Results are shown in Table III. As expected, a lower predicted accuracy ( $Q_3 = 77\%$ ) is observed with a smaller training dataset. Direct application of SPINE trained by 2640 proteins (Experiment 5) yields a  $Q_3$  score of 80% for both sets.

It is of interest to see whether the accuracy of protein structures has an effect on the accuracy of secondary-structure prediction. We analyze the data from Experiment 3 by plotting the  $Q_3$  score as a function of X-ray resolution. Figure 3 shows that indeed the prediction accu-

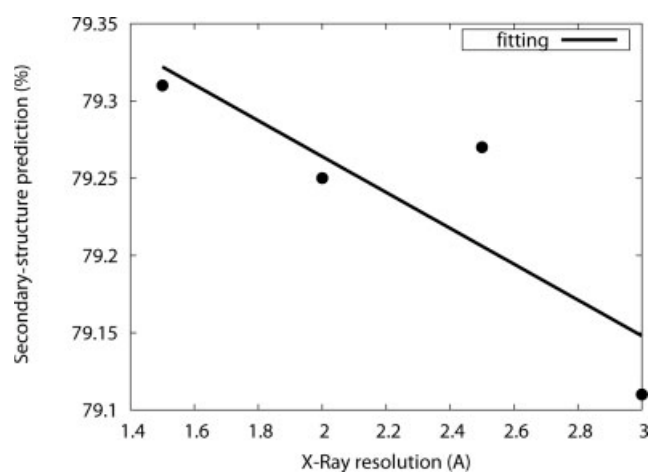


Fig. 3. The accuracy of secondary-structure prediction as a function of the cutoff for the X-ray resolution of protein structures included in the training dataset. Results are from Experiment 3 described in Table 2. The number of residues is 135,378, 410,274, 544,503, and 591,797, for a resolution cutoff of 1.5, 2, 2.5, and 3Å, respectively. The number of chains is 628, 1849, 2427, and 2640, for a resolution cutoff of 1.5, 2, 2.5, and 3Å, respectively.

racy is higher for a structure with higher resolution. The change in  $Q_3$  scores, however, is small. The  $Q_3$  score is 79.3% for all chains with resolution of 1.5 Å or higher and 79.1% for all chains with a resolution of 3 Å or higher.

One can also analyze whether or not the accuracy of secondary-structure prediction is influenced by solvent accessibility of amino acid residues. Figure 4 shows the accuracy of prediction as a function of solvent accessibility. The outstanding feature of this figure is that prediction for 11,583 residues with RSA >95% is the most accurate ( $Q_3 = 87.5\%$ ). This is because exposed residues interact weakly with the rest of proteins and local interactions dominate. As residues became more buried (RSA <90%), the change in accuracy of prediction is small (between 77 and 81%). Similar feature is observed when the accuracy of secondary-structure prediction is analyzed on the individual residue level. Figure 5 shows the results for hydrophobic residue Leu and hydrophilic residue Lys as examples.

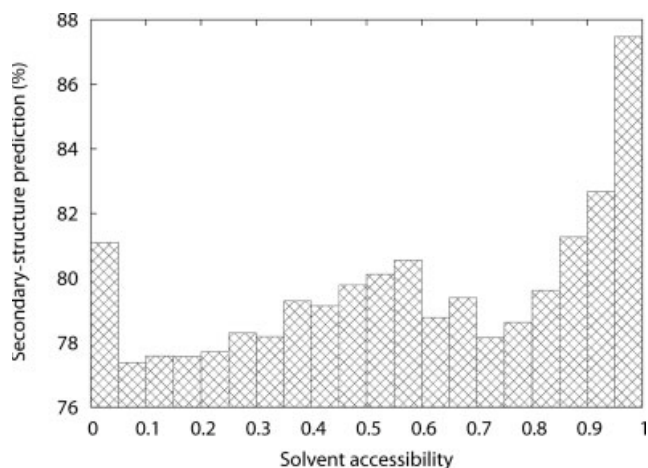


Fig. 4. The accuracy of secondary-structure prediction as a function of solvent accessibility (RSA).

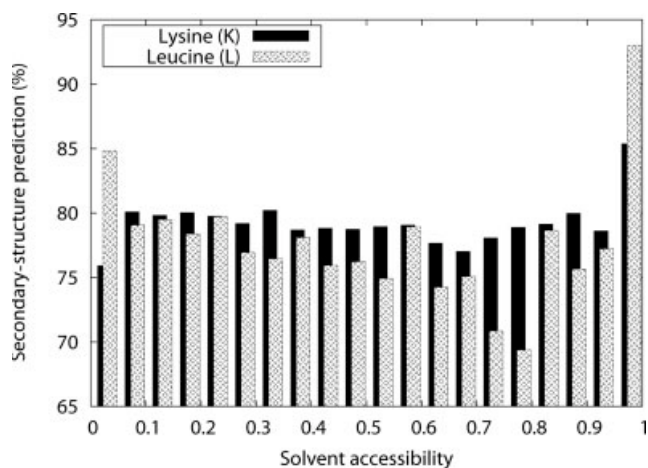


Fig. 5. The accuracy of secondary-structure prediction, for K and L residues, as a function of solvent accessibility (RSA).

The accuracies of predicted secondary structures for 20 amino acid residues (Experiment 5) are shown in Table IV. The prediction accuracy ranges from 75% to 81%. The residue with the lowest accuracy is Cys (75.16%). We found that there is a strong positive correlation (with a correlation coefficient of 0.66) between predicted accuracy for amino acid residues and their abundance in 2640 proteins (or 591,797 residues). (See Fig. S1 in Supplement Materials.) There are two under-performed outliers (hydrophilic Ser and Thr) and two outperformed outliers (hydrophobic Met and Ile). Removing the four residues yields a correlation coefficient of 0.9. It is not clear what makes relatively poorer prediction for Ser and Thr. Nevertheless, a larger dataset for a residue will likely yield a more accurate prediction for that residue. This further highlights the importance of a large dataset for training and test.

What constitutes 20.5% errors in secondary-structure prediction in Experiment 5 described earlier? Table V compares the error rates associated with misclassification between different secondary structural classes. It

**TABLE IV. The Accuracy of Secondary-Structure Prediction ( $Q_3$ ) for 20 Residues Along With Their Compositions in the Training Set of 2640 Proteins**

Amino acid	$Q_3$	Composition (%)
A	81.45	8.21
C	75.16	1.27
D	78.97	5.84
E	80.99	6.8
F	78.16	4.08
G	80.72	7.31
H	75.61	2.33
I	82.07	5.73
K	78.92	5.9
L	81.48	9.2
M	80.42	1.83
N	78.00	4.38
P	80.61	4.54
Q	79.72	3.75
R	79.37	4.98
S	76.06	5.89
T	76.75	5.45
V	81.48	7.07
W	75.67	1.44
Y	76.72	3.58

**TABLE V. Errors Contributed by Misclassification of Residue States**

Predicted	Actual	Error (%)
H	E	1.03
H	C	4.16
E	H	0.85
E	C	3.54
C	H	5.27
C	E	5.61
↓		
Misclassification		Error (%)
H $\leftrightarrow$ C		9.43
H $\leftrightarrow$ E		1.88
E $\leftrightarrow$ C		9.15

shows that the error is mainly due to misclassification between helix and coil (9.4%) and between sheet and coil (9.2%) but is rarely due to misclassification between helix and sheet (1.9%). The result agrees with previous observations<sup>8</sup> and explains the high success rate of prediction of protein classes (all- $\alpha$  proteins, all- $\beta$  proteins, and mixed  $\alpha$ ,  $\beta$  proteins, e.g. Ref. 14).

We also analyze additional details regarding actual secondary structures of the residues surrounding the residue with an incorrectly predicted secondary structure (see Table S1 in Supplement Material). It shows that misidentifying the secondary structure in the middle of a secondary-structure segment is 9% (e.g. 2.4% error involving "H" in the middle of "HHH" that is misidentified as "C") while misidentifying the secondary structure at the edge of a secondary-structure segment

**TABLE VI. Solvent Accessibility (3-States) Prediction**

Experiment	No. of chains	No. of residues	Chain size	Input profile	NN units	$Q_3$ Score (%)
1	2640	591,797	40–2000	PSSM	100	72.2
2	2640	591,797	40–2000	PSSM	200	72.4
3	2640	591,797	40–2000	PSSM+PROP	100	72.8
4	2640	591,797	40–2000	PSSM+PROP	200	73.0 <sup>a</sup>

<sup>a</sup>The partial accuracies are  $Q_{\text{buried}} = 83.66\%$ ,  $Q_{\text{exposed}} = 68.68\%$ , and  $Q_{\text{fully-exposed}} = 16.47\%$ . The fraction of residues are 0.554 for buried (RSA <25%), 0.371 for exposed, and 0.075 for fully exposed (RSA >75%).

**TABLE VII. Solvent Accessibility (2-States) Prediction (25% Cutoff)**

Experiment	No. of chains	No. of residues	Chain size	Input profile	NN units	$Q_2$ Score (%)
1	2640	591,797	40–2000	PSSM	100	78.6
2	2640	591,797	40–2000	PSSM	200	78.8
3	2640	591,797	40–2000	PSSM+PROP	100	79.2
4	2640	591,797	40–2000	PSSM+PROP	200	79.3

**TABLE VIII. Solvent Accessibility (2 and 3 States) Prediction With SPINE Server**

Experiment	Dataset	$Q_{\text{buried}}$	$Q_{\text{exposed}}$	$Q_{\text{fully exposed}}$	$Q_3$ Score (%)	$Q_2$ Score (%)
1	RS-126 <sup>a</sup>	84.30	64.77	16.61	69.57	77.19
2	Manesh-215 <sup>b</sup>	84.77	69.79	20.80	74.39	80.65
3	Carugo-338 <sup>c</sup>	84.67	68.91	19.87	73.26	79.93
4	CB-513 <sup>d</sup>	85.21	66.10	18.64	72.12	78.78

<sup>a</sup>126 Proteins by Rost and Sander.<sup>32</sup>

<sup>b</sup>215 Proteins by Naderi-Manesh et al.<sup>49</sup>

<sup>c</sup>338 Monomeric protein dataset.<sup>50</sup>

<sup>d</sup>513 Protein dataset developed by Cuff and Barton.<sup>51</sup>

is 8% (e.g. 2.7% error involving “E” in the middle of “EEC” or “CEE” that is misidentified as “C”).

### Solvent-Accessibility Prediction

To test the general applicability of the SPINE algorithm, we also use it to predict RSA. The window sizes for the first- and second-level networks are 21 and 1, respectively. (A larger window size for the second-level networks did not improve the results.) Four experiments are conducted on the effects of the use of amino acid properties and the number of hidden layers. The results are shown in Table VI for three-state classification and in Table VII for two-state classification. The reported accuracy confirms a small increment of accuracy with higher number of hidden layers and a relatively larger increment of accuracy with addition of amino acid properties. The best  $Q_3$  score is 73.0% for a three-state classification and 79.3% for a two-state one.

We further evaluate SPINE by applying it directly to several small datasets regardless of whether some or all of proteins were contained in the training/test set of 2640 proteins. Results are shown in Table VIII. The prediction accuracy is stable across different datasets (72–74% for  $Q_3$  and 79–81% for  $Q_2$ ).

## DISCUSSION

In this paper, we have developed a general-purpose neural-network method, called SPINE. The method is

tested for secondary-structure and solvent-accessibility prediction. With a large dataset of 2640 chains and 591,797 residues, SPINE achieves a  $Q_3$  score of 79.5% for secondary-structure prediction and 79.3% for two-state RSA prediction from a ten-fold cross validation. The prediction accuracy given by SPINE can be attributed to a large-scale learning of more than 200,000 weights with a slow learning rate and overfit protection. Moreover, the optimal weights in SPINE are those increasing  $Q_3$  instead of choosing the lowest error functions—an approach commonly used by other methods.

This work represents one of a few studies reporting a rigorous ten-fold cross validation on a large dataset for secondary-structure prediction. A probabilistic-based model<sup>55</sup> yields a 77% accuracy in a 10-fold cross validation on a small test set of 174 proteins. Wood and Hirst<sup>45</sup> reported a 77.2% accuracy in a ten-fold cross validation on a set of 2245 proteins and further improved to 79.4% after iterative use of predicted  $\psi$  dihedral angles. A conditional-random-field consensus-based method<sup>29</sup> produced a 77% accuracy on a dataset of 513 proteins<sup>51</sup> but for a sevenfold cross validation. Porter, a bidirectional recurrent neural network, reported 79.0% accuracy in a fivefold cross validation on a dataset of 2171 proteins. HYPROSP II achieves a greater than 80% success rate for a ten-fold cross validation on large datasets of 3925 and 2217 proteins. This method, however, combines the prediction of PROSP (based on short-fragment matching) and PSIPRED. PSIPRED<sup>12</sup> was

trained separately. Thus, it is difficult to know the actual accuracy from cross validation when all methods involved in HYPROSP II are trained and tested in the same manner. However, it should be emphasized that comparison made between this work and other published works is not a strict one. This is because it is impossible to have an exact comparison between the methods that use the time-dependent sequence and/or structural libraries.

Ten-fold cross validation for RSA prediction is rare. The only work we found is by Yuan et al.,<sup>40</sup> who obtained 74.6% accuracy based on 25% cutoff for a dataset of 531 proteins. On the other hand, the SVMpsi method<sup>43</sup> yields 78.7% in a sevenfold cross validation for two-state RSA prediction with the same 25% cutoff for a dataset of 480 proteins. Yuan and Huang<sup>41</sup> reported a 74% accuracy in a threefold cross validation for a dataset of 1277 proteins. Other studies focused on tests on a few datasets. The best reported accuracy for two-state RSA prediction with the same 25% cutoff is 77–78%.<sup>38,54,56,57</sup> Thus, SPINE yields the most accurate prediction of RSA.

We have examined the effects of amino acid properties, number of hidden layers, and size of database. All are found to positively contribute the accuracy of prediction. The effect is more significant for adding seven amino acid properties and doubling the size of database (by 0.5%) but is less so for doubling the number of hidden layers (by 0.2%). We also show that there is an optimal window size for secondary-structure prediction around 21. However, in the range from 15 to 25, the prediction accuracy could vary only in 0.1–0.15%, as shown previously.<sup>58</sup>

We found that the X-ray resolution has a small but positive effect on the accuracy of prediction. This suggests that the gap between 80% accuracy achieved by SPINE and a suggested theoretical upper limit of 88–90%<sup>8,59</sup> is likely not caused by the errors associated with the accuracy of X-ray structures.

One source of errors for secondary-structure prediction is the lack of adequate accounting of nonlocal interactions.<sup>8,59,60</sup> This is reflected by the following. Selecting a group of proteins based on chain sizes for learning and predicting can change the prediction as much as 0.5%. That is, there is a significant size dependence. This is likely because longer chains are associated with more nonlocal interactions (the interaction between residues are separated by long sequences). Moreover, SPINE can achieve 87.5% for fully exposed residues, which is very close to theoretical limit of 88–90%<sup>8,59</sup>—the accuracy of secondary-structure assignment. This indicates that SPINE provides a near perfect prediction for secondary structures if local interactions dominate as in the case of fully exposed residues. Fully exposed residues only interact with neighboring residues and solvent molecules.

We have made several attempts to further improve the accuracy of secondary-structure prediction. These attempts are inspired by studies indicating that pre-

dicted RSA<sup>44</sup> and  $\psi$  backbone angles<sup>45</sup> can be used to improve secondary-structure predictions. Adamczak et al. improved  $Q_3$  accuracy from 76.6–77.9% to 80.3–81.8% for four small datasets of 135–163 chains by using real-value predicted RSA while Wood and Hirst increased  $Q_3$  accuracy from 77.2–79.4% in a ten-fold cross validation on a set of 2245 proteins via iterative use of predicted  $\psi$  dihedral angles. Similar attempts to improve SPINE were not successful. This happens despite our ability to make more accurate prediction of real-value RSA and  $\psi$  backbone angles (Dor and Zhou, submitted). It is not clear whether this signals that SPINE has reached the upper bound obtainable from sequence profiles and the properties of amino acid residues for a large dataset.

## CONCLUSION

The strong dependence of  $Q_3$  accuracy on the size of the training database suggests that an upper limit for the existing technology of secondary-structure prediction is not yet reached. That is, one can further improve the accuracy of secondary-structure prediction by using a larger dataset when it is available. The datasets, executable versions, and web servers are available at <http://sparks.informatics.iupui.edu>. In the server, predicted results are displayed along with a reliability index, that is 0 if two predictors ( $A$  and  $B$ ) disagree and  $(A + B)/2$  if the two predictors predict the same state.

## ACKNOWLEDGMENTS

We gratefully thank Dr. Dong Xu and Mr. Rajkumar Bondugula for sending us a preprint of their paper and for helpful discussion. We also thank the authors who made their programs and databases available for comparison.

## REFERENCES

1. Scheraga HA. Structural studies of ribonuclease III. A model for the secondary and tertiary structure. *J Am Chem Soc* 1960;82:3847–3852.
2. Finkelstein AV, Ptitsyn OB. Statistical analysis of the correlation among amino acid residues in helical,  $\beta$ -structural and non-regular regions of globular proteins. *J Mol Biol* 1971;62:613–624.
3. Chou PY, Fasman UD. Prediction of protein conformation. *Biochem* 1974;13:211–215.
4. Kabat EA, Wu TT. The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of  $\beta$ -sheets in conavalin A. *Proc Natl Acad Sci USA* 1973;70:1473–1477.
5. Maxfield FR, Scheraga HA. Status of empirical methods for the prediction of protein backbone topography. *Biochem* 1976;15:5138–5153.
6. Holley HL, Karplus M. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA* 1989;86:152–156.
7. Arnold GE, Dunker AK, Johns SJ, Douthart RJ. Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. *Proteins* 1992;12:382–399.
8. Rost B. Review: protein secondary structure prediction continues to rise. *J Struct Biol* 2001;134:204–218.
9. Simossis VA, Heringa J. Integrating secondary structure prediction and multiple sequence alignment. *Curr Protein Pept Sci* 2004;5:1–15.

10. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 1988;202:865–884.
11. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 1993;90:7558–7562.
12. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
13. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;40:502–511.
14. Chandonia G, Karplus M. New methods for accurate prediction of protein secondary structure. *Proteins* 1999;35:293–306.
15. Petersen T, Lundegaard C, Nielsen M, Boher H, Boher J, Brunak S, Gippert G, Lund O. Prediction of protein secondary structure at 80% accuracy. *Proteins* 2000;41:17–20.
16. Pollastri G, Przybylski D, Rost B, Baldi B. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 2002;47:228–235.
17. Pollastri G, McLysaght A, Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* 2005;21:1719–1720.
18. Lin K, Simossis V, Taylor W, Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* 2005;21:152–159.
19. Pan X. Multiple linear regression for protein secondary structure prediction. *Proteins* 2001;43:256–259.
20. Qin S, He Y, Pan X. Predicting protein secondary structure and solvent accessibility with an improved multiple linear regression method. *Proteins* 2005;61:473–480.
21. Sim J, Kim S, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics* 2005;21:2844–2849.
22. Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Eng* 2003;16:553–560.
23. Hu H, Pan Y, Harrison R, Tai P. Improved protein secondary structure prediction using support vector machine with a new encoding scheme and an advanced tertiary classifier. *IEEE Trans Nanobiosci* 2004;3:265–271.
24. Wang L, Li Y, Liu J, Zhou H. Predicting protein secondary structure by a support vector machine based on a new coding scheme. *Genome Inform* 2004;15:181–190.
25. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. *J. Mol Biol* 1992;225:1049–1063.
26. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584–599.
27. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. Jpred: a consensus secondary structure prediction server. *Bioinformatics* 1998;14:892–893.
28. King RD, Ouali M, Strong AT, Aly A, Elmaghraby A, Kantardzic M, Page D. Is it better to combine predictions? *Protein Eng* 2000;13:15–19.
29. Liu Y, Carbonell J, Klein-Seetharaman J, Gopalakrishnan V. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics* 2004;20:3099–3107.
30. Lin H, Chang J, Wu K, Sung T, Hsu W. Hyprosp II-a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 2005;21:3227–3233.
31. Midic U, Dunker K, Obradovic Z. Improving protein secondary structure prediction by predicting ends of secondary-structure Segments. In: *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB*, 2005;490–497.
32. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;20:216–226.
33. Hobrook S, Mushal S, Kim S. Predicting surface exposure of amino acids from protein sequence. *Protein Eng*. 1990;3:659–665.
34. Pascarella S, De Persio R, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins* 1999;32:190–199.
35. Li X, Pan X-M. New method for accurate prediction of solvent accessibility from protein sequence. *Proteins* 2001;42:1–5.
36. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 2002;47:142–153.
37. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 2002;18:819–824.
38. 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.
39. Raih M, Ahmad S, Zheng R, Rahmah M. Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability. *Biophys Chem* 2005;114:63–69.
40. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. *Proteins* 2002;48:566–570.
41. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins* 2004;57:558–564.
42. Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. *Protein Eng* 2003;16:987–992.
43. 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.
44. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins* 2005;59:467–475.
45. Wood MJ, Hirst JD. Protein secondary structure prediction with dihedral angles. *Proteins* 2005;59:476–481.
46. Dunbrack R. A protein sequence culling server (cullpdb\_pc25\_res3.0). [http://dunbrack.fccc.edu/Guoli/pisces\\_download.php](http://dunbrack.fccc.edu/Guoli/pisces_download.php) 2006.
47. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577–2637.
48. Ahmad S, Gromiha M, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* 2003;50:629–635.
49. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. *Proteins* 2001;42:452–459.
50. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng* 2000;13:607–609.
51. Cuff J, Barton G. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;40:502–511.
52. 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. *Nucl Aci Res* 1997;25:3389–3402.
53. 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.
54. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004;56:753–767.
55. Boden M, Yuan Z, Bailey T. Prediction of protein continuum secondary structure with probabilistic models based on NMR solved structures. *BMC Bioinformatics* 2006;7:68–91.
56. Nguyen M, Rajapakse J. Prediction of protein relative solvent accessibility with a two-stage svm approach. *Proteins* 2005;59:30–37.
57. Nguyen M, Rajapakse J. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins* 2006;63:542–550.
58. Sadeghi M, Parto S, Arab S, Ranjbar B. Prediction of protein secondary structure based on residue pair types and conformational states using dynamic programming algorithm. *FEBS Lett* 2005;1:3397–3400.
59. Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci* 2005;14:1955–1963.
60. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA* 2003;100:12105–12110.