PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SANN: Solvent accessibility prediction of proteins by nearest neighbor method

Keehyoung Joo,[1,2] Sung Jong Lee,[1,3]* and Jooyoung Lee[1,4]*

[1] Center for In Silico Protein Science, Korea Institute for Advanced Study, 130-722, Korea
[2] Center for Advanced Computation, Korea Institute for Advanced Study, 130-722, Korea
[3] Department of Physics, University of Suwon, Hwaseong-Si 445-743, Korea
[4] School of Computational Sciences, Korea Institute for Advanced Study, 130-722, Korea

ABSTRACT

We present a method to predict the solvent accessibility of proteins which is based on a nearest neighbor method applied to the sequence profiles. Using the method, continuous real-value prediction as well as two-state and three-state discrete predictions can be obtained. The method utilizes the *z*-score value of the distance measure in the feature vector space to estimate the relative contribution among the *k*-nearest neighbors for prediction of the discrete and continuous solvent accessibility. The Solvent accessibility database is constructed from 5717 proteins extracted from PISCES culling server with the cutoff of 25% sequence identities. Using optimal parameters, the prediction accuracies (for discrete predictions) of 78.38% (two-state prediction with the threshold of 25%), 65.1% (three-state prediction with the thresholds of 9 and 36%), and the Pearson correlation coefficient (between the predicted and true RSA's for continuous prediction) of 0.676 are achieved An independent benchmark test was performed with the CASP8 targets where we find that the proposed method outperforms existing methods. The prediction accuracies are 80.89% (for two state prediction with the threshold of 25%), 67.58% (three-state prediction), and the Pearson correlation coefficient of 0.727 (for continuous prediction) with mean absolute error of 0.148. We have also investigated the effect of increasing database sizes on the prediction accuracy, where additional improvement in the accuracy is observed as the database size increases. The SANN web server is available at http://lee.kias.re.kr/~newton/sann/.

Proteins 2012; 80:1791–1797.

Key words: solvent accessibility prediction; sequence analysis; sequence profile; machine learning.

INTRODUCTION

As Lee and Richards[1] introduced the concept of solvent accessibility, it has been considered as an important quantitative measure for studying the three-dimensional (3D) structures of proteins. The solvent accessibility is particularly important in that it is related to the spatial arrangement and packing of amino acid residues during the process of protein folding. It is also closely related to the protein–protein or protein–ligand interactions and thus to the protein functions.

Even without proper understanding of the protein folding mechanism, useful insights on possible 3D conformations of a target protein can be obtained from its predicted solvent accessibility during the process of protein structure modeling[2–4] including, for example, quality assessment of protein models (for finding the near-native structures),[5,6] and detection as well as threading of remote homologous proteins in template based modeling.[7]

The prediction of protein solvent accessibility has been performed mostly in a discrete fashion as in the two-state or three-state classification of the degree of exposure of amino acids. Proposed methods are based on neural networks,[8–17] support vector machines,[18–19] information theory-based method,[20] multiple linear regression method,[21] bayesian statistics,[22] and nearest neighbor method,[23] and so forth.

As the prediction accuracy of a discrete method depends on the threshold criterion between, for example, exposed, intermediate and buried states, attempts were also made on direct prediction of the continuous real

Grant sponsor: Creative Research Initiatives of MEST/KOSEF (Center for *in-silico* Protein Science; Grant number: 2009-0063610; Grant sponsor: National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST); Grant number: 2009-0090085.
*Correspondence to: J. Lee. E-mail: jlee@kias.re.kr or S. J. Lee.
E-mail: sjree@suwon.ac.kr
Received 19 August 2011; Revised 8 February 2012; Accepted 23 February 2012
Published online 20 March 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24074

value of the solvent accessibility.[13–15,17,24] It is now well-established that inclusion of profile information from homologous proteins considerably enhances the prediction accuracy.[14,15,17,25]

In this work, we present a method to predict the solvent accessibility of proteins which is based on a nearest neighbor method applied to the sequence profile.[23] A database is constructed from a set of existing PDB proteins where each residue is represented by a feature vector generated from the sequence profile of the residue and its neighbors in sequence. After a distance measure between two feature vectors are properly defined, the solvent accessibility of a target residue can be predicted from its top k nearest residues in the database. One important characteristics of our method is that we use the z-score value of the distance measure to evaluate the relative contributions from the k nearest residues for discrete as well as continuous prediction of solvent accessibility (see below for details). The z-score has an advantage that it can be used as a reliability measure of the prediction.

METHODS

To build the solvent accessibility database, we used PISCES culling server[26] with 25% sequence identity cutoff including X-ray (less than 3.0 Å resolution and 0.3 of R-factor) and NMR structures which contain more than 50 residues. As a result, we have collected 5717 protein chains with 12,42,356 residues where we filtered residues corresponding to chain breaks (April 2008).

The relative solvent accessibility (RSA) of a residue in a protein chain is calculated by dividing the accessible surface area from DSSP[27] by the maximum solvent accessibility according to Chothia's work[28] which uses Gly-X-Gly extended tripeptides. In units of Å^2, these are 210 (Phe), 175 (Ile), 170 (Leu), 155 (Val), 145 (Pro), 115 (Ala), 75 (Gly), 185 (Met), 135 (Cys), 255 (Trp), 230 (Tyr), 140 (Thr), 115 (Ser), 180 (Gln), 160 (Asn), 190 (Glu), 150 (Asp), 195 (His), 200 (Lys), and 225 (Arg). With the RSA value, the classification is carried out either in two states [buried (B) and exposed (E)] or in three states [buried (B), intermediate (I) and exposed (E)] as in the literature. We have tried four thresholds of 0, 5, 16, and 25% in the two-state classification, and one set of thresholds of 9% for B/I and 36% for I/E in the three-states one.

For the 5717 protein chains, we generated profiles by using PSI-BLAST[29] with default parameters (3 iterations and 0.001 of E-value cutoff). Using these profiles, we constructed a database of feature vectors from each sliding window of 15 residues centered on the target residue resulting in a matrix of dimensions 15×20.[19,30,31] Each feature vector is labeled with its RSA value as well as with its discrete states classified according to various threshold criteria mentioned above.

The distance measure between two feature vectors A and B is defined as

$$d^{AB} = \sum_{i,j} w_i |P_{ij}^A - P_{ij}^B| \quad (1)$$

where P_{ij}^A ($i = 1,2,\ldots,15$; $j = 1,2,\ldots,20$) is the component of the feature vector A, and w_i is the weight parameter corresponding to the residue position i (with the central residue corresponding to $i = 8$) along the sequence. We choose w_i as $w_i = (8 - |8 - i|)^2$ which gives maximal weight to the central residue, and which decreases in proportion to the square of the distance towards the ends of the window.[23,31]

To predict the solvent accessibility of a given residue, we compute all pairwise distances between the feature vector of the target residue and all the feature vectors in the database. Among these feature vectors of the database, k nearest neighbors are selected based on the distance measure. And then, we calculate the z-value defined as $z = (d_{\text{ave}} - d)/\sigma$ (where d_{ave} is the average distance between the target residue and the residues in the database and σ is the standard deviation) for each distance d. The z-value is the negative of the standard z-score in the usual statistical distribution. We note that z represents the degree of relative significance of the neighbor under inspection as compared with all the other samples. That is, z can be considered as a universal relative measure of the closeness between the query residue and residues in the database in terms of the solvent accessibility.

For discrete prediction (two states or three states), the contribution of each nearest neighbor to state s (B/E for two states, B/I/E for three states) is defined as

$$u_s = \sum_{i=1}^{k} z_i^{\alpha} \delta_{s_i s}, \quad (2)$$

where the summation is over all selected k neighbors (for $z > 0$), the exponent α is an additional parameter of the method for tuning the relative importance of each neighbor, and $\delta_{s_i s}$ is 1 (if $s_i = s$) or 0 (if $s_i \neq s$).

The probability of the target residue to be in state s is estimated as $p_s = u_s / \sum_s u_s$, and the prediction is assigned to the maximum u_s state. The parameters k and α of the method can be determined by grid search using the leave-one-out procedure so that optimal average accuracy is obtained (see the Result section below). We note that the special case of $\alpha = 0$ corresponds to the standard k-nearest neighbor method where a simple majority rule determines the predicted state s.

For the continuous real-value prediction of solvent accessibility, the predicted RSA value for a target residue is defined as the weighted average

$$\text{RSA} = \frac{\sum_{i=1}^{k} \text{RSA}_i z_i^{\alpha}}{\sum_{i=1}^{k} z_i^{\alpha}}, \quad (3)$$

where RSA_i denotes the RSA (via DSSP) of the ith nearest residue in the database. The real value solvent accessi-

bility for a target residue is estimated as RSA multiplied by the maximum accessible surface area of the residue. The parameters k and α in the continuous prediction can be determined by a similar procedure as in the discrete state prediction by optimizing the Pearson's correlation coefficient between predicted RSA values and the observed values.

Table I
The Maximum Accuracy Values of Q_3 and MCC Scores from their Respective Optimal Parameters (k,α) are Shown for the Three-State Prediction (Buried/Intermediate/Exposed) of Solvent Accessibility

Thresholds (3 state)	Q_3(%) (k,α)	MCC (k,α) (B)	(I)	(E)	
9%, 36%	65.10 (64, 6.31)	0.546	0.203	0.517	(109, 6.83)

RESULTS AND DISCUSSION

Determination of optimal parameters by grid search using leave-one-out procedure

To determine optimal parameter values of k and α, we have used grid search method and leave-one-out procedure. We considered the integer k value in the range 1–150 and α value in the range of 0–10 with an interval of 0.01. For a set of k and α, we predict all the discrete states (one three-state and four two-state predictions according to various threshold criteria mentioned above) and RSA of a residue by using the Eqs. (2) and (3). The leave-one-out procedure is carried out in the fashion that the solvent accessibility prediction of all residues from each protein chain is performed using the database containing all residues from the remaining 5716 chains. Accuracy of prediction is evaluated using the DSSP value as the gold standard. For discrete prediction of each chain, the accuracy score (Q_2 for two states or Q_3 for three states) and Matthew's correlation coefficient (MCC) are calculated as

$$Q_i = N_c/N, \tag{4}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{5}$$

where N is the total number of residues in a chain, and N_c is the correctly predicted number of residues (i is 2 for two-state prediction or 3 for three-state prediction). TP, TN, FP, and FN are the numbers of the the true positives, true negatives, false positives, and false negatives, respectively. We note that there are three MCC scores for three states (B/I/E), and the MCC score for buried (B) is the same with the corresponding one for exposed (E) for two-state prediction. For continuous real-value prediction of each chain, the Pearson's correlation coefficient (Corr) between the predicted RSA

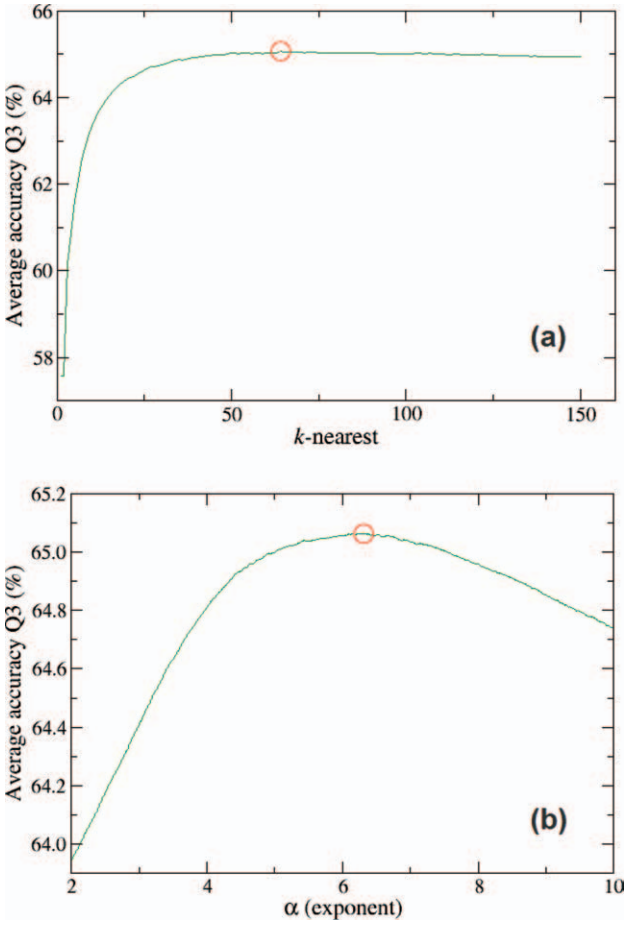

Figure 1
The grid search result for overall accuracies of Q_3 scores is shown as a function of parameters k and α. The optimal result of Q_3 = 65.10% is obtained with k = 64 (**a**) and α = 6.31 (**b**). We observe that Q_3 is more sensitive with α than k, and for k larger than about 50, Q_3 is almost constant with the value of 65%. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table II
The Maximum Accuracy Values of Q_2 and MCC Scores from their Respective Optimal Parameters (k, α) are Shown for the Two-State Prediction (Buried/Exposed)

Thresholds (2 state)(%)	Q_2(%) (k,α)	MCC (k,α)
0	90.88 (55, 5.07)	0.509 (150, 6.71)
5	85.50 (90, 6.41)	0.490 (100, 7.33)
16	80.36 (44, 5.95)	0.533 (66, 6.42)
25	78.38 (49, 6.36)	0.562 (63, 5.90)

Four separate threshold criteria are used.

Table III
The Overall Pearson's Correlation Coefficient Score for Optimal Parameters k and α in the Continuous Real-Value Prediction of RSA

Correlation	(k,α)
0.676	(83, 6.79)

values x and the observed RSA values y is used for accuracy measure as follows,

$$\mathrm{Corr}(x,y) = \frac{\sum(x_i - < x >)(y_i - < y >)}{\sqrt{\sum(x_i - < x >)^2}\sqrt{\sum(y_i - < y >)^2}}. \quad (6)$$

We have calculated average scores for all chains in the database to find optimal values of k and α for all discrete predictions as well as the RSA-value prediction.

Figure 1 shows the results of Q_3 optimization to obtain the optimal values of k and α, 64 and 6.31, respectively. From the figure, we observe that the accuracy is more sensitive with α than k. For values of k greater than about 50, overall accuracy is almost constant with about 65%. Table I summarizes the optimal accuracies (Q_3) and the Matthew correlation coefficients (MCC) for three-state prediction together with the corresponding optimal parameters of k and α. We get the overall accuracy of about 65.1% for Q_3. From the MCC values, we observe that, out of the three states, the buried (B) states can be predicted with the highest accuracy. In contrast, the prediction of the intermediate (I) states is least accurate, which can be probably expected from the arbitrariness of the threshold between the three states. Table II shows the overall accuracies and MCC values for two-state predictions with four separate definitions of the threshold value between the buried and exposed states. The four separate thresholds are 0, 5, 16, and 25, respectively. Here, we note that the optimal parameters k and α are separately determined for each value of the threshold. We observe that the MCC value increases as the threshold value increases. We see that there exists a strong dependence of the prediction accuracy on the value of the boundary thresholds between the two states. With the threshold of 25%, we obtain the accuracy of 78.38% for Q_2.

Table V
The Efficiency of SANN for Continuous Real-Value Prediction of RSA is Compared with Existing Methods

	Full chain (121 targets)		Domain (165 domains)	
Methods	Corr	MAE	Corr	MAE
SARpred	0.589	0.190	0.578	0.200
Real-Spine3	0.678	0.189	0.665	0.194
SANN	**0.727**	**0.148**	**0.707**	**0.153**

Pearson's correlation score (Corr) and mean absolute error (MAE) are calculated using all 121 CASP8 targets. Best values are shown in the bold face.

For the case of continuous real-value prediction of the RSA, Table III shows the value of 0.676 for the overall Pearson's correlation coefficient for the optimal parameters of $k = 83$ and $\alpha = 6.79$.

Accuracy comparisons from the CASP8 benchmark test

During the CASP8 experiment (summer of 2008), a total of 128 protein sequences were released to public for testing the prediction capability of various methods for tertiary structure as well as other structural features. During the CASP8 season, we predicted the solvent accessibility values for all targets. After the season, we compared the accuracies of our method SANN presented in this work for two/three state predictions as well as the real-value RSA with other existing methods.

Table IV summarizes the results of five existing methods of discrete state predictions (FKNN,[23] SABLE,[14] PROF,[32] ACCpro,[11] and NETASA[12] together with those of SANN as applied to the CASP8 121 targets (seven out of 128 are canceled) containing 165 domains. We observe that for almost all the categories, SANN outperforms the other methods, the only exception being the MCC value of the three-state prediction for the exposed (E) state in comparison with PROF. Average accuracy of SANN is 67.58% for Q_3 evaluation and 80.89% for Q_2 using the criterion of 25%.

For the case of continuous real-value prediction, we compared SANN with SARpred[15] and Real-Spine3.[17] In addition to the Pearson's correlation coefficient, we also

Table IV
The Efficiency of SANN for Two-State and Three-State Prediction of Solvent Accessibility is Compared with Existing Methods

		Q_2 (%)				MCC (Three state)			MCC (Two state)			
Methods	Q_3 (%)	0%	5%	16%	25%	(B)	(I)	(E)	0%	5%	16%	25%
FKNN	63.70	89.44	83.18	79.07	78.29	0.527	0.205	0.500	0.465	0.576	0.584	0.590
SABLE	58.67	–	82.83	78.22	76.21	0.495	0.217	0.442	–	0.444	0.573	0.561
PROF	60.79	–	–	77.96	77.69	0.475	0.165	**0.680**	–	–	0.549	0.558
ACCpro	–	89.92	81.83	79.40	76.21	–	–	–	0.395	0.395	0.584	0.567
NETASA	–	88.46	77.23	–	69.33	–	–	–	0.036	0.310	–	0.393
SANN	**67.58**	**91.15**	**86.95**	**82.88**	**80.89**	**0.593**	**0.273**	0.538	**0.469**	**0.627**	**0.638**	**0.625**

Q values and MCC values are calculated using all 121 CASP8 targets. Best values are shown in the bold face.

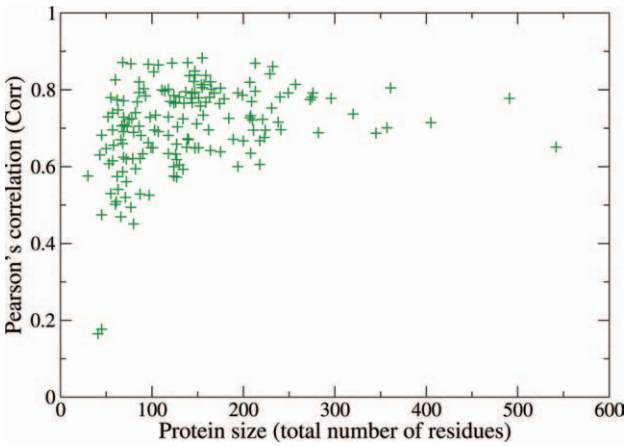

Figure 2

Pearson's correlation coefficient versus domain size (number of residues) is shown for the 165 domains of CASP8 dataset. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

calculated the mean absolute error (MAE) defined as $\mathrm{MAE} = \sum_i |RSA_i - RSA_i^{\mathrm{DSSP}}|/N$, where N is the total number of residues, i is the residue index for the CASP8 dataset and $\mathrm{RSA}^{\mathrm{DSSP}}$ is the RSA calculated by DSSP program using the native structure. From Table V, we observe that the proposed method SANN outperforms the other methods. Average correlation scores of SANN are 0.727 and 0.707 for 121 full-length chains and 165 domains, respectively.

Figure 2 shows the scatter plot of Pearson's correlation coefficients for all 165 target domains from the CASP8 dataset. We observe that the correlation score fluctuates more widely for small domains probably due to the statistics of small number of residues. In SANN, there is no consideration of the size of the protein chain. All profiles are equally treated when determining the nearest neighbors and also during the final prediction stage. However, we expect that the fraction of the surface residues decreases as the protein chain size increases. More rigorous prediction methods would take this size dependence of the average solvent exposure into account in the prediction procedure.

Effect of increasing database size on the prediction accuracy

In this section, we examined the effect of the database size on the prediction accuracy. By using higher sequence identity cutoff levels of the PISCES culling server, the database size can be effectively increased. We have measured the prediction accuracies of SANN using various levels of the sequence identity cutoff of 30, 35, and 40% by collecting effectively additional protein chains as the percentage increases. We did not consider higher than 40% sequence identity cutoffs since it requires much more computational costs to find the optimal parameters in the grid search. For example, there are 22,51,422 residues in the 40 % protein chain database. With the database constructed using the 40 % sequence identity cutoff value, SANN attained the accuracy of 69.15% for Q_3 score, the Pearson's correlation coefficient of 0.718, and the mean absolute error of 0.149. Although the redundancy of database is increased, additional improvements (although small) are obtained for both discrete and continuous predictions (see Table VI). This probably indicates that additional information can be extracted from the increased database.

A highlight for target T0486-D1 from the CASP8 test set

Figure 3 shows a successful example of the solvent accessibility prediction for T0486-D1 from CASP8, where the results of three-state prediction together with the actual (DSSP) result is shown in Figure 3(a). Q_3 score is 76.53% and the correlation score is 0.88 using the database constructed with the 35% sequence identity cutoff value. In Figure 3(c), the continuous real-value prediction value of RSA and the actual continuous value obtained by DSSP are shown. Good correlation between the predicted values and the true values is achieved. We observe that the hydrophobic core regions shown in grey from Figure 3(b) are well predicted. In the actual 3D structure prediction of this target (as well as other CASP8 targets), near the final stage of structure modeling, quality assessment procedure (based on SVM method) was performed with scoring functions depending on structural features including the predicted solvent

Table VI

Prediction Accuracies on the 121 CASP8 Targets are Measured While the Database Size is Varied using Various Sequence Identity Cutoff Values to Construct the Database

SANN (%)		Q_2 (%)				MCC (3 state)			MCC (2 state)				(165 domains)	
(Seq. Id.)	Q_3 (%)	0%	5%	16%	25%	(B)	(I)	(E)	0%	5%	16%	25%	Corr	MAE
25	67.58	91.15	86.95	82.88	80.89	0.593	0.273	0.538	0.469	0.627	0.638	0.625	0.707	0.153
30	68.29	91.08	87.09	83.30	81.35	0.607	0.292	0.553	0.474	0.635	0.648	0.636	0.713	0.151
35	68.61	91.22	87.22	83.56	81.57	0.614	0.300	0.558	0.482	0.639	0.654	0.638	0.715	0.150
40	**69.15**	**91.30**	**87.63**	**83.90**	**81.88**	**0.627**	**0.315**	**0.571**	**0.503**	**0.656**	**0.665**	**0.648**	**0.718**	**0.149**

The best result is obtained with the largest database constructed using 40% sequence identity cutoff. Best values are shown in the bold face.

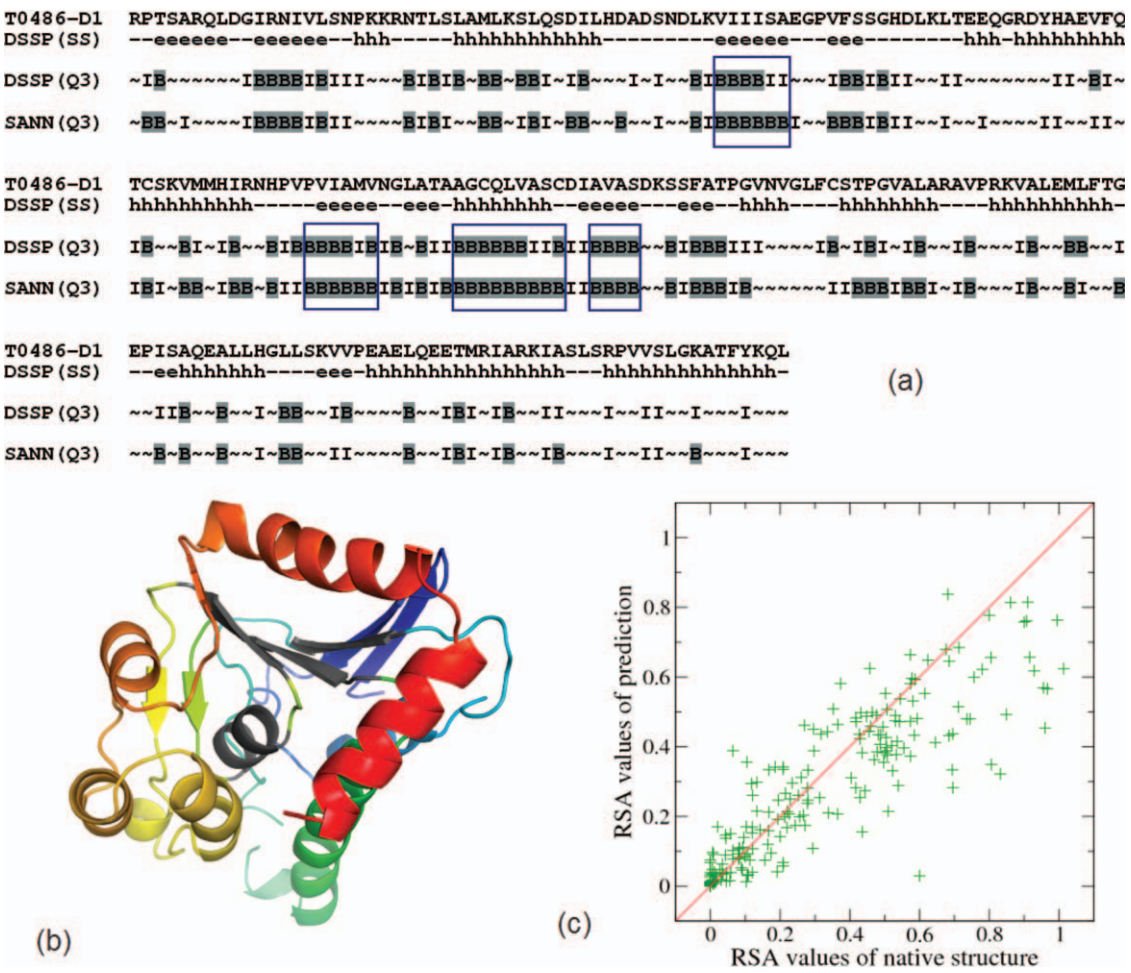

Figure 3

SANN prediction results of CASP8 T0486-D1 in three states are shown. The database is constructed using 35% sequence identity. (**a**) Protein sequence, secondary structure elements (-:coil, e:extended, h:helix), three-state classification (B:buried, I:intermediate, ~:exposed) of solvent accessibility from DSSP, and the three-state prediction of solvent accessibility by SANN are shown. The Q_3 score is 76.53%. The boxed regions are buried which are shown in grey from the cartoon figure (**b**) of the native structure. The RSA values of SANN prediction and the actual value calculated by DSSP using the native structure are shown together (**c**). The Pearson's correlation coefficient score is 0.88 and most buried residues are well predicted with the RSA values near zero.

accessibility (as shown above) in order to discriminate more native-like conformations out of many candidate model structures.

CONCLUSION

In this work, we have presented SANN to predict the discrete state (two and three states) as well as continuous value of the solvent accessibility of a target residue. The method is based on a *k*-nearest neighbor method combined with *z*-value distance statistics in the feature vector space. The overall performance of SANN is excellent and comparable to the best ones currently available.

We also performed additional benchmark tests of SANN on CASP9 targets. On 116 CASP9 targets, the average Q_3 score of SANN is 66.44% to be compared with 62.13% of FKNN. Similary, the Q_2 accuracy (with 16% threshold) of SANN is 81.54% which is better than 72.2% from ACCpro. Pearson's correlation coefficient for real value prediction of SANN is 0.686 that can be compared with 0.661 from Real-Spine3.

SANN has been useful in screening 3D models during protein structure modeling in CASP.[2,33] By increasing the size of the database via varying the sequence identity cutoff value, further increase of the prediction accuracies was observed.

In this work, we adopted DSSP program as the standard to calculate accessible surface areas from protein structures to directly compare our results to those from existing methods. However, we note that there is another program called NACCESS.[34] NACCESS calculates RSA

values by setting Ala-X-Ala tri-peptide as the reference of the maximally accessible surface area. We found that the Pearson correlation coefficient between the maximum surface areas of twenty amino acids for DSSP and NACCESS is 0.979 and that the average Pearson correlation between the absolute surface areas for 121 CASP8 protein native structures (of the two methods) is 0.995. Therefore, we expect that the results from this work would hold with NACCESS.

Currently, the size of the protein chain is not considered as a factor in SANN. As the size of a protein chain increases, the fraction of the surface residues (and also the average accessible surface area per residue) is expected to decrease. More rigorous prediction methods, thus, would require to take this size dependence of the average solvent exposure into account in the prediction procedure, which might further improve the sensitivity of the method in the future. It is also emphasized that *z*-value statistics of the distance measure exercised in this work can be utilized as a useful reliability measure of the prediction in the structure modeling.

ACKNOWLEDGMENTS

The authors thank Korea Institute for Advanced Study for providing computing resources (KIAS Center for Advanced Computation Linux Cluster) for this work.

REFERENCES

1. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
2. Joo K, Lee J, Lee S, Seo J-H, Lee SJ, Lee J. High accuracy template based modeling by global optimization. Proteins 2007;69(Suppl 8):83–89.
3. Xu J, Peng J, Zhao F. Template-based and free modeling by raptor++ in casp8. Proteins 2009;77:133–137.
4. Wang Z, Eickholt J, Cheng J. Multicom: a multi-level combination approach to protein structure prediction and its assessments in casp8. Bioinformatics 2010;26:882–888.
5. Benkert P, Tosatto SCE, Schwede T. Global and local model quality estimation at casp8 using the scoring functions qmean and qmeanclust. Proteins 2009;77(Suppl 9):173–180.
6. Cheng J, Wang Z, Tegge AN, Eickholt J. Prediction of global and local quality of casp8 models by multicom series. Proteins 2009;77(Suppl 9):181–184.
7. Peng J, Xu J. Low-homology protein threading. Bioinformatics 2010;26:i294–i300.
8. Holbrook SR, Muskal SM, Kim S-H. Predicting surface exposure of amino acids from protein sequence. Protein Eng 1990;3:659–665.
9. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
10. 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.
11. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
12. Ahmad S, Gromiha MM. Netasa: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
13. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
14. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753–767.
15. Garg A, Kaur H, Raghava GPS. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005;61:318–324.
16. Dor O, Zhou Y. Real-spine: an integrated system of neural networks for real value prediction of protein structural properties. Proteins 2007;68:76–81.
17. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 2009;74:857–871.
18. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566–570.
19. 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.
20. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi A. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
21. Li X, Pan X-M. New methods for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
22. Thompson MJ, Godstein RA. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
23. Sim J, Kim S-Y, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. Bioinformatics 2005;21:2844–2849.
24. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;57:558–564.
25. Rost B, Sander C. Prediction of protein secondary structure at better than 70 % accuracy. J Mol Biol 1993;232:584–589.
26. Wang G, Dunbrack RL. Pisces: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
27. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
28. Chothia C. The nature of accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.
29. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
30. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices1. J Mol Biol 1999;292:195–202.
31. Joo K, Kim I, Kim S-Y, Lee J, Lee J, Lee SJ. Prediction of the secondary structures of proteins by using predict, a nearest neighbor method on pattern space. J Kor Phys Soc 2004;45:1441–1449.
32. Rost B, Yachdav G, Liu J. The predictprotein server. Nucleic Acids Res 2004;32 Web Server issue: W321–W326.
33. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in casp8. Proteins 2009;77(Suppl 9):114–122.
34. Hubbard SJ, Thornton JM. NACCESS, computer program, Department of Biochemistry and Molecular Biology, University College London, 1993.