BIOINFORMATICS **ORIGINAL PAPER**

Vol. 21 no. 12 2005, pages 2844–2849
doi:10.1093/bioinformatics/bti423

Structural bioinformatics

Prediction of protein solvent accessibility using fuzzy *k*-nearest neighbor method

Jaehyun Sim[1], Seung-Yeon Kim[2] and Julian Lee[1,*]

[1]Department of Bioinformatics and Life Science, Bioinformatics and Molecular Design Technology Innovation Center, and Computer Aided Molecular Design Research Center, Soongsil University, Seoul 156-743, South Korea and [2]School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, South Korea

Received on October 14, 2004; revised on January 25, 2005; accepted on March 30, 2005
Advance Access publication April 6, 2005

ABSTRACT

Motivation: The solvent accessibility of amino acid residues plays an important role in tertiary structure prediction, especially in the absence of significant sequence similarity of a query protein to those with known structures. The prediction of solvent accessibility is less accurate than secondary structure prediction in spite of improvements in recent researches. The *k*-nearest neighbor method, a simple but powerful classification algorithm, has never been applied to the prediction of solvent accessibility, although it has been used frequently for the classification of biological and medical data.

Results: We applied the fuzzy *k*-nearest neighbor method to the solvent accessibility prediction, using PSI-BLAST profiles as feature vectors, and achieved high prediction accuracies. With leave-one-out cross-validation on the ASTRAL SCOP reference dataset constructed by sequence clustering, our method achieved 64.1% accuracy for a 3-state (buried/intermediate/exposed) prediction (thresholds of 9% for buried/intermediate and 36% for intermediate/exposed) and 86.7, 82.0, 79.0 and 78.5% accuracies for 2-state (buried/exposed) predictions (thresholds of each 0, 5, 16 and 25% for buried/exposed), respectively. Our method also showed slightly better accuracies than other methods by about 2–5% on the RS126 dataset and a benchmarking dataset with 229 proteins.

Availability: Program and datasets are available at http://biocom1.ssu.ac.kr/FKNNacc/

Contact: jul@ssu.ac.kr

INTRODUCTION

Predicting the three-dimensional (3D) structure of a protein from its sequence is an important issue because the gap between the enormous number of protein sequences and the number of experimentally determined structures has increased (Rost and Sander, 1994; Thompson and Goldstein, 1996). However, the prediction of the complete 3D structure of a protein is still a big challenge, especially in the case where there is no significant sequence similarity of a query protein to those with known structures (Ginalski and Rychlewski, 2003; John and Sali, 2004; Moult *et al*., 2003; Sander and Schneider, 1991). The prediction of solvent accessibility and secondary structure has been studied as an intermediate step for predicting the tertiary structure of proteins, and the development of knowledge-based approaches has helped to solve these problems (Cuff and Barton, 2000; Frishman and Argos, 1997; Jones, 1999; Przybylski and Rost, 2002; Wohlfahrt *et al*., 2002).

Secondary structures and solvent accessibilities of amino acid residues give a useful insight into the structure and function of a protein (Eyal *et al*., 2004; Russell *et al*., 2003; Totrov, 2004; Wohlfahrt *et al*., 2002). In particular, the knowledge of solvent accessibility has assisted alignments in regions of remote sequence identity for threading (Rost and Sander, 1994; Rost *et al*., 1997). However, in contrast to the secondary structure, there is no widely accepted criterion for classifying the experimentally determined solvent accessibility into a finite number of discrete states such as buried, intermediate and exposed states. Also, the prediction accuracies of solvent accessibilities are lower than those for secondary structure prediction, since the solvent accessibility is less conserved than secondary structure (Rost and Sander, 1994), although there has been some progress recently.

The prediction of solvent accessibility, as well as that of the secondary structure, is a typical pattern classification problem. The first step for solving such a problem is the feature extraction, where the important features of the data are extracted and expressed as a set of numbers, called feature vectors. The performance of the pattern classifier depends crucially on the judicious choice of the feature vectors. In the case of the solvent accessibility prediction, using evolutionary information such as multiple sequence alignment and position-specific scoring matrix generally has given good prediction results (Gianese *et al*., 2003; Pei and Grishin, 2004). Once an appropriate feature vector has been chosen, a classification algorithm is used to partition the feature space into disjoint regions with decision boundaries. The decision boundaries are determined using feature vectors of a reference sample with known classes, which are also called the reference dataset or training set. The class of a query data is then assigned depending on the region it belongs to.

Various classification algorithms have been developed. Bayesian statistics is a parametric method where the functional form of the probability density is assumed for each class, and its parameters are estimated from the reference data.

In nonparametric methods, no specific functional form for the probability density is assumed. There are various nonparametric methods such as, for example, neural networks, support vector machines and nearest neighbor methods. In the neural network

[*]To whom correspondence should be addressed.

methods, the decision boundaries are set up before the prediction using a training set. Support vector machines are similar to neural networks in that the decision boundaries are determined before the prediction, but in contrast to neural network methods where the overall error function between the predicted and observed class for the training set is minimized, the margin in the boundary is maximized.

In the k-nearest neighbor method, the decision boundaries are determined implicitly during the prediction, where the prediction is performed by assigning the query data the class most frequently represented among the k-nearest reference data. The standard k-nearest neighbor rule is to place equal weights on the k-nearest reference data for determining the class of the query, but a more general rule is to use weights proportional to a certain power of distance. Also, by assigning the fuzzy membership to the query data instead of a definite class, one can estimate the confidence level of the prediction. The method employing these more general rules is called the fuzzy k-nearest neighbor method (Keller *et al*., 1985; see Methods section for details).

Neural network methods are very popular and have been widely used for solvent accessibility prediction (Adamczak *et al*., 2004; Ahmad and Gromiha, 2002; Ahmad *et al*., 2003; Cuff and Barton, 2000; Pollastri *et al*., 2002; Rost and Sander, 1994), and support vector machines, a recently developed method, shows comparable results to neural network methods (Kim and Park, 2004; Yuan *et al*., 2002; Yuan and Huang, 2004). Bayesian statistics has also been used by Thompson and Goldstein (1996).

The k-nearest neighbor method has been frequently used for the classification of biological and medical data, and despite its simplicity, the performances are competitive compared to many other methods. However, the k-nearest neighbor method has never been applied for predicting solvent accessibility, although it has been used to predict protein secondary structure (Salamov and Solovyev, 1995, 1997; Salzberg and Cost, 1992; Yi and Lander, 1993).

In this work, we apply the fuzzy k-nearest neighbor method to the prediction of solvent accessibility where PSI-BLAST (Altschul *et al*., 1997) profiles are used as the feature vectors. We obtain relatively high accuracy on various benchmark tests.

MATERIALS AND METHODS

Definition and thresholds of solvent accessibility

Amino acid solvent accessibility is the degree to which a residue in a protein is accessible to a solvent molecule. The relative solvent accessibility can be calculated by dividing the DSSP accessibility (Kabsch and Sander, 1983) by the maximum accessibility of amino acid residues (Rost and Sander, 1994) corresponding to the accessibility for a Gly-X-Gly extended tripeptide conformation.

Various thresholds have been used to classify residues as buried (B) and exposed (E) (2-state prediction) or buried (B), intermediate (I) and exposed (E) (3-state prediction) in previously published results. In this paper, thresholds of 9% for B/I and 36% for I/E (9%; 36% thresholds) in the 3-state prediction and thresholds of 0, 5, 16 and 25% for B/E in the 2-state predictions are used.

Feature vector and distance measure

PSI-BLAST (Altschul *et al*., 1997) generates the profile of a protein in the form of a $20 \times N$ position-specific scoring matrix, where N is the length of the sequence. PSI-BLAST is run with default options (−j 3 −h 0.001 −e 10.0) and the non-redundant protein sequence database (ftp://ncbi.nlm.nih.gov/blast/db) filtered by PFILT (Jones, 1999) to mask out regions of low complexity sequence, the coiled coil regions and transmembrane spans. The BLOSUM62 (Henikoff and Henikoff, 1992) substitution matrix is used for PSI-BLAST.

We construct a window of size 15 centered on a target residue (Jones, 1999; Kim and Park, 2004; Yuan *et al*., 2002), and use the profile that falls within this window, a 15×20 matrix, as a feature vector. Then, the distance between two feature vectors A and B is defined as

$$D_{AB} = \sum_{i,j} W_i |P_{ij}^{(A)} - P_{ij}^{(B)}|$$

where $P_{ij}^{(A)} (i = 1, 2, \ldots, 15;\ j = 1, 2, \ldots, 20)$ is a component of the feature vector A, and W_i is a weight parameter. Since we expect the profile elements for residues nearer to the target residue to be more important in determining the local environment of the target residue, we use weights $W_i = (8-|8-i|)^2$.

Reference dataset for predictions

A reference dataset was constructed by clustering the protein chains in ASTRAL SCOP (version 1.63) chain-select-90 subset (Brenner *et al*., 2000). Since ASTRAL SCOP provides only three kinds of chain sequence sets (100, 95 and 90% sequence identity), we used BLASTCLUST (NCBI BLAST 2.2.5, http://www.ncbi.nlm.nih.gov/BLAST/) for sequence clustering to make a non-redundant dataset. BLASTCLUST was run with 25% sequence identity option (−S 25) over an area covering 90% of the length (−L 0.9). We also removed the proteins equal or shorter than 50 residues, and excluded unusual residues or those corresponding to chain breaks. The first protein was selected from each cluster and placed in the reference dataset. Therefore no two proteins have more than 25% sequence identity in the dataset. The dataset clustered with the −L 0.9 option is a more stringent non-redundant set than datasets constructed by normal pairwise sequence identity. The sequence clustering with the option of −L 1.0 in BLASTCLUST is similar to that based on normal pairwise matches. The resulting reference dataset consists of 3644 non-redundant proteins with 854 876 feature vectors.

The RS126 set and the dataset with 229 proteins

The solvent accessibility was predicted for two sets of proteins in order to evaluate the performance of our method. In the first test, the protein residues in the RS126 dataset (Rost and Sander, 1994) were used as queries. The proteins in the RS126 dataset have less than 25% pairwise sequence identity. This set was used to evaluate different methods of solvent accessibility prediction, for example, PHDacc (Rost and Sander, 1994) and other methods (Kim and Park, 2004; Thompson and Goldstein, 1996; Yuan *et al*., 2002). For a rigorous benchmark test, the chains of the reference dataset and the RS126 dataset were split into domains according to the SCOP annotation, and the chains that have any domain with >25% identity to those of the RS126 dataset were excluded from the reference dataset. The resulting set consisted of 3460 proteins with 819 090 feature vectors, which was used as the reference dataset for the benchmark test on the RS126 dataset.

The second set of query proteins was constructed from newly added proteins, >50 residues, in the 90% identity chain set of the ASTRAL SCOP (version 1.65). We removed from the test set the chains having domains with >25% sequence identity to those of the reference dataset. We also removed one of the chains for any pair in the dataset containing domains with >25% sequence identity. The resulting set consisted of 229 protein chains that had <25% sequence identity with the reference dataset and with each other.

Algorithm

The nearest neighbor algorithm is a simple classification algorithm; a query data is classified according to the classification of the nearest neighbor from a database of known classifications, i.e. a reference dataset. A natural generalization of the nearest neighbor algorithm is the so-called k-nearest neighbor algorithm, where the k-nearest samples are selected and the query data is assigned the class most frequently represented among them. A further extension is to weight the k-nearest samples with a certain power of the distance from the query data. Also, instead of assigning a definite class to the query data, one can calculate the fuzzy membership (see below), which can be used to estimate the confidence level of the prediction. The algorithm incorporating these generalizations is called the fuzzy k-nearest neighbor algorithm (Keller *et al*., 1985).

Despite its simplicity, nearest neighbor methods can give competitive performance compared to many other methods. The nearest neighbor methods have been used to predict protein secondary structure (Salamov and Solovyev, 1995, 1997; Salzberg and Cost, 1992; Yi and Lander, 1993) and classify biological and medical data (Kauffman and Jurs, 2001; Singh *et al*., 1996; Vaidyanathan *et al*., 1997). Also it has been reported that performances of classification were improved by using fuzzy k-nearest neighbor algorithms (Bezdek *et al*., 1993; Cabello *et al*., 1991; Huang and Li, 2004; Leszczynski *et al*., 1999; Seker *et al*., 2003; Sokolowska *et al*., 2003). However the k-nearest neighbor method has never been used to predict protein solvent accessibility.

In this work, we apply the fuzzy k-nearest neighbor method to the solvent accessibility prediction. In the fuzzy k-nearest neighbor method, the fuzzy class membership $u_i(x)$ to the class i is assigned to the query data x according to the following equation:

$$u_i(x) = \frac{\sum_{j=1}^{k} u_i(x^{(j)}) D_j^{-2/(m-1)}}{\sum_{j=1}^{k} D_j^{-2/(m-1)}}, \quad i = 1, \dots, c,$$

where m is a fuzzy strength parameter, which determines how heavily the distance is weighted when calculating each neighbor's contribution to the membership value, k is the number of nearest neighbors, and c is the number of classes. Also, D_j is the distance between the feature vector of the query data x and the feature vector of its jth nearest reference data $x^{(j)}$, and $u_i(x^{(j)})$ is the membership value of $x^{(j)}$ to the ith class, which is 1 if $x^{(j)}$ belongs to the ith class, and 0 otherwise. The advantage of the fuzzy k-nearest neighbor algorithm over the standard k-nearest neighbor method is quite clear. The fuzzy class membership $u_i(x)$ can be considered as the estimate of the probability that the query data belongs to class i, and provides us with more information than a definite prediction of the class for the query data. Moreover, the reference samples which are closer to the query data are given more weights, and an optimal value of m can be chosen along with that for k, in contrast to the standard k-nearest neighbor method with a fixed value of $2/(m-1) = 0$. In fact, the optimal value of k and m are found from the leave-one-out cross-validation procedure (see Results section), and the resulting value for $2/(m-1)$ is indeed nonzero.

Performance measures

In this work, two measures are used to evaluate the performance of prediction methods. One is the accuracy, the percentage of correctly classified residues, and the other is the Matthew's correlation coefficients (MCC). These measures can be calculated by the following equations:

$$\text{accuracy} = \frac{\sum_i^c p_i}{N},$$

$$\text{MCC}_i = \frac{p_i n_i - o_i u_i}{\sqrt{(p_i + o_i)(p_i + u_i)(n_i + o_i)(n_i + u_i)}},$$

where N is the total number of residues, and c is the class number. Also, p_i, n_i, o_i and u_i are the number of true positives, true negatives, false positives and false negatives for class i, respectively. The MCCs have the same value for the two classes in the case of the 2-state prediction, i.e. $\text{MCC}_\text{B} = \text{MCC}_\text{E}$.

Implementation of fuzzy k-nearest neighbor algorithm

The program implementing the fuzzy k-nearest neighbor algorithm for protein solvent accessibility prediction was written in ANSI C and run on a Linux machine with the CPU of AMD Athlon MP2400. It occupies 69 MB of disk space including the reference data. The prediction program has two running modes: a normal mode and a fast mode. The normal mode generates the feature vectors from the PSI-BLAST profiles of the reference dataset whenever it calculates the distance between two feature vectors. It requires the execution time of 8.1 s per query residue on average and 69 MB memory. In the fast mode feature vectors are loaded in memory all at once in the beginning. It takes 4.7 s to predict the solvent accessibility of a query residue. The fast mode costs less computational time than the normal mode, but costs more memory, which is 285 MB.

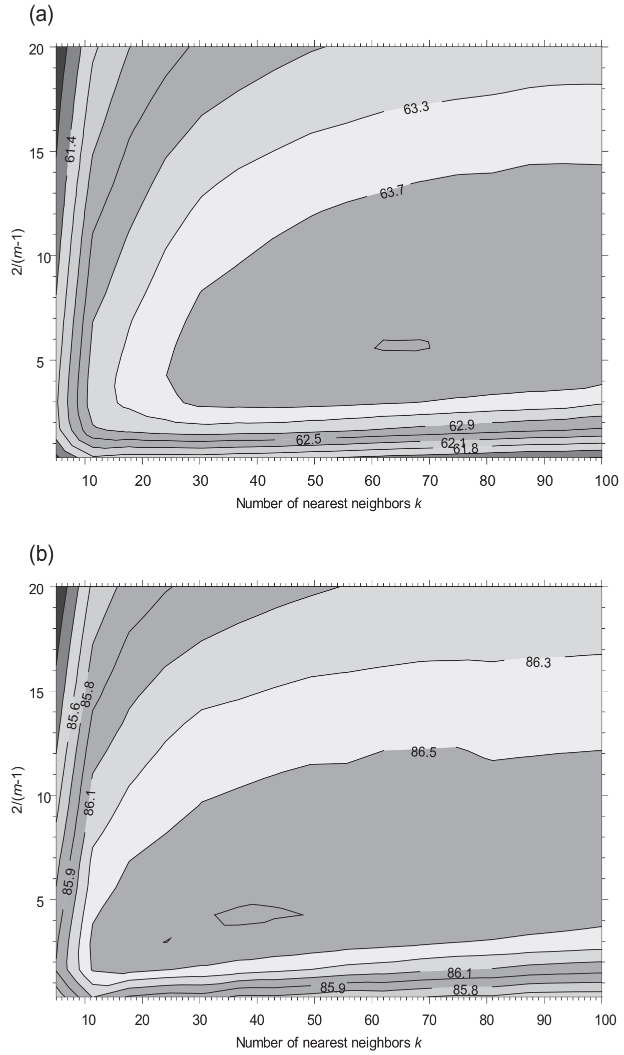

Fig. 1. The contour diagrams of prediction accuracies of leave-one-out cross-validation on the reference dataset derived from ASTRAL SCOP. The contour diagrams of the prediction accuracies of the fuzzy k-nearest neighbor method are shown for (**a**) the 3-state (9%; 36% thresholds) and for (**b**) 2-state (0% threshold) solvent accessibility prediction. The prediction accuracies were calculated by leave-one-out cross-validation on the reference dataset derived from ASTRAL SCOP. The optimal values of the fuzzy strength parameter m and the number of nearest neighbors k are (1.33, 65) for the 3-state prediction (9%; 36% thresholds) and $(m, k) = (1.50, 40)$ for the 2-state prediction with a 0% threshold.

RESULTS AND DISCUSSION

Prediction accuracy of the fuzzy k-nearest neighbor method of leave-one-out cross-validation

First, leave-one-out cross-validation on the ASTRAL SCOP dataset of 3644 proteins (see Methods section) was performed, where we

Table 1. The optimal values of the fuzzy parameter m, the number of nearest neighbors k and the corresponding prediction accuracies of leave-one-out cross-validation on the reference dataset derived from ASTRAL SCOP

	State threshold (%) 3-state (9%; 36%)	2-state (0%)	2-state (5%)	2-state (16%)	2-state (25%)
Fuzzy k-NN					
m	1.33	1.5	1.25	1.29	1.33
k	65	40	75	65	65
Accuracy	64.1	86.7	82.0	79.0	78.5
MCC	B:0.577, I:0.245, E:0.528	0.464[a]	0.564[a]	0.578[a]	0.570[a]
Standard k-NN ($m = \infty$)					
k	35	15	20	20	30
Accuracy	61.2	85.8	80.2	77.3	76.9
MCC	B:0.534, I:0.180, E:0.496	0.414[a]	0.521[a]	0.542[a]	0.537[a]

[a] For the 2-state predictions, the MCC values are the same for the two classes, i.e. $\mathrm{MCC_B} = \mathrm{MCC_E}$.

Table 2. The prediction results on the RS126 dataset[a]

	Accuracy (%) 3-state (9%; 36%)	2-state (0%)	2-state (5%)	2-state (16%)	2-state (25%)
Fuzzy k-NN	63.8	87.2	82.2	79.0	78.3
PHDacc	57.5	86.0	—	75.0	—
SVMpsi	59.6	86.2	79.8	77.8	76.8
Thompson and Goldstein	57.9	—	—	75.0	—

[a] NN means nearest neighbors. PHDacc (Rost and Sander, 1994) used neural networks, SVMpsi (Kim and Park, 2004) was based on support vector machines, and Thompson and Goldstein (1996) applied Bayesian statistics for the solvent accessibility prediction on the RS126 dataset. These accuracies are from their published results.

selected one of the 3644 chains and predicted its solvent accessibility, using the remaining 3643 chains as the reference dataset. This procedure was repeated for each of the chains in the dataset. Tests have been done with various values of the fuzzy strength parameter m and the number of nearest neighbors k, to obtain the optimal values of m and k. The contour diagrams of prediction accuracies as functions of $2/(m-1)$ and k are shown in Figure 1, for the 3-state prediction (9%; 36% thresholds) and the 2-state prediction with 0% threshold. The optimal values are $(m, k) = (1.33, 65)$ for the 3-state prediction (9%; 36% thresholds) and $(m, k) = (1.50, 40)$, (1.25, 75), (1.29, 65) and (1.33, 65) for the 2-state predictions (0, 5, 16 and 25% thresholds), respectively. For these values of (m, k), the prediction accuracies are 64.1% for the 3-state prediction and 86.7, 82.0, 79.0 and 78.5% for the 2-state prediction (0, 5, 16 and 25% thresholds), respectively. Table 1 shows the optimal values of k and m and corresponding prediction accuracies for the fuzzy k-nearest neighbor method. These optimal values of k and m are used for all the benchmark tests in the following. For comparison, the results from the standard k-nearest neighbor method, corresponding to the $m = \infty$, are also shown in Table 1. The results indicate that performance can be improved by allowing a finite value of m.

Prediction accuracies of the benchmark tests on the RS126 dataset and the dataset with 229 proteins

The first benchmark test on the RS126 dataset was performed with the optimal values of m and k determined by the leave-one-out cross-validation on the reference dataset derived from ASTRAL SCOP (see Methods section). The fuzzy k-nearest neighbor method shows 63.8% accuracy for the 3-state prediction (9%; 36% thresholds) and 87.2, 82.2, 79.0 and 78.3% for the 2-state prediction with thresholds of 0, 5, 16 and 25% on the RS126 dataset, respectively. The fuzzy k-nearest neighbor method shows slightly better prediction accuracies than other methods on the RS126 dataset as shown in Table 2. PHDacc used a neural network method using evolutionary profiles of amino acid substitutions derived from multiple sequence alignments, and reported 57.5% for the 3-state prediction (9%; 36% thresholds), and 86.0 and 75.0% for the 2-state predictions (thresholds of 0 and 16%), respectively. SVMpsi (Kim and Park, 2004) was based on a support vector machine using the position-specific scoring matrix generated from PSI-BLAST, and reported 59.6% accuracy for the 3-state prediction (9%; 36% thresholds) and 86.2, 79.8, 77.8 and 76.8% accuracies for the 2-state predictions (thresholds of 0, 5, 16 and 25%), respectively. Thompson and Goldstein (1996) applied Bayesian statistics and optimized residue substitution classes, and reported 57.9% accuracy for the 3-state prediction (9%; 36% thresholds) and 75.0% accuracy for the 2-state prediction (threshold of 16%). These prediction accuracies are obtained from their published results.

For the second benchmark tests on the dataset with 229 proteins, we have used the PredictProtein Server (Rost *et al.*, 2004), which provides PHDacc and PROFacc predictions, and Jpred server (Cuff and Barton, 2000; Cuff *et al.*, 1998) to compare the prediction performance directly. The predictions with PredictProtein Sever (http://cubic.bioc.columbia.edu/pp/index.html)

Table 3. The prediction results on the benchmarking dataset with 229 proteins[a]

	State threshold 3-state (9%; 36%)	2-state (0%)	2-state (5%)	2-state (16%)	2-state (25%)
Fuzzy *k*-NN					
Accuracy (%)	62.6	85.5	80.8	78.1	77.8
MCC	B:0.560, I:0.199, E:0.508	0.431	0.541	0.560	0.554
PHDacc					
Accuracy (%)	57.1	—	—	—	—
MCC	B:0.489, I:0.127, E:0.419	—	—	—	—
PROFacc					
Accuracy (%)	62.0	—	—	—	—
MCC	B:0.551, I:0.200, E:0.508	—	—	—	—
Jpred					
Accuracy (%)	—	84.8	80.7	—	76.6
MCC	—	0.388	0.535	—	0.525

[a] PHDacc and PROFacc results were obtained from PredictProtein Server (http://cubic.bioc.columbia.edu/pp/index.html), and Jpred results were obtained from Jpred sever (http://www.compbio.dundee.ac.uk/∼www-jpred/). The predictions with PredictProtein Sever and Jnet were performed with default options.

and Jpred (http://www.compbio.dundee.ac.uk/∼www-jpred/) were performed with default options. It should be noted that any query protein has <25% sequence identity with those in the reference dataset in our method, whereas no such restriction was placed on the sequence similarity between the query protein and training set in other methods. Therefore, our method has no advantage over other methods in this respect. However, it should also be admitted that the other methods were trained on a much smaller dataset, and database growth can give an indirect advantage to newer methods (Przybylski and Rost, 2002) like ours.

The prediction accuracies of our method are slightly better than other methods on the benchmarking dataset with 229 proteins, as shown in Table 3. For the 3-state prediction (9%; 36% thresholds), PHDacc and PROFacc give 57.1 and 62.0% accuracies, and our method gives a 62.6% accuracy. For the 2-state predictions (thresholds of 0, 5 and 25%), Jpred shows 84.8, 80.7 and 76.6% prediction accuracies, and our method shows slightly better prediction accuracies of 85.5, 80.8 and 77.8%, respectively. Our result is better also in terms of MCCs as shown in Table 3.

CONCLUSION

In this work, we applied the fuzzy *k*-nearest neighbor method to solvent accessibility prediction, using PSI-BLAST profiles as feature vectors. We achieved better prediction accuracies than other methods such as neural network methods and support vector machines.

ACKNOWLEDGEMENTS

We thank Jooyoung Lee for useful discussions. This work was supported by grant no. R01-2003-000-10199-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

REFERENCES

Adamczak,R. *et al*. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.

Ahmad,S. and Gromiha,M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, **18**, 819–824.

Ahmad,S. *et al*. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bezdek,J.C. *et al*. (1993) Review of MR image segmentation techniques using pattern recognition. *Med. Phys.*, **20**, 1033–1048.

Brenner,S.E. *et al*. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Cabello,D. *et al*. (1991) Fuzzy *k*-nearest neighbor classifiers for ventricular arrhythmia detection. *Int. J. Biomed. Comput.*, **27**, 77–93.

Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.

Cuff,J.A. *et al*. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.

Eyal,E. *et al*. (2004) Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J. Comput. Chem.*, **25**, 712–724.

Frishman,D. and Argos,P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329–335.

Gianese,G. *et al*. (2003) Improvement in prediction of solvent accessibility by probability profiles. *Protein Eng.*, **16**, 987–992.

Ginalski,K. and Rychlewski,L. (2003) Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins*, **53** (Suppl. 6), 410–417.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Huang,Y. and Li,Y. (2004) Prediction of protein subcellular locations using fuzzy *k*-NN method. *Bioinformatics*, **20**, 21–28.

John,B. and Sali,A. (2004) Detection of homologous proteins by an intermediate sequence search. *Protein Sci.*, **13**, 54–62.

Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kauffman,G.W. and Jurs,P.C. (2001) QSAR and *k*-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically based numerical descriptors. *J. Chem. Inf. Comput. Sci.*, **41**, 1553–1560.

Keller,J.M. *et al*. (1985) A fuzzy *k*-nearest neighbor algorithm. *IEE Trans. Syst. Man Cybern.*, **15**, 580–585.

Kim,H. and Park,H. (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins*, **54**, 557–562.

Leszczynski,K. *et al*. (1999) Application of a fuzzy pattern classifier to decision making in portal verification of radiotherapy. *Phys. Med. Biol.*, **44**, 253–269.

Moult,J. *et al*. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53** (Suppl. 6), 334–339.

Pei,J. and Grishin,N.V. (2004) Combining evolutionary and structural information for local protein structure prediction. *Proteins*, **56**, 782–794.

Pollastri,G. *et al*. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.

Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Rost,B. *et al*. (1997) Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, **270**, 471–480.

Rost,B. *et al*. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.

Russell,S.J. *et al*. (2003) Stability of cyclic beta-hairpins: asymmetric contributions from side chains of a hydrogen-bonded cross-strand residue pair. *J. Am. Chem. Soc.*, **125**, 388–395.

Salamov,A.A. and Solovyev,V.V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, **247**, 11–15.

Salamov,A.A. and Solovyev,V.V. (1997) Protein secondary structure prediction using local alignments. *J. Mol. Biol.*, **268**, 31–36.

Salzberg,S. and Cost,S. (1992) Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.*, **227**, 371–374.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Seker,H. *et al*. (2003) A fuzzy logic based-method for prognostic decision making in breast and prostate cancers. *IEEE Trans. Inf. Technol. Biomed.*, **7**, 114–122.

Singh,R.K. *et al*. (1996) Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J. Comput. Biol.*, **3**, 213–221.

Sokolowska,B. *et al*. (2003) A fuzzy-classifier system to distinguish respiratory patterns evolving after diaphragm paralysis in the cat. *Jpn. J. Physiol.*, **53**, 301–307.

Thompson,M.J. and Goldstein,R.A. (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. *Proteins*, **25**, 38–47.

Totrov,M. (2004) Accurate and efficient generalized born model based on solvent accessibility: derivation and application for LogP octanol/water prediction and flexible peptide docking. *J. Comput. Chem.*, **25**, 609–619.

Vaidyanathan,M. *et al*. (1997) Normal brain volume measurements using multispectral MRI segmentation. *Magn. Reson. Imaging*, **15**, 87–97.

Wohlfahrt,G. *et al*. (2002) Positioning of anchor groups in protein loop prediction: the importance of solvent accessibility and secondary structure elements. *Proteins*, **47**, 370–378.

Yi,T.M. and Lander,E.S. (1993) Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.*, **232**, 1117–1129.

Yuan,Z. and Huang,B. (2004) Prediction of protein accessible surface areas by support vector regression. *Proteins*, **57**, 558–564.

Yuan,Z. *et al*. (2002) Prediction of protein solvent accessibility using support vector machines. *Proteins*, **48**, 566–570.