COMBINING SEQUENCE AND STRUCTURAL PROFILES FOR PROTEIN SOLVENT ACCESSIBILITY PREDICTION

Rajkumar Bondugula[†]

Digital Biology Laboratory, 110 C.S. Bond Life Sciences Center, University of Missouri
Columbia, MO 65211, USA
Email: raj@bioanalysis.org

Dong Xu[*]

Digital Biology Laboratory, 201 Engineering Building West, University of Missouri
Columbia, MO 65211, USA
[*]*Email: xudong@missouri.edu*

Solvent accessibility is an important structural feature for a protein. We propose a new method for solvent accessibility prediction that uses known structure and sequence information more efficiently. We first estimate the relative solvent accessibility of the query protein using fuzzy mean operator from the solvent accessibilities of known structure fragments that have similar sequences to the query protein. We then integrate the estimated solvent accessibility and the position specific scoring matrix of the query protein using a neural network. We tested our method on a large data set consisting of 3386 non-redundant proteins. The comparison with other methods show slightly improved prediction accuracies with our method. The resulting system does need not be re-trained when new data is available. We incorporated our method into the MUPRED system, which is available as a web server at *http://digbio.missouri.edu/mupred*.

1. INTRODUCTION

Predicting the three-dimensional structure of a protein from its sequence has been an open challenge in bioinformatics for more than three decades. In many cases, three-dimensional structures cannot be predicted accurately and researchers like to obtain structure features such as secondary structures and solvent accessibility (SA). While secondary structure captures some aspects of the protein structure, the SA characterizes different structural features. The concept of the SA was introduced by Lee and Richards[1] and can be defined as the extent to which water molecules can access the surface of a protein. The knowledge of SA helped to further the understanding of protein structure classification[2-4], protein interaction[5-7], etc.

A number of approaches such as information theory[8], support vector machines[9], neural networks[10-12], nearest-neighbor methods[13], and energy optimization[14] have been proposed for SA prediction. Almost all of these methods rely on protein position specific scoring matrix (PSSM)[15] from multiple sequence alignments. There are at least two drawbacks of these approaches. First, they predict the structural features of the proteins without using the structural information available in the Protein Data Bank[16] (PDB). Second, when proteins do not have close homologs in the database of known sequences (for example, *nr* at *http://www.ncbi.nlm.nih.gov*), the PSSM will not be well defined, making the predictions unreliable[17].

In our approach, both the structural information and the sequence profile information are used. We first build a structural profile by estimating the relative solvent accessibility of the query protein using a fuzzy mean operator (FMO) from the solvent accessibilities of proteins with known structures that have similar sequence fragments to the query protein. We then integrate the estimated solvent accessibility and the PSSM using a neural network (NN). We choose a NN as the approproiate scheme for combining information from profiles and FMO is automatically learned by the network from the training data. The output of the NN is the predicted relative solvent accessibility of each residue. The user may either obtain real solvent accessibility values (in terms of $Å^2$) or classify solvent accessibility into multiple classes using any thresholds based on his/her specific needs. The proposed approach has the advantage of simplicity and transparency. Also,

[*]Corresponding author.
[†]Current address: Bldg 363 Miller Drive, Biotechnology HPC Software Applications Institute, US Army Medical Research and Materiel Command Ft. Detrick, MD 21702, USA

most of the existing methods were tested on small data sets containing up to a few hundred sequences. These results on small sets have significant variations in prediction accuracies. To overcome this problem, we tested our method on a large-scale data set of non-redundant proteins to obtain stable performance. The prediction program has been implemented into the MUPRED package as a public web server at *http://digbio.missouri.edu/mupred* along with the secondary structure prediction capacity.

2. METHOD AND MATERIALS

In our method, the relative solvent accessibility of each amino acid in the query protein is first estimated using the FMO. The calculated fuzzy means are used as the initial set of features. The second set of features is derived from the PSSM of the query protein. These two features are integrated using a neural network. In Section 2.1, we introduce the features and the data sets used in this work. The estimation of the relative solvent accessibilities using FMO is explained in Section 2.2. In Section 2.3, the process of deriving the second set of features and integrating these two feature sets using a neural network is described. In Section 2.4, the metrics used for performance assessment are presented.

2.1. Feature Inputs and Data Sets

The PSSM of the query protein is the starting point in generating input features. We use PSI-BLAST[15] and the *nr* database to generate the PSSM. We used the following parameters for generating the PSI-BLAST: *j* (number of iterations) = 3, *M* (substitution matrix) = BLOSUM90 with other parameters set at default values. We use the BLOSUM90 substitution matrix as we want only the hit fragments that are close subsequences of the query protein to contribute to the PSSM being generated. The parameters were experimentally determined on the training set. Similar results were obtained for a wide range of parameters (data not shown).

A database of representative protein set (RPS), whose three-dimensional structures (and hence, solvent accessibilities) are known is required to estimate the relative solvent accessibility of the query protein. We used the March 2006 release of the PDBSelect[18] database to prepare RPS. The PDBSelect database consists of representative proteins such that the sequence identity between any two proteins in the database is not more than 25%. Initially, the database had 3080 chains. We only selected the proteins whose structures are determined by X-ray crystallography method with a resolution of less than 3 Å and lengths of more than 50 residues. We further restricted our selection to proteins which have at least 90% of their residues composed of regular amino acids. The selection process has resulted in RPS that contains 1998 proteins with 310,114 residues.

First, we present the performance of our method on the RPS using a jack-knife procedure (query sequence eliminated from the RPS during prediction). We employed two widely used data sets (benchmark sets) to compare the performance of MUPRED with other methods. The first database used in reference [10] contains 126 representative proteins with 23,426 residues (hereafter referred as RS126). The second data set was introduced by Naderi-Manesh et al. in Reference [8]. The database consists of 215 representative proteins with 51,939 residues (hereafter referred as MN215). The proteins in RPS that are similar to any proteins in the benchmark sets are eliminated using the following procedure: each sequence in the RPS database was queried against proteins in the benchmark sets using the BLAST[19] program. If a hit with an e-value less than 0.01 is found, the query sequence was eliminated from the RPS. This procedure further reduced the number of proteins in RPS to 1657. In addition to testing our method on the RPS and the two standard benchmark sets, we employed a fourth data set derived from the Astral SCOP domain database[20] version 1.69. The original database with 25% maximum identity between any two sequences consists of 5457 protein domains. The proteins in the Astral SCOP data set that are similar to the proteins in the RPS are discarded using the same procedure outlined above (i.e., each sequence in the Astral SCOP database was queried against RPS using the BLAST program. If a hit with an e-value less than 0.01 is found, the sequence was eliminated from the Astral SCOP database). Similar to the procedure used to generate the RPS, domain sequences shorter than 40 residues were removed. If less than 90% of a domain sequence is composed of regular amino acids, it is discarded as well. The remaining 3386 domain sequences with 636,693 residues after the filtering make up the independent benchmark set.

2.2. Fuzzy Mean Operator

The profile of the query protein is used to search for the similar fragments in RPS by running the PSI-BLAST second time. The threshold value of *e* was set to 11,000 when searching the RPS. The higher the threshold, the larger the number of fragments returned by the PSI-BLAST. However, if the threshold is too high, the PSI-BLAST returns large number of informative hits as well as noises from the database. The best compromise was experimentally determined. The relative solvent accessibility (RSA) of each residue in the query protein is calculated using the hit fragments that have a residue aligned with the current residue using FMO. The process is explained in the following paragraphs.

The hit fragments returned by the PSI-BLAST program are scored using the following equation:

$$S = \max\{1, 7 + \log_{10}(e\text{-}value)\} \quad (1)$$

This score is formulated as a dissimilarity measure. For instance, the fragments of proteins in RPS that have high sequence similarity with the subsequences of the query protein have high statistical significance (or low e-value), therefore have low scores.

The RSA of each residue of the query protein is calculated from the RSAs of hits that have a residue aligned with the current residue. The SA of the hit fragments are calculated using the DSSP[21] program. For each residue, the absolute SA retuned by the DSSP program is transformed into RSA by dividing it with the maximum SA given in Reference [10]. The RSA of the query protein is calculated using the following expression for FMO:

$$RSA(r) = \frac{\sum_{j=1}^{K} RSA_j \left(\frac{1}{S^{\frac{2}{m-1}}} \right)}{\sum_{j=1}^{K} \left(\frac{1}{S^{\frac{2}{m-1}}} \right)} \quad (2)$$

where r is the current residue, K is number of hits that have residue aligned with the current residue, RSA_j is the relative solvent accessibility of the residue in the j^{th} hit that is aligned with the current residue, S is the score defined in Equation (1), and m is a fuzzifier[22] that controls the weight of the dissimilarity measure S. The optimal value of fuzzifier was experimentally determined to be 1.5. Note that the Equation (2) is a special case of the fuzzy *k*-nearest neighbor algorithm[22].

2.3. Profile Feature Set and Integration of the Two Feature Sets

The second set of features is generated from the PSSM of the query protein. In the PSSM, each residue is represented by a 20 dimensional vector representing the likelihood of each of the 20 amino acids in that position. The profiles are first normalized and then rescaled into [-1 1] before converting them into vectors suitable for neural network training. We found that the maximum and minimum values in the profiles of all proteins in the RPS were -10 and 12, respectively. Therefore, the profiles were normalized and rescaled using the following expressions:

$$PSSM(i, j) \leftarrow 2x - 1,$$
$$\text{where } x \leftarrow \frac{(PSSM(i, j) + 10)}{22} \quad (3)$$

where $i \in [1,\ldots,n]$ (n is the length of the query protein) and $j \in$ {A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}. An additional bit is used to represent if the current residue lies in the termini of the query protein. We arbitrarily choose 1 to represent the termini, while 0 is used for representing the interior of the protein. The transformed PSSM values, along with the additional bit are converted into vectors suitable for neural network training using a sliding window scheme, i.e., a vector representing the current residue is flanked by the vectors representing the neighbors on the both sides. This scheme allows us to capture that idea that a particular residue's solvent accessibility is dependent on the solvent accessibility states of its neighbors[28,36]. The number of neighbors on each side is determined by parameter W. We experimentally determined that the optimal number of neighbors on each side of the current residue to consider for this feature set is 7 and therefore, the total number of features in this set is (20+1)x15=315.

Similar to the features generated from the PSSM, the fuzzy means that originally lie in [0 1] are rescaled to lie in [-1 1] using the following transformation:

$$RSA(r) \leftarrow 2 \times RSA(r) - 1 \quad (4)$$

The rescaled fuzzy means are converted into vectors suitable for training the neural network using the

sliding window scheme. Again, we use an extra bit to indicate the termini of the protein using the same encoding method as the PSSM feature set. We experimentally determined that the optimal window size is 13 and therefore, the total number of features in this feature set is 2x13=26. These two feature sets together (26+315 = 341 features/residue) are used to train the neural networks. The neural network used to integrate the fuzzy means and PSSM is a fully connected feed-forward network with one hidden layer, trained using standard back-propagation learning. We trained the networks with different number of nodes, starting at 170 and increased 10 units at a time. We found that 240 nodes resulted in an optimal performance. The output layer consists of a single neuron that produces the predicted RSA. The neural network has the following architecture 341×240×1 (input nodes × hidden nodes × output node). We randomly selected 50 of RPS proteins for generating the validation vectors and used the rest for training the neural networks. The networks were trained until the performance using the validation vectors started to decline. A total of 100 networks were trained using random initialization and the top six networks (networks with lowest re-substitution error the on the training data) were retained for prediction. Each query protein is simulated on all six networks and the average of the 6 networks is taken as the output of the prediction system. The block diagram of the MUPRED solvent accessibility prediction system is illustrated in Figure 1.

2.4. Prediction Accuracy Assessment

If the system is used as a classifier to group the residues into two classes (*buried* and *exposed*), the following two metrics are used to assess the performance:
Accuracy (Q_2):

$$Q_2 = \frac{p+n}{t} \tag{5}$$

Matthew's correlation coefficient[23] (MCC):

$$MCC = \frac{pn - uo}{\sqrt{(p+u)(p+o)(n+u)(n+o)}} \tag{6}$$

where p is the number of correctly classified *exposed* residues (true positives), n is the correctly classified *buried* residues (true negatives), o is the number of residues that were incorrectly classified as *exposed* residues (false positives), u is the number of residues that were incorrectly classified as the *buried* residues (false negatives), and $t = p + n + o + u$ (total number of residues).

To asses the performance of the RSA prediction ability of the system, the mean absolute error (MAE) as defined below is used:

$$MAE = \frac{1}{N}\sum \left| RSA_{observed} - RSA_{predicted} \right| \tag{7}$$

where $RSA_{observed}$ is the experimental RSA of a residue from the DSSP file divided by its maximum SA while $RSA_{observed}$ is the predicted RSA, and the summation is over all N residues in the protein.

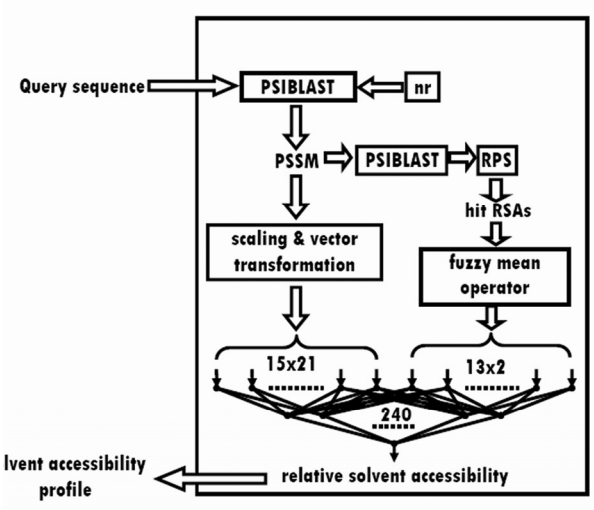

Fig. 1. MUPRED solvent accessibility prediction system. The profile of the query protein is first calculated and used to generate two feature sets. The first set consists of vectors derived from the normalized and rescaled PSSM using a sliding window scheme with window length (W) 15. This set consists of 15x21 =315 features/residue. The second feature set is generated by searching the local database of representative proteins based on profile-sequence alignment. The homologous fragments returned by the search process are used to estimate the relative solvent accessibility of each residue using the fuzzy mean operator. The vectors representing the second feature set are derived from the fuzzy means, using the sliding window of length (W) 13. Similar to the first feature set, an additional bit is used to represent the termini of the query protein. This feature set consists of 13x2 = 26 features, resulting in 341 features for each residue altogether. The neural network consists of 240 hidden units and a single output neuron that produces the predicted solvent accessibility.

3. RESULTS

In this section, we discuss the performance of the FMO alone, FMO with a neural network and finally,

MUPRED that uses both FMO and PSSM on the RPS and independent SCOP derived set. We then compare MUPRED with some existing methods for prediction accuracies on the two benchmarking data sets.

When we tested the SA profile generated by the FMO alone, we noticed that the trend of predicted SA profile often resembles the actual SA profile, except that the dynamic range of the predicted SA profile is consistently smaller. This may be due to the averaging effects over the neighboring residues when building the SA profile using Equation (2), although such an average reduces the noise for better prediction accuracy overall. Since the neural networks function well as the signal amplifiers, we trained a neural network using the sliding window scheme described in Section 2.2 with the window size 13. This network was not used in the final MUPRED as there appears to be no practical advantage in amplifying signals while integrating the feature sets. The performances of our systems as a two class-classifiers on the various data sets are given in Figure 2 (a-d). The plot on the left illustrates the distribution of the RSA in the corresponding data set, while the plot on the right contains the classification accuracies and the Matthew's correlation coefficients at various classification thresholds. The plots show that integrating FMO and PSSM using a neural network significantly improves the prediction accuracy over the FMO prediction alone or the FMO prediction with a neural network.

We compare MUPRED with existing methods on the two most widely used data sets. The comparison in terms of two-state accuracy on the RS126 data set is presented in Table 1, while the comparison on the MN215 is presented in Table 2. The MAEs of MUPRED on RPS, the SCOP derived independent set, RS126 and MN215 are 14.17%, 15.29%, 14.31% and 13.6%, respectively. The Pearson correlation coefficients of our method on RPS, the SCOP derived independent set, RS126 and MN215 are 0.72, 0.69, 0.71 and 0.72, respectively. Garg et al.[12] reported the Pearson correlation coefficient of 0.67 on the MN215 data set. In both the comparisons, MUPRED has the highest prediction accuracy in most cases. The MAE and the Pearson correlation coefficient on the RPS and the SCOP derived set indicate that the overtraining did not occur when we trained our neural networks.

The program is implemented in the ANSI compatible C programming language. The regression analysis performed on the computation time of our method on a Pentium-4, 3 GHz machine with 2 GB of RAM indicates that the prediction time is a linear function of the sequence length and requires 0.55 sec/residue, including the time required for calculating the profile using the PSI-BLAST. The peak memory requirement is under 20 MB.

Table 1. The comparison of MUPRED with existing methods on the RS126 data set. The performance reported is the two-state accuracy obtained by using different threshold values.

Threshold/Method	A	B	C	D	E
0	87	86	-	86	87
5	77	-	-	80	82
9	78	75	78	-	-
16	79	75	78	78	79
23	79	-	77	-	-
25	79	-	-	77	78

A- Current work; B-Rost and Sander, 1994; C-Manesh et al., 2001; D-Kim and Park, 2004;E-Sim et al., 2005. The '-' indicates that no information is available.

Table 2. The comparison of MUPRED with existing methods on the MN215 data set. The performace reported is the two-state accuracy obtained by using different threshold values.

Threshold/Method	A	C	F	G	H
4	77	75	-	-	-
5	77	-	75	77	75
9	78	76	-	-	-
10	78	-	71	78	77
16	79	76	-	-	-
20	79	-	-	78	78
25	79	74	70	78	-
30	79	-	-	-	78
36	80	74	-	-	-
40	80	-	-	-	78
49	81	80	-	-	-
50	2	-	76	-	81
60	86	-	-	-	85
64	88	97	-	-	-
70	91	-	-	-	91
80	95	-	-	-	95
81	96	81	-	-	-

A-current work; C- Manesh et al., 2001; F- Ahmed and Gromiha, 2002; G- Adamczak et al., 2004 ; H- Garg et al., 2005. The '-' indicates that no information is available.

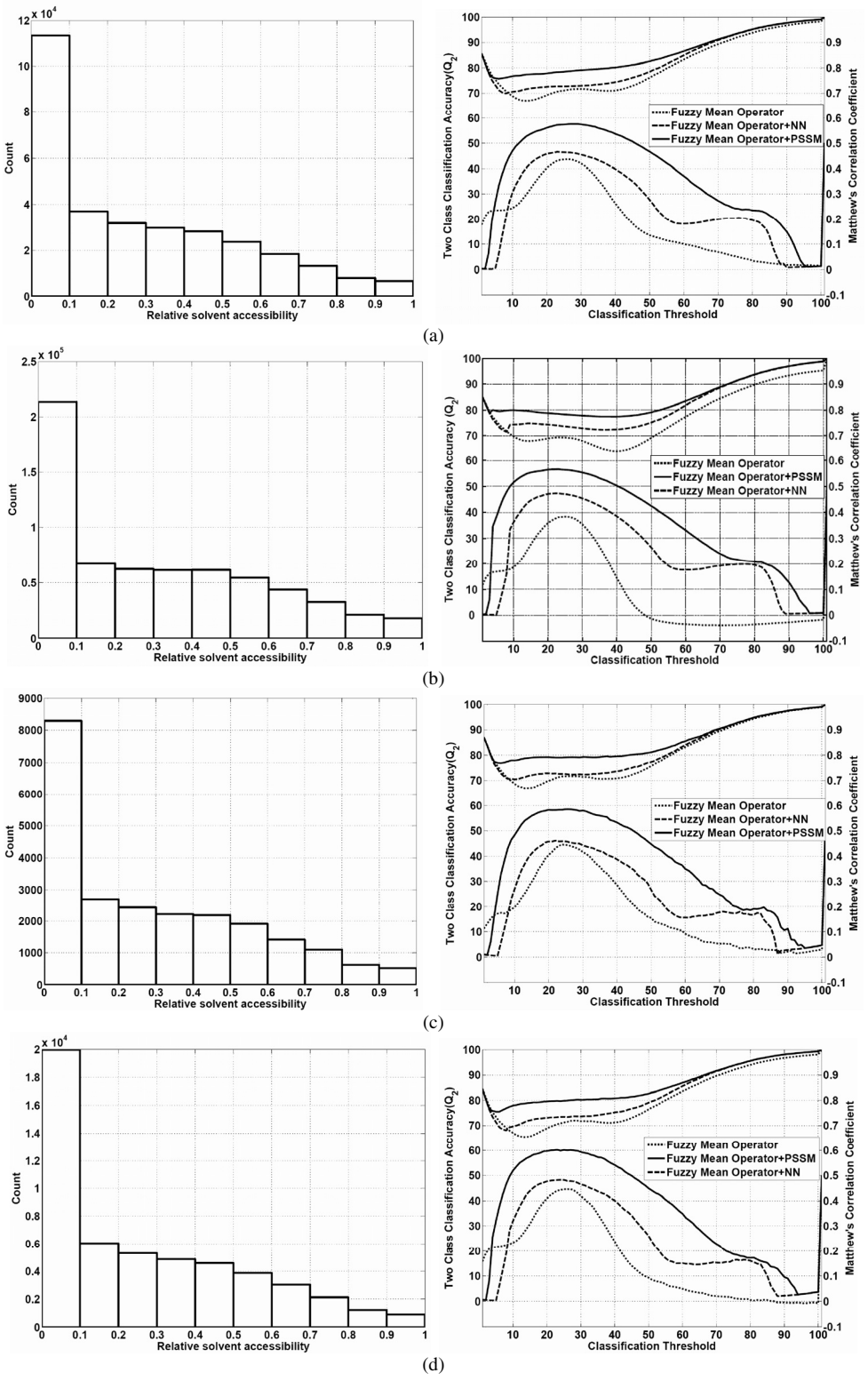

Fig. 2. The histograms showing the compositions of the RSAa in various data sets (Left) and performance of our methods on each of the data sets (Right). The classification threshold is varied along the x-axis, while the two-class classification accuracy (the top three curves) is plotted using the y-axis on the left, while the Matthew's correlation coefficient (the bottom three curves) is plotted using the y-axis on the right. (a) Training set of 1657 proteins; (b) SCOP data set with 3457 proteins; (c) Rost and Sander 126 protein set; (d) Manesh 215 protein set.

4. DISCUSSION

The proposed SA prediction system has some similarity to our secondary structure prediction system[24]. The key difference is that the former is a function approximation, while the later is a classification problem. Our method uses the structural information in the PDB more efficiently than the existing methods and therefore, reduces the dependence on availability of homologous sequences in a sequence database for building a well defined profile. At one extreme, the query sequence has many close homologs in the database of known sequences resulting in a well-defined PSSM. In such cases, our procedure uses profile-sequence alignment for finding similar fragments (exploiting both local and global similarities) in the RPS. Therefore, both PSSM and FMO contribute well for the final prediction. At other extreme where the sequence does not have close homologs, the PSSM is just the scoring matrix used in the alignment procedure. In such situations, our procedure is equivalent to searching for similar fragments in RPS using a sequence-sequence alignment. The homologous fragments (exploiting local similarities only) found by sequence-sequence alignment are effectively used by the FMO and therefore, has the protein structure contribution to the prediction with little or no help from PSSM. The latter case is emulated by the system with FMO followed by a neural network, which provides an estimate of the lower bound of accuracy. Since the output of the neural network is RSA (in [0 1]) of the protein, the system allows a user to choose the number of states and related thresholds, if a classification of residues is desired. The users can multiply the RSA by their maximum solvent accessible areas of respective amino acids to obtain the real solvent accessibility values in terms of $Å^2$. Unlike earlier methods, our system is transparent, weather it succeeds or fails. The predicted solvent accessibility for a given query protein can be traced back to proteins in the RPS that contributed for that prediction, giving additional insight to the users. One of the appealing features of our systems is that it need not be re-trained. As more and more representative structures are solved, their sequences just need to be added to the RPS and the algorithm will use the new information immediately. Over time, we expect our system increases the prediction accuracies automatically by having expanded *nr* and PDB databases, relieving the users or us from the burden of re-training the system in the future.

5. CONCLUSIONS

We developed a new and unique system for effective SA prediction. We use PSSM and fuzzy mean operator to seamlessly integrate sequence profile and structural information into one system, which has not been achieved before. This combination enables successful predictions for the sequences with or without homologs in the database of protein sequences. Our results prove that the additional, complementary information provided by using the structural information has slightly improved the prediction accuracy. Our system will have increased performance accuracy as more protein structures are added to PDB and the expansion of the *nr* databases.

Acknowledgements

This work was supported by a Research Board grant from University of Missouri and by an NIH grant (1R21GM078601). The authors would like to thank Travis McBee for his assistance in the project and Dr. James Keller for discussion on the fuzzy mean operator.

References

1. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 1971, **55(3)**:379–400.
2. Gromiha MM, Suwa M. Variation of amino acid properties in all-beta globular and outer membrane protein structures. *Int J Biol Macromol* 2003, **32(3-5)**:93-8.
3. Sujatha MS, Balaji PV. Identification of common structural features of binding sites in galactose-specific proteins. *Proteins* 2004, **55(1)**:44-65
4. Yu ZG, Anh VV, Lau KS, Zhou LQ. Clustering of protein structures using hydrophobic free energy and solvent accessibility of proteins. Phys Rev E Stat Nonlin Soft Matter Phys. *Physical Review* 2006, **E73(3.1)**:031920.
5. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 2004, **20(4)**:477-86.
6. Chen H, Zhou HX. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* 2005, **61(1)**:21-35.

7. Hoskins J, Lovell S, Blundell TL. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci* 2006, **15(5)**:1017-29
8. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. *Proteins* 2001, **42(4)**: 452-459.
9. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* 2004, **54(3)**:557–562.
10. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994, **20(3)**:216–226.
11. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 2002, **18(6)**:819-24.
12. Garg A, Kaur H, Raghava G. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* 2005, **61(2)**:318–324.
13. Sim J, Kim SY, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics* 2005, **21(12)**:2844-9.
14. Xu Z, Zhang C, Liu S, Zhou Y. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins* 2006, **63(4)**:961-6.
15. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, **25**:3389-3402.
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res* 2000, **28**:235-242
17. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004, **56(4)**:753-67.
18. Hobohm U, Sander C. Enlarged representative set of protein structures. *Protein Science* 1994, 3(3):522-524.
19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990, **215**:403-410.
20. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res* 2000, **28**:254-256.
21. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features. *Biopolymers* 1983, **22**:2577-637.
22. Keller JM, Gray MR, Givens JA. A fuzzy K-Nearest Neighbor Algorithm. *IEEE Trans on SMC* 1985, **15**:580-585.
23. Mathews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975, **405**:442-451.
24. Bondugula R, Xu D. MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins* 2007, **66(3)**:664-670.