American Journal of Biomedical and Life Sciences
2015; 3(2-3): 21-24
Published online August 6, 2015 (http://www.sciencepublishinggroup.com/j/ajbls)
doi: 10.11648/j.ajbls.s.2015030203.14
ISSN: 2330-8818 (Print); ISSN: 2330-880X (Online)

Protein solvent accessibility prediction systems

Ritta Shaheen[1], Hani Amasha[2], Majd Aljamali[3]

[1]Department of Biomedical Engineering, Faculty of Mechanical and Electrical Engineering, Damascus University, Damascus, Syria.
[2]Department of Biomedical Engineering, FMEE, Damascus University and Faculty of Informatics and Communication Engineering, Arab International University, Damascus, Syria
[3]Faculty of Pharmacology, Damascus University, Damascus, Syria

Email address:

Ritta.shaheen@gmail.com (R. Shaheen), haniamasha@gmail.com (H. Amasha), maljamali@gmail.comcom (M. Aljamali)

To cite this article:

Ritta Shaheen, Hani Amasha, Majd Aljamali. Protein Solvent Accessibility Prediction Systems. *American Journal of Biomedical and Life Sciences*. Special Issue: Spectral Imaging for Medical Diagnosis "Modern Tool for Molecular Imaging". Vol. 3, No. 2-3, 2015, pp. 21-24.
doi: 10.11648/j.ajbls.s.2015030203.14

Abstract: Background: Prediction of protein solvent accessibility, also called accessible surface area (ASA) prediction, is an important step for tertiary structure prediction directly from one-dimensional sequences. Traditionally, predicting solvent accessibility is regarded as either a two- (exposed or buried) or three-state (exposed, intermediate or buried) classification problem. However, the states of solvent accessibility are not well-defined in real protein structures. Thus, a number of methods have been developed to directly predict the ASA based on information such as amino acid composition. Results: In this study we use physicochemical properties of amino acid such as hydrophobicity for ASA prediction by considering amino acid composition. We propose a systematic method for identifying residue groups with respect to protein solvent accessibility. The hydrophobicity of amino acid are used to generate features. Finally, Adaptive neuro fuzzy inference system (anfis) is adopted to construct a ASA predictor. Experimental results demonstrate that the features produced by the proposed selection process are informative for ASA prediction. Conclusion: Experimental results based on a widely used benchmark reveal that the proposed method performs good among several of existing packages for performing ASA prediction depending on amino acid sequence only .The program and data are available from the authors upon request.

Keywords: Protein Structure, Protein Solvent Accessibility, Accessible Surface Area, Structure Prediction, Adaptive Neuro Fuzzy Inference, Hydrophobicity

1. Introduction

Predicting protein tertiary structures directly from one-dimensional sequences remains a challenging problem (1).The studies of solvent accessibility have shown that the process of protein folding is driven to maximal compactness by solvent aversion of some residues (2). Therefore, solvent accessibility is considered as a crucial factor in protein folding and prediction of protein solvent accessibility, also called accessible surface area (ASA) prediction, is an important step in tertiary structure prediction (3). Traditionally, predicting solvent accessibility is regarded as either a two- (exposed or buried) or three state(exposed, intermediate or buried) classification problem. Various machine learning methods have been adopted, including neural networks (4) (5) (6) (7) (8) (9) (10) (11), Bayesian statistics (12),logistic functions (13), information theory (14) (15) (16) and support vector machines (SVMs) (17) (18) (19). Among these machine learning methods, neural networks were the first technique used in predicting protein solvent accessibility and are still extensively adopted in recent works. In addition, SVMs were also effective for ASA prediction. Several features were used to train these machine learning methods, such as local residue composition (4) (5), probability profiles (20) and position specific scoring matrix (PSSM) (21). Ahmad et al. developed a method, RVP-net, to predict the real values of relative solvent accessibility (RSA) (22). The RVP-net used the local amino acid composition to train a neural net-work and yielded an accuracy of 74.1%. Yuan and Huang (23), also used the local amino acid composition and adopted support vector regression (SVR) (the regression version of SVM) to achieve an accuracy of 74%. Wang et al. (24) proposed a real value ASA predictor with an accuracy of 78 % by combining the amino acid composition with multiple linear regression. Table 1 summarizes the recent developments in predicting ASA.

Neural networks and SVRs were extensively adopted and outperformed other machine learning methods. This study proposes a systematic process to predict ASA. ANFIS is used to construct an ASA predictor. The present method is compared with three ASA.

Table 1. *The recent developments, in chronological order, for real value ASA prediction.*

Q (%)[1]	Description of features	Regression tool	Work
74.1	Amino acid composition	NN[2]	Ahmad et al., 2003
74	Amino acid composition	SVR[3]	Yuan and Huang, 2004
78	Amino acid composition, PSSM and sequence length	MLR[3]	Wang et al., 2005

2. Datasets

This study collects two independent datasets, first data set for training ASA predictors. The second, small data-sets, (R126) are used for the evaluating the predictor.

2.1. TRAIN Dataset

This dataset contains all proteins in Protein Data Bank (PDB) which have at least 30 amino acids long withe no chain breaks.this set consists of 1180 sequences corresponding to 282,303 amino acids.

2.2. Evaluating Dataset (RS126)

The This is one of the oldest datasets created for evaluating secondary structure prediction schemes. The dataset contains 126 proteins which did not share sequence identity more than 25% over a length of at least 80 residues.

3. Practical Study

Solvent accessibility problem can be considered as a pattern recognition problem, where an artificial neural network is trained to identify the solvent accessibility corresponding to each amino acid in the protein sequence.

In this study we use, Adaptive neuro fuzzy inference system network available in MATLAB R2011a Fuzzy toolbox, with one input layer ,one output layer.

We applied a sliding window of size 15 (an odd number of respectively amino acids) as the input to the network to predict the solvent accessibility of the residue in the middle of the window; this will add the influence of the neighbors into the prediction. ach amino acid in the input window encoded with is hydrophobicity of amino acid represented in table (2).

The output layer is two units, each one corresponds to a solvent accessibility state of amino acid and encoded using a binary system to build the target matrix of the neural network (representing the corresponding solvent accessibility to each amino acid in the input matrix) as following: 1 0 for buried residue. 0 1 for exposed one.

Thus, using the previous input and output matrices , we have created an anfis network shown in Fig (1).

Table 2. *The a hydrophobicity of amino acid.*

Amino acid	hydrophobicity
Phe	100
Ile	99
Trp	97
Leu	97
Val	76
Met	74
Tyr	63
Cys	49
Ala	41
Thr	13
His	8
Gly	0
Ser	-5
Gln	-10
Arg	-14
Lys	-23
Asn	-28
Glu	-31
Pro	-46
Asp	-55

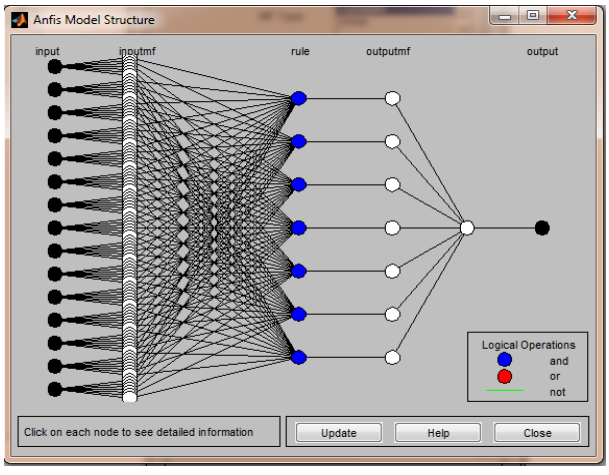

Figure 1. *Shows the anfis structure.*

4. System Specifications

The table(3) demonstrate the anfis specifications.

Table 3. *Anfis specification.*

Unit		Sub. Clustering				Optim. Method	Error Tolerance	Epochs
		Range of Influence	Squash Factor	Accept Ratio	Reject Ratio			
Unit	1	1.3	1.45	0.5	0.15	Hybird	0	300
	2	1.3	1.45	0.5	0.15	Hybird	0	300

5. Results

This section displays the results of system, to be compared, and the comparison depends on the accuracy of each system Q, which is calculated according to the following equation:

$$Q = \frac{P_e + P_b}{N} * 100\%$$

Where P_e,P_b, are the number of amino acids of solvent accessibility class buried and exposed respectively that were correctly predicted, and N is the total number of amino acids.

The total accuracy for predicting of the solvent accessibility is Q= 70.9%, with an accuracy of Qa= 74.4 for buried residue and Qb= 66.92 for exposed residue.

6. Conclusion

In this paper we developed a system to predict the solvent accessibility relying solely on the amino acid sequence of the protein chain without using any additional information, which was used train data set which selected and encoded using hydrophobic values for the training the ANFIS system. System consists of two units each unit is predicting only one type in types of solvent accessibility, then the highest value among the two output of ANFIS units is consider the final output, which is the solvent accessibility of the amino acid located in the middle- of the income window . The accuracy of the system has reached to 71%, which is good accuracy. The following is a table demonstrate the comparison between the prediction accuracy of solvent accessibility that have been reached in this research of other systems that depending on the amino acids sequence only as input.

Table 4. *Comparison between our research and other researches.*

Accuracy	Description of features	Regression tool	Research
74.1	Amino acid sequence	Neural Network	Ahmad et al
74	Amino acid sequence	Support Vector Machine	Yuan and Huang
78	Amino acid sequence	Multiple Linear Regression	Wang et al
70.9	Amino acid sequence	Adaptive Neuro Fuzzy Inference System	Suggested System

Acknowledgements

The authors would like the Damascus university for their support.

References

[1] DW, Mount. Bioinformatics: sequenceand genome analysis. s.l. : Cold Spring Harbor,N.Y., 2004. Vol. 2nd edition.

[2] Chan HS, Dill KA. Origins of Structure in Globular-Proteins. s.l. : Proc Natl Acad Sci, 1990.

[3] Raih MF, Ahmad S, Zheng R, Mohamed R. Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability. Biophys Chem. 2005.

[4] Holbrook SR, Muskal SM, Kim SH. Predicting Surface Exposure of Amino-Acids fromProtein-Sequence. Protein Eng. 1990.

[5] Rost B, Sander C. Conservation and Prediction of Solvent Accessibility in Protein Families. Proteins. 1994.

[6] Pascarella S, De PersioR, Bossa F, Argos P. Easy method to predict solvent accessibility frommultiple protein sequence. Proteins. 1998.

[7] Cuff JA, Barton GJ. Application of multiplesequence alignment profiles to improve protein secondary structure prediction. Proteins. 2000.

[8] Fariselli P, Casadio R. RCNPRED: prediction of the residue co-ordination numbers in proteins. Bioinformatics. 2001.

[9] Li X, Pan XM. New method for accurate prediction of solvent. Proteins. 2001.

[10] Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics. 2002.

[11] Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins. 2002.

[12] Thompson MJ, Goldstein RA:. Predicting solvent accessibility:Higher accuracy using Bayesian statistics and optimized resdue substitution classes. Proteins. 1996.

[13] Mucchielli-Giorgi MH, Hazout S, Tuffery P:. PredAcc: prediction of solvent accessibility. Bioinformatics. 1999.

[14] Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. Protein Eng. 1999.

[15] O, Carugo. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng. 2000.

[16] Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibilitywith information theory. Proteins. 2001.

[17] Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins. 2002.

[18] Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins. 2004.

[19] Nguyen MN, Rajapakse JC. Prediction of proteinrelative solvent accessibility with a two-stage SVM approach. Proteins. 2005.

[20] Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. Protein Eng. 2003.

[21] Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W,Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997.

[22] Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins. 2003.

[23] Yuan Z, Huang BX. Prediction of protein accessible surface areas by support vector regression. Proteins. 2004.

[24] Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. protein solvent accessibility using multiple. 2005.

[25] Garg A, Kaur H, Raghava GPS. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins. 2005.

[26] Nguyen MN, Rajapakse JC. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins. 2006.

[27] Predicting the protein disordered region using modified position specific scoring matrix. Shimizu K, Hirose S, Noguchi T, Muraoka Y. Yokohama Pacifico, Japan : s.n., December 16–18 2004. 15th International Conference on Genome Informatics.

[28] Su CT, Chen CY, Ou YY. Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinformatics. 2006.

[29] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins. 2004.