EXCLI Journal 2009;8:211-217 ISSN 1611-2156
Received: June 28, 2009, accepted: August 16, 2009, published: August 18, 2009

Original article:

PREDICTION OF RELATIVE SOLVENT ACCESSIBILITY USING PACE REGRESSION

Alireza Meshkin[1,*], Mehdi Sadeghi[2,3], Nasser Ghasem-Aghaee[4]

[1] Department of Computer Engineering, University of Sheikh Bahaee, Isfahan, Iran
[2] National Institute for Genetic Engineering and Biotechnology, P.O. Box 14155-6343, Tehran, Iran
[3] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[4] Department of Computer Engineering, University of Isfahan, Isfahan, Iran
[*] Corresponding author: Email: meshkin@nigeb.ac.ir

ABSTRACT

In this paper, a new approach for prediction of protein solvent accessibility is presented. The prediction of relative solvent accessibility gives helpful information for the prediction of native structure of a protein. Recent years several RSA prediction methods including those that generate real values and those that predict discrete states (buried vs. exposed) have been developed. We propose a novel method for real value prediction that aims at minimizing the prediction error when compared with existing methods. The proposed method is based on Pace Regression (PR) predictor. The improved prediction quality is a result of features of PSI-BLAST profile and the PR method because pace regression is optimal when the number of coefficients tends to infinity. The experiment results on Manesh dataset show that the proposed method is an improvement in average prediction accuracy and training time.

Keywords: pace regression, relative solvent accessibility, PSI-BLAST

INTRODUCTION

Due to the strict relation between protein function and structure, the prediction of protein 3D-structure has become one of the most important tasks in bioinformatics.

Prediction of the 3D structure from protein sequence should be feasible based on the well-established credo that protein sequence uniquely determines protein structure (Anfinsen, 1973). Despite several decades of extensive researches in tertiary structure prediction, this task is still a big challenge, especially for sequences that do not have a significant sequence similarity with known structures (Ginalski & Rychlewski, 2003).

The prediction of solvent accessibility (Garg et al., 2005) is an intermediate step in prediction of the protein tertiary structure.

The relative solvent accessibility (RSA) reflects the degree to which a residue interacts with the solvent molecules. Since protein-protein and protein-ligand interactions occur at the protein surface, only the residues that have a large surface area exposed to the solvent can possibly bind to the ligands and other proteins. As a result, prediction of solvent accessibility provides useful information for prediction of binding sites (Huang & Schroeder, 2006) and is vitally important for understanding the binding mechanism of proteins (Chou, 1988). It has been pointed that the burial of core residues is the driving force in protein fold-

EXCLI Journal 2009;8:211-217 ISSN 1611-2156
Received: June 28, 2009, accepted: August 16, 2009, published: August 18, 2009

ing, which suggests that knowledge of localization of individual residues (surface vs. buried) provides useful information to reconstruct the tertiary structure of proteins (Chan & Dill, 1990; Wang et al., 2005; Arauzo-Bravo et al., 2006).

The existing solvent accessibility prediction methods use the protein sequence, which is converted into a fixed-size feature-based representation, as an input to predict the RSA for each of the residues. These methods can be divided into two main groups:

Real value predictors predict RSA value (the definition is given in the Material's section). The representative existing methods are based on linear regression (Wagner et al., 2005), neural network based regression (Adamczak et al., 2004), neural networks (Ahmad et al., 2003), support vector regression (Yuan & Huang, 2004; Xu et al., 2005), and look up table (Wang et al., 2004). In the study of Ahmad et al. (2003) binary coding of the sequence was taken as the input features, while all other studies use the evolutionary information in the form of the PSSM profile derived with PSI-BLAST as the input features (Wagner et al., 2005; Yuan & Huang, 2004; Adamczak et al., 2004; Wang et al., 2004; Xu et al., 2005).

Discrete value predictors classify each residue into a predefined set class. The classes are usually defined based on a threshold and include buried, intermediate, and exposed classes (in most cases the predictions concern only two classes, i. e., buried vs. exposed).

The corresponding prediction methods apply fuzzy-nearest neighbor (Sim et al., 2005), neural network (Cuff & Barton, 2000; Ahmad & Gromiha, 2002; Gianese & Pascarella, 2006), support vector machine (Kim & Park, 2004; Yuan et al., 2002), two stage support vector machine (Nguyen & Rajapakse, 2005), information theory (Naderi-Manesh et al., 2001) and probability profile (Gianese et al., 2003). Early studies only use sequence to generate features (Ahmad & Gromiha, 2002; Naderi-Manesh et al., 2001), while recent studies use the evolutionary information in the form of the PSSM profile to generate features (Nguyen & Rajapakse, 2005; Kim & Park, 2004).

The PSI-BLAST profile (Altschul et al., 1997) was recently introduced as an efficient sequence representation that improves classification accuracy (Cuff & Barton, 2000).

This paper investigates whether pace regression method could lead to improving the RSA predictions. In prediction of protein solvent accessibility with evolutionary information, the dimensions of features are high, i. e. N*20, where N is the size of the window. The idea of this paper is based on this hypothesis that if we use regression method which could be optimal with high number of features, then prediction accuracy would be improved. This result in a simplified prediction model, reduced computational time and optimized prediction quality.

A web based program of this algorithm (RSA-PRP) is available at http://bioinf.cs.ipm.ac.ir/rsa. It requires a protein sequence as input and reports the relative solvent accessibility or two states accessibility (exposed or buried) depending on predefined threshold by user.

METHODS

Datasets

The dataset used in this paper is referred to as the Manesh dataset (Naderi-Manesh, et al., 2001) and consists of 215 low-similarity proteins, i. e., < 25 %. The sequences are available online at http://gibk21.bse.kyutech.ac.jp/rvp-net/all-data.tar.gz. The Manesh dataset was widely used by researchers to benchmark prediction methods (Garg et al., 2005; Ahmad et al., 2003; Wang et al., 2004; Xu et al., 2005; Ahmad & Gromiha, 2002; Gianese et al., 2003), and this motivated us to use it to design and validate our method.

Relative solvent accessibility

RSA reflects the percentage of the surface area of a given residue that is accessible to the solvent. RSA value, which is normalized to [0,1] interval, is defined as

EXCLI Journal 2009;8:211-217 ISSN 1611-2156
Received: June 28, 2009, accepted: August 16, 2009, published: August 18, 2009

the ratio between the solvent accessible surface area (ASA) of a residue within a three-dimensional structure and ASA of its extended tri-peptide (Ala-X-Ala) conformation

$$RSA = \frac{ASA\ in\ \ 3-Dimensional\ structure}{ASA\ in\ an\ extended\ tripeptide} \quad (1)$$

Feature representation

PSI-BLAST profile. PSI-BLAST is used to compare different protein sequences to find similar sequences and to discover evolutionary relationships (Altschul et al., 1997). PSI-BLAST generates a profile representing a set of similar protein sequences in the form of a 20 × *N* position-specific scoring matrix, where *N* is the length of the sequence (window) and where each amino acid in the sequence (window) is described by 20 features. We used PSI-BLAST with the default parameters and the BLOSUM62 substitution matrix. The profile was computed for a 13 residues wide window centered on a target residue.

Terminals feature. The amino acids that are located at the two terminals of the sequence have larger probability of being exposed to the solvent. This fact was implemented during RSA prediction by using one feature with two states (0,1) that indicates if a given residue is located close to either terminus or not. *Terminals feature* set to 1, for the amino acids that are located at the first six positions at the N terminal and the last six position at the C terminal, otherwise this feature set to 0.

Linear regression

Linear regression is a well-known method of mathematical modeling of the relationship between a dependent variable and one or more independent variables (Wagner et al., 2005). Regression uses existing (or known) values to forecast the required parameters.

A linear regression with *p* coefficients and *n* data points (number of samples), assuming that *n*>*p*, corresponds to the construction of the following expression:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1\ x_{11}\ x_{12}\ \dots .\ x_{1p} \\ 1\ x_{21}\ x_{22}\ \dots .\ x_{2p} \\ 1\ \dots\ \dots\ \dots .\ \dots \\ 1\ x_{n1}\ x_{n2}\ \dots .\ x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

where y_i is the predicted RSA value, $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ is the vector *of p* features representing i_{th} protein sequence, β_i (constant) is parameter to be estimated, and ε_i is the standard error. The above formula can be written in vector-matrix form as:

$$y = X.\beta + \varepsilon \quad (3)$$

The solution to minimize the mean square error $||\varepsilon_i||$ is

$$\beta = (X^T X)^{-1} X^T \vec{y}$$
$$\vec{\varepsilon} = \vec{y} - X.\beta \quad (4)$$

Pace regression

Projection adjustment by contribution estimation (Pace) regression is a recent approach to fitting linear models, based on considering competing models (Wang, 2000).

The basic idea of regression analysis is to fit a linear model to a set of data. The classical ordinary least square's estimator is simple and has well-established theoretical justification. Nevertheless, the models produced are often not satisfactory. Pace regression improves the classical ordinary least square's regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regressions. Under regularity conditions, pace regression is provably optimal when the number of coefficients tends to infinity. Wang and Witten (2002) developed pace regression, and showed that it performs the best compared with other regression models for high dimensional data.

As with other forms of linear regression, our model for the relative solvent accessibility is a linear combination of the features in the following format:

EXCLI Journal 2009;8:211-217 ISSN 1611-2156
Received: June 28, 2009, accepted: August 16, 2009, published: August 18, 2009

$$RSA=\alpha_1*A_{i-6}+\alpha_2*N_{i-6}+\alpha_3*D_{i-6}+\alpha_4*Q_{i-6}+\alpha_5*E_{i-6}+\ldots \quad (5)$$

where the values of α_i are assigned by the regression process. It is important to note that the resulting model is a linear combination of the features. This results in a simplified prediction model and reduced computational time.

RESULTS AND DISCUSSION

The pace regression predictor was implemented in Weka, which is a comprehensive open-source library of machine learning methods (Witten & Frank, 2005). The evaluation was performed using two test types to allow for a comprehensive comparison with previous studies. To compare with Garg et al., (2005) and Ahmad et al., (2003), 5-folds cross-validation was executed. On the other hand, following several other prior studies (Wang et al., 2004; Gianese & Pascarella, 2002; Gianese et al., 2003), Manesh dataset was divided into two subsets, 30 sequences were used for training and the remaining 185 as independent test set. The results of both tests, i. e., 5 folds cross-validation and independent test, are reported in Tables 1 and 2. In total, the proposed method was compared with seven RSA prediction methods (Garg et al., 2005; Ahmad et al., 2002; Ahmad et al., 2003; Wang et al., 2004; Adamczak et al. 2004; Xu et al., 2005; Gianese et al. 2003).

Table 1: Comparison between our method and other reported methods; the results were reported based on 3 or 5-folds cross-validation test; the real value predictions were converted to two states prediction (buried vs. exposed) with different threshold (5 %~50 %); unreported results are denoted by "-".

cross-validation methods		MAE (%)	Correlation coefficient (*r*)	Accuracy for two-states (buried vs. exposed) prediction						
				5%	10%	20%	25%	30%	40%	50%
NETASA	(Ahmad et al. 2002)	-	-	74.6%	71.2%	-	70.3%	-	-	75.9%
NN	(Ahmad et al., 2003)	18.0	0.50	-	-	-	-	-	-	-
PP	(Gianese et al., 2003)	-	-	75.7%	73.4%	-	71.6%	-	-	76.2%
NN	(Garg et al., 2003)	15.2	0.67	74.9%	77.2%	77.7%	-	77.8%	78.1%	80.5%
SABLE	(Adamczak et al., 2004)	-	-	76.8%	77.5%	77.9%	77.6%	-	-	-
SVR	(Xu et al., 2005)	16.3	0.58	-	-	-		-	-	-
PR	This paper	13.14	0.6403	76.82%	74.84%	75.35%	76.7%	77.75%	79.86%	86.32%

Table 2: Experimental comparison between our method and other reported methods; the results were reported based on a test on the independent dataset (30 sequences for training and 185 sequences for test); the real value predictions were converted to two states prediction (buried vs. exposed) with different threshold (5 %~50 %); unreported results are denoted by "-".

Prediction method	Reference	MAE (%)	Correlation coefficient (*r*)	Accuracy for two-states (buried vs. exposed) prediction					
				5%	10%	20%	30%	40%	50%
Look-up table	(Wang et al., 2004)	18.8	0.48	-	-	-	-	-	-
Neural Network	(Ahmad & Gromiha, 2002)	-	-	74.6%	71.2%	-	-	-	75.9%
Neural Network	(Gianese et al., 2003)	16.3	0.58	75.7%	73.4%	-	-	-	76.2%
Pace Regression	This paper	13.40	0.6264	76.21%	74.14%	74.70%	77.30%	79.20%	86.10%

EXCLI Journal 2009;8:211-217 ISSN 1611-2156
Received: June 28, 2009, accepted: August 16, 2009, published: August 18, 2009

Evaluation measures

Two widely used measures for real value ASA prediction are adopted in this study to evaluate existing ASA predictors.

The first measure, mean absolute error (MAE), is defined as follows:

$$MAE = \frac{\sum_{for\ each\ residue} \left| RSA_{predicted} - RSA_{observed} \right|}{n} \quad (6)$$

where n is the total number of residues to be predicted, and MAE is the absolute difference between predicted and observed (from experiments) RSA values. The second measure is Pearson's correlation coefficient, which is defined for a pair of variables (X, Y) as follows:

$$r = \frac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{(x_i - \bar{x}_i)^2}\sqrt{(y_i - \bar{y}_i)^2}} \quad (7)$$

where $\bar{X}$ is the mean of X and $\bar{Y}$ is the mean of Y. The value of r is bounded within [-1, 1] interval. Higher absolute value of r corresponds to stronger correlation between X and Y.

Residues were classified into two states (buried – exposed) by different thresholds. The prediction accuracy was evaluated by the percentage of correctly predicted residues divided by the total number of residues in the test dataset. For example, for the two states we have

$$Q\% = \left[\frac{N_B + N_E}{N_{total}} \right] \quad (8)$$

where $Q\%$ is the percentage of correctly predicted residues, N_B and N_E represent the number of residues correctly predicted as buried and exposed, respectively.

Comparison with competing prediction methods

Table 3 shows the results of our approach for its effectiveness by three cross-validation tests and independent test. For the five folds cross-validation test, the mean absolute error (MAE) and the corresponding Pearson's correlation coefficient (r) values of the proposed method are equal to 13.14 and 0.6403 respectively with Manesh dataset.

Figure 1 shows the experimental and predicted values for each residue in thioredoxin (PDB code: 1ABA). We selected this protein as an example, because residues fall within different ranges of RSA values which are indicative of the high degree of accuracy of this prediction across a wide range of RSAs and amino acid residues. It shows good linear relationship between the experimental and predicted values.

Table 3: Evaluation of our approach with Manesh dataset

cross-validation methods	MAE (%)	Correlation coefficient (r)	Accuracy for two-states (buried vs. exposed) prediction					
			5%	10%	20%	30%	40%	50%
7 fold cross-validation	13.13	0.6401	76.79%	74.80%	75.29%	77.71%	79.83%	86.32%
5 fold cross-validation	13.14	0.6403	76.82%	74.84%	75.35%	77.75%	79.86%	86.32%
3 fold cross-validation	13.15	0.6394	76.83%	74.81%	75.24%	77.77%	79.77%	86.31%
Independent dataset	13.40	0.6264	76.21%	74.14%	74.70%	77.30%	79.20%	86.10%

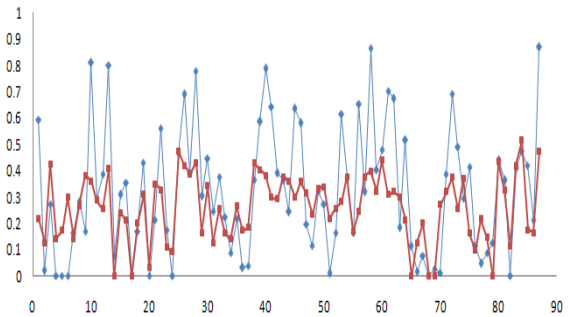

Figure 1: Example of predicted (red) and experimental (blue) RSA values for a protein (PDB code 1ABA)

Table 1 shows the comparison between the proposed PR model and recent methods for RSA prediction, which include neural network and support vector regression models (Ahmad et al., 2002; Ahmad et al., 2003; Gianese et al., 2003; Garg et al., 2003; Adamczak, et al., 2004; Xu, et al., 2005). MAE of the proposed method is 2.06 to 4.86 lower than above mentioned methods.

Since some methods predict discrete valued classes (exposed vs. buried), we also examined the performance of our method by converting the real value prediction into the two states prediction. We followed the standard approach in which the state is defined based on the predicted RSA value and a predefined threshold. For instance, a 5 % threshold means that the residues having an RSA value (%) greater or equal to 5 % are defined as exposed, and otherwise they are classified as buried.

For the independent test, the MAE value for the PR method is equal to 13.4 and the corresponding Pearson's correlation coefficient (*r*) is equal to 0.63. Table 2 shows the comparison of the PR model with recent methods for RSA prediction, which include neural network and look-up table based methods (Wang et al., 2004; Ahmad & Gromiha, 2002; Gianese et al., 2003). MAE of the proposed method is 2.9 to 5.4 lower than above mentioned methods. Similar to the 5-folds cross-validation test, we also evaluated the performance of our method by converting the real value prediction into the two states prediction. The threshold value was adjusted between 5 and 50 %, see Table 2. When compared with the best performing, competing method based on neural network (Gianese et al., 2003), our prediction results have higher accuracy over all thresholds, and also better MAE and correlation coefficient values.

As a final remark on experimental results, it should be pointed out that the focus of the paper was mainly on the impact of using a pace regression method that performs well for high dimensional data, on the overall performance of RSA prediction. The fact that Garg et al. (2003) obtained better results in some thresholds is not surprising, because that system being based on a two stage method and therefore it uses more features to predict real value for accessible surface area and they are computationally more expensive than our simple method.

Acknowledgements: This wok was supported in part by a grant from IPM (No. CS 1385-1-02).

REFERENCES

Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56:753-67.

Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18: 819-24.

Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629-35.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;17:3389-402.

Anfinsen CB. Principles that govern the folding of protein chains. Science 1973; 181:223–30.

Arauzo-Bravo MJ, Ahmad S, Sarai A. Dimensionality of amino acid space and solvent accessibility prediction with neural networks. Comput Biol Chem 2006;30:160-8.

Chan HS, Dill KA. Origins of structures in globular proteins. Proc Natl Acad Sci USA 1990;87:6388-92.

Chou KC. Low-frequency collective motion in biomacromolecules and its biological functions. Biophys Chem 1988; 30:3-48.

Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502-11.

Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005;61: 318-24.

Gianese G, Pascarella S. A consensus procedure improving solvent accessibility prediction. J Comput Chem 2006;27:621-6.

Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. Protein Eng 2003;16: 987-92.

Ginalski K, Rychlewski L. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. Proteins 2003;53:410-7.

Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;6:19.

Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54:557-62.

Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Predicting of protein surface accessibility with information theory. Proteins 2001;42:452-9.

Nguyen MN, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 2005; 59:30-7.

Sim J, Kim SY, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. Bioinformatics 2005; 21:2844-9.

Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 2005;12:355-69.

Wang Y. A new approach to fitting linear models in high dimensional spaces. PhD thesis. Department of Computer Science, University of Waikato, New Zealand, 2000.

Wang Y, Witten IH. Modeling for optimal probability prediction. In: Proceedings of the 19th International Conference in Machine Learning, Sydney, Australia, 2002, pp. 650-7.

Wang JY, Ahmad S, Gromiha MM, Sarai A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. Biopolymers 2004;75:209-16.

Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. Proteins 2005;61:481-91.

Witten I, Frank E. Data mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2005.

Xu WL, Li A, Wang X, Jiang ZH, Feng HQ. Improving prediction of residue solvent accessibility with SVR and multiple sequence alignment profile. In: Proceedings of the 27th IEEE Annual Conference on Engineering in Medicine and Biology, Shanghai, China, 2005, pp. 2595-8.

Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;57:558-64.

Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48: 566-70.