Improved Prediction of Relative Solvent Accessibility Using Two-stage Support Vector Regression

Ke Chen[1], Michal Kurgan[1], Lukasz Kurgan*[1]

[1]University of Alberta, Department of Electrical and Computer Engineering, Edmonton, CANADA, T6G 2V4
*lkurgan@ece.ualberta.ca (corresponding author)

Abstract

Predicted relative solvent accessibility (RSA) provides useful information for prediction of binding sites and reconstruction of the 3D-structure based on a protein sequence, which are at the very core of proteomics. Several RSA prediction methods including those that generate real values and those that predict discrete states (buried vs. exposed) have been published. We propose a novel method for real valued prediction that aims to improve the prediction quality when compared with the existing methods. The proposed method combines Support Vector Regression (SVR) predictors into a two-stage architecture. The improved prediction quality comes from a composite sequence representation, which includes a custom-selected subset of features from the PSI-BLAST profile, secondary structure predicted with PSI-PRED, and binary code that indicates position of a given residue with respect to sequence termini. Based on empirical evaluation with a standard benchmark dataset, the proposed method obtains the mean absolute error (MAE) equal 0.143, which corresponds to 6% error rate reduction when compared with the best performing competing method that obtains 0.152 MAE on this dataset.

Keywords: relative solvent accessibility; support vector regression; PSI-BLAST; PSI-PRED; secondary protein structure

1. Introduction and Related Work

The knowledge of protein structure is invaluable in understanding protein's function. Computational prediction of the tertiary protein structure is one of the most important topics in proteomics due the large and exponentially growing gap between the number of known protein sequences and the number of known structures. Despite several decades of extensive research 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 [1]. The predictions of the secondary structure and the solvent accessibility were proposed as an intermediate step in prediction of the tertiary structure; several such in-silico prediction methods have been already proposed [6, 20].

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 is helpful for prediction of binding sites [2]. Chan and Dill showed that the burial of core residues is a strong driving force in protein folding, which means that knowledge of localization of individual residues (surface vs. buried) provides useful information to reconstruct the 3D-structure of proteins [3].

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

- *real valued* predictors that predict RSA value. They apply linear regression, multiple linear regression, neural networks, support vector regression, and neural network based regression [4-10].
- *discrete valued* predictors which classify each residue into a predefined set classes. The classes are usually defined based a threshold and include buried, intermediate, and exposed (in most cases the predictions concerns only two classes, i.e., buried vs. exposed). They apply fuzzy-nearest neighbor, neural network, probability profile, support vector machine, and two stage support vector machine [11-18].

A PSI-BLAST profile [19] was recently introduced as an efficient sequence representation that improves classification accuracy [11]. Subsequently, researchers have found that secondary structure predicted using the PSI-PRED method [20] helps to improve the real value predictions [6].

To this end, we propose a novel real value predictor which is based on a two-stage support vector regression. As suggested in previous work, the PSI-BLAST profile, secondary structure predicted with PSI-PRED, and additional features that indicate termini of the sequence were adopted to represent the input sequence. In contrast to prior works, we do not use all features from the PSI-BLAST profile, but instead we use a linear correlation based feature selection method to select a subset of best-performing features. This approach results in a simplified prediction model, reduced computational time, and optimized predictive quality.

2. Materials and Methods

2.1. Dataset

1-4244-1120-3/07/$25.00 ©2007 IEEE

The dataset used in this paper is referred to as the Manesh dataset [18] and consist of 215 non-homologous, i.e., < 25% sequence homology, protein chains. The sequences are available online at http://gibk21.bse.kyutech.ac.jp/rvp-net/all-data.tar.gz. The Manesh dataset was used by several researchers to benchmark their prediction methods, and this motivated us to use it in our research.

2.2. Relative Solvent Accessibility

RSA reflects the percentage of the surface area of a given residue that is accessible to the solvent. RSA values, which are normalized to [0, 1] interval, are defined as the ratio between the solvent accessible surface area (ASA) of a residue within a three-dimensional structure and ASA of its extended tripeptide (Ala-X-Ala) conformation

$$RSA = \frac{ASA \text{ in a three-dimensional structure}}{ASA \text{ in an extended tripepetide}} \quad (1)$$

2.3. Feature Generation

PSI-BLAST profile. PSI-BLAST is used to compare different protein sequences to find distant relatives and to discover evolutionary relationships [19]. PSI-BLAST generates a profile representing a set of similar proteins in the form of a 20×*N* position-specific scoring matrix, where *N* is the length of the sequence (window) and 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 15 residues wide window centered on a target residue. The selected size is motivated by previous studies that adopted this window size [14] and good results obtained for the secondary structure prediction with this window [20]. The selected window corresponds to the 15×20 scoring matrix, which gives total of 300 features.

Secondary structure predicted with PSI-PRED. The quality of secondary structure prediction significantly improved in the last decade and nowadays it is successfully used in prediction of tertiary structure. Recently, secondary structure predicted with the PSI-PRED algorithm was shown to improve prediction of solvent accessibility [6]. We used PSI-PRED25 with default parameters to predict secondary structure from the protein sequences. PSI-PRED assigns three probabilities for each residue, which correspond to probability of assuming helix, strand, and coil conformation, respectively. These probabilities were taken as features for the proposed RSA prediction method.

Binary code. The amino acids that are located at the two termini of the sequence have larger probability of being exposed to the solvent. This fact is implemented during RSA prediction by using a simple binary code that indicates position of a given residue that is located close to either terminus. The following binary vector

$$(a_1, a_2, a_3, a_4, a_5, b_1, b_2, b_3, b_4, b_5)$$

is used to encode the first five positions at the N terminus (denoted by a_i) and the last five position and the C terminus (denoted by b_i). For instance, the third residue in the sequence is encoded as (0,0,1,0,0,0,0,0,0,0), while a residue that is outside of the first and the last five residues in the sequence is encoded as (0,0,0,0,0,0,0,0,0,0).

2.4. Feature Selection

PSI-BLAST profile includes 300 features, and thus a feature selection method was used to reduce the dimensionality and potentially improve the prediction accuracy. We applied the *correlation-based feature selection* [21], which is based on Pearson correlation coefficient *r* computed for a pair of variables (*X*, *Y*) as

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

where $\overline{x_i}$ is the mean of *X*, and $\overline{y_i}$ is the mean of *Y*. The value of *r* is bounded within [-1, 1] interval. Higher absolute value of *r* corresponds to higher correlation between *X* and *Y*. This method ranks individual features based on the correlation coefficient between each feature and the actual RSA values. A subset of features with the highest absolute *r* value is selected.

The 300 features corresponding to the PSI-BLAST profile, 3 features corresponding to the predicted secondary structure and 10 binary code values were processed with the feature selection method. As a result, the best 70 features were selected, i.e., the lower ranked feature did not improve RSA prediction. The selected features include 65 features from the PSI-BLAST profile, all 3 predicted secondary structure features, and 2 binary code values that correspond to the first and last position in the sequence, see Tables 1 and 2.

Table 1. Summary of the feature selection results.

Features set	Total # features	# selected features
PSI-BLAST profile	300	65
Binary code	10	2
Predicted secondary structure	3	3
Total	313	70

The feature selection exercise shows that most of the 300 features generated by PSI-BLAST are either redundant and have little or no impact on the RSA predictions. Table 2 shows that when predicting RSA for A_i, the features to encode the first 2 amino acids (A_{i-7} A_{i-6}) and the last amino acids (A_{i+7}) were not selected, i.e., these amino acids have no impact on the prediction of the central amino acid. Therefore, a sliding window of size 13 would be sufficient for the RSA prediction. Additionally, the two amino acids that are adjacent to A_i, i.e., A_{i-1} and A_{i+1}, have the most significant impact on the prediction since they correspond to

Table 2. Summary of feature selection results for the PSI-BLAST profile.

15-wide window	A_{i-7}	A_{i-6}	A_{i-5}	A_{i-4}	A_{i-3}	A_{i-2}	A_{i-1}	A_{i}	A_{i+1}	A_{i+2}	A_{i+3}	A_{i+4}	A_{i+5}	A_{i+6}	A_{i+7}
Total # of features	20	20	20	20	20	20	20	20	20	20	20	20	20	20	20
# of selected features	0	0	2	4	5	0	8	19	7	1	6	6	4	3	0

Table 3. Experimental comparison between the proposed two-stage SVR and other reported methods; the real valued predictions were also converted to two state prediction (buried vs. exposed) with different threshold (5%~50%); unreported results are denoted by "-"; best results are shown in bold.

Reference	Prediction method	MAE (%)	Correlation coefficient *r*	Accuracy for two-states (buried vs. exposed) prediction					
				5%	10%	20%	30%	40%	50%
[6]	Neural Network	15.2	0.67	74.9%	77.2%	77.7%	77.8%	78.1%	80.5%
[10]	Neural Network	18.0	0.50	-	-	-	-	-	-
This paper	Two-stage SVR	**14.3**	**0.68**	**81.1%**	**79.7%**	**78.8%**	**78.6%**	**78.8%**	**80.8%**

the largest number of the selected features. Interestingly, residues at *i*-2 and *i*+2 positions have relatively small influence on the prediction.

2.5. Prediction Method

Support Vector Regression (SVR) was already applied in the RSA prediction [8]. In this paper, we propose a better-performing two-stage SVR model. Due to the page limits, we do not describe the SVR model (the reader is referred to [8]), but instead we focus on describing differences between the one-stage SVR and the applied two-stage SVR.

First, 70 features were generated using a 15 wide window for each residue in the input sequence. In the first stage, the 70 features are used as input and SVR predicts a real value (predicted RSA value) for each residue. The second stage aims to refine the first stage predictions. Similarly to a two-stage neural network designs, the second stage smoothes the predictions. It takes the three predicted secondary structure features and a 7 wide window of the first stage predicted values centered over the being predicted residue as the input to provide the final real valued predictions. The detailed procedure is shown in Figure 1.

The optimization of the prediction, through adjustment of internal parameters of SVR (kernel type and parameters and complexity parameter *C*) and selection of the window size for the second stage SVR, was performed by dividing the Manesh dataset into two subsets, one used to compute the prediction model and the other to perform test. Similarly to [15], 30 sequences were used for training and the remaining 185 as the test set. As a result, RBF kernel was used for both stages. The parameters for the first stage SVR are γ=0.01 and *C*=1, and for the second stage γ=0.15 and *C*=1. The mean absolute error (MAE) of the final prediction for the second stage windows sizes of 5, 7, 9, 11, 15, and 21 equals 0.149, 0.148, 0.148, 0.148, 0.148, and 0.148, respectively. This shows that the window size of 7 is sufficient to provide accurate predictions.

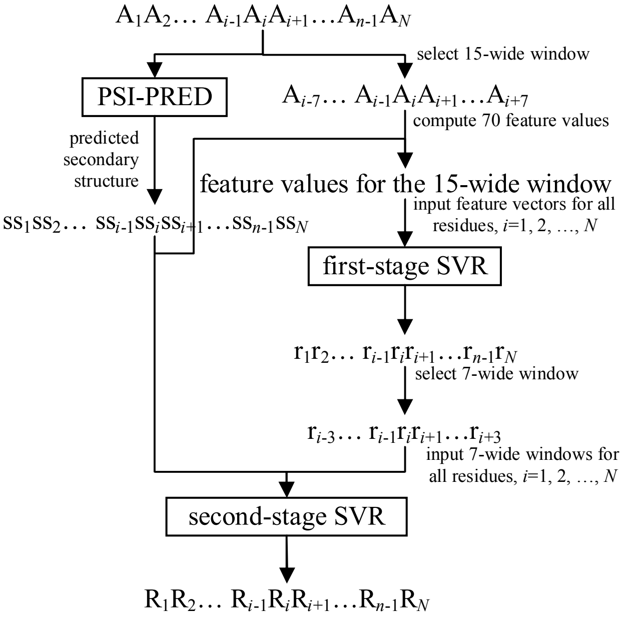
$A_1A_2\ldots A_{i-1}A_iA_{i+1}\ldots A_{n-1}A_N$
select 15-wide window
PSI-PRED
$A_{i-7}\ldots A_{i-1}A_iA_{i+1}\ldots A_{i+7}$
compute 70 feature values
predicted secondary structure
feature values for the 15-wide window
input feature vectors for all residues, i=1, 2, …, N
$ss_1ss_2\ldots ss_{i-1}ss_iss_{i+1}\ldots ss_{n-1}ss_N$
first-stage SVR
$r_1r_2\ldots r_{i-1}r_ir_{i+1}\ldots r_{n-1}r_N$
select 7-wide window
$r_{i-3}\ldots r_{i-1}r_ir_{i+1}\ldots r_{i+3}$
input 7-wide windows for all residues, i=1, 2, …, N
second-stage SVR
$R_1R_2\ldots R_{i-1}R_iR_{i+1}\ldots R_{n-1}R_N$

Figure 1. RSA prediction with the proposed system; the RSA value for the i^{th} residue is predicted based on the 70 feature values (see Table 1) that are computed over a 15 residues wide window centered on i^{th} residue; the feature values are inputted into the first-stage SVR; next, the first-stage predictions are aggregated into 7 residue wide windows and inputted, together with the predicted secondary structure of the central residue, into the second-stage SVR that provides the RSA values.

3. Results and Discussion

The SVR predictor [22] was implemented in Weka [23], which is a comprehensive open-source library of machine learning methods. The Manesh dataset consists of 50682 instances (individual residues), and evaluation was done by 5-folds cross validation to assure objective comparison with previous works.

The MAE value for the first stage of the proposed method equals 0.146 and the corresponding Pearson's correlation coefficient (*r*) equals 0.67. After the second stage, the MAE value is reduced to 0.143 and *r* is improved to 0.68. Table 3 compares the proposed two-stage SVR with recent methods for real valued RSA prediction by Gard and colleagues [6] and Ahmad and colleagues [10]. The proposed method obtains 0.9% and 3.7% lower MAE when compared with methods proposed in [6] and [10], respectively. This translates into 6% and 20% error reduction, respectively.

Since some methods predict discrete valued classes (exposed vs. buried), we also examined the performance of our method by converting the real valued prediction into the two states prediction. We followed the standard procedure, in which the state is defined based on the predicted RSA value and a pre-defined threshold. For instance, a 5% threshold means that the residues having an RSA value (%) greater or equal 5 are defined as exposed, and otherwise they are classified as buried. The threshold's value is usually adjusted between 5 and 50%. When compared with the best-performing, recent method from [6], see Table 3, our predictions obtain higher accuracies over all thresholds, i.e., the differences range between 0.3% and 6.2%. We also note that results from one-stage SVR model introduced in [8] are comparable with results from the neural network model presented in [10], which in turn are worse than results shown in [6]. Since one-stage SVR was not tested on the Manesh dataset, we could not include these results in Table 3.

4. Conclusions

This paper proposes a novel method for the real valued RSA prediction, which is based on a two-stage SVR and a custom-designed sequence representation. Empirical tests with the Manesh dataset show that the proposed method is characterized by lower prediction error when compared with competing methods for the real valued predictions. We also show that the PSI-BLAST profile that is commonly used to represent sequences can by largely reduced by using feature selection, which would result in reduction of the computational time required to develop the prediction model. Our results indicate that window size of 13 is sufficient and only about 22% of the PSI-BLAST features are useful for the RSA prediction.

References

[1] 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.

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

[3] Chan HS, Dill KA. Origins of structures in globular proteins. *Proc Natl Acad Sci USA* 1990;87: 6388 – 6392.

[4] Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins*. 2005; 61(3):481-91.

[5] Arauzo-Bravo MJ, Ahmad S, Sarai A. Dimensionality of amino acid space and solvent accessibility prediction with neural networks. *Comput Biol Chem*. 2006 (2):160-8.

[6] Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*. 2005; 61(2):318-24.

[7] Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. *J Comput Biol*. 2005; 12(3):355-69.

[8] Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins*. 2004; 57(3):558-64.

[9] Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*. 2004; 56(4):753-67.

[10] Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins*. 2003; 50(4):629-35.

[11] Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*. 2000; 40(3):502-11.

[12] Sim J, Kim SY, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. *Bioinformatics*. 2005; 21(12):2844-9.

[13] Nguyen MN, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stage SVM approach. *Proteins*. 2005; 59(1):30-7.

[14] 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):557-62.

[15] Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*. 2002; 18(6):819-24.

[16] Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins*. 2004; 57(3):558-64.

[17] Gianese G, Pascarella S. A consensus procedure improving solvent accessibility prediction. *J Comput Chem*. 2006; 27(5):621-6.

[18] Naderi-Manesh H, Sadeghi M, Araf S, Movahedi AAM. Predicting of protein surface accessibility with information theory. *Proteins* 2001;42:452–459.

[19] 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, 17:3389–402

[20] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*. 1999; 292(2):195-202.

[21] Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the 10th International Conference on Machine Learning*. 2003; 856-863

[22] Alex J. Smola, Bernhard Scholkopf. "A Tutorial on Support Vector Regression". NeuroCOLT2 Technical Report Series. 1998

[23] Witten I, Frank E, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005