EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Original article:

PREDICTION OF RELATIVE SOLVENT ACCESSIBILITY BY SUPPORT VECTOR REGRESSION AND BEST-FIRST METHOD

Alireza Meshkin[1,*], Hossein Ghafuri[1]

[1] Department of Computer Engineering, Payam Nour University of Damavand, Tehran, Iran
[*] Corresponding author: Email: meshkin@nigeb.ac.ir

ABSTRACT

Since, it is believed that the native structure of most proteins is defined by their sequences, utilizing data mining methods to extract hidden knowledge from protein sequences, are unavoidable. A major difficulty in mining bioinformatics data is due to the size of the datasets which contain frequently large numbers of variables. In this study, a two-step procedure for prediction of relative solvent accessibility of proteins is presented. In a first "feature selection" step, a small subset of evolutionary information is identified on the basis of selected physicochemical properties. In the second step, support vector regression is used to real value prediction of protein solvent accessibility with these custom selected features of evolutionary information. The experiment results show that the proposed method is an improvement in average prediction accuracy and training time.

Keywords: Feature selection method; physicochemical properties of amino acids; PSI-BLAST; support vector regression

Abbreviations: RSA: Relative Solvent Accessibility; SVR: Support Vector Regression; PSSM: Position Specific Scoring Matrix

INTRODUCTION

Protein native structure strongly influences the protein's biological function, thus it is relevant to study protein functions, knowing the protein tertiary structure and thus its solvent accessibility. Because knowledge of the solvent accessibility of a protein plays a vital role in predicting the tertiary structure of the protein.

Accessible Surface Area (ASA) reflects the percentage of the surface area of a given residue that is accessible to the solvent. Relative Solvent Accessibility (RSA) was computed by the ASA of a residue normalized by the ASA of this residue in its extended tripeptide (Ala-X-Ala) conformation.

This paper investigates whether improved sequence representation, which is based on the custom selected features harvested from evolutionary information, could lead to improving the accuracy of RSA prediction. In prediction of protein solvent accessibility with evolutionary information, the dimensions of features are too high, i. e. N*20, where N is the size of the window. The idea of this paper is based on the hypothesis that if data mining features selection methods are used for selecting subset of best-performing features, then prediction accuracy and training time would be improved. This idea results in a simplified prediction model, reduced computational time, and optimized prediction quality.

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

The goals of this paper are achieved by designing a two-step procedure for prediction of relative solvent accessibility of proteins. In a first "feature selection" step, a relatively small subset of evolutionary information is identified on the basis of selected physicochemical properties in each position of the given window. In the second step, support vector regression method is used to real value prediction of protein solvent accessibility with these custom selected features of evolutionary information.

PREVIOUS RELATED WORKS

The existing solvent accessibility prediction methods can be divided into two main groups:

- *Real valued predictors* that predict real-value of solvent accessibility. The representative existing methods are based on linear regression (Wagner et al., 2005), neural network based regression (Adamczak et al., 2004), neural networks (Shandar et al., 2003; Faraggi et al. 2009; Petersen et al. 2009; Dor and Zhou, 2007), support vector regression (Yuan and Huang, 2004; Xu et al., 2005), pace regression (Meshkin et al., 2009) and look up table (Wang et al., 2004). In the study of Shandar et al. (2003), binary coding of the sequence was taken as the input features, while all other studies use the evolutionary information (Wagner et al., 2005; Adamczak et al., 2004; Yuang and Huang,, 2004; Xu et al., 2005; Wang et al., 2004).
- *Discrete valued 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 and Barton, 2000; Shandar and Gromiha, 2002; Gianese and Pascarella, 2006), support vector machine (Kim and Park, 2004; Yuan et al., 2002), two stage support vector machine (Nguyen and Rajapakse, 2005), information theory (Naderi-Manesh et al., 2001), and probability profile (Gianese et al., 2003). Early studies only used sequence to generate features (Shandar and Gromiha, 2002; Naderi-Manesh et al., 2001), while recent studies have used the evolutionary information (Kim and Park, 2004; Nguygen and Rajapakse, 2005).

Some conformational structures are mainly determined by local interactions between near residues, whereas others are due to distant interactions in the same protein. Therefore, with reducing number of feature in each position of window, we can enlarge the window size and then the effects of more neighbors can be considered for better prediction of RSA values. In addition, reducing dimensionality and removing irrelevant data has further advantages such as reducing the costs of data acquisition, better understanding of the prediction model, and a decrease in training time.

Considering the advantages that are mentioned above, it seems to be important to investigate the idea of this paper. With regard to the too high number of PSSM profile features (in a window with size N), the main practical aim of this work is to find an optimal subset of features among a set of N*20 features which enables an efficient prediction of relative solvent accessibility of proteins.

MATERIALS

In this section, the dataset is introduced, then qualitative and quantitative features are described.

Dataset

In this study, the Manesh dataset (Naderi-Manesh et al., 2001) is used and it 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 has been widely used by researchers to benchmark prediction methods Adamczak et al., 2004; Meshkin et al., 2009; Shandar and Gromiha, 2002; Garg et al., 2005; Gianese et al., 2003), and this motivated us to use it to design and validate our method.

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Qualitative features

As shown in Table 1, 48 qualitative properties of amino acids are applied for encoding each of 20 amino acids. Qualitative features for a window surrounding the given amino acid are represented by a bipolar vector. Instead of using the physicochemical values, for a given property, the amino acids are grouped based on the binary classification, assigning 1 for those residues having or strongly showing the property and -1 for those without the property. According to this grouping scheme, each amino acid is encoded and represented by a 48-dimensional vector.

The bipolar vector was produced for a 13 residues wide window centered on a target residue. There are 13*48+1 features in a bipolar vector for each residue in a sequence. The pattern of input vector is shown in (1).

$$(f_1, f_2, \ldots, f_{623}, f_{624}, RSA_{\text{for a given residue}}) \quad (1)$$

For instance, physicochemical features for a window surrounding the given amino acid are encoded as (2).

$$(-1, +1, \ldots, -1, -1, 0.87) \quad (2)$$

After creating qualitative input vectors for all residues of proteins in Manesh dataset (Naderi-Manesh et al., 2001), subset of physicochemical properties which have a strong correlation with the relative solvent accessibility of proteins is selected by feature selection method.

Quantitative features

For a protein sequence, the position specific scoring matrix (PSSM) describes the likelihood of a particular residue substitution at a specific position based on evolutionary information. 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* PSSM matrix, where *N* is the length of the sequence and where each amino acid in the sequence is described by 20 features. Since the profile features created by sequence alignment and quantitative criterions, we called them quantitative features. We used PSI-BLAST with the default parameters and the BLOSUM62 substitution matrix in this study.

METHODS

Figure 1 shows a detailed overview of the prediction procedure that consists of two steps, the first is aimed for creating input vector by subset selection of evolutionary features, the second is responsible for model building.

The proposed two-step method works as follows: The task of the first step is grouped into two subtasks: “Physicochemical Feature Selection” and “Evolutionary Information Selection”. In “Physicochemical Feature Selection” subtask, we select subset of physicochemical properties of amino acids in each position of a window which have a strong correlation with relative solvent accessibility of proteins.

Whenever, the subset of physicochemical features is selected, in “Evolutionary Information Selection” subtask, amino acids that have those selected physicochemical properties are chosen in each position of window. Finally, we have subset of best-performing features from PSI-BLAST profile, which are used in the next step for training the model.

The second step is responsible for building model. This step performs core ability and explores unknown relationships between selected PSSM features and RSA by learning from training data. It creates model for RSA prediction of protein sequences. Support vector regression with RBF kernel applied in this step.

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Table 1: 48 physicochemical properties of amino acid

id	property	AA	id	property	AA	id	property	AA
1	**aromatic**	HFWY	17	**carboxyl**	DE	33	**polar/hydrophilic**	RNDEQHKSTWY
2	**UV absorbance**	FWY	18	**carbonyl**	NDEQ	34	**hydrophobic**	ACILMFPWYV
3	**single aromatic ring**	FY	19	**imidazole**	H	35	**very hydrophobic**	ACILMFV
4	**heteroaromatic**	HW	20	**guanidino**	R	36	**weak hydrophobic**	PWY
5	**aliphatic**	GAILVP	21	**amino**	RK	37	**H-bonding**	RNDCEQHKSTWY
6	**branched**	ILTV	22	**symmetrical alpha-C**	G	38	**H-acceptor**	NDCEQHSTY
7	**branched beta-carbon**	ITV	23	**alkyl**	AILV	39	**H-donor**	RNCQHKSTWY
8	**flexible**	G	24	**achiral**	G	40	**tiny**	GA
9	**inflexible**	P	25	**2 chiral centers**	IT	41	**very small**	GASC
10	**alpha imino**	P	26	**ionizable**	RDCEHKY	42	**medium small**	VTNDP
11	**hydroxyl**	STY	27	**charged (pH 6.5-7)**	RDEHK	43	**small**	GASCVTNDP
12	**hydroxyl straight chain**	ST	28	**acidic**	DE	44	**large (bulky)**	KRFYW
13	**phenol**	Y	29	**basic**	RKH	45	**long**	KREQ
14	**sulfur**	CM	30	**strong basic**	RK	46	**very long**	KR
15	**sulfhydryl**	C	31	**weak hydrophilic**	STWY	47	**medium-long**	EQ
16	**amide**	NQ	32	**very hydrophilic**	RNDEQHK	48	**short**	GASCT

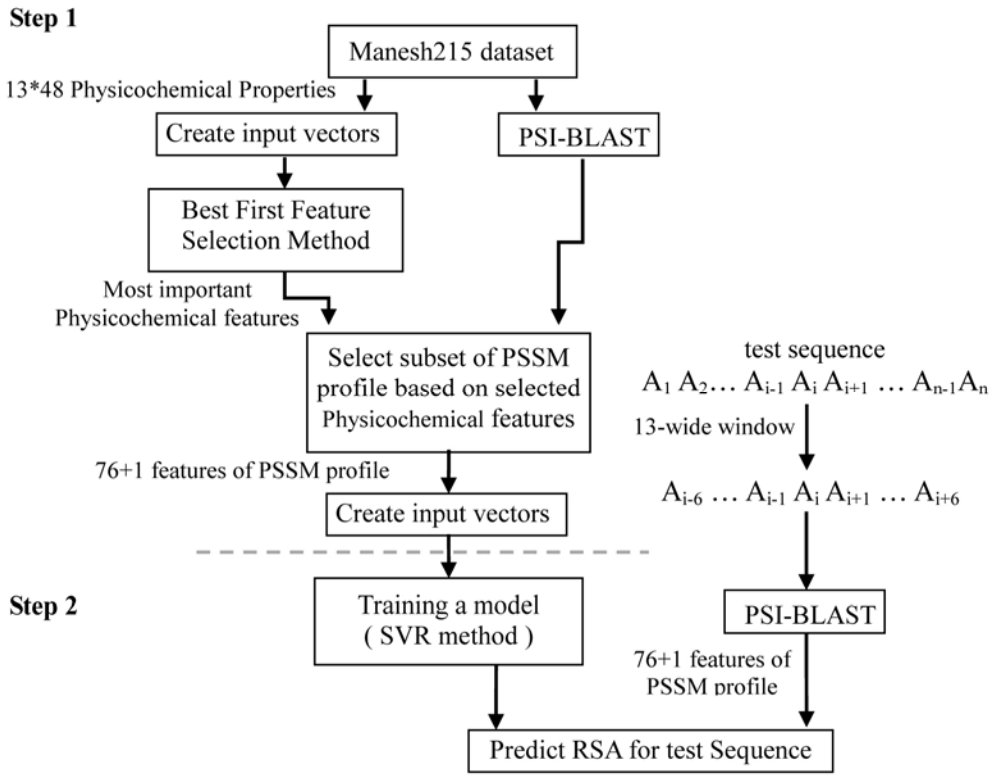

Figure 1: A detailed overview of the proposed method

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Feature selection

Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility that performed in the first step of our proposed method.

Feature selection method was used to find subset of physicochemical properties of amino acids which have a strong correlation with relative solvent accessibility of proteins. We applied the best-first method for selecting a subset of physicochemical features. The best-first method searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility.

We applied the best-first (Korf, 1993) method with forward direction and use CfsSubsetEval (Hall, 1998) method to evaluate the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of physicochemical features that are highly correlated with the RSA values while having low intercorrelation are preferred. The best-first method filters the redundancy among the physicochemical features and selects the final number of selected features, which in our case were 31 features. Table 2, shows the selected physicochemical features which have strong relationship with RSA value of the residue A_i that is located in the center of the window with size 13.

Whenever, the subset of qualitative features are produced, a set of amino acids that have those selected properties are chosen in each position of window, for example, in position A_{i+3}, if inflexibility or very hydrophobic properties are selected, we select only amino acids that have at least one of those properties in that position. Finally, we have a subset of PSI-BLAST profile features, which used for training a model in the second step, see Table 3.

Table 2: Results of subset selection of physicochemical features

13-wide window	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}
Total # of features	48	48	48	48	48	48	48	48	48	48	48	48	48
# of selected features	1	1	1	2	0	3	12	3	2	2	2	1	1
Selected Physicochemical Features	flexible	flexible	long	polar_hydrophilic		branched_beta_carbon	very_hydrophobic very_hydrophilic polar_hydrophilic	flexible	flexible	inflexible	sulfur	flexible	very_hydrophobic
				medium_long		inflexible	single_aromatic_ring Hydrophobic Branched	2_chiral_centers	inflexible	very_hydrophobic	long		
						alpha_imino	Carboxyl Charged Carbonyl	hydrophobic					
							Sulfhydryl Amino Long						

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Table 3: Results of subset selection of evolutionary information

13-wide window	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}
Total # of features	20	20	20	20	20	20	20	20	20	20	20	20	20
# of selected features	1	1	4	11	0	4	19	12	2	8	6	1	7
The selected features	G	G	R	RY		I	C D	AY	G	A	R	G	A
			Q	N		P	EF	CV	P	C	C		C
			E	D		T	AH	G		I	Q		I
			K	Q		V	IK	I		L	E		L
				E			LM	L		M	K		M
				H			NP	M		F	M		F
				K			QR	F		P			V
				S			ST	P		V			
				T			VW	T					
				W			Y	W					

The selected features include 76 features from the PSSM profile and one binary value that corresponds with the residue that is located close to either terminus of the sequence. We add this binary feature; because the amino acids that are located at the two terminus of the sequence have larger probability of being exposed to the solvent, see Table 4.

Among the 13*48 qualitative features, only 31 physicochemical features deemed more significant for prediction of RSA in a given window. The first step of our method discovered all the valuable knowledge about which qualitative features deemed more interesting for prediction of RSA, such as:

- The physicochemical features of the central residue i. e. A_i have the strongest correlation on the prediction. Interestingly, features of other residues have relatively small influence at the prediction.

- The residues that are located in A_{i-6}, A_{i-5}, A_{i-2}, A_{i+2}, A_{i+5} positions, have too low impact on the RSA prediction of the central amino acid.

- The features of A_{i-2} amino acid were not selected, i. e. this residue has no impact on the RSA prediction of the central amino acid.

- Hydrophilicity, hydrophobicity, long, flexibility and inflexibility features of amino acids have strong correlation with RSA values because these features are mentioned in many positions of a window in Table 1.

- Among the 48 physicochemical features of amino acids, only 20 distinct physicochemical features have strong correlation with protein solvent accessibility.

Support vector regression

Given a training set of *n* data point pairs $(x_i, y_i), i = 1,2, \ldots, n$, where x_i denotes the vector of features representing i^{th} protein sequence, y_i denotes the predicted RSA value, finding the optimal SVR is achieved by solving:

$$\min \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*) \quad (3)$$

Such that

$$y_i - w.x_i - b \leq \varepsilon + \xi_i \quad (4)$$

$$w.x_i + b - y_i \leq \varepsilon + \xi_i^* \quad (5)$$
$$\xi_i, \xi_i^* \geq 0$$

Where w is a vector $w.x - b = 0$ perpendicular to hyperplane, C is a user defined complexity constant, ξ_i and ξ_i^* are slack variables that measure the degree of prediction error of x_i for a given hyperplane, and $z = \Phi(x)$ where

$k(x, x') = \Phi(x).\Phi(x')$ is a user defined kernel function.

The SVR was trained using sequential minimal optimization algorithm (Smola and Scholkopf, 1998) that was further optimized by Shevade and colleagues (1999). The proposed SVR uses RBF kernel (6).

$$k(x_i, x_i') = e^{-\gamma\|x - x'\|^2} \quad (6)$$

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Table 4: The total count of selected features

Feature set	Number of features (without feature selection)	Number of features (with feature selection)
Evolutionary information	13*20=260	76
Terminus feature	1	1
The total count of features	261	77

RESULTS AND DISCUSSION

The SVR and best-first methods were implemented in weka, which is a comprehensive open-source library of machine learning methods (Witten and Frank, 2005). The evaluation was performed using 10 fold cross validation test type to allow for a comprehensive comparison with previous studies.

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

$$Q_{\%} = \left[\frac{N_E + N_B}{N_{total}}\right] \quad (7)$$

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

Comparison with other prediction methods

Figure 2 shows the experimental and predicted values for each residue in thioredoxin. 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.

Since the model training in our method is done in one stage, our method should be compared with methods that their training is done in one stage.

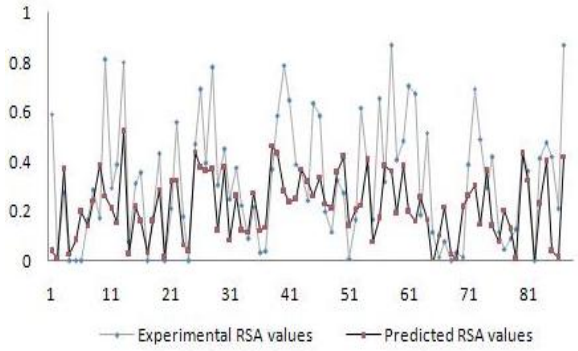

Figure 2: Example of predicted RSA values for a protein (PDB code 1ABA)

Table 3 shows the comparison between this paper and one stage methods for RSA prediction, which include neural network and SVR models (Adamczak et al., 2004; Meshkin et al., 2009; Shandar and Gromiha, 2002; Garg et al., 2005; Gianese et al., 2003).

Since methods predict discrete valued classes (exposed vs. buried), we 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 5 are defined as exposed, and otherwise they are classified as buried. The threshold's value is usually adjusted between 5 % and 50 %. We note that for most of thresholds, our method provides more accurate two states predictions, see Table 5.

EXCLI Journal 2010;9:29-38 ISSN 1611-2156
Received: January 11, 2010, accepted: January 26, 2010, published: February 08, 2010

Table 5: Comparison between our method and other reported methods; unreported results are denoted by "-"

Methods	5 %	10 %	20 %	25 %	30 %	40 %	50 %
NETASA (Shandar and Gromiha, 2002)	74.6	71.2	-	70.3	-	-	75.9
PP (Gianese et al., 2003)	75.7	73.4	-	71.6	-	-	76.2
NN (Garg et al., 2005)	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	-	-	-
RSAPRP (Meshkin et al., 2009)	76.82	74.84	75.35	76.7	77.75	79.86	86.32
This paper	77.13	77.01	77. 49	77.44	78.09	80.62	85.14

The two main remarks based on the performed experimental evaluation include: the proposed method obtains favorable error rates when compared with five competing methods; and the reduced number of features (i. e. 76+1 attributes instead of 13*20+1 attributes) result in a simplified prediction model, reduced computational time, and optimized prediction quality.

CONCLUSION

In this paper, an approach for predicting protein relative solvent accessibility has been presented, which relies on a two-step procedure, consisting of subset selection of evolutionary information, followed by a real-value predictor of relative solvent accessibility.

As shown in our study, feature selection is effective to reduce dimensionality, removing irrelevant features and increasing prediction accuracy in prediction of relative solvent accessibility of proteins.

We have recently proposed an approach for prediction of RSA (Meshkin et al., submitted) with scatter search technique. Results of this paper achieve more improvement in training time by smaller size of feature set rather than research (Meshkin et al., submitted).

We can conclude from this research that most of features in evolutionary information profile do not have any significant impact on prediction of RSA for a central residue in a given window. Despite of choosing subset of features, prediction accuracy has not decreased, and in some thresholds, prediction accuracy has improved in comparison with methods that their training is done in one stage.

For future works we will widen our scope to consider more feature selection and classification algorithms such as boosting, genetic algorithm, evolutionary algorithm, and neural networks, so that we can find an optimal approach to determining discriminatory features. To find common features from different feature selection methods is another interesting task.

REFERENCES

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

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;17:3389-402.

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

Dor O, Zhou Y. Achieving 80 % ten-fold cross-validated accuracy for secondary structure prediction by large scale training. Proteins 2007;66:838-45.

Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 2009;74:847-56.

Garg A, Kaur H, Raghava G. 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.

Hall MA. Correlation-based feature subset selection for machine learning. Hamilton, New Zealand. 1998.

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.

Korf RE. Linear-space best-first search. Artif Intelligence 1993;62:41-78.

Meshkin A, Sadeghi M, Ghasem-Aghaee N. Prediction of relative solvent accessibility using pace regression. EXCLI J 2009;8: 211-7.

Meshkin A, Sadeghi M, Ghafuri H. Prediction of relative solvent accessibility based on qualitative and quantitative features selection by Support Vector Regression (submitted for publication).

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

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

Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 2009;31:9-51.

Shandar A, Gromiha M. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819-24.

Shandar A, Gromiha M, Akinori S. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629-35.

Shevade S, Keerthi S, Bhattacharyya C, Murthy K. Improvements to SMO algorithm for SVM regression. National University of Singapore, 1999. (Technical Report CD-99-16).

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

Smola AJ, Scholkopf B. A tutorial on Support Vector Regression. NeuroCOLT2 Technical Report Series, 1998.

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

Witten I, Frank E. Data mining: Practical machine learning tools and techniques. San Francisco, CA: 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 2005, Shanghai, China, 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 J. Prediction of protein solvent accessibility using support vector machines. Proteins 2002:48;566-70.