EXCLI Journal 2013;12:52-63 – ISSN 1611-2156
Received: October 24, 2012, accepted: January 09, 2013, published: January 21, 2013

Original article:

SCATTER-SEARCH WITH SUPPORT VECTOR MACHINE FOR PREDICTION OF RELATIVE SOLVENT ACCESSIBILITY

Amir Hosein Kashefi[1], Alireza Meshkin[2,*], Mina Zargoosh[2], Javad Zahiri[3], Mohsen Taheri[2], Saman Ashtiani[3]

[1] Young researchers Club, South Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Department of Computer Engineering, Islamic Azad University, Damavand Branch, Damavand, Iran
[3] Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran
* corresponding author: e-mail address: meshkin@nigeb.ac.ir (Alireza Meshkin)

ABSTRACT

Proteins have vital roles in the living cells. The protein function is almost completely dependent on protein structure. The prediction of relative solvent accessibility gives helpful information for the prediction of tertiary structure of a protein. In recent years several relative solvent accessibility (RSA) prediction methods including those that generate real values and those that predict discrete states have been developed. The proposed method consists of two main steps: the first one, provided subset selection of quantitative features based on selected qualitative features and the second, dedicated to train a model with selected quantitative features for RSA prediction. The results show that the proposed method has an improvement in average prediction accuracy and training time. The proposed method can dig out all the valuable knowledge about which physicochemical features of amino acids are deemed more important in prediction of RSA without human supervision, which is of great importance for biologists and their future researches.

Keywords: Physicochemical properties of amino acids, evolutionary information, PSI-BLAST, feature selection methods, support vector regression

INTRODUCTION

Predicting the relative solvent accessibility (RSA) of a protein is an important step for determining its native structure and thus its function (Lee and Richards, 1971).

Even without proper understanding of the protein folding mechanism, useful insights on possible 3D conformations of a target protein can be obtained from its predicted solvent accessibility during the process of protein structure modeling (Joo et al., 2007; Xu et al., 2009; Wang et al., 2010) including, for example, quality assessment of protein models (for finding the near-native structures) (Benkert et al., 2009; Cheng et al., 2009), and detection as well as threading of remote homologous proteins in template based modeling (Peng and Xu, 2010) sequence profile; machine learning.

Due to the immense time and economic costs of experimental methods in determining protein structure, predicting tertiary structure of proteins have great importance. 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 and Rychlewski, 2003). Thus, the simplification of the problem, reducing 3D structure to 1D features, may be useful and regarded as the first-step in understanding the protein-folding problem. The prediction of secondary structures is the most familiar and well-defined aspect of the problem (Jones, 1999). The prediction of solvent accessibility is another aspect of the problem (Garg et al., 2005).

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 of each residue. These methods can be divided into two main groups:

- Real valued predictors that predict RSA value. 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), support vector regression (Yuan and Huang, 2004; Xu et al., 2005; Meshkin and Ghafuri, 2010), pace regression (Meshkin and Sadeghi, 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 in the form of the PSSM profile derived with PSI-BLAST as the input features (Wagner et al., 2005; Adamczak et al., 2004; Yuan 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 (Meshkin and Ghafuri, 2010; Kim and Park, 2004; Yuan et al., 2002), two stage support vector machine (Nguyen and Rajapakse, 2005), information theory (Naderi-Manesh and Sadeghi, 2001), and probability profile (Gianese et al., 2003). Early studies only used sequence to generate features (Shandar and Gromiha, 2002; Naderi-Manesh and Sadeghi, 2001), while recent studies have used the evolutionary information in the form of the PSSM profile to generate features (Kim and Park, 2004; Nguyen and Rajapakse, 2005). The PSI-BLAST profile (Altschul et al., 1997), was applied as an efficient sequence representation that improves classification accuracy (Cuff and Barton, 2000).

Since, it is believed that the 3D-structure of most proteins is defined by their sequences (Anfinsen, 1973), utilizing data mining methods to extract hidden knowledge and information from protein sequences, is unavoidable. Due to the immense time costs of data mining on large training data, the necessity of investigation in feature selection seems to be essential.

This paper investigates whether improved sequence representation, which is based on the custom selected features harvested from evolutionary information, could lead to improving the RSA predictions. We also investigate whether it would be possible to disclose all the valuable knowledge about which qualitative features are highly correlated with the solvent accessibility of proteins without human supervision. 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 the hypothesis that if we use data mining feature selection methods for selecting subset of best-performing features, then prediction accuracy would be improved. This idea results in a simplified prediction model, reduced computational time, and optimized prediction quality. The goals of this paper are achieved by designing a method that operates in two steps.

In the 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. Then in the second step, support vector regression

(SVR) is used for building an appropriate model with selected subset of evolutionary features to real value prediction of RSA.

RESULTS AND DISCUSSION

The SVR and Scatter Search 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 5 fold cross validation test type to allow for a comprehensive comparison with previous studies.

Residues were classified into two states (buried/exposed). Seven thresholds of 5, 10, 20, 25, 30, 40 and 50 % are tried in the two-state classification. 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

$$Q_{\%} = \left[\frac{N_B + N_E}{N_{total}}\right] \quad (1)$$

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.

Figure 1 shows the actual and predicted values for residues 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 residues. It shows good linear relationship between the actual and predicted values.

Comparison with other methods

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

Table 1 shows the comparison between the results of this work and one stage methods for RSA prediction, including neural network and support vector regression models (Garg et al., 2005; Adamczak et al., 2004; Shandar et al., 2003; Xu et al., 2005; Meshkin and Sadeghi, 2009; Shandar and Gromiha, 2002; Gianese et al., 2003).

Since these 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 %, see Table 4. Best values are shown in the bold face.

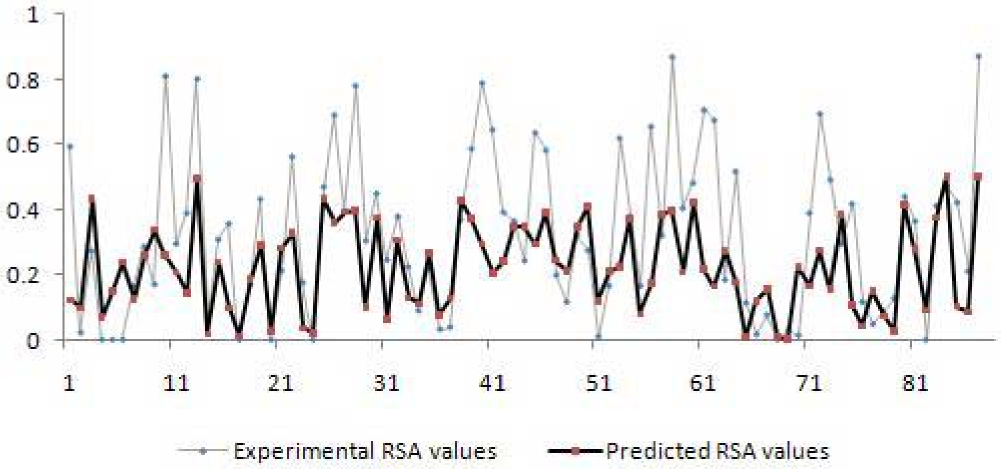


Figure 1: Comparison of actual and predicted RSA values for a thioredoxin (PDB code 1ABA)

Table 1: Comparison between our method and other reported methods; unreported results are denoted by "-".

Methods	MAE (%)	5 %	10 %	20 %	25 %	30 %	40 %	50 %
NETASA (Shandar and Gromiha, 2002)	-	74.6%	71.2%	-	70.3%	-	-	75.9%
NN (Shandar et al., 2003)	18.0	-	-	-	-	-	-	-
PP (Gilanski and Rychlewski, 2003)	-	75.7%	73.4%	-	71.6%	-	-	76.2%
NN (Garg et al., 2005)	15.2	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	-	-	-	-	-	-	-
RSAPRP (Meshkin and Sadeghi, 2009)	13.14	76.82%	74.84%	75.35%	76.7%	77.75%	79.86%	86.32%
SVM-Best (Meshkin and Ghafuri, 2010)	-	77.13%	77.01%	77.49%	77.44%	**78.09%**	**80.62%**	85.14%
SS-SVM (this paper)	**12.31**	**79.19%**	**78.20%**	77. 64%	**77.63%**	77.47%	79.71%	**86.52%**

The two main remarks based on the performed experimental evaluation are the favorable error rates obtained by proposed method when compared with seven competing methods; and the reduced number of features (i.e. 89+1 attributes instead of 13*20+1 attributes) results in a simplified prediction model that reduce computational time, and optimized prediction quality.

CONCLUSIONS

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

As shown in our experiments, many of features in evolutionary information 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 step and using all features of evolutionary information.

The advantage of this work is that we do not apply biological knowledge in selection of qualitative features that have strong relationship with protein solvent accessibility, but instead we use mathematically-based feature selection method to find out which features are highly correlated with the solvent accessibility of proteins. Interestingly, the results of feature selection method adopt with biologically concepts about the relation between physicochemical properties of amino acid and protein solvent accessibility.

MATERIALS AND METHOD

In this section, the definition of relative solvent accessibility, dataset, qualitative features and quantitative features are introduced. Then, the proposed two stage method is described.

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 calculated by dividing the accessible surface area from DSSP (Kabsch and Sander, 1983), by the maximum solvent accessibility according to Chothia's work (1976), which uses Gly-X-Gly extended tripeptides that there are shown in Table 2 in units of Å.

Dataset

The set of 215 nonhomologous protein chains with no more than 25 % pairwise-sequence identity used in the experiment of

Manesh dataset (Naderi-Manesh and Sadeghi, 2001). 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 (Garg et al., 2005; Shandar et al., 2003; Xu et al., 2005; Meshkin and Ghafuri, 2010; Meshkin and Sadeghi, 2009; Wang et al., 2004; Shandar and Gromiha, 2002; Gianese et al., 2003), and this motivated us to use it to design and validate our method.

Quantitative features

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 and each position 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 work.

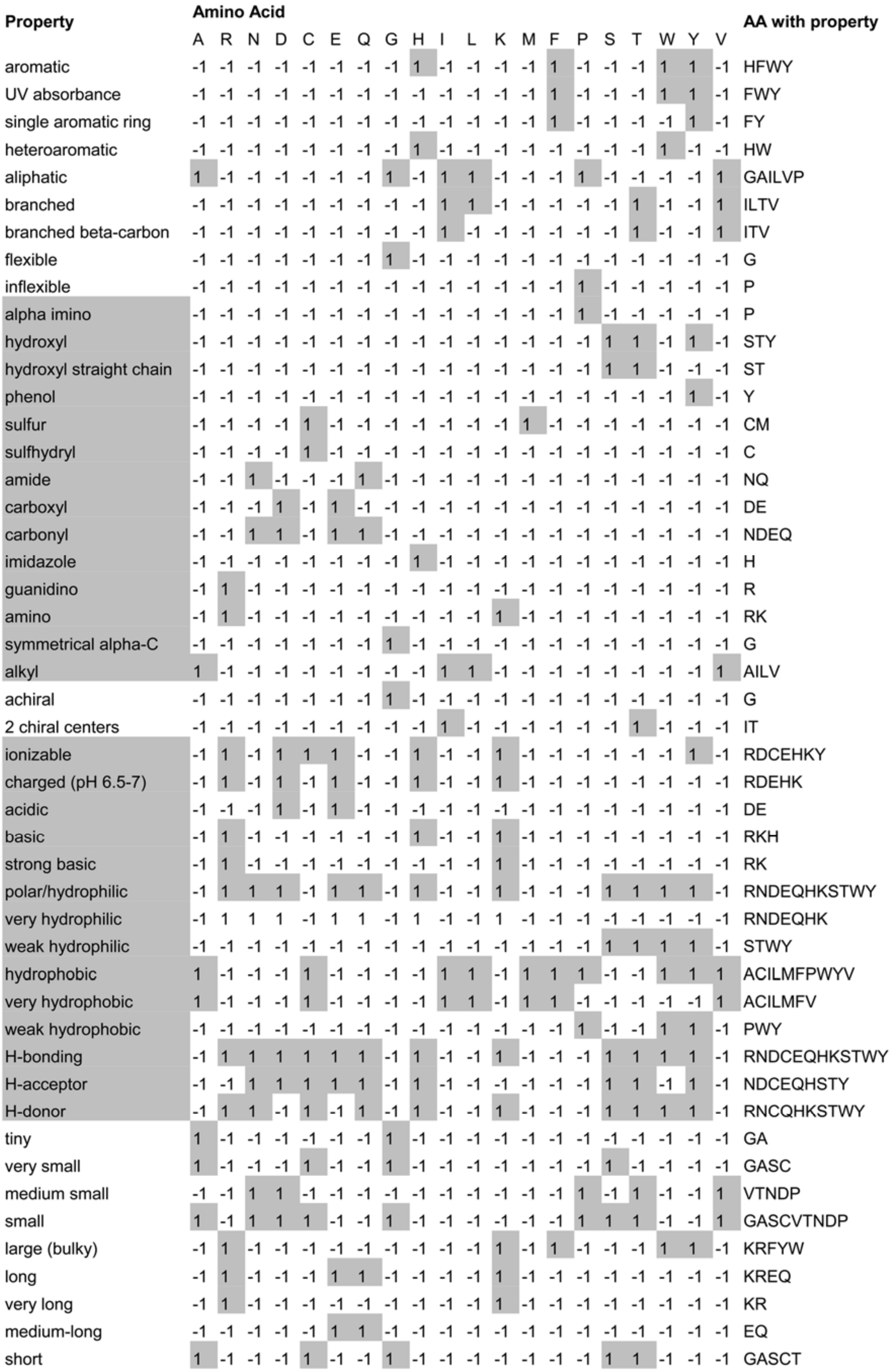
Qualitative features

We utilized 48 qualitative physicochemical features describing side chain structure and functional groups of amino acid. The complete list of these properties and also how the amino acids described based on these qualitative properties are shown in Figure 2. In order to be included as a physicochemical property, a property should be characterized (or at least well-estimated) by theoretical analysis of the amino acid structures. For example, which reactions an amino acid participates in or a comparison of the entropy of formation of the amino acids are not properties we can predict well simply by looking at the amino acid structures, but we can characterize the hydrophobicity of an amino acid by looking at properties of its side chain such as what functional groups it has (Yu, 2001).

Table 2: Maximum Surface Accessibility (Max Acc) of the AAs (Å) in Extended tripeptide Gly-X-Gly Conformation

AA	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
Max Acc	115	225	160	150	135	180	190	75	195	175
AA	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Max Acc	170	200	185	210	145	115	140	255	230	155

EXCLI Journal 2013;12:52-63 – ISSN 1611-2156
Received: October 24, 2012, accepted: January 09, 2013, published: January 21, 2013



Property	Amino Acid																				AA with property
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	
aromatic	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	HFWY
UV absorbance	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	FWY
single aromatic ring	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	FY
heteroaromatic	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	HW
aliphatic	1	-1	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	GAILVP
branched	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	1	-1	-1	1	ILTV
branched beta-carbon	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	ITV
flexible	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
inflexible	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	P
alpha imino	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	P
hydroxyl	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	STY
hydroxyl straight chain	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	ST
phenol	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	Y
sulfur	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	CM
sulfhydryl	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	C
amide	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	NQ
carboxyl	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	DE
carbonyl	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	NDEQ
imidazole	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	H
guanidino	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	R
amino	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RK
symmetrical alpha-C	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
alkyl	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	AILV
achiral	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
2 chiral centers	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	IT
ionizable	-1	1	-1	1	1	1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	RDCEHKY
charged (pH 6.5-7)	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RDEHK
acidic	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	DE
basic	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RKH
strong basic	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RK
polar/hydrophilic	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNDEQHKSTWY
very hydrophilic	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RNDEQHK
weak hydrophilic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	STWY
hydrophobic	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	1	1	1	-1	-1	1	1	1	ACILMFPWYV
very hydrophobic	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	1	ACILMFV
weak hydrophobic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	1	-1	PWY
H-bonding	-1	1	1	1	1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNDCEQHKSTWY
H-acceptor	-1	-1	1	1	1	1	1	-1	1	-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	NDCEQHSTY
H-donor	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNCQHKSTWY
tiny	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	GA
very small	1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	GASC
medium small	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	1	VTNDP
small	1	-1	1	1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	1	GASCVTNDP
large (bulky)	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	1	1	-1	KRFYW
long	-1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	KREQ
very long	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	KR
medium-long	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	EQ
short	1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	GASCT

Figure 2: Matrix of amino acid properties (Yu, 2001)

Received: October 24, 2012, accepted: January 09, 2013, published: January 21, 2013

A bipolar vector is used to represent qualitative features for a window surrounding the given amino acid. Bipolar code set to 1 if amino acids have a specific qualitative feature; otherwise it set to -1.

There are 13*48 features for a 13 residues wide window centered on a target residue in a bipolar vector with format (2). In addition, we added real value of RSA for a centered residue in the given window as a class feature.

$$(f_1, f_2, \ldots, f_{623}, f_{624}, RSA_{\text{for a given residue}} \qquad (2)$$

For instance, qualitative features for a window surrounding the given amino acid are encoded as (3).

$$(-1, +1, \ldots, -1, -1, 0.87) \qquad (3)$$

In the first step of our proposed method, qualitative features vectors were used by feature selection method to disclose which qualitative features have most significant relationship with the solvent accessibility of proteins.

Prediction method

From a logical standpoint, the proposed method can be divided to two main steps. The first one, provide subset selection of quantitative features based on selected qualitative features and the second, dedicated to build model for prediction.

Figure 3 shows a detailed overview of the prediction procedure that consists of two steps. The first is aimed for creating custom selected feature set and the second is responsible for model building.

The proposed two-stage prediction model works as follows:

The task of the first step is grouped into two subtasks: “Qualitative Feature selection” and “Quantitative Feature selection”. In “Qualitative Feature selection” subtask, we want to disclose which physicochemical properties of residues deemed more significant for prediction of RSA with respect to the position of them in a given window. So for this purpose, we create 48*13 (13 is the size of the window) features for each residue in a sequence by considering its neighboring in a given window. After that, we use a data mining feature selection method to find more important features.

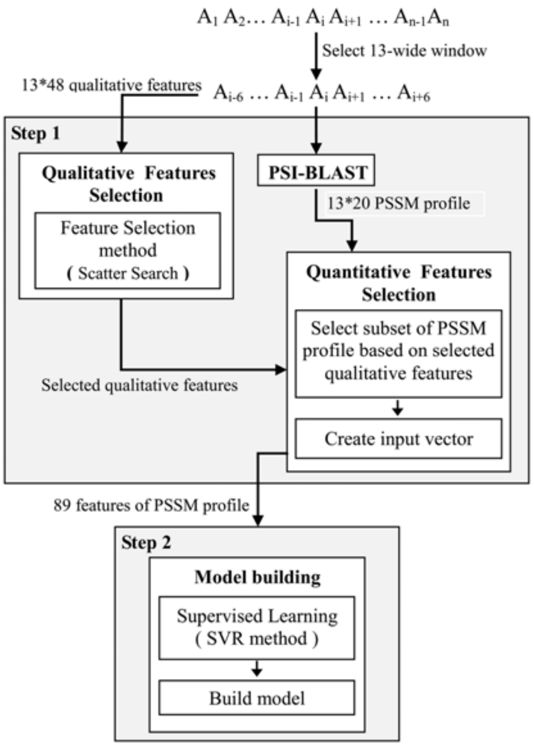

$A_1\,A_2 \ldots A_{i-1}\,A_i\,A_{i+1} \ldots A_{n-1}A_n$
Select 13-wide window
13*48 qualitative features
$A_{i-6} \ldots A_{i-1}\,A_i\,A_{i+1} \ldots A_{i+6}$
Step 1
Qualitative Features Selection
Feature Selection method (Scatter Search)
PSI-BLAST
13*20 PSSM profile
Quantitative Features Selection
Select subset of PSSM profile based on selected qualitative features
Create input vector
Selected qualitative features
89 features of PSSM profile
Step 2
Model building
Supervised Learning (SVR method)
Build model

Figure 3: A detailed overview of the proposed method

Whenever, the subset of qualitative features is found, in “Quantitative Feature selection” subtask, amino acids that have those selected qualitative properties are chosen in each position of window. Finally, we have a subset of best-performing features from PSI-BLAST profile, which are used in the next step for training the model. The sufficient details of feature selection are completely described at subsection 3.1.

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 applies in this step. The SVR is trained using sequential minimal optimization algorithm (Smola and Scholkopf, 1998), which was further optimized by Shevade and colleagues (1999).

Received: October 24, 2012, accepted: January 09, 2013, published: January 21, 2013

FEATURE SELECTION

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.

The knowledge that we reveal in the first step of our method is about which of qualitative features of amino acids have the most significant impact on the RSA prediction of the central amino acid in each position of a given window.

With regard to the too high number of PSI-BLAST profile features (in a window with size N), the main practical aim of the first step of our method is to find a smaller subset of features among a set of N*20 features which enables an efficient prediction of relative solvent accessibility of proteins.

Data mining feature selection method are used to find out which physiochemical features of amino acids are most important for predicting relative solvent accessibility. We applies the Scatter Search method (Garcia-Lopez et al., 2006), with forward direction and use CfsSubsetEval method (Hall, 1998), 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 features that are highly correlated with the RSA values while having low intercorrelation are preferred. Scatter Search is a recent meta heuristic and it has been successfully applied to solve standard problems in three central paradigms of Machine Learning including Clustering, Classification and Feature Selection.

The Scatter Search method filters the redundancy among the features and selects the final number of selected features, which in our case were 32 features. Table 3 shows selected qualitative features for the residue A_i that is located in the center of the window with size 13.

Among the 13*48 qualitative features, only 31 features deemed more significant for prediction of RSA in a given window. The first step of our method discover all the valuable knowledge about correlation between selected qualitative features and the RSA prediction of the centered amino acid, such as:

- The most physicochemical features of central amino acid i.e. A_i have significant impact on the RSA prediction. Interestingly, qualitative features of further residues have relatively small influence at the RSA prediction of the centered amino acid.

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

- Hydrophilicity, hydrophobicity, length, flexibility and inflexibility features of amino acids have strong relationship with RSA values because these features are found in many positions of the window in Table 3.

- Among the 48 physicochemical features of amino acids, only 22 distinct physicochemical features have strong relationships with protein solvent accessibility.

Whenever, the subset of qualitative features is produced, a set of amino acids that have those selected properties is chosen in each position of window. For example, in position $A_{i+1,}$ if flexibility or 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 were used for training model in the second step, see Table 4.

EXCLI Journal 2013;12:52-63 – ISSN 1611-2156
Received: October 24, 2012, accepted: January 09, 2013, published: January 21, 2013

Table 3: Summary of feature selection results in a window with size 13

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	2	2	1	1	3	12	2	1	2	2	1	1
	symmetrical_alpha-C	hydrophobic tiny	amino medium_long	H_bonding	medium_small	branched_beta_carbon inflexible alpha_imino	very_hydrophobic very_hydrophilic polar_hydrophilic single_aromatic_ring Hydrophobic Branched Charged Carbonyl very_long Sulfhydryl Alkyl Long	flexible hydrophobic	alpha_imino	inflexible very_hydrophobic	sulfur long	symmetrical_alpha-C	very_hydrophobic

Table 4. Summary of feature selection results for the PSI-BLAST

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	11	4	3	14	3	19	11	1	8	6	1	7
The selected features	G	A P C Y G V I F L M W	R Q E K	R N D	C D Q P E T H V K S W Y N T	I T V	A R V NDY CQW EHS I LT KMF P	A I C G L M F P V Y W	P	A V C P I L M F	R C Q E K M	G	A C V I L M F

The selected features include 89 features from the PSI-BLAST profile and one binary value that correspond with the first and last residues in the sequence. We add this binary feature; because the amino acids that are located at the terminus of the sequence have larger probability of being exposed to the solvent, see Table 5.

Table 5: Summary of the feature selection results

Feature set	# Features (without feature selection)	# Features (with feature selection)
PSI-BLAST profile	13*20=260	89
Terminus feature	1	1
# Total Features	261	90

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, Schaffer AA et al. 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.

Benkert P, Tosatto SCE, Schwede T. Global and local model quality estimation at casp8 using the scoring functions qmean and qmean-clust. Proteins 2009;77(Suppl 9):173–80.

Cheng J, Wang Z, Tegge AN, Eickholt J. Prediction of global and local quality of casp8 models by multicom series. Proteins 2009;77(Suppl 9):181–4.

Chothia C. The nature of accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.

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

Garcia-Lopez F, García-Torres M, Melián B, Moreno JA, Moreno-Vega J. Solving feature subset selection problem by a parallel scatter search. Eur J Operat Res 2006; 169:477-89.

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.

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.

Hall M. A correlation-based feature subset selection for machine learning. Hamilton, New Zealand, 1998.

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;92:195-202.

Joo K, Lee J, Lee S, Seo J-H, Lee SJ, Lee J. High accuracy template based modeling by global optimization. Proteins 2007;69 (Suppl 8):83–9.

Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.

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.

Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.

Meshkin A, Ghafuri H. Prediction of relative solvent accessibility by support vector regression and best-first method. EXCLI J 2010;9:29-38.

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

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.

Peng J, Xu J. Low-homology protein threading. Bioinformatics 2010;26:i294–i300.

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. Technical Report CD-99-16. Control Division Dept of Mechanical and Production Engineering, National University of Singapore, 1999.

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 Z, Eickholt J, Cheng J. Multicom: a multi-level combination approach to protein structure prediction and its assessments in casp8. Bioinformatics 2010;26:882–8.

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: Morgan Kaufmann, 2005.

Xu J, Peng J, Zhao F. Template-based and free modeling by raptor11 in casp8. Proteins 2009;77:133–7.

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 (pp 2595-8). Shanghai, China, 2005.

Yu K. Theoretical determination of amino acid substitution groups based on qualitative physicochemical properties. http://biochem218.stanford.edu/Projects%2002001/Yu.pdf

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.