J. Biomedical Science and Engineering, 2008, 1, 1-9

Scientific Research Publishing

Sequence based prediction of relative solvent accessibility using two-stage support vector regression with confidence values

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

[1]Department of Electrical and Computer Engineering, University of Alberta, T6G 2V4, Edmonton, CANADA. * Correspondence should be addressed to Lukasz Kurgan (lkurgan@ece.ualberta.ca).

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. Recent years observed development of several RSA prediction methods including those that generate real values and those that predict discrete states (buried vs. exposed). We propose a novel method for real value prediction that aims at minimizing the prediction error when compared with six existing methods. The proposed method is based on a two-stage Support Vector Regression (SVR) predictor. The improved prediction quality is a result of the developed 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. Cross validation tests on a benchmark dataset show that our method achieves 14.3 mean absolute error and 0.68 correlation. We also propose a confidence value that is associated with each predicted RSA values. The confidence is computed based on the difference in predictions from the two-stage SVR and a second two-stage Linear Regression (LR) predictor. The confidence values can be used to indicate the quality of the output RSA predictions.

Keywords: Relative solvent accessibility; Support vector regression; PSI-BLAST; PSI-PRED; Secondary protein structure

1. INTRODUCTION

The knowledge of three dimensional protein structure plays the key role in understanding protein's function. Computational prediction of the tertiary protein structure is one of the central topics in structural biology due to 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]. As a result, the predictions of the solvent accessibility [2] and the secondary structure [3] are addressed as an intermediate step towards the prediction of the 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 [4] and is vitally important for understanding the binding mechanism of proteins [5]. Chan and Dill pointed that the burial of core residues is the driving force in protein folding, which suggests that knowledge of localization of individual residues (surface vs. buried) provides useful information to reconstruct the 3D-structure of proteins [6-8].

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 valued* predictors predict RSA value (the definition is given in the Materials section). The representative existing methods are based on linear regression [9], neural network based regression [11], neural networks [12], support vector regression [10, 13, 15], and look up table [14]. In Ahmad's study, binary coding of the sequence was taken as the input features [12], while all other studies used the evolutionary information in the form of the PSSM profile derived with PSI-BLAST as the input features [9-11, 13-15].
- *discrete valued* predictors classify each residue into a predefined set classes. 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 [17], neural network [16, 20, 22], support vector machine [19, 21], two stage support vector machine [18], information theory [23], and probability profile [24]. Early studies only use sequence to generate features [20, 23], while recent studies use the evolutionary information in the form of the PSSM profile to generate features [18, 19].

The PSI-BLAST profile [25] was recently introduced as an efficient sequence representation that improves classification accuracy [16]. Subsequently, researchers have found that secondary structure predicted using the PSI-PRED method [3] improves the real value RSA predictions [2].

This paper investigates whether improved sequence representation, which is based on the information harvested from the sequence, the PSI-BLAST profile and the predicted secondary structure, could lead to improving the RSA predictions. We also investigate whether it would be possible to build an index that would indicate the quality of the predicted RSA value. The above hypotheses translate into the two following goals: (1) we aim at proposing a prediction method that minimizes the RSA prediction error; (2) the method should provide a confidence value that indicates the quality of the predicted RSA values.

The first goal is achieved by designing a custom-selected set of features, which is based on performing feature selection, to represent the input sequence. As suggested in previous studies, the PSI-BLAST profile, PSI-PRED predicted secondary structure 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 two feature selection methods to select a subset of best-performing features. This results in a simplified prediction model, reduced computational time, and optimized predictive quality.

To address the second goal, the confidence values are computed based on the difference in predictions of RSA by two predictors: a support vector regression and a linear regression. These values can be used to indicate the quality of the output RSA predictions.

2. MATERIALS

2.1. Dataset

The dataset used in this paper is referred to as the Manesh dataset [23] and consists of 215 low-similarity, i.e., < 25%, proteins. 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 [2, 12-15, 20, 24], and this motivated us to use it to design and validate our method.

2.2. Relative solvet 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 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 representation

PSI-BLAST profile. PSI-BLAST is used to compare different protein sequences to find similar sequences and to discover evolutionary relationships [25]. PSI-BLAST generates a profile representing a set of similar protein sequences in the form of a $20 \times N$ 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 15 residues wide window centered on a target residue and thus it consists of 300 features. The selected size is motivated by previous studies that adopted this window size [18] and obtained good secondary structure prediction results [3].

Secondary structure predicted with PSI-PRED. The quality of secondary structure prediction has 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 [2]. 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 the 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 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 at 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 feature selection methods were used to reduce the dimensionality. We applied the *correlation-based feature selection* (CBFS), and another feature selection method, namely correlation-based method for relevance and redundancy analysis (CBRR), which selects a subset of features based on filtering redundancy within the feature set. The CBFS method 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 $\bar{x}_i$ is the mean of X and $\bar{y}_i$ 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. 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 CBRR feature selection method considers both the relevance of the features with respect to the target (RSA values), and the redundancy between the features. It involves two steps: (1) selecting a subset of relevant features, and (2) selecting predominant features from among the relevant features. The details can be found in [26].

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 both feature selection methods. The feature selection was processed using the training set of Manesh dataset, which includes 30 sequences [14, 20].

The CBRR method automatically filters the redundancy among the features and selects the final number of selected features, which in our case was 15. The selected features include 13 features from the PSI-BLAST profile, and 2 predicted secondary structure features, see **Table 1**. In case of CBFS, the number of selected features should be specified by the user. Hence, we tested the performance of different number of selected features using support vector regression model with default parameters to predict RSA values for the test set of the Monash dataset. The mean absolute error (MAE) steadily decreases to 15.6% by adding up to 70 features, and it saturates when adding additional features, see **Figure 1**. As a result, the 70 features with the highest Pearson correlation were selected when using CBFS. 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 **Table 1**.

The two feature sets selected by CBRR and CBFS and the full feature set (313 features) were compared by predicting RSA values for the test set of the Manesh dataset using support vector regression with default parameters. The 15 features selected by CBRR obtain 16.7% MAE, while the 70 features selected by CBFS and the full feature set both result in 15.6% MAE, see **Figure 2**. The features selected by CBFS provide lower MAE than the features selected by CBRR, and they cover only 23% of the full feature set. As a result, the 70 features selected By CBFS were used to design the proposed prediction model. The selected features are summarized in **Table 2**.

The feature selection 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 the residue A_i that is located in the center of the window:

– the features to encode the two leftmost positions (A_{i-7}, A_{i-6}) and the rightmost position (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. 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 the largest number of the selected features. Interestingly,

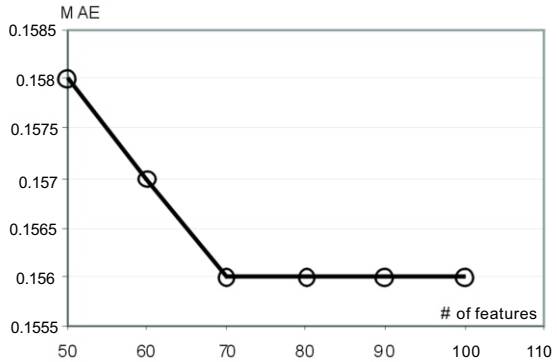

Figure 1. The MAE values against the number of selected features. The MAE is obtained by using support vector regression with default parameters to predict test set of the Monash dataset.

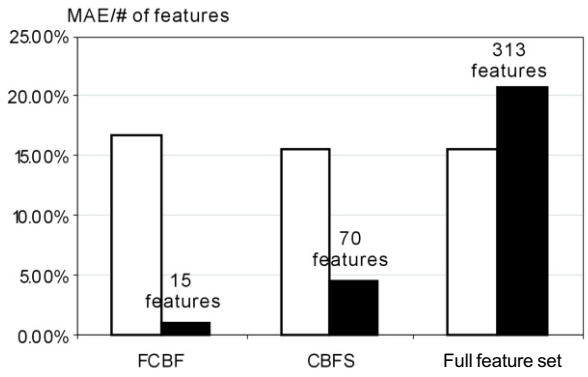

Figure 2. Bar chart of MAE values (white) and number of features (gray) for features selected by CBRR, CBFS, and the full feature set.

Table 1. Summary of the feature selection results.

Features set	Total # features	# selected features by CBFS	# selected features by CBRR
PSI-BLAST profile	300	65	13
Binary code	10	2	0
Predicted second. structure	3	3	2
Total	313	70	15

residues at i-2 and i+2 positions have relatively small influence on the prediction.

– The selected features are almost symmetrically distributed around A_i, e.g., amino acids E, K, Q, R, and D have similar impact on the solvent accessibility of the central residue at the third left position (A_{i-3}) and the third right position (A_{i+3}).

– Hydrophilic residues, which include E, K, Q, R, and D, may have impact on the solvent accessibility of A_i residue which is 3 or 4 positions away from the these residues. This pattern covers 19 of the selected features and we hypothesize that this is related to the α-helical structures due to the following two reasons. Firstly, these 5 hydrophilic residues have larger probability (above 0.5) to form helical structure than strand and coil structures [27]. Secondly, α-helix consists of 3.6 residues per turn, and hence if two residues in a helix are separated by 2 or 3 residues in the sequence then they are spatially close to each other, which in turn may induce some interactions between them. For instance, the hydrogen bond that maintains the helical structure occurs between two residues that are separated in a sequence by three other residues, i.e., A_i and A_{i+4}.

3. METHODS

3.1. Prediction method

Linear Regression (LR) and Support Vector Regression (SVR) were already applied in the RSA prediction [10,13,15]. In this paper, we propose an improved two-stage model, which not only aims at reducing the prediction error, but we also propose and test a confidence value that is associated with each predicted RSA value.

The proposed two-stage prediction model works as follows:

STAGE 1. The input sequences is inputted into PSIPRED to compute predicted secondary structure and into PSI-BLAST to compute the PSI-BLAST profile. Next, the input sequence, the predicted secondary structure, and the PSI-BLAST profile are used to compute the selected 70 features using a 15 residues wide window centered over the being predicted residue, and for each residue in the input sequence. The 70 features are used as an input to the LR model and SVR model that predict a real value (predicted RSA value) for the central residue in a given window.

STAGE 2. The aim of the stage two is to refine the predictions from stage one. Similarly to other two-stage designs [13,18], the second stage "smoothes" the predictions. It takes the three predicted secondary structure features (computed in stage one by PSIPRED) and a 7 residues wide window from the first stage predictions centered over the predicted residue as the input to provide the refined real value predictions.

Since the prediction quality of SVR is better than the quality of LR (results are discussed in the following), the predictions from SVR are taken as the final prediction outcome. The LR results serve as a reference to evaluate quality of SVR predictions. This means that if predictions from SVR and LR are similar then SVR predictions are assumed to be of high quality. On the other hand, if the two predictions are different then the SVR prediction is assumed to be of lower quality. The corresponding confidence value is defined as

$$C = 1 - |R_i - T_i| \qquad (3)$$

where R_i is the predicted RSA from SVR, and T_i is the predicted RSA from LR. A detailed overview of the prediction procedure is shown in **Figure 3**.

The optimization of the prediction, through adjustment of internal parameters of the predictors and selection of the window size for the second stage, was performed by dividing the Manesh dataset into

Table 2. Summary of feature selection results for the PSI-BLAST profile by correlation-based feature selection method.

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
The selected features			I	E	E		E	C D	E	P	E	E	I	I	
			L	K	K		K	E F	K		K	K	L	L	
				Q	Q		Q	G H	Q		Q	Q	V	V	
				R	R		R	I K	H		R	R	F		
					D		D	L M	D		D	D			
							N	N P	N		P	P			
							P	Q R	G						
							S	S T							
								V W							
								Y							

Table 3. Optimization of parameters for two-stage SVR.

First stage			Second stage		
Parameter C	Parameter γ	MAE	Parameter C	Parameter γ	MAE
1	0.001	0.157	1	0.01	0.150
1	0.005	0.153	1	0.08	0.149
1	0.01	0.151	1	0.15	0.148
1	0.02	0.151	1	0.2	0.148
1	0.03	0.152	1	0.3	0.148
1	0.05	0.155	1	0.4	0.149
0.5	0.01	0.152	0.5	0.15	0.148
0.8	0.01	0.151	0.8	0.15	0.148
1	0.01	0.151	1	0.15	0.148
2	0.01	0.151	2	0.15	0.148
3	0.01	0.151	3	0.15	0.148
5	0.01	0.152	5	0.15	0.148

two subsets, one used to compute the prediction model and the other to perform test. Similarly to [14], 30 sequences were used for training and the remaining 185 as the test set. The linear regression is parameterless and thus it does not require optimization. For SVR, 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. These parameters, which were based on experiments summarized in **Table 3**, provide the lowest MAE. We note that the adjustment of C has little impact in the quality of predictions. The MAE of the final prediction for the second stage windows sizes of 5, 7, 9, 11, 15, and 21 equal 0.149, 0.148, 0.148, 0.148, 0.148, and 0.148, respectively. This shows that the window size of 7 is the best choice to provide accurate predictions.

3.2. Linear regreesion

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:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \dots & \dots & \dots & \dots \\ 1 & x_{11} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix} \quad (4)$$

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:

$$y = X.\beta + \varepsilon \quad (5)$$

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

$$\beta = (X^T X)^{-1} X^T \vec{y} \quad (6)$$
$$\vec{\varepsilon} = \vec{y} - X.\beta$$

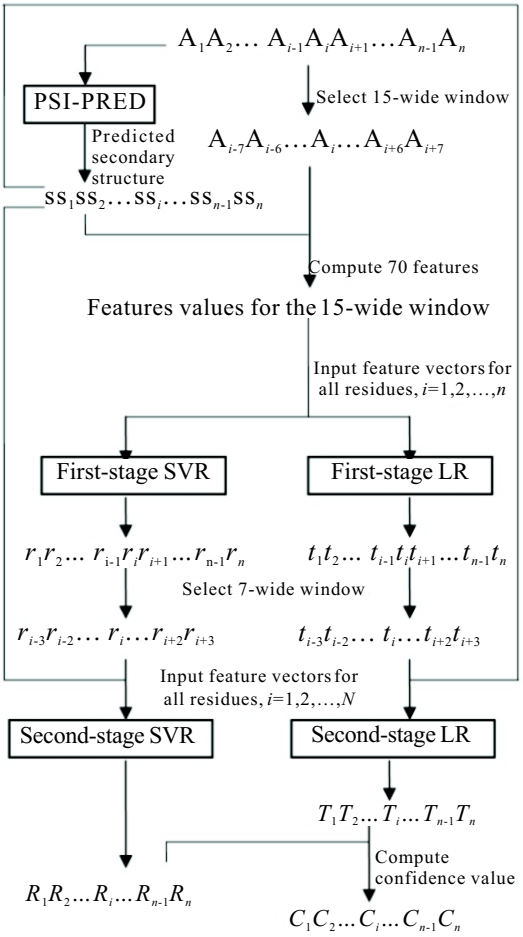

Figure 3. RSA prediction with the proposed system; the RSA value for the i^{th} residue is predicted based on the 70 feature values (see **Table 1**) that are computed over a 15 residues wide window centered on i^{th} residue; the feature values are inputted into the first-stage predictor (LR and SVR); 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 predictor that provides the RSA values. Finally, compare the predictions from SVR and LR, and calculate the confidence value C.

3.3. Support vector regression

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

$$\min \frac{1}{2}\|w\|^2 + C\sum_i (\xi_i + \xi_i^*) \quad (7)$$

such that

$$\begin{aligned} & y_i - w\cdot x_i - b \le \varepsilon + \xi_i \\ & w\cdot x_i + b - y_i \le \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \ge 0 \end{aligned} \quad (8)$$

where w is a vector perpendicular to wx-b=0 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 $k(x,x')=\phi(x)\cdot\phi(x')$ is a user defined kernel function.

The SVR was trained using sequential minimal optimization algorithm [28] that was further optimized by Shevade and colleagues [29]. The proposed SVR uses RBF kernel

$$k(x_i,x_i')=e^{-\gamma\|x-x'\|^2} \quad (9)$$

for both stages.

4. RESULTS AND DISCUSSION

The SVR and LR predictors were implemented in Weka [30], which is a comprehensive open-source library of machine learning methods. The Manesh dataset consists of 50682 instances (individual residues). The evaluation was performed using two test types to allow for a comprehensive comparison with previous studies. To compare with [2] and [12], 5-folds cross validation was executed. On the other hand, following several other prior studies [14, 20, 24], 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, were reported in **Tables 4 and 5**. In total, the proposed method was compared with six real value RSA prediction methods [2, 12-15, 24] and one method that aims at prediction of discrete states [20].

We note that in statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, sub-sampling (such as 5-fold and 7-fold) test, and jackknife test [31]. However, as elucidated by [32] and demonstrated in [33], among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used by investigators to examine the accuracy of various predictors [34-42].

4.1. Comparison with competing prediction methods

For the 5 folds cross-validation test, the mean absolute error (MAE) value of the first stage of the proposed method equals 14.6 and the corresponding Pearson's correlation coefficient (r) equals 0.67. After the second stage, the MAE value is reduced to 14.3 and r is improved to 0.68. **Table 4** compares the proposed two-stage SVR with recent methods for RSA prediction, which include neural network and support vector regression models [2, 12, 13, 15]. The proposed method obtains 0.6 to 3.7 lower MAE when compared with the abovementioned methods. This translates into 4% to 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 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 all thresholds, our method provides the highest accuracy, see **Table 4**. The proposed two-stage model provides 0.3%-0.6% higher accuracies than the prediction coming from the first stage for various thresholds. When compared to the best performing, existing two-stage SVR method [13], our predictions are characterized by lower MAE and more accurate two states predictions.

For the independent test, the MAE value for the first stage of the proposed method equals 15.0 and the corresponding Pearson's correlation coefficient r equals 0.66. After the second stage, the MAE value is reduced to 14.8 and r is improved to 0.67. **Table 5** compares the proposed two-stage SVR with recent methods for RSA prediction, which include neural network and look-up table based methods [14, 20, 24]. The proposed method obtains 1.5 to 4.0 lower MAE when compared with the above three methods. This translates into 9% to 21% error reduction, respectively. Similarly to the 5-folds cross validation test, we also examined the performance of our method by converting the real value prediction into the two states prediction. The threshold's value was adjusted between 5 and 50%.

For all thresholds our method consistently provides the highest accuracy, see **Table 5**. The two-

Table 4. Experimental comparison between the proposed two-stage SVR and other reported methods; the results were reported based on 3 or 5-folds cross validation test; the real valued predictions were 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%
[2]	Neural Network	15.2	0.67	74.9%	77.2%	77.7%	77.8%	78.1%	80.5%
[11]	Neural Network	18.0	0.50	-	-	-	-	-	-
[12]	Two-stage SVR	14.9	0.68	81.1%	78.5%	77.6%	-	-	79.5%
[14]	SVR	16.3	0.58	-	-	-	-	-	-
This paper	One-stage SVR	14.6	0.67	80.5%	79.1%	78.3%	78.3%	78.3%	80.5%
This paper	Two-stage SVR	14.3	0.68	81.1%	79.7%	78.8%	78.6%	78.8%	80.8%

Table 5. Experimental comparison between the proposed two-stage SVR 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 valued predictions were 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%
[13]	Look-up table	18.8	0.48	-	-	-	-	-	-
[19]	Neural Network	-	-	74.6%	71.2%	-	-	-	75.9%
[23]	Neural Network	16.3	0.58	75.7%	73.4%	-	-	-	76.2%
This paper	One-stage SVR	15.0	0.66	79.8%	78.7%	77.7%	77.7%	77.5%	79.8%
This paper	Two-stage SVR	14.8	0.67	80.3%	79.2%	78.1%	78.0%	78.0%	80.2%

stage model provides 0.3%-0.5% higher accuracies than the one-stage model for various thresholds. When compared with the best-performing, competing method based on neural network [24], our predictions result in higher accuracies over all thresholds, i.e., the differences range between 4% and 5.8%, and better MAE and correlation coefficient value.

The three main observations based on the performed empirical evaluation include: (1) the proposed two-state predictor obtains favorable (lower) error rates when compared with six competing methods; (2) the improvements are obtained for both real value and two-state predictions; and (3) the introduction of the second stage in our design allows for obtaining improved predictions when compared with a one stage design.

4.2. Confidence value for RSA prediction

As one of the goals of this work, we defined confidence values to measure the quality of the predicted RSA. The confidence values are based on the difference of predictions made by the two-stage SVR and the two-stage LR. The following discussion is based on results of five folds cross-validation tests.

The MAE for two-stage SVR is 0.143 and for two-stage LR is 0.155. The difference between the predictions from SVR and LR for the same residues ranges between 0 and 0.294. As a result, the confidence value C distributed in the interval [0.706, 1] for the Manesh dataset. Higher C values indicate that the predictions from SVR and LR are more consistent, and therefore the corresponding predictions from the two-stage SVR are assumed to be more accurate.

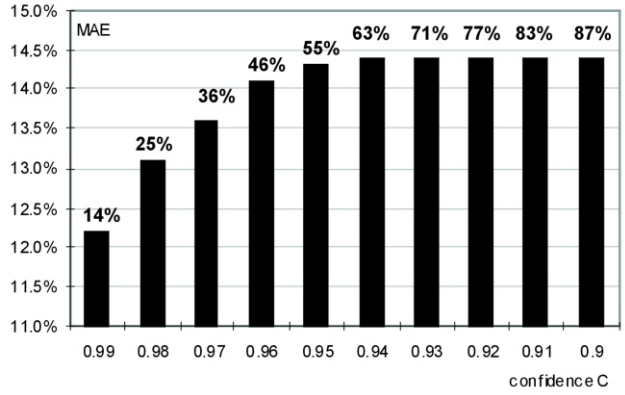

Figure 4. Bar chart of MAE values for the corresponding thresholds of confidence value C. The numbers above the bar show the corresponding coverage, i.e., number of residues for which the predictions had confidence value above the threshold. For example, for residues predicted with which $C > 0.99$ the MAE equals 12.2, and these residues cover 14% of the dataset.

The C value of 7101 samples, which covers 7101/50682= 14% of the dataset, are greater than 0.99, and the corresponding MAE of these samples equals 0.122, see **Figure 4**. The C value of 12846 samples, which covers 12846/50682= 25.3% of the dataset, are greater than 0.98, and the corresponding MAE of these samples equals 0.131. The C value of 18174 samples, which covers 18174/50682= 35.9% of the dataset, are greater than 0.97, and the MAE of these samples is 0.136. When the threshold for C value is set equal or lower than 0.96, the MAE saturates at 0.143, see **Figure 4**, which is equal to the MAE for the entire dataset (without using the confidence values). This shows that the confidence values can be used to identify a subset of the predictions which on average have better quality than the remaining predictions. This way, the user could select a desired fraction of best performing predictions. Additionally, the user could inspect quality of prediction for specific amino acids or groupings of amino acids that share certain properties such as hydrophobicity, charge, size, etc.

5. CONCLUSIONS

This paper proposes a novel method for the real value RSA prediction. The proposed method addresses two goals, which include improving the quality of RSA prediction, and development of a confidence value that allows for selection of better performing RSA predictions.

Empirical tests with the Manesh dataset show that the proposed method is characterized by lower prediction error when compared with six competing real value RSA prediction methods. We also show that the PSI-BLAST profile that is commonly used to represent sequences can by largely reduced by using feature selection, which results a simpler, interpretable model and in reduction of the computational time required to develop the prediction model. Our model indicates that window size of 13 is sufficient and only about 22% of the PSI-BLAST features are useful for

the RSA prediction. The selected features are symmetrically distributed around the predicted residue and include hydrophilic resides when considering the distance of 3 or 4 positions from the predicted residue. The confidence value *C* allows the user to select a subset of the predictions which on average are characterized by better quality than the remaining predictions.

The knowledge of the surface residues, which are predicted by the proposed method and which are directly involved in the interaction with other biological molecules, was used, for instance, for identifying protein function and stability [43, 44], for prediction of binding sites [4], understanding the binding mechanism of proteins [5], reconstruction of the 3D-structure of proteins [6-8], and to aid fold recognition [45, 46]. Therefore, improved prediction of the surface residues would have impact on improving quality of solutions for these associated tasks.

ACKNOWLEDGMENTS

This work was supported in part by NSERC Canada. K.C. also acknowledges support provided through scholarship sponsored by Alberta Ingenuity Fund.

REFERENCE

[1] 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(Suppl. 6):410-417.

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

[3] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999, 292(2):195-202.

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

[5] Chou, K. C. Review: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophysical Chemistry* 1988, 30: 3-48

[6] Chan, H. S. & Dill, K. A. Origins of structures in globular proteins. *Proc Natl Acad Sci USA* 1990, 87: 6388-92.

[7] Wang, J. Y., Lee, H. M. & Ahmad, S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins* 2005, 61(3):481-91.

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

[9] 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.

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

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

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

[13] Nguyen, M. N.& Rajapakse, J. C. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins* 2006, 63(3):542-50.

[14] Wang, J. Y., Ahmad, S., Gromiha, M. M. & Sarai, A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. *Biopolymers* 2004, 75(3):209-16.

[15] Xu, W. L., Li, A., Wang, X., Jiang, Z. H. & Feng, H. Q. Improving Prediction of Residue Solvent Accessibility with SVR and Multiple Sequence Alignment Profile. *Proceedings of the 27th IEEE Annual Conference on Engineering in Medicine and Biology,* Shanghai, China, 2005.

[16] Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000, 40(3):502-11.

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

[18] Nguyen, M. N. & Rajapakse, J. C. Prediction of protein relative solvent accessibility with a two-stage SVM approach. *Proteins* 2005, 59(1):30-7.

[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(3):557-62.

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

[21] Yuan, Z., Burrage, K. & Mattick, J. S. Prediction of protein solvent accessibility using support vector machines. *Proteins* 2002, 48(3):566-70.

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

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

[24] Gianese, G., Bossa, F. & Pascarella, S. Improvement in prediction of solvent accessibility by probability profiles. *Protein Eng.* 2003, 16(12):987-92.

[25] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 1997, 17:3389-402.

[26] Yu, L. & Liu, H. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research.* 2004, 5:1205-24.

[27] Chen, K., Kurgan, L. & Ruan, J. Optimization of the Sliding Window Size for Protein Structure Prediction, *IEEE Symposium on Comp Intelligence in Bioinformatics and Computational Biology,* 2006, 366-72.

[28] Smola, A. J. & Scholkopf, Bernhard. *A Tutorial on Support Vector Regression.* NeuroCOLT2 Technical Report Series, 1998.

[29] Shevade, S. K., Keerthi, S. 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.

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

[31] Chou, K. C. & Zhang, C. T. Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 1995, 30:275-349.

[32] Chou, K. C. & Shen, H. B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nature Protocols* 2008, 3:153-162.

[33] Chou, K. C. & Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry* 2007, 370:1-16.

[34] Diao, Y., Ma, D., Wen, Z., Yin, J., Xiang, J. & Li, M. Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. *Amino Acids* 2008, 34:111-117.

[35] Tan, F., Feng, X., Fang, Z., Li, M., Guo, Y. & Jiang, L. Prediction of mitochondrial proteins based on genetic algorithm partial least squares and support vector machine. *Amino Acids* 2007, 33:669-675.

[36] Li, F. M. & Li, Q. Z. Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. *Amino Acids* 2008, 34:119-125.

[37] Fang, Y., Guo, Y., Feng, Y. & Li, M. Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 2008, 34:103-109.

[38] Zhang, S. W., Zhang, Y. L., Yang, H. F., Zhao, C. H. & Pan, Q. Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00010-00729.

[39] Shi, J. Y., Zhang, S. W., Pan, Q. & Zhou, G. P. Using Pseudo Amino Acid Composition to Predict Protein Subcellular Location: Approached with Amino Acid Composition Distribution. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00623-z.

[40] Zhou, X. B., Chen, C., Li, Z. C. & Zou, X. Y. Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. *Amino Acids* 2007, DOI 10.1007/s00726-00007-00608-y.

[41] Nanni, L. & Lumini, A. Combing Ontologies and Dipeptide composition for predicting DNA-binding proteins. *Amino Acids* 2008, DOI 10.1007/s00726-00007-00018-00721.

[42] Nanni, L. & Lumini, A. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* 2008, DOI 10.1007/s00726-00007-00016-00723.

[43] Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* 1986, 319:199-203.

[44] Gromiha, M. M., Motohisa, O., Hidetoshi, K., Hatsuho, U. & Akinori, S. Role of structural and sequence information in the prediction of protein stability changes, comparison between buried and partially buried mutations. *Protein Engineering* 1999, 12(7):549-555.

[45] Cheng, J. & Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics* 2006, 22(12):1456-63.

[46] Liu, S., Zhang, C., Liang, S. & Zhou, Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins* 2007, 68:636-645.

