Fan *et al. BMC Bioinformatics* 2015, **17**(Suppl 1):8
DOI 10.1186/s12859-015-0851-2

BMC Bioinformatics

PROCEEDINGS Open Access

PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility

Chao Fan[1], Diwei Liu[1], Rui Huang[1], Zhigang Chen[1] and Lei Deng[1,2*]

From The Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016)
San Francisco, CA, USA. 11 - 13 January 2016

Abstract

Background: Protein solvent accessibility prediction is a pivotal intermediate step towards modeling protein tertiary structures directly from one-dimensional sequences. It also plays an important part in identifying protein folds and domains. Although some methods have been presented to the protein solvent accessibility prediction in recent years, the performance is far from satisfactory. In this work, we propose PredRSA, a computational method that can accurately predict relative solvent accessible surface area (RSA) of residues by exploring various local and global sequence features which have been observed to be associated with solvent accessibility. Based on these features, a novel and efficient approach, Gradient Boosted Regression Trees (GBRT), is first adopted to predict RSA.

Results: Experimental results obtained from 5-fold cross-validation based on the Manesh-215 dataset show that the mean absolute error (MAE) and the Pearson correlation coefficient (PCC) of PredRSA are 9.0 % and 0.75, respectively, which are better than that of the existing methods. Moreover, we evaluate the performance of PredRSA using an independent test set of 68 proteins. Compared with the state-of-the-art approaches (SPINE-X and ASAquick), PredRSA achieves a significant improvement on the prediction quality.

Conclusions: Our experimental results show that the Gradient Boosted Regression Trees algorithm and the novel feature combination are quite effective in relative solvent accessibility prediction. The proposed PredRSA method could be useful in assisting the prediction of protein structures by applying the predicted RSA as useful restraints.

Keywords: Solvent accessibility, Sequence features, Gradient boosted regression trees

Background

Since the concept of solvent accessibility was first introduced by Lee and Richards [1], defined as the surface area of a protein that is accessible to a spherical solvent while probing the surface of that molecule, it has been considered as a key factor for understanding protein structure and function [2]. Predicting the three-dimensional (3D) structures of proteins from their one-dimensional sequences is a challenging issue because of the increasing gap between the enormous number of protein sequences and the number of known structures. Studies of solvent accessibility in proteins have provided many useful insights into the 3D structures of proteins [3]. Furthermore, knowledge of solvent accessibility has proved useful for structural domains identification [4], fold recognition [5], binding region identification [6–8] and protein intrinsic disorder [9]. The solvent accessibility is particularly important because it is associated with the spatial arrangement and packing of amino acids during the process of protein folding. It also plays an important role in predicting the active sites of protein-protein or protein-ligand binding [10].

*Correspondence: leideng@csu.edu.cn
[1]School of Software, Central South University, No.22 Shaoshan South Road, 410075 Changsha, China
[2]Shanghai Key Laboratory of Intelligent Information Processing, No.220 Handan Road, 200433 Shanghai, China

In many earlier studies, the solvent accessibility prediction was taken as a classification problem with varying thresholds, two-state (exposed or buried) or three-state (exposed, intermediate or buried) [11–15]. However, there is no standard definition for the thresholds of solvent accessibility states. For instance, a residue may be predicted to be exposed state based on a relative solvent accessibility threshold of 10 %, but the same residue may be predicted to be buried state based on a threshold of 20 %. In view of this, it is necessary to predict the real values of solvent accessibility. Some representative machine learning techniques have been proposed to predict the real values of solvent accessibility, including multiple linear regression [16], support vector regression [17–19], neural network [20, 21], energy optimization [22] and nearest neighbor method [23].

For the real-valued solvent accessibility prediction, Ahmad et al. [20] proposed a neural network method with only single sequence information as the input features. The result showed that this method achieved a MAE of 18.0–19.5 % on different data sets. Adamczak et al. [21] employed evolutionary information in the form of position-specific scoring matrix (PSSM) profile to train a neural network-based regression for the prediction. Compared with the single sequence based neural network [20], the prediction performance was improved and the MAE decreased by about 5 % on the PFAM database [24]. Subsequently, Lee et al. [16] applied PSSM profile by constructing a correlation matrix different window positions to train a multiple linear regression method. The result showed a performance of 16.6 % MAE and 0.63 PCC on the Barton-502 dataset. Garg et al. [25] took multiple sequence alignment and secondary structure as input features to predict RSA based on a feed-forward neural network. The result indicated that a lower MAE achieved on CASP6 was 15.9 % and a higher PCC was 0.68.

Although these methods for surface accessibility prediction were developed, several issues still exist and make surface accessibility prediction a very challenging task. Mainly, there are three reasons: (1) specific biological properties for precisely predicting surface accessibility are not fully exploited, and no single parameter can definitely estimate the accessible surface area, various combinations of different feature types, including PSSM profiles, secondary structure features, native disorder features as well as other global sequence features [26], need to be investigated comprehensively; (2) the performance of the existing methods is still unsatisfactory, especially in terms of independent testing and (3) high-performance ensemble learning algorithms such as boosted regression trees haven't been intensively used in this area.

In this article, we propose a new and efficient approach, PredRSA (Prediction of Relative Solvent Accessible surface area), that integrates gradient boosted regression trees (GBRT) algorithm with multiple sequence-based features (position-specific scoring matrix, secondary structure, conservation score, native disorder) and a global feature (side-chain environment) to predict RSA. We have benchmarked PredRSA using the Manesh training dataset and an independent dataset. Results show that PredRSA significantly outperforms the state-of-the-art methods and indicate that the GBRT algorithm and the novel feature combination are important determinants in the prediction of RSA.

Methods

The GBRT algorithm

Our approach utilizes an ensemble regression algorithm for predicting RSA values of amino acid residues in a protein sequence. Generally, a target residue in sequence can be described as an n-dimension vector. Let us denote an amino acid residue by $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ where $x_i \in \boldsymbol{R}$ and the corresponding real-valued RSA by y. The goal of predicting RSA real value of the amino acid residue in sequence is to find a function $F^*(\boldsymbol{x})$ that maps $\boldsymbol{x}$ to y, such that over the joint distribution of all $(y, \boldsymbol{x})$-values, the expected value of some specified loss function $\Psi(y, F(\boldsymbol{x}))$ is minimized as follows:

$$\begin{aligned} F^*(\boldsymbol{x}) &= \underset{F(\boldsymbol{x})}{arg\ min}\ E_{y,\boldsymbol{x}}\Psi(y, F(\boldsymbol{x})) \\ &= \underset{F(\boldsymbol{x})}{arg\ min}\ E_{\boldsymbol{x}}[E_y(\Psi(y, F(\boldsymbol{x}))|\boldsymbol{x}] \end{aligned} \tag{1}$$

Let $\{y_i, x_i\}_1^N$ be a set of training data, N is the number of all amino acid residues in the training set. The GBRT algorithm iteratively constructs M different weak learners $h(\boldsymbol{x}, \Theta_1), \cdots, h(\boldsymbol{x}, \Theta_M)$ which consist of regression trees of fixed size from training set and constructs the following additive function $F(\boldsymbol{x})$:

$$F(\boldsymbol{x}) = \beta_0 + \sum_{m=1}^{M} h(\boldsymbol{x}, \Theta_m) \tag{2}$$

where β_m and Θ_m are a weight and vector of parameters for the mth weak regression tree $h(\boldsymbol{x}, \Theta_m)$, respectively, and β_0 is an initial value. Both the weight β_m and the parameters Θ_m are iteratively determined from $m = 1$ to $m = M$ so that a loss function $\Psi(y, F(\boldsymbol{x}))$ is minimized. That is, β_m and Θ_m for the mth regression tree are determined as follows:

$$(\beta_m, \Theta_m) = \underset{\beta,\Theta}{arg\ min} \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(\boldsymbol{x}_i) + \beta h(\boldsymbol{x}_i, \Theta)) \tag{3}$$

where $F_0(\boldsymbol{x})$ is an initial value and given by $F_0(\boldsymbol{x}) = \beta_0 = arg\ min_\beta \sum_{i=1}^{N} \Psi(y_i, \beta)$, $F_{m-1}(\boldsymbol{x})$ is the $(m-1)$th additive function combined from the first to the $(m-1)$th weak regression tree.

However, in general, it is not straightforward to solve Eq. (3). Therefore, GBRT separately and approximately estimates (β_m, Θ_m) with a simple two-step fashion [27]. For the estimation of the parameters Θ_m, we determine them so that the function defined by the regression tree approximates a gradient with respect to the current function $F_{m-1}(\boldsymbol{x})$ in the sense of least-square error as follows:

$$\Theta_m = \underset{\Theta}{arg\ min} \sum_{i=1}^{N} (\tilde{y}_{im} - h(\boldsymbol{x}_i, \Theta))^2 \quad (4)$$

where $\tilde{y}_{im}$ is the gradient and is given by

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi(y_i, F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)}\right]_{F(\boldsymbol{x})=F_{m-1}(\boldsymbol{x})} \quad (5)$$

When the mth regression tree using the Θ_m has L_m leaf nodes, the regression tree is given by

$$h\Big(\boldsymbol{x}, \{R_{lm}\}_{l=1}^{L_m}\Big) = \sum_{l=1}^{L_m} \bar{y}_{lm} I(\boldsymbol{x} \in R_{lm}) \quad (6)$$

where R_{lm} is a disjoint region that the lth leaf node of the mth regression tree defines. $I(.)$ is a Boolean function that outputs 1 in case the argument of the function is true. $\bar{y}_{lm}$ is a constant for the R_{lm}th region, defined as the mean of training data that belongs to the lth leaf node of the mth regression tree. The weight β_m can be straightforwardly chosen using line search:

$$\beta_m = \underset{\beta}{arg\ min} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}(\boldsymbol{x}_i) - \beta \frac{\partial \Psi\left(y_{i,F_{m-1}(\boldsymbol{x}_i)}\right)}{\partial F_{m-1}(\boldsymbol{x}_i)}\right) \quad (7)$$

Then, a new additive function $F_m(\boldsymbol{x})$ is updated as follows:

$$F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \nu \sum_{l=1}^{L_m} \beta_m \bar{y}_{lm} I(\boldsymbol{x} \in R_{lm}) \quad (8)$$

where $0 < \nu < 1$ is a shrinkage parameter, also called the learning rate to scale the step length the the gradient descent procedure. In this work, we take Huber loss function [28] as the loss function given by

$$\Psi(y, F) = \begin{cases} \frac{1}{2}(y - F)^2 & if\ |y - F| \leqslant \delta \\ \delta(|y - F| - \delta/2) & if\ |y - F| > \delta \end{cases} \quad (9)$$

Hence, in Eq. (5), $\tilde{y}_{im}$ becomes:

$$\tilde{y}_{im} = \begin{cases} y_i - F_{m-1}(\boldsymbol{x}_i) & if\ |y_i - F_{m-1}(\boldsymbol{x}_i)| \leqslant \delta \\ \delta.sign(y_i - F_{m-1}(\boldsymbol{x}_i)) & if\ |y_i - F_{m-1}(\boldsymbol{x}_i)| > \delta \end{cases} \quad (10)$$

The value of the transition point δ depends on the iteration number m.

Finally, the resulting RSA value y corresponding to the amino acid residue $\boldsymbol{x}$ is given by: $y = F_M(\boldsymbol{x})$.

Sequence encoding schemes

Selecting appropriate features is a crucial step because it directly determines the prediction performance. In this article, we explore various sequenced-based features which have been shown to be related to the solvent accessibility or ever applied in the similar issues. These features include PSSM profiles [29–31], PSIPRED-predicted secondary structure [32], DISOPRED-predicted native disorder [33], conservation score and side-chain environment compositions [34]. In this section, a more detail description about how to extract and encode these different sequence-based features as follows.

PSI-BLAST-based profiles

Position-specific scoring matrix (PSSM) of a residue which is achieved by the PSI-BLAST program contains important evolutionary information that determines whether this residue is conserved in its family of related proteins. Each element in the PSSM represents the probability of each residue position in the multiple sequence alignment. Plenty of previous studies have shown that multiple sequence alignments in the form of PSSM can substantially improve overall prediction performance [35–38]. In this article, the PSSM profile for each protein sequence is generated with default parameters (3 iterations and 0.001 of E-value cutoff) against the non-redundant (nr) dataset obtained from the NCBI. We encode each residue using a local sliding window approach based on the PSSM profiles. The PSSM profile generated by PSI-BLAST consists of the likelihood of a particular residue substitution at a specific position. These likelihood values are normalized to [0,1] by standard logistic function:

$$x' = \frac{1}{1 + exp\,(-x)} \quad (11)$$

where x is the score derived from the PSSM profile and x' is the standardized value of x. For a given residue, its local sequence fragment is extracted and encoded as a $20 \times (2l + 1)$-dimensional vector by using a sliding window scheme where l denotes the half window size and $L = 2l + 1$ is the whole window length. Furthermore, the predictive performance of a variety of different local window sizes L (from 3–17) has been evaluated to select the optimal local window size L for the RSA prediction. Finally, in this encoding scheme, a residue is encoded by a $20 \times L = 20 \times (2l + 1)$-dimensional vector.

In addition, we try to introduce residue conservation score for the solvent accessibility prediction. The value of sequence conservation for residue is a measure of how often a given residue is seen at an equivalent position in an equivalent protein across different species. Generally, residue conservation score is proportional to its buried degree. The conservation score is obtained by PSI-BLAST search as well [39, 40].

PSIPRED-predicted secondary structure information

In this work, we use the PSIPRED program to predict the secondary structure information. PSIPRED provides highly accurate prediction for protein secondary structures by applying a feed-forward neural network. The outputs of PSIPRED are encoded by the probability profiles of three secondary structures (C for coil, H for helix and E for strand). Some previous works have shown that incorporation of PSIPRED-predicted secondary structure information can significantly improve the prediction performance [25, 41].

Analogously, for a given residue, its three-state secondary structure profiles are extracted and encoded using a sliding window of $L = 2l + 1$ consecutive residues. Therefore, in this encoding scheme, a residue is composed of a $3 \times L = 3 \times (2l + 1)$-dimensional vector.

DISOPRED-predicted native disorder information

In the past decade, protein disorder or unstructured regions have received considerable attention in that they are commonly responsible for important protein function. As such, there has been an increasing interest in studying such regions in proteins. Unstructured regions are found to be associated with molecular assembly, protein modification and molecular recognition [42–44]. Research shows unstructured regions have a large solvent accessible area, which explains why polar and charged residues which favorably interact with water are prevalent in these regions [45]. The conclusion is that disordered regions are strongly correlated with local solvent accessibility areas. Local solvent accessibility values are often used to find the disordered regions as well [46, 47].

In order to further improve the performance, in this study, we use DISOPRED program to output the predicted possibility of each residue being natively disordered or ordered. Similarly, a residue is encoded by a $2 \times L = 2 \times (2l + 1)$-dimensional vector in this encoding scheme.

Side-chain environment

The concept of side-chain environment was first purposed by Eisenberg et al. [34] and used to identify protein sequences that fold into a known three-dimensional structure. Then Li et al. [39] utilized it for prediction of protein-protein binding site.

The side-chain environment of a residue is typically defined as buried, partially buried, or exposed based on its solvent accessible surface area. The buried and partially buried residue environments can be further subdivided according to the fraction of side-chain area covered by polar atoms. Based on this, we divide the side-chain environment of a residue into six classes (see Fig. 1). The detailed definition of the side-chain environment were described in the work of Eisenberg et al. [34].

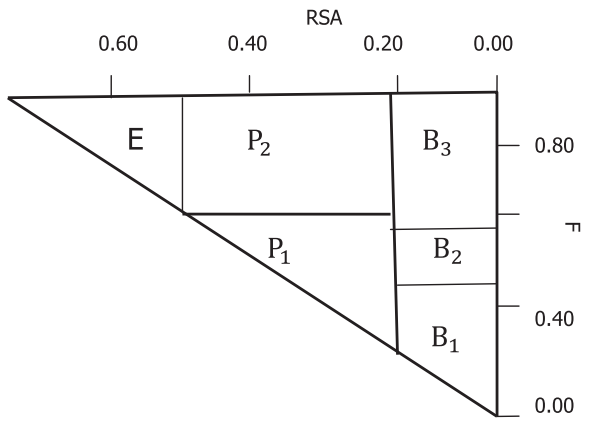

Fig. 1 The definition of the six side-chain environment categories. This figure shows the classification method of side-chain environment. *RSA* represents the relative accessible surface areas and *F* represents the fraction of the whole side-chain area covered by polar atoms. If $RSA < 0.09$, the residue will be placed into class B(buried). If $0.09 \leqslant RSA < 0.36$, the residue will be placed into class P(partial buried). If $RSA \geqslant 0.36$, the residue will be placed into class E(exposed). Within class B, if $F < 0.45$, the residue will be placed into B_1, if $0.45 \leqslant F < 0.58$, the residue will be placed into B_2, and if $F \geqslant 0.58$, the residue will be divided into class B_3. In class P, if $F < 0.67$, the reside will be divided into class P_1, and if $F \geqslant 0.67$, the reside will placed into class P_2

Framework of PredRSA

In this subsection, we describe the PredRSA framework that uses an accurate and effective ensemble computational approach for real values of relative solvent accessibility prediction from protein primary sequences. We are interested in investigating the influence of various sequence-based features and their combinations on the prediction performance of solvent accessibility. In order to fully exploit the sequence-derived features for RSA prediction, we propose a novel PredRSA approach which incorporates five different types of sequence-derived features as inputs. They are four local features (position-specific scoring matrix, secondary structure, conservation score, native disorder) and a global feature (side-chain environment). Figure 2 illustrates the flowchart of our proposed approach.

To determine the optimal local sliding window size L and the iterative tree number M, we calculate the prediction performance for L in the range of 3–17 with a step of 2 and M in the range of 100–1500 with a step of 50 using a grid search method. With $L = 7$ and $M = 800$, the PredRSA approach achieves the best performance for the RSA prediction.

Results and discussion

Datasets

Two non-homologous datasets of proteins chains with pair-wise sequence similarity less than 25 % have been

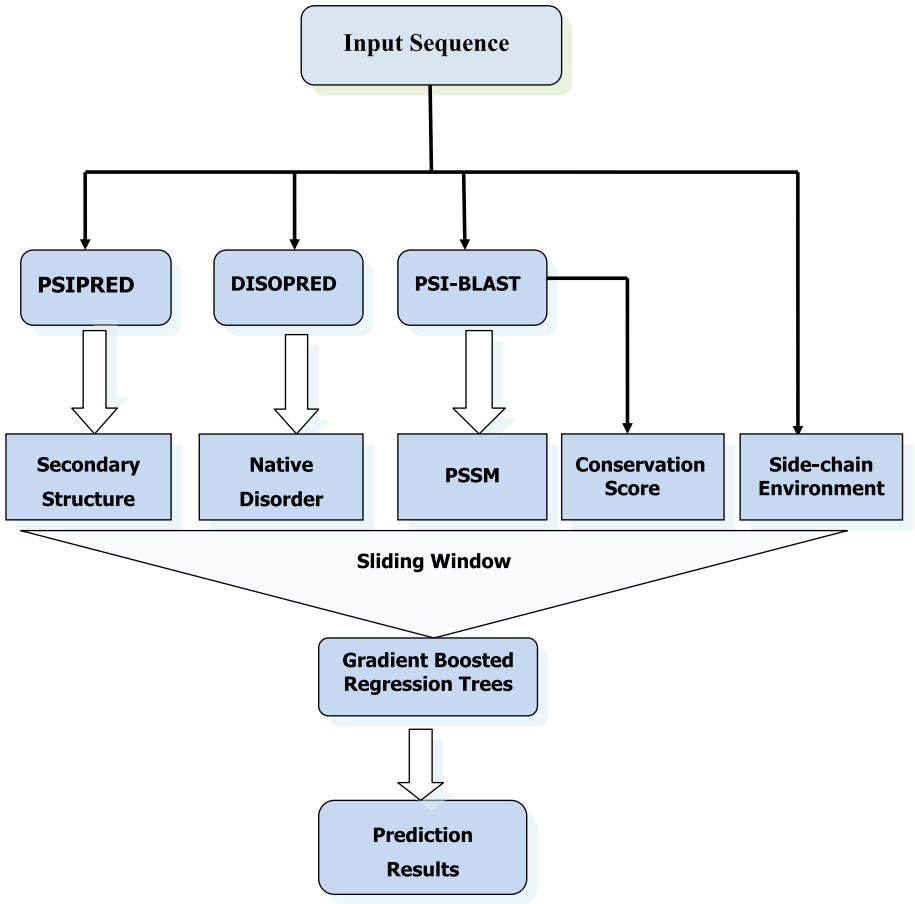

Fig. 2 The framework of PredRSA for protein relative solvent accessibility prediction. Five different types of sequence-derived features are generated and used as input to build the GBRT model. These features consist of PSSM in the form of PSI-BLAST profiles, predicted secondary structure information by PSIPRED, predicted native disorder information by DISOPRED, conservation score and side-chain environment

used in order to objectively compare our approach with other available methods developed previously. One dataset is consisted of 215 proteins, which was also used earlier by Manesh et al. [11] for solvent-accessible surface area of residues prediction. The other dataset is consisted of 502 proteins, obtained from the Cuff and Barton [48] dataset of 513 proteins , selected by removing those sequences, which have less than 30 residues. These two datasets have been referred to as Manesh-215 and CB-502, respectively. However, since the Manesh-215 dataset was widely used by researchers to benchmark prediction methods [18, 20, 25, 49], taking into account comparative purposes, we use Manesh-215 as the main data set for evaluation and analysis.

To further evaluate the performance of existing methods and the method developed in the present study, we also generate an independent dataset of CASP10 proteins. Originally, it contains 85 proteins [50], and we have removed 17 structures (containing chains) by using PISCES culling sever [51] with 25 % sequence similarity cutoff including X-ray (less than 3.0 Å resolution and 0.3 of R-factor) and NMR structures which contain more than 50 residues. Finally, the remaining 68 proteins are used for independent test.

Calculation of RSA

In this work, we take relative solvent accessibility, also called relative solvent accessible surface area (RSA) as the prediction of solvent accessibility. The RSA of a residue in a protein chain is a normalized value from 0–1. It is calculated as the ratio by dividing the solvent accessible surface area (ASA) by the maximum solvent accessibility according to Manesh's work [11] which uses Gly-X-Gly extended tripeptides. The values of ASA are calculated using DSSP [52] for all considered protein structures.

Evaluation measures

To measure the performance of real-valued RSA predictions, three widely used measures for real value RSA prediction are adopted in this study.

The first measure, mean absolute error (MAE), is defined as the average difference between the predicted and experimental RSA values of all residues:

$$MAE = \frac{\sum |RSA_{predicted} - RSA_{experimental}|}{N} \quad (12)$$

The second measure is the root mean square error (RMSE), which is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{N}^{i=1}\left(RSA_{predicted} - RSA_{experimental}\right)^2} \quad (13)$$

The third measure, Pearson correlation coefficient (PCC), the ratio of the covariance between the predicted and experimental RSA values which is given by:

$$PCC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \quad (14)$$

where N is the total number of residues in a protein sequence to predict; x_i and y_i are the experimental and predicted RSA values of the i-th residue, respectively; $\bar{x}$ and $\bar{y}$ are their corresponding means. $PCC = 1$ indicates that the two sets of values are fully correlated, while $PCC = 0$ indicates that they are completely uncorrelated.

Two-state (buried or exposed) predictions are evaluated according to various thresholds of RSA. Prediction accuracy which is defined by the percentage of correctly predicted residues divided by the total number of residues and Matthews correlation coefficient (MCC) are given as follows:

$$ACC = \frac{N_b + N_e}{N} \quad (15)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (16)$$

where N is the total number of residues in a chain, N_b and N_e represent the number of residues correctly predicted as buried and exposed, respectively. TP, TN, FP and FN are the numbers of the true positives, true negatives, false positives and false negatives, respectively.

Effect of different sequence encoding schemes on the prediction performance

We analyze the importance or contribution for each individual feature, which is useful to identify those features that have the most significant influence on overall prediction performance. The performance of each individual predictive is shown in Fig. 3. The feature of side-chain environment is first introduced to predict RSA and it is strongly related to solvent-accessible surface areas.

Table 1 compares the prediction performance of five different combinations of sequence-based features on Manesh-215 with 5-fold cross-validation. As shown in Table 1, the prediction performance of combining all five types of features is the best. It suggests that comprehensive sequence encoding schemes can improve the predictive performance. More importantly, incorporating side-chain environment into the model can significantly increase the prediction performance.

Performance comparison with other regression approaches

In this section, we compare the performance of PredRSA with that of other five existing real value RSA predictors,

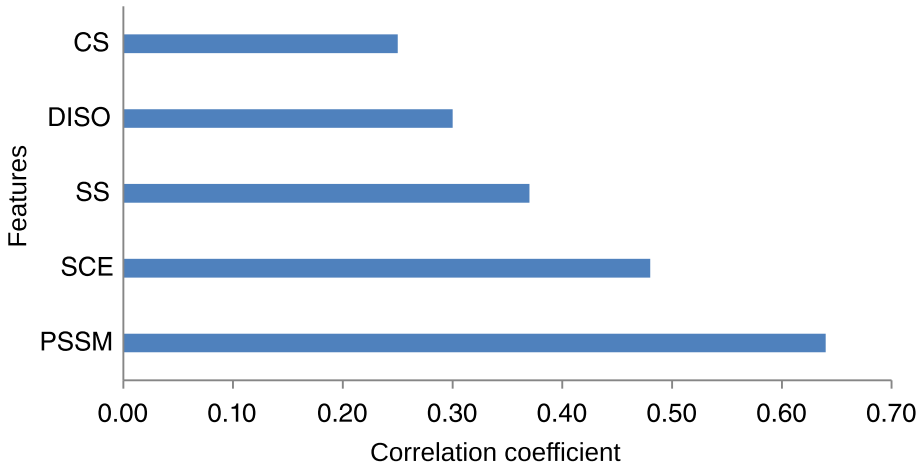

Fig. 3 The importance of the five relevant features used in PredRSA. PSSM, SS, DISO, SCE and CS stand for position specific scoring matrix, protein secondary structure, protein native disorder, side-chain environment and conservation score, respectively

Table 1 Prediction of real-valued RSA using the GBRT algorithm based on five different sequence encoding schemes that incorporate various combinations of sequence features

Feature	RMSE(%)	MAE(%)	PCC
PSSM	13.87	10.26	0.67
PSSM+DISO	13.38	9.93	0.69
PSSM+DISO+SS	13.04	9.64	0.71
PSSM+DISO+SS+SCE	12.23	8.99	0.74
PSSM+DISO+SS+SCE+CS	12.07	8.86	0.75

including a quadratic programming and buriability energy function for solvent accessibility prediction (QBES) [22], a neural network-based method using multiple sequence alignment and secondary structure (SARpred) [25], an improved two-layer neural network (Real-SPINE) [53], a support vector regression using enhanced PSSM features (SVR) [54] and an ensemble of artificial neural networks method (NetSurfP) [55]. Table 2 summarizes the results of these methods. We observe that our method achieves a significantly better performance over the compared predictors. Particularly, the PCC value of PredRSA is approximately 5 % higher than that of the previous predictors on Manesh-215. It is worth to point out that experimental maximum solvent accessibility scores are varied based on different references [11, 56, 57]. A higher maximum solvent accessibility score will lead to a lower RSA value, and thus a relatively lower MAE is obtained according to the definition of MAE. One reason for the differences of MAE between PredRSA and the other methods is that these methods may use different maximum solvent accessibility scores. On the other hand, the prediction precision of PredRSA is higher than that of the other methods and yields a lower MAE.

Performance comparison for two-state prediction

In the past, a plenty of approaches have been proposed for predicting the states (exposed or buried) of residues. Here we examine the performance of our method in terms of two-state prediction. We assign the label of a residue based on its predicted RSA value and a chosen threshold. Table 3 shows the performance of the two-state classification prediction.

Table 2 Performances comparison in predicting real values: PredRSA vs. other existing methods

	Manesh-215		CB-502	
Method	MAE(%)	PCC	MAE(%)	PCC
QBES	-	0.52	-	0.49
SARpred	14.9	0.68	15.9	0.66
SVR	14.2	0.69	14.8	0.68
Real-SPINE	13.8	0.70	14.5	0.68
NetSurfP	13.6	0.70	14.3	0.71
PredRSA	9.0	0.75	9.4	0.73

Table 3 Prediction performance of two-state classification based on different thresholds

	Manesh-215		CB-502		CASP10	
Threshold(%)	ACC(%)	MCC	ACC(%)	MCC	ACC(%)	MCC
5	80.1	0.54	77.9	0.50	78.5	0.48
10	81.7	0.63	79.0	0.58	79.1	0.57
20	81.0	0.61	80.5	0.60	78.3	0.56
25	81.2	0.58	81.0	0.57	79.7	0.56
30	82.4	0.54	82.1	0.52	80.5	0.51
40	87.1	0.42	86.8	0.39	85.0	0.40
50	93.2	0.25	93.0	0.23	91.2	0.30

We also compare the classification accuracy of PredRSA with that of other approaches by different thresholds. The threshold is used to determine the state (exposed or buried) of a predicted real value. For example, a 5 % threshold means a residue is defined as buried if its RSA value is less than 5 %. The methods for comparison include SARpred [25], pace regression algorithm (PR) [58], two-stage SVR [19] and SVR [54]. The prediction accuracy is showed in Table 4. Our method yields more than 80 % classification accuracy at any thresholds and obtains almost the highest accuracy across all the thresholds.

Independent test on the CASP10 dataset

An independent test (CASP10) is constructed to further validate the usability of our PredRSA method. We train the classifiers based on the Manesh-215 dataset and test against the CASP10 dataset which contains 68 proteins. Other state-of-the-art methods including SPINE-X [59] and ASAquick [60] are also evaluated. SPINE-X uses a multistep neural-network algorithm by coupling secondary structure prediction with prediction of solvent accessibility and backbone torsion angles in an iterative manner, while ASAquick utlizes solely sequential widow information and global features with a general neural network method. The Pearson correlation coefficient of PredRSA is 0.71, which outperform the results of SPINE-X and ASAquick by a rate of 2 % (0.69) and 4 % (0.67).

Table 4 Performance comparison of two-state classification: PredRSA vs. other existing predictors

	Accuracy for two-states prediction(%)						
Method	5 %	10 %	20 %	25 %	30 %	40 %	50 %
SARpred	74.9	77.2	77.7	-	77.8	78.1	80.5
PR	76.8	74.8	75.3	76.7	77.7	79.8	86.3
SVR	80.9	80.1	78.7	-	-	-	80.8
Two-stageSVR	81.1	78.7	77.6	77.3	-	-	79.5
PredRSA	80.0	81.6	80.9	81.1	82.2	87.1	93.2

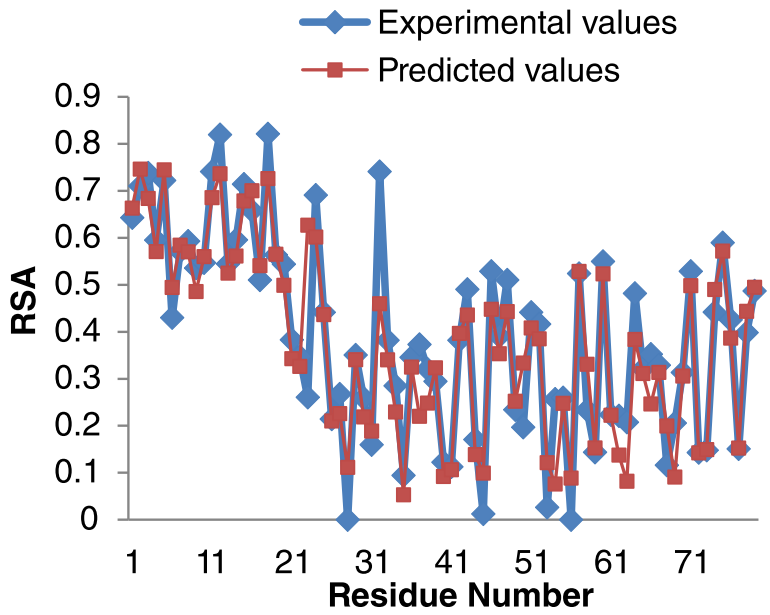

Fig. 4 Predicted and experimental values (%) of RSA for each residue of CASP10 T0675

Case study

For a better understanding of the power of our proposed PredRSA approach and illustrating the significance of PCC, RMSE and MAE measures used in this work, an example of the real-valued RSA for T0675 (Insulinoma-associated protein) from CASP10 is shown in Fig. 4. For this protein, our method gives a MAE of 5.31 %, a RMSE of 7.91 % and a PCC of 0.92. From Fig. 4, we can see that the majority of its predicted RSA values are in good agreement with the corresponding experimental RSA values calculated by DSSP, except for several separate positions.

In Fig. 5, the continuous real-value prediction of RSA and the actual continuous values are shown. Significant correlation between the true values and the predicted values is obtained.

Residue-specific variation in prediction error

In order to assess the prediction performance of various types of residues, we further calculate the average RSA values on the Manesh-215 dataset for all 20 amino acids (Fig. 6) from the PredRSA predictor. As can be seen from Fig. 6, an overwhelming majority of types of amino acids

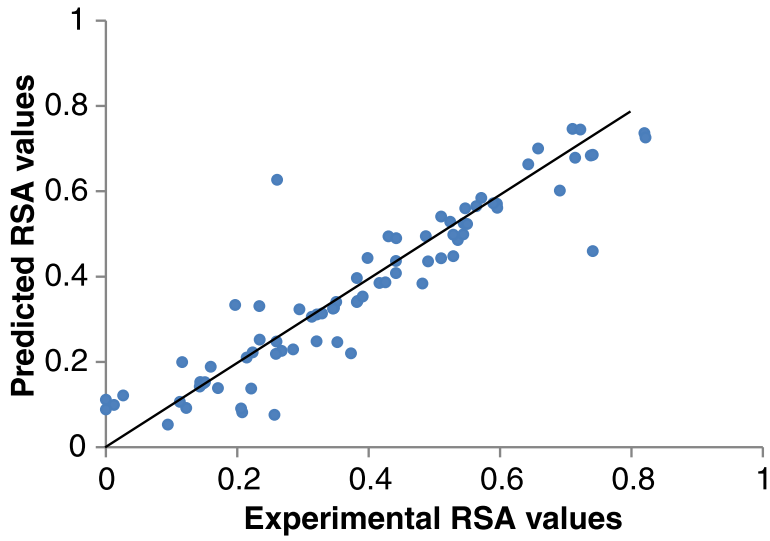

Fig. 5 Correlation between experimental RSA values and predicted RSA values of CASP10 T0675. The Pearson correlation coefficient score is 0.92 and the most buried residues are well predicted with the RSA values near zero

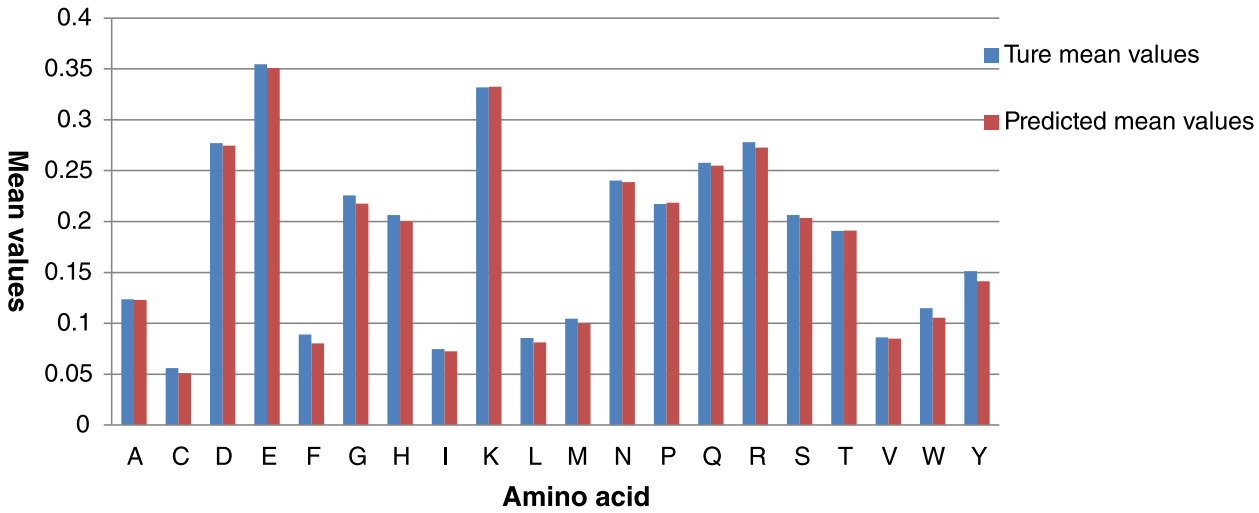

Fig. 6 Comparison between true mean values and predicted mean values for 20 amino acids on the Manesh-215 dataset

are predicted with <1 % mean error. All types of amino acids are predicted with <2 % mean error in our method. In particularly, we find that the true mean RSA values are in highly accord with the predicted mean RSA values for these amino acids, such as A (Ala), K (Lys), N (Asn), T (Thr).

Furthermore, we calculate the prediction errors of 20 amino acids on the Manesh-215 dataset. Figure 7 shows the mean absolute error (MAE) and the standard root mean square error (RMSE) of 20 amino acids. As expected, G (Gly) shows the highest MAE and RMSE due to its flexibility, and other polar residues show similar behavior. Hydrophobic amino acids including C (Cys), F (Phe), M (Met) and W (Trp) are better predicted than less hydrophobic amino acids. These results are also in good agreement with our PredRSA method.

Conclusions

Knowledge of residue solvent accessibility gives useful insights into protein structure and function prediction. In this work, we have presented PredRSA to predict real-valued relative solvent accessibility as well as classification state (buried or exposed) of a target residue. The method is based on a gradient boosted regression trees (GBRT)

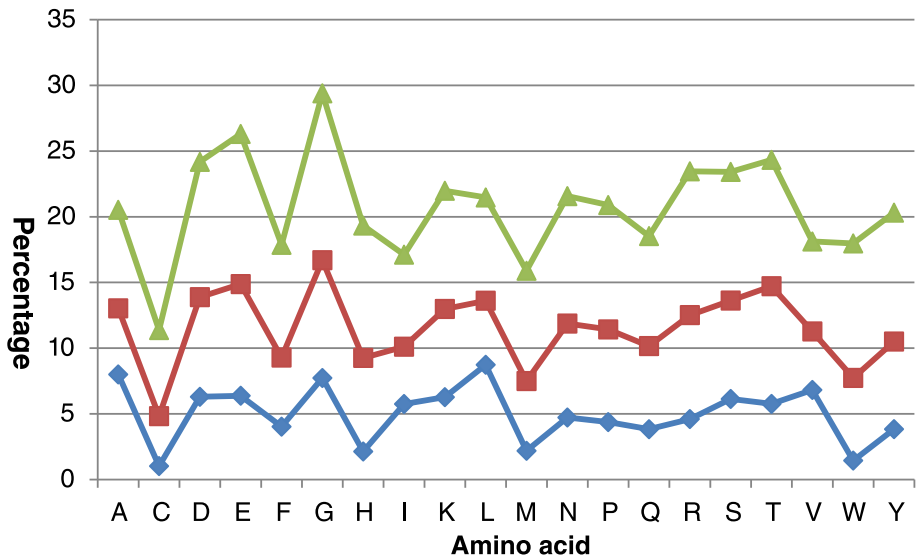

Fig. 7 Mean predicted errors of 20 amino acids on the Manesh-215 dataset. The green line represents standard root mean square error, the red line represents mean absolute error and the blue line represents the corresponding data distribution of 20 amino acids

algorithm combined with a novel set of features. The 5-fold cross-validated correlation coefficient between predicted and experimental RSA (0.75) is significantly better than existing methods on the Manesh-215 dataset. We also performed additional independent benchmark tests of PredRSA on the CASP10 set containing 68 proteins where we find that the proposed method outperforms existing methods. Furthermore, for prediction of discrete state, our method is able to achieve an accuracy of 79.7 % with an MCC value of 0.56 using two states classifications at a threshold of 25 %, which defines an approximately balanced division into the two classes.

Experimental results show GBRT is an efficient machine learning approach for continuous values of the solvent accessibility of a target residue. Compared with other traditional techniques, GBRT has several obvious advantages such as high prediction accuracy and stronger generalization capability.

On the other hand, PredRSA utilizes a variety of multiple sequence-derived features, including the position-specific scoring matrices and conservation score in the form of PSI-BLAST profiles, predicted secondary structure, predicted natively disordered region and side-chain environment. We have comprehensively assessed the effects of different sequence encoding schemes on the prediction performance of RSA, and the results show the prediction performance of RSA outperforms previous methods. Our work provides a complementary and useful approach towards the more accurate prediction of protein solvent accessibility.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CF carried out the literature study, developed the new method and drafted the manuscript. CF and DL participated in several independent tests. LD participated in its design and coordination. LD, RH and ZC helped to draft and revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by National Natural Science Foundation of China under grants No. 61309010 and No. 61379057, China Postdoctoral Science Foundation under grant no. 2015T80886, Specialized Research Fund for the Doctoral Program of Higher Education of China under grant no. 20130162120073 and Shanghai Key Laboratory of Intelligent Information Processing under grant no. IIPL-2014-002.

Declarations
The publication fee of this article is funded by National Natural Science Foundation of China under grant No.61309010. This article has been published as part of BMC Bioinformatics Volume 17 Supplement 1, 2016: Selected articles from the Fourteenth Asia Pacific Bioinformatics Conference (APBC 2016). The full contents of the supplements are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/17/S1.

Published: 11 January 2016

References

1. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J mole biol. 1971;55(3):379–4.
2. Eyal E, Najmanovich R, Mcconkey BJ, Edelman M, Sobolev V. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. J comput chem. 2004;25(5):712–24.
3. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins Struct Funct Genet. 1994;20(3):216–26.
4. Wodak SJ, Janin J. Location of structural domains in proteins. Biochem. 1981;20(23):6544–52.
5. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins Struct Funct Genet. 2007;68(3):636–45.
6. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature. 1986;319(6050):199–203.
7. Mooney C, Pollastri G, Shields DC, Haslam NJ. Prediction of short linear protein binding regions. J mol biol. 2012;415(1):193–204.
8. Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D. Predus: a web server for predicting protein interfaces using structural neighbors. Nucleic acids res. 2011;39(suppl 2):283–7.
9. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell res. 2009;19(8):929–49.
10. Huang B, Schroeder M. Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC structural biol. 2006;6(1):19.
11. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins Struct Funct Bioinforma. 2001;42(4):452–9.
12. Ahmad S, Gromiha MM. Netasa: neural network based prediction of solvent accessibility. Bioinforma. 2002;18(6):819–24.
13. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins Struct Funct Bioinforma. 2002;48(3):566–70.
14. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3d local descriptor. Proteins Struct Funct Bioinforma. 2004;54(3):557–62.
15. Sim J, Kim SY, Lee J. Prediction of protein solvent accessibility using fuzzy k-nearest neighbor method. Bioinforma. 2005;21(12):2844–9.
16. Wang JY, Lee HM, Ahmad S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. Proteins Struct Funct Bioinforma. 2005;61(3):481–91.
17. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins Struct Funct Bioinforma. 2004;57(3):558–64.
18. Xu W, Li A, Wang X, Jiang Z, Feng H. Improving prediction of residue solvent accessibility with svr and multiple sequence alignment profile. Conf Proc IEEE Eng Med Biol Soc. 2005;3:2595–8.
19. Nguyen MN, Rajapakse JC. Two-stage support vector regression approach for predicting accessible surface areas of amino acids. Proteins Struct Funct Bioinforma. 2006;63(3):542–50.
20. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins Struct Funct Bioinforma. 2003;50(4):629–35.
21. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks–based regression. Proteins Struct Funct Bioinforma. 2004;56(4):753–67.
22. Xu Z, Zhang C, Liu S, Zhou Y. Qbes: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. Proteins Struct Funct Bioinforma. 2006;63(4):961–6.
23. Joo K, Lee SJ, Lee J. Sann: solvent accessibility prediction of proteins by nearest neighbor method. Proteins Struct Funct Bioinforma. 2012;80(7): 1791–7.
24. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The pfam protein families database. Nucleic acids res. 2002;30(1):276–80.
25. Garg A, Kaur H, Raghava G. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins Struct Funct Bioinforma. 2005;61(2):318–24.
26. Song J, Tan H, Wang M, Webb GI, Akutsu T. Tangle: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences. PloS ONE. 2012;7(2):30361.
27. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232.
28. Huber PJ. Robust estimation of a location parameter. Ann Math Stat. 1964;35(1):73–101.

29. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids res. 1997;25(17):3389–402.
30. Deng L, Guan J, Wei X, Yi Y, Zhang QC, Zhou S. Boosting prediction performance of protein–protein interaction hot spots by using structural neighborhood properties. J Comput Biol. 2013;20(11):878–91.
31. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S. PredHS: a web server for predicting protein–protein interaction hot spots by using structural neighborhood properties. Nucleic acids res. 2014;42:W290–295.
32. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J mol biol. 1999;292(2):195–202.
33. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J mol biol. 2004;337(3):635–45.
34. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science. 1991;253(5016):164–70.
35. Zhang J, Zhao X, Sun P, Ma Z. Psno: predicting cysteine s-nitrosylation sites by incorporating various sequence-derived features into the general form of chous pseaac. Int J Mol Sci. 2014;15(7):11204–19.
36. Song J, Burrage K, Yuan Z, Huber T. Prediction of cis/trans isomerization in proteins using psi-blast profiles and secondary structure information. BMC bioinforma. 2006;7(1):124.
37. Chen K, Kurgan L. Pfres: protein fold classification by using evolutionary information and predicted secondary structure. Bioinforma. 2007;23(21): 2843–50.
38. Mizianty MJ, Kurgan L. Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure, and evolutionary information. Proteins Struct Funct Bioinforma. 2011;79(1):294–303.
39. Li N, Sun Z, Jiang F. Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC bioinforma. 2008;9(1):553.
40. Deng L, Guan J, Dong Q, Zhou S. Prediction of protein-protein interaction sites using an ensemble method. BMC bioinforma. 2009;10(1):426.
41. Pugalenthi G, Kumar Kandaswamy K, Chou KC, Vivekanandan S, Kolatkar P. Rsarf: prediction of residue solvent accessibility from protein sequence using random forest method. Protein and peptide letters. 2012;19(1):50–6.
42. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6(3):197–208.
43. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol. 2006;2(8):100.
44. Gsponer J, Futschik ME, Teichmann SA, Babu MM. Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science. 2008;322(5906):1365–8.
45. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B. Improved disorder prediction by combination of orthogonal approaches. PLoS ONE. 2009;4(2):4433.
46. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins Struct Funct Bioinforma. 2009;76(3):617–36.
47. Marsh JA. Buried and accessible surface area control intrinsic protein flexibility. J mol biol. 2013;425(17):3250–63.
48. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins Struct Funct Bioinforma. 2000;40(3):502–11.
49. Wang JY, Ahmad S, Gromiha MM, Sarai A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. Biopolymers. 2004;75(3):209–16.
50. The CASP10 Database. http://predictioncenter.org/casp10/groups_analysis.cgi. Accessed 2012.
51. Wang G, Dunbrack RL. Pisces: a protein sequence culling server. Bioinforma. 2003;19(12):1589–91.
52. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
53. 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 Struct Funct Bioinforma. 2009;74(4):847–56.
54. Chang DT, Huang HY, Syu YT, Wu CP. Real value prediction of protein solvent accessibility using enhanced pssm features. BMC bioinforma. 2008;9(Suppl 12):12.
55. 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;9(1):51.
56. Chothia C. The nature of the accessible and buried surfaces in proteins. J mol biol. 1976;105(1):1–12.
57. Oobatake M, Ooi T. Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol. 1993;59(3):237–84.
58. Meshkin A, Sadeghi M, Ghasem-Aghaee N. Prediction of relative solvent accessibility using pace regression. EXCLI J. 2009;8:211–7.
59. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J comput chem. 2012;33(3):259–67.
60. Faraggi E, Zhou Y, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. Proteins Struct Funct Bioinforma. 2014;82(11):3170–6.