Journal of Theoretical Biology 441 (2018) 44–57

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

RBSURFpred: Modeling protein accessible surface area in real and binary space using regularized and optimized regression

Sumit Tarafder[a], Md. Toukir Ahmed[a], Sumaiya Iqbal[b], Md Tamjidul Hoque[c], M. Sohel Rahman[a,*]

[a] Department of CSE, BUET, ECE Building, West Palasi, Dhaka 1205, Bangladesh
[b] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[c] Department of Computer Science, University of New Orleans, LA, USA

ARTICLE INFO

Article history:
Received 12 October 2017
Revised 11 December 2017
Accepted 28 December 2017
Available online 2 January 2018

Keywords:
Accessible surface area
PSEE
Protein structure
Relative solvent accessibility
Metaheuristics

ABSTRACT

Accessible surface area (ASA) of a protein residue is an effective feature for protein structure prediction, binding region identification, fold recognition problems etc. Improving the prediction of ASA by the application of effective feature variables is a challenging but explorable task to consider, specially in the field of machine learning. Among the existing predictors of ASA, REGAd^{3}p is a highly accurate ASA predictor which is based on regularized exact regression with polynomial kernel of degree 3. In this work, we present a new predictor RBSURFpred, which extends REGAd^{3}p on several dimensions by incorporating 58 physicochemical, evolutionary and structural properties into 9-tuple peptides via Chou's general PseAAC, which allowed us to obtain higher accuracies in predicting both real-valued and binary ASA. We have compared RBSURFpred for both real and binary space predictions with state-of-the-art predictors, such as REGAd^{3}p and SPIDER2. We also have carried out a rigorous analysis of the performance of RBSURFpred in terms of different amino acids and their properties, and also with biologically relevant case-studies. The performance of RBSURFpred establishes itself as a useful tool for the community.

1. Introduction

Proteins are large macromolecules which consist of polymer chain of amino acids known as polypeptide. The monomers of this polymer chain are amino acids. Proteins can also have secondary, tertiary and quaternary structures. The appropriate determination of these protein structures as well as the properties are fundamental in structural and functional biology. Among the properties of protein, *accessible surface area* (ASA) is one of the most important properties (Lee and Richards, 1971). The ASA of a protein residue refers to the area of that residue exposed to the solvent molecule while dissolved in the solvent (usually water) surrounding the protein. Hydrophobic residues are the ones that reside at the core of the protein structure whereas the hydrophilic (or polar) residues remain on the surface of the protein structure. Solvent exposure can be numerically described by several measures and the most popular measures among them are ASA and *relative solvent accessibility* (RSA). ASA works as an important feature for secondary structure prediction (Faraggi et al., 2012; Wang et al., 2016; Zhou and Troyanskaya, 2014), disordered residue prediction (Iqbal and Hoque, 2015; Eickholt and Cheng, 2013; Jones and Cozzetto, 2014; Wang et al., 2015), hotspot residue prediction (Cho et al., 2009) and protein fold recognition (Liu et al., 2007). RSA has been considered as an essential measure for spatial arrangement of protein folding, determination of protein domain boundary (Eickholt et al., 2011) etc. Surface areas being in the form of exposed residues, are engaged in inter protein interactions. The conformational dynamics of proteins, characterized by flexible regions and thermal fluctuations (B-factor) of a protein, is important for their functionality and is found to be correlated with the ASA of every single residue of a protein (Marsh, 2013; Zhang et al., 2009). Accurate prediction of ASA improved the accuracy of *ab initio* protein structure prediction (Bonetti et al., 2014) and energy function development for correct discrimination of native conformation from the decoys (Khashan et al., 2012; Wang and Hou, 2012) as well. By the correct analysis of ASA, the existence of low-frequency phonons in proteins was revealed (KUO-CHENG and NIAN-YI, 1977), which opened a new area for studying the internal motion of bio-macromolecules and functions thereof (Chou et al., 1981). A comprehensive review (Chou, 1988) reports that, each unit of protein accessible hydrophobic surface contributes a certain amount of free energy when such an area is buried within a protein, which eventually

* Corresponding author.
E-mail addresses: sumaiya@broadinstitute.org (S. Iqbal), thoque@uno.edu (M. Tamjidul Hoque), msrahman@cse.buet.ac.bd (M. Sohel Rahman).

https://doi.org/10.1016/j.jtbi.2017.12.029
0022-5193/

indicates the existance of low frequency vibrations in the protein molecule. Thus, it is evident that, accurate prediction of real-valued ASA and binary classification of RSA from primary protein sequence is a crucial but rewarding task in proteomics.

The solvent accessibility prediction has been studied in two forms: firstly, multiclass classification problem (Ahmad and Gromiha, 2002; Gianese et al., 2003; Holbrook et al., 1990; Kim and Park, 2014; Li and Pan, 2001; Rost and Sander, 1994; Yuan et al., 2002) and secondly, real-value prediction problem (Ahmad et al., 2013; Faraggi et al., 2009a; 2012; Wang et al., 2007). However, multiclass classification problems are often transformed into a binary classification problem according to a defined threshold of RSA values. The state-of-the-art works for real value prediction of ASA includes several pattern recognition algorithms, such as, multiple linear regression, support vector machines (SVM) (Wang et al., 2007; 2005), artificial neural network (ANN) (Ahmad et al., 2013; Faraggi et al., 2009a; 2012) and deep neural-network learning with parallel multistep iterative algorithm (Khashan et al., 2012). All these state-of-the-art predictors use various variables like position specific scoring matrix(PSSM), physical properties, secondary structure probabilities of the amino acids etc. Among the works of binary prediction, one of the most recent work is Wu et al. (2017) which uses a defined threshold and a Conditional Random Field (CRF) (Lafferty et al., 2001; Wang and Sauer, 2008) modeling routine. Most proteins have significant functions that are incorporated with protein–protein interactions that affect the biological processes in a living cell. To really understand these interactions, it is necessary to acquire the information of Protein–Protein Binding Sites (PPBSs). To intuitively identify the PPBS based only on the sequence information, a predictor named iPPBS-PseAAC was proposed recently (Jia et al., 2016a). iPPBS-PseAAC predicts the binding site based on two layer ensemble classifier for training data and feature selection. A similar predictor, called iPPBS-Opt (Jia et al., 2016b), was also proposed recently in which K-Nearest Neighbors Cleaning (KNNC) and Inserting Hypothetical Training Samples (IHTS) treatments were used to optimize the training dataset and the ensemble voting approach was used to select the most relevant features. Various studies show that hydrophobic residues tend to occur in protein binding regions more often than hydrophilic residue (DeLano, 2001; Glaser et al., 2001). The conservation scores of amino acid are often used as features, because the protein binding sites are more conserved than other surface residues (Zhou and Shan, 2001). The binding sites also show higher ASA values than those of the other surface residues (Chen and Zhou, 2005). So an accurate prediction of ASA can significantly help in identifying which residues constitute the binding sites and thus contribute in improving applications stated above.

The real-value prediction of ASA is often preferred over the binary-state prediction since the residues surface area tends to vary largely due to their free movement in three-dimensional space. With a view to this, a real-value prediction framework was proposed by Iqbal et al., called REGAd^{3}p (Iqbal et al., 2015), which used machine learning technique, such as regularized exact regression and genetic algorithm. However, there are applications where recognizing the exposed residues at the protein's surface has critical implications, e.g., to find the potential binding-sites on the surface of peptide-recognition domain that mainly stay in the largest pocket of the protein's surface. In this work, we propose the comprehensive surface area predictor that computes the real-valued ASA of protein residues as well as classifies them as buried or exposed base on the normalized ASA value. We call this predictor as "**R**eal and **B**inary space **SURF**ace Area **pred**ictor" or **RBSURFpred** in short.

The proposed predictor uses the real-value predictor framework with higher-order polynomial kernel proposed in Iqbal et al. (2015). However, we extended our model in several aspects. We introduced three new features, one energy-based property (Iqbal and Hoque, 2016) and two dihedral torsion angle fluctuations (Zhang et al., 2010), to improve the prediction accuracy. Unlike Iqbal et al. (2015) that used genetic algorithm for optimization purpose, we implemented 4 different optimization techniques and selected differential evolution algorithm as it performed the best for this application. To perform the two-state binary predictions, we chose an appropriate threshold of value 0.18. We compared both real and binary space predictions with state-of-the-art predictors, such as REGAd^{3}p (Iqbal et al., 2015) and SPIDER2 (Khashan et al., 2012). Lastly, we carried out a rigorous analysis of the performance of RBSURFpred in terms of different amino acids and their properties, and also with biologically relevant case-studies. The overall promising performance of RBSURFpred establishes it as a useful tool for the community.

In developing a really useful sequence-based statistical predictor for a biological system as reported in a series of recent publications (Chen et al., 2017; Cheng et al., 2018; Feng et al., 2017a; Liu et al., 2016), one should observe the Chou's 5-step rule (Chou, 2011); i.e., (*i*) construct or select a valid benchmark dataset to train and test the predictor, described in Section 2.1; (*ii*) formulate the biological sequence samples with an effective mathematical expression or a set of features, that can truly reflect their intrinsic correlation with the target to be predicted as discussed in Section 2.2; (*iii*) introduce or develop a powerful algorithm (or engine) to operate the prediction, discussed in Sections 3; (*iv*) the performance of the proposed predictor is reported, analyzed and compared with statistical measures and case-studies which is discussed in Section 4; (*v*) establish a user-friendly web-server for the predictor that is accessible to the public. Although we leave establishing the web-server as a future work, the RBSURFpred code has been made publicly accessible[1] for researchers and practitioners. Below, we are going to describe how to deal with these steps one-by-one.

2. Materials

In this section, we describe the datasets used to train and validate the model, feature set preparation along with the extraction of three new features, and the model evaluation criteria.

2.1. Dataset information

We used the dataset prepared by Hoque-BML Lab and reported and used in Iqbal et al. (2015). The dataset was prepared from Protein Data Bank (PDB) (Berman et al., 2000) which is referred to as the Secondary Structure Dataset (SSD1299), consisting of 1299 protein sequences. Initially, 2793 protein chains (both single and multiple chain) were collected from PDB with following specifications: (*a*) solved by X-ray crystallography; (*b*) resolution<= 1.5 Å; (*c*) chain length>=40 residues and (*d*) 30% sequence identity cut-off. The dataset was further refined in three steps: (*i*) no chain with pair wise sequence similarity greater than 25% was allowed in the set; (*ii*) the protein sequences that contained unknown amino acid residues labeled as 'X' or 'Z' were discarded as the physiochemical properties for this amino acid were unknown and (*iii*) the sequences containing residues of unknown coordinates were removed as the actual ASA cannot be computed for those residues. This resulted in a dataset of 1299 sequences (SSD1299) with 272,800 residues. For this study, we considered 1296 sequences as we could not generate evolutionary features for 3 chains in the SSD1299 dataset. Randomly selected 295 sequences were separated from 1296 chains and used as the test set, named

[1] https://github.com/Sumit46/RBSURFpred.

SSD_TS295. The remaining 1001 sequences were used as the training dataset (SSD_TR1001).

SSD_TR1001 contains 210,967 residues which combines 69,253 helix (32.8%), 51,859 beta (24.5%) and 89,856 coil (42.5%) residues and SSD_TS295 consists of 61,074 residues which combines 19,792 helix (32.4%), 16,052 beta (26.28%) and 25,230 coil (41.32%) residues. The annotation of secondary structure and ASA were determined by the DSSP program (Kabsch and Sander, 1983).

2.2. Input feature set

To develop a model to predict the protein surface area in both real and binary space, we used a set of 58 features which were carefully chosen to be able to include useful properties such as the sequence information, evolutionary information as well as the structural information. These features are: (*i*) one amino acid (AA) indicator; (*ii*) seven physiochemical properties (PP); (*iii*) twenty position specific scoring matrix (PSSM) values; (*iv*) one monogram (MG) and twenty bigram (BG) values; (*v*) three predicted secondary structure (SS) probabilities (helix, beta and coil); (*vi*) two predicted disorder probabilities (short and long) (IUS and IUL); (*vii*) one position specific estimated energy (PSEE); (*viii*) two torsion angle fluctuation ($\Delta\Phi$, $\Delta\Psi$) and (*ix*) one per-residue terminal tag (T).

To enhance the quality of the proposed predictor, we used 3 additional features along with the 55 features used by REGAd^{3}p (Iqbal et al., 2015). These new features are the two backbone dihedral torsion angle fluctuations(AFs), $\Delta\Phi$ and $\Delta\Psi$ and one position specific estimated energy, PSEE (Iqbal and Hoque, 2016). The previous work (Iqbal et al., 2015) found all 55 features important for ASA prediction. However, there are always provisions for extracting better features which motivated us to use new features in this study. The three newly introduced features give us additional value irrespective of the existence of the previous 55 features, which is discussed in Section 4.1.1. Therefore, we believe that this set of 58 features can represent a single residue of an amino acid much more appropriately. For the 55 features borrowed from Iqbal et al. (2015), we followed the same methodology of extraction as described therein. We omit the details here for the sake of brevity. The interested readers are referred to Iqbal et al. (2015) for details. However, we elaborately discuss the new features and their extraction process in the following subsections.

2.2.1. Extraction of position specific estimated energy (PSEE) from the primary structure

Position specific estimated energy (PSEE) is a relatively new concept which was introduced in Iqbal and Hoque (2016) as a feature for disordered protein residue prediction. PSEE was extracted to characterize the favorable state (negative energy gain) of the folded protein residues and the otherwise neutral state (non-negative energy) of disordered protein residues. The study showed that PSEE can be regarded as an effective feature for the development of tools to predict disordered vs ordered residues, residues of different types of secondary structure, solvent exposure of protein residues and so on, where 1D sequence information to 3D structural mapping is essential. Specifically, the intrinsic disorder property of protein has a strong correlation with its solvent exposure. A protein that is accessible to the partners with a larger interaction pane is more likely to achieve the heterogeneous conformations required to be a disordered protein. With a view to these observations, in this research, we intend to use PSEE as a feature for both real-values and 2-state exposure prediction.

PSEE of a protein residue is computed from the protein's primary sequence without having any knowledge about its 3-dimensional structure, as described in Iqbal and Hoque (2016). This estimation of energy requires three concepts: (*i*) the residues that are in contact with the target residue, (*ii*) the contact energies in the neighborhood of primary protein sequence and (*iii*) relative solvent accessibility of the target residue and its contact residues.

PSEE of a residue (AA_i) is formulated as :

$$PSEE(AA_i) = pBur(AA_i)\left[\frac{\sum_{AA_j \epsilon N_i} P(AA_i, AA_j) \times pBur(AA_j)}{2CR}\right] \quad (1)$$

where,

AA_j = residues in the neighborhood N_i of AA_i

CR = Contact radius number of residues on the either side of AA_i

$P(AA_i, AA_j)$ = predicted pairwise contact energy (Dosztanyi et al., 2005) between AA_i and AA_j

$pBur(AA_j)$ = proportional burial of AA_j

The proportional exposure, $pExp(AA_i)$, measures to what extent an amino acid residue is accessible to the solvent which can be expressed as follows:

$$pExp(AA_i) = \frac{predicted\ ASA\ (AA_i)}{ASA\ (AA_i)\ in\ the\ conformation\ Gly - AA_i - Gly} \quad (2)$$

Thus, proportional burial of a residue AA_i is computed as follows:

$$pBur(AA_i) = 1 - pExp(AA_i) \quad (3)$$

The ASA normalization values for 20 different types of amino acids are collected from Tien et al. (2013). We used the contact radius as 9 as specified in Iqbal and Hoque (2016). PSEE was computed by running the DisPredic2 software (Iqbal and Hoque, 2016).

2.2.2. Extraction of angle fluctuations

We considered two other new features that are the torsion angle fluctuations $\Delta\Phi$ and $\Delta\Psi$. The two backbone torsion angles, Φ and Ψ, can define the fine-grained description of the backbone of a protein structure. Therefore, the angle fluctuations (AFs), which are computed from the ensemble of structures solved by Nuclear Magnetic Resonance (NMR) spectroscopy (Zhang et al., 2010), represent the flexibility of protein backbone. Solvent exposure is correlated with the flexibility or the atomic motion of a residue as the core of a protein (buried area) is mostly solid-like and the exposed surface is mostly molten-like (Zhou et al., 1999).

To calculate $\Delta\Phi$ and $\Delta\Psi$, we used a software called DAVAR (Zhang et al., 2010). For each residue, we prepared a 34-dimensional feature vector to predict the angle fluctuations. These features include the seven representative physical parameters (Meiler et al., 2001) and the 20-dimensional PSSM vector calculated from the PSI-BLAST profiles by querying a given sequence with three iterations against the NCBI's non-redundant (nr) protein sequence database. Moreover, six predicted structural properties were generated using SPINE X (Faraggi et al., 2009b). They are 3 secondary structure probabilities, 1 solvent accessibility and 2 torsion angles, as required by the DAVAR software. The predicted solvent accessibility was normalized by the solvent ASA of an extended conformation (Gly-X-Gly) (Ahmad and Gromiha, 2002; Zhang et al., 2009) and the two predicted torsion angles were normalized by 180°. Furthermore, the per-residue disorder probability output given by IUpred (Dosztanyi et al., 2005) was used. After collecting these 34 features we used the DAVAR tool to predict the angle fluctutations. The predictor is a two hidden-layer (51 hidden neurons) neural network with a hyperbolic activation function and guided learning technique. We performed 5 rounds of computation for predicting each of the angle fluctuations using the software and finally used the average output.

To understand how the torsion-angle fluctuations differ for different amino acid types we plotted the mean torsion-angle fluctuation values of each amino acid type which is shown in Fig. 1. Glycine (G) and two hydrophilic residues serine (S) and histidine (H) in general are the top three most flexible residues with high

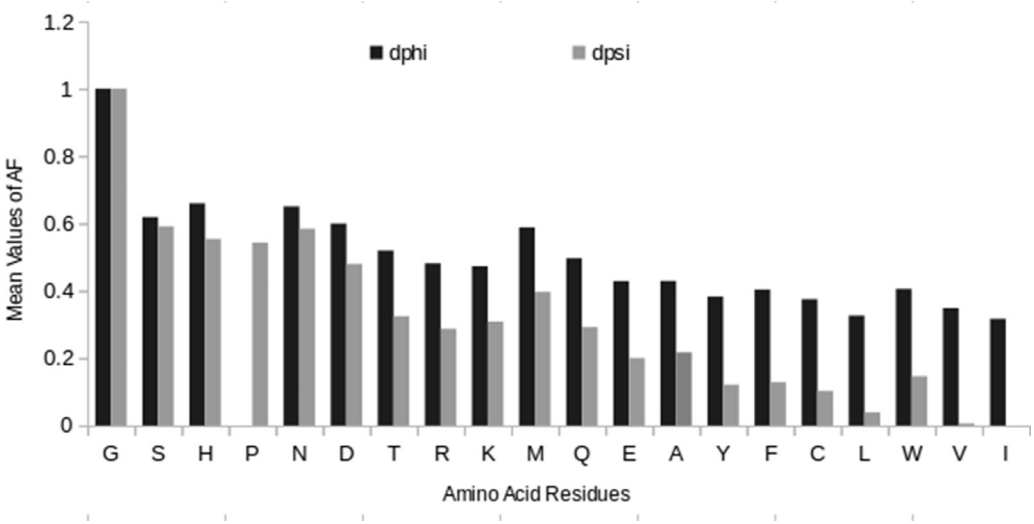

Fig. 1. Mean torsion-angle fluctuations of Φ and Ψ for the 20 amino acid types.

mean values of both ΔΦ and ΔΨ, whereas three hydrophobic residues tryptophan (W), valine (V), and isoleucine (I) are the least flexible residues with very low amount of ΔΦ and ΔΨ.

Hydrophobic residues are more likely to be buried and less flexible and hydrophlic residues are in generally exposed residues with high flexibility. Because of this correlation we are inspired to use these two angle fluctuation features in our high accuracy ASA predictor. We can observe from the Fig. 1 that ΔΦ and ΔΨ are significantly closer for all amino acids except proline (P), which is characterized by significantly lower ΔΦ. This is due to the fact that the last atom of the proline side chain is bonded to the main chain, forming a ring which restricts the available conformational space and results in a nearly fixed ΔΦ angle.

We further checked the correlation of these 3 new features with rest of the 55 features and evaluated the performance of the predictor with different feature combinations to understand their contribution (see Section 4.1.1). Guided by the results, we used all 58 features and further included the information of neighboring residues within the features of each residue by using a sliding window, keeping the target residue at the center of the window. For this application, window size 9 gave the best performance, reported in Section 4.1.2. Moreover, we applied a 3rd degree polynomial function kernel(check Section 3.1 for the details), which made the total number of features per residue: $58 \times 9 \times 3 = 1566$.

2.3. Performance evaluation metrics

Our proposed tool provides both real and binary space prediction outputs which were evaluated and compared using different set of metrics.

2.3.1. Evaluation measures of real value prediction

We calculated two measures named Mean Absolute Error (MAE) and Pearsons Correlation Coefficient (PCC). We defined a multi-objective function as in Iqbal et al. (2015) to combine the effects of both MAE and PCC. Our target is to achieve high performance by ensuring low MAE and high PCC. So the equation of the multi objective function is defined as follows:

$$OBJ = PCC + (1 - MAE) \tag{4}$$

where,

$$PCC = \frac{\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})(ASAp_i - \overline{ASAp})}{\sqrt{[\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})^2][\sum_{i=1}^{N}(ASAp_i - \overline{ASAp})^2]}} \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{N}|ASAr_i - ASAp_i|}{N} \tag{6}$$

In Eqs. (5) and (6), N is the total number of residues in both train and test datasets combined.

2.3.2. Evaluation measures for binary prediction of RSA

To evaluate the prediction quality of binary classification of RSA by our predictor, we have used a set of 4 metrics defined in Lin et al. (2014). The following set of four equations was inspired by the formulation used by Chou (2001), which provides a more intuitive method to identify the quality of a predictor. According to Chou's formulation, the sensitivity, specificity, overall accuracy and matthews correlation coefficient can be expressed as follows (Chen et al., 2013; Qiu et al., 2014; Xu et al., 2013a; 2013b).

$$\begin{cases} Sensitivity, Sn = 1 - \frac{N_-^+}{N^+} & 0 \le Sn \le 1 \\ Specificity, Sp = 1 - \frac{N_+^-}{N^-} & 0 \le Sp \le 1 \\ Accuracy, ACC = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le ACC \le 1 \\ MCC = \frac{1 - (\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-})}{\sqrt{(1 + \frac{N_+^- - N_-^+}{N^+})(1 + \frac{N_-^+ - N_+^-}{N^-})}} & -1 \le MCC \le 1 \end{cases} \tag{7}$$

We have labeled the buried state of a residue as positive and exposed state as negative. Based on this assumption, the quantities used in Eq. (7) are defined as follows:

N_-^+ = Total number of buried residues inccorrectly predicted as exposed residue.

N_+^- = Total number of exposed residue inccorrectly predicted as buried residue.

N^+ = Total number of buried residues investigated.

N^- = Total number of exposed residues investigated.

In addition to the 4 metrics introduced in Eq. (7), we measured two other statistical quantities namely precision and F1 score. Precision indicates the proportion of positive predictive value (PPV)

and F1 score is the harmonic mean of sensitivity and precision as suggested by equation Eqs. (8) and (9).

$$Precision(PPV) = \frac{N^{+} - N^{+}_{-}}{N^{+} - N^{+}_{-} + N^{-}_{+}} \tag{8}$$

$$F1\ \ score = \frac{2 \times (Sensitivity \times Precision)}{Sensitivity + Precision} \tag{9}$$

From Eq. (7) we can find out that, the sensitivity of a predictor refers to the proportion of correctly predicted positive values which in our case is buried residues. Similarly, specificity of a predictor indicates the proportion of correctly predicted negative class of values which is exposed residue classification in our study. The accuracy denotes the overall correct prediction of both positive and negative class values. So, if $N^{+}_{-} = N^{-}_{+} = 0$, then Eq. (7) suggests that the ACC will become 1. The Matthews correlation coefficient MCC is used for measuring the quality of binary (two-class) classifications. The value of MCC spans the range from -1 to $+1$, where $+1$ means perfect prediction, 0 means random prediction and -1 means totally different result of prediction from observation. The 4 metrics stated in Eq. (7) explian the behavior of the predictor without any complications specially for its Mathew's correlation coefficient. We have used a single-label system in our work which gives only one label (positive or negative) to a single residue, hence Eq. (7) can be applied appropriately in assessing the quality of our predictor. For multi-label systems (Chou and Shen, 2007; Chou et al., 2012; Shen and Chou, 2007; Xiao et al., 2013), a set of more complicated metrics should be used as suggested in Chou (2013).

3. RBSURFpred framework

The proposed RBSURFpred tool predicts the per-residue ASA from protein sequence as well as classifies each residue as either buried or exposed. In this section, we describe the methodology of RBSURFpred.

3.1. Real-value prediction of ASA

As established in REGAd^{3}p tool (Iqbal et al., 2015), we built the real-value predictor model using regularized exact regression. However, we carefully checked different metaheuristics for optimizing the weights and used differential evolution algorithm unlike REGAd^{3}p (Iqbal et al., 2015) that uses genetic algorithm. The equation to compute the weights by exact regression (Hastie et al., 2001) is:

$$\beta = (X^T X)^{-1} X^T Y \tag{10}$$

where,

X = input feature matrix with dimensions of $N_{residue} \times N_{feature}$
$N_{residue}$ = Number of residues in training dataset
$N_{feature}$ = Number of features per residue
X^T = Transpose of the feature matrix X
Y = Matrix of all the actual values of ASA of each residue
β = Weights of the model determined by the equation

With the weights, we computed the ASA using the following equation :

$$\hat{Y} = X\beta \tag{11}$$

where, $\hat{Y}$ = predicted value of ASA of a residue and X = feature vector of a residue

However, Eq. (11) is for basic linear regression model. We extended the kernel of this regression method to degree 3 polynomial within the feature matrix using basis expansion. We inserted two extra column vectors for each features which are the squares and cubes of the original feature values as suggested in Iqbal et al. (2015). This extension is expressed by the following equation, where p is the number of features.

$$X = [1\ x_1 x_2 x_3 \ldots x_p] \tag{12}$$

$$X^3 = [1\ x_1\ x_1^2\ x_1^3 \ldots x_p\ x_p^2\ x_p^3] \tag{13}$$

Here, X^3 is the extended feature matrix where each feature x_i is accompanied by its square and cube value.

The extension of the kernel gave us better results than the linear model. However, increasing the degree of polynomial can cause overfitting in the model due to highly fluctuating weights. An overfitted model towards training data can give poor prediction on test dataset. To overcome this problem, we implemented regularization as suggested in Iqbal et al. (2015). To implement regularization, we introduced a penalty term in the error estimate to shrink the value of the weights. The equation of REGAd^{3}p including regularization is as follows:

$$\beta = (X^T X + \lambda M_{\lambda_{(p+1)(p+1)}})^{-1} X^T Y \tag{14}$$

where,

λ = regularization parameter to control the coefficients of the model and

M_λ = identity matrix of dimension $(p+1) \times (p+1)$ with the first diagonal element equal to zero to avoid affecting the bias term.

To find the best value of λ, we experimented with different values of λ ranging from -100 to $+100$ with a step size of 2. We compared all the results from 100 weight sets obtained by using 100 different λ as stated above and recorded the best result as non-optimized result of the model. This set of 100 weights was later used as seeds for the different metaheuristics that we explored to optimize our model.

3.1.1. Optimization by metaheuristics

Here, we briefly discuss the four metaheuristic techniques that we explored to optimize the initial weight set generated by the regression so that it can better fit the training points. These metaheuristics are population-based optimization algorithms that start with a population of individuals (or, candidate solutions) and then try to find an optimal solution by tweaking the available solutions in multiple iterations by operations, such as crossover and mutation. For this application, each individual is a real valued vector of dimension equal to the number of weights. We used 58 features with an optimized window size of 9 and a kernel of degree 3. So there are in total $58 \times 9 \times 3 = 1567$ elements or float values in one vector or individual. From the 100 weight sets (individuals), we have chosen 20 weight sets based on the PCC and MAE values as our initial population in our meta-heuristic algorithms. We also experimented with different number of weight sets as our initial population, but the results were similar. So we opted for the best twenty individuals as our initial population. As the number of features and the total number of residues in the dataset are lot higher, so it takes a lot of time to finish one iteration of these algorithms. As a result, we limited our algorithms to 50 iterations. Below we present a brief discussion on four metaheuristics we implemented for optimization.

Genetic algorithm: The parameter values of our genetic algorithm (Holland, 1992) implementations are: (*i*) population size = 50, (*ii*) elitism rate = 10%, (*iii*) mutation rate = 10% and (*iv*) crossover rate = 80%. We implemented roulette wheel selection method to select individuals for crossover operation. We performed one point crossover, two point crossover, uniform crossover and recombination. The two point crossover operation worked best for our case. The reason behind this could be that the two point crossover eliminates the problem of possible linkage among the

first and last elements of a real valued vector called epistasis (Beasley et al., 1993), caused by the one-point crossover. Finally, we performed mutation by implementing gaussian convolution on floating-point weight values. Gaussian convolution is controlled largely by the distribution variance σ^2, which is known as the mutation rate and determines the noise in the mutate operation (Luke, 2013). We selected a low variance of 0.0004 due to the nature of the problem under consideration and its solution space. In particular, in our experiments, a little change in the weights caused a huge deviation in the results. We limited the mutation to 100 randomly selected values over the range of 1567 values in an individual because a large number tweaks in an individual can cause high fluctuation due to excessive mutation.

Differential evolution: Differential evolution (DE) determines the size of mutates largely based on the current variance in the population (Luke, 2013). If the population is spread out, mutate will make major changes. If the population is condensed in a certain region, mutates will be small. It is called an adaptive mutation algorithm (Abbass, 2002).

DE's mutation operators employ vector addition and subtraction, so it only works in floating point vector spaces, which is why we have applied this metaheuristics on our model. For each member $\vec{i}$ of the population, we generate a new child by picking three individuals from the population and performing some vector additions and subtractions among them. The idea is to mutate away from one of the three individuals ($\vec{a}$) by adding a vector to it. This vector is created from the difference between the other two individuals $\vec{b}-\vec{c}$. If the population is spread out, $\vec{b}$ and $\vec{c}$ are likely to be far from one another and this mutation vector is large, otherwise it is small. This way, if the population is spread throughout the space, mutations will be much bigger than when the algorithm has later converged on fit regions of the space.

Spatial breeding: Spatial breeding was also one of our implemented metheuristics. Spatially embedded models not only promote exploration and diversity in the population but also brings the notion of physical locations of individuals in the population (Luke, 2013). For example, the population may be laid out in a 3D grid, or a 1D ring, and each individual occupies a certain point in that space. Such models are mostly used to maintain diversity in the population, and so promote exploration. Individuals are only allowed to breed with "nearby" individuals, so a highly fit individual cannot spread as rapidly through a population as it could if there were no breeding constraints. We organized our population in a 2D grid based on the fitness of the individuals. While selecting an individual for crossover with base individual, we chose one individual from the neighborhood of the base individual randomly. The neighborhood of an individual is defined as maximum possible four positions around it i.e. immediate left,right,up and down. This method of choosing individual for tweaking helps the population to grow in a diverse way than any other algorithm.

The (μ, λ) evolution strategy: The (μ, λ) strategy is the simplest of all evolution strategies (ES) (Beyer and Schwefel, 2002; Luke, 2013). ES employs a simple procedure for selecting individuals called truncation selection, and (usually) only uses mutation as the tweak operator. We began our implementation with λ number of (seeded) individuals for our algorithm. From this population μ fittest individuals are chosen as parents. We deleted from the population all but the μ fittest individuals and each of them generated λ/μ individuals to form the next population of next iteration. The size of λ controls the size of the population. The higher the value of λ, the more explorative the search will be. The weights to be optimized in our study were highly fluctuating and generated degraded results whenever the mutation was large or the search was explorative. So to keep it less explorative and more exploitative, we have chosen a value of 10 for λ. This smaller value of λ also helped in reducing the time for computation as each iteration takes a fair amount of time to complete. The size of μ on the other hand controls how selective the algorithm is; low value of μ with respect to λ pushes the algorithm more towards exploitative search as only the best individuals will survive. So to keep a balance between exploration and exploitation in this case, we chose a value of 5 for μ which is just half of the value of λ, thus preserving both the search characteristics. The degree of mutation performed was also kept low due to the fluctuating behavior of the weight vector as mentioned earlier.

We evaluated the performance of these 4 optimization techniques in Section 4.1.3 and found that the differential evolution works the best for this application. Thus, we used this technique in the final RBSURFpred framework. The overall process of predicting ASA with our new RBSURFpred predictor is graphically depicted using flow-chart in Fig. 2.

3.2. Binary prediction of ASA

In this work, we have not only performed real-value prediction but also performed binary prediction of ASA based on appropriate threshold of RSA values. Intuitively, the RSA value of a residue is an indicator of the percentage of the residue surface area that is exposed. RSA of a protein residue is calculated by normalizing the ASA of that residue (X) by the surface area of the same type of residue in a reference state. We used the ASA normalizing values from Tien et al. (2013) using Gly-X-Gly tripeptide as the reference state for a given residue X. The authors in Tien et al. (2013) recommend to use their theoretical MaxASA values as they were obtained from a systematic enumeration of all possible conformations and likely to represent a true upper bound of observable ASA values. We have calculated Relative Solvent Accessibility (RSA) of each residue in our entire dataset with the following formula:

$$RSA = \frac{Predicted\ ASA\ of\ a\ residue}{ASA\ in\ the\ Gly-X-Gly\ conformation} \tag{15}$$

Now, to do binary prediction we have chosen two classes for a residue namely: Buried (true) and Exposed (false). We have classified each residue in either of these two classes based on RSA values of each residue by applying a threshold of 0.18.

3.2.1. Choice of exposure threshold for binary prediction

Our predictor classifies residues into buried or exposed by means of the exposure threshold on RSA values associated to query residues. In this work, we set the value of the exposure threshold to 18%. This value of threshold has been chosen so that it allows an even distribution of residues, with respect to solvent accessibility (RSA) value, of the sequences in the considered dataset (Rose et al., 1985). This threshold value has often been considered as a reference in later works (Adamczak et al., 2004; Carugo, 2000; Garg et al., 2005; Gianese et al., 2003; Pollastri et al., 2002; Rost and Sander, 1994; Thompson and Goldstein, 1996). There are 272,041 residues in our entire dataset. By choosing a threshold of 0.18 we have almost evenly classified those residues in buried and exposed classes.

4. Results and performance analysis

In this section, we represent the results obtained from the proposed predictor. We further analyze and compare the results with the exiting predictor.

4.1. Results of real-value ASA prediction

4.1.1. Evaluations of features

We added three new variables in our feature set than that used by REGAd^{3}p tool that are: PSEE, $\Delta\Phi$ and $\Delta\Psi$. To understand the

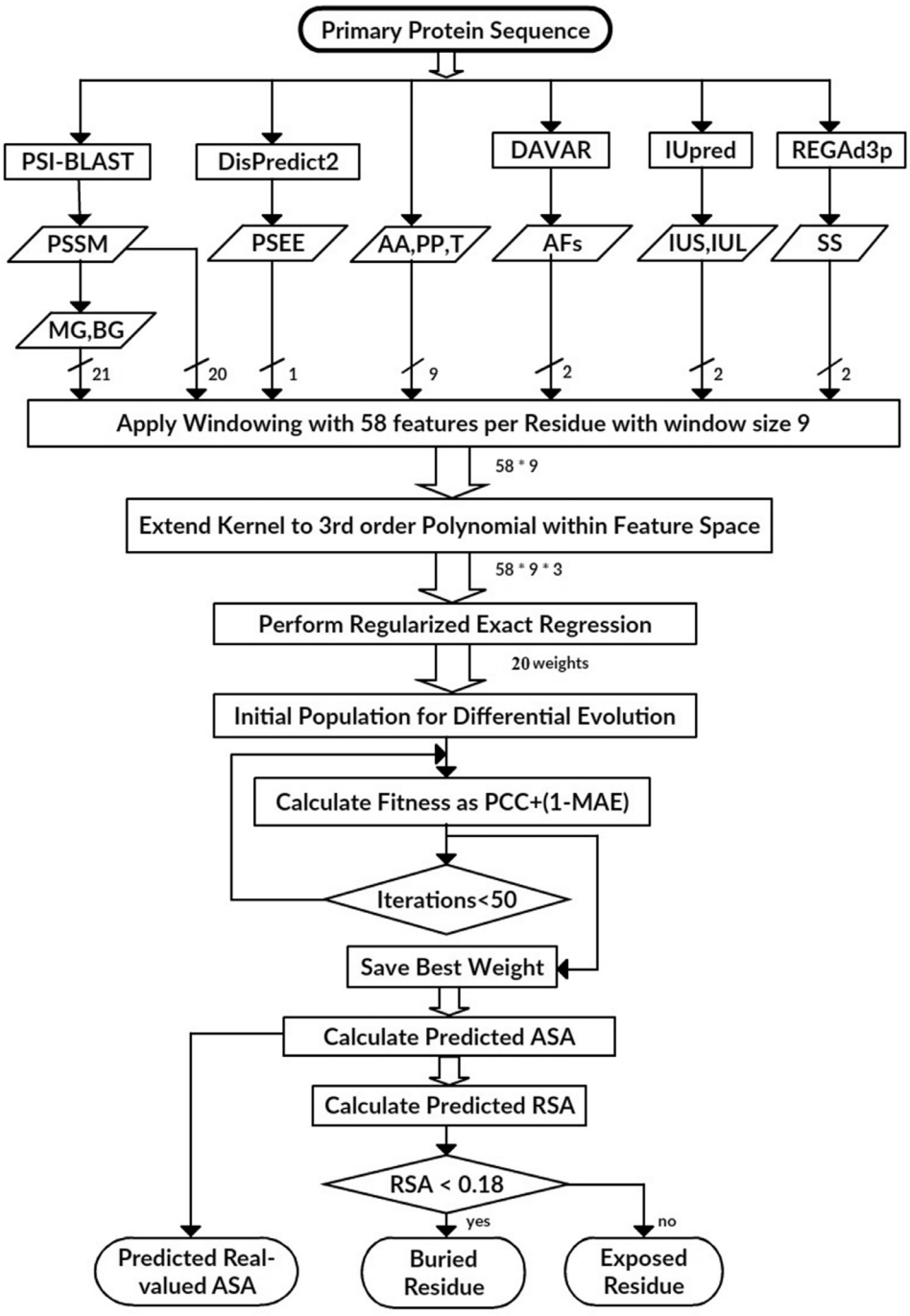

Fig. 2. Overview of RBSURFpred framework including feature generation, real and binary state predictions, and optimization.

Table 1
Correlation coefficients of 3 newly added features with the existing ones.

	AA(1)	PP(7)	PSSM(20)	MG(1)	BG(20)	SS(3)	IUS(1)	IUL(1)
PSEE	−0.04	−0.17	−0.003	−0.44	0.68	0.27	0.05	−0.15
$\Delta\Phi$	0.002	−0.20	−0.10	−0.52	0.78	0.29	0.05	−0.11
$\Delta\Psi$	−0.17	−0.37	0.02	−0.22	0.54	0.21	0.05	−0.01

uniqueness of these features we computed Pearson Correlation Coefficient (PCC) of each of these three features against the rest of the 55 features and listed them in Table 1. It ranges from −1 to 1. The higher the coefficient, the more alike two variables are, the lower the coefficient the more unlike they are. At first, we computed the PCC of these three features against rest of the 55 features for all the sequences in our dataset and averaged them. We further averaged the PCC values for features having more than one values such as PP(7), PSSM(20), BG(20) etc.

From the results reported in Table 1, we can observe that all these newly added features have a slightly high correlation with monogram (MG) and bigram (BG) values. However, they have very

Table 2
List of features used in ASA prediction according to different feature plan.here √ and – imply that the corresponding feature-set is included and excluded, respectively in the feature-plan.

Feature description	Feature count	Plan #1	Plan #2	Plan #3	Plan #4	Plan #5
Amino acid (AA)	1	√	√	√	√	√
Physical Properties (PP)	7	√	√	√	√	√
Position specific scores matrix (PSSM)	20	√	√	√	√	√
Monogram (MG)	1	√	√	√	–	–
Bigram (BG)	20	√	√	√	–	–
Short and long probabilities (IUS/IUL)	2	√	√	√	–	√
Secondary structural probabilities (SS)	3	√	√	√	√	√
Position specific estimated energy (PSEE)	1	√	–	–	–	√
Torsion Angle fluctuation ($\Delta\Phi$)	1	–	√	–	–	√
Torsion Angle fluctuation ($\Delta\Psi$)	1	–	–	√	–	√
Terminal tag (T)	1	√	√	√	√	√
Total feature count	58	56	56	56	32	37

Table 3
Prediction quality of ASA for five different feature plans each using 3rd order polynomial as kernel.

Dataset	SSD_TR1OO1		SSD_TS295	
measures	MAE	PCC	MAE	PCC
Plan #1. non-optimized	23.29	0.744	23.53	0.742
Plan #1. optimized	22.92	0.747	23.23	0.745
Plan #2. non-optimized	23.28	0.744	23.52	0.745
Plan #2. optimized	22.94	0.746	23.21	0.747
Plan #3. non-optimized	23.30	0.744	23.54	0.745
Plan #3. optimized	22.98	0.745	23.24	0.746
Plan #4. non-optimized	23.75	0.73	23.70	0.742
Plan #4. optimized	23.45	0.734	23.35	0.743
Plan #5. non-optimized	23.63	0.736	23.61	0.742
Plan #5. optimized	23.35	0.738	23.29	0.744

low correlation with rest of the features. To further understand the importance of these three features, we prepared various feature combinations as depicted in Table 2. Plan#1, plan#2 and plan#3, respectively, include PSEE, $\Delta\Phi$, and $\Delta\Psi$ with the 55 features, therefore separately evaluate the importance of these 3 variables. In addition to that, we created plan#4 which includes only one dimensional sequence information along with structural probabilities. Finally, we performed predictions using plan#5 that includes all the features except monograms and bigrams to show how much additive information the three new features are providing in spite of having a high correlation with monograms and bigrams. Table 3 shows the performance of the predictor using these 5 feature plans. As has been mentioned before, throughout our experiments, we have used the dataset SSD_TR1001 for training and dataset SSD_TS295 for testing.

From the results in Table 3, we can follow that each of the new added features have equally contributed to the performance of the predictor and improved the MAE by a margin of 3.2% and PCC by 1.1% individually. Plan#4 consists of only one-dimensional information of protein along with structural probabilities. The results show that without the three dimensional information, i.e., monograms, bigrams, disorder probabilities, and the 3 new features, the predictor suffers a lot of performance. The results of plan#5 show that despite having a high correlation with the added features both MG and BG are still needed for high accuracy of the predictor. Therefore, we used all 58 features for the proposed RBSURFpred predictor.

4.1.2. Evaluations of different window sizes

Here, we search for a suitable value of the sliding window size. The value of window size approximates the number of residues that may form the necessary local environment around a target residue. We evaluated the performance of our predictor on real-value ASA prediction using 10 different window sizes (3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23). Table 4 shows the performance on SSD_TR1001 and SSD_TS295 dataset for both with and without optimization. The best results are shown in bold which were achieved with window size 9. Therefore, we adopted windowing with 4 residues on the either side of the target residue in the final predictor framework. Notably, this window size is different than the one found to be best for REGAd^{3}p (Iqbal et al., 2015).

4.1.3. Selection of optimization technique

Here, we report the results that we found using 4 different metaheuristic techniques, genetic algorithm (GA), different evolution (DE), spatial breeding (SB) and evolutionary strategy (ES), to optimize the weight set generated by the regularized regression. Table 5 shows the results. Among the four metaheuristics, DE worked best for us. We achieved best optimization of our model using DE in very few iterations (8–12). The reason behind the top performance of differential evolution is that it is an adaptive mutation algorithm and it can successfully control the change in the individuals and quickly converge to the global optima through the search space. The use of DE has optimized our prediction accuracy by 2.7% in terms of MAE and 2.5% in terms of PCC for training dataset and for test dataset the improvement is 3.2% in terms of MAE and 1.35% in terms of PCC. Thus, we used DE for optimization in the proposed RBSURFpred model.

4.1.4. Comparison RBSURFpred with other predictors

We compared the prediction performance of RBSURFpred with REGAd^{3}p that uses a similar framework including regularized regression with degree-3-polynomial kernel.

We report the performance on both training dataset (SSD_TR1001) and test dataset (SSD_TS295) in Table 6. The result show that the RBSURFpred (optimized) model gave higher PCC on both the dtasets than those given by the non-optimized model. Moreover, we can observe that the addition of three new features in our work improved the performance than that of REGAd^{3}p by reducing the MAE by a margin of 13.22% and improving the PCC by 7.15% on training dataset. In case of test dataset, it improved the MAE and PCC by 7.5% and 2.22%, respectively.

To compare the proposed predictor with another top performing predictor, SPIDER2, we downloaded the software and ran it on both SSD_TR1001 and SSD_TS295 datasets. From the results reported in Table 6, we can observe that RBSURFpred has outperformed SPIDER2 and made better prediction in terms of MAE in case of both training and test datasets. The PCC measure for training dataset has been almost the same for both predictors. SPIDER2 has only been able to outperform RBSURFpred in case of PCC of test dataset. So, the overall comparative performance of RBSURFpred versus SPIDER2 is noteworthy. If we consider a multi objective function as the performance measure such as PCC + (1 −

Table 4
Analysis of RBSURFpred using various window sizes for the search of optimal window size.

Window size	Non-optimized				Optimized			
	SSD_TR1001		SSD_TS295		SSD_TR1001		SSD_TS295	
	MAE	PCC	MAE	PCC	MAE	PCC	MAE	PCC
3	23.41	0.73	23.37	0.735	23.14	0.74	23.2	0.742
5	23.27	0.73	23.31	0.73	22.95	0.74	23.1	0.74
7	22.97	0.742	23.09	0.741	22.81	0.741	22.91	0.742
9	**22.85**	**0.732**	**23.04**	**0.741**	**22.25**	**0.75**	**22.30**	**0.75**
11	22.75	0.743	23.05	0.741	22.35	0.75	22.53	0.741
13	22.72	0.748	23.06	0.742	23.21	0.745	22.70	0.744
15	23.18	0.738	23.31	0.732	22.38	0.748	22.54	0.748
19	23.25	0.734	23.29	0.741	22.55	0.741	22.61	0.742
21	22.65	0.742	23.25	0.738	22.11	0.75	22.65	0.741
23	38.47	0.08	38.57	0.08	38.11	0.09	38.21	0.10

Bold indicates best obtained values.

Table 5
Results of different metaheuristics applied on newly proposed RBSURFpred.

Metaheuristics	SSD_TR1OO1		SSD_TS295	
	MAE	PCC	MAE	PCC
Genetic algorithm	22.54	0.742	22.72	0.743
Differential evolution	**22.25**	**0.75**	**22.30**	**0.75**
Spatial breeding	22.57	0.742	22.77	0.734
Evolutionary strategy	22.63	0.74	22.75	0.738

Bold indicates best obtained values.

Table 6
List of predictions of ASA for different predictor frameworks.

Dataset	SSD_TR1OO1		SSD_TS295	
measures	MAE	PCC	MAE	PCC
RBSURFpred (non-optimized)	22.85	0.732	23.04	0.741
RBSURFpred (optimized)	**22.25**	**0.75**	**22.30**	0.75
REGAd^{3}p	25.19	0.702	23.97	0.734
SPIDER2	23.3	0.75	22.89	**0.77**

Bold indicates best obtained values.

Table 7
Performance of RBSURFpred and SPIDER2 on Moulder dataset.

Independent dataset measures	Moulder MAE	PCC
RBSURFpred	**23.97**	0.73
SPIDER2	24.12	**0.74**

Bold indicates best obtained values.

MAE), then RBSURFpred achieved higher performance than SPIDER2 over the course of both datasets. Thus, RBSURFpred performed with accuracy compared with the state-of-the-art ASA predictors.

To test the performance of RBSURFpred further, we collected another dataset named Moulder as used in Iqbal et al. (2015). It is a challenging decoy dataset consisting of 20 native proteins with 300 comparative decoy models generated using homologous template for each protein. The results of our predictor vs SPIDER2 is listed in Table 7. The results show that RBSURFpred is competitive with SPIDER2. RBSURFpred achieved a minimum MAE of 23.97 between them whereas SPIDER2 achieves higher PCC. Therefore, the performance of RBSURFpred on a blind dataset shows the robustness of the proposed model.

4.1.5. Case study of individual proteins

For further comparison between RBSURFpred and SPIDER2, we have selected two protein chains, (*i*) PDB ID: 2EI5 (chain: B) and (*ii*) PDB ID: 7FD1 (chain: A). We plotted the residue wise predicted ASA by RBSURFpred and SPIDER2 along with the actual values calculated using DSSP in Figs. 3 and 4. Both of the predictors failed to predict the ASA correctly when the values are too high or too low. However, RBSURFpred predicted better in some ranges of residues than SPIDER2 for example, for the residue index: 5–15, 23, 25–31, 57–64, 73, 86–88, 91–94, 97–102, 103 of protein chain 2EI5B. Over the specified ranges, RBSURFpred could get much closer with the actual ASA values from DSSP even if the values were too low or high.

In case of protein chain 7FD1A (Fig. 4), the overall performance of RBSURFpred is much better than SPIDER2 over most of the residue ranges. Some of these ranges, where RBSURFpred's performance is noticeably better in Fig. 4, are indices 3–9, 10–17, 19, 30–34, 49–57, 64–68, 102–106 etc. In both Figs. 3 and 4, we mentioned the sequence wise MAE and PCC measures of both the predictors reported in the top right corner of the plots. In case of 2EI5B, the scores of RBSURFpred are slightly better than SPIDER2 but in case of the sequence 7FD1A, RBSURFpred outperformed SPIDER2 by quite a large margin.

Table 8
Comparison of binary prediction performance given by different predictor frameworks.

Performance measures	RBSURFpred	SPIDER2	REGAd^{3}p
Sensitivity	**0.74**	0.63	0.60
Specificity	0.84	**0.92**	0.90
Accuracy	**0.79**	0.77	0.75
MCC	**0.58**	**0.58**	0.53
Precision	0.83	**0.88**	0.81
F1 score	**0.78**	0.74	0.71

Bold indicates best obtained values.

4.2. Results of binary ASA prediction

The results of 2-state predictions, buried (positive) and exposed (negative), are reported in Table 8. We evaluated all the three predictor frameworks, i.e. REGAd^{3}p, RBSURFpred and SPIDER2 for binary classification of the residues as buried or exposed. We performed the binary prediction on SSD_TS295 dataset which consists of 61,074 residues in total. Among them, the total number of buried residues, N^+ is 31,218 and the total number of exposed residues, N^- is 29,856. By applying RBSURFpred on SSD_TS295 dataset, we found out the total number of buried residues predicted correctly are 22,990, total number of buried residues predicted incorrectly i.e. N^+_- is 8,228, total number of exposed residues predicted correctly is 24,962 and total number of exposed residues predicted incorrectly i.e. N^-_+ is 4894. We can observe from Table 8 that RBSURFpred outperformed REGAd^{3}p in all of the five performance measure categories except for the specificity score. On the other hand, comparing the results of RBSURFpred with SPIDER2 we can

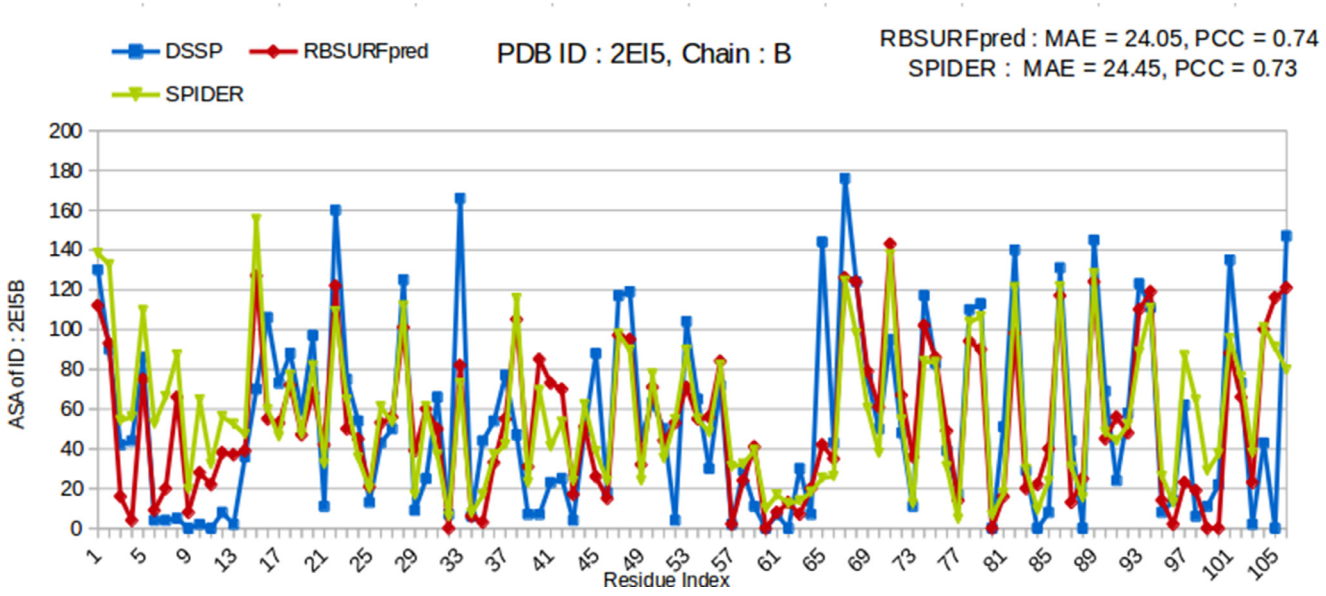

Fig. 3. Comparison of residue wise ASA values for PDB ID: 2EI5, Chain: B. The *x*-axis and *y*-axis shows the residue index and ASA values, respectively.

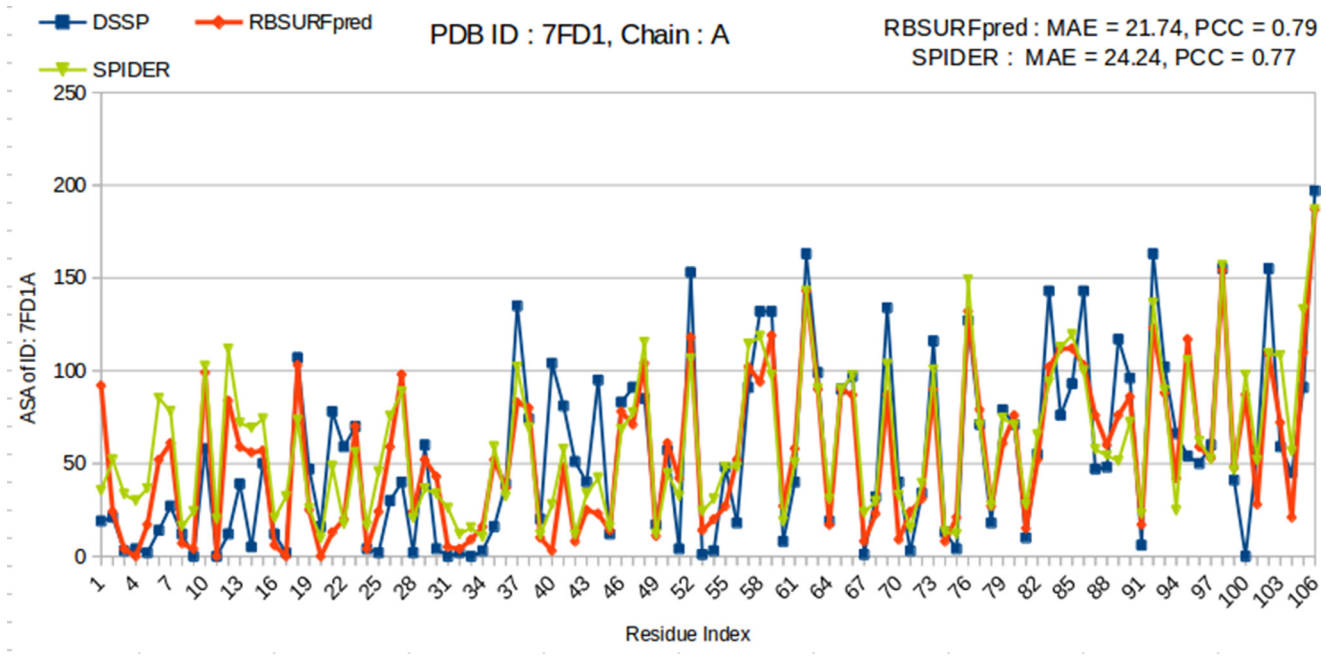

Fig. 4. Comparison of residue wise ASA values for PDB ID: 7FD1, Chain: A. The *x*-axis and *y*-axis shows the residue index and ASA values, respectively.

see that RBSURFpred has achieved a better recall, accuracy and f1 score, a similar MCC score, and a lower specificity and precision than those of SPIDER2.

According to Wu et al. (2017),the existing methods for the prediction of ASA are unbalanced and fails to predict potential exposed state of the residues than the buried residues. One potential problem associated with the existing prediction methods may be the unbalanced training sets. Prediction of ASA requires large non-redundant training sets. But the tools that collect this training set reserve the longest sequences to represent a clustered group, while shorter sequences are removed from the training sets. Differing from other one-dimensional structural characteristics, RSA value is impacted not only by its own orientation and that of its neighbors, but also by other residues located elsewhere in the protein structure. Due to spatial contacts, a residue within a longer sequence is more easily buried relative to one found in a shorter sequence. So, for a long sequence of residues there is a higher chance of a residue to be buried and so the prediction of 2-state will also give correct result for buried residue. Thus, an unbalanced training set with high percentage of long length sequences generally fails to predict exposed residue state correctly and a training set with low percentage of long length or high percentage of short length sequences fails to predict buried residue state correctly.

The training set we used namely SSD_TR1001 has in total 1001 protein IDs. We made three groups of these sequences based on their chain lengths. They are named short-length (40–100), medium-length(101–250) and long-length (>250). There are 105 sequences in short-length group, 571 sequences in medium-length group and 325 sequences in long-length group. Therefore, we included a small amount of short-length sequences and failry large amount of long length sequences in the training set. We observed the significance of such a training dataset in the outputs of RB-

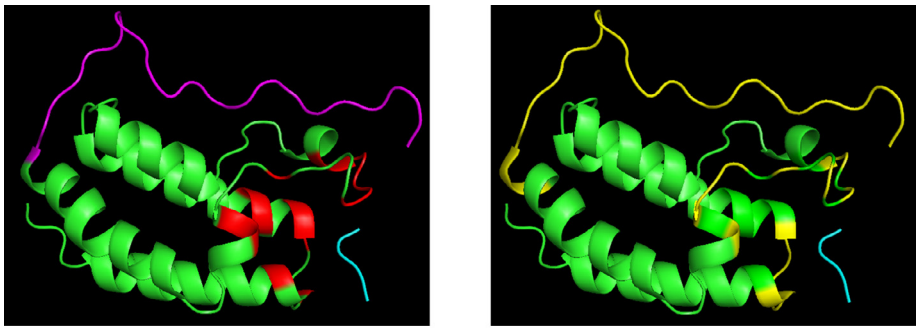

Fig. 5. Case study with PDB ID: 3JVK with crystal structure of bromodomain 1 of mouse Brd4 (green) in complex with histone H3-K (cyan). In (a), the N-terminal region with coil residues and peptide-binding residues are highlighted in pink and red. The plot in (b) shows the predicted exposed residues (yellow) in the two regions that are marked in (a). The images are generated using PyMOL (https://pymol.org/). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

SUFpred. Specifically, the sensitivity score (true positive rate) given by RBSURFpred shows that the rate of correctly predicted buried residues is 17.5% higher than that of SPIDER2 at a cost of only 9.1% lower specificity (true negative rate), indicating the rate of correctly predicted exposed residues. This lower value of specificity also justifies the relatively lower precision given by RBSURFpred because of higer value of falsely predicted exposed residue. However, the best performance RBSURFpred in terms of F1 score, which is the hermonic mean of precision and sensitivity, endorses the well-balanced prediction ability of the proposed tool.

4.2.1. Case study of individual proteins

In this section, we apply RBSURFpred on biologically relevant test proteins and study the binary prediction performance of the proposed tool. Especially, we focus on the ability of the predictor in recognizing exposed residues of critical regions in proteins, such as disordered regions, terminal regions with marginal secondary structure and peptide-binding regions with larger interaction surface.

PDB ID - 3JVK: Fig. 5 shows the crystal structure of the bromodomain 1 of mouse Brd4 protein (green) bound to a peptide (cyan) with acetylated lysine, H3-K(ac). In Fig. 5a, we highlight the extended coil-like N-terminal region of the Brd4 protein (chain: A) in pink and the region that binds to the acetylated peptide in red. We label the residues that stay within the 6 Angstrom distance from a residue of the peptide as interaction or binding with that peptide. Both the coil-like region and peptide-binding region are likely to be composed of exposed residues. The terminal region with loose secondary structure (coil-like) is usually highly flexible. Moreover, a binding region is potentially composed of exposed residues that can recognize peptides and form transient interaction with peptides. It is important to identify such peptide-binding residues that can recognize crucial phosphorylated peptides like the one discussed here. Fig. 5b shows the prediction output of RBSURFpred where we highlight the predicted exposed residues in yellow in the two regions discussed above. The two plots in the figure shows that the RBSURFpred tool was quite effective in predicting the exposed residues in two crucial regions.

PDB ID - 3PQZ: Here, we picked the growth factor receptor-bound protein 7 (grb7) with SH2 domain (chain: A) bound to a peptide to study the usefulness of RBSURFpred in predicting exposed residues in critical regions. The PDB files store the disordered residues information (REMARK 465) which are the residue with missing coordinates in X-ray crystallography, thus have no structure. The Grb7 protein in the structure under consideration has two disordered regions, an 8-residue-long (indices: 1–8) in N-terminal and a 4-residue-long (indices: 114–117) C-terminal regions. The disordered regions are highly flexible and adopt heterogeneous conformations through interactions with different partners using the constituent exposed residues. The output of RBSURFpred on this protein shows that the tool predicted the disordered region in the N-terminal (indices: 1–7) as exposed residues as well as labeled 3 residues of the 4-residue-long disordered regions in the C-terminal as exposed. This study shows effectiveness of the proposed tool in identifying exposed residues in disordered regions.

4.3. Analysis of prediction using RBSURFpred

In this section, we organized the analysis on the prediction of ASA by RBSURFpred. We performed analysis for real value prediction on SSD_TS295 dataset.

4.3.1. Amino acid specific analysis

The analysis with respect to 20 different amino acids is performed to test whether RBSURFpred has any significant under prediction or over prediction problem for any of the residues or not. The analysis is performed with some statistical measures such as mean actual ASA, mean predicted ASA, standard deviation of actual ASA and MAE of prediction. The predicted and actual ASA values for each amino acid is highly correlated, with a PCC value equal to 0.999.

From this analysis in Fig. 6, we can note that for some amino acids such as Alanine(A), Asparagine(N), Histidine(H) the mean prediction is exactly same with the mean actual ASA values of these amino acids. For the rest of the amino acids, some of them has over prediction and some of them has under prediction, but all of these deviations in prediction are of small margins.

4.3.2. Length specific analysis

Here, we conducted an analysis based on the length of the protein sequences and observed how RBSURFpred can perform on protein with different chain length. We presented the analysis results in Fig. 7. For this analysis, we distributed 295 proteins of SSD_TS295 into 15 buckets or ranges of the form $[\ x_1 - x_2)$ where $x_1 <=$ chain-length $< x_2$. We observed that the overall performance of RBSURFpred is very much similar to SPIDER2. For the sequences of length $<= 200$ which constitues 45% of total dataset of SSD_TS295, the prediction performance of SPIDER2 was slightly better than RBSURFpred. But for the rest of the sequences (55%) of length > 200 RBSURFpred outperformed SPIDER2 by a fair margin. Specially from the range of [225–250) to [300–325), we can see a fair amount of spike in the MAE curve of SPIDER2 which shows a relatively lower prediction ability of SPIDER2 for sequences of

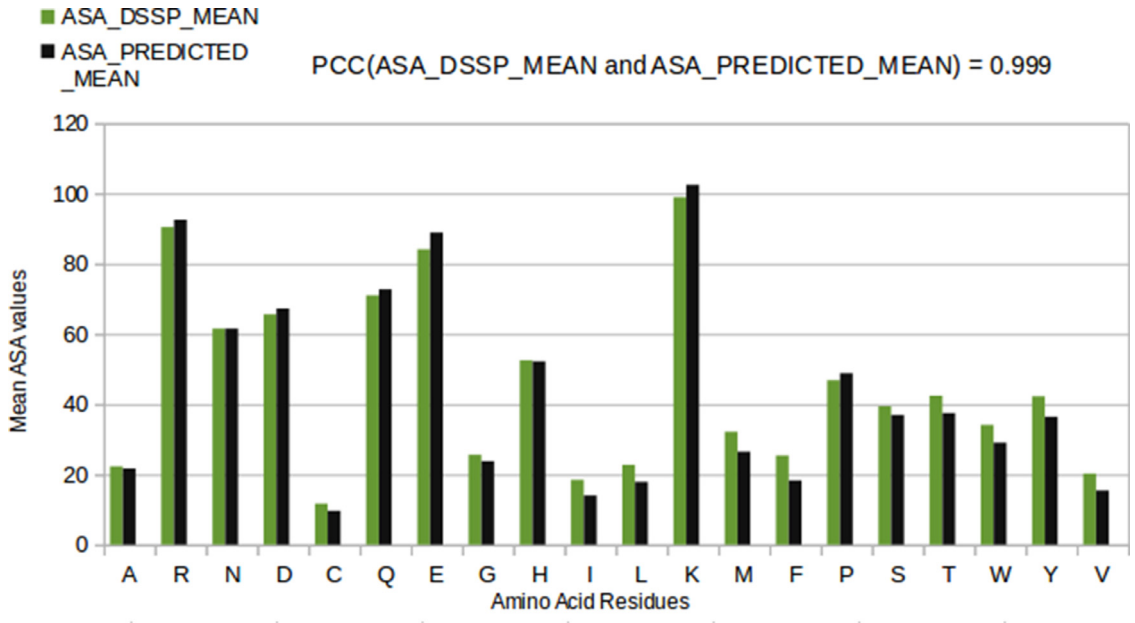

Fig. 6. Amino acid specific comparison between mean actual ASA and mean predicted ASA values. The *x*-axis and *y*-axis show the amino acid and ASA values, respectively.

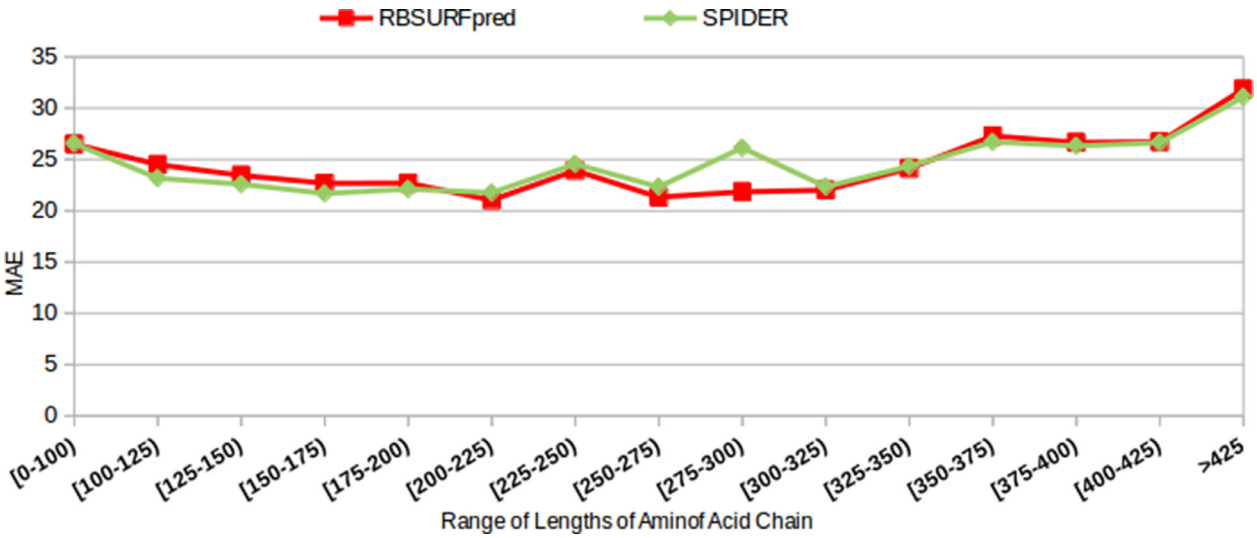

Fig. 7. Length specific analysis of predicted ASA by RBSURFpred, based on fifteen different range of length values.

large lengths. In this region of length, RBSURFpred performed better with fairly low range MAE measures($<= 22.5$). Thus, overall we can say that the prediction accuracy of RBSURFpred is competitive for a protein chain of any length or any number of residues.

5. Conclusions

In this work, we presented a new framework for protein ASA prediction that outputs per-residue ASA as well as classifies each residue as exposed or buried by applying a threshold on the normalized ASA (relative solvent accessible surface area). The proposed tool named RBSURFpred is built using the regularized exact regression technique with higher-order polynomial function as kernel to fit non-linear feature space. We have incorporated 3 important features of a protein residue to predict its exposure to solvent that have not been explored before for this application. These features are position specific estimated energy (*PSEE*) (Iqbal and Hoque, 2016) and two torsion angle fluctuations ($\Delta\Psi$, $\Delta\Phi$). PSEE is a feature that can measure a residue's stability by approximating its free energy contribution to the folded state of the protein, thus can effectively identify the structured or unstructured state of a protein. Torsion angle fluctuations measure the flexibility of protein residues in their three-dimensional structure which is related to the possible location of a residue, such as in the core or on the surface of the respective protein structure.

The performance of RBSURFpred showed that the addition of these 3 new features improved the prediction performance of an existing predictor without these 3 features. We also tried to optimize the output of the model by applying several metaheuristic algorithms i.e., genetic algorithm, differential evolution, spatial breeding and evolution strategy. The differential evolution algorithm effectively improved the performance of the predictor. Finally, the predictor resulted in promising performance when compared with two other existing state of the art predictors in the literature. Thus, we believe our proposed real and binary space surface area predictor will be useful in related applications of bioinformatics.

Finally, as demonstrated in a series of recent publications (Feng et al., 2017b; Liu et al., 2017) in developing new prediction methods or bioinformatics tools, user-friendly and publicly accessible web-servers represent the future's trend (Chou and Shen, 2009). And these web-servers significantly enhance the impacts (Chou, 2017) of the predictors. Hence our immediate future work will include establishing a web-server for RBSURFpred.

Acknowledgment

MTH gratefully acknowledges the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2016-19)-RD-B-07. This research was done as part of an undergraduate thesis work of Tarafder and Ahmed under the supervision of Rahman at the Department of CSE, BUET.

References

Abbass, H.A., 2002. The self-adaptive pareto differential evolution algorithm. Evol. Comput. 1, 831–836.
Adamczak, R., Porollo, A., Meller, J., 2004. Accurate prediction of solvent accessibility using neural networks based regression. Proteins 56, 753–767.
Ahmad, S., Gromiha, M., 2002. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18, 819–824.
Ahmad, S., Gromiha, M., Sarai, A., 2013. Real value prediction of solvent accessibility from amino acid sequence. Proteins 50, 629–635.
Beasley, D., Bull, D.R., Martin, R.R., 1993. An overview of genetic algorithms: part 2, research topics. Univ. Comput. 15, 170–181.
Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242.
Beyer, H.-G., Schwefel, H.-P., 2002. Evolution strategies–a comprehensive introduction. Nat. Comput. 1, 3–52.
Bonetti, D., Pérez-Sánchez, H., Delbem, A., 2014. An efficient solvent accessible surface area calculation applied in ab initio protein structure prediction. Proceedings of International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO.
Carugo, O., 2000. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Proteins Eng. 13, 607–609.
Chen, H., Zhou, H.-X., 2005. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins Struct. Funct. Bioinform. 61, 21–35.
Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., Chou, K.-C., 2017. IRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget 8(3), 4208-4217.
Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013. iRSpot-pseDNC: Identify Recombination Spots with Pseudo Dinucleotide Composition. Nucleic Acids Res. 41 (6), e68.
Cheng, X., Xiao, X., Chou, K.-C., 2018. Ploc-meuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general pseAAC. Genomics 110 (1), 50–58.
Chou, K.-C., 1988. Low-frequency collective motion in biomacromolecules and its biological functions. Biophys. Chem. 30, 3–48.
Chou, K.-C., 2001. Using subsite coupling to predict signal peptides. Protein Eng. 14, 75–79.
Chou, K.-C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273, 236–247.
Chou, K.-C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.
Chou, K.C., 2017. An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr. Top. Med. Chem. 17 (21), 2337–2358.
Chou, K.C., Chen, N.Y., Forsen, S., 1981. The biological functions of low-frequency phonons. 2. Cooperative effects. Chem. Scr. 18, 126–132.
Chou, K.-C., Shen, H.-B., 2007. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.
Chou, K.-C., Shen, H.-B., 2009. Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 1, 63.
Chou, K.-C., Wu, Z.-C., Xiao, X., 2012. Iloc-hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.
Cho, K.I., Kim, D., Lee, D., 2009. A feature-based approach to modeling protein–protein interaction hot spots. Nucleic Acids Res. 37, 2672–2687.
DeLano, W.L., 2001. Unraveling hot spots in binding interfaces: progress and challenges. Curr. Opin. Struct. Biol. 12, 14–20.
Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I., 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 347 (4), 827–839.
Eickholt, J., Cheng, J., 2013. DNdisorder: predicting protein disorder using boosting and deep networks. BMC Bioinform. 14, 88.
Eickholt, J., Deng, X., Cheng, J., 2011. Dobo: protein domain boundary prediction by integrating evolutionary signals and machine learning. BMC Bioinform. 12 (43).
Faraggi, E., Xue, B., Zhou, Y., 2009a. 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 74, 847–856.
Faraggi, E., Yang, Y., Zhang, S., Zhou, Y., 2009b. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. Structure 17, 1515–1527.
Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., Zhou, Y., 2012. SPINE x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Comput. Chem. 33, 259–267.
Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., Chou, K.-C., 2017a. IRNA-psecoll: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into pseKNC. Mol. Therapy – Nucleic Acids 7, 155–163.
Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., Chou, K.-C., 2017b. IRNA-psecoll: identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into pseKNC. Mol. Therapy – Nucleic Acids 7, 155–163.
Garg, A., Kaur, H., Raghava., 2005. G.p.s.: real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 61, 318–324.
Gianese, G., Bossa, F., Pascarella, S., 2003. Improvement in prediction of solvent accessibility by probability profiles. Proteins 16, 987–992.
Glaser, F., Steinberg, D.M., Vakser, I.A., Ben-Tal, N., 2001. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins Struct. Funct. Bioinf. 43, 89–102.
Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 978-0-387-84858-7 (Online).
Holbrook, S., Muskal, S., Kim, S., 1990. Predicting surface exposure of amino acids from protein sequence. Protein Eng. 3 (8), 659–665.
Holland, J.H., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press.
Iqbal, S., Hoque, M.T., 2015. Dispredict: a predictor of disordered protein using optimized RBF kernel. PloS One 10 (10), E0141551.
Iqbal, S., Mishra, A., Hoque, M.T., 2015. Improved prediction of accessible surface area results in efficient energy function application. J. Theor. Biol. 380, 380–391.
Iqbal, S., Hoque, M.T., 2016. Estimation of Position Specific Energy as a Feature of Protein Residues from Sequence Alone for Structural Classification. PloS One 11 (9), E0161452.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016a. Identification of protein–protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. 34, 1946–1961.
Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K.-C., 2016b. iPPBS-opt: a sequence-based ensemble classifier for identifying protein-protein binding sites by optimizing imbalanced training datasets. Molecules 21, 95.
Jones, D.T., Cozzetto, D., 2014. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 31 (6), 857–863.
Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.
Khashan, R., Zheng, W., Tropsha, A., 2012. Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. Proteins 80, 2207–2217.
Kim, H., Park, H., 2014. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3d local discriptor. Proteins 54, 557–562.
Kuo-Cheng, C., Nian-Yi, C., 1977. The biological functions of low-frequency phonons. Sci. Sin. 20, 447–457.
Lafferty, J., McCallum, A., Pereira, F.C.N., 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289.
Lee, B., Richards, F., 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55 (3), 379–400.
Li, X., Pan, X., 2001. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 42, 1–5.
Liu, S., Zhang, C., Liang, S., Zhou, Y., 2007. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins. 68, 636–645.
Lin, H., Deng, E.-Z., Ding, H., Chen, W., Chou, K.-C., 2014. Ipro54-pseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42, 12961–12972.
Liu, B., Wang, S., Long, R., Chou, K.-C., 2016. IRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics 33, 35–41.
Liu, L.M., Xu, Y., Chou, K.C., 2017. iPGK-pseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general pseAAC. Med. Chem. 13 (6), 552–559. Shariqah, United Arab Emirates
Luke, S., 2013. Essentials of Metaheuristics, second ed lulu.com.
Marsh, J.A., 2013. Buried and accessible surface area control intrinsic protein flexibility. Proteins. 425, 3250–3263.

Meiler, J., Muller, M., Zeidler, A., Schmäschke, F., 2001. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J. Mol. Model. 7, 360–369.
Pollastri, G., Baldi, P., Fariselli, P., Casadio, R., 2002. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47, 142–153.
Qiu, W.-R., Xiao, X., Chou, K.-C., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. 15, 1746–1766.
Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H., 1985. Hydrophobicity of amino acid residues in globular proteins. Science 229, 834–838.
Rost, B., Sander, C., 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20, 216–226.
Shen, H.-B., Chou, K.-C., 2007. Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem. Biophys. Res. Commun. 355, 1006–1011.
Thompson, M.J., Goldstein, R.A., 1996. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. Proteins 25, 38–47.
Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., Wilke, C.O., 2013. Maximum allowed solvent accessibilites of residues in proteins. PLoS One 8 (11), e80635.
Wang, J., Hou, T., 2012. Develop and test a solvent accessible surface area-based model in conformational entropy calculations. J. Chem. Inf. Model. 52 (5), 1199–1212.
Wang, J., Lee, H., Ahmad, S., 2007. SVM-cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. Proteins 68, 82–91.
Wang, J.-Y., Lee, H.-M., Ahmad, S., 2005. Prediction and evolutionary information analysis of proteins solvent accessibility using multiple linear regression. Proteins 61 (3), 481–491.
Wang, L., Sauer, U.H., 2008. Ond-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. Bioinformatics 24, 1401–1402.
Wang, S., Peng, J., Ma, J., Xu, J., 2016. Protein secondary structure prediction using deep convolutional neural fields. Sci. Rep. 6, 18962.
Wang, S., Weng, S., Ma, J., Tang, Q., 2015. DeepCNF-d: predicting protein order/disorder regions by weighted deep convolutional neural fields. Int. J. Mol. Sci. 16 (8), 17315–17330.
Wu, W., Wang, Z., Cong, P., Li, T., 2017. Accurate prediction of protein relative solvent accessibility using a balanced model. BioData Min. 10, 1.
Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H., Chou, K.-C., 2013. IAMP-2l: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436, 168–177.
Xu, Y., Ding, J., Wu, L.-Y., Chou, K.-C., 2013a. ISNO-pseAAC: predict cysteine s-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS One 8, e55844.
Xu, Y., Shao, X.-J., Wu, L.-Y., Deng, N.-Y., Chou, K.-C., 2013b. ISNO-AAPair: incorporating amino acid pairwise coupling into pseAAC for predicting cysteine s-nitrosylation sites in proteins. PeerJ 1, e171.
Yuan, Z., Burrage, K., Mattick, J., 2002. Prediction of protein solvent accessibility using support vector machines. Proteins 48, 566–570.
Zhang, H., Zhang, T., Chen, K., Shen, S., Ruan, J., Kurgan, L., 2009. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 76, 617–636.
Zhang, T., Faraggi, E., Zhou, Y., 2010. Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. Proteins 78, 3353–3362.
Zhou, H.-X., Shan, Y., 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins Struct. Funct. Bioinform. 44, 336–343.
Zhou, J., Troyanskaya, O.G., 2014. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. Comput. Sci. 32, 745–753.
Zhou, Y., Vitkup, D., Karplus, M., 1999. Native proteins are surface-molten solids: application of the lindemann criterion for the solid versus liquid state. J. Mol. Biol. 285, 1371–1375.