Note: Accepted and published in 2015 in *Journal of Theoretical Biology*, doi: 10.1016/j.jtbi.2015.06.012.

Improved Prediction of Accessible Surface Area Results in Efficient Energy Function Application

Sumaiya Iqbal, Avdesh Mishra and Md Tamjidul Hoque[*]
Computer Science, University of New Orleans, Louisiana, 70148, USA

Abstract

An accurate prediction of real value accessible surface area (ASA) from protein sequence alone has wide application in the field of bioinformatics and computational biology. ASA has been helpful in understanding the 3-dimensional structure and function of a protein, acting as high impact feature in secondary structure prediction, disorder prediction, binding region identification and fold recognition applications. To enhance and support broad applications of ASA, we have made an attempt to improve the prediction accuracy of absolute accessible surface area by developing a new predictor paradigm, namely REGAd^{3}p, for real value prediction through classical *Exact Regression* with *Regularization* and *p*olynomial kernel of *d*egree *3* which was further optimized using *Genetic Algorithm*. ASA assisting effective energy function, motivated us to enhance the accuracy of predicted ASA for better energy function application. Our ASA prediction paradigm was trained and tested using a new benchmark dataset, proposed in this work, consisting of 1001 and 298 protein chains, respectively. We achieved maximum Pearson Correlation Coefficient (PCC) of 0.76 and 1.45% improved PCC when compared with existing top performing predictor, SPINE-X, in ASA prediction on independent test set. Furthermore, we modeled the error between actual and predicted ASA in terms of energy and combined this energy linearly with the energy function 3DIGARS which resulted in an effective energy function, namely 3DIGARS2.0, outperforming all the state-of-the-art energy functions. Based on Rosetta and Tasser decoy-sets 3DIGARS2.0 resulted 80.78%, 73.77%, 141.24%, 16.52%, and 32.32% improvement over DFIRE, RWplus, dDFIRE, GOAP and 3DIGARS respectively.

1. INTRODUCTION

Function of a protein is found to be closely coupled with the *accessible surface area* (ASA), which is the surface area of a biomolecule (atoms) that is accessible to a spherical solvent while probing the surface of that molecule. The wide conformational dynamics of proteins, which is often exemplified by intrinsic flexible (disorder) regions and thermal fluctuations (B-factor) of a protein, is crucial for their diverse functionalities and is found to be strongly correlated with the ASA of each of the residue of a protein (Marsh, 2013; Zhang et al., 2009). Surface areas, often in the form of exposed residues, are directly involved in the protein-protein interaction (Connoly, 1983; Lee and Richards, 1971). ASA is also found to play an important role in the binding mechanism of proteins in the literature (Chou and Chen, 1977). Thus, the measure of the ASA is essential in understanding the 3-dimensional structure and function of a protein (Butler et al., 2013; Raquel Requejo, 2010). Moreover, ASA has been found to be an important feature for secondary structure prediction, intrinsic disorder prediction, binding region identification, fold recognition and protein function identification (Cheng and Baldi, 2006; Eisenberg and McLachlan, 1986; Liu et al., 2007; Marsh and Teichmann, 2011; Rost, 1995). Importantly, accurate prediction of surface area of protein residues elevates the success in *ab initio* protein structure prediction (Bonetti et al.) and accurate energy function development for correct discrimination of native conformation from the decoys (Khashan et al., 2012; Wang and Hou, 2012). Needless to say, the prediction of real valued accessible surface area from primary protein sequences alone is challenging, yet rewarding in the field of structural biology. We respond to this challenge by developing tools to find accurate ASA from a protein sequence alone and validate the outcome with test dataset as well as by significantly improving an energy function application.

The solvent accessibility prediction has been studied in two forms: firstly, binary or, 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., 2009; Faraggi et al., 2012; Wang et al., 2007; Yuan and Huang, 2014). However, the later approach is

[*] Corresponding author
Email address: thoque@uno.edu

preferred over the former since the residue's solvent accessible surface area tends to vary largely due to their free movement in 3-dimensional space (Ahmad et al., 2013; Zhang et al., 2009). Direct prediction of a continuously varying ASA as a real value reduces the inherent error introduced within the approaches, like binary state classification of the residues (exposed or, buried) or, multi-class classification using different choice of thresholds. The state-of-the-art works for real value prediction of accessible surface area includes several pattern recognition algorithms, such as, multiple linear regression (Wang et al., 2005), support vector machines (SVM) (Wang et al., 2007; Yuan and Huang, 2014) and artificial neural network (ANN) (Ahmad et al., 2013; Faraggi et al., 2009; Faraggi et al., 2012). In this article, we propose a new predictor framework, namely REGAd^{3}p, for real value prediction which involves classical *Exact Regression* with *Regularization* to avoid overfitting phenomenon and *p*olynomial kernel of *d*egree *3*. Furthermore, we applied *Genetic Algorithm* (GA) to optimize the weights computed by regularized regression. We selected the kernel with optimization on accuracy in terms of *Mean Absolute Error* (MAE) and *Pearson Correlation Coefficient* (PCC). Our approach achieved maximum PCC of 0.76 on the challenging Tasser benchmark test dataset.

Effective energy function is an essential component of protein's structure prediction for which homologous templates are absent. The major theme of the energy function developed till date are based on the fact that protein in their native state gains the lowest free energy compared to its other possible states. The developed Energy functions can be categorized into two different types (Hao and Scheragat, 1999; Lazaridis and Karplus, 2000; Miyazawa and Jernigan, 1999; Moult, 1997; Vajda et al., 1997): first, physical-based potential, based on empirical molecular mechanics force fields (Brooks et al., 1983; Cornell et al., 1995) and second, knowledge-based potentials or empirical potential energy function, based on statistical analysis of known proteins (Jernigan and Bahar, 1996; Koretke et al., 1996; Samudrala and Moult, 1997; Tanaka and Scheraga, 1976; Tobi and Elber, 2000; Zhou and Zhou, 2002). Knowledge-based potentials can be more successful over physical-based potential (Skolnick, 2006) as it uses growing number of experimental (known) protein structures, can capture unrecognized forces and the execution is much faster compared to the molecular mechanics based tools. In this article, we calculate predicted accessible surface area based energy component and integrate it with hydrophobic-hydrophilic model (HP model) based 3-Dimensional Ideal Gas Reference State (3DIGARS) potential (Mishra and Hoque, 2014) towards a better energy function application.

It is common in the literature to express and predict ASA in the form of relative accessible surface area (RSA) which is calculated by normalizing the absolute ASA by residue specific maximum values of ASA found in the dataset or, ASA of the extended tripeptide conformation, such as, Ala-X-Ala or, Gly-X-Gly. However depending on different normalizing factors, RSA values vary for same amino acid which makes the comparison of performance with existing predictors inconsistent. To overcome such inconsistencies, we avoided normalizing the ASA values. Instead, we directly predicted the absolute accessible area of the protein residues. We introduced a new benchmark dataset in this work collected from Protein Data Bank (PDB) consisting of 1299 protein sequence, called as Secondary Structure Dataset (SSD1299), with 25% sequence identity cut-off. We tested our predictor (REGAd^{3}p) with three blind harder test datasets and compared our predictor's performance on ASA prediction with SPINE-X (Faraggi et al., 2012). The improved performance of our REGAd^{3}p in all cases suggests that integrating GA optimization with regression resulted a robust real value predictor. Furthermore, we developed a secondary structure predictor model for generating three dimensional secondary structure profile (helix, beta and coil probabilities) which is used as features for the ASA prediction using support vector machine package, the LIBSVM (Chang and Lin, 2011).

In the rest of the article, we presented a rigorous analysis of the quality of the predicted ASA by REGAd^{3}p in terms of different amino acids and their physical properties, secondary structure components and range of ASA values. Finally, we applied the predicted ASA values to improve the accuracy of the energy function, 3DIGARS, which actually resulted in outperforming all the state-of-the-art energy functions significantly.

As demonstrated by a series of recent publications ((Chen et al., 2013), (Lin et al., 2014), (Ding et al., 2014), (Xu et al., 2014), (Liu et al., 2015), (Jia et al., 2015)), to establish a really useful sequence-based statistical predictor for a biological system, we aligned the outline of our paper accordingly towards the steps of Chou's 5-step rule (Chou, 2011) for the two different parts (i.e. ASA predictor REGAd^{3}p and Energy function 3DIGARS-2.0) as: *(a)* construct or select a valid benchmark dataset to train and test the predictor, described in sections 2.1 and 3.5.1; *(b)* formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, described in sections 2.3 and 3.5.3; *(c)* introduce or develop a powerful algorithm (or engine) to operate the prediction, described in sections 2.3 and 3.5.3; *(d)* properly perform cross-

validation tests to objectively evaluate the anticipated accuracy of the predictor, described in sections 3.2 and 3.5.3 *(e)* establish a user-friendly web-server for the predictor that is accessible to the public, if not available immediately, then at least the stand alone code is provided. This has been described in the conclusions section.

2. MATERIALS AND METHODS

Here we describe datasets, feature set, methods and evaluation-steps to measure the effectiveness of our approach.

2.1 Datasets

We prepared a new dataset 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, we collected 2793 protein chains (both single and multiple chain) 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. We further carried out three step refinement of this dataset: (*i*) we filtered the dataset so that the pair wise sequence similarity is no more than 25% using BLASTCLUST; (*ii*) we discarded the protein sequences that contain unknown amino acids labelled as 'X' as the physical properties of this amino acid is unknown and (*iii*) we removed the sequences containing amino acids of unknown coordinates. This resulted a dataset of 1299 sequences (SSD1299) and 272,800 residues. We separated randomly selected 298 sequences from this dataset as the test dataset (SSD_TS298), and the remaining 1001 sequences are used as the training dataset (SSD_TR1001). SSD_TR1001 contains 211,048 residues which combines 69,333 helix (32.8%), 51,859 beta (24.5%) and 89,856 coil (42.5%) residues and SSD_TS298 comprised of 61,752 residues which combines 20,470 helix (33.1%), 16,052 beta (25.9%) and 25,230 coil (40.8%) residues. We determined the real or the actual annotation of secondary structural and surface area by the DSSP program (Kabsch and Sander, 1983).

2.2 Feature Set

We gathered a comprehensive set of residue level features for predicting the secondary structures as well as the accessible surface area. The residue level information includes: (*a*) one amino acid (AA) indicator; (*b*) seven physical properties (PP); (*c*) twenty Position Specific Scoring Matrix (PSSM) values; (*d*) one monogram (MG) and twenty bigram (BG) values; (*e*) two predicted disorder probabilities (short and long) (IUS and IUL); (*f*) three predicted secondary structure (SS) probabilities (helix, beta and coil) and (*g*) one terminal tag (T) to indicate five residues from N and C terminal as (-1.0, -0.8, -0.6, -0.4, -0.2) and (+0.2, +0.4, +0.6, +0.8, +1.0) while others as 0.0. Feature AA is one numerical value as indices, sequentially from 1 to 20, which correspond to the twenty different amino acids and PP are the seven physical properties of amino acids described in (Meiler et al., 2001). ASA is found to vary largely with different amino acids, and is correlated with properties, such as hydrophobicity and isoelectric point (Zhang et al., 2009). Thus we used these features to predict real ASA values. PSSM values accommodate evolutionary information of protein residues and generated by executing 3 iterations of PSI-BLAST (Altschul et al., 1997) against NCBI's non-redundant database. A feature extraction technique by (Sharma et al., 2013) suggests that MG and BG values contains useful 3-dimentional evolutionary information of protein residues. Therefore, we calculated BG and MG values from PSSMs as described in (Sharma et al., 2013) and used as features in our proposed ASA predictor. Disorder residues are often characterized to have large ASA values (Schlessinger et al., 2009). To incorporate this correlation into feature set, we computed disorder probabilities (IUS and IUL) using IUPRED (Dosztányi et al., 2005) and incorporated into our feature set. Protein secondary structure is also closely coupled with ASA of the protein residues (Momen-Roknabadi et al., 2008). Thus we developed our SVM model to generate predicted SS probabilities and used as features to predict ASA values. We separately reported our predictor's performance using two feature sets to depict the importance of 3 dimensional structural features in ASA prediction. Feature Plan #1 contains sequence based one dimensional features (*a* – *c*, *g*). In addition with the features in Feature-Plan #1, Feature-Plan #2 includes 3-dimensional feature (*d*) and predicted structural features (*e* – *f*). **Table 1** illustrates an overview of different feature plans used for secondary structure (SS) and ASA prediction along with the total counts of features in a given features plan. Finally, we included the neighboring residue information within the residue specific features set by applying a window size of 21 to incorporate the effect of residue contacts.

Table 1: List of features used in secondary structure and ASA prediction according to different feature plans.

Feature description (abbreviation)	Feature Count	Feature-Plan #1		Feature-Plan #2	
		SS	ASA	SS	ASA
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	-	√	-	√
Terminal tag (T)	1	√	√	√	√
Total	55	29	32	52	55

'√' and '-' imply that the corresponding feature-set is included and excluded, respectively in the feature-plan.

2.3 Real Value Predictor Framework (REGAd3p) and Evaluation Measures

REGAd^{3}p is a real value predictor framework that combines the exact regularized regression with optimization of weights by genetic algorithm. The equation of basic exact regression (Hastie et al., 2001) is:

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

Here, X = input feature matrix having dimensions: number of residues ($N_{residue}$) × number of features ($N_{feature}$), X^T = transpose of the feature matrix X, Y = actual value of ASA (*ASAr*) and β= weights. Equation (1) analytically determines the best coefficients (weights) for the regression predicting the ASA value. After having the weights, Equation (2) is followed to predict the ASA. Here, $\hat{Y}$ = predicted values of ASA (*ASAp*).

$$\hat{Y} = X\beta \tag{2}$$

However, this basic equation is for linear regression model which can give poor fits to the data. We extended the kernel of this regression method to degree 3 polynomial within the feature matrix using basis expansion by inserting two extra column vectors for each features which are the squares and cubes of the original feature values. This extension is expressed by the following equation, where p is the number of features given:

$$X = [1\ x_1\ x_2\ x_3 \ldots\ x_p] \tag{3}$$

$$X^3 = [1\ x_1\ {x_1}^2\ {x_1}^3\ \ldots\ x_p\ {x_p}^2\ {x_p}^3] \tag{4}$$

Here, X^3 is the extended feature matrix which is used in place of X is the basic Equation (1) and (2) to determine weights and calculate predicted ASA values, respectively. Extension of the kernel gave us the flexibility of model selection with higher order polynomial to select the best fit model. However, increasing the degree of polynomial can cause overfitting, resulted from highly fluctuating weights. An overfitted model towards training data can give poor performance on test dataset. To overcome this overfitting problem, we implemented regularization which involves adding a penalty term to the error to shrink the value of weights. Therefore, the modified Equation (5) that includes regularization, where, λ = regularization parameter to control weight values and M_λ is the identity matrix of dimension (p+1) by (p+1) with the first diagonal element equal to 0 to avoid affecting the bias term. We performed a search for the best value of λ within the range [-100.0 to 100.0] with an interval size equal to 2.0 and reported the best result (*see* **Section 3**) to compare the best result from regularized exact regression with and without GA optimization.

$$\beta = (X^T X + \lambda M_{\lambda_{(p+1).(p+1)}})^{-1} X^T Y \tag{5}$$

This search for λ gave us 100 sets of weights which is then used as seeds for genetic algorithm. The parameter values of our genetic algorithm implementation are: (*i*) population size = 200, (*ii*) number of generations = 2000, (*iii*) chromosome length = number of weights ($N_{feature}$) × number of bits for each weight (19 in our implementation), (*iv*) elite rate = 10%, (*v*) crossover rate = 80% and (*vi*) mutation rate = 10%. While generating initial population, 100 individuals are taken from the output of regularization and the rests were generated randomly. 10% ((200 × 0.1) or, 20) best performing weight sets are always forwarded towards next generation population from current one. To select the candidates for crossover, we implemented roulette wheel selecting algorithm to sample highly fitted individuals to be utilized for next generation population. We performed crossover on (200 × 0.8)/2 or, 80 pairs of chromosomes and filled up next 160 positions of next generation with 160 chromosomes resulted from crossover. Finally, we filled up last 20 positions of next generation population with 20 least fitted chromosomes of current generation. Then, we randomly selected a mutation candidate from these 20 chromosomes with repetition for (200 × 0.7) or, 140 times. In this process, a single chromosome can be selected multiples times (at most 70% of the total number of chromosomes in a population), and get mutated at multiple positions. Literature suggests that the best value of mutation rate is problem specific (Kühn et al., 2013; Ochoa et al., 2000). The problem space of real valued accessible surface area is large and complicated. Moreover, we had highly diversified population of 200 chromosomes and the length of each chromosome is very high. We used 55 unique features per residue. Including the features of neighborhood residues within the window of size 21, we had (55 × 21) or, 1155 features per residue. The coefficient of each feature is encoded by 19 bits within a chromosome. Therefore, a chromosome of GA is (1155 × 19) or, 24255 bits long. In a single mutation process, we flipped only one randomly chosen bit within the chromosome. Therefore, Mutation at multiple position was desirable to significantly change a long chromosome which aided in finding new improves solution within the large and complex search space of real values ASA. **Figure 1** illustrates an overview of the REGAd^{3}p real ASA prediction framework.

In contrast to the practice in the state-of-the-art predictors, we did not guide our predictor to achieve only low mean absolute error (MAE) or, high Persons correlation coefficient (PCC). Rather, we defined a multi-objective function (OBJ), combining PCC and MAE, and carried out the optimization of weights, β, for maximizing OBJ to achieve high performance both in terms of PCC and MAE. The equations for OBJ, PCC and MAE are as follows:

$$OBJ = PCC + (1 - MAE) \tag{6}$$

where,

$$PCC = \frac{\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})(ASAp_i - \overline{ASAp})}{\sqrt{[\sum_{i=1}^{N}(ASAr_i - \overline{ASAr})^2][\sum_{i=1}^{N}(ASAp_i - \overline{ASAp})^2]}} \tag{7}$$

and,

$$MAE = \frac{\sum_{i=1}^{N}|ASAr_i - ASAp_i|}{N} \tag{8}$$

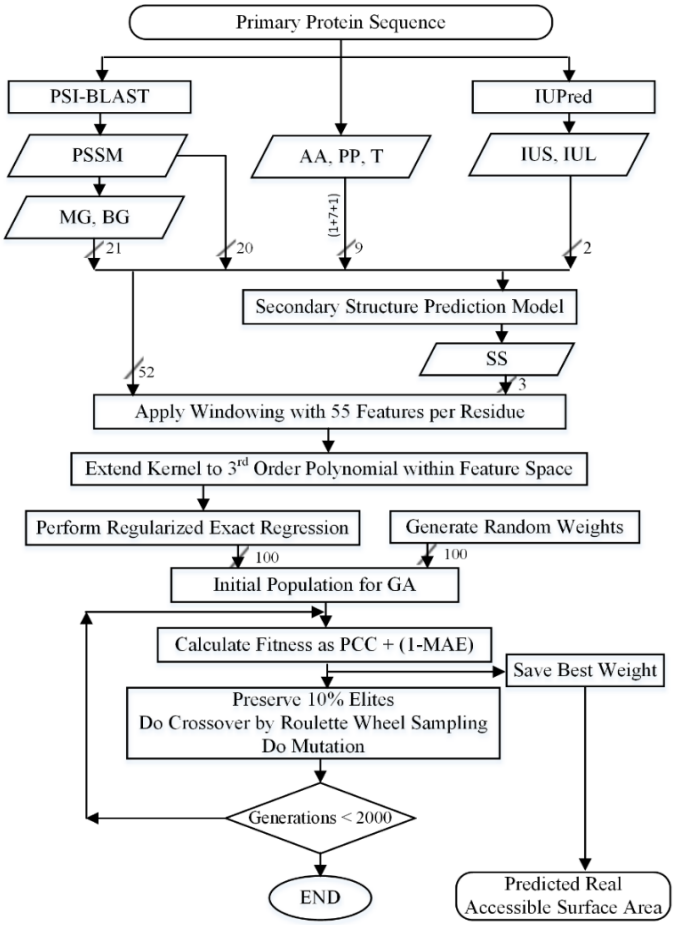

Figure 1: Overview of REGAd^{3}p real value accessible surface area prediction framework, including feature aggregation, secondary structure prediction, and regularized exact regression and GA optimization. The features are represented by their abbreviations introduced in **Table 1**.

Here, N is the total number of residue in the dataset. As the prediction obtains higher accuracy the PCC value increases and the MAE value decreases. Furthermore, we integrated a post processing of predicted ASA values within GA to avoid negative values of the predicted ASA. To keep the ASA values practicable, the predicted negative values (as a result of the natural extension of the equation towards the non-admissible region) were replaced by zero.

3. RESULTS AND DISCUSSION

Here we discuss the robustness of our approach based on obtained results and analysis.

3.1 Results of Secondary Structure Prediction

At first, we predicted secondary structure of a residue of our dataset so that we can use the predicted secondary structure probabilities as features for ASA prediction. We explored four classifiers: (*i*) Logistic Regression (LogReg) using LIBLINEAR (Fan et al., 2008), (*ii*) Random Forest (RF), (*iii*) Artificial Neural Network (ANN) using WEKA (Hall et al., 2009), and (*iv*) support vector machine (SVM) using LIBSVM (Chang and Lin, 2011). Note that, the main objective of this work is to predict ASA and utilize the predicted ASA to improve 3DIGARS (Mishra and Hoque, 2014) energy function. We collected the eight state secondary structure annotation from DSSP program (Kabsch and Sander, 1983) which includes α-helix (H), 3-helix or, 310-helix (G), 5-helix or, π-helix (I), residue in isolated β-bridge (B), residue in extended beta or, β-ladder (E), hydrogen bonded turn (T), bend (S) and random coil or, loop (blank) for all residues in SSD1299 dataset. We converted this eight state annotation into three state annotation (Faraggi et

al., 2012) by coding H, G and I as H (helix), B and E as E (beta) and T, S and blank as C (coil). **Table 2** gives an illustration of the performance of the four classifiers for both the feature plans (***see* Table 1**) in terms of accuracy (total number of correctly predicted residues) of three class classification. All the classifiers were trained on SSD_TR1001 dataset for three class (helix, beta and coil) classification and tested on SSD_TS298 dataset. The superior performance of SVM model on both the feature plans motivated us to select it as our predictor of secondary structure. We used *radial basis function* (RBF) as the kernel for the applied SVM. We used LIBSVM (Chang and Lin, 2011) package for building SVM model and used the default parameter values provided within the package. The default values of misclassification cost of SVM and gamma parameter of RBF were one and ($1/N_{feature}$), respectively.

Table 2: Performance of secondary structure prediction by four classifiers on SSD_TS298 dataset.

Classifier	LogReg(%)	RF(%)	ANN(%)	SVM(%)
Feature-Plan # 1	73.46	72.3	72.8	**75.31**
Feature-Plan # 2	73.51	72.7	72.53	**74.86**

Bold: indicates the obtained best values.

3.2 Results of Accessible Surface Area Prediction

We evaluated the performance of REGAd^{3}p on both the training (SSD_TR1001) and test (SSD_TS298) dataset. **Table 3** presents the performance of REGAd^{3}p in predicting absolute ASA values in terms of PCC and MAE for feature plan # 1 and plan # 2. As a result of inclusion of three dimensional features (MG and BGs) and structural features (short, long disorder probabilities), PCC value is increased (**0.28%**) as well as MAE is decreased (**0.16%**). This result validates the correlation of residue exposure with protein's flexibility (disorder) and usefulness of these features in ASA prediction. In addition, it also motivates us to do further experiments only on feature plan # 2.

We then extended the kernel from degree 1 polynomial (linear) up to 4 to determine the optimal polynomial function to be utilized so that the model best fits the ASA values with feature plan # 2. **Table 4** summarizes the results. PCC is increased by **2.03%** and MAE is decreased by **2.3%** as we extended the kernel from 1st order polynomial to 3rd order polynomial function. **Table 4** also shows **4.63%** and **5.13%** fall in performance in terms of PCC and MAE, respectively, when the kernel is extended beyond 3rd order polynomial. This behavior indicates that the predictor's performance can suffer from high dimensionality as a result of making the model too complex and motivated us to select 3rd order polynomial as the optimal kernel function.

Table 3: Prediction quality of ASA for different feature plans with 1st order polynomial as kernel.

Dataset	SSD_TR1001		SSD_TS298	
Features	**MAE**	**PCC**	**MAE**	**PCC**
Plan # 1, non-optimized	27.27	0.655	25.44	0.711
Plan # 1, optimized	26.53	0.661	24.57	0.717
Plan # 2, non-optimized	27.05	0.665	24.45	0.711
Plan # 2, optimized	**26.31**	**0.670**	**24.53**	**0.719**

Bold: indicates the obtained best values.

Table 4: Prediction accuracy of ASA due to the extension of kernel function from linear to higher order polynomial (Feature-Plan # 2).

Dataset	SSD_TR1001		SSD_TS298	
Polynomial kernel	**MAE**	**PCC**	**MAE**	**PCC**
Degree 2, non-optimized	26.24	0.683	25.05	0.717
Degree 2, optimized	25.86	0.686	24.53	0.723
Degree 3, non-optimized	25.29	0.699	25.54	0.727
Degree 3, optimized	**25.19**	**0.702**	**23.97**	**0.734**
Degree 4, non-optimized	26.61	0.676	25.98	0.685
Degree 4, optimized	26.21	0.679	25.21	0.700

Bold: indicates the obtained best values.

Table 3 and **Table 4** compare the performances of REGAd^{3}p for both with and without the optimization and signify the importance of weight optimization for improving the performance. For the best model (with Feature-Plan # 2 and 3rd order polynomial kernel), the improvement due to optimization over un-optimized model was **0.96%** for PCC and **6.14%** for MAE. Genetic algorithm successfully enhanced the performance of classical regression method to make it competitive with several complex pattern recognition algorithm, like artificial neural network and support vector regression that we had tested extensively. To further verify the robustness of our best model, we carried out 10-fold cross validation on the training dataset, SSD_TR1001. In statistical prediction, the following three cross-validation methods are often used to examine a predictor's effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in (Chen et al., 2013). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (e.g., (Lin et al., 2014), (Ding et al., 2014), (Xu et al., 2014), (Liu et al., 2015), (Jia et al., 2015), (Chou, 2011), (Guo et al., 2014), (KC, 2015)). However, to reduce the computational time, we adopted the 10-fold cross-validation in this study, as done by many investigators with SVM as the prediction engine. The result of 10-fold cross validation test is **0.69** and **25.43** in terms of PCC and MAE, respectively, which is consistent with the result of independent test (indicated with bold in **Table 4**). Finally, the overall performance comparison among the models (for different feature plans as well as kernels) is shown by the **Figure 2** on SSD_TS298 dataset with optimization, where the best model is indicated.

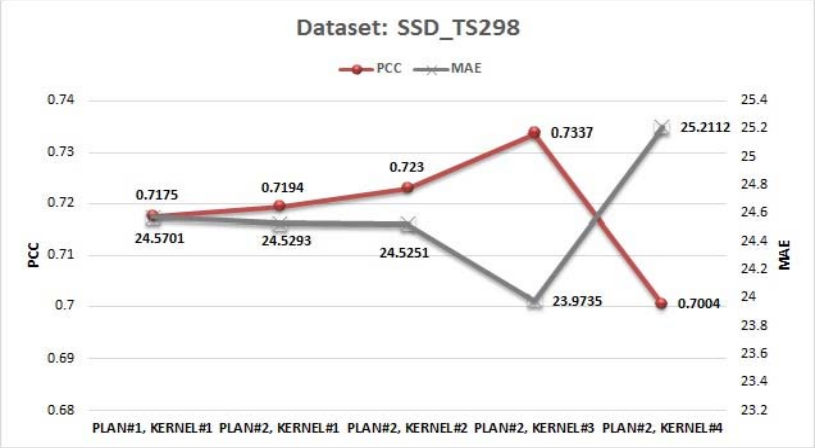

Figure 2: Overall comparison of performance for different feature plans and kernel functions on test dataset with GA optimization. The *x*-axis and *y*-axis shows the model description and performance measure scores (PCC and MAE), respectively. The best model is marked.

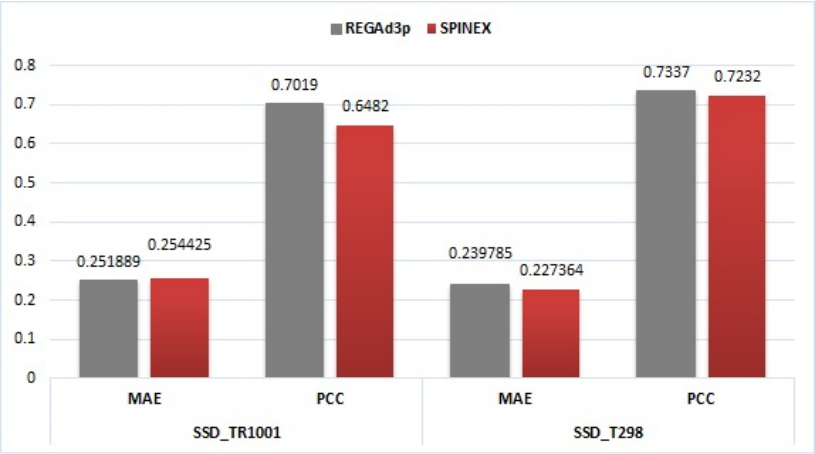

Figure 3: Comparison between our predictor and SPINE-X (Faraggi et al., 2012) in absolute ASA prediction on SSD_TR1001 and SSD_TS298 dataset separately. The x-axis and y-axis represents the dataset and performance measure scores (MAE and PCC), respectively.

3.3 Comparison of Predicted ASA with Existing Methods

We compared our regularized regression technique with optimization against top performing artificial neural network based predictor, SPINEX (Faraggi et al., 2012). To avoid the inconsistencies raised in comparison due to different datasets, normalizing factors and evaluation measures, we ran SPINE-X on SSD1299 dataset and collected the absolute ASA for a fair comparison. **Figure 3** summarizes the result. It shows that except in case of MAE for SSD_TS298 dataset, REGAd^{3}p outperformed SPINE-X (Faraggi et al., 2012) in absolute ASA prediction. REGAd^{3}p gave **8.2%** and **1.45%** improved PCC score than SPINE-X (Faraggi et al., 2012) for SSD_TR1001 and SSD_TS298 datasets, respectively. Further, we evaluated the statistical significance of these improvements by t-test with R package (Revelle, 2015) which shows that both of the improvements are significant. Moreover, we executed support vector regression (SVR) from LIBLINERA package (Fan et al., 2008) on SSD1299 dataset and found the better performances of REGAd^{3}p. SVR resulted a PCC value of 0.51 when trained on SSD_TR1001 dataset and tested on SSD_TS298 dataset for Feature-Plan # 2. To compare, REGAd^{3}p gave PCC value of **0.73** in case of the same dataset and feature plan.

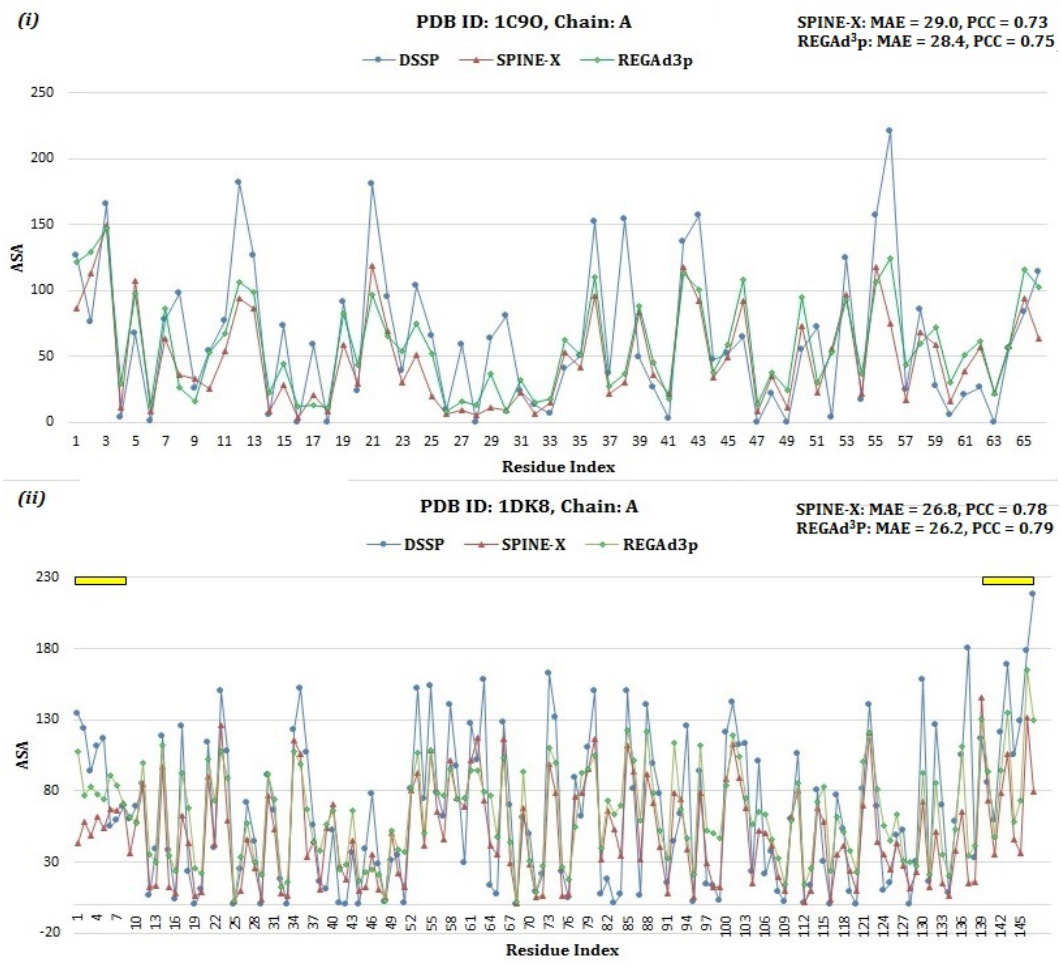

Figure 4: Plot of ASA for each residues of protein (i) PDB ID: 1C9O, Chain: A and (ii) PDB ID: 1DK8, Chain: A, given by DSSP (blue line with circle marker), SPINE-X (Faraggi et al., 2012) (red line with triangle marker) and REGAd^{3}p (green line with diamond marker). For each plot, MAE and PCC scores between predicted ASA given by SPINE-X and REGAd^{3}p with actual ASA from DSSP are shown on the top right corner. In PDB ID: 1DK8, the yellow bars represent two disorder regions *(Iqbal and Hoque, 2014)* at the terminals. The x-axis and y-axis shows the residue index and ASA values, respectively.

3.3.1 Case Study of Individual Proteins

We selected two protein chains, (*i*) 1C9OA and (*ii*) 1DK8A of length 66 and 147, respectively, for investigating the sequence wise performance of REGAd^{3}p versus SPINE-X (Faraggi et al., 2012) in predicting absolute ASA values. We plotted the actual annotation of ASA calculated from DSSP with residue wise predicted ASA values from REGAd^{3}p and SPINE-X (Faraggi et al., 2012) in **Figure 4**. It is clear from the plots that both of the predictors lack in accuracy when exposure is high. To be specific, for the residue index: 11 – 13, 21, 36, 38, 43 – 44, 55 – 56 of protein chain 1C9OA, both predictors under predict. To compare, REGAd^{3}p could result better prediction than SPINE-X for the residue index: 11 – 13, 15, 19, 23 – 3, 36 and 55 – 56 of protein chain 1C9OA. For protein chain 1DK8A, we marked the predicted disorder region (residues 1 – 8 and 139 – 142) at the terminals collected from DisPredict (Iqbal and Hoque, 2014). Disorder regions are often characterized by dynamic conformation and high residue accessible surface areas. It is evident from the plots that REGAd^{3}p could better predict the ASA values at disorder regions than SPINE-X. We further summarized the residue wise performance by sequence wise MAE and PCC scores given by both the methods, reported in the top right corner of the plots. For both the proteins as well as measures (MAE and PCC), REGAd^{3}p outperformed SPINE-X. In case of 1C9OA, REGAd^{3}p gave **2.1%** lower MAE and **2.7%** higher PCC than SPINE-X. At the same time, for 1DK8A, REGAd^{3}p outperformed SPINE-X by **1.5%** decrease in MAE and **1.3%** increase in PCC score.

3.4 Prediction Analysis

We performed extensive analysis on the predicted ASA by REGAd^{3}p and actual ASA obtained from DSSP (Kabsch and Sander, 1983) to assess the quality of our proposed real value prediction framework. We used the prediction output on SSD_TS298 for the following analysis.

3.4.1. Amino Acid Specific Analysis

This analysis is executed in terms of total number of residues within dataset (COUNT) per amino acid, maximum value of actual ASA within dataset (ASAr(MAX)), mean actual ASA (ASAr (MEAN)), standard deviation of actual ASA (ASAr(SD)), mean predicted ASA (ASAp(MEAN)) and MAE resulted from prediction. **Figure 5** illustrates that how the predicted and actual ASA values for each amino acid is highly correlated, with a PCC value equal to 0.99, without no significant over prediction or, under prediction.

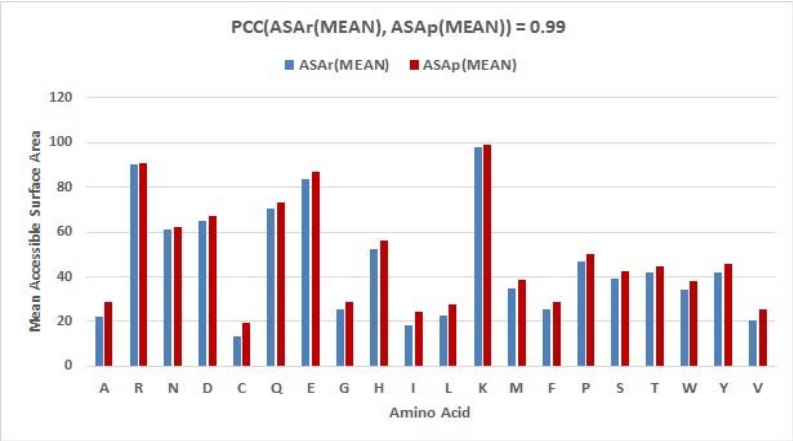

Figure 5: Amino acid specific comparison between mean actual ASA (blue bar) and mean predicted ASA (red bar) values. The x-axis and y-axis show the amino acid and ASA values, respectively.

Figure 6 represents the correlation between the mean absolute errors in prediction with the inherent variability (computed by standard deviation) of ASA values within the dataset. It shows a high correlation value of 0.97 which indicates that the internal fluctuation of ASA within the dataset determines the prediction quality. To this end,

identification of flexible (disorder) residues with high net charge and low hydrophobicity is challenging, which are also characterized by large ASA values (Zhang et al., 2009). It is also evident from **Figure 6**, as the disorder promoting amino acids (R, Q, E, K) (Szilagyi et al., 2008) are likely to have highly variable ASA, therefore contributing in high MAE values. However, **Figure 5** shows that the average predicted ASA given by REGAd^{3}p for these amino acids are very close to the average actual ASA. Furthermore, **Figure 6** shows the prediction error for each amino acid with its respective hydrophobicity index and isoelectric point (charge index) (Meiler et al., 2001). The high errors in prediction are resulted for Arg (R), Lys (K), Glu (E), which have high charge and low hydrophobicity. These residues usually reside in the exterior of proteins and incorporate flexible ASA values with high deviation. **Figure 6** also shows that the low errors are found for Cys (C), Gly (G), Ile (I), Val (V), Ala (A), which are normally buried or, partially buried inside the core of the protein with higher hydrophobicity values.

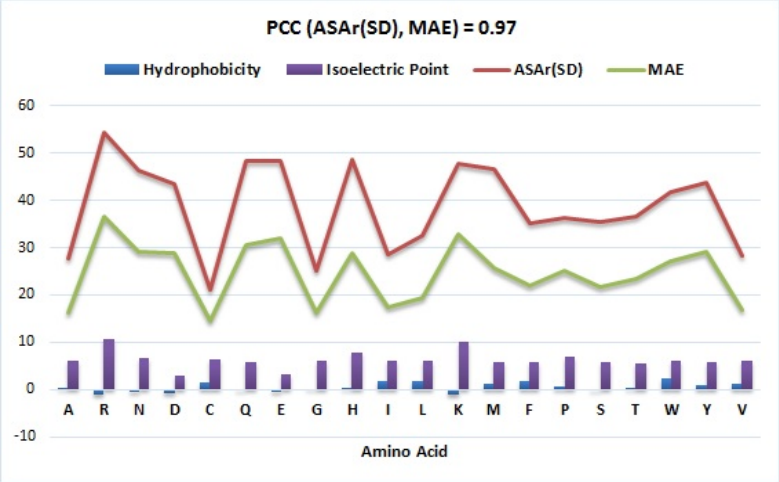

Figure 6: Correlation between MAE (green line) with actual standard deviation (red line) of ASA value within dataset, hydrophobicity (blue bar) and isoelectric point (purple bar) for each amino acid type.

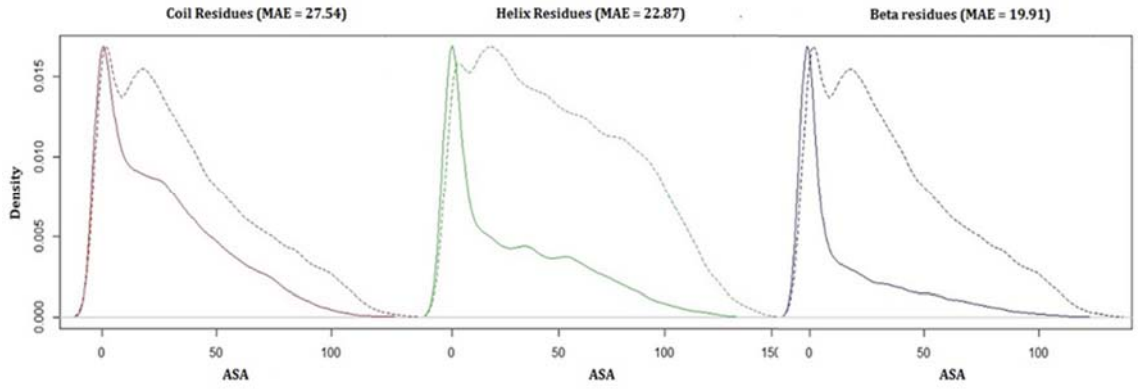

Figure 7: Distribution of actual (solid line) and predicted (dotted line) ASA values for coil (red), helix (green) and beta (blue) residue. The *x*-axis and *y*-axis shows the ASA values and corresponding density. MAE is reported in the title of the each type of residue's curve.

3.4.2. Secondary Structure Specific Analysis

Figure 7 presents the distribution of actual and predicted ASA values for three different secondary structure types of residues (coil, helix and beta) with their corresponding MAE. It is evident from the distributions that for each type of residues, prediction of ASA is relatively easier for the unexposed residues, as the distributions overlap with low ASA values. Coil residues have highest exposure and beta residues have lowest exposure (mean actual ASA for coil, helix

and beta residues are 27.49, 45.87 and 56.61, respectively) which indicates that beta residues are mostly structured whereas coil residues are mostly flexible. This is also clear from the MAE value which decreases from coil to helix to beta residues.

3.4.3. ASA Range Specific Analysis

Figure 8 shows the mean absolute error in prediction resulted by REGAd^{3}p for different range, $[x_1 – x_2)$ where $x_1 \leq ASAr < x_2$ of actual ASA values. It also represents the count of total residues (COUNT), coil residues (Coil), helix residues (Helix) and beta residues (Beta) within a range. The graphical overview illustrates that for a wide range of actual ASA values, [0 – 105), REGAd^{3}p could predict with consistently low MAE which incorporates 87% of the total dataset. For the rest of the residues (13%), the MAE increases with the range which is justified, since these residues are highly exposed and often involved in protein-protein interactions which results in dynamic conformations. **Figure 8** also shows that only within the ASA range [0 – 15), there are more beta residues than other (helix or, coil) residues which are relatively easy to predict as beta residues are mostly structured. For rest of the range, there are more coil residues. However, REGAd3p could still predict the flexible coil residues for a wide range, [15 – 105), resulting consistent MAE with the structured beta residues. Thus, the prediction output of our REGAd^{3}p can be utilized as useful feature for flexible region prediction in future.

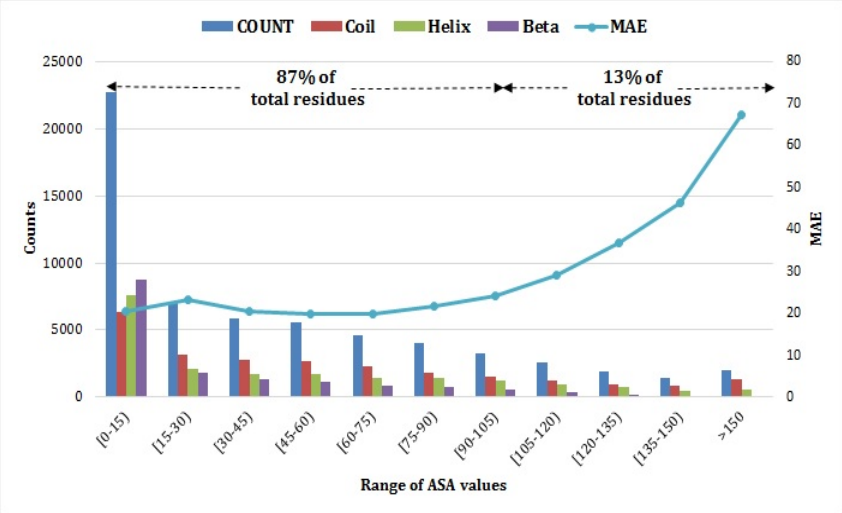

Figure 8: Variation of error in ASA prediction depending on the range of actual ASA values. The x-axis shows the range of actual ASA values in the form $[x_1 – x_2) = x_1 \leq ASAr < x_2$. The primary y-axis (left side) shows the count of all types of residues (blue bar), coil residues (red bar), helix residues (green bar) and beta residues (purple bar) within a range and secondary y-axis (right side) shows MAE (light blue line), respectively.

3.5 Application of Predicted ASA in the Energy Function

An efficient energy function is the one which can identify more native from their decoy sets. Towards this goal, we have developed a new energy function (3DIGARS2.0) which is an improvement over the previous version (3DIGARS[1]) (Mishra and Hoque, 2014) for protein structure prediction. 3DIGARS2.0 is different from 3DIGARS in terms of using sequence based solvent-accessibility information. We obtained the sequence-based solvent-accessibility energy by modelling the error between actual and predicted accessible surface area (ASA). 3DIGARS2.0 is a linear combination of energy from 3DIGARS and energy from sequence-specific solvent-accessibility which is further optimized using Genetic Algorithm (GA). 3DIGARS2.0 outperforms DFIRE (Zhou and Zhou, 2002), RWplus (Zhang and Zhang, 2010), dDFIRE (Yang and Zhou, 2008), GOAP (Zhou and Skolnick, 2011) and 3DIGARS (Mishra and Hoque, 2014) energy functions based on the most challenging Rosetta and I-Tasser decoy sets with an improvement on weighted average of 80.78%, 73.77%, 141.24%, 16.52%, and 32.32% respectively based on count of correct identification of native from decoy sets (see Table 6).

[1] http://cs.uno.edu/~tamjid/TechReport/3DIGARS.pdf, http://cs.uno.edu/~tamjid/Software.html

3.5.1 The 3DIGARS Potential

3-Dimensional Ideal Gas Reference State (3DIGARS) (Mishra and Hoque, 2014) potential is based on hydrophobic-hydrophilic model (HP model) which computes three different energy score interaction tables, specifically, *i*) hydrophobic-hydrophilic (HP), *ii*) hydrophobic-hydrophobic (HH); and *iii*) hydrophilic-hydrophilic (PP). The interaction distance bin size is kept Δr = 0.5 Å each, with a cutoff distance of 15 Å, where *r* represents each distant bin with values ranging from 0.5 Å to 15 Å. The value of Δr_{cut} = 0.5 Å as all bin size are same. The exception consider in 3DIGARS is based on the fact that – amino acid, based on their types are not distributed equally over the 3D structure of a protein to consider them in the same scale on an average by a single dimensional parameter, which can rather be segregated into at least 3 different categories based on the regular distributions within native conformations. 3DIGARS implements Genetic Algorithm (GA) to obtain the best fitted value of six parameters: 3D alpha (three different values of alpha represents three different interaction mentioned above and generate three different energy score tables) and 3D beta (three different values of beta represents the contributions of each of the group along with the z-score).

3.5.2 Decoy Sets

The performance of 3DIGARS2.0 has been compared with other state-of-art energy functions based on three most challenging decoy datasets Moulder, Rosetta and I-Tasser. Modular[2] dataset consists of 20 native proteins with 300 comparative decoy models generated using homologous template for each protein. Rosetta[3] dataset includes 58 proteins generated by Baker Lab. Each of the proteins contains 20 random models and 100 lowest scoring models from 10,000 decoys generated by ROSETTA *de novo* structure prediction (Tsai et al., 2003) followed by all-atom refinement. I-Tasser[4] decoy set-II consist of 56 proteins. Each of these proteins contains 300 to 500 decoys. These decoys were generated first by using Monte Carlo Simulations and then refined by GROMACS4.0[5] MD simulation.

3.5.3. The 3DIGARS2.0 Potential

3DIGARS2.0 combines the sequence-specific solvent-accessibility energy, E^{ASA}, computed from the error between real and predicted accessible surface area (ASA) linearly with 3DIGARS. E^{ASA} is based on probability $P(\Delta ASA_i \mid AA_i)$ computed from the error modelling of our predicted solvent-accessibility $(\Delta ASA_i = ASA_i^{\text{Real}} - ASA_i^{\text{Pred}})$ for a given amino acid type AA_i . Here, for each residue *i* within each protein, ASA_i^{Real} (real accessible surface area) is computed using DSSP and ASA_i^{Pred} is computed using REGAd^{3}p methodology. **Table 5** summarizes the performance achieved by our SVM model on secondary structure prediction in terms of accuracy and REGAd^{3}p predictor on absolute accessible surface area prediction in terms of MAE as well as PCC. The best result was found for I-Tasser dataset in ASA prediction with PCC value equal to **0.76**.

Table 5: Performance of secondary structure prediction and ASA prediction by REGAd^{3}p for Moulder, Rosetta and I-Tasser datasets.

Dataset	SS prediction	ASA prediction	
	Accuracy (%)	MAE	PCC
Moulder	66.56	23.07	0.68
Rosetta	72.04	24.93	0.71
I-Tasser	74.23	24.73	0.76

Now, keeping the prediction accuracy of REGAd^{3}p in mind as in **Table 5**, we wanted to model the error pattern in a useful way to aid in enhancing the accuracy of the energy function. With a view to this, we computed the error in ASA prediction for each of the residues of 1299 proteins and obtained the frequency distribution (FDT) of the error between real and predicted accessible surface area. While building frequency distribution we first calculated the max error ΔASA from the dataset of 1299 protein which was found to be 240. The error range from 0 to 195 was then divided by bin width of 5 to obtain 39 bins of equal size. Remaining of the error ranging from 195 to 240 is considered

[2] http://salilab.org/john_decoys.html

[3] http://www.bakerlab.org/

[4] http://zhanglab.ccmb.med.umich.edu/decoys/

[5] http://www.gromacs.org/Downloads

to fall in the last bin or the 40th bin of the frequency distribution. Thus, the 40th bin has the width of 45 (240-195 = 45). In our implementation of having each bin size of equal width of 5, we normalized the values of last bin by dividing each cell count by 9 (45/9 = 5). Thus, the final frequency distribution table consist of 20 rows (for 20 different types of amino acid) and 40 bins of equal size of 5. For each residue, frequency distribution table is updated as Equation (9):

$$FDT(AA_i, bin_j) = FDT(AA_i, bin_j) + 1.0 \tag{9}$$

Here, AA_i is the i^{th} amino acid and bin_j is the j^{th} bin. Index i ranges from 1 to 20 indicating twenty different amino acids and j ranges from 1 to 40 indicating bins. bin_j is defined by Equation (10):

$$bin_j = abs(\Delta ASA_i = ASAr_i - ASAp_i)/bin\_size \tag{10}$$

where, $bin\_size = 5$. Once the frequency table is obtained, cell whose frequency count is zero are replaced with a small value of 10^{-6}. After the frequency is computed, probability table is obtained by Equation (11):

$$P(AA_i, bin_j) = FDT(AA_i, bin_j)/Tot\_Freq \tag{11}$$

where Tot_Freq is the sum of the count of each amino acid type in frequency table. Finally, the energy score library for sequence specific solvent accessibility is obtain by Equation (12):

$$E(AA_i, bin_j) = -\ln(Bin\_Count \times P(AA_i, bin_j)) \tag{12}$$

Energy associate with each native protein as well as decoy protein is given by Equation (13):

$$E^{ASA} = \sum_{K=1}^{N} E_k(AA_i, bin_j) \tag{13}$$

The combined energy $E^{3DIGARS2.0}$ for each of the proteins including native as well as decoy sets are calculated using Equation (14):

$$E^{3DIGARS2.0} = E^{3DIGARS} + (w \times E^{ASA}) \tag{14}$$

3DIGARS2.0 potential combines 3DIGARS energy with the sequence-specific solvent accessible energy for each of the protein with weight w_1. The optimal value of weight w, ranging from 0 to 2, is obtained from using Genetic Algorithm (GA). The Genetic Algorithm (GA) parameters were of population size 300, elite-rate 5%, crossover-rate 90% and mutation-rate 50%. The stopping criteria to stop the optimization was set to maximum number of generations 2000. **Figure 9** shows the performance of GA where the value remains stable after around 3rd generation.

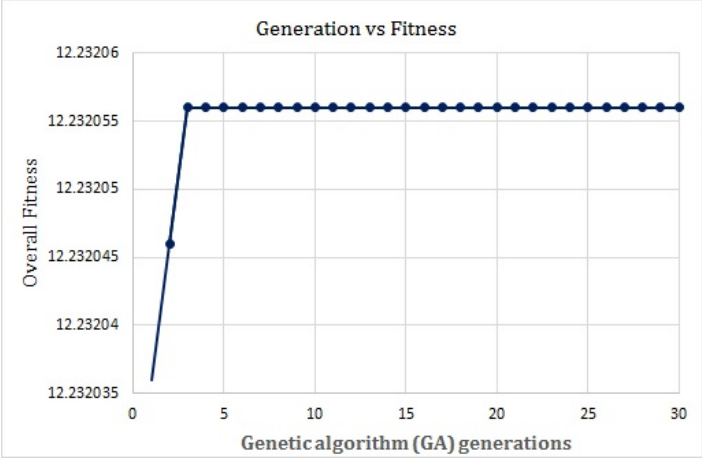

Figure 9: Generations versus fitness graph. Fitness increases sharply and remains constant over number of iterations indicating stable outcome.

3.5.4. Performance of 3DIGARS2.0

We compare the performance of 3DIGARS2.0 with the state-of-art-approaches DFIRE, RWplus, dDFIRE and GOAP using the most challenging three different decoy datasets, Moulder, Rosetta and I-Tasser. 3DIGARS2.0 is found to outperform all of the state-of-art approaches including the previous version of 3DIGARS with high number of correctly identified native proteins from their decoy datasets. For example, based on Rosetta and Tasser decoy-sets 3DIGARS2.0 improved on an average over DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS are **79.64**%, **72.5**%, **162.48**%, 17.77%, and 31.86% respectively. In **Table 6** the results for DFIRE, RWplus, dDFIRE and GOAP are obtained from (Zhou and Skolnick, 2011) and 3DIGARS from (Mishra and Hoque, 2014).

Table 6: Comparison between DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS and 3DIGARS2.0 based on correct selection of native from their decoy-set and z-score.

Decoy Sets (No. of targets)	DFIRE	RWplus	dDFIRE	GOAP	3DIGARS	3DIGARS 2.0
Moulder	**19**	**19**	18	**19**	**19**	**19**
(20)	(-2.97)	(-2.84)	(-2.74)	**(-3.58)**	(-2.998)	(-2.6728)
Rosetta	20	20	12	45	31	**49**
(58)	(-1.82)	(-1.47)	(-0.83)	**(-3.70)**	(-2.023)	(-2.9871)
Tasser	49	**56**	48	45	53	**56**
(56)	(-4.02)	**(-5.77)**	(-5.03)	(-5.36)	(-4.036)	(-4.2964)

Bold indicates best score. Values within the parenthesis are average z-scores of the native structure.

4. CONCLUSIONS

In this article, we proposed a new methodology, namely REGAd^{3}p, to predict absolute accessible surface area of residues from protein sequence alone. The accessible surface area is often used as an important measure related to proteomic study for describing the biophysical properties of a protein. Our method combines canonical exact regression technique together with the optimization by the genetic algorithm. The superior performance achieved by our proposed framework proves that integration of optimization by genetic algorithm can successfully enhance the capability of classical pattern recognition methods. We introduced a comprehensive feature set which could better characterize absolute ASA as we achieved better accuracy (PCC: 0.73) in case of independent test compared to existing independent test results (PCC equal to 0.49 (Ahmad et al., 2013), 0.61 (Yuan and Huang, 2014), 0.66 (Wang et al., 2007)). However, these result are subject to different datasets and different normalizing factors which make the comparison often inconsistent. In this work, we introduced a new benchmark dataset, SSD1299 and compared the

performance of our methodology with the existing state-of-the-art predictor, SPINE-X (Faraggi et al., 2012) by running it on SSD1299 dataset. Our test results for multiple datasets (SSD_TR1001, SSD_TS298, Moulder, Rosetta, I-Tasser) further prove our method is robust. As demonstrated in a series of recent publications (e.g., (Chen et al., 2013), (Lin et al., 2014), (Ding et al., 2014), (Xu et al., 2014), (Liu et al., 2015), (Guo et al., 2014)) in developing new prediction methods, user-friendly and publicly accessible web-servers significantly enhance their impacts (KC, 2015). We will make efforts in our future work to provide a web-server for the prediction method presented in this paper.

For this work, we avoided the common practice of predicting normalized ASA (Faraggi et al., 2009) and then de-normalizing since the normalized ASA varies depending on normalizing factor. Current research is still determining a suitable reference value for normalizing ASA (Matthew et al., 2013) which indicates that inaccurate normalizing factor can lead to misinterpretation of the absolute accessible surface area of residue within a protein conformation. We followed the normalized ASA calculation adopted in (Faraggi et al., 2009) and computed the MAE for normalized ASA prediction with respect to the SSD1299 dataset by our proposed framework. We found that the correlation between the error in prediction in case of with and without normalization for twenty different amino acids is poor with PCC value equal to 0.41 which also proves the inconsistency between absolute and normalized ASA values. Therefore, we predicted absolute ASA values and justified the quality of our prediction through exhaustive analyses of different amino acids along with their physical properties, residues with different type of secondary structure and multiple range of ASA values. Through these analyses, we were able to show that flexible and higher ASA values are harder to predict. However, for a wide range of ASA values, [0 – 105), REGAd^{3}p can predict with consistently lower error and higher correlation which suggests that our methodology can be useful and consistent in measuring parameters in protein's dynamic or, flexible structure prediction and function identification. As to come up with a concrete instance of the claim, we extended 3DIGARS energy functions by optimally combining the ASA error model based energy generated by REGAd^{3}p, namely 3DIGARS2.0, which significantly outperformed all the state-of-the-art energy functions based on the most challenging decoy datasets.

Energy function is one of the key component of *ab initio* protein structure prediction. *Ab initio* protein structure prediction is the method of predicting proteins 3D structure from the amino acid sequence only. In cases where homologous proteins are absent, the problem of *ab initio* protein structure prediction becomes essential. Therefore, we believe that our real value predictor of ASA, REGAd^{3}p and energy function, 3DIGARS2.0 will be very useful for emerging research of proteomics and related fields.

ACKNOWLEDGEMENTS

SI, AM and MTH gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19.

References

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.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

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.

Bonetti, D., Pérez-Sánchez, H., Delbem, A., An Efficient Solvent Accessible Surface Area calculation applied in Ab Initio Protein Structure Prediction.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., Karplus, M., 1983. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J. Comput. Chem. 4, 187-217.

Butler, E., Davis, R., Bari, V., PA, P. N., Ruiz, N., 2013. Structure-Function Analysis of MurJ Reveals a Solvent-Exposed Cavity Containing Residues Essential for Peptidoglycan Biogenesis in Escherichia coli. Journal of Bacteriology 195, 4639-4649.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1--27:27.

Chen, W., Feng, P. M., Lin, H., Chou, K. C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.

Cheng, J., Baldi, P., 2006. A machine learning information retrieval approach to protein fold recognition. Bioinformatics 22, 1456-1463.

Chou, K. C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273, 236 - 47.

Chou, K. C., Chen, N. Y., 1977. The biological functions of low-frequency phonons. Scientia Sinica 20, 447 - 457.

Connoly, M., 1983. Solvent accessibility surfaces of protein and nucleic acids. Science 221, 709 - 713.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., Kollman, P. A., 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 117, 5179-5197.

Ding, H., Deng, W. Z., Yuan, L. F., Liu, L., Lin, H., Chen, W., Chou, K. C., 2014. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International 2014.

Dosztányi, Z., Csizmok, V., and, P. T., Simon, I., 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433-3434.

Eisenberg, D., McLachlan, A., 1986. Solvation energy in protein folding and binding. Nature 319, 199 - 2013.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: A Library for Large Linear Classification, Journal of Machine Learning Research. Journal of Machine Learning Research 9, 1871--1874.

Faraggi, E., Xue, B., Zhou, Y., 2009. 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., 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. J Comput Chem 33, 259 - 267.

Gianese, G., Bossa, F., Pascarella, S., 2003. Improvement in prediction of solvent accessibility by probability profiles. Proteins 16, 987 - 992.

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., CChou, K., 2014. iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. Bioinformatics 30, 1522 - 9.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11.

Hao, M.-H., Scheragat, H. A., 1999. Designing Potential Energy Functions for Protein Folding. Curr Opin in Str Bio. 9, 184-188.

Hastie, T., Tibshirani, R., Friedman, J. H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Holbrook, S., Muskal, S., Kim, S., 1990. Predicting surface exposure of amino acids from protein sequence. Protein Eng 3.

Iqbal, S., Hoque, M., 2014. DisPredict: A Fine Disorder-Protein Predictor. Tech. Report TR-2014/1.

Jernigan, R. L., Bahar, I., 1996. Structure-Derived Potentials and Protein Simulations. Curr Opin in Str Bio. 6, 195-209.

Jia, J., Liu, Z., Xiao, X., Liu, B., Chou, K. C., 2015. iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol 377, 47 - 56.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features Biopolymers 22, 2577 - 2637.

KC, C., 2015. Impacts of bioinformatics to medicinal chemistry. Med Chem 11, 218 - 34.

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 - 2012.

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.

Koretke, K. K., Luthey-Schulten, Z., Wolynes, P. G., 1996. Self-Consistently Optimized Statistical Mechanical Energy Functions for Sequence Structure Alignment. Protein Sci. 5, 1043-1059.

Kühn, M., Severin, T., Salzwedel, H., 2013. Variable Mutation Rate at Genetic Algorithms: Introduction of Chromosome Fitness in Connection with Multi-Chromosome Representation International Journal of Computer Applications 72, 31 - 38.

Lazaridis, T., Karplus, M., 2000. Effective Energy Functions for Protein Structure Prediction. Curr Opin in Str Bio. 10, 139-145.

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

Li, X., Pan, X., 2001. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 42, 1 - 5.

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 - 72.

Liu, S., Zhang, C., Liang, S., Zhou, Y., 2007. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 68, 636 - 664.

Liu, Z., Xiao, X., Qiu, W. R., Chou, K. C., 2015. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem 474, 69 - 77.

Marsh, J. A., 2013. Buried and Accessible Surface Area Control Intrinsic Protein Flexibility. Journal of Molecular Biology 425, 3250 - 3263, doi:10.1016/j.jmb.2013.06.019.

Marsh, J. A., Teichmann, S. A., 2011. Relative Solvent Accessible Surface Area Predicts Protein Conformational Changes upon Binding. Structure 19, 859–867.

Matthew, Z. T., Austin, G. M., Dariya, K. S., Stephanie, J. S., Claus, O. W., 2013. Maximum Allowed Solvent Accissibilities of Residues in Proteins PLOS ONE 8, e80635.

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.

Mishra, A., Hoque, M., 2014. Three-Dimensional Ideal Gas Reference State based Energy Function. Tech. Report TR-2014/2.

Miyazawa, S., Jernigan, R. L., 1999. An Empirical Energy Potential with a Reference State for Protein Fold and Sequence Recognition. Proteins: Struct., Funct., Genet. 36, 357-369.

Momen-Roknabadi, A., Sadeghi, M., Pezeshk, H., Marashi, S. A., 2008. Impact of residue accessible surface area on the prediction of protein secondary structures. BMC Bioinformatics 9.

Moult, J., 1997. Comparison of Database Potentials and Molecular Mechanics Force Fields. Curr Opin in Str Bio. 7, 194-199.

Ochoa, G., Harvey, I., Buxton, H., 2000. Optimal Mutation Rates and Selection Pressure in Genetic Algorithms. Proc. Genetic and Evolutionary Computation Conference (GECCO}.

Raquel Requejo, T. R. H., Nikola J. Costa and Michael P. Murphy, 2010. Cysteine residues exposed on protein surfaces are the dominant intramitochondrial thiol and may protect against oxidative damage. The Febs Journal 277, 1465–1480.

Revelle, W., 2015. psych: Procedures for Psychological, Psychometric, and Personality Research.

Rost, B., 1995. TOPITS: Threading one-dimensional predictions into three-dimensional structures. Third International Conference on Intelligent Systems for Molecular Biology, 314-312.

Rost, B., Sander, C., 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20, 216 - 226.

Samudrala, R., Moult, J., 1997. An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. J. Mol. Biol., 895-916.

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., Rost, B., 2009. Improved Disorder Prediction by Combination of Orthogonal Approaches. PLOS ONE 4, e4433 - e4442.

Sharma, A., Lyons, J., Dehzangi, A., Paliwal, K., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. J Theor Biol. 320, 41 - 46.
Skolnick, J., 2006. In quest of an empirical potential for protein structure prediction. Curr Opin in Str Bio. 16, 166-171.
Szilagyi, A., Györffy, D., Zavodszky, P., 2008. The Twilight Zone between Protein Order and Disorder. Biophysical Journa 95, 1612 - 1626.
Tanaka, S., Scheraga, H. A., 1976. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules 9, 945-950.
Tobi, D., Elber, R., 2000. Distance-Dependent, Pair Potential for Protein Folding: Results From Linear Optimization. Proteins: Struct., Funct., Bioinf. 41, 40-46.
Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, B., Rohl, C. A., Baker, D., 2003. An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction. Proteins: Struct., Funct., Bioinf. 53, 76-87.
Vajda, S., Sippl, M., Novotny, J., 1997. Empirical Potentials and Functions for Protein Folding and Binding. Curr Opin in Str Bio. 7, 222-228.
Wang, J.-Y., Lee, H.-M., Ahmad, S., 2005. Prediction and evolutionary information analysis of proteins solvent accessibility using multiple linear regression. Proteins 61.
Wang, J., Hou, T., 2012. Develop and Test a Solvent Accessible Surface Area-Based Model in Conformational Entropy Calculations. Journal of Chemical Information and Modeling 52.
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.
Xu, Y., Wen, X., Wen, L. S., Wu, L. Y., Deng, N. Y., Chou, K. C., 2014. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS One 9, e105018.
Yang, Y., Zhou, Y., 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72, 793-803.
Yuan, Z., Huang, B., 2014. Prediction of protein accessible surface areas by support vector regression. Proteins 57, 558 - 564.
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 - 36, doi:10.1002/prot.22375.
Zhang, J., Zhang, Y., 2010. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. Plos One 5.
Zhou, H., Zhou, Y., 2002. Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction. Protein Sci., 2714–2726.
Zhou, H., Skolnick, J., 2011. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophys. J . 101, 2043-2052.