Journal of Theoretical Biology 380 (2015) 380–391

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Improved prediction of accessible surface area results in efficient energy function application

Sumaiya Iqbal, Avdesh Mishra, Md Tamjidul Hoque *

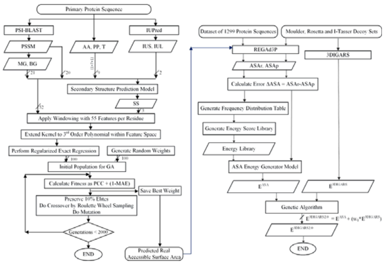
Computer Science, University of New Orleans, Louisiana 70148, USA

HIGHLIGHTS

- Regularized exact regression with 3rd order polynomial kernel based predicted ASA.
- Kernel based prediction performance is further optimized by genetic algorithm (GA).
- ASA prediction error is modeled into the basic Energy function optimistically.
- Test results based on multiple benchmark datasets outperformed all others.

GRAPHICAL ABSTRACT

Accessible Surface Area (ASA) Prediction Framework (REGAd^{3}p). Energy Function (3DIGARS-2.0) Development Framework.


Primary Protein Sequence
PSI-BLAST
IUPred
PSSM
AA, PP, T
IUS, IUL
MG, BG
21
20
[1+7+1]
9
2
Secondary Structure Prediction Model
SS
32
3
Apply Windowing with 55 Features per Residue
Extend Kernel to 3rd Order Polynomial within Feature Space
Perform Regularized Exact Regression
Generate Random Weights
100
100
Initial Population for GA
Calculate Fitness as PCC + (1-MAE)
Save Best Weight
Preserve 10% Elites
Do Crossover by Roulette Wheel Sampling
Do Mutation
Generations < 2000
END
Predicted Real Accessible Surface Area
Dataset of 1299 Protein Sequences
Moulder, Rosetta and I-Tasser Decoy Sets
REGAd3P
3DIGARS
ASAr, ASAp
Calculate Error ΔASA = ASAr-ASAp
Generate Frequency Distribution Table
Generate Energy Score Library
Energy Library
ASA Energy Generator Model
E^ASA
E^3DIGARS
Genetic Algorithm
E^3DIGARS2.0 = E^ASA + (w1*E^3DIGARS)
E^3DIGARS2.0
END

ARTICLE INFO

Article history:
Received 10 April 2015
Received in revised form
15 May 2015
Accepted 2 June 2015
Available online 17 June 2015

Keywords:
Accessible surface area
Protein structure
Energy function
Regression
Benchmark
Decoysets

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 *polynomial* kernel of *degree 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 decoysets 3DIGARS2.0 resulted 80.78%, 73.77%, 141.24%, 16.52%, and 32.32% improvement over DFIRE, RWplus, dDFIRE, GOAP and 3DIGARS respectively.

Published by Elsevier Ltd.

* Corresponding author.
E-mail address: thoque@uno.edu (M.T. Hoque).

http://dx.doi.org/10.1016/j.jtbi.2015.06.012
0022-5193/Published by Elsevier Ltd.

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 et al., 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., 2014) 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, 2012; Wang et al., 2007; Yuan and Huang, 2014). However, the later approach is 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, 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 structure prediction of a protein 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 the performance of our predictor 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, 2015); (Ding et al., 2014); (Xu et al., 2014); (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