RESEARCH ARTICLE

Three-Dimensional Ideal Gas Reference State Based Energy Function

Avdesh Mishra and Md Tamjidul Hoque*

Department of Computer Science, University of New Orleans, LA-70148, USA

ARTICLE HISTORY

Received: January 26, 2016
Revised: August 5, 2016
Accepted: August 24, 2016

DOI:
10.2174/1574893611666161109093137

Abstract: ***Background****:* Energy functions of proteins are developed to quantitatively capture the desirable features of physical interaction that determines the protein folding and structure prediction processes.

Objective*:* It is vital to develop an accurate energy function to discriminate native-like proteins from decoys. Along the same line, we develop an accurate energy function, which involves careful modelling of the reference state.

Method*:* Here we propose a novel three-dimensional ideal gas reference state based energy function, which is based on three distinct hydrophobic-hydrophilic interactions of amino acids. The three distinct group of interactions, namely hydrophobic versus hydrophilic, hydrophobic versus hydrophobic and hydrophilic versus hydrophilic are controlled via three-dimensional optimized values of alpha. Using Genetic Algorithm, we optimized the contributions of each of the three groups along with the z-score to discriminate the native from the decoys.

Results*:* The approach allows us to segregate the statistics, which in turn enables us to model the interactions more accurately without grossly averaging the impact as done in well-known ideal gas reference state based approach. To compute the energy scores we use a database of 4332 known protein structures obtained from the Protein Data Bank.

Conclusion*:* Our energy function is found to be very competitive compared to the state-of-the-art approaches, and outperforms the nearest competitor by 40.9% for the most challenging Rosetta decoy-set.

Keywords: Decoy-set, energy function, genetic algorithm, optimization, protein structure.

1. INTRODUCTION

The history of protein structure prediction is based on the thermodynamic hypothesis that the native structure gains the lowest free energy compared to the other decoys or the intermediate conformations under same physiological conditions [1]. A good energy function model that can discriminate between native and a nearly infinite number of possible decoy structures is vital for protein structure prediction. So far, many attempts have been made towards development of better energy functions which can be categorized into two different types [2-6] i) physical-based potential, based on molecular dynamics or computation of atom-atom forces [7, 8]; and ii) knowledge-based potentials, obtained from statistical analysis of known protein structure [9-14]. Some of the energy functions are modeled based on a simplified representation of the amino acids which consider a few heavy atoms and a few major forces. Others are based on all atom (167 heavy atoms), knowledge based, distance dependent potential. For example, Kortemme *et al.* [15] obtained a knowledge-based hydrogen-bonding potential. Yang and Zhou incorporated polar-polar and polar-nonpolar orientation dependence to the distance-dependent knowledge-based potential that is based on a distance-scaled, finite-ideal gas reference (DFIRE) state [16] by treating polar atoms as a dipole (dDFIRE) [17]. Lu *et al.* [18] defined side-chain orientation according to rigid blocks of atoms (OPUS-PSP).

Zhang and Zhang [19] employed orientation angles between two vector pairs predefined for each side-chain (RWplus). Zhou and Skolnick improved over the DFIRE energy function by incorporating relative orientation of the planes associated with each heavy atom (GOAP) [20]. For obvious reasons, the detailed and relatively complete approaches are the all atom based approaches. The efficacy of the all-atom based approach relies heavily on the formulation of the more accurate reference state [21]. Our proposed work in this paper focuses on all-atom as well as knowledge based approach that derives an effective energy function from known protein structures with multidimensional reference states.

In the seminal work of DFIRE, reference state is formulated by placing the neighboring residues on a modified spherical space. The appropriate shape of the sphere is determined by a single parameter alpha, where the alpha value implies a constant factor (assuming amino-acids are distributed in a protein conformation as a finite system) [10]. On the

*Address correspondence to this author at the Department of Computer Science, University of New Orleans, New Orleans, USA; Tel: +1 (504) 280-2406; Fax: +1 (504) 280-7228; E-mail: thoque@uno.edu

1574-8936/17 $58.00+.00 © 2017 Bentham Science Publishers

contrary, our motivation towards this work comes from the fact that amino acids, based on their types, are not distributed equally over the 3D structure of a protein in order to be able to consider them in the same scale on an average by a single dimensional parameter (Fig. **1a**). Rather, they can be segregated into at least 3 different categories based on the usual distribution with the protein conformations [22] (Fig. **1b**). Related to this, is one of the dominating properties of protein folding, modeled in a hydrophobic-hydrophilic or hydrophobic-polar (HP) model. This model considers that the hydrophobic (H) amino acids have a fear of solvents like water so, they want to keep themselves away from the aqueous solvents forming the core or inner-kernel [23] of a protein and thus remain inside of the protein. On the other hand, the hydrophilic or the polar (P) amino acids or residues, being attracted to water, try and remain outside the core, forming the outer-kernel (Fig. **1b**). Thus Ps are often found on the outside of the folded structure [24, 25], and in between these two layers, there is a thin HP-mixed-layer [23]. Following these aforementioned properties, we proposed our multidimensional reference states based energy function 3DIGARS to improve prediction accuracy.

For an application point of view, an energy function can be crucial. For example, the energy function can be extended to identify appropriate MicroRNA (miRNA) which can play an important role as a regulator in biological processes [26, 27], can be applied in the SNP interaction studies [28] as well as in studying and identifying appropriate DNA-binding proteins [29]. Altogether, they provide important scope of studying the cellular functions and interactions [30].

The rest of the paper is organized as follows. Section 2 discusses the evolution of the relevant theories and underpins theoretical aspects of our proposed approaches. Section 3 discusses our approach for training data collections as well as the collections of the most challenging decoy-datasets to be used for measuring performances of our approach compared to other state-of-the-art approaches. Section 4, discusses the obtained results and lastly section 5 concludes the proposed energy functions.

2. MATERIALS AND METHOD

2.1. Residue Specific All-Atom Probability Discriminatory Function Based Potential

Initially, the residue specific all-atom probability discriminatory function (RAPDF) based energy function was proposed by Samudrala and Moult [9] which was based on averaging reference state. In this approach, the energy score of a conformation was computed in two different ways: as a conditional probability based approach and as a free energy based approach. It was found that these two approaches are equivalent. Although, we would like to note that it is easier to work with the conditional probability based approach because of the Boltzmann assumption on three different aspects of it: *i)* an equilibrium distribution of atom pairs, *ii)* the physical nature of the reference state and *iii)* the probability of observing a system in any given state is also subject to change with respect to the temperature [2].

Conditional probabilities of pairwise atom-atom interactions in proteins can be computed using statistical observations of native structures [9] from protein-databases such as PDB [31]. The conditional probabilities are based on two different types of structures one which is native (N) and the other is the near native or decoy (D). Energy potentials are developed based on the pairwise atom-atom interactions of native structures. Pairwise atom-atom distance is a set of intra-atomic separation within a structure represented as $\{S_{ab}^{ij}\}$, where $\{S_{ab}^{ij}\}$ is the distance between atom *i* and *j* of amino acid type *a* and *b*, respectively. The probability that the atom pairs separated by distance $\{S_{ab}^{ij}\}$ belong to native conformation can be represented by $P(N|S_{ab}^{ij})$. Therefore, we write the general formula of conditional probability such that atom pairs separated by distance $\{S_{ab}^{ij}\}$ belong to the

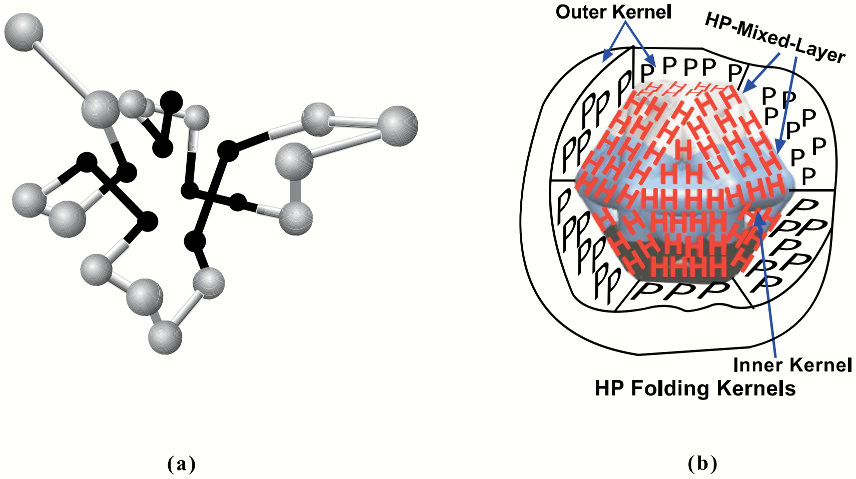

Fig. (1). (a) Native like protein conformation [23], presented in a 3D hexagonal-close-packing (HCP) configuration using hydrophobic (H) and hydrophilic or polar (P) residues. The H-H interactions space is relatively smaller than P-P interactions space, since hydrophobic residues (black ball) being afraid of water tends to remain inside of the central space. (b) 3D metaphoric HP folding kernels, depicted based on HCP configuration based HP model, showing the 3 layers of distributions of amino-acids [22, 23].

native conformation as:

$$P(N \mid S_{ab}^{ij}) = (P(S_{ab}^{ij} \mid N) * P(N)) / P(S_{ab}^{ij}) \qquad (1)$$

Assuming that all distances are independent of each other, conditional probabilities can be expressed as the probabilities of observing the set of distances as products of the probabilities of observing each individual distance [9]

$$P(S_{ab}^{ij} \mid N) = \prod_{ij} P(S_{ab}^{ij} \mid N) \text{ and } P(S_{ab}^{ij}) = \prod_{ij} P(S_{ab}^{ij}) \qquad (2)$$

Substituting the Eq. 1 by Eq. 2 we get Eq. 3:

$$P(N \mid S_{ab}^{ij}) = P(N) * \prod_{ij} P(S_{ab}^{ij} \mid N) / P(S_{ab}^{ij}) \qquad (3)$$

$P(N)$ in the above equation is a constant and independent of conformation of a given structure. Hence, it can be omitted from further consideration. Omission of the $P(N)$ implies that scores from different sequences cannot be compared. Thus, the log form of Eq. 3 is used to both scale the quantities to a small range and give a form similar to that of a potential of mean force. Scoring Function (*SF*) proportional to the negative log conditional probability shows that the structure is correct and can be written as:

$$SF(\{S_{ab}^{ij}\}) = \begin{pmatrix} -\sum_{ij} \ln P(S_{ab}^{ij} \mid N) / P(S_{ab}^{ij}) K \\ -\ln P\left(N \mid \{S_{ab}^{ij}\}\right) \end{pmatrix} \qquad (4)$$

Therefore, given a protein structure or conformation, using Eq. 4, we can calculate the distance separation between all pairs of atom types and compute the total energy by summing up the probability ratios assigned to each separation between a pair of atom types. Thus, we can compute the probability of observing atom type *a* and *b* in a particular bin which is *S* distance apart in a native conformation $P(S_{ab} \mid N)$ as:

$$P(S_{ab} \mid N) = N(S_{ab}) / \sum_{d} N(S_{ab}) \qquad (5)$$

where $N(S_{ab})$ is the frequency of observation of atom types *a* and *b* in a particular bin which is *S* distance apart. The denominator is the number of such observation for all bins.

The denominator in Eq. 4 is the average over different atom types in the experimental conformations which represent the random reference state. Thus the probability of seeing any two atom types *a* and *b* in a bin which is *S* distance apart can be represented as:

$$P(S_{ab}) = \sum_{ab} N(S_{ab}) / \sum_{S} \sum_{ab} N(S_{ab}) \qquad (6)$$

where, $\sum_{ab} N(S_{ab})$ is the total number of counts summed over all pairs of atom types in a particular distance *S*, and the denominator is the total number of counts summed over all pairs of atom types summed over all bins.

As RAPDF energy function is based on averaging reference state, it does not consider the distribution of amino acids in 3D conformational space. Whereas, DFIRE based potential considers the protein as a sphere comprised of uniformly distributed points. Also, it suggests that the radius of such spheres does not increase in r^2 as in an infinite system. Rather, the protein is a finite system and so the increase in the radius is represented by α (a variable which can be $\leq$ 2). This motivated our work towards a more detailed study into a DFIRE based potential.

2.2. DFIRE Based Potential

Distance-scaled, finite ideal-gas reference (DFIRE) state is a distance-dependent, all atom, knowledge-based potential [10]. The reference state formulation in DFIRE is more appealing and effective over RAPDF. For the reference state, RAPDF uses an averaged distribution over all residue or atom pairs. Whereas, DFIRE uses a pair distribution function from statistical mechanics to formulate the finite ideal-gas reference state.

The basis of a finite ideal-gas reference state can be explained by exploring the fundamental equation of statistical mechanics for an infinite system. For an infinite system, the observed number of atom pairs, namely i^{th} and j^{th} atoms, denoted as $N_{obs}(i,j,d)$, at spatial distance *d* with tolerance ±Δ*d* are related to the pair distribution function $g_{ij}(d)$. It describes how density varies as a function of distance from a reference particle and can be represented as:

$$N_{obs}(i,j,d) = \frac{1}{v^s} N_i^s N_j^s g_{ij}(d)(4\pi d^2 \Delta d) \qquad (7)$$

where the volume of the system is represented as v^s, and N_i^s and N_j^s are the number of i^{th} and j^{th} atoms in a system respectively. The potential, based on pairwise distance $P(i,j,d)$, can be written as:

$$P(i,j,d) = \frac{-RT \ln((N_{obs}(i,j,d) * V^s)}{(N_i^s N_j^s (4\pi d^2 \Delta d)))} \qquad (8)$$

In case there is no interaction between the atoms, we can write: $P(i,j,d) = 0$, thus from Eq. 8 we can have:

$$N_{\exp}(i,j,d) = N_{obs}(i,j,d) = N_i^s N_j^s (4\pi d^2 \Delta d / v^s) \qquad (9)$$

Above equations, from statistical mechanics can be directly applied in infinite systems. However, proteins are finite systems. Therefore, the pair density will not be increased by a square factor (i.e., d^2), rather it will increase by some factor α (i.e., d^α) which was determined by the best fit of d^α considering the number of points in 1011 finite protein size spheres.

Thus, Eq. 9 can be written as:

$$N_{\exp}(i,j,d) = N_i^s N_j^s (4\pi d^\alpha \Delta d / v^s) \qquad (10)$$

Furthermore, Eq. 8 can be rewritten as:

$$P(i,j,d) = \frac{-RT \ln((N_{obs}(i,j,d) * V^s)}{(N_i^s N_j^s (4\pi d^\alpha \Delta d)))} \qquad (11)$$

Assuming that there is no interaction beyond a cutoff distance of d_{cut} or $P(i,j,d) = 0$ at $d \geq d_{cut}$, d is replaced by d_{cut}. This leads Eq. 11 to form Eq. 12:

$$P(i,j,d) = -\eta RT \ln \frac{N_{obs}(i,j,d)}{\left(\frac{d}{d_{cut}}\right)^{\alpha} \frac{\Delta d}{\Delta d_{cut}} N_{obs}(i,j,d_{cut})} \tag{12}$$

Here, a constant η is placed for mutation induced changes and is also needed since temperature is a free parameter in potentials derived from static structures. Eq. 12 implies new equation for $N_{\exp}(i,j,d)$:

$$N_{\exp}(i,j,d) = \left(\frac{d}{d_{cut}}\right)^{\alpha} \frac{\Delta d}{\Delta d_{cut}} N_{obs}(i,j,d_{cut}) \tag{13}$$

It is to be noted that the Eq. 13 does not contain any distance dependent terms. Rather, it is a formulation obtained from ideal gas reference state and implementable for finite system.

Similar to the approaches in Samudrala and Moult [9], DFIRE also uses 167 heavy atom types. The cutoff distance d_{cut} is = 14.5 Å. The bin width Δ*d* has different widths for *d* < 2 Å, Δ*d*=2 Å, for 2 Å < *d* < 8 Å, Δ*d*=0.5 Å and for 8 Å < *d* < 15 Å, Δ*d*=1 Å. Thus, the total number of bins is 20. Bin width and d_{cut} were not optimized.

2.3. 3DIGARS, the Proposed Approach

Based on the hydrophobic-hydrophilic model, we constructed three different energy score libraries with bin size, Δ*r* = 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 sizes are the same. We name these libraries as *i*) hydrophobic-hydrophilic (HP); *ii*) hydrophobic-hydrophobic (HH); and *iii*) hydrophilic-hydrophilic (PP) interactions libraries. Each of these libraries are comprised of its independent reference states. A reference state corresponding to the HP group can be written as:

$$N_{i,j}^{EXP-HP}(r) = \left(\frac{r}{r_{cut}}\right)^{\alpha_{hp}} \frac{\Delta r}{\Delta r_{cut}} (N_{obs-HP}(i,j,r_{cut}) + N_{obs-HH}(i,j,r_{cut}) + N_{obs-PP}(i,j,r_{cut})) \tag{14}$$

where $N_{i,j}^{EXP-HP}(r)$ represents the expected number of atom pairs at distance r for the HP group, $N_{obs-HP}(i,j,r_{cut})$ represents the number of observations of atom pairs i^{th} and j^{th} at the cutoff distance from the HP library, $N_{obs-HH}(i,j,r_{cut})$ represents the number of observations of atom pairs i^{th} and j^{th} at the cutoff distance from the HH library, $N_{obs-PP}(i,j,r_{cut})$ represents the number of observations of atom pairs i^{th} and j^{th} at the cutoff distance from the PP library and α_{hp} is the parameter that belongs to the hydrophobic versus hydrophilic group which is obtained by the GA.

Similarly, reference states corresponding to the HH group can be written as:

$$N_{i,j}^{EXP-HH}(r) = \left(\frac{r}{r_{cut}}\right)^{\alpha_{hh}} \frac{\Delta r}{\Delta r_{cut}} (N_{obs-HP}(i,j,r_{cut}) + N_{obs-HH}(i,j,r_{cut}) + N_{obs-PP}(i,j,r_{cut})) \tag{15}$$

where $N_{i,j}^{EXP-HH}(r)$ represents the expected number of atom pairs at distance *r* for the HH group, α_{hh} is the parameter that belongs to the HH group which is also obtained by the GA and the rest of the terms are as defined under Eq. 14.

Finally, reference state corresponding to the PP group can be written as:

$$N_{i,j}^{EXP-PP}(r) = \left(\frac{r}{r_{cut}}\right)^{\alpha_{pp}} \frac{\Delta r}{\Delta r_{cut}} (N_{obs-HP}(i,j,r_{cut}) + N_{obs-HH}(i,j,r_{cut}) + N_{obs-PP}(i,j,r_{cut})) \tag{16}$$

where $N_{i,j}^{EXP-PP}(r)$ represents the expected number of atom pairs at distance *r* for the PP group, α_{pp} is the parameter that belongs to PP group which is also obtained by the GA and the rest of the terms are as defined under Eq. 14.

While generating energy score libraries, residue-atom pairs are categorized to identify which group (HP, HH or PP) mentioned above they fall in. E.g., while considering interaction between two Nitrogen (N) atoms of the amino acid Alanine (ALA:N versus ALA:N), we categorize this interaction as belonging to the HH group as ALA (Alanine) is hydrophobic in nature. Similarly, while considering the interaction between a Nitrogen (N) atom of amino acid Arginine (ARG) and a Carbon (C) atom of amino acid Serine (SER); (ARG:N versus SER:C), we categorize this interaction as belonging to the PP group as both residues Arginine (ARG) and Serine (SER) are hydrophilic in nature. The categorization of an amino acid into the HP group is obtained from (Hoque *et al.*) [25]. Along with the categorization of residue-atom pairs, the frequency counts of the specific group is updated simultaneously. Furthermore, these energy score libraries are used for total energy or minimum energy calculations. Once we compute frequencies of all 3 groups, we generate energy scores corresponding to each group. Energy scores for the HP group can be written as:

$$ES_{i,j,r}^{HP} = -\ln(N_{obs-HP}(i,j,r) / N_{i,j}^{EXP-HP}(r)) \tag{17}$$

where $ES_{i,j,r}^{HP}$ is the energy score of atom pair i^{th} and j^{th} at distant bin *r* for group HP, $N_{obs-HP}(i,j,r)$ is the observed number of atom pair i^{th} and j^{th} at distant bin *r* for HP group and $N_{i,j}^{EXP-HP}(r)$ is the expected number of atom pairs at distance *r* for HP group.

Similarly, energy scores for HH group can be written as:

$$ES_{i,j,r}^{HH} = -\ln(N_{obs-HH}(i,j,r) / N_{i,j}^{EXP-HH}(r)) \tag{18}$$

where $ES_{i,j,r}^{HH}$ is the energy score of atom pair i^{th} and j^{th} at distant bin *r* for group HH, $N_{obs-HH}(i,j,r)$ is the observed number of atom pair i^{th} and j^{th} at distant bin *r* for HH group and $N_{i,j}^{EXP-HH}(r)$ is the expected number of atom pairs at distance *r* for HH group.

Finally energy scores for PP group can be written as:

$$ES_{i,j,r}^{PP} = -\ln(N_{obs-PP}(i,j,r) / N_{i,j}^{EXP-PP}(r)) \tag{19}$$

where $ES_{i,j,r}^{PP}$ is the energy score of atom pair i^{th} and j^{th} at distant bin r for group PP, $N_{obs-PP}(i,j,r)$ is the observed number of atom pair i^{th} and j^{th} at distant bin r for PP group and $N_{i,j}^{EXP-PP}(r)$ is the expected number of atom pairs at distance r for PP group.

Later minimum energy or total energy of decoy-set as well as native proteins are calculated from these energy score libraries. We use weight factors $\beta_{hp,}\ \beta_{hh}$ and β_{pp} to optimize the contribution of each of the three energy score libraries. So, total energy (TE) of the protein conformation can be written as:

$$TE = \beta_{hp}E_{hp} + \beta_{hh}E_{hh} + \beta_{pp}E_{pp} \tag{20}$$

where β_{hp}, β_{hh}, and β_{pp} are 3D weights of contribution and E_{hp}, E_{hh}, and E_{pp} are the energy scores obtained from three groups HP, HH and PP. Here E_{hp} can be written as:

$$E_{HP} = \sum_{i,j,r} ES_{i,j,r}^{HP} \tag{21}$$

Similarly, E_{hh} can be written as:

$$E_{HH} = \sum_{i,j,r} ES_{i,j,r}^{HH} \tag{22}$$

And, E_{pp} can be written as:

$$E_{PP} = \sum_{i,j,r} ES_{i,j,r}^{PP} \tag{23}$$

We use a GA [32] for determining the best possible values of alpha ($\alpha_{hp,}\ \alpha_{hh}$ and α_{pp}), and optimized the contributions of each of the three groups by determining their appropriate weights ($\beta_{hp,}\ \beta_{hh}$ and β_{pp}) along with the z-score to discriminate the native from their decoys. The z-score of a native structure is defined as:

$$Z = \frac{E_{native} - E_{average}}{E_{SD}} \tag{24}$$

where E_{native} is the energy of the native protein, $E_{average}$ and E_{SD} are the average and standard deviation, respectively, of

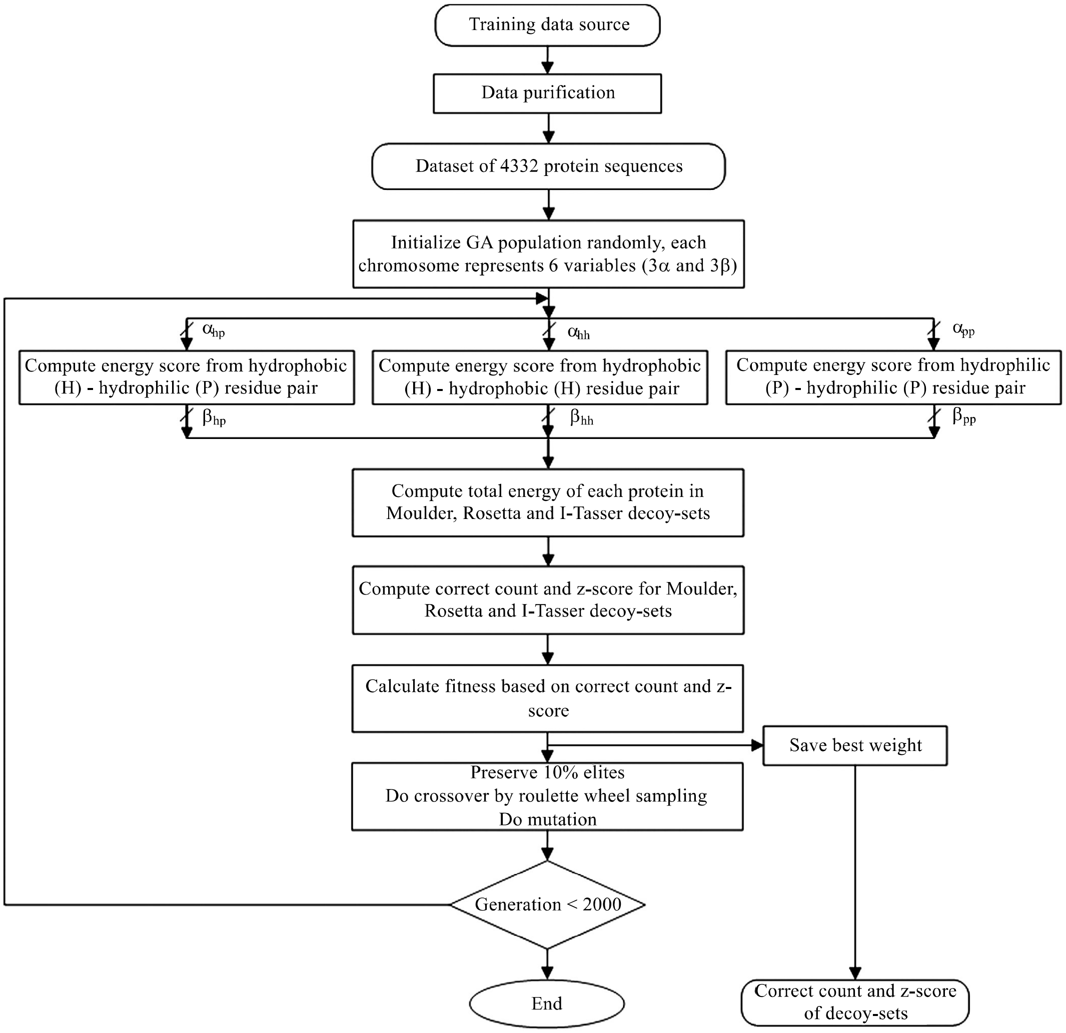
Training data source
Data purification
Dataset of 4332 protein sequences
Initialize GA population randomly, each chromosome represents 6 variables (3α and 3β)
α_hp
α_hh
α_pp
Compute energy score from hydrophobic (H) - hydrophilic (P) residue pair
Compute energy score from hydrophobic (H) - hydrophobic (H) residue pair
Compute energy score from hydrophilic (P) - hydrophilic (P) residue pair
β_hp
β_hh
β_pp
Compute total energy of each protein in Moulder, Rosetta and I-Tasser decoy-sets
Compute correct count and z-score for Moulder, Rosetta and I-Tasser decoy-sets
Calculate fitness based on correct count and z-score
Save best weight
Preserve 10% elites
Do crossover by roulette wheel sampling
Do mutation
Generation < 2000
End
Correct count and z-score of decoy-sets

Fig. (2). Overview of 3DIGARS energy function framework, including dataset collection, 3-dimensional energy score libraries creation, GA optimization based on 3 decoy-sets: Moulder, Rosetta and I-Tasser.

the energies of all the decoy-sets. (Fig. **2**) illustrates an overview of the 3DIGARS energy function framework.

In the GA optimization, the value of alpha and beta ranges from 0 to 2 and -2 to 2 respectively. Population size was set to 50 and single-point crossover and mutation were applied. The elite, crossover and mutation rates were 5%, 90% and 50% respectively. Scores were optimized based on 3 decoy-sets: Moulder, Rosetta and I-Tasser.

The best values obtained for alphas are: α_{hp} = 1.3802541, α_{hh} = 1.6832844 and α_{pp} = 1.9315737. The obtained best beta values are β_{hp} = 1.4921875, β_{hh} = 0.55859375 and β_{pp} = 0.265625. Plots of obtained fitness versus the $\alpha_{hp,}$ α_{hh} and α_{pp} values at each generation show the GA performance on selecting best fitness and also consistency of the obtained fitness with values of $\alpha_{hp,}$ α_{hh} and α_{pp} (Fig. **3**).

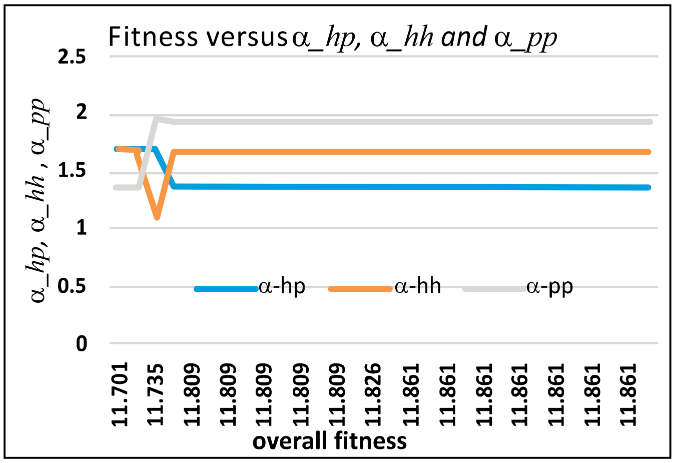

Fig. (3). Fitness versus α_{hp}, α_{hh} and α_{pp} values. The values remain stable during optimization.

To access the performance of our 3DIGARS energy function, we tested 3DIGARS using the most challenging decoy-sets. The performance of 3DIGARS is compared against the state-of-the-art approaches DFIRE, RWplus, dDFIRE and DFIRE2.0 based on the number of correctly identified proteins and average z-score for three different decoy-sets (see Table **1**).

3. DATASET COLLECTION AND DECOY DATASETS

3.1. Training Dataset

Datasets to generate energy scores were obtained from three different sources, the PDB [31] server, the ccPDB [33] (Compilation and Creation of datasets from PDB) server and the PISCES [34] server. We collected datasets with a maximum resolution ≤ 1.5, similarity cutoff 30%, single chain and with a maximum chain length of 500.

Furthermore, we removed proteins with unknown residues as well as missing residues anywhere except for 5 terminal residues on either side. We generated the purified dataset keeping all other specifications common besides maximum resolution, ranging from 1.5 to 2.5, and sequence identities cutoff, of 25%, 30%, 40%, 50%, 70% and 100%. The overall best result is obtained from a collection of 4332 proteins from PDB which are single chains with 100% sequence identity cut-off. Selecting proteins with 100% identity cutoff means we are not discarding proteins even if they are structurally similar because they would represent true frequency distribution. None of the structures from the training dataset were used in the test dataset (Decoy Datasets). Likewise, none of the structures from test datasets were included in training dataset.

3.2. Decoy Datasets

We used 6 decoy datasets to evaluate the performance of the proposed energy function, 3DIGARS, which are described in brief as follows:

3.2.1. Moulder Decoy-Set

The Moulder [35] decoy-set consists of 20 proteins for which 300 comparative models were built using homologous templates. The models were build based on the following criteria: i*)* alignment of the models should not share more than 95% of identically aligned positions or *ii)* alignment of models should have at least 5 different alignment positions. These decoys were build using the MODELLER-6 program which applied a default model building routine with the fastest refinement. This keeps most of the template structure unchanged and different from decoys that are generated by ab initio folding which have all structure regions reassembled from scratch.

3.2.2. Rosetta Decoy-Set

A decoy-set for 58 proteins was generated by the Baker Lab. It contains 20 random models and 100 lowest scoring models from 10,000 decoys using ROSETTA de novo structure prediction followed by all-atom refinement [36]. The current Rosetta decoy-set has been improved over the original Rosetta decoy-set by adding side chains to the centroid/backbone models and refining the structures to remove steric clashes. The improvements over the original Rosetta were based on four important points required to generate optimal decoy-sets: 1) the decoy-set should contain conformations for a wide variety of different proteins to avoid over fitting; 2) the decoy-set should contain conformation close to (< 4Å) to the native structure; 3) the decoy-set should consist of conformations that are at least near local minima of energy potential; and 4) the decoy-set should be produced without using information from the native structure [37].

3.2.3. I-Tasser Decoy-Set-II

I-Tasser [38] decoy-set-II was generated first by using Monte Carlo Simulations and then refined by GROMACS4.0 MD simulation to remove steric clashes and improve hydrogen-bonding networks [38]. This set contains of 56 proteins each of which contains 300-500 decoys generated by both template-based modeling and atomic-level structure refinement.

3.2.4. 4state_Reduced

The 4state_reduced [39] decoy-set consists of 7 proteins. A program called segmod was used to build all atom models from the CA (alpha carbon) atoms. The CA positions for these decoys were generated by choosing 10 residues in each protein using a 4-state off-lattice model.

Table 1. Comparison between DFIRE, RWplus, dDFIRE, DFIRE2.0 and our energy function, 3DIGARS, based on correct selection of native from their decoy-set and z-score.

Decoy-Sets	DFIRE	RWplus	dDFIRE	DFIRE2.0	3DIGARS	No. of Targets
Moulder	**19** (-2.97)	**19** (-2.84)	18 (-2.74)	**19** (-2.71)	**19** **(-2.998)**	20
Rosetta	20 (-1.82)	20 (-1.47)	12 (-0.83)	22 (-1.76)	**31** **(-2.023)**	58
I-Tasser	49 (-4.02)	**56** (-5.77)	48 (-5.03)	53 (-4.548)	53 (-4.036)	56
4state_reduced	6 (-3.48)	6 (-3.51)	**7** (-4.15)	6 (-3.16)	6 (-3.37)	7
Fisa_casp3	4 (-4.80)	4 (-5.17)	4 (-4.83)	4 (-5.08)	**5** (-4.31)	5
Lmds	**7** (-0.88)	**7** (-1.03)	6 (-2.44)	**7** (-0.71)	**7** (-1.96)	10

Bold indicates best score and underline indicates competitive score. Values close to the best results are indicated by underscore ('_').

3.2.5. Fisa_Casp3

Fisa_casp3 [40] decoy-set consists of 5 proteins. It contains decoys for proteins predicted by the Baker group for CASP3 (Critical Assessment of protein Structure Prediction). The main chains for these decoys were generated using a fragment insertion simulated annealing procedure whereas side chains were modelled with a SCWRL package.

3.2.6. lmds

lmds [41] stands for local minima decoy-set. It contains 10 proteins derived from experimental secondary structures from diverse structural classes. Two of the proteins among 10 are from CASP3 (Critical Assessment of protein Structure Prediction) targets.

4. RESULTS

In addition to Moulder [35], Rosetta [36] and I-Tasser [38] decoy-sets (used in Genetic Algorithm (GA) optimization), we tested our energy function, 3DIGARS, on 3 additional decoy-sets, namely 4state_reduced [39], Fisa_casp3 [40] and lmds [41]. It is important to note that the Moulder, Rosetta and I-Tasser sets are considered to be the most challenging decoy-sets whereas the other 3 decoy-sets: 4state_reduced, Fisa_casp3, and lmds are only considered moderate to less challenging for a computational energy function in terms of identifying native out of decoys. The performance of various energy functions on the 6 decoy-sets for native structure selection is compared in Table **1**. 3DIGARS, appears to consistently perform better compared to the state-of-the-art methods. In Table **1**, the values within the parenthesis are average z-scores of the native structures, and the values outside of parenthesis are the number of correct counts. Here the term correct count can be described as the number of correctly identified native proteins from its decoy-sets. A good energy function is one that assigns the lowest free energy to the native proteins within its decoy-set. Thus, it is able to classify native proteins from its decoy-set more effectively. In other words, correct count implies that an efficient energy function can identify more native proteins from the collection of natives within their decoy-sets. Results for DFIRE, RWplus and dDFIRE are obtained from the GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential from Protein Structure Prediction [42]. Correct count and z-score for DFIRE2.0 is computed from the DFIRE2.0 package freely available from the Sparks Lab online server [43]. Correct counts by 3DIGARS is calculated using energy score libraries generated using the dataset with resolution 2.5, sequence similarity cutoff of 100%, and keeping all other parameters used for data collection common as described in DATASET section above. Table **1** clearly shows that the hydrophobic and hydrophilic based energy function outperforms DFIRE, RWplus, dDFIRE and DFIRE2.0 based energy functions for the most challenging Rosetta and the moderately challenging Fisa_casp3 decoy-set. It is to be noted that both RWplus and I-Tasser are designed by the same author. Hence, the rule is most likely embedded within RWplus. Thus, the correct count of 56 out of 56 total targets could be a special case. It is also evident from Table **1** that the result of RWplus for the most challenging Rosetta decoy-set is only 20 correct out of 58, which is a relatively poor performance with respect to other energy functions. Therefore, the performance of RWplus over I-Tasser may not be considered as a very important achievement.

Furthermore, we are often interested in determining whether the mean from more than two populations or groups are equal or not. Therefore, we also conducted Analysis of Variance 1 (ANOVA 1) test to see if one energy function is significantly better than other. ANOVA 1 is a statistical test which is useful to compare the means of two, or more groups. Comparison among the group means is done by estimating comparisons of variance estimates [44].

To test whether the difference in means is statistically significant we can perform ANOVA 1. If the ANOVA 1 F-test shows a significant difference in means between the groups, we would further want to perform pair-wise comparisons between all the groups to determine how they differ [45].

Before testing statistical significance, we graphically compared the means of the energy functions across all of the energy functions (DFIRE, RWplus, dDFIRE, DFIRE2.0 and

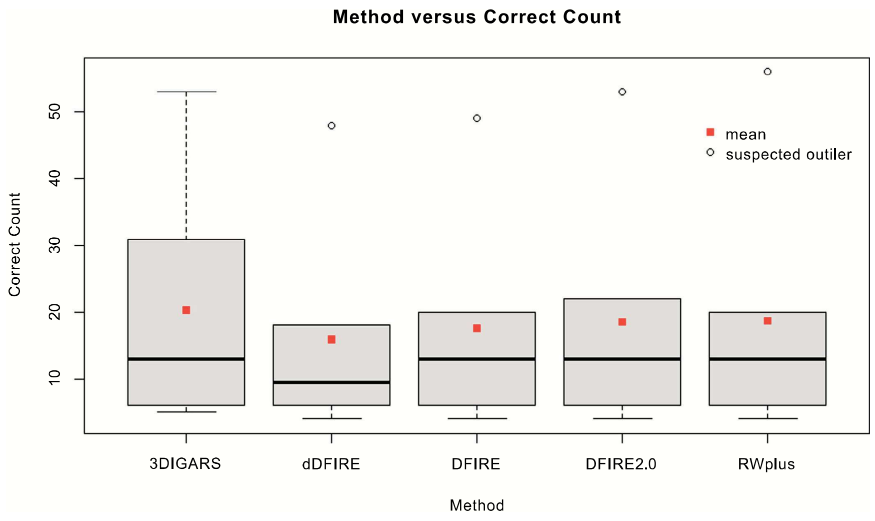

Fig. (4). Comparisons of mean correct count of proposed method with the state-of-the-art methods.

3DIGARS). We used the *boxplot()* function in R to create side-by-side boxplots of measurements organized by groups (energy functions) (Fig. **4**). The data point located outside the fences ("whiskers") shown by the white circles are the outliers. Outliers are an observation that are numerically distant from the rest of the data. Data points that are 1.5 times outside the interquartile range, above the upper quartile and below the lower quartile, are marked as outliers in the *boxplot()* function with default parameter setup in R.

From Fig. (**4**), it appears that the mean correct count for 3DIGARS is highest. We further explain the difference between the boxplots of different methods. From Table **1** we can see that the sample size for each group is 6. As an illustration, for the method 3DIGARS, the sample points sorted in ascending order are 5, 6, 7, 19, 31 and 53, and the median value (Q_2) of the sample space is 13. This splits the data set into two halves (5, 6, 7) and (19, 31, 53). Since, each of the halves of the data set contain an odd count, the sub-medians (Q_1) is 6 for the first subset (5, 6, 7) and the sub-median (Q_3) is 31 for the second subset (19, 31, 53). Thus, the interquartile range (IQR = $Q_3 - Q_1$) is 25. The minimum and maximum values of the sample space are 5 and 53 respectively. Any data point that lies 1.5*IQR above the third quartile (upper inner fence value = (Q_3+(1.5*IQR)), which, for the 3DIGARS method is (31+(1.5*25) = 68.5), is marked as a suspected outlier. As the highest data point for 3DIGARS is 53, it is less than the value of the upper inner fence 68.5. There are no points in the sample space that fall in upper suspected outliers range and thus the upper whisker is of 53.

In contrast, for the method dDFIRE, we can see from Table **1** that the sample points sorted in ascending order are 4, 6, 7, 12, 18 and 78. The median value of the sample space is 9.5. Again, this splits the data set into two halves, (4, 6, 7) and (12, 18, 48). Since each half of the data set contains an odd count, the sub-median (Q_1) is 4 for the first subset (4, 6, 7) and the sub-median (Q_3) is 18 for the second subset (12, 18, 48). Therefore, the interquartile range (IQR = Q_3 - Q_1) is 12. The minimum and the maximum values of the sample space are 4 and 48 respectively. As per the definition of suspected outliers, any data point that lies 1.5*IQR above the third quartile (upper inner fence value = (Q_3+(1.5*IQR)), which for dDFIRE method is (18+(1.5*12) = 36) is marked as an outlier. This discussion implies that the highest data point 48 is greater than the value of the upper inner fence of 36 and so the data point 48 is marked as a suspected outlier (see unfilled circle for the method dDFIRE in Fig. **4**). The outliers (unfilled circles) for the rest of the methods, DFIRE, DFIRE2.0 and RWplus, are also assigned in similar fashion by the *boxplot()* function in R.

Additionally, for the ANOVA 1 test, we used the *aov()* function in R to test if the means of the energy functions are statistically significant. The test resulted in the *f*-value of 0.047 and the *p*-value of 0.996. As the obtained *p*-value of 0.996 is greater than confidence level of 0.05, we accept the null hypothesis of no difference among the mean values of energy function methods with respect to ANOVA 1. However, in reality it requires significant efforts to improve these energy functions and once they are used within various protein computational methods, they magnify the outcomes, which may not be reflected at all by ANONA 1.

CONCLUSION

Identifying native proteins from their decoy-sets has always been a challenging task. While exercising with two different reference state implementations, RAPDF and DFIRE, we formulated a better energy function based on the training dataset, hydrophobic and hydrophilic properties of the amino acids and their role in 3D structure formation, 3D values of alpha based on hydrophobic and hydrophilic residues spatial distributions, and by optimizing the weights of each of the three combinations along with the z-score for discriminating the native from the decoys.

The most important contribution we made is the extension of the concept of ideal gas reference state by constructing three energy score libraries based on hydrophobic and hydrophilic residue's spatial distribution within protein conformations. Each of the three categories of residues are given their independent and more appropriate semi-spherical distribution parameter alphas. Then, we determine their best values instead of having a single parameter based gross average interaction model.

The performance of the training dataset, with sequence similarity cutoff of 100%, gave consistent results over several different datasets obtain by varying the parameters. This indicates that keeping a 100% similar dataset helps us maintain the natural frequency distributions and helps develop a better energy function.

We compare the performance of our proposed 3DIGARS with the state-of-the-art approaches, DFIRE, RWplus, dDFIRE and DFIRE2.0, using six commonly used decoy datasets. 3DIGARS is found to be very competitive and based on the most challenging dataset Rosetta, 3DIGARS outperforms the nearest competitor by 40.9%.

AUTHOR CONTRIBUTIONS

AM and MTH developed the plan and concepts of the work. MTH supervised and AM implemented as well as extended the methodologies. All authors have given approval to the final version of the manuscript.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19. We acknowledge the useful discussion with Sumaiya Iqbal and Md Nasrul Islam. We are also thankful to Aaron Maus, Glenn Robert McLellan and Denson Smith for their help in reviewing the correctness of grammar as well as the structure of the sentences.

SUPPLEMENTARY CONTENT

The software, dataset and related material are available at http://cs.uno.edu/~tamjid/Software/3DIGARS/3DIGARS. zip

REFERENCES

[1] Lu H, Skolnick J. A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection. Proteins Struct Funct Genet 2001; 44: 223-32.

[2] Moult J. Comparison of Database Potentials and Molecular Mechanics Force Fields. Curr Opin in Str Bio 1997; 7: 194-9.

[3] Vajda S, Sippl M, Novotny J. Empirical Potentials and Functions for Protein Folding and Binding. Curr Opin in Str Bio 1997; 7: 222-8.

[4] Hao M-H, Scheragat HA. Designing Potential Energy Functions for Protein Folding. Curr Opin in Str Bio 1999; 9: 184-8.

[5] Miyazawa S, Jernigan RL. An Empirical Energy Potential with a Reference State for Protein Fold and Sequence Recognition. Proteins Struct Funct Genet 1999; 36: 357-69.

[6] Lazaridis T, Karplus M. Effective Energy Functions for Protein Structure Prediction. Curr Opin in Str Bio 2000; 10: 139-45.

[7] Cornell WD, Cieplak P, Bayly CI *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc 1995; 117: 5179-97.

[8] Brooks BR, Bruccoleri RE, Olafson BD *et al.* CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem 1983; 4: 187-217.

[9] Samudrala R, Moult J. An All-atom Distance-dependent Conditional Probability Discriminatory Function for Protein Structure Prediction. J Mol Biol 1998; 275(5): 895-916.

[10] Zhou H, Zhou Y. Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction. Protein Sci 2002: 11(11): 2714–26.

[11] Tanaka S, Scheraga HA. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules 1976; 9: 945-50.

[12] Jernigan RL, Bahar I. Structure-Derived Potentials and Protein Simulations. Curr Opin in Str Bio 1996; 6: 195-209.

[13] Koretke KK, Luthey-Schulten Z, Wolynes PG. Self-Consistently Optimized Statistical Mechanical Energy Functions for Sequence Structure Alignment. Protein Sci 1996; 5: 1043-59.

[14] Tobi D, Elber R. Distance-Dependent, Pair Potential for Protein Folding: Results From Linear Optimization. Proteins Struct Funct Bioinf 2000; 41: 40-6.

[15] Kortemmea T, Morozova AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. Journal of Molecular Biology 2003; 326: 1239-59.

[16] Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 2002; 11: 2714-26.

[17] Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. PROTEINS: Structure, Function, and Bioinformatics 2008; 72: 793-803.

[18] Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of Molecular Biology 2008; 376: 288-301.

[19] Zhang J, Zhang Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. PLoS ONE 2010; 5: e15386.

[20] Zhou H, Skolnick J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal 2011; 101: 2043–52.

[21] Deng H, Jia Y, Wei Y, Zhang Y. What is the Best Reference State for Designing Statistical Atomic Potentials in Protein Structure Prediction? Proteins Struct Funct Bioinf 2012; 80: 2311–22.

[22] Hoque MT, Chetty M, Lewis A *et al.* DFS Generated Pathways in GA Crossover for Protein Structure Prediction. Neurocomputing 2010; 73: 2308-16.

[23] Hoque MT, Chetty M, Sattar A. Protein Folding Prediction in 3D FCC HP Lattice Model Using Genetic Algorithm In: Bioinformatics special session, IEEE Congress on Evolutionary Computation (CEC). Singapore: 2007.

[24] Fidanova S. An Improvement of the Grid-based Hydrophobic-Hydrophilic Model. Int J Bioautomation 2010; 14: 147-56.

[25] Hoque T, Chetty M, Sattar A. Extended HP Model for Protein Structure Prediction. J Comput Biol 2009; 16: 85-103.

[26] Wei L, Liao M, Gao Y *et al.* Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2013; 11: 192-201.

[27] Zou Q, Mao Y, Hu L *et al.* miRClassify: An Advanced Web Server for miRNA Family Classification and Annotation. Computers in Biology and Medicine 2014; 45: 157-60.

[28] Zou Q, Li X-B, Jiang W-R *et al.* Survey of MapReduce Frame Operation in Bioinformatics. Brefings in Bioinformatics 2013: bbs088.

[29] Song L, Li D, Zeng X *et al.* nDNA-Prot: Identification of DNA-binding Proteins based on Unbalanced Classification. BMC Bioinformatics 2014; 15: 298.

[30] Lin C, Zou Y, Qin J *et al.* Hierarchical Classification of Protein Folds using a Novel Ensemble Classifier. PLOS ONE 2013; 8: e56499.

[31] PDB R. Advanced Search Interface. In: p. Web. February 2014. http://www.rcsb.org/pdb/search/advSearch.do.

[32] Hoque MT, Chetty M, Sattar A. Protein Folding Prediction in 3D FCC HP Lattice Model using Genetic Algorithm. In: IEEE Congress on Evolutionary Computation (CEC) Singapore. Singapore: 2007: 4138-45.

[33] Singh H, Chauhan JS, Gromiha MM *et al.* ccPDB: Compilation and Creation of Data Sets from Protein Data Bank. Nucleic Acids Res 2012; 40: D486-9.

[34] Lab D. Taking Input Parameters for Culling Whole PDB. In: 1969. p. Web. February 2014. http://dunbrack.fccc.edu/Guoli/PISCES_ChooseInputPage.php.

[35] Sali A. Decoy Models. In: p. Web. July 2014. http://salilab.org/john_decoys.html.

[36] Zhang J, Zhang Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. Plos One 2010; 5 (10): e15386.

[37] Tsai J, Bonneau R, Morozov AV *et al.* An Improved Protein Decoy-set for Testing Energy Functions for Protein Structure Prediction. Proteins Struct Funct Bioinf 2003; 53: 76-87.

[38] Lab Z. Protein Structure Decoys. In: Zhang Lab; p. Web. July 2014. http://zhanglab.ccmb.med.umich.edu/decoys/.

[39] Park B, Levitt M. Energy Functions that Discriminate X-ray and Near-native Folds from Well-constructed Decoys. J Mol Biol 1996; 258: 367-92.

[40] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. J Mol Biol 1997; 268: 209-25.

[41] Keasar C, Levitt M. A Novel Approach to Decoy-set Generation: Designing a Physical Energy Function Having Local Minima with Native Structure Characteristics. J Mol Biol 2003; 329: 159-74.

[42] Zhou H, Skolnick J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophys J 2011; 101: 2043-52.

[43] Yang Y. Download Page. In: pp. Web. June 2014. http://sparks-lab.org/yueyang/download/index.php.

[44] Bandyopadhyay S, Mallik S, Mukhopadhyay A. A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 2014; 11: 95-115.

[45] Maulik U, Mallik S, Mukhopadhyay A, Bandyopadhyay S. Analyzing Large Gene Expression and Methylation Data Profiles Using StatBicRM: Statistical Biclustering-Based Rule Mining. PLOS ONE 2015; 10: e0119448.

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.