Journal of Theoretical Biology 398 (2016) 112–121

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Discriminate protein decoys from native by using a scoring function based on ubiquitous Phi and Psi angles computed for all atom

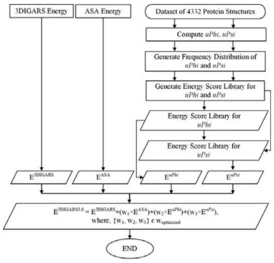
Avdesh Mishra, Sumaiya Iqbal, Md Tamjidul Hoque*

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

HIGHLIGHTS

- Ubiquitous dihedral angles (uD) are mined to generate energy components.
- Regularized exact regression based predicted ASA is modeled into energy score.
- Weight-optimized linear sum of core, ASA and uD energies formed 3DIGARS3.0.
- The new Energy function 3DIGARS3.0, outperforms state-of-the-art methods.

GRAPHICAL ABSTRACT


3DIGARS Energy
ASA Energy
Dataset of 4332 Protein Structures
Compute uPhi, uPsi
Generate Frequency Distribution of uPhi and uPsi
Generate Energy Score Library for uPhi and uPsi
Energy Score Library for uPhi
Energy Score Library for uPsi
E3DIGARS
EASA
EuPhi
EuPsi
E3DIGARS3.0 = E3DIGARS+(w1×EASA)+(w2×EuPhi)+(w3×EuPsi),
where, {w1, w2, w3} ϵ woptimized
END

ARTICLE INFO

Article history:
Received 26 September 2015
Received in revised form
26 February 2016
Accepted 17 March 2016
Available online 28 March 2016

Keywords:
Protein structure prediction
Decoy structure
Native structure
Genetic Algorithm
Optimization

ABSTRACT

The success of solving the protein folding and structure prediction problems in molecular and structural biology relies on an accurate energy function. With the rapid advancement in the computational biology and bioinformatics fields, there is a growing need of solving unknown fold and structure faster and thus an accurate energy function is indispensable. To address this need, we develop a new potential function, namely 3DIGARS3.0, which is a linearly weighted combination of 3DIGARS, mined accessible surface area (ASA) and ubiquitously computed Phi (*uPhi*) and Psi (*uPsi*) energies – optimized by a Genetic Algorithm (GA). We use a dataset of 4332 protein-structures to generate *uPhi* and *uPsi* based score libraries to be used within the core 3DIGARS method. The optimized weight of each component is obtained by applying Genetic Algorithm based optimization on three challenging decoy sets. The improved 3DIGARS3.0 outperformed state-of-the-art methods significantly based on a set of independent test datasets.

Published by Elsevier Ltd.

1. Introduction

Energy function is one of the most important component of protein folding and structure prediction problem. We need an accurate energy function that can assign the lowest global or, local energy to the native protein. Although, there exist fields of quantum mechanics (Cornell et al., 1995; Brooks et al., 1983), that can solve the existing problem, the equations are tedious to work with as the scope of the space or domain is fairly complex. Thus, statistical-based, empirical or knowledge-based energy functions have been developed (Samudrala and Moult, 1997; Zhou and Zhou, 2002; Tanaka and Scheraga, 1976; Jernigan and Bahar, 1996; Koretke et al., 1996; Tobi and Elber, 2000) and found to be more successful than potentials based on quantum mechanics. One of the major reason for the success of knowledge-based potential is the growing number of experimental protein structures.

The knowledge of 3D (three-dimensional) structures of the target proteins and their binding sites with ligands is important for rational drug design. Although, X-ray crystallography is a powerful tool in this regard, it is time-consuming and expensive, and not all proteins can be successfully crystallized. Membrane proteins are difficult to crystallize and most of them will not

* Corresponding author.
E-mail addresses: amishra2@uno.edu (A. Mishra), siqbal1@uno.edu (S. Iqbal), thoque@uno.edu (M.T. Hoque).

http://dx.doi.org/10.1016/j.jtbi.2016.03.029
0022-5193/Published by Elsevier Ltd.

dissolve in normal solvents. Thus, so far, very few membrane protein structures have been determined. Further, although recent breakthrough in high resolution NMR has indicated that it is indeed very powerful tool in determining the 3D structures of membrane proteins and their complexes (Brüschweiler et al., 2015; Berardi et al., 2011; OuYang et al., 2013), it is also time-consuming and costly. To acquire the structural information in a timely manner, a series of 3D protein structures and their binding sites with ligands are usually derived using various structural bioinformatics tools (Chou and Wei, Zhong; Carter and Chou, 1998; Wang et al., 2009; Chou, 2004).

Various automated method are also developed to rapidly and effectively identify these binding sites (Jia et al., 2015a; 2015b). Additionally, the progress of knowledge-based statistical energy function and their applicability in the field of protein–ligand, protein–protein and protein–DNA binding affinity prediction are found to be widely accepted and useful in drug design (Zhang et al., 2005; Muegge and Martin, 1999; Gohlke et al., 2000; Mitchell et al., 1999; Mitchell et al., 1999). For example, Zhang et al. showed that the protein–ligand binding affinities predicted by the DFIRE energy function has correlation coefficient of 0.63 with the experimentally measured protein–ligand binding affinities (Zhang et al., 2005; Gohlke et al., 2000). Thus, intra-disciplinary application of statistical potential and their performance have drawn the attention of numerous researchers and drug industries.

It has been demonstrated that all-atom based potentials outperform the residue-based potentials (Samudrala and Moult, 1997; Yang and Zhou, 2008; Zhou and Skolnick, 2011). All-atom based potentials incorporate both the backbone information as well as the side chain information. In the recent past, we have seen numerous efforts to enhance the accuracy of the all-atom based energy function. Several of these energy functions take advantage of inherent properties of atoms or residues such as generalized orientation angles (Zhou and Skolnick, 2011), orientation dependent interactions by considering each polar atom as a dipole with a direction (Yang and Zhou, 2008), error modeling between real and predicted torsion angles (Hoque et al., 2016) and hydrophobic and hydrophilic (*HP*) properties (Mishra, 2015). Moreover, some energy functions use accurately predicted structural properties with some restraints to improve the performance of protein structure prediction. As an example, the predicted secondary structure is frequently utilized to limit the conformational space which results in increased protein structure prediction accuracy (Chou et al., 1985; Chou and Scheraga, 1982; Chou et al., 1992; Chou and Carlacci, 1991; Carlacci et al., 1991; Chou et al., 1992). Likewise, the generalized orientation-dependent all-atom potential (GOAP) (Zhou and Skolnick, 2011), computes the relative orientation of the planes associated with each heavy atom in interacting pairs. In addition, the dDFIRE (Yang and Zhou, 2008; Zhou and Skolnick, 2011; Hoque et al., 2016) potential augments orientation dependence of polar–polar and polar–nonpolar atom interactions (dipoles) with the distance-dependent knowledge-based, finite ideal-gas reference state potential (DFIRE) (Zhou and Zhou, 2002). The orientation vector of a given polar atom is described by the sum of the bond vectors that covalently bond the polar atom with other heavy atoms. The dDFIRE potential computes three angles (namely, θ_p, θ_q and θ_{pq}, in Fig. 1) in dipole–dipole interactions made by the two polar atoms (p and q in Fig. 1). Adding the orientation dependency of the polar atoms to the distance dependent DFIRE potential, has improved the performance over DFIRE in refolding the protein, especially at the terminal regions. The improvement could be because the orientation angles can capture three dimensional (3D) features, while the distance-dependent core potential is computed purely from the pair-wise distance between two atoms which may not contain enough tertiary information.

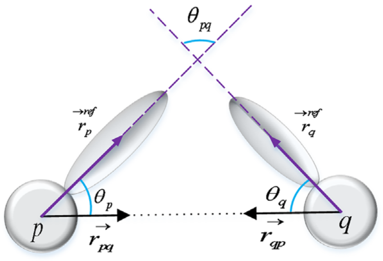

θ_{pq}
$\overrightarrow{r_p^{ref}}$
$\overrightarrow{r_q^{ref}}$
θ_p
θ_q
p
q
$\overrightarrow{r_{pq}}$
$\overrightarrow{r_{qp}}$

Fig. 1. Definition of the orientation angles θ_p, θ_q and θ_{pq}. The vectors $\overrightarrow{r_p^{ref}}$ and $\overrightarrow{r_q^{ref}}$ are the reference directions for polar atoms p and q, respectively (Hoque et al., 2016).

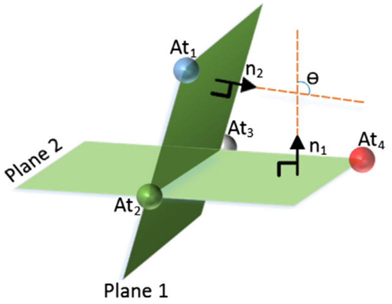


Fig. 2. Definition of the angle θ formed by four atoms (At_1, At_2, At_3 and At_4). Planes 1 and 2 are two planes passing through At_1, At_2, At_3 and At_2, At_3, At_4 respectively. n_1 and n_2 are the normal vectors on the respective planes. *uPhi* is computed using At_1 belonging to one residue and a set of atoms, At_2, At_3, At_4 belonging to some other residues. Similarly, *uPsi* is computed using a set of atoms, At_1, At_2, At_3 belonging to one residue and an atom At_4 belonging to some other residue.

Furthermore, dihedral or torsion angles contain important local structural information that help manifest the folding (Borguesana et al., 2015). A dihedral angle is the angle between two intersecting planes (see Fig. 2). These angles indicate the rotational displacement required for the backbone of the amino acid sequence to sample a certain structure or folds. The well-known Ramachandran plots (Ramachandran et al., 1963), which are the sampling distribution of the two dihedral angles, namely Phi and Psi is reliably used in computing and predicting possible folds (Hooft et al., 1997; Yang and Zhang, 2015). The third dihedral angle, omega essentially only varies in between two values, 0° or 180°. Dihedral angles are defined by four points in the space and provide orientation information of the secondary structure, such as helix, sheet or coil, of a protein.

In their regular usage, the aforementioned dihedral angles convey only the local structural information pertaining to the backbone conformation of the protein. Mining dihedral angles ubiquitously over all-atom orientations, could improve the structure prediction and the energy function. This idea motivated us to integrate all-atom dihedral angle based potential that can map one dimensional linear information of interacting atoms to three dimensional features which can be incorporated into our existing core 3DIGARS2.0 energy function.

With a goal to develop a more accurate all-atom distance-dependent knowledge-based potential to discriminate native structures from decoys, we propose a potential in this article which is a linearly weighted combination of 3DIGARS, sequence-specific solvent-accessibility, mined ubiquitous Phi (*uPhi*) and Psi (*uPsi*) based energies – optimized by genetic algorithm (GA). In our prior work, to build 3DIGARS2.0, we combined the 3DIGARS and sequence-specific solvent-accessibility energies to improve 3DIGARS. In this work, we combine 3D structural features, *uPhi* and *uPsi*, in the form of energy component to improve the accuracy further. The *uPhi* angle is computed in similar way as the dihedral angle, Phi, using the coordinates of four atoms. Likewise, the *uPsi* is an angle computed in the same way as dihedral angle, Psi, using the coordinates of four atoms. We name the angle as *uPhi* or *uPsi* because, we compute these angles ubiquitously for all the atoms in a given structure. Training using Protein Data Bank (PDB) dataset, linearly optimizing using most challenging decoy datasets and testing using several independent decoy datasets, we ensured the superior performance of the proposed energy function, which significantly outperformed the state-of-the-art methods.

To establish an *uPhi* and *uPsi* based energy function for a biological system, we aligned the outline of our paper following Chou's 5-step rule (Chou, 2011) as: (a) details of the underpin theoretical aspect and the formulation of our proposed approach, described in Section 2, (b) construction or selection of a valid benchmark dataset to train, optimize and test the method, described in Section 3, (c) brief discussion of the evolution of the relevant theories based on which the proposed method is evolved, described in Section 4, (d) conclude the proposed method, in Section 5, (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.[1] A series of recent publications following such steps can be found here (Jia et al., 2015a, 2015b; Chen et al., 2013; Lin et al., 2014; Ding et al., 2014; Xu et al., 2014; Zi Liu et al., 2015).

2. Material and methods

This section discusses the proposed energy function to provide a clear picture of the proposed advancements.

2.1. uPhi, uPsi based energy (E^{uPhi}, E^{uPsi})

Protein macromolecule is a linear chain of amino acid residues, where the residues are linked together by peptide bond. The local conformation of the amino acid residues along the backbone is regarded as the secondary structure of a protein (Lehninger et al., 2005). The next level of complexity is regarded as tertiary or 3D structure of a protein, which essentially provides the functional proteins (Borguesana et al., 2015; Lesk, 2004). Torsion angles are one of the ways to represent a protein structure. The conformation of a peptide backbone is mainly described by two torsion angles, Phi and Psi. Furthermore, the torsion angle omega is not involved in the molecular rotation, because it is restricted by the strong double bond (Lehninger et al., 2005). Phi involves the backbone atoms $C(O)_{n-1}$–N_n–$C(\alpha)_n$–$C(O)_n$ and Psi involves the backbone atoms N_n–$C(\alpha)_n$–$C(O)_n$–N_{n+1}. Thus, Phi controls the $C(O)_{n-1}$ versus $C(O)_n$ distance and Psi controls the N_n versus N_{n+1} distance and orientation, where *n* is the current atom for which the dihedral angles are calculated (Lodish et al., 1990).

In our implementation, we capture 3D structural information described by the torsion angles Phi and Psi computed for all atoms, named as *uPhi* and *uPsi* respectively, in the form of energy components. To compute the *uPhi*, a set of four atoms say, "{At_m, At_n, At_o, At_p}" and their corresponding residue index say, "{RI_m, RI_n, RI_o, RI_p}" are considered as follows:

(i) RI_n, RI_o, and RI_p can either be of same or different residues,
(ii) $RI_m \neq (RI_n$ or RI_o or $RI_p)$, i.e., RI_m must be of the different residue w.r.t to the other 3 atoms,
(iii) $m < n < o < p$, and
(iv) {*n*, *o*, *p* are consecutive} or, {o–$n=1$ and p–$o=1$}.

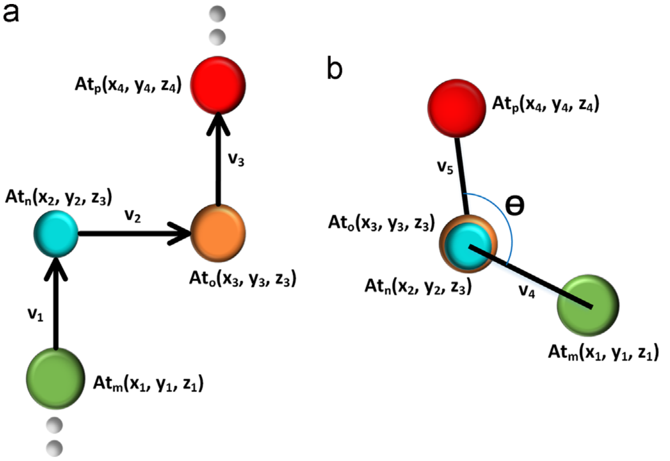


Fig. 3. (a) Atoms arrangement as well as vectors created using the Cartesian coordinates of the atoms. (b) The dihedral angle θ involving the four atoms.

Similarly, to compute the *uPsi*, a set of four atoms and their corresponding residue indexes are used which satisfies the following conditions:

(i) RI_m, RI_n, and RI_o can either be of same or different residue,
(ii) $RI_p \neq (RI_m$ or RI_n or $RI_o)$, i.e., RI_p must be of the different residue w.r.t to the other 3 atoms,
(iii) $m < n < o < p$, and
(iv) {*m*, *n* and *o* are consecutive} or, {n–$m=1$ and o–$n=1$}.

We compute the *uPhi* and *uPsi* using the $\langle x, y, z\rangle$ or, the Cartesian coordinates of the four atoms. Two planes are defined using four atoms. Plane 1 is formed by the atoms At_m, At_n and At_o and plane 2 is formed by the atoms At_n, At_o and At_p (see Fig. 2). The angle between these two planes is define as the dihedral angle, θ. To compute the dihedral angle, we first calculate three vectors v_1, v_2 and v_3 as shown in Fig. 3. Next, we calculate the normal vectors to both of the planes. The first normal vector is calculated by cross product of v_1 and v_2 (i.e., $v_1 \times v_2$) and named as v_4. In the same way, second normal vector is calculated by cross product of vector v_2 and v_3 (i.e., $v_2 \times v_3$) and named as v_5. The angle between these two normal vectors (i.e., v_4 and v_5) is then calculated via their dot product which provides the dihedral angle, θ. For mining the *uPhi* and *uPsi* information into energy score libraries, we use the training dataset of 4332 proteins as it was in (Mishra, 2015).

To compute the *uPhi* and *uPsi* energies, first, we obtain two different frequency distribution (FD) tables, namely FD_{uPhi} and FD_{uPsi}. Second, the range of *uPhi* and *uPsi* values, computed as the cosine value of the angle are mapped from -1 to 1, this range is divided into 20 bins, each with an equal width of 0.1. We considered 14,028 possible atom pairs that can be obtained from 167 different heavy atom types to represent the rows. Thus, both the FD_{uPhi} and FD_{uPsi} consist of 14,028 rows and 20 bins of equal width. The FD_{uPhi} for *uPhi* is updated using (1)

$$FD_{uPhi}(At_m, At_n, BI_{uPhi}) = FD_{uPhi}(At_m, At_n, BI_{uPhi}) + 1.0 \tag{1}$$

where, At_m and At_n are the two atoms from different residues in a protein structure and the BI_{uPhi} is the bin index computed from *uPhi*. While updating the FD_{uPhi}, we do not take the atom-pairs into consideration whose *distance* $(At_m, At_n) > 15$ Å. The bin index in (1) is expressed as in (2)

$$BI_{uPsi} = uPsi/Bin\_Width \tag{2}$$

where *Bin_Width* = 0.1. In the same way, the FD_{uPsi} for *uPsi* is updated using (3)

$$FD_{uPsi}(At_o, At_p, BI_{uPsi}) = FD_{uPsi}(At_o, At_p, BI_{uPsi}) + 1.0 \tag{3}$$

[1] http://cs.uno.edu/~tamjid/Software/3DIGARS/3DIGARS3.0.zip.

where, At_o and At_p are the two atoms from different residues in a given protein and the BI_{uPsi} is the bin index computed from *uPsi*. While updating the $FD_{uPsi,}$ we do not take into consideration atom-pairs whose *distance* $(At_o,\ At_p) > 15$ Å. The bin index in (3) is expressed as in (4)

$$BI_{uPsi} = uPsi/Bin\_Width \tag{4}$$

where, $Bin\_Width = 0.1$. After the corresponding FD is updated for all the 4332 protein structures from the dataset, we replaced the zero entries with a small value, 10^{-6}. The FD_{uPhi} and FD_{uPsi} are further utilized to compute the energy score e^{uPhi} and e^{uPsi} respectively for each cell given by (5) and (6) respectively.

$$e^{uPhi}(At_m, At_n, BI_{uPhi}) = \frac{\left\{FD_{uPhi}(At_m, At_n, BI_{uPhi})/\sum_{BI_{uPhi}} FD_{uPhi}(At_m, At_n, BI_{uPhi})\right\}}{\left\{\sum_{At_m,At_n} FD_{uPhi}(At_m, At_n, BI_{uPhi})/\sum_{BI_{uPhi}} \sum_{At_m,At_n} FD_{uPhi}(At_m, At_n, BI_{uPhi})\right\}} \tag{5}$$

$$e^{uPsi}(At_o, At_p, BI_{uPsi}) = \frac{\left\{FD_{uPsi}(At_o, At_p, BI_{uPsi})/\sum_{BI_{uPsi}} FD_{uPsi}(At_o, At_p, BI_{uPsi})\right\}}{\left\{\sum_{At_m,At_n} FD_{uPsi}(At_o, At_p, BI_{uPsi})/\sum_{BI_{uPsi}} \sum_{At_m,At_n} FD_{uPsi}(At_o, At_p, BI_{uPsi})\right\}} \tag{6}$$

where, $\sum_{BI_{uPhi}}$ indicates the summation over 20 bins, and $\sum_{At_m,\ At_n}$ indicates the summation over 14,028 atom pairs and $\sum_{BI_{gPhi}}$ $\sum_{At_m,\ At_n}$ indicates total of all the atom pairs summed over all the bins. Eqs. (5) and (6) are similar, except the computations are done using FD_{uPhi} and FD_{uPsi} respectively. Finally, the *uPhi* and *uPsi* based energies E^{uPhi} and E^{uPsi} for a given protein structure are computed using (7) and (8) respectively

$$E^{uPhi} = \sum_{uPhi} e^{uPhi}(At_m,\ At_n,\ BI_{uPhi}) \tag{7}$$

$$E^{uPsi} = \sum_{uPsi} E^{uPsi}(At_o,\ At_p,\ BI_{uPsi}) \tag{8}$$

where, $\sum_{uPhi}$ and $\sum_{uPsi}$ represents the summation over all possible *uPhi* and *uPsi* values respectively. Fig. 4 illustrates an overview of the computation of *uPhi* and *uPsi* based energies and then optimized incorporation to form 3DIGARS3.0.

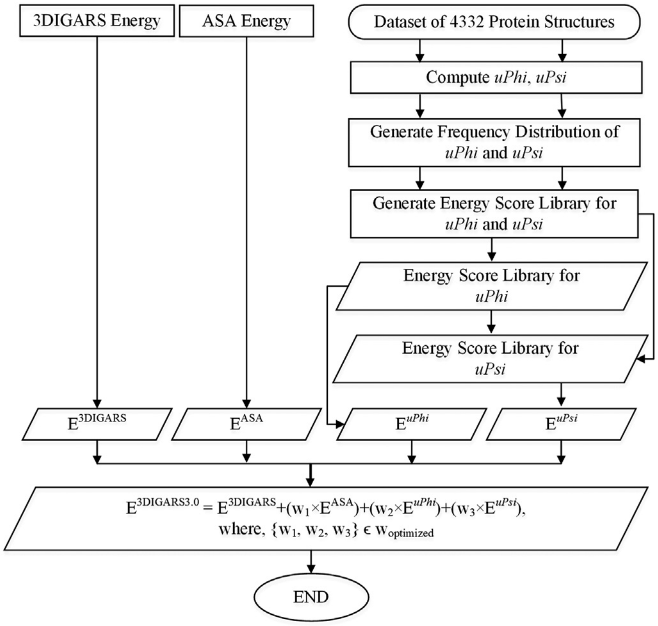

3DIGARS Energy
ASA Energy
Dataset of 4332 Protein Structures
Compute *uPhi*, *uPsi*
Generate Frequency Distribution of *uPhi* and *uPsi*
Generate Energy Score Library for *uPhi* and *uPsi*
Energy Score Library for *uPhi*
Energy Score Library for *uPsi*
$E^{3DIGARS}$
E^{ASA}
E^{uPhi}
E^{uPsi}
$E^{3DIGARS3.0} = E^{3DIGARS} + (w_1 \times E^{ASA}) + (w_2 \times E^{uPhi}) + (w_3 \times E^{uPsi})$,
where, $\{w_1, w_2, w_3\} \in w_{optimized}$
END

Fig. 4. Process flow of the design and development of 3DIGARS3.0 energy function.

2.2. The 3DIGARS3.0 energy function

3DIGARS3.0 is a linearly weighted combination of 3DIGARS, sequence-specific solvent-accessibility, *uPhi* and *uPsi* based energies which is expressed as in (9)

$$E^{3DIGARS3.0} = E^{3DIGARS} + (w_1 \times E^{ASA}) + (w_2 \times E^{uPhi}) + (w_3 \times E^{uPsi}) \tag{9}$$

where, the optimized values of the weights w_1, w_2 and w_3 are obtained by applying GA (Hoque et al., 2010; Bhandari et al., 1996; Hoque et al., 2009) over optimization dataset (see Section 3.1.2).

3. Datasets and results

This section first discusses the training, optimization and independent test datasets, then presents the results obtained by the proposed method including the analysis.

3.1. Datasets

The training datasets for 3DIGARS, sequence-specific solvent-accessibility, *uPhi* and *uPsi* energies were obtained from the PDB (Berman et al., 2000). We optimized the advanced energy function with three decoy datasets (Moulder, Rosetta and I-Tasser). Then we rigorously test our method on the independent test datasets (4state_reduced, fisa_casp3, hg_structal, ig_structal, ig_structal_hires).

3.1.1. Training datasets

The following datasets were used to compute the energy score libraries for 3DIGARS, sequence-specific solvent-accessibility, *uPhi* and *uPsi* based energies.

3.1.1.1. Training dataset for 3DIGARS ($E^{3DIGARS}$). 3DIGARS energy score libraries were created using the dataset obtained from the PDB (PDB, 2014) server. Multiple datasets, which differ based on the dataset collection parameters such as maximum resolution and sequence identity cutoff, were initially generated. Furthermore, proteins with unknown residue as well as with missing residues anywhere except for five terminal residues on either side were removed to avoid unstable statistics. Two different energy functions, RAPDF and DFIRE, were trained on these multiple datasets with varying configurations and tested using the three most challenging decoy sets. Then the dataset with the best performance was selected as the training dataset for the 3DIGARS energy function. In our implementation, we found that the dataset of 4332 proteins with resolution ≤ 2.5, single chain and sequence identity cutoff of 100% provided the best results. We believe that selecting proteins with 100% identity cutoff (which means we are not ignoring proteins even if they are structurally similar) provided better results because they provide us the true representation of the natural frequency-distributions.

3.1.1.2. Training dataset for sequence-specific solvent-accessibility energy (E^{ASA}). We prepared a new dataset from 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 chains) from PDB with the following specification: (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 the following 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 labeled '*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 in a dataset of 1299 sequences (SSD1299) and 272,800 residues. We determined the actual value of the surface area by the DSSP program (Kabsch and Sander, 1983) and the predicted value of surface area by REGAd^{3}p.

3.1.1.3. Training dataset for uPhi and uPsi energies (E^{uPhi}, E^{uPsi}). To generate *uPhi* and *uPsi* based score libraries we use the same dataset of 4332 known protein structures used by 3DIGARS method.

3.1.2. Optimization datasets

Following datasets were used to optimize the proposed energy function, 3DIGARS2.0.

3.1.2.1. Moulder decoy dataset. The Moulder (Sali, 2014) decoy set consists of 20 proteins for which 300 comparative models were built using a homologous template. The program, called MODELLER-6, was used to build the decoys. MODELLER-6 uses a default model building routine with fast refinement. Fast refinement keeps most of the template structure unchanged and is different from decoys generated by ab initio folding that have all structure regions reassembled from scratch.

3.1.2.2. Rosetta decoy dataset. The Rosetta datasets consists of 58 protein sets generated by the Baker Lab. Each set contains a native structure, 20 random models and the 100 lowest scoring models obtained from 10,000 decoys using ROSETTA de novo structure prediction followed by all-atom refinement (Zhang and Zhang, 2010; Tsai et al., 2003).

3.1.2.3. I-Tasser decoy dataset. The I-Tasser datasets consist of 56 protein sets. Each contains a native structure and around 300–500 decoys generated by both template-based modeling and atomic-level structural refinement. The I-Tasser (Lab, 2014) decoy set-II was generated by first using Monte Carlo Simulations and then refined by GROMACS4.0 MD simulation in order to remove steric clashes and improve hydrogen-bonding networks (Lab, 2014).

3.1.3. Independent test datasets

Five independent test decoy sets were used to evaluate the performance of the proposed energy function, 3DIGARS3.0 which are described below.

3.1.3.1. 4state_reduced decoy dataset. The 4state_reduced (Park and Levitt, 1996) decoy set consist of 7 proteins. The alpha-carbon positions for these decoys were generated by selecting ten residues in each protein using a 4-state off-lattice model. The all atom models were then built from the alpha-carbon atoms with the segmod package (Levitt, 2014).

3.1.3.2. fisa_casp3 decoy dataset. The fisa_casp3 (Simons et al., 1997) decoy set consist of 5 proteins. These decoys are the structures predicted by the Baker group for CASP3. The main chain for these decoys was generated using a fragment insertion simulated annealing procedure whereas, the side chains were modeled with the SCWRL package (Krivov et al., 2009).

3.1.3.3. hg_structal decoy dataset. The hg_structal (Levitt, 2014) set contains decoys for 29 globins (hg). Each globin is built through comparative modeling by using 29 other globins as template (Fogolari et al., 2007) applying segmod program (Levitt, 2014).

3.1.3.4. ig_structual decoy dataset. The ig_structal (Levitt, 2014) decoy set contains 61 immunoglobulins (ig). Each decoys in this set is built by comparative modeling or homology modeling using all the other immunoglobulins as templates. Most of the models have very low RMSD from the native (Fogolari et al., 2007).

3.1.3.5. ig_structual_hires decoy dataset. This set contains 20 immunoglobulins which are a high resolution subset of the ig_structal decoy set. The resolution range for this set is 1.7–2.2 Å compared to full set of 61 which has a resolution range from 1.7–3.1 Å. Also, the decoys in this set are built by comparative modeling or homology modeling using all the other immunoglobulins as templates and by applying program segmod (Levitt, 2014). Most of the models in this set also have very low RMSD from the native (Fogolari et al., 2007).

3.2. Results

The performance of 3DIGARS3.0 potential is evaluated based on the five independent test datasets (see Section 3.1.3). None of the proteins from the independent test dataset were either used in training or in optimization. Each of the decoy sets consists a of protein structure very close to the native one and the native one is also included within the set. Our objective here is to correctly identify the native structure out of the decoy structures present within each set. All the decoys, including the native structure, are first scored using the energy function. Next, the one with the minimum negative score is picked. If the structure picked is the native, we conclude the energy function is able to correctly identify the native protein out of its decoys.

To examine the effectiveness of a statistical prediction, the following three cross-validation methods are often used in practical application: independent dataset test, subsampling or *k*-fold cross validation test, and jackknife test. However, among these test methods, the jackknife test is deemed the least arbitrary and can always yield a unique result for a given benchmark dataset as elaborated in (Chou and Zhang, 1995). Though, the jackknife test has been widely recognized and increasingly used by researchers to examine the quality of various predictors (Cai, 2003; Dehzangi et al., 2015; Shen and Chou, 2007; Khan et al., 2015; Chou and Cai, 2005; Kumar et al., 2015; Mandal et al., 2015), however, it could be computationally very expensive. To reduce the computational time, we adopted the independent dataset test in this study.

We primarily validated the usefulness of the proposed energy components (*uPhi* and *uPsi*) used in formulating our energy function 3DIGARS3.0, by separate tests, shown in Table 2. There, to compare the combined effect, we tested the effects of the three orientation dependent energy components (θ_p, θ_q, θ_{pq}) proposed in dDFIRE with the 3DIGARS2.0. First, the orientation dependent energy components were added to all the decoy sets. Next, optimization was performed using optimization datasets and testing was done over independent test datasets. In Table 1, we show that the performance of 3DIGARS3.0 is -3.94%, -0.81% and -1.61% less than (3DIGARS2.0+dDFIRE), (3DIGARS2.0+*uPhi*) and (3DIGARS2.0+*uPsi*) respectively based on the optimization dataset. However, in Table 2, we show that the 3DIGARS3.0 outperforms (3DIGARS2.0+dDFIRE), (3DIGARS2.0+*uPhi*) and (3DIGARS2.0+*uPsi*) methods by 495%, 29.348% and 440.91% respectively based on the independent test dataset. The percentage of weighted average improvements are calculated using (10).

$$\%WA = \frac{\sum_1^n (y_i - x_i) * 100}{\sum_1^n x_i} \qquad (10)$$

where, y_i represents new value and x_i represents old value that are to be compared. Additionally, from Tables 1 to 4, the values outside of the parenthesis are the number of *correct counts* and the values within the parenthesis are the average *z*-score of the native structures. Correct counts are generated based on the assignment of the lowest energy scores to the number of native protein structures. For

Table 1
Performance comparison of several combined methods on optimization datasets based on correct native count.

Decoy Sets (no. of targets)	Methods			
	3DIGARS2.0 + dDFIRE	3DIGARS2.0 + uPhi	3DIGARS2.0 + uPsi	3DIGARS2.0 + uPhi + uPsi(3DIGARS3.0)
Moulder (20)	19 (−2.625)	19 (−3.239)	19 (−2.672)	**20** (−**3.851**)
Rosetta (58)	**52** (−**3.080**)	48 (−2.870)	49 (−2.987)	46 (−2.683)
I-Tasser (56)	**56** (−3.992)	**56** (−4.972)	**56** (−4.295)	**56** (−**5.573**)
Weighted average in %	−3.94	−0.81	−1.61	

Legend: entry format is native-count (*z*-score). **Bold** indicates best scores. Underscore indicates close to best scores.

Table 2
Performance comparison of several combined methods on independent test datasets based on correct-native count.

Decoy sets (no. of targets)	Methods			
	3DIGARS2.0 + dDFIRE	3DIGARS2.0 + uPhi	3DIGARS2.0 + uPsi	3DIGARS2.0 + uPhi + uPsi(3DIGARS3.0)
4state_reduced (7)	4 (−2.421)	6 (−3.063)	4 (−2.641)	**7** (−**3.456**)
fisa_casp3 (5)	**5** (−4.553)	**5** (−4.543)	**5** (−4.681)	4 (−4.076)
hg_structal (29)	11 (−1.576)	25 (−2.674)	12 (−1.583)	**28** (−**3.678**)
ig_structal (61)	0 (0.412)	40 (−1.120)	0 (0.267)	**60** (−**2.526**)
ig_structal_hires (20)	0 (0.219)	16 (−1.162)	1 (0.030)	**20** (−**2.378**)
Weighted average in %	495	29.348	440.91	

Legend: entry format is native-count (*z*-score). **Bold** indicates best scores. Underscore indicates close to best scores.

example, for the fisa_caps3 decoy set of five proteins sets, native structures of four protein-sets were assigned the lowest energy score among their respective decoy structures. Whereas for one of the protein sets, the lowest energy score was assigned to a decoy structure rather than a native structure. So, the correct count for fisa_caps3 decoy set is 4 out of 5. The rest of the decoys were similarly scored and their correct counts were obtained. The results of the addition of orientation dependent components (i.e., the dDFIRE) to 3DIGARS2.0 based on the optimization datasets and independent test datasets are shown in Tables 1 and 2 respectively (see second column). We can see that the addition of orientation dependent energy resulted in impressive results for optimization. But, it performed poorly for the independent test dataset. Following similar procedures, we tested the addition of *uPhi* and *uPsi* based energies to the 3DIGARS2.0 energy. It is clear that addition of *uPsi* performed slightly better than *uPhi* during optimization (see the results shown in Table 1, third column versus fourth column). However, the performance of *uPhi* on the independent test dataset was very impressive and outperformed *uPsi* with larger differences (see the results shown in Table 2, third column versus fourth column). Also, the performance of *uPhi* and *uPsi* during optimization are slightly less than dDFIRE (see the results shown in Table 1 to compare second, third and fourth columns). On the other hand, while testing the addition of *uPhi* and *uPsi*, we found that the performance of *uPhi* is significantly better than both *uPsi* and dDFIRE (see the results shown in Table 2 to compare second, third and fourth columns). Whereas, the performance of *uPsi* is slightly better than dDFIRE (see the results shown in Table 2, second column versus fourth column). Next, we combined *uPhi* and *uPsi* based energies components with 3DIGARS2.0 and we optimized and tested. The optimization resulted in slight decrement of correct count for Rosetta. Nevertheless, the correct count for Moulder increased and resulted in 20 out of 20 (see the results shown in fifth column of Table 1). Additionally, the test of this combination which we finally named 3DIGARS3.0, outperforms all the other methods significantly (see the results shown in fifth column of Table 2).

In addition, we compare the performance of 3DIGARS3.0 to its prior versions (3DIGARS and 3DIGARS2.0) as well as various state-of-the-arts approaches such as DFIRE, RWplus, dDFIRE and GOAP. We first optimize our method by GA using the optimization datasets and then perform the independent dataset test. However, overfitting was not a concern because our objective was to optimize the linear combinations of the energy components. Irrespectively, independent test were used and the outcome confirms the robustness of our method. Table 3 shows the performance comparison based on the optimization datasets. 3DIGARS3.0 outperforms DFIRE, RWplus, dDFIRE, GOAP and 3DIGARS by 38.64%, 28.42%, 56.41%, 11.93% and 18.45% respectively. Whereas, it was seen that the performance of 3DIGARS3.0 decreased by −1.61% while comparing with 3DIGARS2.0 which is a minor decrement compared to the improvement made with respect to other methods. In addition, these results are from the GA optimization and so, to have a more reliable evaluation, we compare the performance of 3DIGARS3.0 with respect to the same methods based on the independent test datasets, as shown in Table 4. Based on the independent test datasets, 3DIGARS3.0 outperforms DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS as well as 3DIGARS2.0 by 440.91%, 440.91%, 72.46%, 20.20%, 417.49% and 440.91% respectively. Note that the percentage of weighted average improvement while comparing 3DIGARS3.0 with 3DIGARS2.0 for optimization dataset is −1.61% whereas, in case of independent test dataset comparison is 440.91%.

Note that 3DIGARS is a pair wise distance based energy function and 3DIGARS2.0 is a linear combination of 3DIGARS with the sequence-specific solvent-accessibility based energy component included. Thus, only DFIRE and RWplus (Zhang and Zhang, 2010) may be directly comparable with 3DIGARS and 3DIGARS2.0. Furthermore, the dDFIRE combines DFIRE with the orientation dependent polar atom interactions. Also, the GOAP combines DFIRE with the relative orientation of the planes associated with each heavy atom in the interacting pairs. Thus, these energy functions may not be directly compared with 3DIGARS and 3DIGARS2.0 but with 3DIGARS3.0. Since, 3DIGARS3.0 uses the *uPhi* and *uPsi* which are similar 3D features as used by dDFIRE and GOAP. Results for DFIRE, RWplus, dDFIRE and GOAP are obtained from Zhou and Skolnick (2011). Likewise, the results for optimization dataset for 3DIGARS and 3DIGARS2.0 are obtained from Mishra (2015) and Iqbal et al. (2015) respectively. Additionally, the results for independent test for 3DIGARS and 3DIGARS2.0

Table 3
Performance comparison of different energy functions on optimization datasets based on correct native count.

Decoy sets (no. of targets)	Methods						
	DFIRE	*RWplus*	*dDFIRE*	*GOAP*	*3DIGARS*	*3DIGARS2.0*	*3DIGARS3.0*
Moulder (20)	$\underline{19}$ (−2.97)	$\underline{19}$ (−2.84)	18 (−2.74)	$\underline{19}$ (−$\underline{3.58}$)	$\underline{19}$ (−2.99)	$\underline{19}$ (−2.68)	**20** (−**3.851**)
Rosetta (58)	20 (−1.82)	20 (−1.47)	12 (−0.83)	$\underline{45}$ (−**3.70**)	31 (−2.023)	**49** (−$\underline{2.987}$)	$\underline{46}$ (−2.683)
I-Tasser (56)	$\underline{49}$ (−4.02)	**56** (−5.77)	48 (−5.03)	45 (−$\underline{5.36}$)	53 (−4.036)	**56** (−4.296)	**56** (−**5.573**)
Weighted average in %	38.64	28.42	56.41	11.93	18.45	−1.61	

Legend: entry format is native-count (*z*-score). **Bold** indicates best scores. Underscore indicates close to best scores.

Table 4
Performance comparison of different energy functions on independent test datasets based on correct native count.

Decoy sets (no. of targets)	Methods						
	DFIRE	*RWplus*	*dDFIRE*	*GOAP*	*3DIGARS*	*3DIGARS2.0*	*3DIGARS3.0*
4state_reduced (7)	$\underline{6}$ (−3.48)	$\underline{6}$ (−3.51)	**7** (−$\underline{4.15}$)	**7** (−**4.38**)	$\underline{6}$ (−3.371)	4 (−2.642)	**7** (−3.456)
fisa_casp3 (5)	$\underline{4}$ (−4.80)	$\underline{4}$ (−$\underline{5.17}$)	$\underline{4}$ (−4.83)	**5** (−**5.27**)	**5** (−4.319)	**5** (−4.682)	$\underline{4}$ (−4.076)
hg_structal (29)	12 (−1.97)	12 (−1.74)	16 (−1.33)	$\underline{22}$ (−$\underline{2.73}$)	12 (−1.914)	12 (−1.589)	**28** (−**3.678**)
ig_structal (61)	0 (0.92)	0 (1.11)	26 (−1.02)	$\underline{47}$ (−$\underline{1.62}$)	0 (0.645)	0 (0.268)	**60** (−**2.526**)
ig_structal_hires (20)	0 (0.17)	0 (0.32)	16 (−2.05)	$\underline{18}$ (−$\underline{2.35}$)	0 (−0.002)	1 (0.030)	**20** (−**2.378**)
Weighted average in %	440.91	440.91	72.46	20.20	417.39	440.91	

Legend: entry format is native-count (*z*-score). **Bold** indicates best scores. Underscore indicates close to best scores.

were obtained by running these methods on the test sets under this work.

4. Discussions

This section discusses different energy functions in a chronological order based on which the proposed method, 3DIGARS3.0 is evolved.

4.1. Average reference state

The residue specific all-atom probability discriminatory function (RAPDF) based potential was proposed by Samudrala and Moult (Samudrala and Moult, 1997), and uses averaging reference state. It involves the computation of the conditional probabilities for pairwise atom–atom interactions in proteins using statistical observation of the native structures. RAPDF reference state computes the probability of seeing any two atom types a and b in a distant bin, S distance apart which can be represented as

$$P(S_{ab}) = \sum_{ab} N(S_{ab}) / \sum_{S} \sum_{ab} N(S_{ab}) \tag{11}$$

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 $\sum_S \sum_{ab} N(S_{ab})$ is the total number of counts summed over all pairs of atom types summed over all the bins. As an averaging reference state, it does not consider experimental structures as 3-dimensional space where as DFIRE based potential considers proteins as a 3-dimensional sphere having radius r^{α} where α is a variable which can be ≤ 2.

4.2. Finite ideal-gas reference state

In the distance-scaled, finite ideal-gas reference, (Zhou and Zhou, 2002) acquired a pair-wise distribution function from statistical mechanics to formulate finite ideal-gas reference state. The expected number of atom pairs in a spherical system was computed by

$$N_{\exp}(i,j,d) = \left(\frac{d}{d_{cut}}\right)^{\alpha} \frac{\Delta d}{\Delta d_{cut}} N_{obs}(i,j,d_{cut}) \tag{12}$$

where $N_{obs}(i,j,d)$ represents the observed number of pairs of atoms, namely ith and jth atoms, at spatial distance d. The $d_{cut} = 14.5$ Å is a cut off distance and α represents the radius of the sphere which was determined by the best fit considering a number of points in 1011 finite protein size spheres. Eq. (12) is a formulation obtained from the ideal gas reference state that is implementable for a finite system.

4.3. Three-dimensional ideal gas reference state based energy ($E^{3DIGARS}$)

The 3-dimensional ideal reference state based energy function (Mishra, 2015) is an all-atom knowledge based potential based on the hydrophobic–hydrophilic model (*HP* model). It computes three different interaction energy libraries, namely, (i) hydrophobic–hydrophilic (*HP*), (ii) hydrophobic–hydrophobic (*HH*), and (iii) hydrophilic–hydrophilic (*PP*). Each interaction library maintains a uniform bin size of $\Delta r = 0.5$ Å for all bins, and a cutoff distance r_{cut} equal to 15 Å, where r represents each distant bin with values ranging from 0.5 Å to 15 Å. The value of $\Delta r_{cut} = 0.5$ Å as all bin sizes are the same. These three different libraries are computed using three different reference states. Reference state corresponding to the hydrophobic–hydrophilic group can be written as

$$N_{i,j}^{EXP-HP}(r) = \left(\frac{r}{r_{cut}}\right)^{\alpha_{hp}} \frac{\Delta r}{\Delta r_{cut}} (N_{obs-HP}(i,j,r_{cut}) + N_{obs-HH}(i,j,r_{cut}) + N_{obs-PP}(i,j,r_{cut})) \tag{13}$$

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

Similarly, HH reference state, $N_{i,j}^{EXP-HH}(r)$, and the *PP* reference states, $N_{i,j}^{EXP-PP}(r)$, varies from$N_{i,j}^{EXP-HP}(r)$, only in terms of the parameter α_{hh} and α_{pp} respectively. 4332 known protein structures were used to obtain the frequency computations. While computing the frequency distributions, residues corresponding to atom pairs are identified and classified to find the associated group (*HP*, *HH* or *PP*) and simultaneously update the frequency count of the group. Once the frequency distribution is computed, energy scores are obtained. Energy scores for *HP* group can be written as

$$ES_{i,j,r}^{HP} = -\ln\left(N_{obs-HP}(i,j,r)/N_{i,j}^{EXP-HP}(r)\right) \tag{14}$$

where $ES_{i,j,r}^{HP}$ represents the energy score of the atom pairs *i*th and *j*th at a distance bin *r* for group *HP*, $N_{obs-HP}(i,j,r)$ and $N_{i,j}^{EXP-HP}(r)$ are the observed and expected number of atom pairs *i*th and *j*th respectively at a distance bin *r* for the *HP* group. Energy scores for the other two groups *HH* and *PP* are also computed in a similar fashion as in (4) (see (Mishra, 2015)). Finally, the minimum energy of a protein structure can be obtained by (15)

$$TE = \beta_{hp}E_{HP} + \beta_{hh}E_{HH} + \beta_{pp}E_{PP} \tag{15}$$

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 the three groups *HP*, *HH* and *PP* respectively, where E_{HP} can be written as

$$E_{HP} = \sum_{i,j,r} ES_{i,j,r}^{HP} \tag{16}$$

Additionally, E_{HH} and E_{PP} are also calculated in a similar fashion as in (16). Furthermore, a GA is used to determine the best possible values of alpha (α_{hp}, α_{hh} and α_{pp}), and optimized the contributions of each of the three groups by determining their appropriate weights β_{hp}, β_{hh} and β_{pp} along with the *z*-score to discriminate the natives from their decoys, where the *z*-score of native structure is defined as

$$Z = \frac{E_{\text{native}} - E_{\text{average}}}{E_{SD}} \tag{17}$$

where, E_{native} is the energy of native protein, E_{average} is the average energy of all the decoys corresponding to its native protein excluding the native protein itself and E_{SD} is the standard deviation of the energies of all decoy sets.

4.3.1. *Predicted accessible surface area using REGAd^{3}p*

REGAd^{3}p is a real value predictor framework that combines the exact regularized regression with the optimization of weights by GA (Iqbal et al., 2015). The classical linear regression model can result in a poor fit to the data. Therefore, the kernel of the basic regression method is extended to a degree 3 polynomial equation. This basis expansion is performed by inserting two extra column vectors for each of the features which are the squares and cubes of the original feature values. Extending the kernel in such a way gives us the flexibility of model selection with higher order polynomials. However, increasing the degree of polynomial can cause overfitting, because of highly fluctuating weights. A model overfit to training data can give poor performance on test datasets. To overcome this overfitting problem, we implemented regularization, which involves adding a penalty term (regularization parameter) to the error in order to shrink the value of the weights. We performed a search for the best value of the regularization parameter within the range [−100.0 to +100.0] with an interval size equal to 2.0. The weights computed from regularization are further optimized using a GA. The optimization is carried out to minimize the minimization of Mean Absolute Error (MAE) as well as to maximize the Pearson Correlation Coefficient (PCC) between actual and predicted ASA values. The parameter values of our GA implementation are: (i) population size=200, (ii) number of generations=2000, (iii) chromosome length=number of weights × number of bits for each weight=1155 × 18 bits, where the weights 1155 come from features (55) × window-size (21), (iv) elite rate=10%, (v) crossover rate=80% and (vi) mutation rate=70%. While generating the initial population, 100 individuals were taken from the output of regularization and the rest were generated randomly. 10% best performing weight sets are always forwarded to the next generation's population from the current one. The high rate of mutation aided in finding new and improved solutions within the large and complex search space of real ASA. To select the candidates for crossover, we implemented the roulette wheel selecting algorithm to sample highly fitted individuals to be utilized for the next generation's population. Furthermore, we integrated a post processing of predicted ASA values within our GA to avoid the 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.

4.3.2. *Sequence-specific solvent-accessibility energy (E^{ASA})*

Solvent accessible energy is closely related to the hydrophobic interaction, stimulating the investigation into the low-frequency internal motion of proteins and their biological function (Chou, 1988; Chen, 1977). We computed sequence-specific energy from accessible surface area (ASA), E$^{\text{ASA}}$, by modeling the error between the actual and predicted ASA. We obtained the actual ASA (ASA$^{\text{Actl}}$) and predicted ASA (ASA$^{\text{Pred}}$) of each residue for 1299 proteins using the software DSSP (Wolfgang and Christian, 1983) and REGAd^{3}p (Iqbal et al., 2015) respectively. The error between the actual and predicted ASA ($\Delta ASA_k = ASA_k^{\text{Actl}} - ASA_k^{\text{Pred}}$) of a given residue for a specific amino acid type is used to obtain the frequency distribution (FD). To compute the frequency distribution, we first calculate the max error ΔASA from the dataset of 1299 proteins which was found to be 240. We then divided the error ranging from 0 to 195 with a bin width of 5 to obtain 39 bins of equal size. The error range remaining after subtracting 195 from 240, which is equal to 45, is considered as the 40th bin of the frequency table. To get all the 40 bins to be of equal size, we normalized the values of the last bins by dividing them by 9 (45/9=5). This resulted in a frequency distribution table of 20 rows (for 20 different amino acid types) and 40 bins, each with an equal size of 5. Mathematically, for each residue *"k"* of a specific amino acid type, the frequency distribution table is updated as in (18)

$$FD(AA(k), Bin\_Index) = FD(AA(k), Bin\_Index) + 1.0 \tag{18}$$

where, amino acid type of the *k*th residue is obtained by *AA(k)* and *Bin_Index* of the *k*th residue can be expressed as in (19)

$$Bin\_Index = abs(\Delta \text{ASA}_k = \text{ASA}_k{}^{\text{Actl}} - \text{ASA}_k{}^{\text{Pred}})/Bin\_Width \tag{19}$$

where, *Bin_Width*=5. Once we update the frequency table for all the 1299 proteins residues of the dataset, cells for which the frequency count is found to be zero are replaced by a small value of 10^{-6}. The foregoing process provides us with a frequency table which is further used to compute the probability *P* of each cell given by (20)

$$P = FD(AA(k), Bin\_Index)/TF \tag{20}$$

where, TF is the sum of the counts of each amino acid type of all the bins in the frequency table. The energy score library (ESL) for sequence-specific solvent-accessibility is obtained by (21)

$$ESL = -\ln(Bin\_Value \times P) \tag{21}$$

where, *Bin_Value* is the frequency count of each cell. Described above is the method to obtain the ESL. We use this ESL to compute the sequence-specific solvent-accessible energy E^{ASA} for a given protein structure to compute, the E^{ASA} of a given protein sequence, we first compute the actual and predicted ASA of each residue (R_i)

using DSSP and REGAd^{3}p software respectively. Next, we pick the energy associated with each reside from ESL based on the difference between the actual and predicted ASA values. We use the error between the actual and predicted ASA to compute the *Bin_Index* which is obtained from (19) and then obtain the energy for each residue "*k*" by (22)

$$E_K = ESL(AA(k), Bin\_Index) \tag{22}$$

Finally, the energy of the complete protein sequence of length *N* is computed by (23)

$$E^{\mathrm{ASA}} = \sum_{K=1}^{N} E_k \tag{23}$$

4.3.3. Energy function, 3DIGARS2.0

3DIGARS2.0 (Iqbal et al., 2015) is an optimized combination of the above mentioned 3DIGARS energy, E$^{\mathrm{3DIGARS}}$, and sequence-specific solvent-accessibility energy, E^{ASA}, given by

$$E^{\mathrm{3DIGARS2.0}} = E^{\mathrm{3DIGARS}} + (w \times E^{\mathrm{ASA}}) \tag{24}$$

where, *w* is the variable ranging from 0 to 2, whose best possible value is obtained using a GA. The GA parameters were population size = 300, elite rate = 5%, crossover rate = 90% and mutation rate = 50%. The stopping criteria, *max_iteration* was set to 2000 iterations. The objective function was a linear combination of the correct count and the average *z*-score of the three most challenging decoy sets: Moulder, Rosetta and I-Tasser. *Correct count* is defined as the number of correctly identified native protein structures from its decoy sets. A better energy function assigns highest negative energy to the native protein compared to its decoy sets and thus is able to classify native proteins from its decoy sets more efficiently. The count of such correctly identified native protein structures is termed as correct count.

5. Conclusions

Predicting the 3D structure of a protein from its amino acid sequence alone has established significant popularity in the past two decades because of its wide spread importance in drug design as well as design of novel enzymes. Energy functions are one of the fundamental components for solving protein structure and folding prediction problems. Hence, we develop a new energy function, 3DIGARS3.0 to improve the accuracy of tertiary protein structure prediction or protein folding methods. Being motivated by the fact that the 3D structural information assists the advancement of the accuracy of the energy function, we introduce two new 3D structural features *uPhi* and *uPsi*. We linearly combine these *uPhi* and *uPsi* based energies with our prior energy components, 3DIGARS and predicted ASA based energies. This linear combination was further optimized using three groups of challenging decoy data-sets using a GA and tested on five independent test datasets. The importance of individual features *uPhi* and *uPsi* was analyzed by taking different combinations of *uPhi*, *uPsi* and our prior energy component and code function. In addition, we also analyzed the effect of addition of orientation angles by dDFIRE. The addition of *uPhi* and *uPsi* to 3DIGARS and ASA outperformed all other combinations based on the independent test datasets. 3DIGARS3.0 outperformed the state-of-the-arts approaches such as DFIRE, RWplus, dDFIRE, GOAP, 3DIGARS and 3DIGARS2.0 by 440.91%, 440.91%, 72.46%, 20.20%, 417.39% and 440.91% respectively, based on the independent test datasets.

Supplementary Content

The software is available online http://cs.uno.edu/~tamjid/Software/3DIGARS/3DIGARS3.0.zip.

Acknowledgments

We gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19. We would also like to thank Mr. Glenn Robert McLellan for his helpful reviews for English Language.

References

Berardi, M.J., et al., 2011. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476, 109–113.

Berman, H., et al., 2000. The protein data bank. Nucl. Acids Res. 28, 235–242.

Bhandari, D., Murthy, C.A., Pal, S.K., 1996. Genetic algorithm with elitist model and its convergence. Int. J. Pattern Recognit. Artif. Intell. 10 (06), 731–747.

Borguesana, B., et al., 2015. APL: Anangleprobabilitylisttoimproveknowledge-based metaheuristics forthethree-dimensionalproteinstructureprediction. Comput. Biol. Chem. 59, 142–157.

Brüschweiler, S., et al., 2015. Substrate-modulated ADP/ATP-transporter dynamics revealed by NMR relaxation dispersion. Nat. Struct. Mol. Biol. 22, 636–641.

Brooks, B.R., et al., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187–217.

Cai, Y.-D., 2003. Predicting protein quaternary structure by pseudo amino acid composition. Protein: Struct. Func. Genet. 53, 282–289.

Carlacci, L., Chou, K.-C., Maggiora, G.M., 1991. A heuristic approach to predicting the tertiary structure of bovine somatotropin. Biochemistry 30, 4389–4398.

Carter, D.B., Chou, K.-C., 1998. A model for structure-dependent binding of Congo red to Alzheimer β-amyloid fibrils. Neurobiol. Aging 19, 37–40.

Chen, N.-Y., 1977. The biological functions of low-frequency phonons. Sci. Sin. 20, 447–457.

Chen, W., et al., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucl. Acids Res., 41 (p. gks1450).

Chou, K.C., et al., 1992. An energy-based approach to packing the 7-helix bundle of bacteriorhodopsin. Protein Sci. 1, 810–827.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273 (1), 236–247.

Chou, K.-C., et al., 1985. Interactions between an alpha-helix and a beta-sheet. Energetics of alpha/beta packing in proteins. J. Mol. Biol. 186, 591–609.

Chou, K.-C., 1988. Low-frequency collective motion in biomacromolecules and its biological functions. Biophys. Chem. 30, 3–48.

Chou, K.-C., 2004. Structural bioinformatics and its impact to biomedical science. Curr. Med. Chem. 11, 2105–2134.

Chou, K.-C., Scheraga, H.A., 1982. Origin of the right-handed twist of beta-sheets of poly (LVal) chains. Proc. Nat. Acad. Sci. 79, 7047–7051.

Chou, K.-C., Carlacci, L., 1991. Energetic approach to the folding of α/β barrels. Protein: Struct. Funct. Bioinf. 9, 280–295.

Chou, K.-C., Zhang, C.-T., 1995. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.

Chou, K.-C., Cai, Y.-D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Model. 45, 407–413.

Chou, K.-C., Maggiora, G.M., Scheraga, H.A., 1992. Role of loop–helix interactions in stabilizing four-helix bundle proteins. Proc. Nat. Acad. Sci. 89, 7315–7319.

Chou, K.-C., Wei, D.-Q., Zhong, W.-Z., 2003. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochem. Biophys. Res. Commun. 308, 148–151.

Cornell, W.D., et al., 1995. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. 117, 5179–5197.

Dehzangi, A., et al., 2015. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. 364, 284–294.

Ding, H., et al., 2014. iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed. Res. Int. 2014.

Fogolari, F., et al., 2007. Scoring predictive models using a reduced representation of proteins: model and energy definition. BMC Struct. Biol.

Gohlke, H., Hendlich, M., Klebe, G., 2000. Knowledge-based scoring function to predict protein–ligand interactions. J. Mol. Biol. 295, 337–356.

Hooft, R.W., Sander, C., Vriend, G., 1997. Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci. 13, 425–430.

Hoque, M.T., et al., 2010. DFS generated pathways in GA Crossover for protein structure prediction. Neurocomputing 73, 2308–2316.

Hoque, Md. Tamjidul, et al., 2016. sDFIRE: sequence-specific statistical energy function for protein structure prediction by decoy selections. J. Comput. Chem.

Hoque, T., Chetty, M., Sattar, A., 2009. Extended HP model for protein structure prediction. J. Comput. Biol. 16, 85–103.

Iqbal, S., Mishra, A., Hoque, T., 2015. Improved prediction of accessible surface area results in efficient energy function application. J. Theor. Biol. 380, 380–391.

Jernigan, R.L., Bahar, I., 1996. Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6, 195–209.

Jia, J., et al., 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.

Jia, J., et al., 2015. Identification of protein–protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn., 1–16.

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

Khan, Z.U., Hayat, M., Khan, M.A., 2015. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. 365, 197–203.

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.

Krivov, G.G., Shapovalov, M.V., Dunbrack, R.L., 2009. Improved prediction of protein side-chain conformations with SCWRL4. Protein: Struct. Func. Bioinf. 77, 778–795.

Kumar, R., et al., 2015. Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. 365, 96–103.

Lab, Z., 2014. Protein structure decoys (July). Available from: ⟨http://zhanglab.ccmb.med.umich.edu/decoys/⟩.

Lehninger, A.L., Nelson, D.L., Cox, M.M., 2005. Principles of Biochemistry. W.H. Freeman and Company, New York, USA.

Lesk, A.M., 2004. Introduction to Protein Science, 2nd ed. Oxford University Press, New York, p. 310.

Levitt, M., 2014. Accurate Modeling of Protein Conformation by Automatic Segment Matching. [cited 2014; Web (July). Available from: ⟨http://www.ncbi.nlm.nih.gov/pubmed/1640463]⟩.

Lin, H., et al., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucl. Acids Res. 42, 12961–12972.

Lodish, H., et al., 1990. Molecular Cell Biology, 5th ed. Scientific American Books, W. H. Freeman, New York, USA.

Mandal, M., Mukhopadhyay, A., Maulik, U., 2015. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med. Biol. Eng. Comput. 53, 331–344.

Mishra, A., 2015. Three-Dimensional Ideal Gas Reference State based Energy Function, Department of Computer Science, University of New Orleans ⟨http://scholarworks.uno.edu/⟩.

Mitchell, J.B., et al., 1999. BLEEP—potential of mean force describing protein–ligand interactions: II. Calculation of binding energies and comparison with experimental data. J. Comput. Chem. 20, 1177–1185.

Mitchell, J.B., et al., 1999. BLEEP—potential of mean force describing protein–ligand interactions: I. generating potential. J. Comput. Chem. 20, 1165–1176.

Muegge, I., Martin, Y.C., 1999. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J. Med. Chem. 42, 791–804.

OuYang, B., et al., 2013. Unusual architecture of the p7 channel from hepatitis C virus. Nature 498, 521–525.

Park, B., Levitt, M., 1996. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J. Mol. Biol. 258, 367–392.

PDB, R. Advanced Search Interface. February 2014; Available from: ⟨http://www.rcsb.org/pdb/search/advSearch.do'⟩.

Ramachandran, G.N., Ramachandran, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7, 95–99.

Sali, A., 2014. Decoy Models (July). Available from: ⟨http://salilab.org/john_decoys.html⟩.

Samudrala, R., Moult, J., 1997. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J. Mol. Biol. 275, 895–916.

Shen, H.-B., Chou, K.-C., 2007. A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.

Simons, K.T., et al., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. 268, 209–225.

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. Protein: Struct. Funct. Bioinf. 41, 40–46.

Tsai, J., et al., 2003. An improved protein decoy set for testing energy functions for protein structure prediction. Protein: Struct. Funct. Bioinf. 53, 76–87.

Wang, S.-Q., et al., 2009. Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus. Biochem. Biophys. Res. Commun. 386, 432–436.

Wolfgang, K., Christian, S., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637.

Xu, Y., et al., 2014. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. Plos One, 9, p. e105018.

Yang, J., Zhang, Y., 2015. I-TASSER server: new development for protein structure and function predictions. Nucl. Acids Res. 43, W174–W181.

Yang, Y., Zhou, Y., 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72, 793–803.

Zhang, C., et al., 2005. A knowledge-based energy function for protein–ligand, protein–protein, and protein-DNA complexes. J. Med. Chem. 48, 2325–2335.

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 (10).

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. 11, 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.

Zi Liu, et al., 2015. iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition. Anal. Biochem. 474, 69–77.