Three-Dimensional Ideal Gas Reference State based Energy Function

Avdesh Mishra[a], Md Tamjidul Hoque[*a]

Note: Accepted in Current Bioinformatics, Bentham Journal in 2015.

[a]*Department of Computer Science, University of New Orleans, LA-70148, USA*

Abstract: The major improvements in energy function that can discriminate a native-like protein structure from the decoys are done based on the development of better reference state. We propose an improved 3-dimensional energy function based on hydrophobic-hydrophilic properties of amino acid, where instead of having the neighboring residues placed in a modified spherical environment controlled only by a single dimensional parameter (alpha) as done in well know ideal gas reference state based approach, our 3D values for alpha represent three different groups of interactions, namely hydrophobic versus hydrophilic, hydrophobic versus hydrophobic and hydrophilic versus hydrophilic to segregate the statistics which in turn enable us to model the interactions more accurately instead of grossly averaging the impact. Using a Genetic Algorithm (GA), we optimized the contributions of each of the three group along with the z-score to discriminate the native from the decoys. Our all-atom based energy function generates scores from a database of 4332 protein structures obtained from Protein Data Bank (PDB). Our energy function, 3 Dimensional Ideal Gas Reference State based (3DIGARS) is found to be very competitive compared to the state-of-art-approaches, and for the most challenging Rosetta decoy-set 3DIGARS outperforms the nearest competitor by 40.9%.

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

1. INTRODUCTION

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 decent potential that can discriminate between native and 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 function 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 modelled based on simplified representation of the amino acid considering few heavy atoms and few major forces, some 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 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 relatively complete and detail approaches are the all atom based approaches. The efficacy of the all-atom based approach relay 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 fundamental work of DFIRE of proposing ideal gas reference state based approach, to formulate the reference state, the neighboring residues are placed in a modified spherical environment controlled by a single dimensional parameter (alpha), where the alpha value implicates a constant factor assuming amino-acid distribution in a protein conformation as a finite system [10]. On the 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 to consider them in the same scale on an average by a single dimensional parameter Fig. (1(a)). Rather they can be segregated into at least 3 different categories based on the usual distribution with the protein conformations Fig. (1(b)). Related to this, in one of the dominating properties of protein folding, modeled in a hydrophobic-hydrophilic or hydrophobic-polar (HP) model, considers hydrophobic (H) amino acids having fear of solvent like water, they want to keep themselves away from aqueous solvent forming the code or inner-kernel [22] of protein and thus remain inside of

*Address correspondence to this author at the (Md Tamjidul Hoque) Department of Computer Science, University of New Orleans, Louisiana, USA. E-mails: thoque@uno.edu

a protein. And, on the other hand, the hydrophilic or the polar (P) amino acid or residues being the attracted to water of water, try to remain outside the core, forming the outer-kernel Fig. (1(b)). Thus Ps are often found on the outside of their folded structure [23, 24], and in between this two layer there is a thin HP-mixed-layer [22]. Following these aforementioned properties, we proposed our multidimensional reference states based energy function 3DIGARS (3 Dimensional Ideal Gas Reference State) for improved accuracy.

The rest of the paper is organized as follows. Section 2 discusses the evolution of the relevant theories and underpins theoretical aspect 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-art-approaches. Section 4, discussed the obtained results and finally 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: conditional probability based approach and free energy based approach. It was found that these two approaches are equivalent for all practical purposes while it is more easier to work with conditional probability based approach because, of the Boltzmann assumption on three different aspect of it: an equilibrium distribution of atom pairs, the physical nature of the reference state and the probability of observing a system in any given state which 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 observation of native structures [9] from protein-databases such as PDB (Protein Data Bank) [25]. The conditional probabilities are based on two different type 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 $\{S_{ab}^{ij}\}$ belongs to native conformation can be represented by $P(N \mid S_{ab}^{ij})$. Therefore, we can write the general formula of conditional probability that, atom pairs separated by distance $\{S_{ab}^{ij}\}$ belongs to native conformation as:

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

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]

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

$$\text{and } P(S_{ab}^{ij}) = \prod_{ij} P(S_{ab}^{ij})$$

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

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

$P(N)$ in above equation is a constant and independent of conformation of given structure and so it can be omitted from further consideration. Omission of $P(N)$ implicates 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 to give a form similar to that of potential of mean force. Scoring function *SF* proportional to the negative log conditional probability that the structure is correct can be written as:

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

Therefore, given a protein structure or conformation, using Eq. 4, we can calculate all 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:

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

where $N(S_{ab})$ is the frequency of observation of atom type *a* and *b* in a particular bin which is of *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 represents 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:

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

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 acid in 3D conformational space whereas DFIRE based potential considers protein as a sphere comprising of uniformly distributed points and also suggest that the radius of such spheres does not increase in r^2 as in an infinite system rather protein is a finite system and so the increase in

the radius is represented by α (a variable which can be ≤ 2). This involved our concerns toward more detailed study into 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. The reference state RAPDF uses an averaged distribution over all residue or atom pairs whereas, DFIRE uses pair distribution function from statistical mechanics to formulate finite ideal-gas reference state.

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

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

where volume of the system is represented as v^s, 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:

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

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

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

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

Thus, Eq. 9 can be written as:

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

Further, Eq. 8 can be rewritten as:

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

Assuming that there is no interaction beyond 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:

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

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

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

It is to be noted that the Eq. 13 does not contains any distance dependent term rather it is a formulation obtained from ideal gas reference state 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* have different width 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 (HP 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 size are same. We name these libraries as $i)$ hydrophobic-hydrophilic (HP); $ii)$ hydrophobic-hydrophobic (HH); and $iii)$ hydrophilic-hydrophilic (PP) interactions libraries and each of these libraries comprises of its independent reference state. Reference state corresponding to hydrophobic-hydrophilic group can be written as:

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

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

Similarly, reference state corresponding to hydrophobic-hydrophobic group can be written as:

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

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

Finally, reference state corresponding to hydrophilic-hydrophilic group can be written as:

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

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

While generating energy score libraries, residue-atom pairs are categorized to identify which of the group (HP, HH or PP) mentioned above they fall in e.g. while considering interaction between two Nitrogen (N) atom of amino acid Alanine (ALA:N versus ALA:N), we categorize this interaction as hydrophobic-hydrophobic (HH) group as amino acid ALA (Alanine) is hydrophobic in nature. Similarly, while considering interaction between Nitrogen (N) atom of amino acid Arginine (ARG) and Carbon (C) atom of amino acid Serine (SER); (ARG:N versus SER:C), we categorize this interaction as hydrophilic-hydrophilic (PP) as both residues Arginine (ARG) and Serine (SER) are hydrophilic in nature. The categorization of amino acid into hydrophobic and hydrophilic group is obtained from (Hoque et al.)[24]. Along with the categorization of residue-atom pairs the frequency counts of the specific group is updated simultaneously. Further these energy score libraries are used for total energy or minimum energy calculation. Once we compute frequencies of all the 3 groups, we generate energy scores corresponding to each group. Energy score for HP group can be written as:

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

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 expected number of atom pairs at distance *r* for HP group.

Similarly energy score for HH group can be written as:

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

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 expected number of atom pairs at distance *r* for HH group.

Finally energy score for PP group can be written as:

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

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 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 β_{hp}, β_{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:

$$TE = \beta_{hp}E_{hp} + \beta_{hh}E_{hh} + \beta_{pp}E_{pp} \quad (20)$$

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:

$$E_{HP} = \sum_{i,j,r} ES_{i,j,r}^{HP} \quad (21)$$

Similarly, E_{hh} can be written as:

$$E_{HH} = \sum_{i,j,r} ES_{i,j,r}^{HH} \quad (22)$$

And, E_{pp} can be written as:

$$E_{PP} = \sum_{i,j,r} ES_{i,j,r}^{PP} \quad (23)$$

We use Genetic Algorithm (GA) for determining the best possible values of alpha (α_{hp}, α_{hh} and α_{pp}), and optimized the contributions of each of the three group by determining their appropriate weights β_{hp}, β_{hh}, and β_{pp} along with the z-score to discriminate the native from their decoys, where z-score of native structure is defined as:

$$Z = \frac{E_{native} - E_{average}}{E_{SD}} \quad (24)$$

where E_{native} is the energy of native protein, $E_{average}$ and E_{SD} is the standard deviation of the energies of all the decoy sets.

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

The obtained best values of 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. Plot of obtained fitness versus α_{hp}, α_{hh} and α_{pp} values at each generation shows the GA performance on selecting best fitness and also constancy of obtained fitness with values of α_{hp}, α_{hh} and α_{pp}. Fig. (2).

To access the performance of our 3DIGARS energy function we tested it with most challenging decoy-sets in Table **1** against the state of 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.

3. DATASET COLLECTION AND DECOY DATASETS

3.1. Training dataset

Datasets to generate energy score were obtained from three different sources, PDB [25] (Protein Data Bank) server, ccPDB [26] (Compilation and Creation of datasets from PDB) server and PISCE [27] server. We collected dataset with maximum resolution ≤ 1.5, similarity cutoff 30%, single chain and maximum chain length of 500.

Furthermore, we remove proteins with unknown residue as well as missing residues anywhere except for 5 terminal residues on either sides. We generated purified dataset keeping all other specification common besides maximum resolution ranging from 1.5 to 2.5 and sequence identities cutoff, of 25%, 30%, 40%, 50%, 70% and 100%. The best result overall of these combination is obtained from collection of 4332 proteins from PDB which are single chain with 100% sequence identity cut-off. Selecting proteins with 100% identity cutoff mean we are actually not ignoring proteins even if they are structurally similar because they provide us the true representative sample of the natural distribution of the frequency. Neither any of the structures from the training dataset were used into the test dataset (Decoy Datasets) nor were any of the structures from test datasets included in training dataset.

3.1. Decoy datasets

We tested the performance of 3DIGARS with respect to the most challenging 3 decoy datasets, which are described in brief as follows:

3.1.1. Moulder Decoy-set:

Moulder [28] decoy set consist of 20 proteins for which 300 comparative models were built using homologous template. The models were build using alignment that did not shared more than 95% of identically aligned positions or had at least 5 different alignment positions. These decoys were build using MODELLER-6 program which applied default model building routine with fastest refinement which keeps most of the template structure unchanged and are different from decoys that are generated by ab initio folding that have all structure regions reassembled from scratch.

3.1.2. Rosetta Decoy-set:

Decoy set for 58 proteins were generated by Baker Lab which contains 20 random models and 100 lowest scoring models from 10,000 decoys using ROSETTA de novo structure prediction followed by all-atom refinement [29]. Current Rosetta decoy set has been improved than the original Rosetta decoy set by adding side chains to the centroid/backbone models and refining the structures to remove steric clashes. The improvement over original Rosetta were based on four important points required to generate optimal decoy sets 1) decoy set should contain conformations for a wide variety of different proteins to avoid over fitting; 2) decoy set should contain conformation close to (< 4Å) to the native structure; 3) decoy set should consist of conformations that are at least near local minima of energy potential; and 4) decoy set should be produced without using information from native structure [30].

3.1.3. I-Tasser Decoy-set-II:

I-Tasser [31] decoy set-II were generated first by using Monte Carlo Simulations and then refined by GROMACS4.0 MD simulation to remove steric clashes and improve hydrogen-bonding network [31]. This set contains of 56 proteins each of which contains 300-500 decoys generated by both template-based modeling and atomic-level structure refinement.

4. RESULTS:

In Table **1** value within the parenthesis are average z-scores of the native structures and values outside of parenthesis are number of correct count. Here the term correct count can be described as number of correctly identified native proteins from its decoy sets. Good energy function is the one which can assign highest energy to the native proteins compared to its decoy sets and thus is able to classify native proteins from its decoy sets more efficiently. In other words correct count implicates that an efficient energy function can identify more native proteins from the collection of native and its 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 [32]. Correct count and z-score for DFIRE2.0 is computed from DFIRE2.0 package freely available from the Sparks Lab online server [33]. Correct counts by (3DIGARS) is calculated using energy score libraries generated using the dataset with resolution 2.5, sequence similarity cutoff of 100%, keeping all other parameters used for data collection common as described in DATASET section above. Table **1** clearly shows that hydrophobic and hydrophilic based energy function outperform DFIRE, RWplus, dDFIRE and DFIRE2.0 based energy functions for most challenging Rosetta decoy-set. It is to be noted that RWplus computed 56 out of 56 for their own designed Tasser decoy-set, which could be a special case, as it is not consistently better in other cases.

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
Tasser	49 (-4.02)	**56** (-5.77)	48 (-5.03)	53 (-4.548)	53 (-4.036)	56

Bold indicates best score and underline indicates competitive score.

CONCLUSION

Identifying native proteins from their decoy sets have always been a challenging task. While exercising with the two different reference state implementation, RAPDF and DFIRE, we formulated a better energy function based on the training dataset, hydrophobic and hydrophilic property of amino acid and their role in 3D structure formation, 3D values of alpha based on hydro-phobic and hydrophilic residues spatial distributions and by optimized the weights 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 category of residues is given their independent and more appropriate semi-spherical distribution parameter alphas, and then we determine their best values instead of having a single parameter based gross average inaction model.

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

We compare the performance of our proposed 3DIGARS with the state-of-art-approaches DFIRE, RWplus, dDFIRE and DFIRE2.0 using the most challenging three different 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%.

CONFLICT OF INTEREST

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

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.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19. We also acknowledge the discussion with Sumaiya Iqbal and Md Nasrul Islam.

SUPPLEMENTARY CONTENT

The software, dataset and related material is available free of charge via the Internet 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; 44:223-232.

[2] Moult J. Comparison of Database Potentials and Molecular Mechanics Force Fields. Curr Opin in Str Bio. 1997; 7:194-199.

[3] Vajda S, Sippl M, Novotny J. Empirical Potentials and Functions for Protein Folding and Binding. Curr Opin in Str Bio. 1997; 7:222-228.

[4] Hao M-H, Scheragat HA. Designing Potential Energy Functions for Protein Folding. Curr Opin in Str Bio. 1999; 9:184-188.

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

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

[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; 117:5179-5197.

[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. 1997:895-916.

[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. 2002:2714–2726.

[11] Tanaka S, Scheraga HA. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules 1976; 9:945-950.

[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; 5:1043-1059.

[14] Tobi D, Elber R. Distance-Dependent, Pair Potential for Protein Folding: Results From Linear Optimization. Proteins: Struct., Funct., Bioinf. 2000; 41:40-46.

[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; 326:1239-1259.

[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; 11:2714-2726.

[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: 2043–2052.

[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; 80:2311–2322.

[22] 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.

[23] Fidanova S. An Improvement of the Grid-based Hydrophobic-Hydrophilic Model. Int. J. Bioautomation 2010; 14:147-156.

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

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

[26] Singh H, Chauhan JS, Gromiha MM *et al.* ccPDB: Compilation and Creation of Data Sets from Protein Data Bank. Nucleic Acids Research 2011; 40.

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

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

[29] 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.

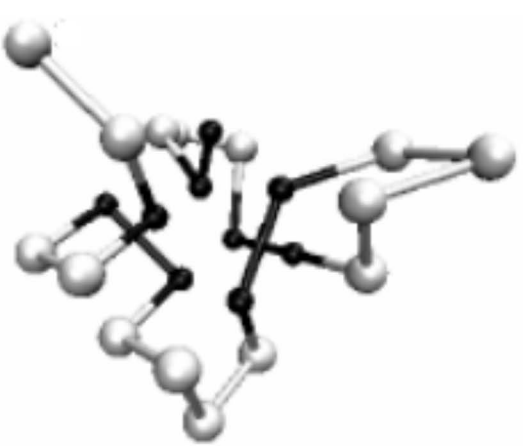
[30] Tsai J, Bonneau R, Morozov AV *et al.* An Improved Protein Decoy Set for Testing Energy Functions for Protein Structure Prediction. Proteins: Struct., Funct., Bioinf. 2003; 53:76-87.

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

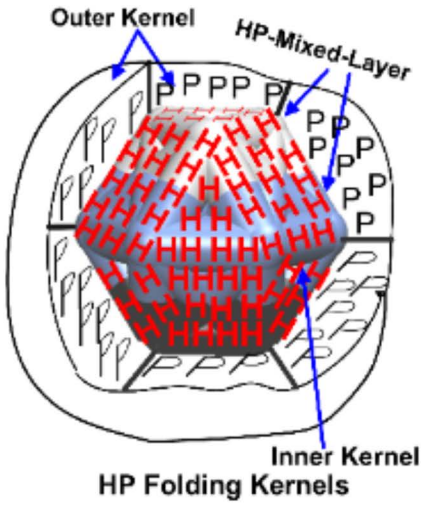
[32] Zhou H, Skolnick J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophys. J . 2011; 101:2043-2052.

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

Received: March 25, 2015 Revised: April 16, 2015 ~~Accepted: April 20, 2015~~



(a)


Outer Kernel
HP-Mixed-Layer
Inner Kernel
HP Folding Kernels

(b)

Figure 1: (a) Native like protein conformation (Hoque et al., 2007), 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 (Hoque et al., 2007), (Hoque et al., 2010).

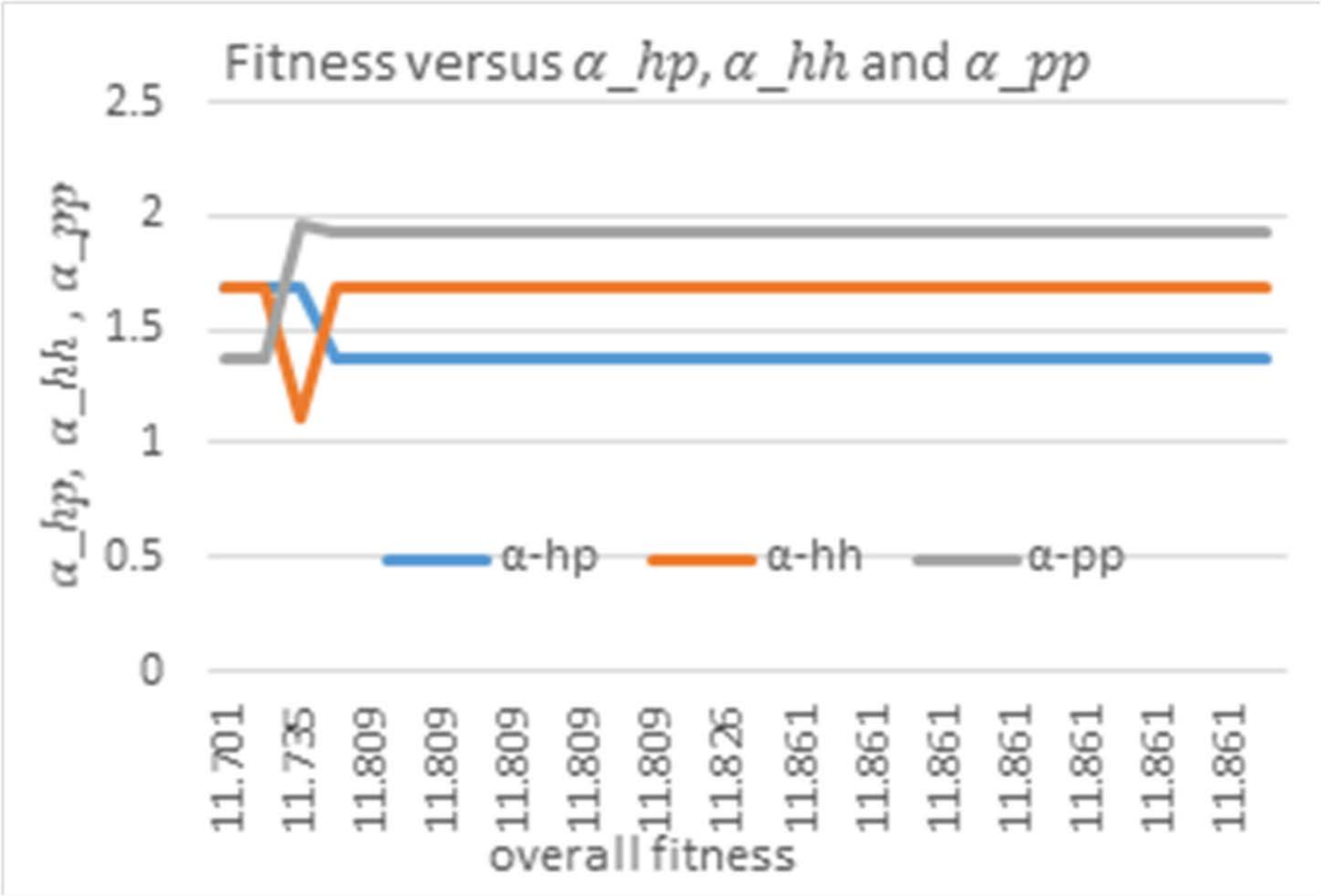

Fitness versus α_hp, α_hh and α_pp
2.5
2
1.5
1
0.5
0
α_hp, α_hh , α_pp
α-hp
α-hh
α-pp
11.701
11.735
11.809
11.809
11.809
11.809
11.809
11.826
11.861
11.861
11.861
11.861
11.861
11.861
11.861
overall fitness

Figure 2: Fitness versus α_{hp}, α_{hh} and α_{pp} values. The values remain stable during optimization.