sDFIRE: Sequence-Specific Statistical Energy Function for Protein Structure Prediction by Decoy Selections

Md Tamjidul Hoque,*[a] Yuedong Yang,[b] Avdesh Mishra,[a] and Yaoqi Zhou*[b]

An important unsolved problem in molecular and structural biology is the protein folding and structure prediction problem. One major bottleneck for solving this is the lack of an accurate energy to discriminate near-native conformations against other possible conformations. Here we have developed sDFIRE energy function, which is an optimized linear combination of DFIRE (the Distance-scaled Finite Ideal gas Reference state based Energy), the orientation dependent (polar-polar and polar-nonpolar) statistical potentials, and the matching scores between predicted and model structural properties including predicted main-chain torsion angles and solvent accessible surface area. The weights for these scoring terms are optimized by three widely used decoy sets consisting of a total of 134 proteins. Independent tests on CASP8 and CASP9 decoy sets indicate that sDFIRE outperforms other state-of-the-art energy functions in selecting near native structures and in the Pearson's correlation coefficient between the energy score and structural accuracy of the model (measured by TM-score).

DOI: 10.1002/jcc.24298

Introduction

To significantly advance protein structure and protein function prediction, we need an accurate energy function to describe water-mediated interactions between amino-acid residues of proteins.[1–4] Although these interactions can be described by quantum mechanical equations,[5,6] their solution is computationally prohibitive because of large number of atoms associated with a protein in water. As a result, physical-based, empirical, or knowledge-based energy functions as well as their combinations have been proposed as described in several reviews.[1–3,7] This article focuses on the knowledge-based approach that derives an effective energy function from known protein structures because knowledge-based potentials have been more successful in practical applications of protein structure prediction, in particular.[2]

The key for an accurate energy function is the specificity. That is, it has to recognize unique protein structure from nearly infinite number of other possible conformations (decoy structures). One approach in recent years to improve the specificity of an energy function is to extract from known protein structures the orientation dependence and/or multi-body effect at residue level,[8–14] as well as at all-atom level.[15–19] For example, Kortemme et al.[18] 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[20] by treating polar atoms as a dipole (dDFIRE).[19] Lu et al.[15] defined side-chain orientation according to rigid blocks of atoms (OPUS-PSP). Zhang and Zhang[16] 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).[17]

Another approach to improve an energy function is to employ restraints from accurately predicted structural properties. For example, the prediction for secondary structure has obtained an accuracy of more than 80%.[21] The predicted secondary structure is frequently utilized to limit the conformational space and thus to increase the accuracy of predicted protein structure.[22–26] More recently, we found that predicted torsion angles in real values[27,28] are more effective restraints for sampling and *ab initio* structure prediction[29] because unlike three-state secondary structures, continuous torsion angles capture non-ideal helical and strand conformations and provide backbone information on coil residues. Using predicted secondary structures, solvent accessible surface area (ASA),[30] and torsion angles was found useful for improving template-based structure prediction by threading or fold recognition techniques.[31–35]

The purpose of this work is to examine the usefulness of employing torsion angles and ASA predicted by SPINE X[21] as sequence-specific energy terms for discriminating near-native structures from decoys. More specifically, we will introduce

[a] *M. T. Hoque, A. Mishra*
Department of Computer Science, University of New Orleans, New Orleans, Louisiana 70148
E-mail: thoque@uno.edu
[b] *Y. Yang, Y. Zhou*
Institute for Glycomics and School of Informatics and Communication Technology, Griffith University, Queensland 4222, Australia
E-mail: yaoqi.zhou@griffith.edu.au
Contract grant sponsor: Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF; Contract grant number: (2013-16)-RD-A-19; Contract grant sponsor: National Health and Medical Research Council of Australia and Australian Research Council's Linkage Infrastructure, Equipment and Facilities funding scheme; Contract grant numbers: 1059775; 1083450; Contract grant sponsor: National Natural Science Foundation of China; Contract grant number: 61271378

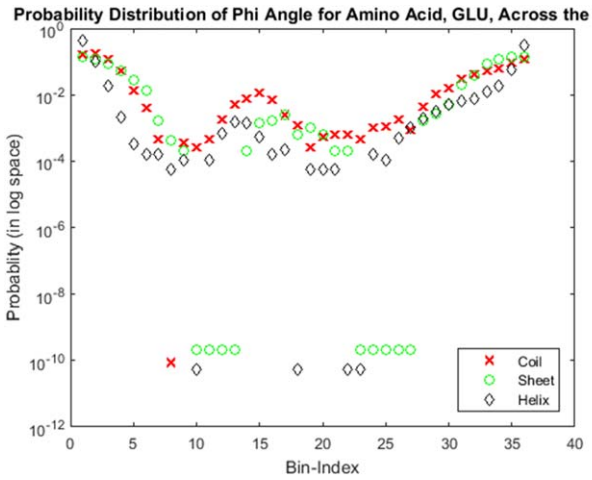

Figure 1. Probability distribution of phi angle for a sample amino-acid (GLU), across the bin for secondary structures: coil, sheet, and helix.

energetic terms based on estimated probability of match between predicted one-dimensional structural properties of a given sequence and actual structural properties of a given structure (decoy and near-native). Similar energetic terms have been found useful in improving template-based structure prediction in SPARKS X.[36] Here, we found that their combinations with the dDFIRE knowledge-based potential are useful for improving decoy discrimination in terms of the Pearson's correlation coefficient between the energy score and model structural accuracy (measured by TM-score) and selection of near-native structures.

Materials and Methods

Sequence-specific torsion energy

We define an energy function E_τ for describing the match between model torsion angles and predicted torsion angles from amino acids sequence as below:

$$E_\tau = -RT\sum_i \ln P(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i) \quad (1)$$

where R is the gas constant, T is the temperature, and $P(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i)$ is observed probability (scaled) of the angle error of prediction ($\Delta\tau_i = \tau_i - \tau_i^{\mathrm{Pred}}$, where τ_i^{Pred} is the predicted torsion angle by SPINEX[21]) for a given amino acid type, $\mathrm{AA_i}$ and predicted secondary structure, $\mathrm{SS_i}$; τ is ϕ or ψ angle from the model (decoy), and the summation is over all the torsion angles along the sequence. The probability function $P(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i)$ was obtained from a database of 2479 high-resolution (resolution <3 Å), non-redundant (<25% sequence identity) proteins that have 500 or less amino acid residues.[36] This database was employed in training and testing the SPINE X server.[21] A bin of ten degrees is employed for $\Delta\tau_i$, which resulted in a table of (20 × 3 × 36) or, 2160 entries coming from 20 different amino-acids, 3 different secondary structures (coil, sheet and helix) and 36 bins. The probability $P(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i)$ in eq. (1) has been computed as the ratio, $\frac{N_{\mathrm{obs}}(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i)}{\frac{1}{k}\sum_k N_{\mathrm{obs}}(\Delta\tau_i|\mathrm{AA}_i, \mathrm{SS}_i)}$, where $k = 36$ (number of bins) and N_{obs} is the observed frequency. Figure 1 shows the sample frequency distribution graphically and the complete tables are provided in the Supporting Information.

Sequence-specific energy for ASA

Similarly, the sequence-specific energy for ASA, E^{SA} is based on probability $P(\Delta\mathrm{SA}_i|\mathrm{AA}_i)$ of the prediction error of ASA ($\Delta\mathrm{SA}_i = \mathrm{SA}_i - \mathrm{SA}_i^{\mathrm{Pred}}$) for a given amino acid type, $\mathrm{AA_i}$ from the decoy or model structure. That is,

$$E^{\mathrm{SA}} = -RT\sum_i \ln P(\Delta\mathrm{SA}_i|\mathrm{AA}_i) \quad (2)$$

where the summation is over all residues along the sequence. The probability function $P(\Delta\mathrm{SA}_i|\mathrm{AA}_i)$ was also obtained from the database of 2479 proteins.[36]

The dipolar DFIRE statistical potential

The dipolar DFIRE statistical potential,[19] E^{dDFIRE}, is composed of the terms described below.

$$E^{\mathrm{dDFIRE}} = E^{\mathrm{DFIRE}} + E^{\mathrm{P_1}} + E^{\mathrm{P_2}} + E^{\mathrm{PN}} \quad (3)$$

where, $E^{\mathrm{DFIRE}} = \sum_{ij} u^{\mathrm{DFIRE}}(r_{ij})$ with the summation over heavy atoms i and j, $E^{\mathrm{P_1}} = \sum_{\mathrm{pq}}\left[u(\theta_{\mathrm{p}}|r_{\mathrm{pq}}) + u(\theta_{\mathrm{q}}|r_{\mathrm{pq}})\right]$ and $E^{\mathrm{P_2}} = \sum_{\mathrm{pq}} u(\theta_{\mathrm{pq}}|r_{\mathrm{pq}})$ with the summation over polar atoms p and q only, and $E^{\mathrm{PN}} = \sum_{\mathrm{pn}} u(\theta_{\mathrm{p}}|r_{\mathrm{pn}})$ with the summation over polar and nonpolar atoms p and n, respectively. The energy applies only to heavy atoms belonging to residues that are not sequence neighbors ($|i-j| > 3$). θ_{p}, θ_{q}, and θ_{pq} are the angles between the reference direction of polar atom p, $\vec{r}_{\mathrm{p}}^{\,\mathrm{Ref}}$ and the distance vector $\vec{r}_{\mathrm{pq}}$, between $\vec{r}_{\mathrm{p}}^{\,\mathrm{Ref}}$ and $\vec{r}_{\mathrm{pq}}$ and between $\vec{r}_{\mathrm{p}}^{\,\mathrm{Ref}}$ and $\vec{r}_{\mathrm{q}}^{\,\mathrm{Ref}}$, respectively (Fig. 2). Reference vectors of polar atoms are defined according to bond vectors of heavy atoms.[19] In addition,

$$u^{\mathrm{DFIRE}}(r_{ij}) = \begin{cases} -RT\ln\dfrac{N^{\mathrm{obs}}(i,j,r)}{\left(\dfrac{r}{r_{\mathrm{cut}}}\right)^{\alpha}\left(\dfrac{\Delta r}{\Delta r_{\mathrm{cut}}}\right)N^{\mathrm{obs}}(i,j,r_{\mathrm{cut}})}, & r < r_{\mathrm{cut}} \\ & \text{and} \\ 0, & r \geq r_{\mathrm{cut}} \end{cases}$$

$u(\theta|r) = -RT\ln\frac{P^{\mathrm{obs}}(\theta|r)}{P^{\mathrm{obs}}(\theta|r_{\mathrm{cut}})}$, where $N^{\mathrm{obs}}(i,j,r)$ is the observed frequency of the atomic pair (i,j) with in a distance shell $\left(r-\frac{\Delta r}{2}\right)$ to $\left(r+\frac{\Delta r}{2}\right)$, $P_{i,j}^{\mathrm{obs}}(\theta|r) = N^{\mathrm{obs}}(i,j,\theta|r)/N^{\mathrm{obs}}(i,j,r)$ with θ as either θ_{p}, θ_{q}, or θ_{pq}. The Δr is the width of the bin, which is set to 0.5 Å uniformly for each bin with the cutoff $r_{\mathrm{cut}} = 15.0$ Å and $\alpha = 1.61$. This energy function was generated by employing a database of 3574 structures having resolution less than 2.0 Å and the corresponding sequences are less than 30% homologous, obtained from the work by Hobohm et al.[37]

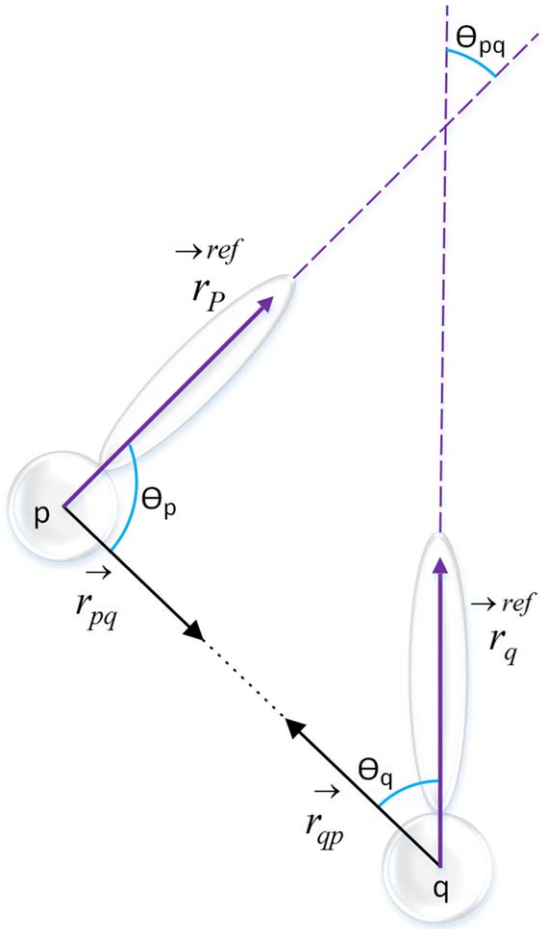

Figure 2. Graphical representation of orientation angles θ_p, θ_q, and θ_{pq}. The vectors $r_p^{\rightarrow ref}$ and $r_q^{\rightarrow ref}$ are the reference directions for polar atoms p and q, respectively.

The sDFIRE energy function

sDFIRE is an optimized combination of the above-mentioned sequence-specific and dDFIRE energy terms. That is,

$$E^{sDFIRE} = E^{DFIRE} + w_1 E^{P_1} + w_2 E^{P_2} + w_3 E^{PN} + w_4 E^{\varphi} + w_5 E^{\phi} + w_6 E^{SA} \quad (4)$$

where w_i (i=1, . . ., 6) are to-be-optimized weights.

Decoy datasets

The weights of sDFIRE were trained with three decoy datasets, MOULDER, ROSETTA, and I-TASSER datasets. The trained sDFIRE energy function was tested on the CASP 8 and the CASP 9 datasets. These training and test datasets are discussed in brief as follows.

MOULDER decoy set comprises of 20 proteins.[38] Each of these proteins contains ~300 comparative models built using homologous template. The models were built using alignment that has at least five different alignment positions or in other words they do not share more than 95% of identically aligned positions. Comparative protein structure modeling program called MODELLER-6 is used to build these decoys. MODELLER-6 program applies default model building routine with fastest refinement which keeps most of the template structure unchanged. The decoys developed by MODELLER-6 are different from decoys that are generated by *ab initio* folding in which all structure regions are reassembled from scratch. MOULDER decoy sets of 20 proteins were obtained from http://salilab.org/decoys/.

ROSETTA decoy set is a collection of 58 proteins generated by Baker-lab. Each of these proteins contains 20 random models and 100 lowest scoring models from 10,000 decoys generated using ROSETTA *de novo* structure prediction followed by all-atom refinement.[39] Improvement of current Rosetta decoys over original decoys is based on addition of side chains to the backbone models and removal of steric clashes. In generating optimal decoy sets following four important points were considered[40]: (1) decoy set should contain conformations for a wide variety of different proteins to overcome the over-fitting problem, (2) native and decoy structures conformation should be less than 4 Å root-mean-squared distance, (3) decoys conformation should be at least close to local minima of energy function, and (4) native structure information should not be used to construct decoys. ROSETTA decoy sets of 58 proteins are from http://depts.washington.edu/bakerpg/decoys/.

I-TASSER decoy set-II is a collection of 56 proteins.[41] Each of these proteins contains 300 to 500 decoys. Decoys are generated using template-based modeling and atomic-level structure refinements. These decoys are first generated by Monte Carlo simulations and then refined using GROMACS4.0 molecular dynamics simulation. GROMACS4.0 simulation is used to remove steric clashes and improve hydrogen-bonding network. We obtained I-TASSER decoy sets of 56 proteins from http://zhanglab.ccmb.med.umuch.edu/.

CASP8 and CASP9 decoy sets are used for testing. The 125 proteins of CASP8 decoy sets and 112 proteins of CASP9 decoy-set were downloaded from the CASP8 website http://predictioncenter.org/casp8/ and CASP9 website http://predictioncenter.org/casp9/ respectively. Most of the structures in decoy-set are generated by homology modeling by all the CASP8 and CASP9 servers. A few decoys were discarded as GOAP failed to process them. Finally, on an average, our test sets consisted of 278 and 293 decoy-sets per protein of CASP8 and CASP9, respectively. The full list is provided in the Supporting Information.

Searching optimal parameters

We applied search algorithms to obtain the optimal parameters or weights described in eq. (4) with I-TASSER, ROSETTA, and MOULDER datasets (native structures are excluded). We used two different ways to cross-validate narrower regions of the optimal parameters with two separate objectives: to minimize the negative average Pearson's correlation coefficient (PCC) of energy score with respect to the TM-score (Template Modeling score)[42] and to maximize the average TM-score of the top-1 model ranked by the energy. TM-score is commonly used to measure similarity between protein structures having either same or, even different sequence length. TM-score ranges from (0, 1], where score 1 means perfect match and 0

Table 1. The best outcome of coarse grid search with the step size 0.05.

Decoy sets (count)	sDFIRE	w_1	w_2	w_3	w_4	w_5	w_6
I-TASSER (56)	−0.526 (0.575)						
ROSETTA (58)	−0.472 (0.546)	0.20	1.95	0.05	0.05	0.20	0.10
MOULDER (20)	−0.915 (0.779)						

Legend for the 2nd column: the average correlation coefficient (the TM-score of the best model).

Table 2. The best outcome of fine grid search with the step size 0.01.

Decoy sets (count)	sDFIRE	w_1	w_2	w_3	w_4	w_5	w_6
I-TASSER (56)	−0.535 (0.581)						
ROSETTA (58)	−0.472 (0.548)	0.31	1.77	0.03	0.03	0.16	0.08
MOULDER (20)	−0.913 (0.779)						

Legend for the 2nd column: the average correlation coefficient (the TM-score of the best model).

indicates no match at all. Practically, score below 0.17 indicates no similarity whereas with a score higher than 0.5 are assumed to have the same structural fold.

We applied *grid-search (GS)* as well as *genetic algorithm (GA)* described as follows.

Grid-Search. We implemented GS to find the best possible values of the six variables ($w_1, w_2, w_3, w_4, w_5, w_6$) of eq. (4). The search landscape was too convoluted and time intensive having six dimensions. It required us to proceed with search in two steps: first to use coarse steps and then to go for fine steps. For the ranges of each variable from 0.01 to 2.0 with step size 0.01 (fine steps) GS took too long, thus, we went for the ranges of each variable from 0.05 to 2.0 with coarse increment factor of 0.05. Then, for the identified potential areas we went for fine grid steps of 0.01 to identify the best parameter-values. The best result found from coarse GS with step size 0.05 is shown in Table 1.

Based on the obtained values, for finer search with step-size 0.01, we looked for better values of w_1 within range 0.01–0.4, w_2 within range 1.75–2.0, w_3 within range 0.03–0.10, w_4 within range 0.03–0.10, w_5 within range 0.01–0.4, and w_6 within range 0.01–0.3. The obtained results are placed in Table 2.

Genetic Algorithms. Six weight parameters of eq. (4) were also optimized by a genetic algorithm search.[43] The GA parameters were of population-size 200, elite-rate 5%, crossover-rate 90%, and mutation-rate 50%. The parameter search stops when there is no improvement.

Because multiple solutions exist with consistently similar performance from either grid search or genetic algorithms, we have chosen the weights from the GA because it yields slightly better performance in our training set (I-TASSER, ROSETTA, and MOULDER). The final weights are 0.182, 1.351, 0.754, 0.18, 0.018, and 0.45, and the relative importance of the corresponding weights are 0.01037, 0.04262, 0.08504, 0.00023, 0.00006, and 0.00028, respectively. Here, the relative importance of the weights is obtained by dividing actual weight by the maximum value found from respective energy term.

Results

Table 3 compares the results from sDFIRE to the other state-of-art energy functions such as DFIRE, dDFIRE, RWPlus, and GOAP. In general, sDFIRE shows a better correlation coefficient (6% improvement over the next best) and a higher average TM-score for the model selected (4% improvement over the next best) for all three datasets (I-TASSER, ROSETTA, and MOULDER). This comparison is for reference only because these three datasets were employed for training six weighting parameters in sDFIRE.

For a more reliable comparison, Table 4 compares the performance of various methods on independent CASP 8 and CASP 9 datasets according to the correlation coefficients between energy score and model structure accuracy (TM-score). sDFIRE makes 16.8% to 21.4% improvements for the CASP8 dataset and 7.4% to 24% improvements for the CASP9 dataset.

Table 3. Performance of several methods in I-TASSER, ROSETTA, and MOULDER datasets.

Decoy sets (no. of proteins)	DFIRE	dDFIRE	RWPlus	GOAP	sDFIRE
I-TASSER (56)	−0.491 (0.575)	−0.525 (0.578)	−0.488 (0.577)	−0.477 (0.567)	−0.530 (0.581)
ROSETTA (58)	−0.438 (0.509)	−0.393 (0.480)	−0.444 (0.505)	−0.476 (0.511)	−0.476 (0.5519)
MOULDER (20)	−0.837 (0.749)	−0.881 (0.748)	−0.840 (0.745)	−0.886 (0.771)	−0.913 (0.780)
Weighted Average	−0.519 (0.605)	−0.521 (0.594)	−0.521 (0.604)	−0.537 (0.607)	−0.563 (0.633)

For the 2nd to last columns, each cell legend: the average correlation coefficient (the TM-score of the best model).

Table 4. Performance of several methods on CASP datasets based on correlation coefficients.

Decoy sets (no. of proteins)	DFIRE	dDFIRE	RWPlus[a]	GOAP	sDFIRE
CASP8 (125)	−0.5544 (16.83%)	−0.5337 (21.36%)	−0.5398 (19.98%)	−0.5526 (17.20%)	**−0.6477**
CASP9 (112)	−0.5099 (23.53%)	−0.5252 (18.60%)	−0.5165 (20.60%)	−0.580 (7.39%)	**−0.6229**

For the 2nd to last columns, each cell entry legend: correlation coefficients (percentage of improvement by sDFIRE). Best values are highlighted. [a] RWPlus was optimized on CASP8 datasets.[16]

Table 5. Performance of several methods on CASP datasets based on the structural accuracy (the average TM-score of the first ranked models).

Decoy sets (no. of proteins)	DFIRE	dDFIRE	RWPlus[a]	GOAP	sDFIRE
CASP8 (125)	0.6458 (6.28%)	0.63836 (7.52%)	0.63832 (7.53%)	0.6777 (1.28%)	**0.6864**
CASP9 (112)	0.5991 (7.34%)	0.5857 (8.92%)	0.6047 (6.35%)	0.621 (3.55%)	**0.6431**

For the 2nd to last columns, each cell entry legend: TM-score (percentage of improvement by sDFIRE). Best values are highlighted. [a] RWPlus was optimized on CASP8 datasets.[16]

Table 5 compares the ability of various methods to locate the best models from decoys based on the average TM-score. A small but consistent improvement (1–9%) is observed by sDFIRE over other methods. It should be noted that GOAP failed to produce results for many decoys. Tables 4 and 5 represent a subset of decoys for which all methods can produce energy scores.

We further examine the validity and robustness of our method by comparing TM-scores of predicted models for CASP 8 (Fig. 3 (Top)) and CASP 9 (Fig. 3 (Bottom)) proteins against those from other methods. It is clear from the figures that sDFIRE offers a substantially improvement over GOAP, DFIRE2, or RWPLUS as the majority of the points is below the $x = y$ line (i.e., the model predicted by sDFIRE has a higher TM-score than that predicted by the other method in comparison). The difference is large for a significant fraction of CASP targets.

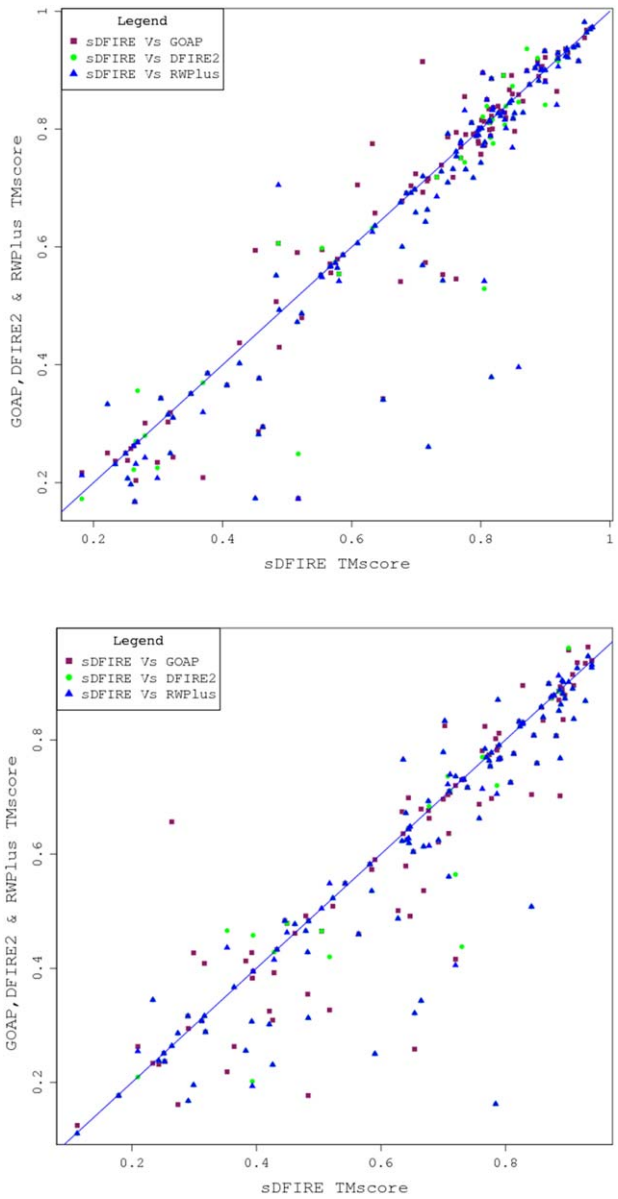

Figure 3. TMscore for CASP targets (CASP 8, top; CASP 9, bottom) given by sDFIRE (x axis) compared to that by other methods (GOAP in maroon, DFIRE2 in lime and RWPlus in blue). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Conclusions

We have developed a new energy function, sDFIRE for protein structure prediction. Unlike most knowledge-based energy functions, sDFIRE incorporates predicted sequence-based structural properties. Basically, we incorporated evolution-driven properties through predicted secondary structure and predicted solvent accessible surface areas. We have also accounted for the orientation-dependent energy components of polar atoms. The performance of sDFIRE was compared to other state-of-the-art energy functions based on CASP8 and CASP9 data-sets, which indicated that 18.18% overall improvements based on average Pearson's correlation coefficient between the predicted energy score and the TM-score of decoy structures.

Keywords: energy function · protein structure · decoy sets · genetic algorithm · optimization

How to cite this article: M. Tamjidul Hoque, Y. Yang, A. Mishra, Y. Zhou. *J. Comput. Chem.* **2016**, *37*, 1119–1124. DOI: 10.1002/jcc.24298

Additional Supporting Information may be found in the online version of this article.

[1] T. Lazaridis, M. Karplus, *Curr. Opin. Struct. Biol.* **2000**, *10*, 139.
[2] J. Skolnick, *Curr. Opin. Struct. Biol.* **2006**, *16*, 166.
[3] Y. Zhou, Y. Duan, Y. Yang, E. Faraggi, H. Lei, *Theor. Chem. Acc.* **2011**, *128*, 3.
[4] S. J. Fleishman, T. A. Whitehead, E. M. Strauch, J. E. Corn1, S. Qin, H. X. Zhou, J. C. Mitchell, O. N. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. A. Bates, M. Zacharias, H. Park, J. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Azé, S. Soner, ŞK. Oval, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Pérez-Cano, C. Pons, J. Fernández-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang, L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G.

Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. J. J. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodríguez, D. Kihara, P. B. Stranges, R. Jacak, B. Kuhlman, S. Y. Huang, X. Zou, S. J. Wodak, J. Janin, D. Baker, *J. Mol. Biol.* **2011**, *414*, 289.

[5] N. Yu, H. P. Yennawar, K. M. Merz, Jr., *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61*, 322.

[6] H Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, W. Yang, *Proteins* **2001**, *44*, 484.

[7] Y. Zhou, H. Zhou, C. Zhang, S. Liu, *Cell Biochem. Biophys.* **2006**, *46*, 165.

[8] Y. Wu, M. Lu, M. Chen, J. Li, J. Ma, *Protein Sci.* **2007**, *16*, 1449.

[9] S. Miyazawa, R. L. Jernigan, *J. Chem. Phys.* **2005**, *122*, 024901.

[10] C. Hoppe, D. Schomburg, *Protein Sci.* **2005**, *4*, 2682.

[11] D. Gilis, C. Biot, E. Buisine, Y. Dehouck, M. Rooman, *J. Chem. Inf. Model* **2006**, *46*, 884.

[12] N. Buchete, J. Straub, D. Thirumalai, *Curr. Opin. Struct. Biol.* **2004**, *14*, 225.

[13] S. E. Moughon, R. Samudrala, *BMC Bioinformatics* **2011**, *12*, 368.

[14] M. T. Zimmermann, S. P. Leelananda, A. Kloczkowski, R. L. Jernigan, *J. Phys. Chem. B* **2012**, *116*, 6725.

[15] M. Lu, A. D. Dousis, J. Ma, *J. Mol. Biol.* **2008**, *376*, 288.

[16] J. Zhang, Y. Zhang, *PLoS One* **2010**, *5*, e15386.

[17] H. Zhou, J. Skolnick, *Biophys. J.* **2011**, *101*, 2043.

[18] T. Kortemmea, A. V. Morozova, D. Baker, *J. Mol. Biol.* **2003**, *326*, 1239.

[19] Y. Yang, Y. Zhou, *Proteins* **2008**, *72*, 793.

[20] H. Zhou, Y. Zhou, *Protein Sci.* **2002**, *11*, 2714.

[21] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, Y. Zhou, *J. Comput. Chem.* **2011**, *33*, 259.

[22] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209.

[23] A. R. Ortiz, A. Kolinski, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 1020.

[24] B. Fain, M. Levitt, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 10700.

[25] M. Nanias, M. Chinchio, J. Pillardy, D. R. Ripoll, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 1706.

[26] H. Li, Y. Zhou, *J. Bioinformatics Comput. Biol.* **2005**, *3*, 1151.

[27] B. Xue, O. Dor, E. Faraggi, Y. Zhou, *Proteins* **2008**, *72*, 427.

[28] E. Faraggi, B. Xue, Y. Zhou, *Proteins* **2009**, *74*, 847.

[29] E. Faraggi, Y. Yang, S. Zhang, Y. Zhou, *Structure* **2009**, *17*, 1515.

[30] S. Iqbal, A. Mishra, M. T. Hoque, *J. Theor. Biol.* **2015**, *380*, 380.

[31] Hargbo, J.; Elofsson, A. *Proteins* **1999**, *36*, 68.

[32] Zhou, H.; Zhou, Y. *Proteins* **2004**, *55*, 1005.

[33] H. Zhou, Y. Zhou, *Proteins* **2005**, *58*, 321.

[34] J. S. Ding, A. Biegert, A. N. Lupas, *Nucleic Acids Res.* **2005**, *33*, W244.

[35] J. Peng, J. Xu, *Bioinformatics* **2010**, *26*, i294.

[36] Y. Yang, E. Faraggi, H. Zhao, Y. Zhou, *Bioinformatics* **2011**, *27*, 2076.

[37] U. Hobohm, M. Scharf, R. Schneider, C. Sander, *Protein Sci.* **1992**, *1*, 409.

[38] A. Sali, *Moulder Decoy Sets* **2014**. Available at: http://salilab.org/decoys.

[39] J. Zhang, Y. Zhang, *PLoS One* **2010**, *5*, 1.

[40] J. Tsai, R. Bonneau, A.V. Morozov, B. Kuhlman, C.A. Rohl, D. Baker, *Proteins* **2003**, *53*, 76.

[41] Y. Zhang, *Protein Structure Decoys* **2014**. Available at: http://zhanglab.ccmb.med.umich.edu/decoys/.

[42] Y. Zhang, J. Skolnick, *Proteins* **2004**, *57*, 702.

[43] M.T. Hoque, M. Chetty, A. Lewis, A. Sattar, *IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB)* **2011**, *8*, 234.

Received: 17 October 2015
Revised: 6 December 2015
Accepted: 13 December 2015
Published online on 5 February 2016