Journal of Molecular Biology, **427**(15): 2562-2576, 2015.

ICOSA: A Distance-dependent, Orientation-specific Coarse-grain Contact Potential for Protein Structure Modeling

Wessam Elhefnawy*, Lin Chen*, Yun Han, and Yaohang Li†

Department of Computer Science

Old Dominion University

Norfolk, VA 23529, USA

* These two authors contribute equally to this work.

† Correspondent Author: yaohang@cs.odu.edu

Journal of Molecular Biology, **427**(15): 2562-2576, 2015.

Abstract

The relative distance and orientation in contacting residue pairs plays a significant role in protein folding and stabilization. We hereby devise a new knowledge-based, coarse-grained contact potential, so-called ICOSA, by correlating inter-residue contact distance and orientation in evaluating pair-wise inter-residue interactions. The rationale of our approach is to establish icosahedral local coordinates to estimate the statistical residue contact distributions in all spherical triangular shells within a sphere. We extend the theory of finite ideal gas reference state to icosahedral local coordinates. ICOSA incorporates long range contact interactions, which is critical to ICOSA sensitivity and is justified in statistical rigor. With only backbone atoms information, ICOSA is at least comparable to all-atom, fine-grained potentials such as Rosetta, DFIRE, I-TASSER, and OPUS in discriminating near-natives from misfold protein conformations in the Rosetta and I-Tasser protein decoy sets. ICOSA also outperforms a set of widely used coarse-grained potentials and is comparable to all-atom, fine-grained potentials in identifying CASP10 models.

Introduction

Amino acids are the basic structural building blocks of proteins. These amino acids exhibit very different physic-chemical properties, which form complicated contact preferences among each other. The inter-residue contacts control the stability of protein structure as well as their biological functions.

The knowledge-based contact potentials (energies) are the most popularly used representation of inter-residue contacts. Although the scope and limitation of the contact potentials are still vigorously debated and disputed [1-4], contact potentials have been successfully used in a variety of applications, including fold recognition [5], protein structure prediction [6], protein design [7], and protein-protein docking [8]. The basic idea of contact potentials is to derive statistical contact preferences from known protein structures presented in Protein Data Banks (PDB) and then estimate the corresponding pseudo-potentials with respect to the defined reference states. Since 1990s, many contact potentials have been proposed and developed. The key variation lies on the definition of a contact. The simple contact definition based on a distance cutoff is popularly used [9-11]. In coarse-grain residue-level modeling, the residue-residue distance is measured by CA-CA, CB-CB, or estimated geometric side chain centers and the cutoff distance usually ranges from around 6A to 8A. Another type of knowledge-based potentials models contacts that are distance-dependent [12, 13], where contact frequencies are observed in particular distance intervals within longer distance cutoff up to 20A. When side chain information is available in atomic detail, the contact definition with distance cutoff can be extended to atomic level [14, 15]. An alternative definition of contacts is based on geometric models such as Voronoi diagram [16, 17], Laguerre tessellation [18], Delaunay triangulation [19, 20], or alpha shape [21, 22]. Compared to simple distance cutoff, geometric models contain richer topological information to more precisely define and parameterize inter-residue contacts [23], such as interaction areas, side chain orientation, and solvent accessibility surface. In spite of the variations in contact definitions and parameterizations, these contact potentials have achieved different levels of success in identifying the native structures from the erroneous models.

Although the fine-grained potentials at atomic level are usually more precise, the coarse-grained contact potentials at residue level attempt to strike a balance between the accuracy of energy representation and computational efficiency. An ideal coarse-grained contact potential is desired to have sensitivity and accuracy comparable to the all-atom potentials while maintaining reduced protein structure representation for computational expedition. One way to enhance the accuracy of coarse-grained potentials is to account for the environment information of the contacting residues. For example, studies have showed that the distribution of the contact distances varies in their amino acid sequence [24, 25], secondary structures [26], and protein sizes and classes [25]. Distance-dependent potentials accounting for contact residues by separation in sequence [27], secondary structures [26], sequence profile context [28], or hydrophobicity [29] have exhibited certain accuracy gain. Alternatively, some multibody potentials [22, 30, 31] modeling high-order inter-residue interactions have been developed and demonstrated advantages over pairwise residue contact potentials. In addition to the contact distance, study of angular distribution of side chains in amino acids by Bahar and Jernigan [32] has shown that the relative orientations in residue pairs play an important role in determining side chain packing in proteins. Incorporating estimated side chain orientation shows another direction to improve the accuracy of coarse-grained potentials. Recent efforts [33-36] start to encode the pair-wise residue orientation information into the protein interaction potentials, usually as additional orientation terms. Moughon and Samudrala [37] divided the contact space into cubic bins in local coordinates defined by N, CA, and C atoms to

evaluate contact favorability at specific position. Feng et al. [30] developed a four-body potential by considering the orientation of a contacting residue to the tetrahedron formed by four sequential residues, which has showed improved accuracy than the four-body potentials using a Delaunay tessellation algorithm.

The majority of the knowledge-based contact potentials relies on the Bayesian formulation or inverse Boltzmann law to convert the observed contact probability into interaction potentials [38]. Consequently, a critical aspect differentiates the effectiveness of various knowledge-based contact potentials is their underlying reference states. Theoretically, the reference state should result from hypothetical systems lack of specific inter-residue interactions. Nevertheless, there is no universal way of constructing the reference states. The early contact potentials [39] adopt a quasi-chemical approximation reference state, assuming that the expected probability of a specific residue pair contact is proportional to the mole fraction of these two residues. Samudrala and Moult [40] used a conditional probability method to derive an average over all amino acid types (coarse-grained) or atom types (fine-grained) to represent the random reference state. Zhang et al. [15] developed random crystal reference states to remove compositional bias. Skolnick et al. [41] proposed reference states accounting for the constraints of chain connectivity and compactness. Rykunov and Fiser [36] used an atom-shuffled reference state in their orientation-dependent potential. Zhang and Zhang [42] used an ideal random walk chain as the reference state for their side-chain orientation dependent potential. There is no clear winner in recent performance assessment in knowledge-based potentials using various reference states [43]. Nevertheless, the generality and specificity of a knowledge-based potential are contradicting and the optimality of a reference state depends rather on the specific applications [43].

In this paper, we present a novel knowledge-based, coarse-grained contact potential, so-called ICOSA, to correlate contact distance and orientation in pair-wise residues interactions. This is based on our analysis indicating that residues relative orientation is strongly correlated with contact distance. Our approach is to set up icosahedral local coordinates by dividing the sphere surface into uniform icosahedrons to estimate the pair-wise residue contact distributions in spherical triangular shells within a sphere. The finite ideal-gas reference state [12, 44] is a theoretical reference state assuming uniformly distributed non-interacting points in finite spheres, which is independent of structures solved by experiments or decoys generated by computational methods. As a result, the finite ideal-gas reference state covers a broad conformation space, which is particularly suitable for general protein structure modeling applications. The uniform icosahedron division of sphere allows us to adopt the finite ideal gas reference state in our contact potential aiming for good generalization. Moreover, the long range interactions included in ICOSA lead to better correlation between the potential and the qualities of the models and thus yield improved accuracy and sensitivity. The cutoff distance in ICOSA is also justified in statistical rigor. The effectiveness of our context-dependent contact potential is demonstrated on the Rosetta [45] and I-Tasser [42] protein decoy sets as well as CASP10 [46] models.

Materials and Methods

Data Sets

We use the protein chain dataset **Cull9791** generated by the PISCES server [47] on 2/15/2015 to collect inter-residue contact samples to generate statistics. **Cull9791** contains 9,791 chains with

at most 25% sequence identity, 3.0A resolution cutoff, and less than 1.0 R-factor. Rosetta [45] and I-Tasser [42] protein decoy sets and CASP10 [46] models are used to benchmark our contact potential. To ensure the correctness of our computational experiments, we exclude all sequences in **Cull9791** with greater than 25% sequence identity to any protein targets in Rosetta or I-Tasser decoy sets or CASP10 targets when the residue contact samples are extracted to generate contact statistics for ICOSA.

Icosahedral Local Coordinates

A regular icosahedron is a polyhedron with 20 identical equilateral triangular faces, 30 edges, and 12 vertices. One of the important geometrical significances of a regular icosahedron is its capability of completely subdividing the surface of a sphere into 20 identically equal spherical polygons, which is the most among the only five Platonic solids. The icosahedron has many attractive characteristics. The regularity in each equilateral spherical triangle allows computational efficiency in generating an icosahedral grid. Moreover, recent studies [33], [48] on residue packing show that icosahedron models fit well for side chain packing geometries and orientational distribution of contact clusters.

In ICOSA, we set up icosahedral local coordinates to correlate contact distance and orientation in pair-wise residues. We use the CA-CA contact as an example to illustrate the icosahedral local coordinates. The contact atom CA is placed in the origin. The CA-N bond extends the positive X axis, the atoms forming the C-CA-N triangle determine the X-Y plane, and rotation axis generated by the right-hand rule of C, CA, N atoms forms the positive Z axis. Then, the unit vectors of the 12 vertices of the regular icosahedron are

$$
\begin{aligned}
v1 &= [\,0 \quad 0 \quad -1\,] \\
v2 &= [\,0.7236 \quad 0.5257 \quad -0.4472\,] \\
v3 &= [\,-0.2764 \quad 0.8507 \quad -0.4472\,] \\
v4 &= [\,-0.8944 \quad 0 \quad -0.4472\,] \\
v5 &= [\,-0.2764 \quad -0.8507 \quad -0.4472\,] \\
v6 &= [\,0.7236 \quad -0.5257 \quad -0.4472\,] \\
v7 &= [\,0.8944 \quad 0 \quad 0.4472\,] \\
v8 &= [\,0.2764 \quad 0.8507 \quad 0.4472\,] \\
v9 &= [\,-0.7236 \quad 0.5257 \quad 0.4472\,] \\
v10 &= [\,-0.7236 \quad -0.5257 \quad 0.4472\,] \\
v11 &= [\,0.2764 \quad -0.8507 \quad 0.4472\,] \\
v12 &= [\,0 \quad 0 \quad 1\,].
\end{aligned}
$$

Correspondingly, the 20 triangular faces of an icosahedron are

$$
\begin{aligned}
t1 &= [\,v1 \quad v3 \quad v2\,] \\
t2 &= [\,v1 \quad v4 \quad v3\,] \\
t3 &= [\,v1 \quad v5 \quad v4\,] \\
t4 &= [\,v1 \quad v6 \quad v5\,] \\
t5 &= [\,v1 \quad v2 \quad v6\,] \\
t6 &= [\,v2 \quad v3 \quad v8\,] \\
t7 &= [\,v3 \quad v4 \quad v9\,] \\
t8 &= [\,v4 \quad v5 \quad v10\,] \\
t9 &= [\,v5 \quad v6 \quad v11\,]
\end{aligned}
$$

$$t10 = [\ v6 \quad v2 \quad v7\ ]$$
$$t11 = [\ v2 \quad v8 \quad v7\ ]$$
$$t12 = [\ v3 \quad v9 \quad v8\ ]$$
$$t13 = [\ v4 \quad v10 \quad v9\ ]$$
$$t14 = [\ v5 \quad v11 \quad v10\ ]$$
$$t15 = [\ v6 \quad v7 \quad v11\ ]$$
$$t16 = [\ v7 \quad v8 \quad v12\ ]$$
$$t17 = [\ v8 \quad v9 \quad v12\ ]$$
$$t18 = [\ v9 \quad v10 \quad v12\ ]$$
$$t19 = [\ v10 \quad v11 \quad v12\ ]$$
$$t20 = [\ v11 \quad v7 \quad v12\ ].$$

Figure 1 shows the icosahedral local coordinates. As a result, in the icosahedral local coordinates, each CA-CA contact can be represented as (r, t), where r is the contact distance and t is the triangle face number in the icosahedron indicating the contact orientation. The implementation of ICOSA is based on CA-CA contacts; however, the icosahedral local coordinates can also be applied to CB-CB or centroid-centroid contacts in a similar setup.

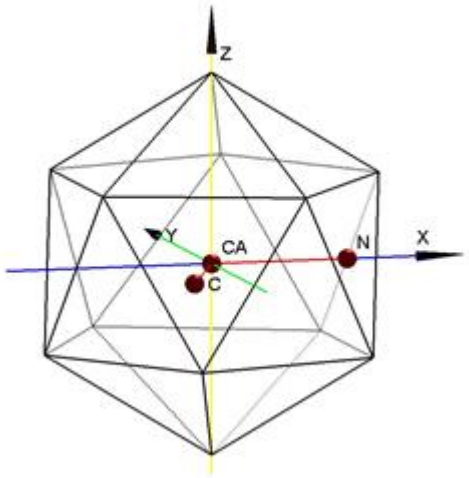


Figure 1. Icosahedral local coordinates with CA at the origin

Residue Contact Orientational Distribution in Spherical Triangular Shells in Icosahedral Local Coordinates

In the icosahedral local coordinates, a spherical shell $S(r, d)$ with inner radius of r and thickness of d is divided into 20 identically equal spherical triangular shells. Figure 2 illustrates a spherical triangular shell $\Delta(r, d, t)$ in a sphere, where r is the inner radius, d is thickness, and t is the icosahedron triangle face number. The uniform division of a spherical shell $S(r, d)$ allows us to conveniently study the residue contact orientational distribution in the 20 spherical triangular shells.

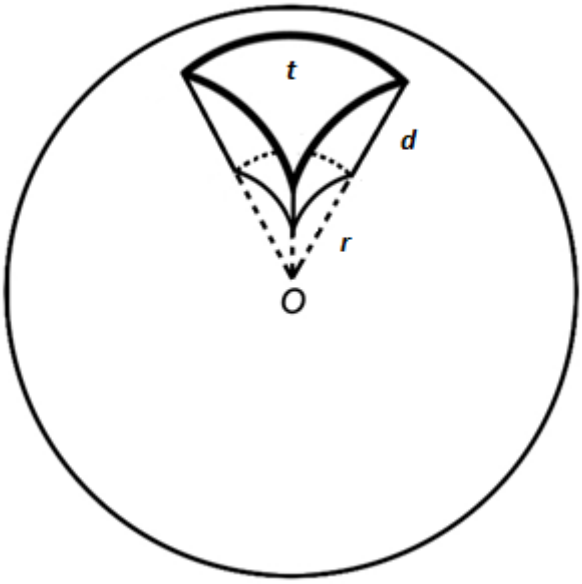
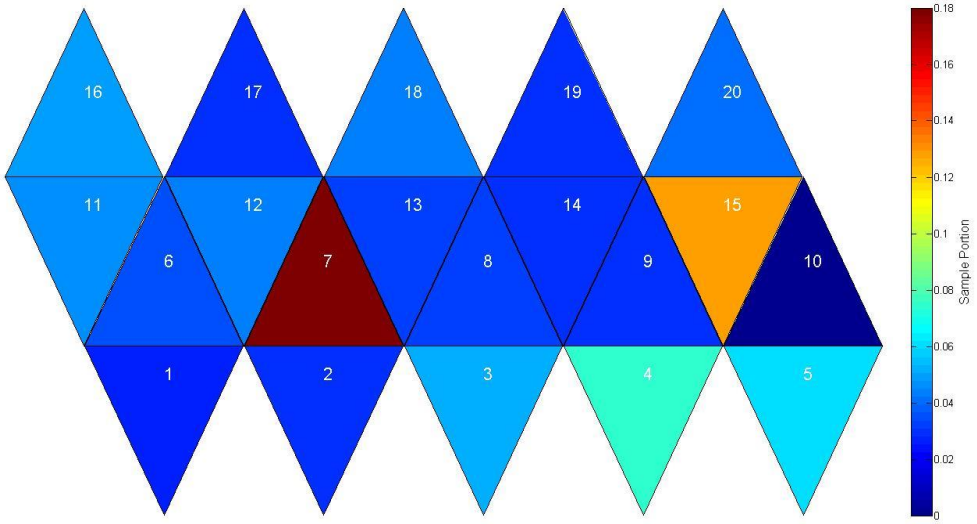
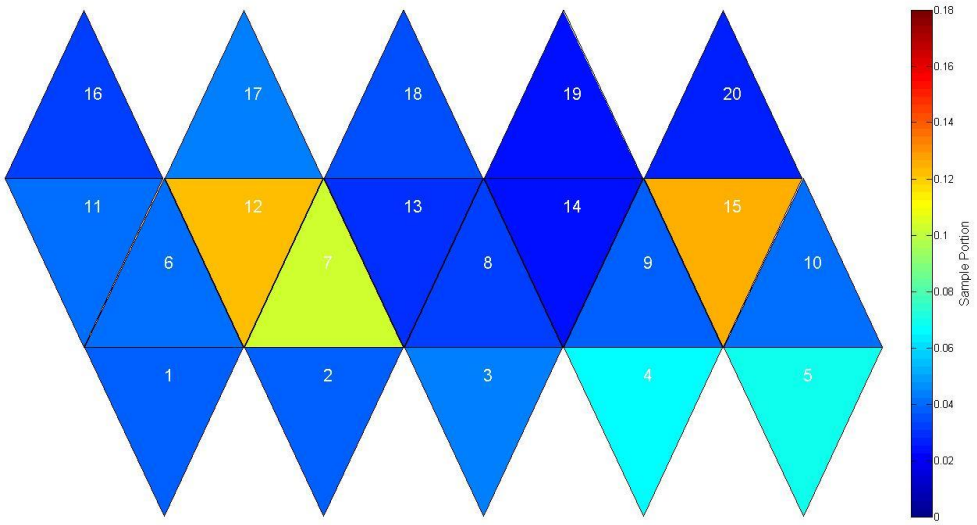


Figure 2. A Spherical Triangular Shell Δ(r, d, t) with Inner Radius r, Thickness d, and Icosahedral Face t in a Sphere

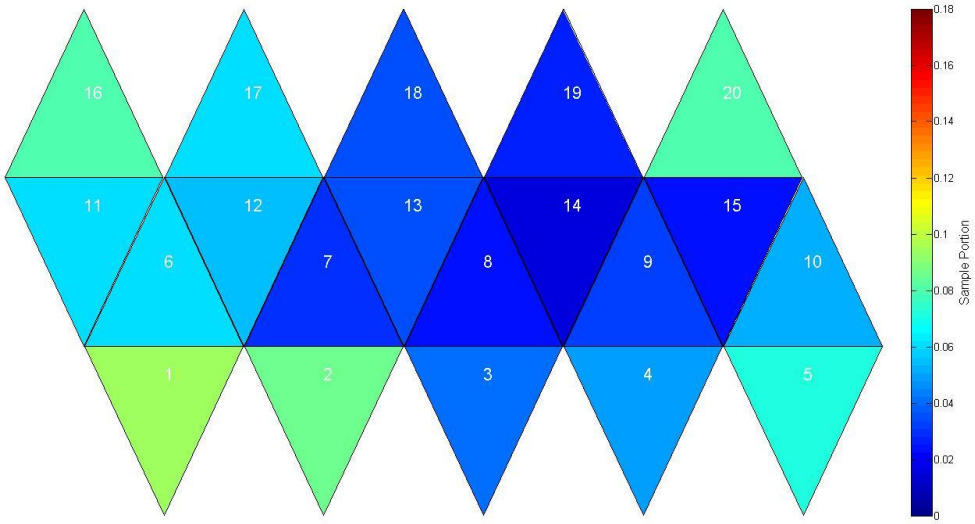
Figure 3 shows the density distribution of Glu-Phe CA-CA contacts on spherical triangular shells in spherical shells with different inner radiuses in a planar icosahedron map. One can find that the contact orientational distribution has two peaks in spherical triangular shells 7 and 15 on sphere shell (r = 5.8A, d = 1.0A), due to hydrogen bonds in forming α-helices and remote β-strand interactions, respectively. For spherical shells with longer inner radiuses, the favorable contact orientations shift to other spherical triangular shells. This indicates that the contact orientational distribution has strong correlation with contact distance, which inspires us to correlate contact distance and orientation to build a more precise contact potential.



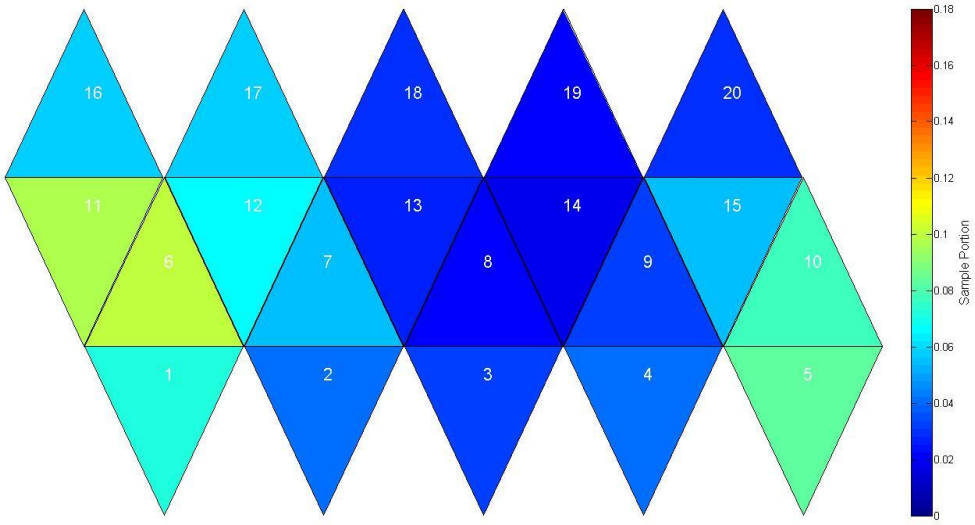
(a) Contact Orientational Distribution on Sphere Shell with r = 5.8A and d = 1.0A. Two distribution peaks in spherical triangular shells 7 and 15, due to formation of α-helices and remote β-strand interactions, respectively. 10 is strongly unfavorable because of being often occupied by the second immediate neighboring residues.



(b) Contact Orientational Distribution on Sphere Shell with r = 7.8A and d = 1.0A. One distribution peak shifts from 7 to 12 while the other peak remains at 15.



(c) Contact Orientational Distribution on Sphere Shell with r = 9.8A and d = 1.0A. Spherical triangular shells 1, 2, 16, and 20 are more favorable than the others.



(d) Contact Orientational Distribution on Sphere Shell with r = 11.8A and d = 1.0A. Spherical triangular shells 6 and 11 are more favorable than the others.

Figure 3. Density Distribution of Glu-Phe Contacts in Spherical Shells with Different Inner Radiuses. The inter-residue contact orientational distribution is strongly correlated with contact distance.

Contact Potential Correlating Contact Distance and Orientation

In ICOSA, the contact potential U_{ij} is to approximate the interaction potential between residues R_i and R_j in forming an inter-residue contact where the center of the CA atom in R_j is located at the spherical triangular shell $\Delta(r_{ij}, d, t_{ij})$ within the icosahedral coordinates of R_i, i.e.,

$$U_{ij} = U\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big). \quad (1)$$

Unlike pair-wise residue distance r_{ij} which is symmetric ($r_{ij} = r_{ji}$), the relative orientations between two contacting residues are asymmetric with respect to their local coordinates. Therefore, $t_{ij} = t_{ji}$ does not generally hold. Consequently, $\Delta(r_{ij}, d, t_{ij}) \neq \Delta(r_{ji}, d, t_{ji})$ and hence $U_{ij} \neq U_{ji}$ in general. Previous studies [49], [50] indicate that the relative orientations of pair-wise residues can be considered independent if they are separated by a sufficiently large number of peptide bonds along the protein backbone. In our contact potential implementation, we adopt a sequence separation gap of 4 as suggested by the previous studies [49], [50], i.e., $|i - j| > 4$, to define R_i and R_j as a contact pair. Summing the potentials of all contact residue pairs together, the overall contact potential of a protein molecule is calculated as

$$U_{protein} = \sum_{i,j}^{|i-j|>4} (U_{ij} + U_{ji}). \quad (2)$$

Finite Ideal Gas Reference State

The contact potential is generated based on Sippl's potentials of mean force method [38]. According to the inverse-Boltzmann theorem,

$$U_{ij} = U\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big) = -kTln\frac{P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)}{P_{exp}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)}. \quad (3)$$

Here k is the Boltzmann constant and T is the temperature. $P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)$ is the observed probability of R_i-R_j contact in spherical triangular shell $\Delta(r_{ij}, d, t_{ij})$ within the local icosahedral coordinates of R_j. $P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)$ is estimated by the fraction of R_i-R_j contacts in $\Delta(r_{ij}, d, t_{ij})$ among all observed R_i-R_j contacts within cutoff distance r_{cutoff}, i.e.,

$$P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big) = \frac{N(R_i, R_j, \Delta(r_{ij}, d, t_{ij}))}{N(R_i, R_j, S\big(r_{min}, r_{cutoff} - r_{min}\big))}, \quad (4)$$

where $N(R_i, R_j, \Delta(r_{ij}, d, t_{ij})$ is the observed R_i - R_j contacts in $\Delta(r_{ij}, d, t_{ij})$ and $N(R_i, R_j, S\big(r_{min}, r_{cutoff} - r_{min}\big))$ is the total number of R_i - R_j contacts in spherical shell $S\big(r_{min}, r_{cutoff} - r_{min}\big)$, which forms the contact space within cutoff distance r_{cutoff} and minimum contact distance r_{min}. $P_{exp}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)$ is the expected probability, which is estimated according to the ideal gas reference state,

$$P_{exp}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big) = \frac{V(\Delta(r_{ij}, d, t_{ij}))}{V(S\big(r_{min}, r_{cutoff} - r_{min}\big))}, \tag{5}$$

where $V(\Delta(r_{ij}, d, t_{ij}))$ is the volume of spherical triangular shell $\Delta(r_{ij}, d, t_{ij})$ and $V(S\big(r_{min}, r_{cutoff} - r_{min}\big))$ is the volume of the spherical shell in the contact space with minimum CA-CA contact distance r_{min} and contact cutoff distance r_{cutoff}. The regularity of icosahedrons allows us to calculate the volume of a spherical triangular shell precisely, which is

$$V\big(\Delta(r_{ij}, d, t_{ij})\big) = (\frac{4}{3}\pi\big(r_{ij} + d\big)^3 - \frac{4}{3}\pi {r_{ij}}^3)/20. \tag{6}$$

The volume of the contact space is

$$V\left(S\big(r_{min}, r_{cutoff} - r_{min}\big)\right) = \tfrac{4}{3}\pi {r_{cutoff}}^3 - \tfrac{4}{3}\pi {r_{min}}^3. \tag{7}$$

Nevertheless, the ideal gas reference state is derived from liquid-state statistical mechanics. In a finite protein system, $N_{contact}(R_i, R_j, S\big(r_{min}, r_{cutoff} - r_{min}\big))$ does not grow with the volume of a spherical shell infinitely. Zhou and Zhou [51] addressed this problem by adjusting the volume growing rate as r^{α} instead and estimated parameter α by a machine learning approach. In ICOSA, we use an alternative approach to remedy this problem by including a parameter σ_{ij} in the denominator of (4) and thus $P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big)$ becomes

$$P_{obs}\big(R_i, R_j, \Delta(r_{ij}, d, t_{ij})\big) = \frac{N_{contact}(R_i, R_j, \Delta(r_{ij}, d, t_{ij}))}{N_{contact}(R_i, R_j, S\big(r_{min}, r_{cutoff} - r_{min}\big)) + \sigma_{ij}}. \tag{8}$$

σ_{ij} can be interpreted as the number of "missing" $R_i - R_j$ contacts if the protein system is infinite. Then, we obtain additional statistics $N_{contact}(R_i, R_j, S\big(r_{cutoff}, d\big))$ from the large protein chain set, which is the number of $R_i - R_j$ contacts in the spherical cell with thickness d beyond r_{cutoff}. By assuming that there is no interaction between R_i and R_j if they are separated by a distance longer than r_{cutoff} in space, i.e., we can get

$$U\big(R_i, R_j, S(r_{ij}, d)\big) = -kTln\frac{P_{obs}\big(R_i, R_j, S(r_{cutoff}, d)\big)}{P_{exp}\big(R_i, R_j, S(r_{cutoff}, d)\big)} = 0, \tag{9}$$

where $P_{exp}\big(R_i, R_j, S(r_{cutoff}, d)\big)$ is estimated as the volume ratio of the correspondent spherical shells such as

$$P_{exp}\big(R_i, R_j, S(r_{cutoff}, d)\big) = \frac{V(S(r_{cutoff}, d))}{V(S\big(r_{min}, r_{cutoff} - r_{min}\big))}. \tag{10}$$

By substituting (8) and (9) into (10), we can estimate σ_{ij} by

$$\sigma_{ij} = \frac{N_{contact}(R_i, R_j, S(r_{cutoff}, d)) * V(S\big(r_{min}, r_{cutoff} - r_{min}\big))}{V(S(r_{cutoff}, d))} - N_{contact}(R_i, R_j, S\big(r_{min}, r_{cutoff} - r_{min}\big)). \tag{11}$$

Contact Cutoff Distance

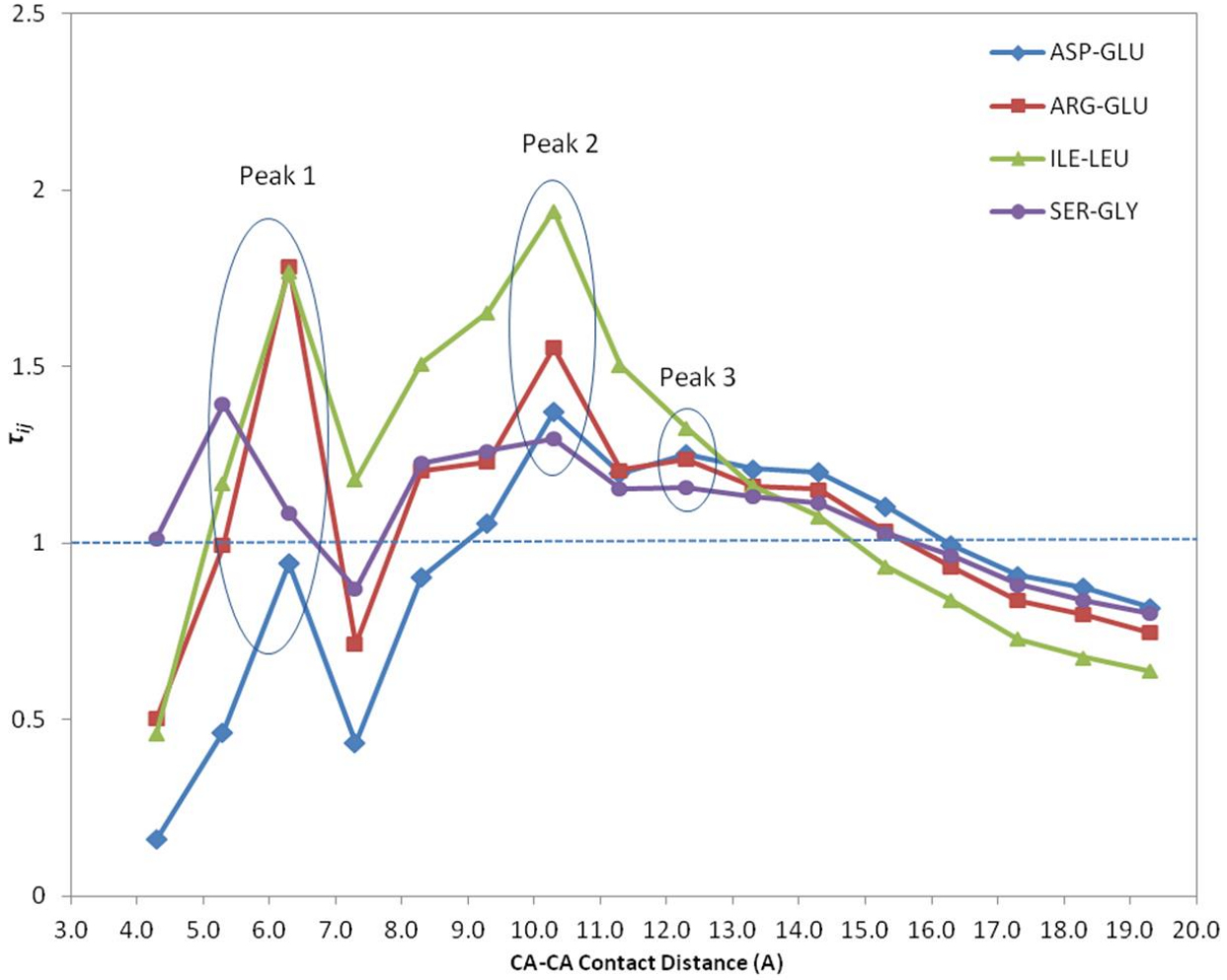


Figure 4. Contact distribution of Arg-Glu, Asp-Glu, Ile-Leu, and Ser-Gly over CA-CA contact distances in every 1A interval. Peak 1 between 5A-7A, peak 2 between 10A-11A, and peak 3 (if exists) between 12A-13A.

In the literature, there has been a wide spectrum in the cutoff distance, r_{cutoff}, ranging from 6A to 16A, for pair-wise residue contact definitions. Nevertheless, there is lack of clear statistical or theoretical justifications on what the most appropriate r_{cutoff} is. In this article, we denote τ_{ij} as the sample ratio over volume ratio to study the contact distribution over cutoff distance in $R_i - R_j$ contacts. τ_{ij} is defined as

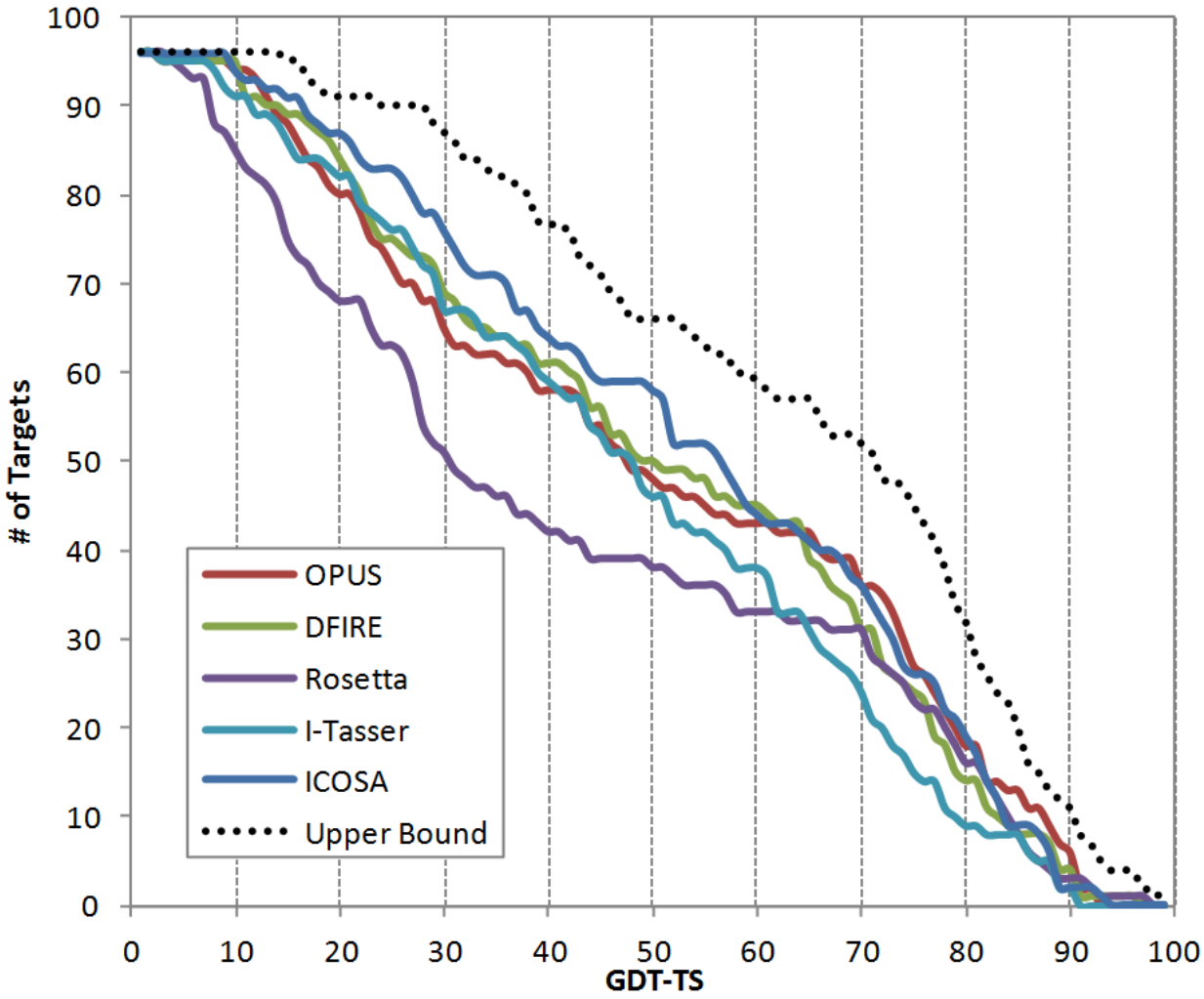
$$\tau_{ij} = \frac{N_{contact}(R_i, R_j, S(r,d))/N_{contact}(R_i, R_j, S(r_{min}, r_{cutoff} - r_{min}))}{V(S(r,d))/V(S(r_{min}, r_{cutoff} - r_{min}))}. \quad (12)$$

Clearly, if there is no interaction occurred between residues R_i and R_j such as the ideal gas state, τ_{ij} is 1. Figure 4 plots the contact distributions of Arg-Glu, Asp-Glu, Ile-Leu, and Ser-Gly over contact distance in every 1A interval. It is interesting to notice that the distributions of contacts between different amino acids share certain common characteristics. The first peak appears at 5A-7A in r_{cutoff}, which is mainly due to formations of hydrogen bonds for secondary structures as well as other short-range interactions. Therefore, the contact definition with r_{cutoff} less than 8A typically captures the first peak only. Ser and Gly are residues with strong backbone affinity, where the first peak of Ser-Gly interaction occurs in shorter range than the others. The second peak appears between 10A to 11A. The first and second peaks appear in the similar contact distance in almost all residue pairs. It is interesting to notice that the first peak is higher than the second peak in Arg-Glu interaction. This is due to the fact that Arg and Glu have opposite charged side chains and thus the majority of Arg-Glu interactions occur in short distances. In contrast, both Asp and Glu have negative side chain charges and therefore the second peak in ASP-GLU interaction are higher than the first one, which indicates that the major Asp-Glu interactions occurs in longer distance, similar to hydrophobic interactions such as Ile-Leu. In certain residue pairs, such as Arg-Glu and Asp-Glu shown in Figure 4, the third peak, although not as obvious as the first two peaks, appears between 12A and 13A. The second and third peaks result from side chain packing and interactions. These peaks represent interresidue interaction concentrations and afterwards τ_{ij} decreases gradually past 1.0. This suggests that a distance cutoff r_{cutoff} capturing all three peaks is most likely an appropriate contact cutoff distance. In ICOSA, we start the distance intervals at $r_{min} = 2.8$A where minimum CA-CA contact distance occurs in our dataset, increase each interval with a step size of 1A, and adopt 12.8A as our CA-CA contact cutoff distance so that all three peaks are include. A longer cutoff does not further improve ICOSA, but demands higher computational cost.

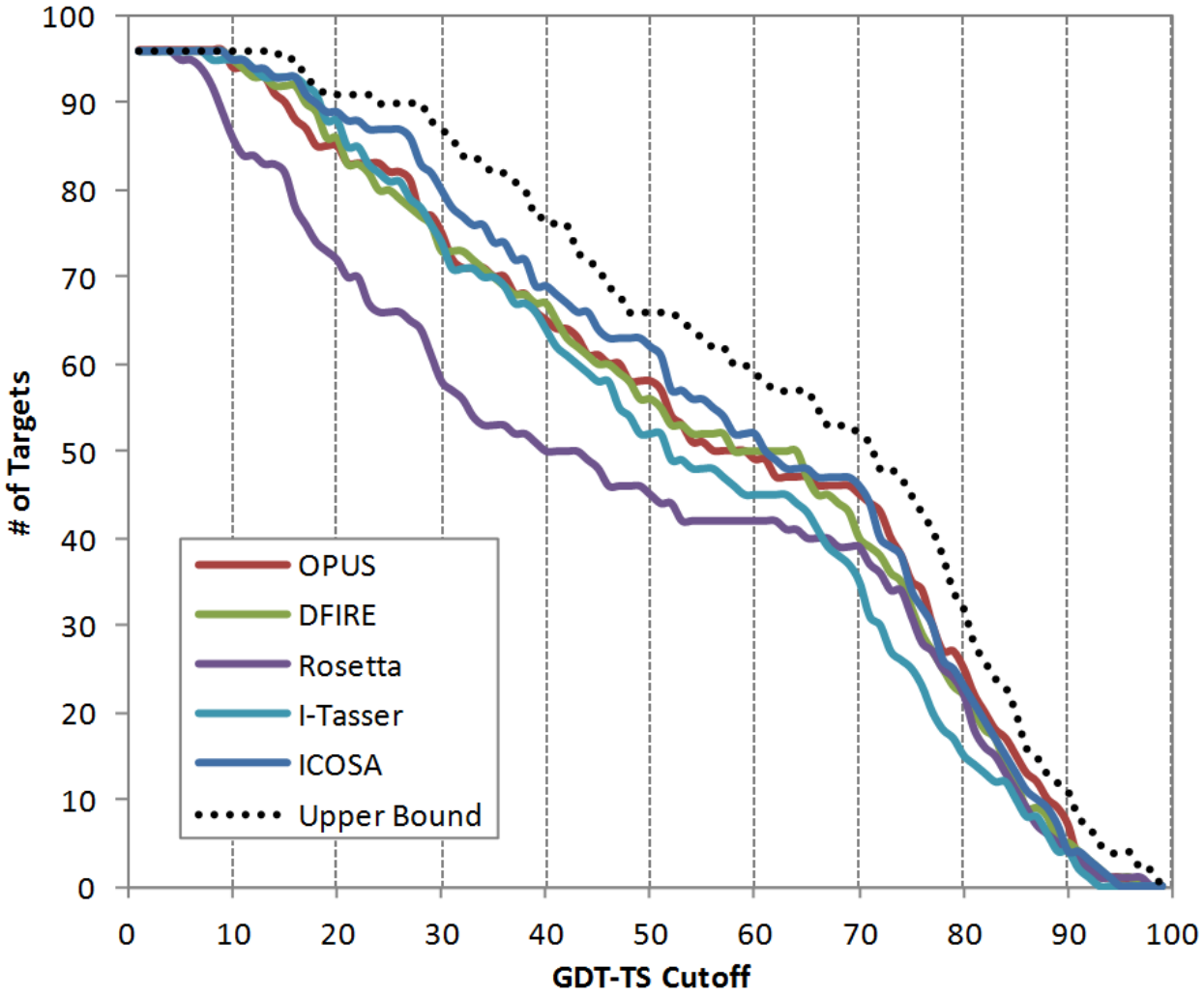
Results

Identifying CASP10 Predicted Models

We compare ICOSA with all-atom energy potentials including OPUS [52], DFIRE [12], Rosetta [53], and I-Tasser [42] on the predicted models in 96 domains in CASP10 [46] targets, where many of them are partially correct models. We use the GDT-TS (Global Distance Test – Total Score) [54], which indicates the percentage of the conformational structure superimposed correctly onto the native, to measure the quality of a model. Figures 5(a) and 5(b) show the number of targets that the top-ranked model and the best top-5 ranked models in each CASP10 target identified by the energy potentials can fit under gradually increasing GDT-TS cutoffs (step size 1), respectively. For GDT-TS cutoff values between 20 and 60, ICOSA yields a higher percentage of targets than the all-atom potentials and is closer to the upper bound, which indicates that ICOSA is more robust in identifying low resolution models with partially correct conformations. For GDT-TS cutoff values exceeding 60, ICOSA starts to be slightly surpassed by OPUS and occasionally by DFIRE. This is because the all-atom potentials carry information on major side chain atoms and thus allow more sensitive identification of highly accurate models than ICOSA which is based on backbone atoms distance and orientation only. Anyway, although slightly worse than the best all-atom potentials, ICOSA is still quite comparable to other all-atom potentials for high GDT-TS cutoffs.



(a) Top ranked models



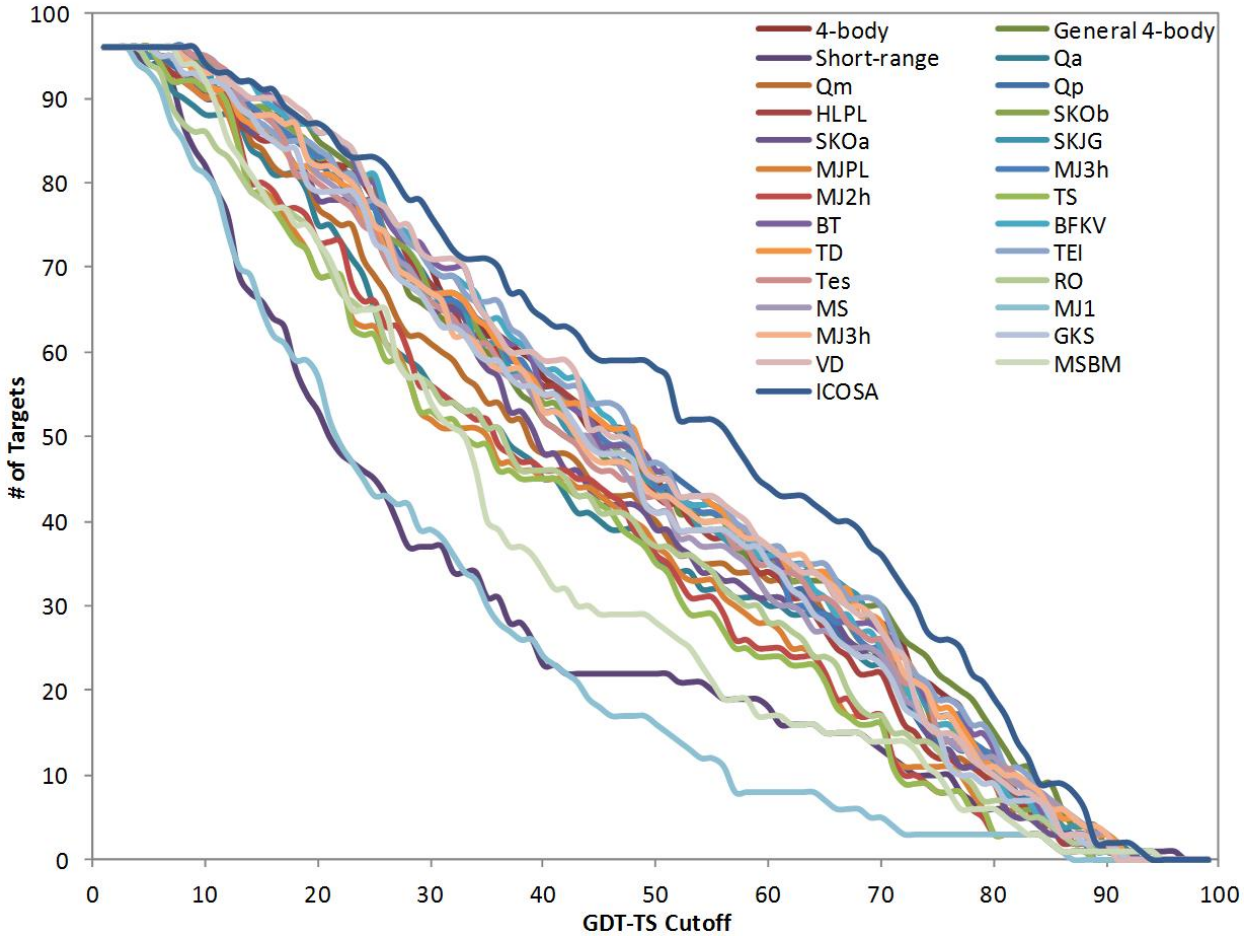
(b) Best top-5 ranked models

Figure 5. Comparison of ICOSA with all-atom energy potentials (OPUS, DFIRE, Rosetta, and I-Tasser) in identifying the best predicted models in CASP10 models. ICOSA outperforms the all-atom energy potentials for GDT-TS cutoff values between 20 and 60 and is comparable to the all atom energy potentials for high GDT-TS (>60) cutoffs.

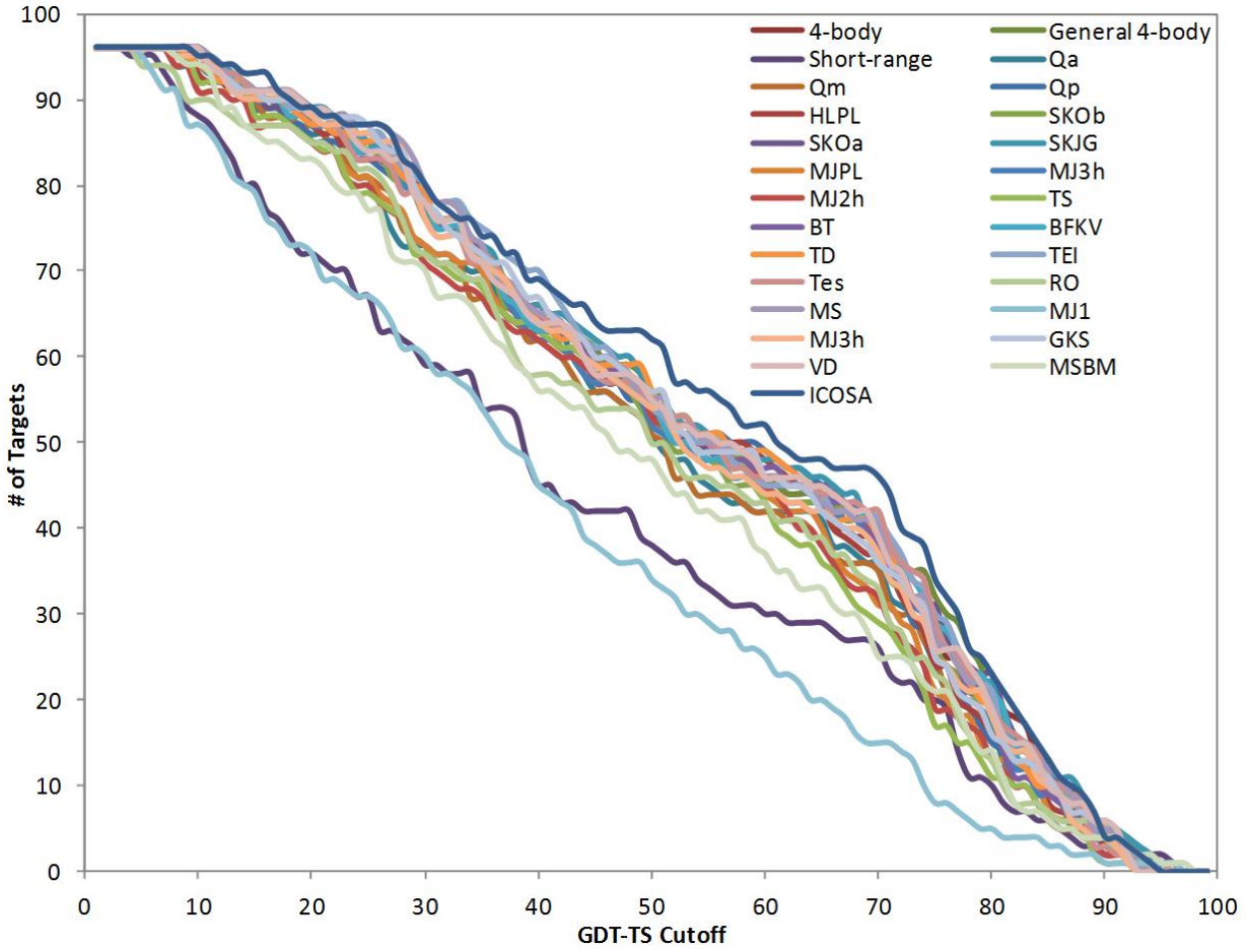
Comparison with other Coarse-grained Knowledge-based Potentials

We also compare the performance of ICOSA with a set of coarse-grained knowledge-based potentials presented in the potentials'R'Us web-server [55] representing a variety of contact potentials widely used in protein structure modeling applications. The potentials set consists of 23 two-body potentials including Qa, Qm, Qp [56], HLPL, MJPL [5], SKOa, SKOb, SJKG [41, 57], MJ1, MJ2h, MJ3, MJ3h [9, 39, 58], TS [59], BT [10], BFKV [60], TD [61], TEl, TEs [57], RO [62], MS [3], GKS[63], VD [1], BL [64], and MSBM [53, 65] as well as a short-range potential [66] and two four-body potentials [30]. Figures 6(a) and 6(b) show the top ranked or best top-5

ranked models identified by different energy potentials with respect to gradually increasing GDT-TS cutoffs on CASP10 models, respectively. ICOSA has the highest overall number of targets in almost all GDT-TS cutoffs in both top ranked and top-5 models.



(a) Top Ranked models



(b) Best top-5 ranked models

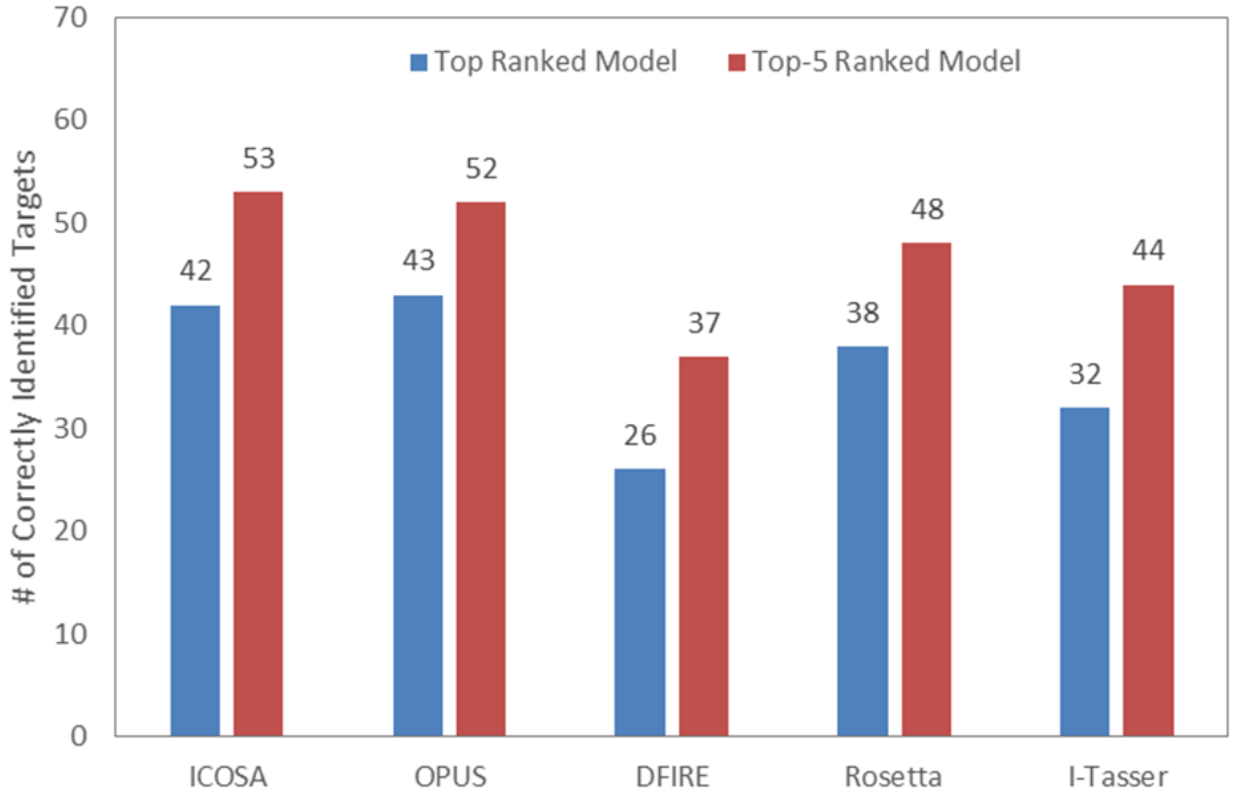
Figure 6. Comparison of ICOSA with 26 coarse-grained knowledge-based potentials in identifying the best predicted models in CASP10 models. ICOSA dominates the coarse-grained potentials for most GDT-TS cutoff values.

Identifying Near-Native Models

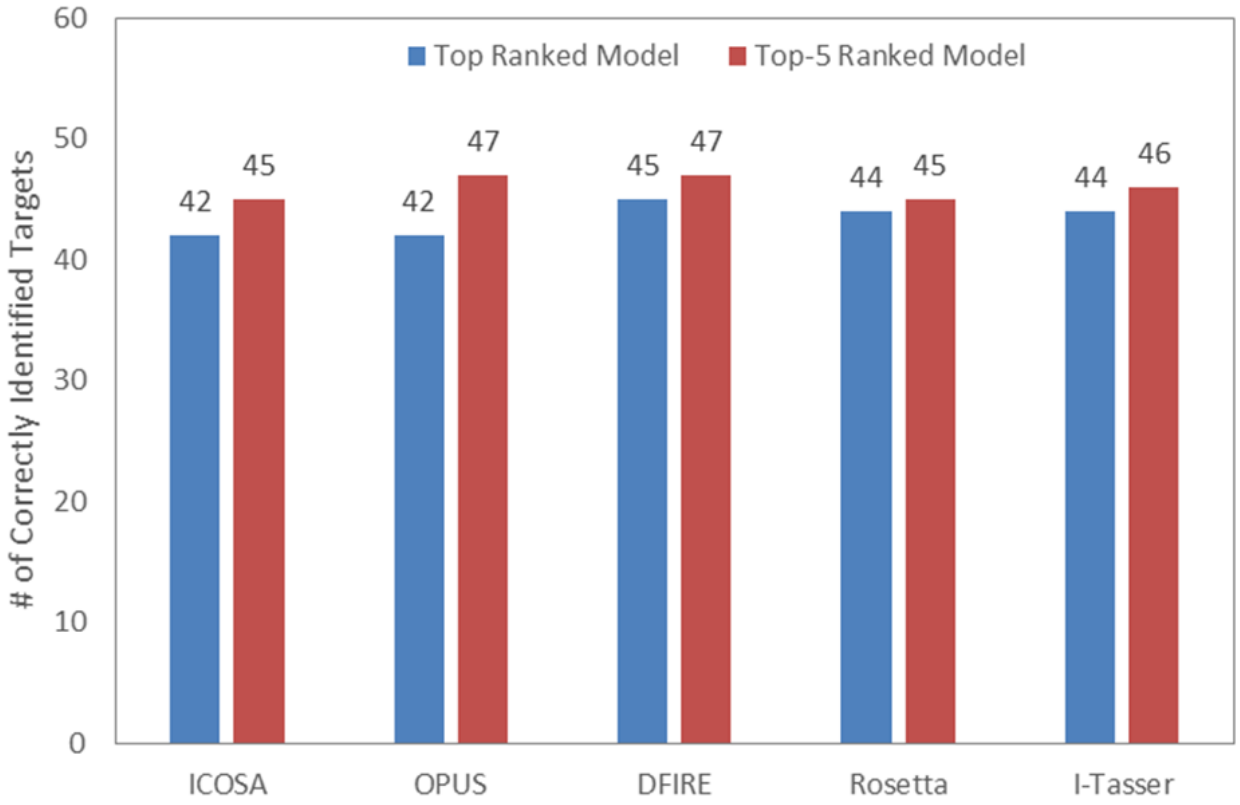
We use Rosetta and I-Tasser decoy sets to compare ICOSA with popular all-atom energy potentials, including Rosetta, I-Tasser, DFIRE, and OPUS, for discriminating near-natives from misfold decoys. To clear the concern that the knowledge-based potentials may bias to the experiment generated models, in addition to the original decoys in the benchmarks, for each protein chain target, we add 20 near-native conformations to the decoy set by relaxing the native under the OPLS-AA potential [67] with SGB solvation [68] and then remove the native. These near-native conformations mostly have GDT-TS scores over 80.

Figures 7(a) and 7(b) show the numbers of targets that the top ranked and best top-5 ranked models have GDT-TS scores over 80, respectively. As one can see, ICOSA is quite comparable to the all-atom potentials. In Rosetta decoy sets, ICOSA has similar performance as OPUS (one less

in top ranked models but one more in best top-5 ranked models) and is better than the other all-atom potentials. In I-Tasser decoy sets, ICOSA is slightly worse than the other all-atom potentials; however, the difference is within three in the numbers of corrected identified targets.



(a) Rosetta Decoy Sets



(b) I-Tasser Decoy Sets

Figure 7. Comparing ICOSA with OPUS, DFIRE, Rosetta, and I-Tasser in identifying near-native models on Rosetta and I-Tasser decoy sets

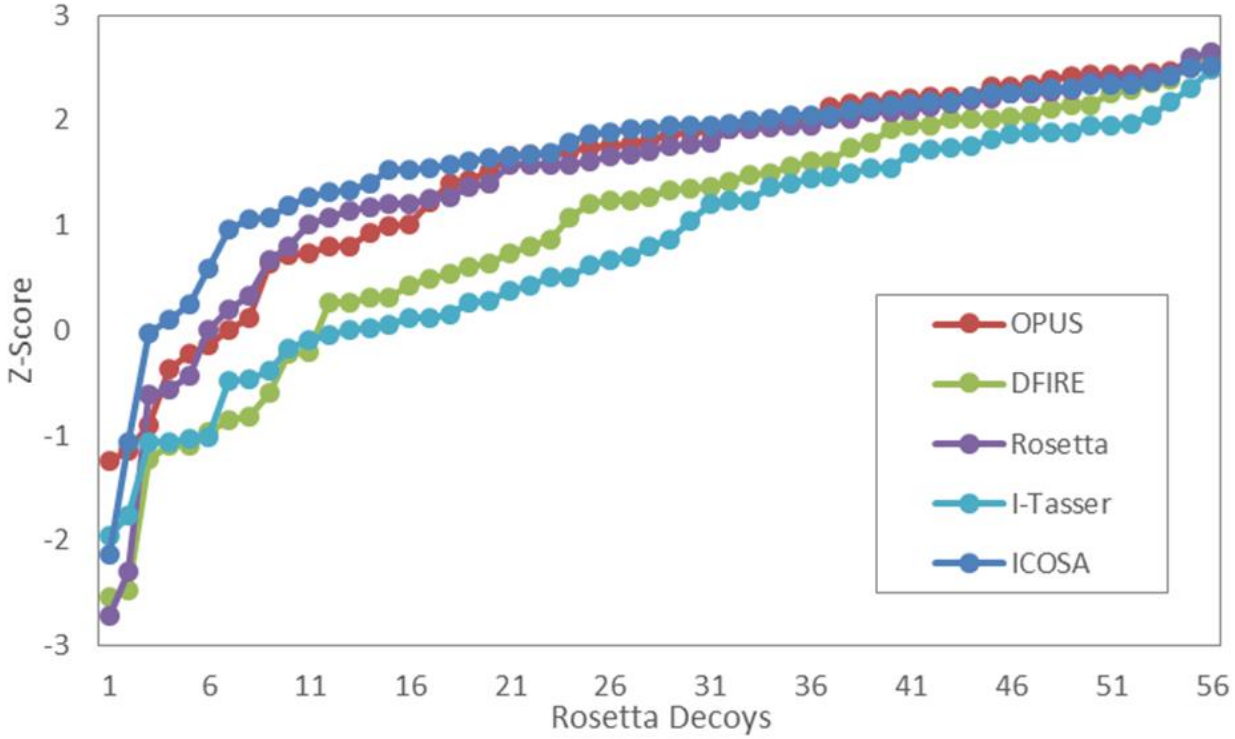
We apply Z-score to measure the performance of energy potentials in identifying near-natives. Considering the conformations with over 80 GDT-TS score as near-natives, the Z-score for a decoy set is calculated as

$$Z\text{-}score = \frac{\overline{U}_{misfolds} - \overline{U}_{near-natives}}{\sigma},$$

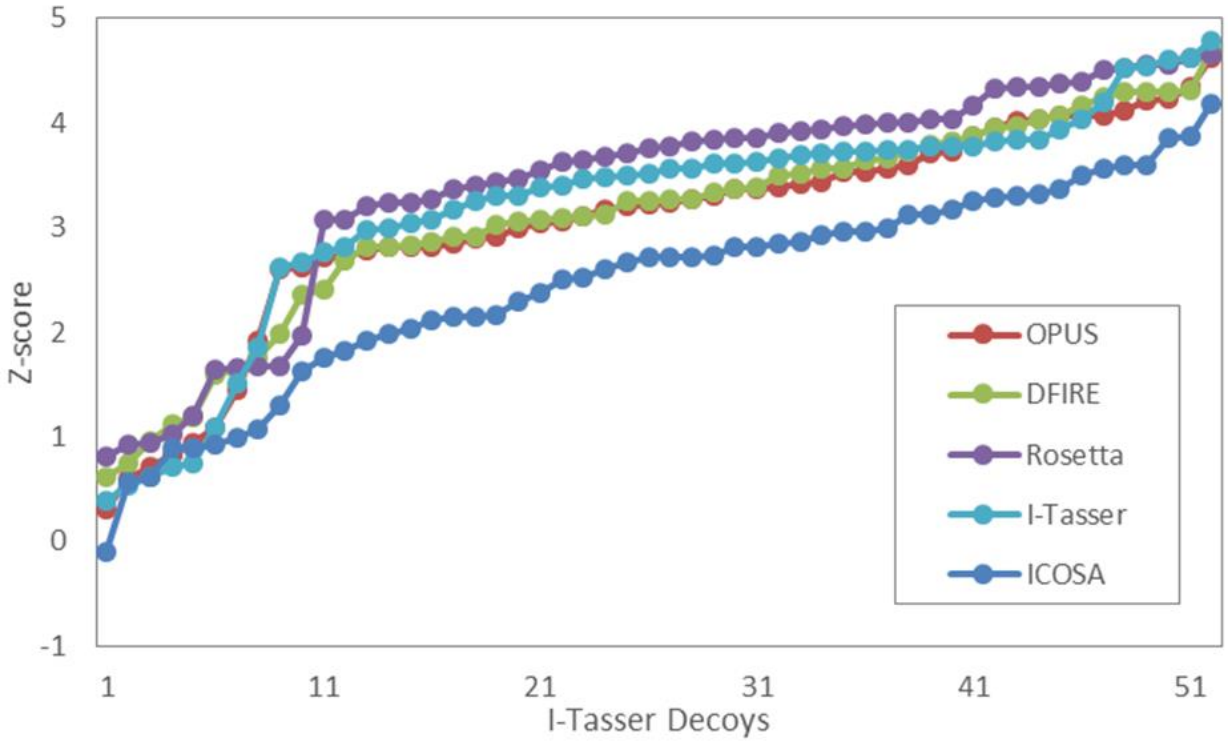
where $\overline{U}_{near-natives}$ and $\overline{U}_{misfolds}$ are the average energy potential values of the near-natives (GDT-TS > 80) and the misfolds, respectively, and σ is the standard deviation. The Z-score measures the separation between the near-natives and misfold conformations in a given energy potential. The higher the Z-score value, the better the energy potential in discriminating the near-natives.

Figures 8(a) and 8(b) show the sorted Z-scores calculated by ICOSA, OPUS, DFIRE, Rosetta, and I-Tasser for targets in Rosetta and I-Tasser decoy sets, respectively. One can find that, in Rosetta decoy sets, in almost all cases, ICOSA either has the highest Z-scores or second highest Z-scores slightly less than OPUS. However, it turns to be the opposite in I-Tasser decoy sets. The main reason is that many decoys presented in I-Tasser benchmark exhibit slight steric clashes in side chain atoms and/or unfavorable backbone torsion conformations, which are easily captured by the all-atom potentials but not by ICOSA which only calculates CA-CA distances and

orientations. In fact, in Figure 11 presented in Discussion section, combined ICOSA with a torsion potential [69] leads to significantly improved decoy discrimination in targets such as 1fo5A and 2cr7A where ICOSA alone does not perform well.



(a) Rosetta decoy sets



(b) I-Tasser decoy sets

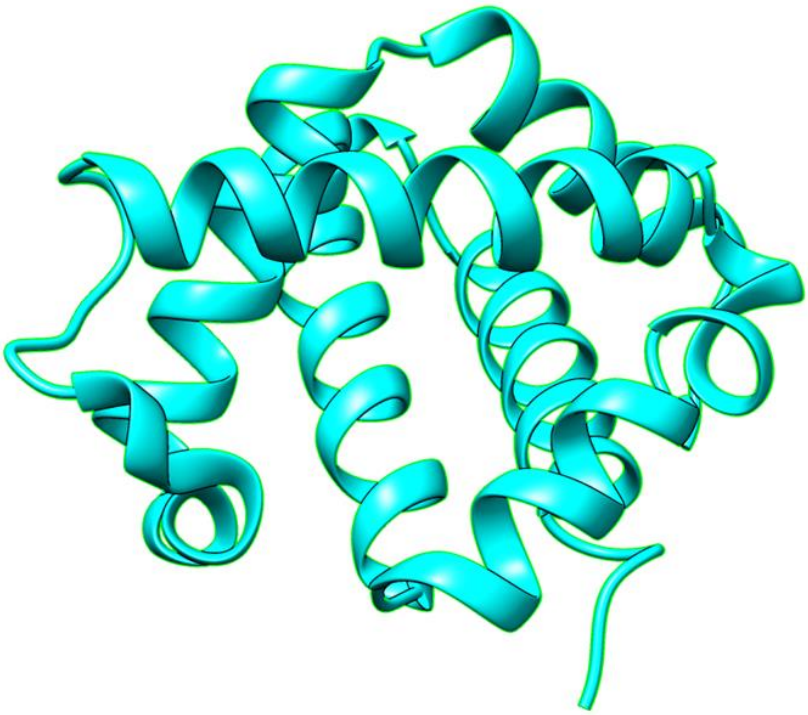
Figure 8. Z-score sorted for targets in (a) Rosetta and (b) I-Tasser decoy sets calculated by ICOSA, OPUS, DFIRE, Rosetta and I-Tasser

It is important to notice that Rosetta, I-Tasser, DFIRE, and OPUS are all-atom energy potentials while ICOSA only relies on backbone atoms information. As shown in Rosetta and I-Tasser decoy sets, the coarse-grained ICOSA contact potential has comparable decoy discrimination effectiveness to identify near-natives as the all-atom energy potentials.

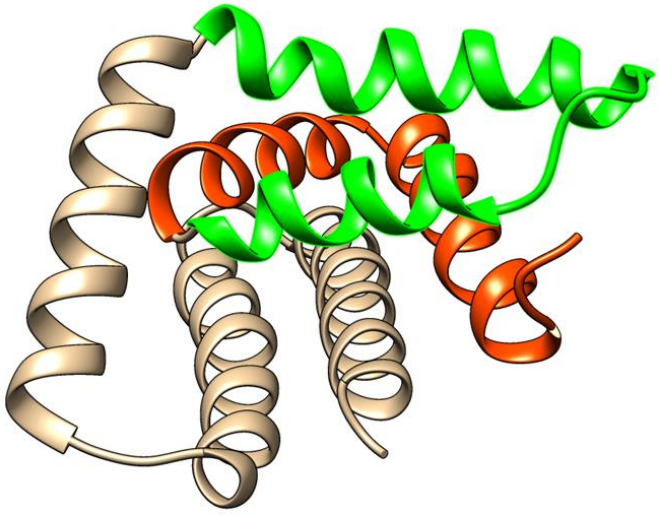
Effectiveness of Correlating Contact Distance and Orientation in Icosahedral Local Coordinates

To analyze the property of ICOSA and its extended finite ideal gas reference state in icosahedral local coordinates, we compare ICOSA and DFIRE on a near-native model (Figure 9(a), GDT-TS = 95.21) and an erroneous misfold (Figure 9(b), GDT-TS = 14.68) of 1cg5 B chain. Compared to the near-native model, one can find that three helices are packed incorrectly in the erroneous one. ICOSA successfully identifies the near-native model by assigning a lower energy value than the erroneous one while DFIRE fails. Figures 9(c) and 9(d) plot ICOSA and DFIRE potential values on each contact residue pair on the contact maps of the near-native and the erroneous model, respectively. The contacts of the incorrectly packed helices in the erroneous model are highlighted in Figures 9(c) and 9(d). ICOSA indicates that some of these contacts are highly unfavorable with

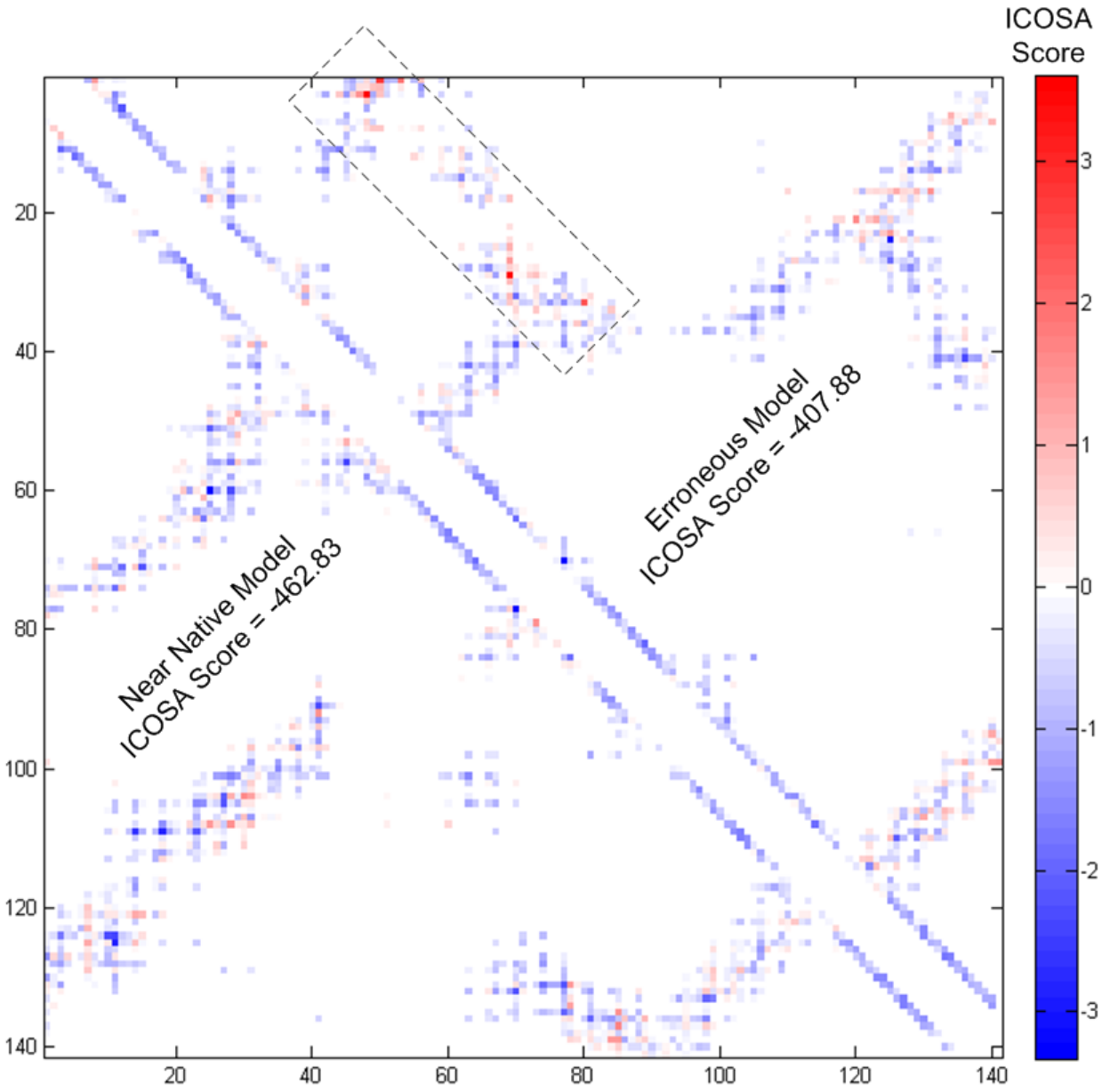
high energy value while such unfavorability is not shown by DFIRE. Although both ICOSA and DFIRE are based on finite ideal gas reference state, correlating contact distance and orientation allows ICOSA to be more sensitive in measuring favorability of interresidue contacts than DFIRE, which is based on contact distance only.



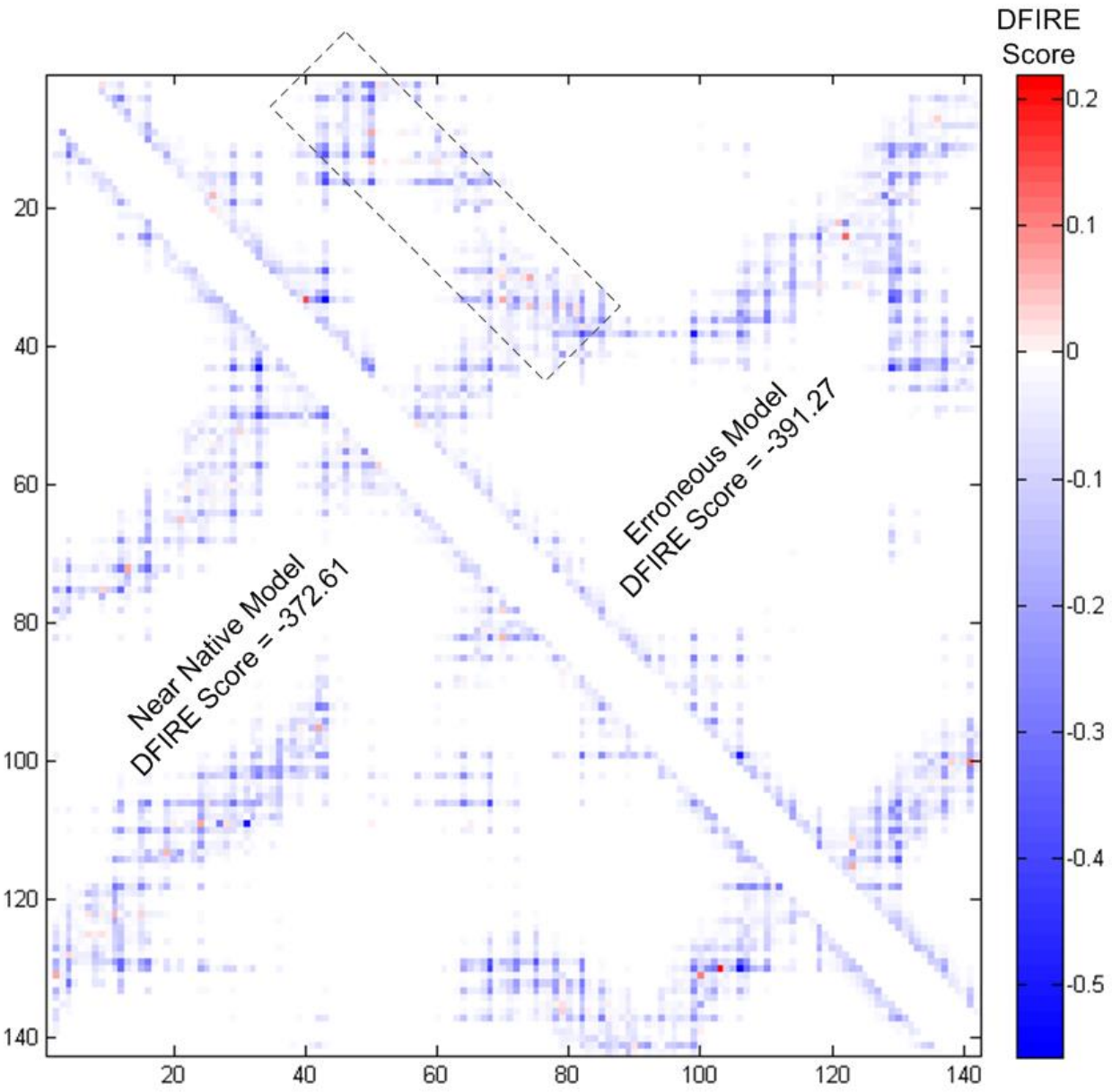
(a) Near Native Model of 1cg5 B Chain (GDT-TS = 95.21)



(b) Erroneous Model of 1cg5 B Chain (GDT-TS = 14.68). The first two helices (1-34) are shown in red and the next two helices (36-72) are shown in green.



(c) ICOSA potential on contact map (upper triangle: contact map of the erroneous model; lower triangle: contact score of the near-native model).



(d) DFIRE potential on contact map (upper triangle: contact map of the erroneous model; lower triangle: contact score of the near-native model).

Figure 9. Comparison of ICOSA and DFIRE potential energy values between a near-native model and an erroneous model of 1cg5 B chain. ICOSA is capable of capturing the unfavorable interresidue contacts between the highlighted helices and thus identifies the near native model with a lower score. In contrast, DFIRE is insensitive to these unfavorable contacts.

Cutoff Distances in ICOSA

Figure 10 compares the ICOSA-RMSD plots for 1a68 and 1bm8 decoys in Rosetta benchmark using 6.8A, 10.8A, and 12.8A contact cutoff distances. 12.8A cutoff distance captures all three peaks in the contact distribution plotted in Figure 4 while 10.8A and 6.8A cutoff distances include the first two and one peaks, respectively. As shown in Figure 10, the long-range contacts between 10.8A to 12.8A play an important role. Only when these long-range contacts are taken into considerations, ICOSA can successfully discriminate the native and near-natives from the incorrect decoys in 1a68 and 1bk2. These computational results agree with our statistical analysis in contact cutoff distance in Materials and Methods section.

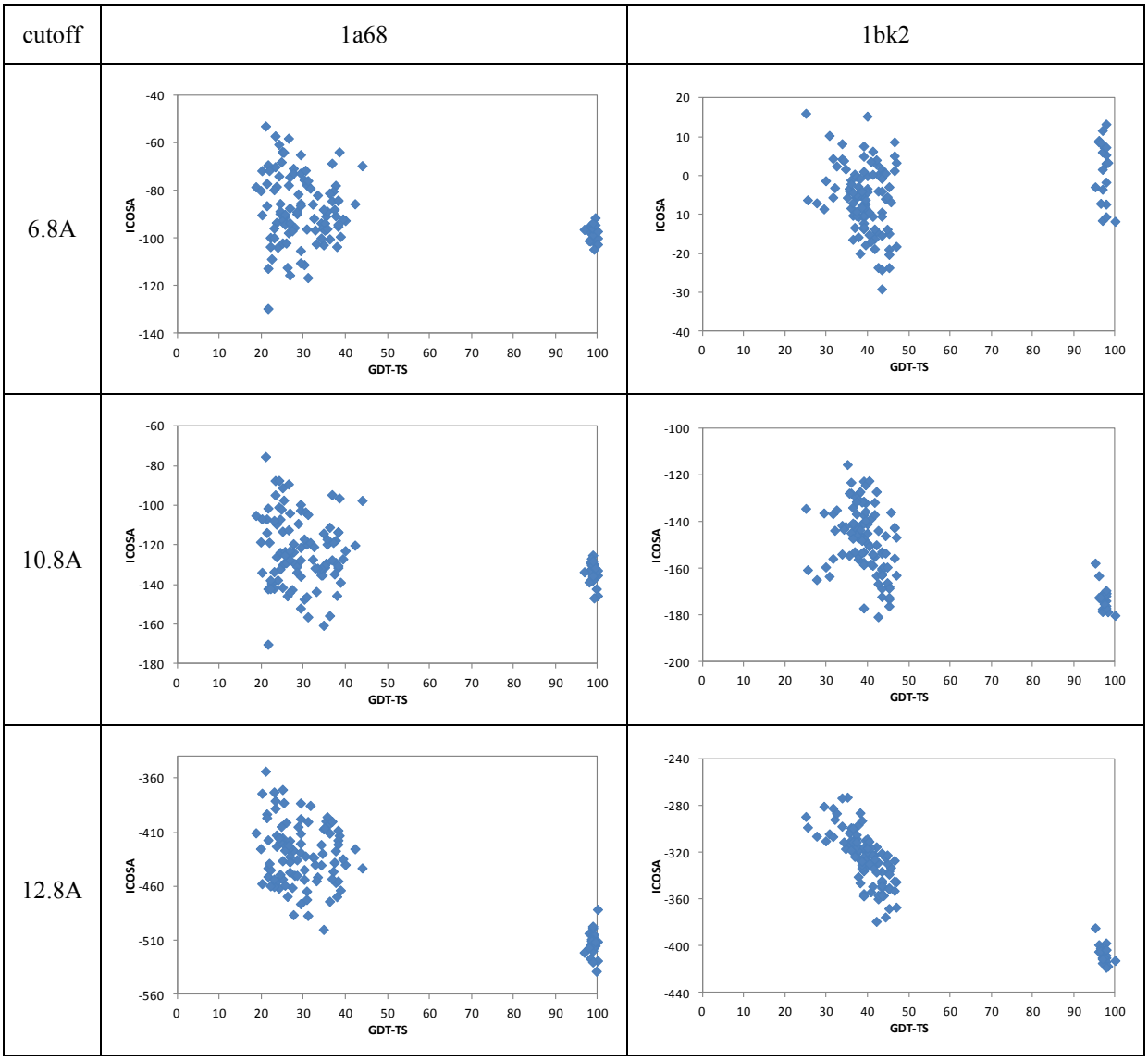


Figure 10. Energy–GDT-TS plots for ICOSA adopting contact cutoff distances of 6.8A, 10.8A, and 12.8A in 1a68 and 1bk2 decoys in Rosetta benchmark. Long-range inter-residue contacts between 10.8A to 12.8A are critical to differentiate near-natives and incorrect models

Discussions

ICOSA has demonstrated its sensitivity and accuracy comparable to all-atom, fine-grained potentials in identifying CASP10 models and discriminating near-natives from misfolds in Rosetta and I-Tasser decoy sets. As a coarse-grained contact potential, ICOSA only needs information of the backbone atoms, which makes it particularly suitable for modeling protein structures with reduced representation.

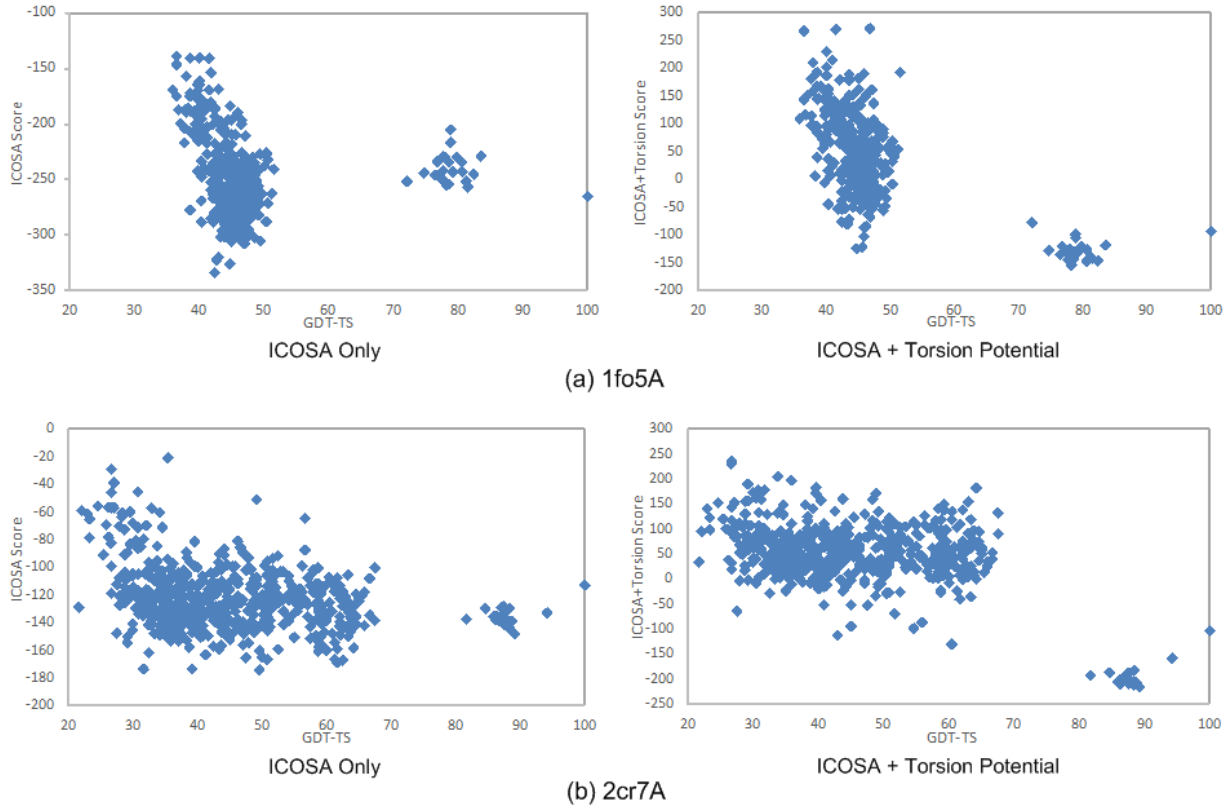


Figure 11. Coupled with the torsion potential, ICOSA is able to identify the near-natives in 1fo5A and 2cr7A decoy sets in I-Tasser benchmark, where ICOSA alone fails.

There are several ways to potentially improve ICOSA in the future. First of all, it is certainly possible to combine ICOSA with other coarse-grained potentials measuring different characteristics of protein molecules, such as torsion angle distributions [69], secondary structures [70], solvent accessibility, beta-sheet pairing, etc., to achieve even better accuracy. Figure 11 shows that coupled with a torsion potential [69], ICOSA can successfully separate most of the near-natives from the other decoys, where ICOSA alone fails. However, determination of appropriate weights to combine these potential terms together requires deliberate considerations. Also, as the number of protein structures deposited in PDB continues to grow and enough samples for certain folding classes become available in the future, the ICOSA approach can be applied to generate fold-specific knowledge-based potentials. Moreover, ICOSA can be extended to an all-

atom potential, where the icosahedral local coordinates are built to correlate orientation and distance in each atom pair interaction.

Furthermore, similar to many other knowledge-based energy potentials, ICOSA is derived from a non-redundant subset of protein chains from the PDB. Recent study from Yanover et al. [71] indicates that using the redundancy-weighted method, which takes advantage of all available structures in the PDB to obtain better estimations of structure-sequence distributions, can improve the accuracy of knowledge-based energy potentials. Using a redundancy-weighted dataset to generate contact statistics may further improve the accuracy and sensitivity of ICOSA.

Acknowledgements

YL acknowledges support from NSF under grant 1066471.

References

[1] Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys. 1998;109:11101-8.
[2] Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins. 2000;38:134-48.
[3] Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol. 1996;264:1164-79.
[4] Khatun J, Khare SD, Dokholyan NV. Can contact potentials reliably predict stability of proteins? J Mol Biol. 2004;336:1223-38.
[5] Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol. 1996;258:367-92.
[6] Wu ST, Szilagyi A, Zhang Y. Improving Protein Structure Prediction Using Multiple Sequence-Based Contact Predictions. Structure. 2011;19:1182-91.
[7] Russ WP, Ranganathan R. Knowledge-based potential functions in protein design. Curr Opin Struc Biol. 2002;12:447-52.
[8] Ravikant DVS, Elber R. Energy design for protein-protein interactions. J Chem Phys. 2011;135.
[9] Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996;256:623-44.
[10] Betancourt MR, Thirumalai D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci. 1999;8:361-9.
[11] Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins. 2000;38:3-16.
[12] Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction (vol 11, pg 2714, 2002). Protein Science. 2003;12:2121-.
[13] Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001;44:223-32.

[14] Berrera M, Molinari H, Fogolari F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. Bmc Bioinformatics. 2003;4.
[15] Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol. 1997;267:707-26.
[16] Zimmer R, Wohler M, Thiele R. New scoring schemes for protein fold recognition based on Voronoi contacts. Bioinformatics. 1998;14:295-308.
[17] McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. P Natl Acad Sci USA. 2003;100:3215-20.
[18] Esque J, Leonard S, de Brevern AG, Oguey C. VLDP web server: a powerful geometric tool for analysing protein structures in their environment. Nucleic Acids Res. 2013;41:W373-W8.
[19] Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. J Comput Biol. 1996;3:213-21.
[20] Reck GM, Vaisman II. Decoy discrimination using contact potentials based on delaunay tessellation of hydrated proteins. The 4th International Symposium on Voronoi Diagrams in Science and Engineering2007. p. 158-67.
[21] Li X, Hu CY, Liang J. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins. 2003;53:792-805.
[22] Li X, Liang J. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. Proteins-Structure Function and Bioinformatics. 2005;60:46-65.
[23] Esque J, Oguey C, de Brevern AG. Comparative Analysis of Threshold and Tessellation Methods for Determining Protein Contacts. J Chem Inf Model. 2011;51:493-507.
[24] Brocchieri L, Karlin S. How are close residues of protein structures distributed in primary sequence? P Natl Acad Sci USA. 1995;92:12136-40.
[25] Faure G, Bornot A, de Brevern AG. Protein contacts, inter-residue interactions and side-chain modelling. Biochimie. 2008;90:626-39.
[26] Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci U S A. 2000;97:2550-5.
[27] Rojnuckarin A, Subramaniam S. Knowledge-based interaction potentials for proteins. Proteins. 1999;36:54-67.
[28] Zhao F, Xu J. A Position-Specific Distance-Dependent Statistical Potential for Protein Structure and Functional Study. Structure. 2012;20:1118-26.
[29] Duan MJ, Zhou YH. A contact energy function considering residue hydrophobic environment and its application in protein fold recognition. Genomics Proteomics Bioinformatics. 2005;3:218-24.
[30] Feng YP, Kloczkowski A, Jernigan RL. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins-Structure Function and Bioinformatics. 2007;68:57-66.
[31] Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Science. 1997;6:1467-81.
[32] Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Fold Des. 1996;1:357-70.
[33] Buchete NV, Straub JE, Thirumalai D. Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures. J Chem Phys. 2003;118:7658-71.
[34] Mukherjee A, Bhimalapuram P, Bagchi B. Orientation-dependent potential of mean force for protein folding. J Chem Phys. 2005;123.

[35] Makino Y, Itoh N. A knowledge-based structure-discriminating function that requires only main-chain atom coordinates. Bmc Struct Biol. 2008;8.
[36] Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. Bmc Bioinformatics. 2010;11.
[37] Moughon SE, Samudrala R. LoCo: a novel main chain scoring function for protein structure prediction based on local coordinates. Bmc Bioinformatics. 2011;12.
[38] Sippl MJ. Calculation of Conformational Ensembles from Potentials of Mean Force - an Approach to the Knowledge-Based Prediction of Local Structures in Globular-Proteins. J Mol Biol. 1990;213:859-83.
[39] Miyazawa S, Jernigan RL. Estimation of Effective Interresidue Contact Energies from Protein Crystal-Structures - Quasi-Chemical Approximation. Macromolecules. 1985;18:534-52.
[40] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998;275:895-916.
[41] Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci. 1997;6:676-88.
[42] 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.
[43] Deng HY, Jia Y, Wei YY, Zhang Y. What is the best reference state for designing statistical atomic potentials in protein structure prediction? Proteins-Structure Function and Bioinformatics. 2012;80:2311-22.
[44] Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Science. 2006;15:2507-24.
[45] Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins. 2003;53:76-87.
[46] Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins-Structure Function and Bioinformatics. 2014;82:7-13.
[47] Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics. 2003;19:1589-91.
[48] Feng Y, Jernigan RL, Kloczkowski A. Orientational distributions of contact clusters in proteins closely resemble those of an icosahedron. Proteins-Structure Function and Bioinformatics. 2008;73:730-41.
[49] Sippl MJ. Knowledge-Based Potentials for Proteins. Curr Opin Struc Biol. 1995;5:229-35.
[50] Thomas PD, Dill KA. Statistical potentials extracted from protein structures: How accurate are they? J Mol Biol. 1996;257:457-69.
[51] Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci. 2002;11:2714-26.
[52] Lu MY, Dousis AD, Ma JP. OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol. 2008;376:288-301.
[53] Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol. 1997;268:209-25.
[54] Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003;31:3370-4.

[55] Feng YP, Kloczkowski A, Jernigan RL. Potentials 'R'Us web-server for protein energy estimations with coarse-grained knowledge-based potentials. Bmc Bioinformatics. 2010;11.
[56] Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J Comput Aid Mol Des. 2003;17:725-38.
[57] Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins. 2000;40:71-85.
[58] Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins. 1999;34:49-68.
[59] Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules. 1976;9:945-50.
[60] Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins. 2001;44:79-96.
[61] Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. P Natl Acad Sci USA. 1996;93:11628-33.
[62] Robson B, Osguthorpe DJ. Refined Models for Computer-Simulation of Protein Folding - Applications to the Study of Conserved Secondary Structure and Flexible Hinge Points during the Folding of Pancreatic Trypsin-Inhibitor. J Mol Biol. 1979;132:19-51.
[63] Godzik A, Kolinski A, Skolnick J. Are Proteins Ideal Mixtures of Amino-Acids - Analysis of Energy Parameter Sets. Protein Sci. 1995;4:2107-17.
[64] Bryant SH, Lawrence CE. An Empirical Energy Function for Threading Protein-Sequence through the Folding Motif. Proteins. 1993;16:92-112.
[65] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins. 1999;34:82-95.
[66] Bahar I, Kaplan M, Jernigan RL. Short-range conformational energies, secondary structure propensities, and recognition of correct sequence-structure matches. Proteins. 1997;29:292-308.
[67] Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc. 1996;118:11225-36.
[68] Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J Phys Chem B. 1998;102:10983-90.
[69] Rata IA, Li YH, Jakobsson E. Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops. J Phys Chem B. 2010;114:1859-69.
[70] Li YH, Liu H, Rata I, Jakobsson E. Building a Knowledge-Based Statistical Potential by Capturing High-Order Inter-residue Interactions and its Applications in Protein Secondary Structure Assessment. J Chem Inf Model. 2013;53:500-8.
[71] Yanover C, Vanetik N, Levitt M, Kolodny R, Keasar C. Redundancy-weighting for better inference of protein structural features. Bioinformatics. 2014;30:2295-301.