PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Detecting Local Residue Environment Similarity for Recognizing Near-Native Structure Models

Hyungrae Kim[1] and Daisuke Kihara[1,2]*

[1] Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47906

[2] Department of Computer Science, Purdue University, West Lafayette, Indiana 47907

ABSTRACT

We developed a new representation of local amino acid environments in protein structures called the Side-chain Depth Environment (SDE). An SDE defines a local structural environment of a residue considering the coordinates and the depth of amino acids that locate in the vicinity of the side-chain centroid of the residue. SDEs are general enough that similar SDEs are found in protein structures with globally different folds. Using SDEs, we developed a procedure called PRESCO (Protein Residue Environment SCOre) for selecting native or near-native models from a pool of computational models. The procedure searches similar residue environments observed in a query model against a set of representative native protein structures to quantify how native-like SDEs in the model are. When benchmarked on commonly used computational model datasets, our PRESCO compared favorably with the other existing scoring functions in selecting native and near-native models.

Proteins 2014; 82:3255–3272.

Key words: protein local structures; residue environment; residue depth; protein structure models; quality assessment; decoy selection.

INTRODUCTION

Structural similarities and commonalities at various levels have been found in protein tertiary structures. At a global structural level, there are abundant folds that arise from proteins with different evolutionary histories (superfolds).[1] At a smaller structural level, commonly occurring secondary structure arrangements have been found,[2] which are often called super-secondary structures.[3] Moreover, common fragment conformations were found in structures from proteins with different folds.[4–6] At the stereochemical structure level, interaction patterns of residues[7] and bond angles in main-chains[8] and side-chains[9] have been extensively studied. Observed structural patterns are not only important for understanding physical nature of the protein structures but are also practically useful as a source of information for validating protein crystal structures[10] as well as computationally predicted protein structure models.

In protein structure prediction, commonly occurring structure units (e.g., fragments) can be used as building blocks of protein structure models.[11,12] Alternatively, observed structural patterns can be represented as scoring functions (e.g., knowledge-based statistical potentials) that guide modeling procedures[11,13–15] and are also used for selecting the native structure or near-native structure models (often called decoys) from a pool of alternative models.[16,17] For example, pairwise residue contact potentials, which are derived from the statistics of physically contacting amino acid residues observed in representative protein structures, are one of the most frequently used scores for structure prediction.[7,18,19] In addition to atom/residue contact potentials, various types of knowledge-based scores have been developed, which capture residues' or atoms' propensities of angles,[20,21] accessible surface area,[22] and number of contacts,[23] to name a few. Tasks of knowledge-based scores are, in

Grant sponsor: National Institute of General Medical Sciences of the National Institutes of Health; Grant number: R01GM097528; Grant sponsor: National Science Foundation; Grant numbers: IIS0915801, DBI1262189, IOS1127027; Grant sponsor: National Research Foundation of Korea Grant funded by the Korean Government; Grant number: NRF-2011-220-C00004.
*Correspondence to: D. Kihara, Department of Biological Sciences, Purdue University, West Lafayette IN 47906. E-mail: dkihara@purdue.edu
Received 15 January 2014; Revised 10 June 2014; Accepted 21 July 2014
Published online 31 July 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24658

principle, twofold: one is to build and select "protein-like" models, i.e., models that have geometrical features that agree with those in known structures. Another purpose is to construct sequence-specific structures by capturing local and global sequence-structure corresponding patterns (the extreme is homology modeling). It is ideal if a scoring function performs well for these two purposes simultaneously, but it is not an easy task. There are instances that the lowest energy states led by physics-based potentials are not close to the native structures of proteins.[24]

In order to generate protein-like models that are also native structures, a scoring function must capture sequence-specific structural patterns while remaining sufficiently general to apply to multiple proteins. This can be achieved by describing structures at a level somewhere between the residue level and the whole protein level. Such structure representation should consider residues in their structural environment, which includes interactions with local and distant residues.

The concept of residue environment has been studied both computationally and experimentally over a decade. Manavalan and Ponnuswamy observed that surrounding residues for a certain residue have a biased distribution that reflects cooperativity of the residue pairs.[25,26] Karlin et al. investigated the atom density, which was defined as the count of atoms within a certain distance from each residue, and found that the densities differ depending on the residue at the center.[27] It was shown that the secondary structure prediction accuracy for residues with a high residue-wise contact order is worse than average, suggesting that distant contacts affect secondary structure formation.[28] Experiments have also demonstrated that the environment affects the secondary structure of the same amino acid sequences.[29,30] Following these observations, several groups developed scoring functions that describe residue environments or multi-residue interactions for protein structure predictions and structure-based function predictions. Along this line, knowledge-based contact potentials that consider interactions between three or four residues were developed.[31–33] DeGrado and his colleagues developed an atom "microenvironment" potential, which consider the number of different types of atoms within a certain radius of a center atom.[34] The Levitt group developed a hydrophobicity score that considers hydrophobic residue interactions within a sphere of 10 Å.[35] Simons et al. defined a residue environment as the number of residues within a 10 Å sphere and used it as a part of definitions of the scoring function for their *ab initio* protein structure prediction method.[36] Mooney et al. used a residue environment representation that captures atoms within concentric spheres around a Cβ atom of a residue to recognize functional sites of proteins.[37]

In this work, we developed a new representation of local amino acid environments in protein structures called the Side-chain Depth Environment (SDE). The SDE considers three important structural features of a residue environment: the side-chain centroids of a chain fragment centered on the target residue, the number of surrounding residues within a sphere around the target residue, and the depth of these surrounding residues. The residue depth quantifies location of a residue relative to the protein surface.[38] We chose the residue depth as one of the structural features because it was shown to be an effective scoring term in a fold recognition method.[22] First, we show that SDEs capture the characteristic environment of each amino acid. Subsequently, we show that SDEs are general enough that similar SDEs are found in protein structures with globally different folds. Because similar SDEs are found in native protein structures of different folds, they can be applied for detecting near-native models from a pool of decoys. Given a structure model to be evaluated, the more residues of the model in SDEs similar to those in known native structures, the more likely the model is close to the native structure.

Using the SDEs, we developed a scoring procedure for selecting native or near-native models from a pool of decoys (decoy selection) named PRESCO (Protein Residue Environment SCOre). The protocol searches similar SDEs and a main-chain environment named the Main-chain Residue Environment (MRE) observed in a query model against a set of representative native protein structures. We benchmarked the ability of native and near-native model selection by PRESCO on decoy datasets that are commonly used for examining scoring functions for protein models. We show that our procedure compared favorably with the other existing scoring functions by significant margins.

MATERIALS AND METHODS

Database of reference protein structures

We used a nonredundant protein dataset for investigating local structure similarity of proteins. The dataset was also used as a reference database for assessing protein structure models (decoys) using the residue environment scores. Totally, 4803 nonredundant protein structures with a resolution better than 2.0 Å and a pairwise sequence identity of less than 30% between each other were downloaded from the PISCES server (Oct. 17, 2008 version).[39] This set was further reduced to 2536 proteins by removing chains with missing residues or missing backbone atoms.

Decoy sets for native/near-native structure recognition tests

To test how well the new residue environment scores perform in discriminating native or near-native models from other models with poorer quality, we used four

decoy sets, the Decoy "R" Us set,[40] the Moulder decoy set,[41] the I-TASSER decoy set,[42] and the Rosetta decoy set.[43] In addition, we also tested the residue environmental scores on the Fiser's CASP model set.[44] Each set consists of subsets with a native protein structure associated with decoy structures. The Decoy "R" Us set include eight sets, 4state_reduced, Fisa, Fisa_casp3, Lmds, Lattice_ssfit, hg_structal, ig_structal, ig_structal_hires that consists of 144 subsets in total, each of which contains the native structure of a protein and on average 271.19 decoys of the protein. The Moulder set has 20 subsets, each of which contains 319.3 decoys on average. The I-TASSER set consists of 56 subsets, which contain on average 438.2 decoys. The Rosetta sets has 58 subsets with 100 decoys. The Fiser set consists of 143 proteins used as prediction targets in the CASP5 to CASP8 and their prediction models. On average there are 18.3 models per target.

Removal of homologous proteins from the representative structure database

When the environment score was computed for a structure model, all database proteins that have more than 25% sequence identity with the target protein are excluded from the database. Clustal Omega[45] was used for computing the sequence identity.

Local structural environment of residues, SDE

Characteristic physicochemical properties of each amino acid should reflect the structural environments of amino acids in native protein structures. To describe the environment of a residue, we wanted to consider the main-chain conformation around the residue of interest, the number of surrounding residues and their relative positions, and the depth of the surrounding residues from the protein surface. These features altogether capture multi-body interactions of residues in a comprehensive manner.

The residue depth is defined as the distance between a given residue and the nearest water molecule located close to the solvent-accessible surface.[38] We used the side-chain to represent a residue position because the residue specificity originates from side-chains and the side-chain packing is a dominant factor of protein folding.[15,46]

For a given residue in a protein structure, residues with a similar structural environment, SDE, in different proteins were identified by the following procedure:

1. Side-chain centroids of the target residue and structures in a reference database are computed. The side-chain centroid of a residue is the average positions of all heavy atoms in the side-chain. Atoms in the backbone are not included.
2. To ensure that residues locate in similar main-chain conformations, side-chains along local fragments are compared. Nine consecutive side-chain centroids along the protein backbone (four residues before and after the query residue) were compared with structures in the database and the 500 most similar fragments having the lowest root mean square deviation (RMSD) with the query fragments are kept for the subsequent steps. The rotation matrix and the translation vector used for aligning the two fragments are stored.
3. Next, neighboring residues within a sphere of an 8.0 Å radius centering on the side-chain centroid of the query residue are compared with those for the 500 fragments stored in Step 2. Fragments are discarded if the numbers of neighboring side-chain centroids in the two spheres are different. 8.0 Å was chosen because an early work showed a residue's cooperativity is effective up to 8.0 Å.[26]
4. For each of the remaining fragments, one-to-one correspondence of neighboring side-chain centroids in the sphere to those in the query residue is computed. This is done by superimposing the neighboring centroids using the stored rotation matrix in Step 2 and pairing side-chain centroids of the two spheres by their distances.
5. Finally, the root mean square distance of the residue depth (residue-depth RMSD) of corresponding neighboring side-chain centroids is computed.

Thus, similarity in residue environment indicates that the residues share similar main-chain conformation and the same number of neighboring residues that are located at similar depth in the protein structures (Fig. 1).

RESULTS

Characterization of SDEs in representative native protein structures

To begin with, we examined SDEs of amino acid residues in representative native protein structures. Figure 2 shows the distribution of the number of side-chain centroids within the sphere of 8.0 Å for each amino acid residues. Consistent with previous works,[47] the number of neighboring residues clearly reflects hydrophobicity of amino acids. Isoleucine and valine, the two most hydrophobic amino acids according to the Kyte-Doolittle hydrophobicity scale,[48] show the highest median, followed by leucine, phenylalanine, and alanine, which are also highly hydrophobic amino acids [Fig. 2(A)]. In contrast, hydrophilic amino acids have the least number of neighboring residues; these include arginine, lysine, asparagine, aspartic acid, glutamine, and glutamic acid. The average number of neighboring residues of each amino acid has a significant Pearson's correlation coefficient of 0.793 to the Kyte-Doolittle hydrophobicity scale.

In Figure 2(B), we examined residues that have similar SDEs for each residue type. The color scale indicates

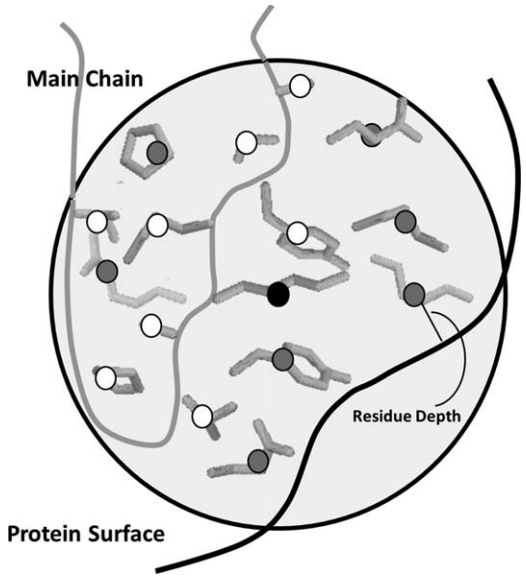

Figure 1

Side-chain depth environment (SDE). The SDE of an amino acid (black circle) is defined as the depth of the side-chain centroids within a sphere of 6.0 or 8.0 Å (gray and white circles) from the center amino acids. To find similar SDEs from a database, the RMSD of nine side-chains (white circles) including the center one, the number of side-chain centroids in the sphere, and the residue depth of the side-chains are considered.

enrichment of the amino acid type among residues with similar SDEs. Concretely, the enrichment is computed as the fraction of each amino acid among the top 40 most similar residues divided by the background fraction of the amino acid. In all the amino acids except for glutamine, the identical amino acids (shown in the diagonal positions in the heat map) have the largest enrichment. In the case of glutamine, glutamic acid came to the top with a ratio of 1.88 and the glutamine itself was the close second with 1.83. Amino acids that are different but have similar physicochemical properties to the query amino acid tend to have an enrichment ratio over 1.0. Such examples include arginine, which had lysine with the second largest enrichment (2.44) and isoleucine, which had valine with the second largest enrichment. The highest enrichment, 6.88, was observed for glycine with itself. Proline also showed a high enrichment of 5.01 with itself. To summarize, Figure 2(A) shows that the number of side-chains within a sphere of 8.0 Å radius, which is used as a filtering step for finding similar SDEs, mainly contains the hydrophobicity information of residues whereas the residue depth [Figure 2(B)] further encodes residue-specific environment features. These results suggested that the SDE can be used for designing a scoring function for quantifying native-likeness of computational protein structure models.

Similar SDEs are found in proteins with globally very different structures (Fig. 3). Figure 3(A) shows the distribution of the residue depth RMSD of surrounding side-chain centroids of SDEs in comparison with the global RMSD of protein structures where the two SDEs were taken from. For a SDE, the depth RMSD and the global RMSD of protein structures for the top 5 most similar

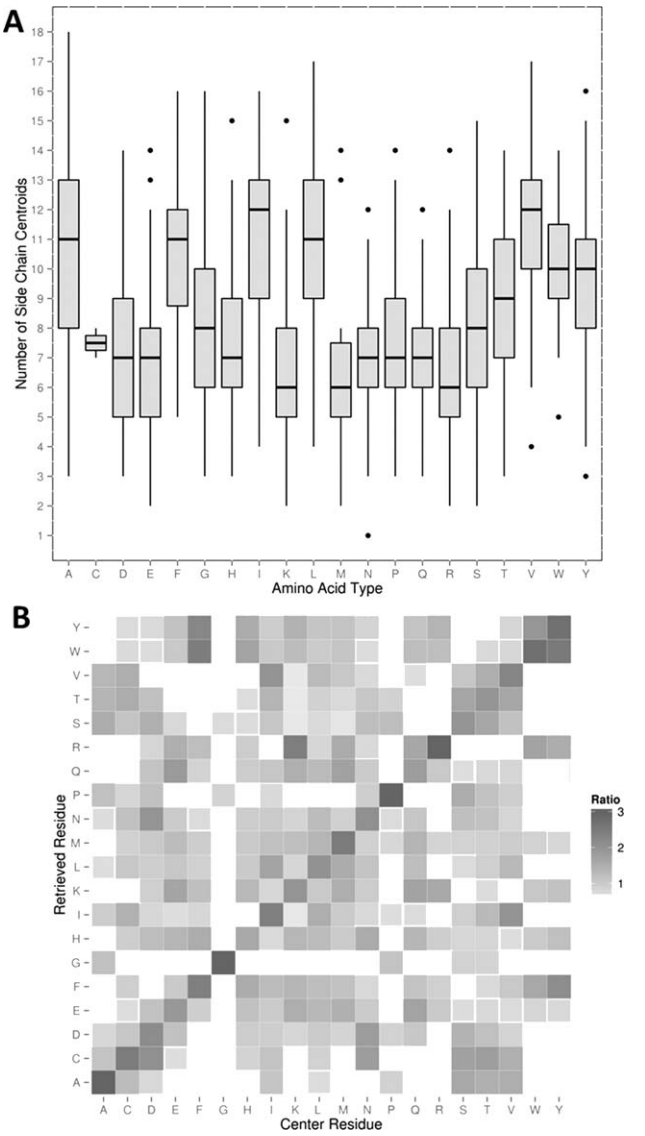

Figure 2

Characteristics of SDEs. **A**: The distribution of the number of side-chain centroids in the sphere of an 8.0 Å radius for each residue type shown in a box-and-whisker plot. The bar inside the box shows the median and the two ends of the box show the first and third quartiles. Outliers (shown as dots) are defined as more than 1.5 times the inter-quartile range (the third quartile minus the first quartile) outside the first or third quartiles. **B**: Residues with similar SDEs identified by database searches for each residue type. For the SDE of each residue in each protein in the reference database, the top 40 most similar SDEs in terms of the residue depth RMSD were retrieved. Then fraction of 20 different residues retrieved for each residue type was computed and normalized by the overall fraction of the residue in the reference database. The color scale shows the enrichment, the darker the higher ratio of the residue relative to the background fraction.

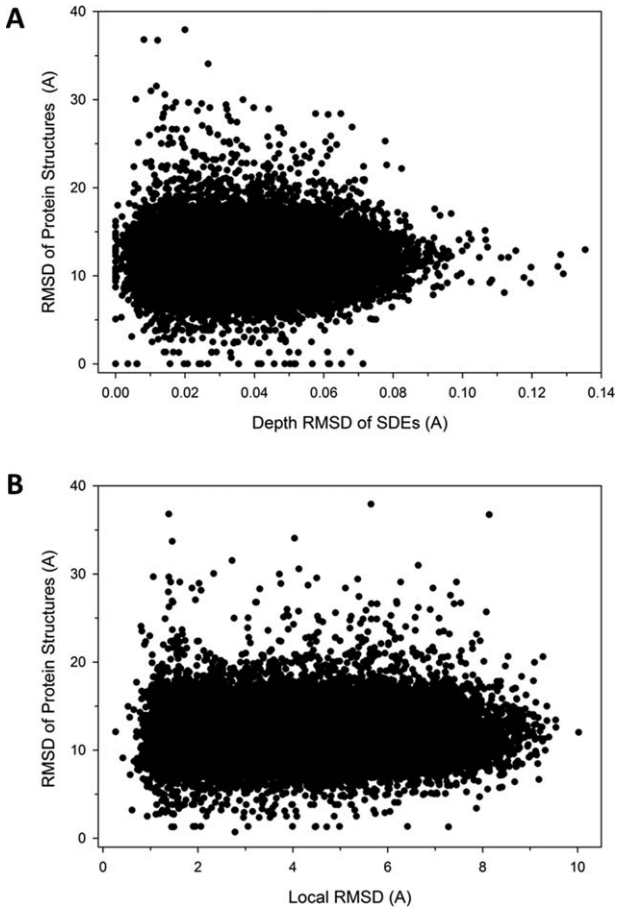

Figure 3

Global structural similarity of proteins that have similar side-chain environments. **A**: Global coordinate RMSD of protein structures whose residue has similar SDEs. For each residue in the reference protein database, five most similar residues to the query residue in terms of SDEs were selected. Their depth RMSD of the side-chains in a sphere of 8.0 Å and the global RMSD of the whole protein structures were computed. Global RMSD was computed with BioShell, which computes RMSD of gapped structure alignment between two proteins. **B**: For the same residue pairs shown in the Plot A, the conventional RMSD of the neighboring side-chain centroids were compared with the global RMSD of the proteins.

SDEs were plotted. Global RMSDs were computed by Bioshell.[49] It is shown that the proteins with similar SDEs have very different global structure (on average around 13 Å) and the depth RSMD and the global RMSD have virtually no correlation. The same conclusion was drawn when we compared the conventional RMSD of side-chain centroids in SDEs against the global RMSD of the protein structures [Fig. 3(B)]. Thus, not only the depth RMSD but also the constellation of side-chains in SDEs does not have correlation to the global RMSD of the protein structures.

In Figure 4 we show examples of similar SDEs found in proteins with different folds. In all cases the pairs have a small depth RMSD, ranging from 0.022 to 0.037 Å. In the first example [Fig. 4(A)], SDEs taken from helices of two different fold class (left: 1iib, an αβ class protein; right: 1ykh, α class) are shown. They have seven side-chains in the SDEs and their depth RMSD is 0.022 Å. In the second example [Fig. 4(B)], similar SDEs are taken from globally different folds. Six residues are included in the SDEs, which are located at an end of a helix. The next SDE pair [Fig. 4(C)] includes 13 residues that are distant in sequences. The SDE in 1e6i consists of residues in three α-helices while the one from 1bqcA have residues from two α-helices and one β-strand. In the last pair [Fig. 4(D)], two SDEs are taken from different secondary structure combinations: on the left, side-chains are taken from β-sheets and two α helices in 1ew4, while the SDE on the right (2bj0A) consist of residues from two β-sheets.

Procedure of selecting native/near-native models

In this section, we describe the procedure to evaluate decoys and to select ones from a pool of decoys that are likely to be the native or close to the native structures. Briefly, in the scoring procedure named PRESCO (Protein Residue Environment SCOre) a count will be accumulated for a query model if residues in the model have similar environments to those of similar amino acid type in representative native structures. In addition to the SDE, we introduce another environment score, the Main-chain Residue Environment (MRE), which considers main-chain fragment conformation around the target residue. In the decoy selection procedure, similarity of both SDEs and MREs of target residues to representative proteins are considered. Before providing the steps of computing PRESCO, we explain the MRE.

The MRE compares the main-chain conformation of the fragment of five residues centered on the target residue against fragments in the database of representative proteins. The similarity of two fragments is quantified by the RMSD of four main-chain heavy atom positions (N, Cα, C, and O) and the Cβ position of all the residues in a fragment. This representation was shown to perform better than a Cα representation of fragments in identifying amino acid patterns that fold into the fragment conformation.[50] For glycine, a pseudo Cβ atom position is computed using the TINKER package.[51]

Figure 5 gives the overall PRESCO procedure. For each of the residues in a query model, MRE and two SDEs of different radii, 8.0 Å and 6.0 Å, are computed. They are then compared with those for residues in a structure database, as represented in the three branches in the flowchart. The left most branch explains steps to compute a score that comes from comparison of MREs of the target model to those in the representative protein structure database. As described in the Method section, the database consists of 2536 non-redundant native

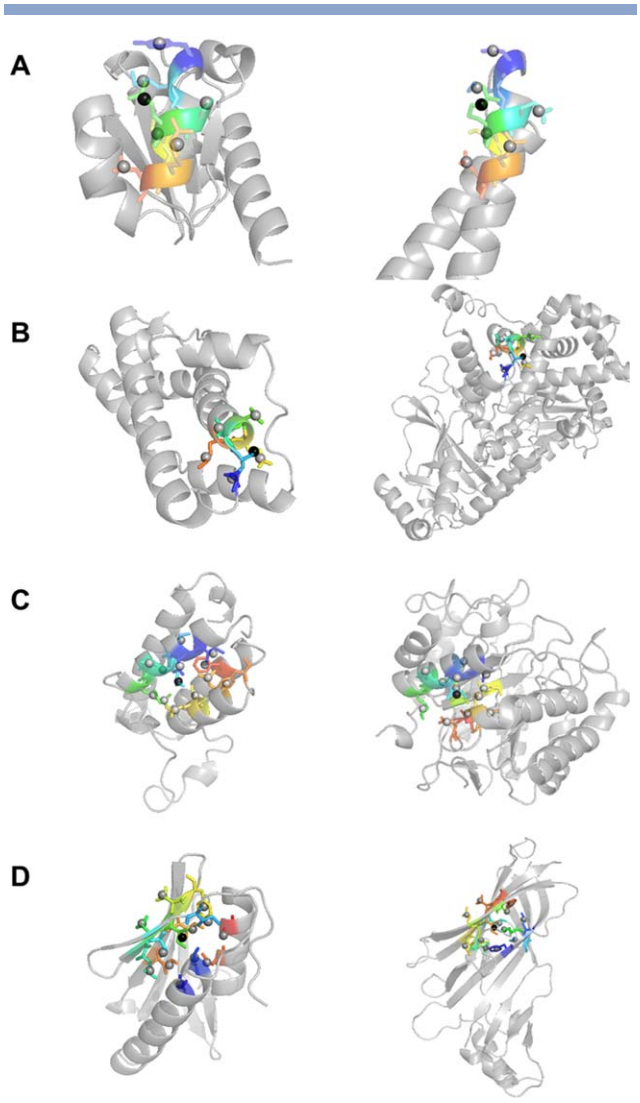

Figure 4

Examples of similar SDEs. Residues included in SDEs are shown in color. **A**: SDE of residue 64 (Glu) of IIBcellobiose (PDB code: 1iib) (left) and residue 146 (Arg) of RNA polymerase II mediator complex protein MED7 (1ykh, chain A) (right). The center side-chain is shown in black spheres. 8.0 Å was used for the sphere to define SDEs. Both SDEs contain seven side-chain centroid points of neighboring residues (shown in the stick representation). The residues are 60, 61, 63, 64, 65, 67, and 68 for 1iib and 32, 33, 35, 36, 37, 39, and 40 from 1ykhA. The depth RMSD (dRMSD) of the two SDEs is 0.022 Å. **B**: Residue 40 (Ser) in fertilization protein (1lis) and residue 312 (Ser) of Anthrax lethal factor (1yqyA). Six residues are included: residue 39, 40, 41, 42, 43, 44 for 1lis and 311, 312, 313, 314, 315, 316 from 1yqyA. dRMSD: 0.023 Å. **C**: SDEs of residue 12 (Leu) in bromodomain of GCN5 (PDB: 1e6i) and residue 71 (Ile) of β-mannose (1bqcA). 13 residues are in the SDEs, 8, 9, 11, 12, 13, 15, 16, 46, 49, 50, 53, 58, and 64 for 1e6i and 67, 68, 70, 71, 72, 74, 75, 81, 83, 113, 114, 117, and 121 for 1bqcA. The depth RMSD is 0.037 Å. **D**: Residue 41 (Ile) of map kinase 14 (1ew4) and residue 196 (Ile) of tetracycline repressor (2bj0A). 13 residues are included. 1ew4: 39, 40, 41, 42, 43, 17, 21, 30, 32, 49, 51, 91, 95; and 2bj0A: 194, 195, 196, 197, 198, 31, 33, 50, 137, 139, 161, 175, 177. dRMSD: 0.035 Å.

protein structures. For a MRE computed for a residue in the query model, the 25 closest (i.e. lowest RMSD) MREs in the database are retrieved and sorted by their RMSD. This process is done for all the fragments in the model (*L-4* fragments, where *L* is the length of the model). Because fragments are taken by a sliding window that moves by one residue at a time, a residue in the middle of the protein will be covered by 25 x 5 = 125 fragments, while residues close to the terminus of the chain have fewer fragments. The MRE-based score given to an amino acid position in a model is the weighted average of the amino acid similarity score between the amino acid in the query fragment and each of retrieved fragments. Amino acid similarity score is defined by an amino acid similarity matrix chosen for MRE. A weight is assigned to each retrieved fragment, which reflects similarity of the fragment to the query fragment. The weights and an amino acid similarity matrix used were determined by a small benchmark study described in the next section. Taken together, a MRE-based score of a query model is computed as

$$\text{MRE\_based\_Score} = \sum_{i=1}^{L}\sum_{j=1}^{K} w_j M_{a_i - a_j(i)}, \qquad (1)$$

where *L* is the length of the model, *K* is the number of fragments that cover the residue *i*. *K*=125 if the position *i* is between 5 to *L-4*, while *K* reduces to *25*i* if $i < 5$ (N-terminus) and *25*(L-i+1)* if $i > L\text{-}4$ (C-terminus). w_j is the weight given to the *j*-th fragment and $M_{ai\text{-}aj(i)}$ is the amino acid similarity score taken from the matrix *M* for the amino acid at the position *i* in the query model and the amino acid in the fragment *j* that is aligned with amino acid *i* of the query.

The second and the third branches represent computation of scores by comparing SDEs in the query models to the representative structure database. For each residue in the query model, two SDEs are constructed with radii of 6.0 Å and 8.0 Å. It was found that the score with the 8.0 Å radius gave more correct native structure selections than 6.0 Å but 6.0 Å worked better for a certain smaller fraction of the cases (data not shown). Then, following the procedure to find similar SDEs as described above, for each residue the 40 most similar (i.e., smallest depth RMSD) SDEs to the query SDE are selected. Then, similar to the branch for MRE, a score is computed for the query residue, which is the weighted sum of the amino acid similarity values between the query residue and residues retrieved in the database search. The score for a query model is the sum of the residue-based score:

$$\text{SDE\_based\_Score} = \sum_{i=1}^{L}\sum_{j=1}^{N} w_j S_{a_i - a_j}, \qquad (2)$$

where $N = 40$, the number of SDEs retrieved from the database and $S_{ai\text{-}aj}$ is the amino acid similarity score taken from a matrix *S* for residue *i* in the query and residue *j* retrieved from the database. The SDE-based scores

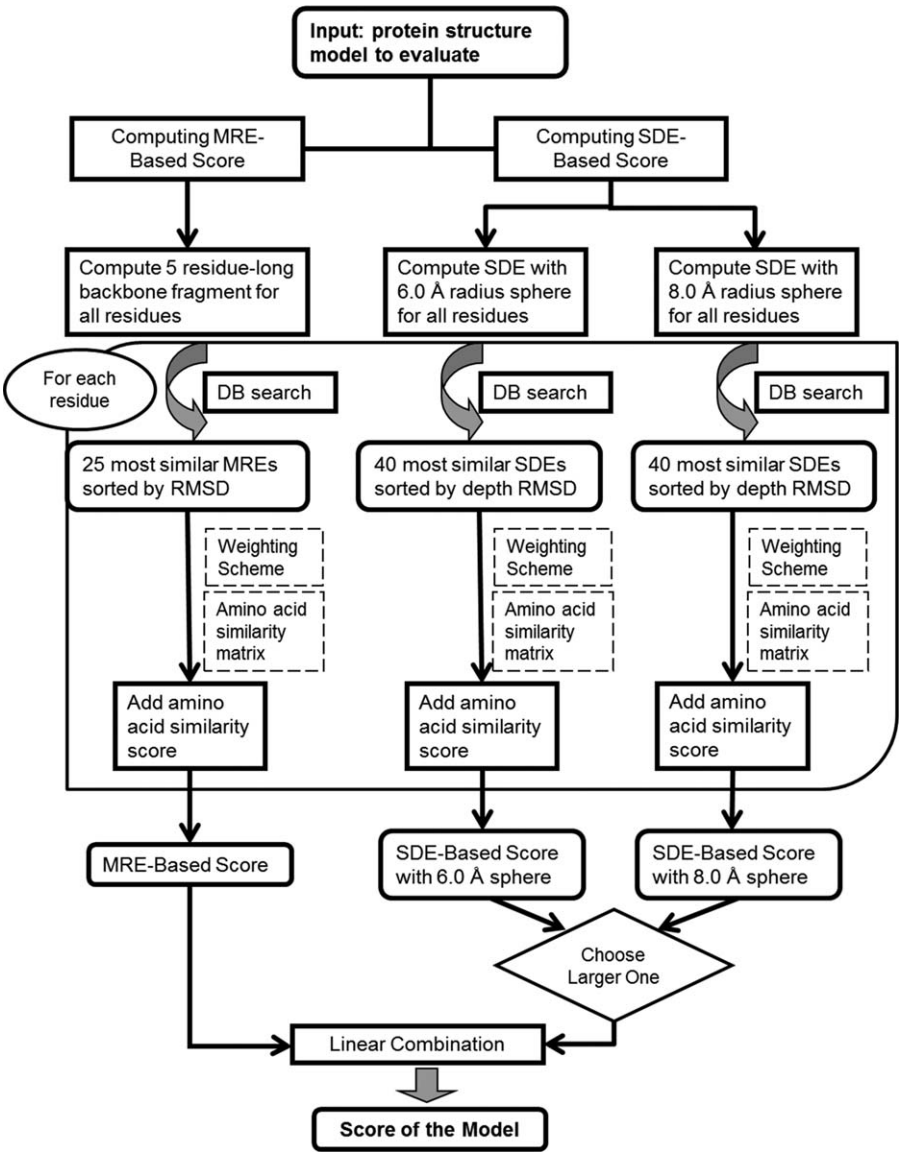

Figure 5

Schematic diagram of the PRESCO scoring system for evaluating decoys**.** For each residue in a decoy to be evaluated, two residue environments, MRE and SDE, are constructed and compared against residues in the database of representative proteins. Two sphere sizes, 8.0 Å and 6.0 Å, are used for SDE. Similar MRE/SDEs found in the database are sorted according to their similarity to that of the target residue. A score for a target residue is a weighted sum of the amino acid similarity score, and the score of the decoy is the sum of the score given to each residue. See text for more details.

for a query model is computed with the 6.0 Å (the middle branch in Fig. 5) and the 8.0 Å (the right branch) radii separately, which are subsequently compared and a larger one is chosen. Finally, the MRE- and the SDE-based scores will be linearly combined to yield the final score of the query model. The weighting scheme for the linear combination is discussed in the next section. The two scores based on MRE and SDE complement each other. The MRE assesses how native the structure is in terms of local main-chain fragment similarity, while the SDE evaluates the structure from the view point of side-chain packing.

Weighting schemes and amino acid similarity matrices

The combinations of weights and amino acid similarity matrices used in the PRESCO model evaluation procedure (Fig. 5) were determined based on the native structure selection test performed on 30 decoy target

sets, which were randomly chosen from the Rosetta decoy set.[43] A decoy target set in the Rosetta set consists of one native structure and 100 decoy structures with varied RMSDs to the native structure of a protein. Using different weighting schemes and amino acid similarity matrices in Eqns. (1) and (2), we examined how many times the native structure was selected with the highest score among the 30 decoy sets.

Three weighting schemes were tested:

1. RMSD-based weights: the weight w_i to a retrieved i-th MRE or SDE is computed based on the RMSD value of the main-chain fragment to the query environment for MRE and side-chain centroids in the sphere for SDE. It is inversely proportional to the power α of the RMSD.

$$w_i = 1/(\mathrm{RMSD})^{\alpha} \quad (3)$$

Thus, retrieved MRE/SDE with a small RMSD has a larger weight. 11 values, 0.005, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.8, 0.9, 1.0, and 2.0, were tested for α.

2. Rank-based weights: MREs/SDEs retrieved from the database are ranked by their RMSD (MREs) or the depth RMSD (SDEs) to the query. Then, each MRE/SDE is weighted according to its rank:

$$w_i = 1 \Big/ \{\lfloor n_i/\beta \rfloor + 1\}, \quad (4)$$

where n_i is the rank of the MRE/SDE. For MRE, n_i will range between 1 to 25 while it ranges between 1 to 40 for SDE. $\lfloor \cdot \rfloor$ is the floor function which takes the largest integer that does not exceed the inside value, and β is a parameter. Four values, 1, 2, 5, 10, and 15 were tested for β.

3. Exponential weights: The weight decays as the rank of MRE/SDE decreases:

$$w_i = \exp(-\gamma^*(n_i - 1)/(N - 1)), \quad (5)$$

where γ is the parameter to be optimized, n_i is the rank of MRE/SDE, N is the total number of retrieved environments for a query, i.e., 25 for MRE and 40 for SDE. Six values between 0.5 and 2.5 were tested for γ: 0.5, 0.7, 0.8, 0.9, 1, 1.5, and 2.5.

Seven amino acid similarity matrices were tested. These matrices are labeled as BLOSUM30,[52] QU_C1 (QU_C930101),[53] QU_C2 (QU_C930102),[53] QUIB (QUIB020101),[54] KOLA (KOLA920101),[55] CCPC,[56] and CC80.[56] Shown in parentheses are the ID in the AAIndex database[57] if the matrix is indexed. Except for BLOSUM30, which is a standard matrix for sequence alignment, the other six matrices were chosen because they performed well in computing aligning distantly related protein sequences.[56] These matrices capture amino acids' preference of structural contexts in protein tertiary structures. QU_C1 and QU_C2 capture amino acid residue contact propensities. KOLA is based on the similarity of the dihedral angles of amino acids. QUIB was numerically optimized to minimize the average RMSD of aligned proteins in benchmark databases. CCPC is based on the correlation coefficients of an amino acid residue contact potential, while CC80 is a linear combination of CCPC and KOLA.

Selecting weight and matrix combinations on the 30 decoy sets

First, we benchmarked the combinations of the weights and matrices on the 30 Rosetta decoy sets in terms of their native structure recognition ability. Figure 6(A,B) show the results of using the MRE and the SDE, respectively. BLOSUM30 showed higher success rates for the MRE than for the SDE with many weight combinations. CCPC is another matrix that performed better for the MRE than the SDE particularly with the exponential weights [Eq. (5)]. On the other hand, QU_C1, QU_C2, and KOLA performed better for the SDE than the MRE. Among the weights tested, RMSD-based weights (α) of 0.5 or larger consistently performed poorly while large exponential weights (γ) of 1.0 or larger gave a high native recognition rate for several matrices for both MRE and SDE. The highest number of successful native recognition was 17, which was achieved by CCPC when applied to the MRE as well as QU_C2 and QUIB applied to the SDE.

From these results, we have chosen all the weight and matrix combinations that gives over 15 native recognitions from both MRE [Fig. 6(A)] and SDE [Fig. 6(B)], and further tested linear combinations of the MRE and the SDE in the native recognition:

$$Combined\_Score = MRE_i + w^* SDE_j, \quad (6)$$

where MRE_i and SDE_j are one of the selected weight-matrix combinations from MRE and SDE. w is a weight value that ranged from 0.05 to 10.0 with an interval of 0.05.

Among the linear combinations tested, combinations of BLOSUM30 (exponential weight, $\gamma = 1.5$) for the MRE and QU_C2 ($\gamma = 1.0$) for the SDE with weight values between 1.15 to 1.45 or 1.7 to 2.05 gave the largest number of native recognition of 24. We found that all the Combined Scores that recognized native structures 22 or more consisted of a limited variety of matrices. BLOSUM30 (ranked-based weight with $\beta = 10$ or exponential weight with $\gamma = 1.5$), CC80 (RMSD-based weight with $\alpha = 0.01$), and QUIB (exponential weight with $\gamma = 1.5$) were selected for the MREs, while the following eight were selected for the SDEs: CC80 (RMSD-based weight with $\alpha = 0.01$), CCPC (RMSD-based weight with

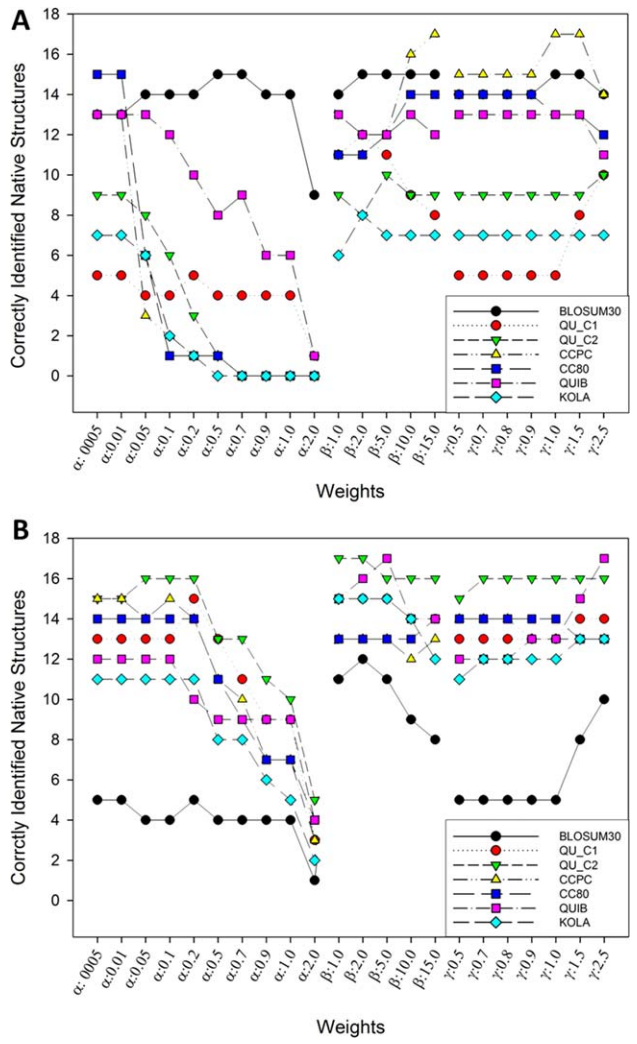

Figure 6

The number of successfully identified native structures in 30 Rosetta decoy sets. Seven matrices were tested in combinations with 22 weights for native structure recognition. α is a parameter for the RMSD-based weight [Eq. (3)]; β is a parameter for the rank-based weight [Eq. (4)]; and γ is a parameter for the exponential weight [Eq. (5)]. **A:** MRE; **B:** SDE.

$\alpha = 0.01$), QU_C1 (ranked-based weight with $\beta = 10$), QU_C2 (RMSD-based weight with $\alpha = 0.01$; ranked-based weight with $\beta = 1.0$ or 2.0; and exponential weight with $\gamma = 1.0$ or 1.5). Therefore, we only used these matrix and weight combinations for MREs and SDEs in the subsequent analyses.

Native structure recognition test

Using the selected MREs and SDEs, we tested the native recognition ability of the scores on a larger dataset. The dataset consists of five previously published decoy sets (see Materials and Methods). This is a standard dataset for testing potentials and scoring functions, which has been used by several recent papers.[33,42,58,59] We tested our scores in two ways. First, we concentrated on the ability to select native structures from decoys as done in a previous work.[42] Then we examined RMSD and TM-score[60] of selected decoys when native structures are not included in the dataset, as this is more close to realistic scenarios in protein structure prediction. As described in Materials and Methods, for each query protein, all similar protein structures were removed from the database if they have more than 25% sequence identity to the query. The performance of the native/decoy selection by our scores was compared with seven other knowledge-based statistical potentials, DFIRE,[61] dDFIRE,[62] DOPE,[63] RW, RWplus,[42] OPUS-PSP,[58] and GOAP.[59] For our residue environment scores, the best performing MRE and SDE in terms of the total number of native structure recognition, CC80 and QUIB, respectively, as well as the three best linear combinations of MRE and SDE are shown. The results are summarized in Table I.

The performance of our residue environment-based scores, particularly the combinations of MRE & SDE, was better than the other potentials in terms of the total number of correctly recognized native structures. Among the 278 decoy targets in total, the Combined Scores successfully recognized native structure for 255 cases. The MRE (CC80) performed better than the SDE and the three Combined Scores gave further improvement, which is consistent with what we observed for the 30 Rosetta decoy set. Noticeable improvement relative to the other scores was made in the native selection for the datasets of Lmds, Moulder, I-TASSER, ig_structal, and ig_structal_hires. For these five sets, our residue-environment scores recognized native structures perfectly; all 10 natives in the Lmds, 20 natives in the Moulder, 56 natives in I-TASSER, 61 natives in ig_structal and 20 natives in ig_structal_hires decoy sets. MRE also successfully identified 28 natives in hg_structal decoy set. Two of our Combined Scores performed also well for the ROSETTA set, recognizing 41 out of 58, second to GOAP.

The performance by MRE on the hg_structal, ig_structal and ig_structal_hires sets (109 native recognitions) are noticeable, which is more than 25% increase from that of the previous best performing potential, GOAP (87 native recognitions). These three decoy sets are products of high accurate homology modeling.[40] The average RMSD of the decoy models to its native are 2.38 Å (ig_structal), 2.55 Å (ig_structal_hires), and 4.10 Å (hg_structal). In Table I, it was observed that the accurate native recognition mainly come from MRE. By a close investigation, interestingly, MREs performance was found to be nearly independent of the choice of substitution matrix and weighting schemes. (CCPC(β =10): 107, blosum30(γ =1.5): 110, CC80(α =0.01): 109, QU_C1(β =2.0): 109, QU_C2(γ =1.5): 109, QUIB(γ =0.9): 110 natives were successfully recognized by MRE).

Table I
Performance of Native Structure Recognition

											Combinations of MRE and SDE[d]			
Decoy sets	DFIRE[a]	dDFIRE	DOPE	RW		RWplus	OPUS-PSP	GOAP	MRE (CC80)[b]	SDE (QUIB)[c]	BLSM30+ QU_C2	BLSM30+ QU_C2	CC80+ QU_C1	# Targets [e)]
4state_ reduced	6 (−3.48)[f)]	**7** (−4.15)	**7** (−3.66)	6 (−3.45)	6 (−3.51)	**7** (−4.49)	**7** (−4.38)	**7** (13.60)	**7** (3.68)	**7** (8.88)	**7** (9.07)	**7** (8.50)	7	
Fisa	**3** (−4.87)	**3** (−3.80)	**3** (−3.91)	**3** (−4.87)	**3** (−4.79)	**3** (−4.24)	**3** (−3.97)	2 (1.16)	2 (4.13)	2 (3.46)	2 (3.44)	**3** (3.51)	4	
Fisa_casp3	4 (−4.80)	4 (−4.83)	3 (−5.06)	4 (−5.22)	4 (−5.17)	**5** (−6.33)	**5** (−5.27)	2 (3.67)	1 (3.10)	3 (4.41)	3 (4.45)	4 (4.03)	5	
Lmds	7 (−0.88)	6 (−2.44)	7 (−1.34)	7 (−1.20)	7 (−1.03)	8 (−5.63)	7 (−4.07)	**10** (8.20)	6 (4.25)	**10** (8.86)	**10** (9.04)	**10** (7.70)	10	
Lattice_ssfit	**8** (−9.44)	**8** (−10.12)	**8** (−7.43)	**8** (−8.15)	**8** (−8.85)	**8** (−6.75)	**8** (−8.38)	**8** (7.38)	**8** (7.77)	**8** (8.66)	**8** (8.69)	**8** (7.09)	8	
hg_structal	12 (−1.97)	16 (−1.33)	—[g)]	—	12 (−1.74)	18 (1.87)	22 (−2.73)	**28** (5.14)	11 (1.27)	27 (3.82)	27 (3.89)	27 (4.13)	29	
ig_structal	0 (0.92)	26 (−1.02)	—	—	0 (1.11)	20 (0.69)	47 (−1.62)	**61** (7.69)	6 (0.28)	**61** (6.93)	**61** (6.97)	**61** (7.11)	61	
ig_structal_ hires	0 (0.17)	16 (−2.05)	—	—	0 (0.32)	14 (−0.77)	18 (−2.35)	**20** (4.35)	6 (0.09)	**20** (4.17)	**20** (4.18)	**20** (4.21)	20	
Moulder	19 (−2.97)	18 (−2.74)	19 (−3.09)	19 (−2.79)	19 (−2.84)	19 (−4.84)	19 (−3.58)	**20** (13.58)	16 (2.87)	**20** (6.73)	**20** (6.90)	**20** (7.20)	20	
ROSETTA	20 (−1.82)	12 (−0.83)	21 (−1.61)	20 (−1.62)	20 (−1.47)	39 (−3.00)	**45** (−3.70)	25 (1.90)	31 (2.20)	41 (2.91)	41 (2.91)	39 (2.74)	58	
I-TASSER	49 (−4.02)	48 (−5.03)	30 (−2.18)	53 (−4.42)	**56** (−5.77)	55 (−7.43)	45 (−5.36)	**56** (9.61)	47 (3.43)	**56** (7.30)	**56** (7.38)	**56** (7.20)	56	
#Total (*Z*-score)	128 (−1.94)	164 (−2.52)	98/168 (−2.47)	120/168 (−3.23)	135 (−2.13)	196 (−2.86)	226 (−3.57)	239 (**6.78**)	141 (2.14)	**255** (5.70)	**255** (5.76)	**255** (5.65)	278	

The largest values for each dataset are shown in bold.
[a]Values for DFIRE to GOAP were taken either from the paper by Zhou and Skolnick (2011) or by Zhang and Zhang (2010).
Results of MRE using CC80 (RMSD-based weight, $\alpha = 0.01$), which showed the largest number of successful native recognition at #Total among preselected four MREs, is shown.
[b]SDE using QUIB (exponential weight, $\gamma = 2.5$), which performed best among eight selected SDEs.
[c]Three combined_scores shown are BLOSUM30 (γ =1.5)+2.05*QU_C2($\gamma = 1.0$), BLOSUM30 ($\beta = 10$)+2.9*QU_C2(γ = 1.5), and CC80 ($\alpha = 0.01$)+0.85*QU_C1($\beta = 5.0$).
[d]The number of subsets in each decoy set is shown. A subset consists of the native structure of a protein and decoy structures.
[e]The number in a parenthesis is the average Z-score of the native structure among decoy structures. Values of DFIRE to GOAP are negative because the potentials are negative values for favorable atom interactions while MRE and SDE have positive scores for native structures.
[f]Values for DOPE and RW are left empty for hg_structal, ig_structal, and ig_stractal_hire because the paper by Zhang and Zhang (2010) which showed the results for DOPE and RW did not have results for those.

Table II
Average RMSD and TM-Score of Decoys

Decoy Sets		DFIRE[a]	DOPE	RW	RWplus	MRE (CC80)[b]	SDE (QUIB)	Combinations of MRE and SDE: BLSM30+ QU_C2	Combinations of MRE and SDE: BLSM30+ QU_C2	Combinations of MRE and SDE: CC80+QU_C1
ROSETTA	**Top-1**	7.36 (0.469)	7.43 (0.466)	7.62 (0.460)	7.48 (0.464)	7.90 (0.452)	**6.63** (0.489)	7.06 (0.488)	7.04 (0.489)	6.82 (**0.492**)
Decoys	**Top-5**	6.08 (0.533)	6.10 (0.536)	6.04 (0.537)	6.01 (0.525)	6.18 (0.517)	**5.63** (**0.541**)	5.97 (0.531)	5.99 (0.531)	5.76 (0.534)
	Top-10	5.79 (0.559)	5.85 (0.555)	5.78 (**0.560**)	5.76 (**0.560**)	5.61 (0.545)	**5.42** (0.553)	5.71 (0.548)	5.63 (0.551)	5.53 (0.553)
I-TASSER	**Top-1**	5.61 (0.558)	5.31 (0.560)	5.22 (0.569)	5.19 (**0.575**)	6.04 (0.539)	5.26 (0.570)	5.25 (0.559)	5.28 (0.558)	**5.18** (0.574)
Decoys	**Top-5**	4.45 (0.612)	**4.21** (0.613)	4.30 (0.616)	4.29 (0.608)	4.62 (0.593)	4.26 (**0.622**)	4.30 (0.608)	4.31 (0.610)	4.36 (0.605)
	Top-10	3.95 (0.632)	**3.89** (0.631)	**3.89** (0.633)	**3.89** (0.625)	4.24 (0.609)	3.98 (**0.635**)	4.10 (0.622)	4.12 (0.619)	4.12 (0.620)

Native structures were excluded from the datasets. The best RMSD (Å) and the TM-score (in parentheses) within Top 1, 5, 10 decoys selected by each scoring function are shown.
[a]Data for DFIRE to RWPlus are taken from the article by Zhang and Zhang (2010).
[b]The same configurations at Table I were used for MRE, SDE, and the combinations.

It is also worthwhile to note that our residue-environment scores recognized native structures with high *Z*-scores compared with the others. For the Lmds set, the second Combind Score showed the highest *Z*-score of 9.04, while the MRE score achieved 13.58 for the Moulder set. The MRE score also showed significantly high *Z*-score of 13.60 for the 4state_reduced set and 9.61 for the I-TASSER set. The *Z*-scores by MRE for the 4_state_reduced set is three times higher than OPUS-PSP, which recognized the same number of natives. In the case of the I-TASSER set, the *Z*-score of 9.61 by MRE is almost twice higher than RWplus that recognized the same number of natives (56). This noticeable *Z*-score difference is equally observed in the hg_structal, ig_structal, and ig_structal_hires decoy sets. Our MRE results showed the highest *Z*-scores, 5.14 (hg_structal), 7.69 (ig_structal), and 4.35 (ig_structal_hires), which is on average more than two times better results than those of the second best one, GOAP. The three combinations of MRE and SDE also presented similar results.

Average global RMSD and TM-score of decoys

Next, we calculated the quality of top scored decoy structures (excluding the native structures) by our residue environment scores (Table II). The Rosetta decoy set, which was the most difficult in the native structure recognition in Table I, as well as the I-TASSER set were used. For the Rosetta set, the SDE with QUIB showed the smallest RMSD for all Top-1, Top-5, and Top-10. It also showed the best (largest) TM-score for Top-1 and Top-5, and the second best TM-score (0.553) for Top-10. The combined score of CC80 and QU_C1 (rightmost column) performed second best to the SDE in terms of RMSD for all Top-1, Top-5, and Top-10. The MRE with CC80, which performed better than the SDE for the native structure recognition (Table I) did not perform well for the decoy selection.

For the I-TASSER decoy set, the combined score with CC80 and QU_C1 showed the smallest RMSD for Top-1 rank. In the case of Top-5 and Top-10, DOPE[63] was the best in the RMSD, and the SDE was a close second. SDE also was the best in terms of TM-score for Top-5 and Top-10. Overall, we can conclude that the SDE performed the best for selecting low RMSD and large TM-score decoys.

Correlation coefficient of energy scores with TM-score of models

This test was done to examine near-native model selection performance of the scoring functions. Therefore, native structures were not included in the decoy set. Pearson's correlation coefficient (CC) of the scores with TM-scores of models as well as the average TM-score of the top-scoring models of each decoy sets were reported in Table III.

The best results in CC among our scores came from SDE results, 0.598. SDE using the QUIB matrix showed the largest CC among all scores in the fisa_casp3, Lattice_ssfit, ig_structal_hires, and Mourlder sets. However, GOAP and the other three potentials showed slightly larger average CC values. In terms of the average TM-score of the top-scoring model, SDE showed the largest value of 0.697, which is a tie with GOAP. This result indicates that SDE is good at selecting near-native decoys on average, which was consistent for SDE with the results in Table II and Table III. Combinations of MRE and SDE gave similar results with the SDE.

Examples of correlations of the residue environment score and the RMSD of decoys are shown in Figure 7. In these three decoy sets, native structures were recognized with significantly higher MRE, SDE, and Combined Scores compared to the decoys. The MRE score (left panels) had weak correlation to RMSD of decoys; however, the score gap between native structures and decoy structures was larger than the SDE. The SDE had larger

Table III
Average Pearson's Correlation Coefficient of Energy Score with TM-Score and Average TM-Score of Selected Models

								Combinations of MRE and SDE			
Decoy sets	DFIRE[a]	RWplus	dDFIRE	OPUS-PSP	GOAP	MRE (CC80)[b]	SDE (QUIB)	BLSM30+QU_C2	BLSM30+QU_C2	CC80+QU_C1	# Targets
4state_reduced	−0.635 0.659	−0.606 0.667	−0.693 0.732	−0.589 0.755	**−0.694 0.818**	0.065 0.353	0.660 0.745	0.423 0.700	0.420 0.700	0.365 0.645	7
Fisa	−0.446 0.449	**−0.462** 0.434	−0.461 0.454	−0.282 0.405	−0.347 0.475	0.310 0.419	0.410 0.447	0.433 **0.478**	0.436 0.478	0.401 0.411	4
Fisa_casp3	−0.243 0.288	−0.240 0.277	−0.149 0.309	−0.095 0.270	−0.221 0.300	0.179 0.295	**0.325** 0.303	0.310 0.324	0.306 0.324	0.312 **0.336**	5
Lmds	−0.118 0.333	−0.147 0.346	**−0.248 0.364**	−0.091 0.339	−0.146 0.339	0.014 0.350	0.148 0.334	0.055 0.350	0.056 0.361	0.061 0.353	10
Lattice_ssfit	−0.094 0.247	−0.097 0.251	−0.070 0.266	−0.051 0.248	−0.058 0.248	0.007 0.242	**0.118 0.274**	0.080 0.240	0.078 0.254	0.069 0.263	8
hg_structal	−0.817 0.890	−0.806 **0.891**	−0.796 **0.891**	−0.752 **0.891**	**−0.825** 0.889	0.265 0.802	0.801 0.889	0.758 0.881	0.756 0.882	0.699 0.879	29
ig_structal	−0.785 0.945	−0.782 0.948	−0.766 0.948	−0.779 **0.953**	**−0.865** 0.946	0.245 0.923	0.832 0.939	0.738 0.935	0.744 0.934	0.715 0.937	61
ig_structal_hires	−0.876 0.947	−0.879 **0.950**	−0.844 0.946	−0.832 0.946	−0.885 0.944	0.343 0.933	**0.897** 0.939	0.823 0.945	0.827 0.943	0.799 0.941	20
Moulder	−0.859 0.734	−0.840 0.745	−0.881 0.748	−0.802 0.738	−0.886 0.771	0.198 0.409	**0.900 0.774**	0.811 0.721	0.809 0.725	0.743 0.707	20
ROSETTA	−0.441 0.507	−0.444 0.505	−0.393 0.480	−0.343 0.506	**−0.476** 0.511	0.163 0.484	0.307 0.525	0.281 0.522	0.277 0.523	0.317 **0.527**	58
I-TASSER	−0.519 0.571	−0.488 0.577	**−0.525 0.578**	−0.284 0.547	−0.477 0.567	0.102 0.541	0.503 0.570	0.395 0.559	0.393 0.558	0.387 0.574	56
All average	−0.613 0.689	−0.605 0.692	−0.601 0.691	−0.521 0.688	**−0.632 0.697**	0.185 0.632	0.598 **0.697**	0.523 0.688	0.523 0.689	0.510 0.690	278

Native structures are excluded from all sets. Bold character entries are the best results in the respective decoy set. The first number in each cell is the Pearson correlation coefficient; the second number is the TM-score of lowest energy (highest score in our case) selected model.

[a]The values for DFIRE, RWplus, dDFIRE, OPUS-PSP, and GOAP were taken from the paper by Zhou and Skolnick (2011).

[b]The same configurations at Table I were used for MRE, SDE, and the combinations.

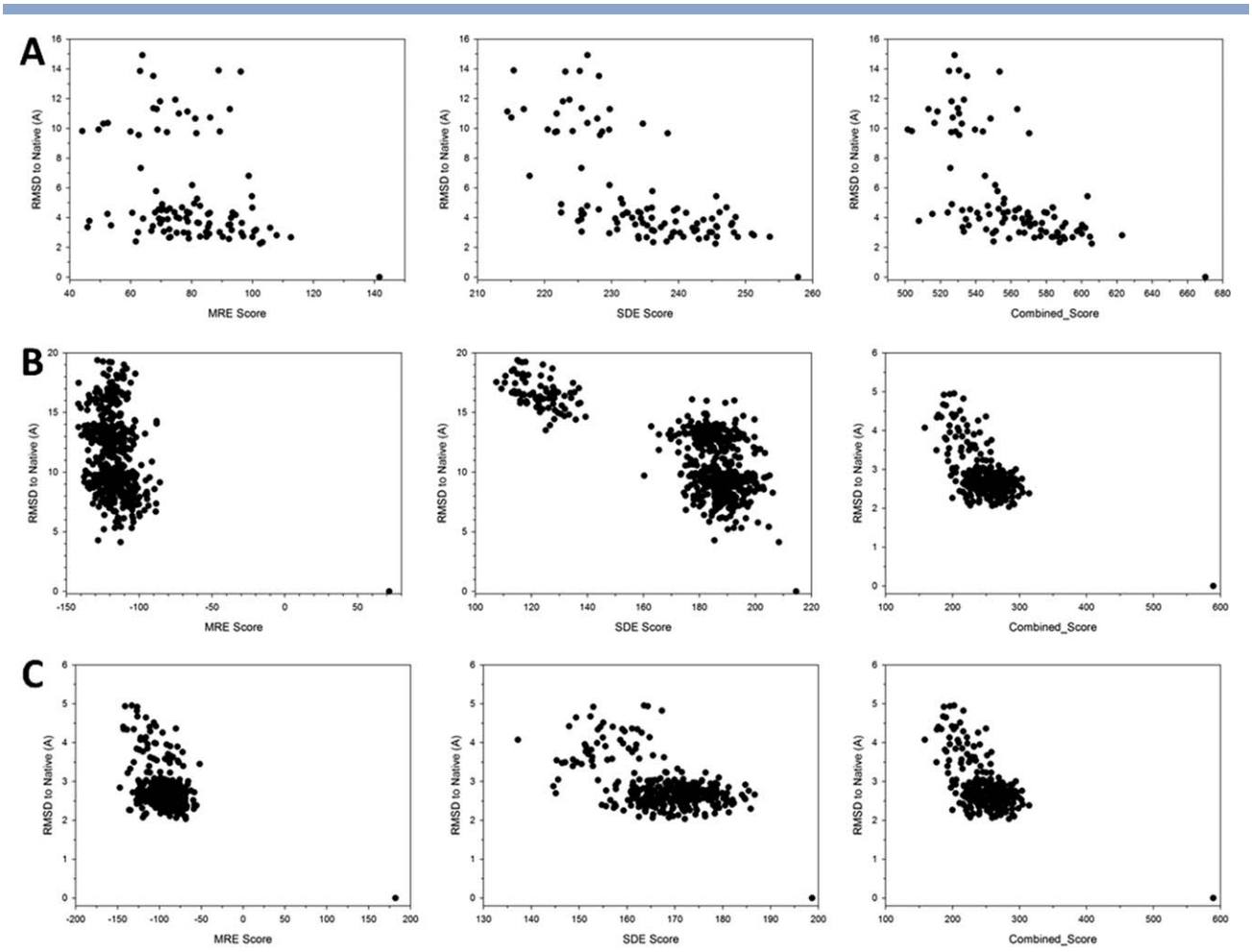

Figure 7

Examples of correlation between the environment scores and RMSD. RMSD of decoys are plotted relative to the residue environment scores for decoy sets of three proteins. Left, MRE (CC80 with $\gamma = 1.5$); middle, SDE (QU_C2 with $\gamma = 2.0$); Combined Score with MRE (CC80 with $\gamma = 1.5$) and SDE (QU_C2 with $\gamma = 2.0$) with a weight value of 2.05 were used. **A**: 2chf from the Rosetta decoy set; Correlation coefficients (CC) are −0.31, −0.67, and −0.62, respectively from left to right. **B**: 1gnuA from the I-TASSER decoy set. CC are −0.31, −0.74, and −0.77, respectively. **C**: 1csp from the I-TASSER decoy set. CC are −0.50, −0.55, and −0.63, respectively.

correlation to RMSD than the MRE, and the combined scores had comparable correlation coefficients to the SDE.

Performance on the Rykunov & Fiser's CASP model sets

We further tested the residue environment scores on prediction models submitted to the Critical Assessment of Techniques for Protein Structure Prediction (CASP) 5 to 8, which were compiled by Rykunov & Fiser[44] (Table IV). We show results of the same MRE and SDE and three combined scores shown in Table I and II. In addition, results of two extra combined scores that exhibited good performance are shown. The first one is a combination of MRE with BLOSUM30 and SDE with QU_C1 and the second one is a linear combination of two SDEs with CC80 and BLOSUM30. We wanted to try a combination of two SDEs because the SDE score performed well in model recognition in the absence of native structures (Table II, Table III, Fig. 7).

When the native structures are included (the right half of the table), the combination of BLOSUM30 and QU_C1 showed the smallest average rank (1.18) and the largest number of successful recognition of native structures (139). MRE performed better than SDE. When native structures were excluded (the left half of the table), different scoring functions showed up as top performing. The two SDE score combinations, CC80 and BLOSUM30, had the best average rank of 2.82 followed by SDE alone. Thus, SDE consistently performed well in the model recognition in the absence of native structures as observed in Tables II–IV.

Table IV
Performance on the Rykunov and Fiser CASP5–8 Decoy Set

Scoring function		Models only		Native included	
		Average Rank[a]	Ranked 1[b]	Average Rank[c]	ranked 1[d]
MRE (CC80)[e]		6.77	29	1.32	131
SDE (QUIB)		2.89	56	1.98	97
Combinations of	BLSM30+QU_C2	5.59	39	1.60	121
MRE and SDE	BLSM30+QU_C2	5.62	38	1.78	118
	CC80+QU_C1	5.07	37	1.92	113
	BLSM30+QU_C1	6.79	31	**1.18**	**139**
	CC80(SDE)+BLSM30(SDE)	**2.82**	**66**	1.99	89
Random		9.72	13.9	10.1	8.3

Values of the best performance are highlighted in bold.
[a]The average rank of lowest energy (highest scored; in the case of our residue environment scores) decoy by GDT_TS score in the absence of native structure.
[b]The number of sets when the best model was ranked as first in the absence of native structure.
[c]The average rank of the lowest energy (highest scored; in the case of our residue environment scores) decoy by GDT_TS when native structures are present.
[d]The number of sets when the best model was ranked as first when native structures are present.
[e]MRE using CC80 (RMSD-based weight, $\alpha = 0.01$) and the SDE using QUIB (exponential weight, $\gamma = 2.5$). In addition, results of five combined scores are shown. The first three are the same combinations of a MRE and a SDE as shown in Tables I and II: BLOSUM30 ($\gamma = 1.5$)+2.05*QU_C2($\gamma = 1.0$), BLOSUM30 ($\beta = 10$)+2.9*QU_C2($\gamma = 1.5$), and CC80 ($\alpha = 0.01$)+0.85*QU_C1($\beta = 5.0$). The last two are new combined scores: Blosum30 (β=10.0)+1.15*QU_C1(β=5.0), and the last one combines two scores from SDEs, CC80(α=0.01)+6.7*Blosum30 (β=2.0).

Examples of different performance of PRESCO and pairwise potentials

Here we show examples from the native selection test (Table I) that illustrate the characteristic PRESCO's performance in comparison with the existing statistical potentials. In Table V, we show native selection results of four decoy sets, for which PRESCO showed significantly better performance than the other potentials. For the Lmds decoy sets of 1bba and 1fc2, both MRE and SDE selected the native at the top rank among 501 decoys in the dataset while the native was ranked almost at the bottom by the other four potentials. Indeed, 1bba and 1fc2 are small proteins and known as difficult decoy sets as several existing potentials[58,61,64–66] failed to recognize the native structures among the decoys. The two decoys sets, 1fbi and 3hfm, in the Ig_structal set present the same story: MRE recognized the native at the top and SDE identified the native at the 4th and 14th for 1fbi and 3hfm, respectively, while the other potentials ranked the native almost at the bottom of the rank.

In Figure 8, we closely investigated the difference between SDE and two potentials, GOAP and DFIRE, which are well-known and widely used potentials. For the 1fc2 decoy sets, GOAP selected a decoy with an RMSD of 4.42 Å (1fc2.60276.pdb) with the lowest (i.e., best) energy, −5445.20 (−126.6 per residue), while the native structure has a GOAP energy of −4072.89 (−94.7 per residue). Figure 8(A) shows the breakdown of the energy difference of individual residues in the decoy and the native by SDE, MRE, GOAP, and DFIRE. Residues with the negative energy difference (y-axis) have a lower (thus preferred) energy for the native over the decoy. It is shown that GOAP has only 13 residues among 43 residues that have a lower energy in the native than in the decoy, while SDE preferred the native over the decoy for 25 residues. We further focused our attention to Residue 12, isoleucine (ILE12) because the evaluation of this residue's conformation by SDE and GOAP were very different: SDE strongly preferred the native structure for this residue over the decoy with the residue-wise energy of −618.3 (native) and 155.3 (decoy), while GOAP gave a preferable, negative residue-wise energy (−197.2) for the decoy. The side-chain centroid position of ILE12 in the decoy is 7.37 Å away from its correct position when the decoy is superimposed with the native, and consequently, ILE12 is in a very different environment in the decoy in comparison with the native: It is exposed outside in the decoy [Fig. 8(B), residues in red] while it is buried in the core in the native [Fig. 8(C)]. Although an exposed hydrophobic residue is in general not preferred, ILE12 in the decoy has negative energies by GOAP [Fig. 8(D)] partly because of interactions with neighboring hydrophobic residues, PHE26, LEU30, and LEU41, all of which deviated from their correct positions by 3.41, 1.55, and 3.82 Å, respectively [Fig. 8(E)]. On the other hand, SDE correctly considered the native, buried environment is more preferable than the exposed environment in the decoy for ILE12. In Table VI, top 5 most similar residue environments for ILE12 in the native and the decoy

Table V
Illustrative Performance of PRESCO

Decoy set	PDB ID	DFIRE	RWplus	OPUS-PSP	GOAP	MRE	SDE
Lmds	1fc2	500/501[a]	501/501	409/501	489/501	1/501	1/501
	1bba	501/501	501/501	501/501	501/501	1/501	1/501
Ig_structal	1fbi	61/61	59/61	57/61	46/61	1/61	4/61
	3hfm	59/61	59/61	57/61	55/61	1/61	14/61

[a]The rank of the native structures (left) among the total number of decoys (right) is shown.

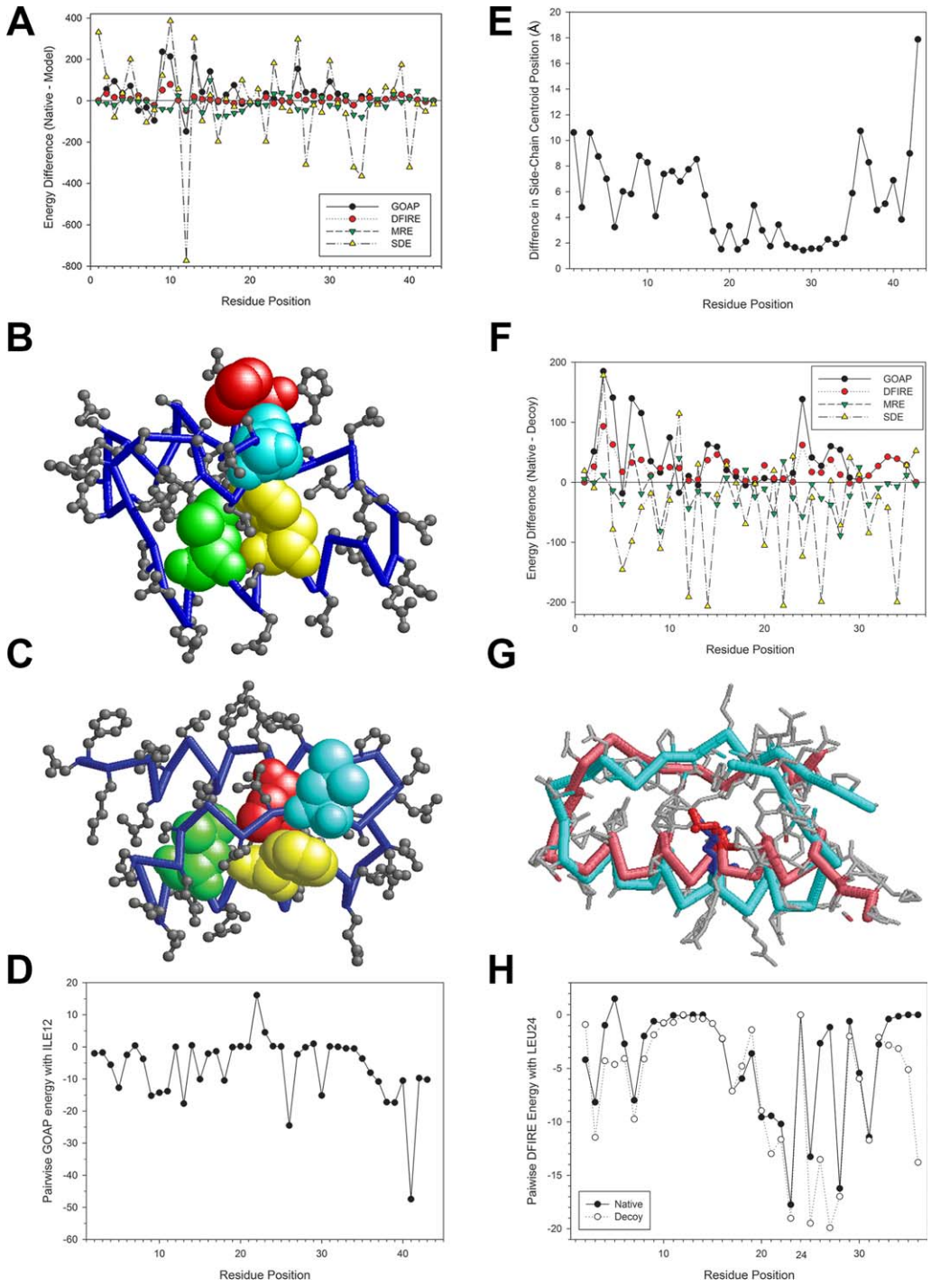

Figure 8

Examples of residue-wise energies by SDE and other potentials. Two decoy sets included in the Lmds set, 1fc2 and 1bba, were used. As shown in Table V, SDE and MRE successfully identified the native structure among 501 decoys for these two decoy sets. **A** to **E** are comparison between the SDE and GOAP energies at individual residues in the native structure of 1fc2 and the lowest GOAP energy decoy, 1fc2.60276.pdb. **F** to **H** are comparison between the SDE and DFIRE energies at each residue in the native of 1bba and the lowest DFIRE energy decoy, 1bba.1697.pdb. A: Energy difference at each residue between the native 1fc2 and the decoy 1fc2.60276.pdb by GOAP, DFIRE, MRE, and SDE. The *y*-axis shows the residue-wise energy difference between the native and the decoy. A negative value indicates that the residue has a lower energy in the native than the decoy. SDE and MRE scores are negated so that they have the same sign as the other two potentials. B: The structure of the decoy 1fc2.60276.pdb. ILE12 is shown in red, and three residues that have preferable GOAP energy between 1LE12, namely, PHE26, LEU30, and LEU41 are shown in yellow, green, and cyan, respectively. C: The native structure of 1fc2. ILE12, PHE26, LEU30, and LEU41 are shown in the same colors as in the Panel B. D: The pairwise GOAP energies between ILE12 and the other residues in the decoy 1fc2.60276.pdb. E: The Euclidean distance of side-chain centroids of each residue in the native and the decoy after the two structures are superimposed by the LGA program. A high distance (*y*-axis) indicates that the residue position in the decoy is far off from its correct position. F: Energy difference at each residue between the native 1bba and the decoy 1bba.1697.pdb by GOAP, DFIRE, MRE, and SDE. G: Superimposition of the native structure (pink) of 1bba and the decoy 1bba.1697.pdb (cyan). LEU24 are shown in the stick representation in red and blue in the native and the decoy, respectively. H: Pairwise DFIRE energies between LEU24 and each residue in the native (filled circles) and in the decoy 1bba.1697.pdb (open circles).

Table VI
Top 5 Most Similar Environments Identified by SDE

		Native		
Target residue	Rank	PDB ID	Res. Pos.[a]	AA Type[b]
ILE12 in 1fc2 (native)	1	2ahmA[c]	44	ILE
	2	2bgiA	173	LEU
	3	1iznA	21	ILE
	4	1p3dA	317	LEU
	5	1m15A	181	LEU
ILE12 in a decoy 1fc2.60276.pdb[d] (lowest energy by GOAP)	1	2b81A	142	ALA
	2	1uddA	93	GLU
	3	2c31A	22	MET
	4	1ykhA	133	LEU
	5	1m5wA	146	GLU
LEU24 in 1bba (native)	1	1ub9A	91	LEU
	2	1ppjC	160	LEU
	3	1fxkC	121	ILE
	4	1bh9A	61	VAL
	5	1n5uA	183	ASP
LEU24 in a decoy 1bba.1697.pdb[e] (lowest energy by DFIRE)	1	1x8yA	370	ASP
	2	1bh9A	62	ILE
	3	2a61A	128	ILE
	4	1txlA	210	GLU
	5	1vf7A	81	GLN

Top five most similar residue environments of the target residues in the native and in the lowest energy decoy (defined by GOAP/DFIRE) were selected by SDE.
[a]Residue position.
[b]Amino acid type.
[c]According to SDE, ILE44 of 2ahmA has the most similar residue environment with the target residue, ILE12 in the native structure of 1fc2.
[d]1fc2.60276.pdb was selected by GOAP as the lowest energy decoy among the decoy set. Because ILE12 has a different residue environment in the decoy from that in the native, the five most similar environments are different from those for ILE12 in the native.
[e]1bba.1697.pdb was selected by DFIRE as the lowest energy decoy among the decoy set.

structure that were identified by SDE are summarized. In the case of ILE12 in the native, all the five identified environments are centered on hydrophobic amino acids, isoleucine or leucine, which locate at packing interfaces of helices and loops in protein structures that are globally different from 1fc2. In contrast, similar environments of ILE12 in the decoy are all exposed, including those which are centered on glutamic acid residues. Thus, SDE clearly distinguished very different environments of ILE12 in the native and in the decoy structures.

In the panels F–H in Figure 8, performance of SDE is compared with DFIRE. A decoy with an RMSD of 6.07 Å (1bba.1697.pdb) is selected as the lowest energy by DFIRE. Figure 8(F) shows that DFIRE (red) considered that the majority of the residues (33 out of 36 residues) have a lower energy in the decoy than in the native. In contrast, SDE energy is lower in the native, often substantially lower, than in the decoy for many residues (21 out of 36 residues). For 1bba, we further investigate energies for LEU24, because SDE and MRE evaluated energy of this residue very differently from DFIRE and GOAP [Fig. 8(F)]: DFIRE and GOAP strongly preferred the decoy over the native while SDE and MRE considered the native is more preferable structure for this residue. From the structure imposition of the native and the decoy [Fig. 8(G)], difference of environment of LEU24 in the native and decoy is not very obvious, although two helices in the native are packed slightly tighter than those in decoys. A close examination of the pairwise DFIRE energy between LEU24 and each residue [Fig. 8(H)] shows that the energy profile for the native and the decoy are almost the same with some difference observed for residue 25 to 27, which have lower energy in the decoy than in the native. On the other hand, SDE preferred the native for LEU24 to the decoy, which is also evident in the top five most similar environments for LEU24 detected by SDE (Table V). The four most similar environments to LEU24 in the native are those from hydrophobic amino acids that locate inside of proteins, while similar environments for LEU24 in the decoy included exposed residues.

The purpose of showing the examples is to illustrate the performance of PRESCO in comparison with pairwise potentials. Of course PRESCO's performance was not always better than the two potentials in the entire benchmark test; nevertheless the examples show the advantage that the residue environment-based score can achieve.

DISCUSSION

We have developed two residue environment sores, the Main-chain Residue Environment (MRE) and the Side-chain Depth Environment (SDE). The Protein Residue Environment Score (PRESCO), which uses MRE and SDE, compared favorably against existing knowledge-based statistical potentials in recognizing native and close-to-native decoys. Notably, MRE and SDE have complementary strength: MRE performs well in identifying native structures (Tables I and V) while the SDE score has better correlation to the RMSD of decoys and thus works better in recognizing near-native decoys when native structure is not included (Tables II–IV).

As scoring decoys is a central problem in protein structure prediction, various potentials have been developed for the native structure/near-native decoy selection. In contrast to most of the knowledge-based statistical potentials that capture preference of pairwise interactions between atom or atom groups including the potentials compared with PRESCO in this work, PRESCO was designed to capture multibody interactions of residue side-chains. It was shown that such residue environments captured by PRESCO exist in proteins of different folds. PRESCO does not need the reference state, which is often problematic in designing statistical potentials. Through this work we have shown that considering residue environments for capturing multibody interactions may be a promising alternative direction to the conventional two-body statistical potentials for protein structure

prediction and modeling. Similar ideas of residue environment scores will be also effectively applied for validating crystal structures of proteins and for protein design.

ACKNOWLEDGMENTS

The authors thank Lenna X. Peterson for proofreading the manuscript.

REFERENCES

1. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
2. Harrison A, Pearl F, Mott R, Thornton J, Orengo C. Quantifying the similarities within fold space. J Mol Biol 2002;323:909–926.
3. Rossmann MG. Super-secondary structure: a historical perspective. Methods Mol Biol 2013;932:1–4.
4. Fernandez-Fuentes N, Fiser A. Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol 2006;6:15.
5. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng 1994;7:953–960.
6. Hvidsten TR, Kryshtafovych A, Fidelis K. Local descriptors of protein structure: a systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. Proteins 2009;75:870–884.
7. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676.
8. Yang YD, Park C, Kihara D. Threading without optimizing weighting factors for scoring function. Proteins 2008;73:581–596.
9. Mitchell JB, Thornton JM, Singh J, Price SL. Towards an understanding of the arginine-aspartate interaction. J Mol Biol 1992;226:251–262.
10. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. Procheck—a program to check the stereochemical quality of protein structures. J Appl Crystallography 1993;26:283.
11. Bonneau R, Strauss C, Rohl C, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65.
12. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001(Suppl 5):127.
13. Zhou H, Skolnick J. Protein structure prediction by pro-Sp3-TASSER. Biophys J 2009;96:2119–2127.
14. Zhang Y, Arakaki AK, Skolnick J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 2005;61 (Suppl 7):91–98.
15. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA 2001;98:10125.
16. Kihara D, Chen H, Yang YD. Quality assessment of protein structure models. Current Protein Pept Sci 2009;10:216–228.
17. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653.
18. Chen H, Kihara D. Effect of using suboptimal alignments in template-based protein structure prediction. Proteins 2011;79:315–334.
19. Sippl MJ. Knowledge-based potentials for proteins. CurrOpinStructBiol 1995;5:229.
20. Betancourt MR. Knowledge-based potential for the polypeptide backbone. J Phys Chem B 2008;112:5058–5069.
21. Zhang W, Liu S, Zhou Y. SP5: improving protein fold recognition by using torsion angle profiles and profile-based gap penalty model. PloS One 2008;3:e2325.
22. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins 2007;68:636–645.
23. Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55:1005.
24. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. J Comput Chem 2007;28:2059–2066.
25. Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature 1978;275:673–674.
26. Manavalan P, Ponnuswamy PK. A study of the preferred environment of amino acid residues in globular proteins. Arch Biochem Biophys 1977;184:476–487.
27. Karlin S, Zhu ZY, Baud F. Atom density in protein structures. Proc Natl Acad Sci USA 1999;96:12500–12505.
28. Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. Protein Sci 2005;14:1955.
29. Zhong L, Johnson WC, Jr. Environment affects amino acid preference for secondary structure. Proc Natl Acad Sci USA 1992;89:4462–4465.
30. Minor DL, Jr., Kim PS. Context-dependent secondary structure formation of a designed protein sequence. Nature 1996;380:730–734.
31. Feng Y, Kloczkowski A, Jernigan RL. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. Proteins 2007;68:57–66.
32. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Sci 1997;6:1467–1481.
33. Sanchez-Gonzalez G, Kim JK, Kim DS, Garduno-Juarez R. A beta-complex statistical four body contact potential combined with a hydrogen bond statistical potential recognizes the correct native structure from protein decoy sets. Proteins 2013;81:1420–1433.
34. Summa CM, Levitt M, Degrado WF. An atomic environment potential for use in protein structure prediction. J Mol Biol 2005;352:986–1001.
35. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 1995;252:709–720.
36. 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;268:209–225.
37. Mooney SD, Liang MH, DeConde R, Altman RB. Structural characterization of proteins using residue environments. Proteins 2005;61:741–747.
38. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 1999;7:723–732.
39. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic acids research 2005;33(Web Server issue):W94–W98.
40. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
41. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
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:e15386.
43. 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.
44. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics 2010;11:128.

45. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol 2014;1079:105–116.
46. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data: an efficient Monte Carlo model. Proteins 1998; 32:475.
47. Nishikawa K, Ooi T. Prediction of the surface-interior diagram of globular proteins by an empirical method. Int J Peptide Protein Res 1980;16:19–32.
48. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
49. Gront D, Kolinski A. BioShell—a package of tools for structural biology computations. Bioinformatics 2006;22:621–622.
50. Dai L, Yang Y, Kim HR, Zhou Y. Improving computational protein design by using structure-derived sequence profile. Proteins 2010;78: 2338–2348.
51. Ponder JW, Richards FM. An efficient newton-like method for molecular mechanics energy minimization of large molecules. J Comp Chem 1987;8:1016–1024.
52. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915.
53. Qu CX, Lai LH, Xu XJ, Tang YQ. Phyletic relationships of protein structures based on spatial preference of residues. J Mol Evol 1993; 36:67.
54. Qian B, Goldstein RA. Optimization of a new score function for the generation of accurate alignments. Proteins 2002;48:605.
55. Kolaskar AS, Kulkarni-Kale U. Sequence alignment approach to pick up conformationally similar protein fragments. J Mol Biol 1992;223:1053.
56. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins 2006;64:587.
57. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res 2000;28:374.
58. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 2008;376:288–301.
59. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101:2043–2052.
60. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302.
61. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714.
62. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72:793–803.
63. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
64. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001;44:223.
65. Wang K, Fain B, Levitt M, Samudrala R. Improved protein structure selection using decoy-dependent discriminatory functions. BMC Struct Biol 2004;4:8.
66. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.