PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SVR_CAF: An integrated score function for detecting native protein structures among decoys

Jianhong Zhou,[†] Wenying Yan,[†] Guang Hu, and Bairong Shen*

Center for Systems Biology, Soochow University, Suzhou, Jiangsu, 215006, China

ABSTRACT

An accurate score function for detecting the most native-like models among a huge number of decoy sets is essential to the protein structure prediction. In this work, we developed a novel integrated score function (SVR_CAF) to discriminate native structures from decoys, as well as to rank near-native structures and select best decoys when native structures are absent. SVR_CAF is a machine learning score, which incorporates the contact energy based score (CE_score), amino acid network based score (AAN_score), and the fast Fourier transform based score (FFT_score). The score function was evaluated with four decoy sets for its discriminative ability and it shows higher overall performance than the state-of-the-art score functions.

Proteins 2013; 00:000–000.

Key words: protein native structure selection; support vector regression; contact energy; amino acid network; fast Fourier transform.

INTRODUCTION

The protein folding problem, which refers to the rules governing the formation of a three-dimensional structure from the amino acid sequence of a protein, is considered as the second half of the genetic code and the major challenge in structural biology.[1] During the last decades, a large body of experimental and computational methods has provided insights into this problem. With the development of high-throughput genome sequencing, there is an increasing realization that computational methods for predicting and modeling of protein structures become very necessary for structural genomics, since the number of protein sequences is increasing rapidly while the experimental approaches for structural determination are expensive and time-consuming. One of the basic ideas of protein structure prediction is to enumerate all possible decoys of the target protein and then design an energy function or a score function to select the most native-like structure from them.[2]

Over these years, a large number of score functions have been designed based on the Anfisen's principle, which assumes the native protein structure to be the lowest energy state.[3] Score functions are mainly grouped into four catalogs: (1) physics-based energy functions, (2) statistics or knowledge-based score functions, (3) consensus scores, (4) machine learning-based scores.[4] The physical energy function,[5–7] which calculates the molecular mechanics force field, has not been widely applied to native protein recognition due to the time consuming problem and computational complexity.[8] The knowledge-based score functions[9–12] are extracted from the statistical distribution of atoms or residues interactions in known native structures. An earlier work[13] has compared various forms of knowledge-based score functions depending on how statistics were calculated and how proteins were modeled, such as distance independent/dependent scores,[10,14] solvent accessible surface score,[15] contact

Grant sponsor: National Natural Science Foundation of China; Grant numbers: 91230117; 31170795; 20872107; Grant sponsor: the Specialized Research Fund for the Doctoral Program of Higher Education of China; Grant number: 20113201110015; Grant sponsor: International S&T Cooperation Program of Suzhou; Grant number: SH201120; Grant sponsor: National High Technology Research and Development Program of China, 863 program; Grant number: 2012AA02A601.

[†] Jianhong Zhou and Wenying Yan are the co-first authors for this article.

*Correspondence to: Bairong Shen, Center for Systems Biology, Soochow University, No1. Shizi Street, Suzhou, Jiangsu 215006, China. E-mail: bairong.shen@suda.edu.cn

Received 22 April 2013; Revised 30 August 2013; Accepted 10 September 2013
Published online 30 September 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24421

energy (CE),[16,17] and composite energy.[18,19] However, the prediction power of knowledge-based score functions may be limited because they only reflect some of the statistical properties, while the folding of proteins is more complicated.[20–22] The hypothesis of consensus approaches is that near-native structures have low energy and thus tend to be most clustered to similar conformations.[4] By using the representatives from the largest clusters in a decoy set as candidates, consensus approaches have been applied in Critical Assessment of protein Structure Prediction (CASP).[23–28] However, it only performs well when most of the decoys in the set are close to native structure, otherwise it is outperformed by knowledge-based approach.[4] Machine-learning methods, like artificial neural networking[29] or support vector machine,[19,30–33] utilize machine learning techniques to learn how to combine multiple features. Because of the intrinsic superiority of nonlinear combination in comparison with linear combination, they are widely used to evaluate the quality of decoys.[33] On the other hand, a novel method of investigating protein folding is to view protein structures as networks of connections due to the interactions between amino acids.[34] This approach captures both the global and the local perspectives of structures or topologies and has become an attractive model in understanding and predicting protein structure and function.[35] In the related works,[22,36–41] some particular network parameters, including degree, average short path length, complexity, clustering coefficient of the largest cluster (CCoe), and the size of the top large communities (CComS) are reported to provide important insights in protein selection.

In this study, we investigated three individual scores that are based on the calculation of (1) environment-dependent residue CE (**C**E_score), the amino acid network (**A**AN_score) and the fast Fourier transform (**F**FT_score). Using support vector regression (SVR) to combine these three normalized scores, we proposed a novel machine learning score function (SVR_CAF) to predict the structural similarity between native structures and their decoys. The advantages of this method were demonstrated by comparative performance assessment on four different decoy sets. This work provides new insight into the future development of the scoring functions for protein structure selection.

MATERIALS AND METHODS

Decoy sets

Seven decoy sets, including 4-state_reduced,[42] fisa, fisa_casp3,[43] lmsd, lattice_ssfit,[44] Rosetta,[45] and I-TASSER decoy set,[46] were used to evaluate the performance of our score function. The first five decoy sets were available from Decoys'R'Us database.[42] The Rosetta decoy set was generated for 59 proteins by Baker and coworkers.[47] For each protein, 20 random models and 100 lowest scoring models from 10,000 decoys were included using Rosetta *de novo* structure predictions followed by all-atom refinement. The I-TASSER decoy set includes decoys generated for 56 nonhomologs small proteins.[46]

To assess the quality of the score functions, RMSD, and TM-score were used as two criteria. RMSD is the root mean squared derivation of all Cα atoms of the decoy to the native structure. In the following evaluation, the lower RMSD will indicate the better selective ability of the score function. Since, RMSD is related to the length of protein but insensitive to the protein structure topology, the template modeling score (TM-score) was calculated to measure the similarity of topologies of two protein structures.[48] The TM-score has a range between (0,1] with better similarity having higher TM-scores.

Environment-dependent residue contact energy calculation and the CE_score

Environment-dependent residue contact energy (ERCE) demonstrates the influence of secondary structural environments on the residue interactions, which is defined as[49]:

$$e_{ij} = -\ln\left(\frac{N_{ij}N_{00}}{N_{i0}N_{j0}}\frac{C_{ij}C_{00}}{C_{i0}C_{j0}}\right) \tag{1}$$

where N_{ij}, N_{i0}, N_{j0}, and N_{00} are the contact numbers from the protein structures, and C_{ij}, C_{i0}, C_{j0}, and C_{00} are the corresponding quantities that are expected in a reference state.

The contact energies between residues and the total CE of a protein were calculated by the software RankViaContact.[50,51] The total CE for each structure is calculated and normalized over the decoy set and the normalized total CE was taken as the contact energy score (**CE**_score) for the next step SVR model input.

The amino acid network-based score (AAN_score or Score$_{AACEN}$)

On the basis of the CE above, protein amino acid network (AAN) was constructed, with residues as nodes and contact energies as edges. When we performed the calculation, the residues in the protein were represented by the centroids of their side chains, two amino acids (i and j) were considered to be in contact, if the distance between the centroids (RC) is less than 6.5 Å.[49–51] Each element in the adjacent matrix AM of the AAN can be defined as:

$$AM_{ij} = \begin{cases} 0, & E_{ij} \geq 0 \\ 1, & E_{ij} < 0 \end{cases} \tag{2}$$

The degree of a node is given by:

$$K_i = \sum_{j=1}^{N} AM_{ij} \tag{3}$$

N here is the number of nodes.

Average short path length of the network is the average shortest path between all pairs of nodes, and it is given by:

$$L = \frac{1}{N(N-1)} \sum_{i \neq j=1}^{N} l_{ij} \tag{4}$$

Where N is the number of nodes, l_{ij} is the minimal number of edges one must transverse to reach a node j from a node i ($i \neq j$). This topology property measures the least link number of all pathways between two nodes and a lower L value implies a more compact form in the network.

On the basis of K_i and L, we defined a new amino acid CE network score function as:

$$\mathrm{Score}_{\mathrm{AACEN}} = -\frac{\sum_{i}^{N} K_i}{L} \tag{5}$$

In the AACEN, the total degree and average short path length are chose to define the AACEN score function, because these two network parameters are the indicators of the well folding of protein structures. Higher total degree of a network indicates better interconnectivity among the residues in the structure, and lower average shortest path length implies the structure is in compact form.[52] Our hypothesis is that native structures, which are perfectly folded, should show higher degree and lower average shortest path length than their non-native decoys, which are usually incorrectly folded or misfolded. The combination of the two network properties is expected to show better performance for the discrimination of native folding from decoys.

Score based on fast fourier transform of hydrophobicity and solvent accessibility profiles (FFT_score or Score$_{\mathrm{FFT}}$)

The sequence of each conformation for a protein is represented as:

$$\mathrm{seq} = \{aa_1,\ aa_2, \ldots aa_N\} \tag{6}$$

N here is the length of the protein sequence.

To characterize the protein folding, both the hydrophobicity profile (which is the main driving force for protein folding) and solvent accessibility surface (SAS) profile (which is the effect of hydrophobicity) are considered. The *seq* is mapped to the Eisenberg hydrophobic values of the 20 amino acids,[53] and the hydrophobic profile is formed:

$$\mathrm{hyd} = \{h_1, h_2, \ldots, h_{\mathrm{N}}\} \tag{7}$$

where h_1, h_2, … h_{N} is the standard measure of hydrophobicity of each amino acid sequence. The absolute SAS was computed using the program DSSP.[54] The relative SAS ($\mathrm{S_x}$) was calculated as the observed absolute SAS of a residue (X) in protein structure divided by that observed in an extended tripeptide (G-X-G or A-X-A) conformation.[55,56] So the structure has a solvent accessibility profile:

$$\mathrm{sas} = \{s_1, s_2, \ldots s_{\mathrm{N}}\} \tag{8}$$

Then, the buried sequence (buried_*seq*), exposed sequence (exposed_*seq*), buried_SAS, and exposed_SAS were defined as in the earlier work.[57] All of the parts are decomposed by fast fourier transform (FFT) formulation,[58] respectively:

$$S(k) = \sum_{j=1}^{N} s(j)\omega_N^{(j-1)(k-1)} \tag{9}$$

where $\omega = \exp(-2\pi i/N)$, with $i = 0, 1, \ldots, N-1$, N is the length of buried or exposed sequence. Given by $S(k) = Y$, the power spectrum is represented as:

$$p = Y. * \mathrm{conj}(Y)/N \tag{10}$$

The FFT here converts protein primary sequence information into components of frequencies, the power spectrum in our case. This conversion is used for the characterization of the relationship between hydrophobic property of amino acids and the solvent accessibility in protein native structures since it is believed that these two profiles of native proteins would show more similarity than those of decoys. Our assumption is that these two profiles of native proteins would show more similarity than those of decoys.[13,59–62]

Finally, a score for describing the similarity between hydrophobicity and SAS profiles is defined:

$$\mathrm{Score}_{\mathrm{FFT}} = -\left(N_{\mathrm{buried}} * CC_{\mathrm{buried}} + N_{\mathrm{exposed}} * CC_{\mathrm{exposed}}\right) \tag{11}$$

where N_{buried} and N_{exposed} are the numbers of buried and exposed residues, respectively, $\mathrm{CC_{buried}}$ and $\mathrm{CC_{exposed}}$ are the correlation coefficients between the power of hydrophobicity and SAS for buried parts and exposed parts at each frequency point, respectively.

The integrated score: SVR_CAF

An integrated score was developed by combining the information from the three score functions described above. For this purpose, we employed the SVR, which was implemented using the LIBSVM.[63] SVR model was trained to predict the RMSD of a decoy given the three normalized scores above. The Gaussian radial basis function (RBF) was selected as the kernel function. The parameters γ (which controls the peak of the Gaussian functions) and C (which controls the cost for the regression errors) were adjusted. A leave-one-out jackknife method was applied to train every decoy set mentioned earlier.[33] For each decoy set, the decoys of one protein were used as the test set and the remaining as the training set. This process was repeated *N* times (*N* is the number of target proteins in the set) to test the performance and the parameters were optimized on the training set in each round. The set of parameters with highest accuracy were chosen to build the SVR model for the testing set. The predicted RMSD for all decoys were then collected for the further analysis.

Assessment of score functions

Four criteria were used to evaluate the integrated score: (1) the ability to discriminate native structures from decoys, including the number of correctly identified natives and *Z*-score of natives; (2) the quality of top 1, top 5, and top 10 decoys while native structures are absent; (3) the correlation coefficient between the score function and the quality (RMSD and TM-score to the native) of decoys.

RESULTS

Discrimination of native protein structures

We first tested the discrimination power of SVR_CAF score and its composed individual terms on Decoys "R" Us sets. The comparison results are listed in Table I, presented in terms of the number of the top 1, top 5, and top 10 ranked native structures within the decoy sets. It is clear that the CE term contributed most to discrimination of SVR_CAF score and select correctly 18 native proteins form the 32 decoys.

The performance of SVR_CAF was then compared with the 7 knowledge-based scoring functions: GOAP,[64] EPAD,[65] KBP,[10] DFIRE,[11] NCACO-score,[66] RAPDF,[9] 4 BOPT POT.[67] The comparison results are shown in Table II and presented in terms of the number of the top 1 ranked native structures within the decoy sets and the average *Z*-scores. The *Z*-score is defined as:

Table I

Performance of SVR_CAF and Its Individual Terms on Decoy'R'Us Sets

	CE	AACEN	FFT	SVR_CAF		
Decoyset/number[a]	Top 1	Top 1	Top 1	Top 1	Top 5	Top 10
4state/7	5	3	1	5	5	6
Fisa/4	1	1	1	1	2	2
fisa_casp3/3	0	1	0	0	0	1
Lmds/10	5	5	5	7	8	10
Lattice/8	7	5	3	8	8	8
Total	18	15	10	21	25	27/32

[a]Number of target proteins in the decoy set.

$$Z-\text{score}=\frac{\text{score}_{\text{native}}-<\text{score}_{\text{decoys}}>}{\sigma_{\text{decoys}}} \quad (12)$$

where $\text{score}_{\text{native}}$ is the score calculated for native structure and $<\text{score}_{\text{decoys}}>$ and σ_{decoys} are the average and standard deviation of scores of decoys. It is clear that the higher the *Z*-score, the better for the function's discriminative ability.

As shown in Table II, native structures of 21 proteins out of 32 proteins (65.63% success rate) were ranked in top-1 by SVR_CAF, the performance of native structure selection of SVR_CAF is better than those of KBP and 4BOPT POT, but not good as other. Among all these score functions, SVR_CAF has the best performance in terms of the *Z*-score (−5.66), followed by GOAP (−5.21), NCACO (−5.06), DFIRE (−4.27), KBP (−2.87), 4BOPT POT (1.87), and RAPDF (0.83). The *Z*-score was not used by EPAD for evaluation on this decoy set, so the performance of EPAD is not comparable here. The reason that SVR_CAF score missed the native state of 4rxn in 4state reduced set, 1fc2 and 1hdd-C in the *fisa* set, 1bl0, and 1jwe in *fisa_casp3* set, will be further discussed in the following sections.

The correlation between scores and decoy qualities

We compared the performance of the best decoy selection with DFIRE, DOPE,[12] RW, RWplus,[68] and SVR_CAF on both the Rosetta and I-TASSER decoy sets, the results are listed in Tables III and IV, respectively.

Both RMSD and TM-score were used for assessing the quality of every decoy. RMSD_top1 is the RMSD of the lowest score decoy, the average RMSD of the best decoy selected by SVR_CAF in Rosetta decoy set is 7.01 Å, which is 0.26 Å lower than that by DFIRE. The average TM-score of the best decoy selected by SVR_CAF is 0.494, which is higher than those obtained by other score functions. The top 5 and top 10 best decoys are also compared and the results are listed in the fourth and fifth columns of Table III. SVR_CAF performed best to select the top 10 decoys with an average RMSD 5.61 and average TM-score 0.564, while RW and RWplus performed best to select the top 10 decoys with an average TM-score 0.560 and average RMSD 5.76 Å, respectively. In Table III, the Pearson correlation coefficients between RMSD (and TM-score) and the scores given by DFIRE, DOPE, RW, RWplus, and SVR_CAF are listed in the sixth column. The SVR_CAF performs best among all these functions for this measurement.

Table II
The Comparison in Native Protein Discrimination for Decoy'R'Us Sets

Decoy set/number	GOAP	EPAD	RAPDF	KBP	DFIRE	NCACO−score	4BOPT POT	SVR_CAF
4state reduced/7	7/−4.38	5	7[a]/−3.01[b]	7/−3.24	6/−3.44	7/−4.67	5/2.56	5/−3.21
Fisa/4	3/−3.97	3	1/−1.27	0/−1.21	3/−4.67	1/−1.68	1/1.04	1/−3.04
Fisa_casp3/3	3/−5.27	2	3/−4.09	0/−2.08	3/−4.93	0/−2.16	0/−1.22	0/−2.13
Lmds/10	7/−4.07	9	3/0.52	3/−0.53	7/−0.99	7/−3.31	1/0.39	7/−4.25
Lattice_ssifit/8	8/−8.38	8	8/−7.18	8/−6.63	8/−8.00	8/−10.35	8/4.71	8/−15.65
Summary								
Top1[c]/total	28/32	27/32	22/32	18/32	27/32	23/32	15/32	21/32
Average *Z*-score	−5.21	–	−2.83	−2.88	−4.27	−5.06	1.87	**−5.66**

Bold signifies the best performance among the methods.
[a] Number of cases when native proteins ranked top 1 predicted by each score function.
[b] The average *Z*-score of native structure in the decoy structures. EPAD does not use *Z*-score for evaluation on this decoy sets so the *Z*-scores are not included.
[c] Total number of cases where native proteins ranked top 1 in multiple decoy sets.

The results for the analysis of I-TASSER decoy sets listed in Table IV show the similar tendency of the native structure selective ability, the SVR_CAF score function performed better than other functions in terms of both RMSD and correlation coefficient. In addition to those methods listed in Table IV, another promising potential is EPAD, which outperforms DFRIE and DOPE. Our SVR_CAF model is slightly better than EPAD with the first-ranked TM-score (SVR_CAF: 0.582; EPAD: 0.579) and the correlation (SVR_CAF:-0.573; EPAD:-0.532).

The discriminative ability of the score functions on the CASP5-8 models

We also evaluated the discriminative ability of the SVR_CAF and other score functions on the 143 decoy sets from CASP5-8.[69] Figure 1 shows the performance of SVR_CAF and other score functions, in terms of the average rank of the best decoy (*x*-axis) and the number of targets when the best decoy is ranked first (*y*-axis), in the absence or presence of native structures. Identification of native structures and best decoys using SVR_CAF is mediocre, correctly selecting 87 native structures and 54 best modes, respectively. Like all the other methods, SVR_CAF works better when native structures are present. In terms of the average ranking of native structures, SVR_CAF (2.58) works better than EPAD (2.71), but it performs worse than EPAD in terms of the average ranking of best decoys when natives are absent (SVR_CAF:4.73; EPAD:3.29).

DISCUSSION

Validation of SVR_CAF score

The **SVR_CAF score** has following advantages:

1. It considers environment-dependent residue CE and protein amino acid CE network (AAN). The discriminatory capability of the CE was tested in the threading experiment, three-dimensional contact prediction and the protein stability analysis.[49,51] As a result, it outperformed other residue pair potentials such as Miyazawa–Jernigan potential (MJ).[70]
2. It is well-known that a protein conformation is stabilized by the noncovalent interactions, which depends on the topology of the folded structure.[71] The merit of AAN is that it takes into account the information of topology of the three-dimensional structure. With the AAN model, a network-based score function (AAN_score) was proposed. Firstly, the degree K_i is an inherent nature of the graph and higher average degree indicates good interconnectivity among residues in protein structure; the average shortest path *L*

Table III
Comparison Results on Rosetta Decoy Set

	Real top 1[a]	RMSD_top 1[b] (TMscore_top 1)	RMSD_top 5[c] (TMscore_top 5)	RMSD_top 10[c] (TMscore_top 10)	Correlation coefficient
DFIRE		7.36 (0.469)	6.08 (0.533)	5.79 (0.559)	0.440 (−0.432)
DOPE		7.43 (0.466)	6.10 (0.536)	5.85 (0.555)	0.421 (−0.427)
RW		7.62 (0.460)	6.04 (0.537)	5.78 (0.560)	0.441 (−0.434)
RWplus		7.48 (0.464)	**6.01** (0.525)	5.76 (0.542)	0.435 (−0.427)
SVR_CAF	4.99 (0.596)	**7.10 (0.494)**	6.61 (**0.543**)	**5.61 (0.564)**	**0.491 (−0.524)**

Bold signifies the best performance among the methods.
[a] The true average RMSD (TMscore) of the best decoys in Rosetta set.
[b] The predicted average RMSD (TMscore) of the best decoys.
[c] The predicted average RMSD (TMscore) of the best 5 and 10 decoys.

Table IV
Comparison Results on I-Tasser Decoy Set

	Real top 1	RMSD_top 1 (TMscore_top 1)	RMSD_top 5 (TMscore_top 5)	RMSD_top 10 (TMscore_top 10)	Correlation coefficient
DFIRE		5.61 (0.558)	4.45 (0.612)	3.95 (0.632)	0.514 (−0.492)
DOPE		5.31 (0.560)	**4.21** (0.613)	3.89 (0.631)	0.319 (−0.317)
RW		5.22 (0.569)	4.30 (**0.616**)	**3.89 (0.633)**	0.520 (−0.500)
RWplus		5.19 (0.575)	4.29 (0.608)	3.89 (0.625)	0.528 (−0.517)
SVR_CAF	3.245 (0.677)	**5.09 (0.582)**	5.46 (0.570)	5.39 (0.569)	**0.567 (−0.573)**

Bold signifies the best performance among the methods.

is the shortest distance between any nodes and lower *L* value implies the protein structure is in a compact form. Dokholyan *et al.*[52] had shown that protein folded conformations had higher degree and lower *L* values than unfold ones. Our hypothesis is that native structures (perfectly folded) and their decoys (incorrectly folded) should have different network properties. This difference can be developed to discriminate natives from decoy sets from a network point of view. Secondly, Wang and coworkers[72] presented a network-based combinatorial score function by using degree and characteristic path length in the application of protein–protein docking. Their results encouraged us to apply network-based score function to the selection of native structures from decoys.

3. We employed the method of Fourier transform, a basic tool for signals processing in mathematical science, to quantify the intrinsic correlation of hydrophobicity and SAS profiles. Fourier transform has been employed to study the influences of hydrophobicity on protein structures previously[73] and is by no means a new attempt to correlate hydrophobicity and solvent accessibility profiles. To discriminate native structures from decoys, our method follows the assumption that solvent accessibility area profiles of the amino acids from native protein structures should have more similarity to the hydrophobicity profiles of amino acid sequences than those from decoys.[57] We found that $Score_{FFT}$ for the native structures tended to be lower than that for decoys, thus supporting our hypothesis.

SVR_CAF score selected better quality decoys than other knowledge-based scores from Rosetta and I-TASSER decoy set in top 1 and top 10, though it is not the best for top-10 RMSD and top-10 TM-score. The significantly higher correlation coefficient between SVR_CAF score and RMSD/TM-score than other scores suggests the probability of using SVR_CAF in improving protein fold selection and structure prediction.

Some specific proteins

The SVR_CAF score was able to rank native proteins in top 10 from their corresponding decoys with an accuracy rate of 84.38% on Decoy'R'Us data set. Five of 32 target proteins were not ranked top 10 among the decoys. It is interesting to find that these five proteins share some common features. All of them are alpha proteins except that 4rxn is a small protein with 54 amino acids. It is relatively difficult to discriminate native structures for small proteins because they usually lack well-packed hydrophobic core.[12] 1fc2 is another challenge. It was not only missed by SVR_CAF score, but also missed by many other functions, such as RAPDF, KBP, DFIRE, and Rosetta. 1fc2 is Immunoglobulin Fc Fragment B of protein A, mainly consisting of two alpha helices. Fragment B binds to Immunoglobulin Fc through a hydrophobic contact, which stabilizes the complex. With an exposed hydrophobic region, Fragment B might not be as stable as decoys that have reduced exposed hydrophobic area.[74] For this reason, it is conceivable that 1fc2 is not a good test case for score functions. 1hdd is an engrailed homeodomain-DNA complex. RAPDF and KBP also failed to rank the native structure of 1hdd in top 10. 1bl0 is also a DNA-binding protein and in this case, SVR_CAF would score the native state of the isolated protein incorrectly due to the incompleteness of the assessed structure. 1jwe is determined by NMR spectroscopy, which is relatively less accurate as compared to X-ray structures, making it more difficult to identify the native structure.[75] These examples adequately reflect the complexity of predicting real proteins, where the size of the proteins, ligand binding, and protein stability is of utmost importance.

CONCLUSION

A novel integrated scoring function SVR_CAF is proposed in this article to discriminate native proteins among decoys. SVR_CAF is a machine learning score based on the integration of CE_score, AAN_score, and FFT_score. The results show that SVR_CAF ranks native structures in the top for Decoys'R'Us decoys and select the native-like structures from both Rosetta and I-TASSER decoy sets effectively. To the best of our knowledge, this is the first attempt to combine these three types of score functions to detect native structures among decoy sets. Compared with its individual score functions and other energy functions,

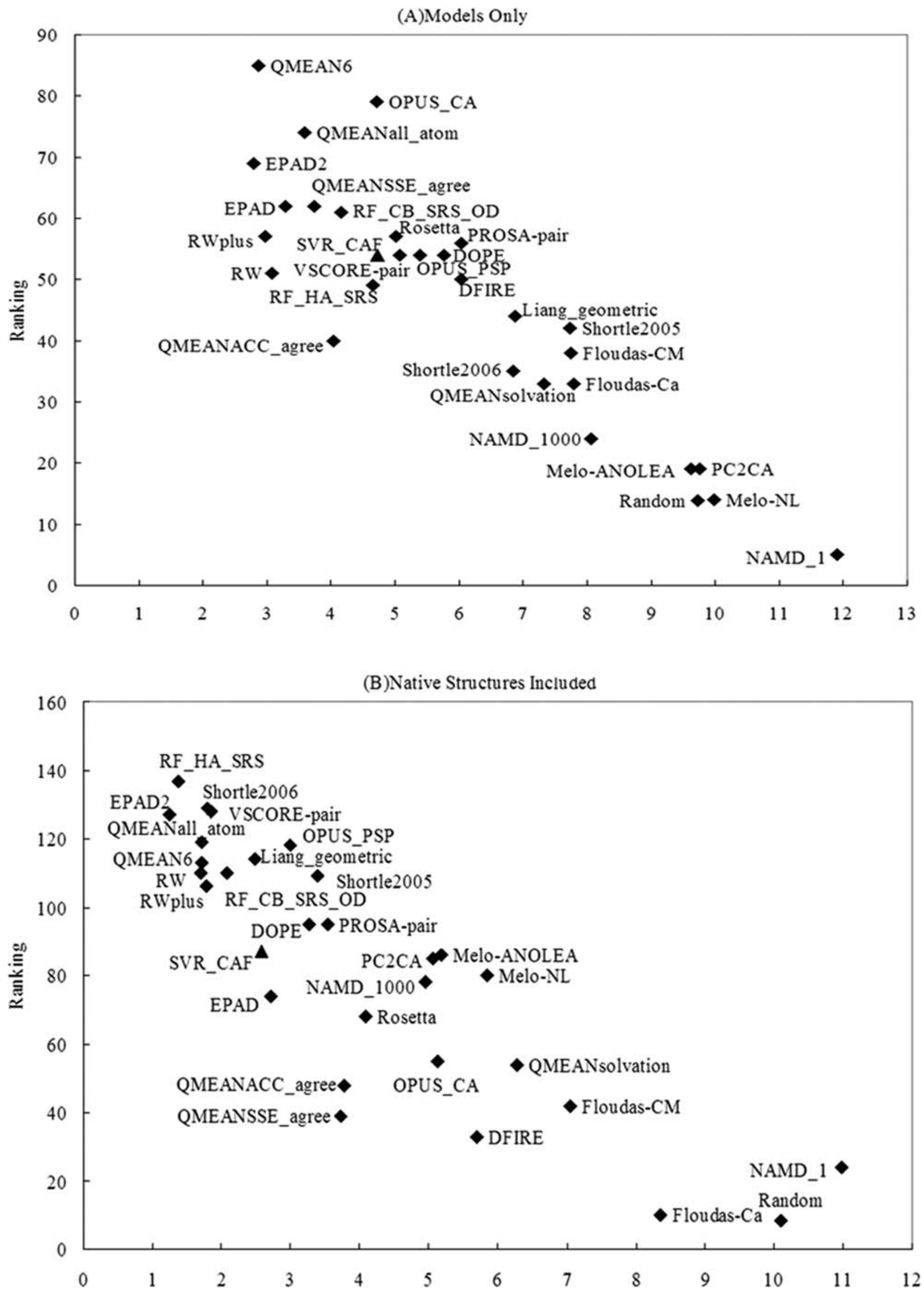

Figure 1

Performance of different potentials on the selected CASP5-8 models. The performance of SVR_CAF is shown in triangle and all the other score functions are shown in diamond. The data of other potentials were taken from Refs. 64, 65. (**A**) Models only: excluding the native structures in the sets; Ranking: the number of targets when the best decoy was ranked first. Average: the average rank of the best decoy according to GDT scores. (**B**) Native structures are included in the sets. Ranking: the number of targets when the native structures are ranked as first. Average: the average rank of native structures.

SVR_CAF discriminates native structures from decoys, not only based on energy considerations, but also including network topology characters, thus provides a novel tool for better discrimination of native and non-native structures.

For the select of the native structures from decoy sets, it should be better for researchers to combine both the score function method with the physical potential method together. Both methods can be complementary to each other, the potential based methods could be applied to all the data sets, but they are computational intensive and also the potential are not accurate, the score function methods could be fast but data specific. At present, most models are developed based on many high-quality decoys. It will become impossible to generate many high-quality decoys in real-world applications. Novel methods are required to extract important features from low-quality decoys for the accurate structure modeling. Although all the methods still need to be improved further, we expect our method can provide novel insight into the future development of the scoring functions for protein structure selection.

REFERENCES

1. Kolata G. Trying to crack the second half of the genetic code. Science 1986;233:1037–1039.
2. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294:93–96.
3. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
4. Shi X, Zhang J, He Z, Shang Y, Xu D. A sampling-based method for ranking protein structural models by integrating multiple scores and features. Curr Protein Pept Sci 2011;12:540–548.
5. Case DA, Cheatham TE, III, Darden T, Gohlke H, Luo R, Merz KM, Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.
6. Brooks BR, Brooks CL, III, Mackerell AD, Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem 2009;30:1545–1614.
7. Roterman IK, Gibson KD, Scheraga HA. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. I. Conformational predictions for the tandemly repeated peptide (Asn-Ala-Asn-Pro)9. J Biomol Struct Dyn 1989;7:391–419.
8. Huang ES, Samudrala R, Park BH. Scoring functions for ab initio protein structure prediction. Methods Mol Biol 2000;143:223–245.
9. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
10. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001;44: 223–232.
11. 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;11:2714–2726.
12. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
13. Mehdi Mirzaie, Sadeghi M. Knowledge-based potentials in protein fold recognition. J Paramed Sci 2010;1:63–73.
14. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
15. 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;55:1005–1013.
16. Chen C, Li L, Xiao Y. All-atom contact potential approach to protein thermostability analysis. Biopolymers 2007;85:28–37.
17. Arab S, Sadeghi M, Eslahchi C, Pezeshk H, Sheari A. A pairwise residue contact area-based mean force potential for discrimination of native protein structure. BMC Bioinformatics 2010;11:16.
18. Zhou H, Skolnick J. Protein model quality assessment prediction by combining fragment comparisons and a consensus C(alpha) contact potential. Proteins 2008;71:1211–1218.
19. 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;15:1653–1666.
20. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009;14:386–393.
21. Kihara D, Chen H, Yang YD. Quality assessment of protein structure models. Curr Protein Pept Sci 2009;10:216–228.
22. Hu G, Zhou J, Yan W, Chen J, Shen B. The topology and dynamics of protein complexes: insights from intra- molecular network theory. Curr Protein Pept Sci 2013;14:121–132.
23. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010;26:882–888.
24. Larsson P, Skwark MJ, Wallner B, Elofsson A. Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins 2009;77(Suppl 9):167–172.
25. Cheng J, Wang Z, Tegge AN, Eickholt J. Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 2009;77(Suppl 9):181–184.
26. Benkert P, Tosatto SC, Schwede T. Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. Proteins 2009;77(Suppl 9):173–180.
27. Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008;71: 261–277.
28. Benkert P, Kunzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res 2009;37(Web Server issue): W510–W514.
29. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
30. Qiu J, Sheffler W, Baker D, Noble WS. Ranking predicted protein structures with support vector regression. Proteins 2008;71:1175–1182.
31. Shirota M, Ishida T, Kinoshita K. Development of a new meta-score for protein structure prediction from seven all-atom distance dependent potentials using support vector regression. Genome Inform 2009;23:149–158.
32. Zhu F, Shen B. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. PLoS One 2012; 7:e39230.
33. Wang Z, Tegge AN, Cheng J. Evaluating the absolute quality of a single protein model using structural features and support vector machines. Proteins 2009;75:638–647.
34. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E 2002;65(6 Pt 1):061910.
35. Galan JF, Gao J, Pabuwal V, Meek PJ, Li Z-J. Application of network theory in understanding and predicting protein structure and function. Curr Proteomics 2008;5:181–190.
36. Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM, Prado-Prado F, Perez-Montoto LG, Concu R, Podda G, Shen B. Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, antiparasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 2010;11:379–406.
37. Yan W, Zhu H, Yang Y, Chen J, Zhang Y, Shen B. Effects of time point measurement on the reconstruction of gene regulatory networks. Molecules 2010;15:5354–5368.
38. K.Muppirala U, Li Z. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. Protein Eng Des Sel 2006;19:265–275.
39. Küçükural A. Discrimination of native folds using network properties of protein structure. BMC Syst Biol 2007;1(Suppl 1):49.
40. Vassura M, Margara L, Fariselli P, Casadio R. A graph theoretic approach to protein structure selection. Artif Intell Med 2008;7:dio: 10.1016.

41. Chatterjee S, Bhattacharyya M, Vishveshwara S. Network properties of protein-decoy structures. J Biomol Struct Dyn 2012;29:606–622.
42. Samudrala R, Levitt M. Decoys ‘R’ Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
43. 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.
44. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
45. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259–264.
46. Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. Proteins 2009;77(Suppl 9):100–113.
47. 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.
48. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics 2010;26:889–895.
49. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci U S A 2000;97:2550–2555.
50. Shen B, Vihinen M. RankViaContact: ranking and visualization of amino acid contacts. Bioinformatics 2003;19:2161–2162.
51. Yang Y, Chen B, Tan G, Vihinen M, Shen B. Structure-based prediction of the effects of a missense variant on protein stability. Amino Acids 2013;44:847–855.
52. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Topological determinants of protein folding. Proc Natl Acad Sci U S A 2002;99:8637–8641.
53. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A 1984;81:140–144.
54. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
55. Zamyatnin AA. Protein volume in solution. Prog Biophys Mol Biol 1972;24:107–123.
56. Chothia C. The nature of the accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–12.
57. Chen M, Liu B, Yan W, Shen B. Wavelet transform based protein decoy discrimination. IEEE 2009:1–4.
58. Harris CM. The Fourier analysis of biological transients. J Neurosci Methods 1998;83:15–34.
59. Muppirala UK, Li Z. A simple approach for protein structure discrimination based on the network pattern of conserved hydrophobic residues. Protein Eng Des Sel 2006;19:265–275.
60. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. Adv Protein Chem 1988;39:191–234.
61. Shen B, Vihinen M. Conservation and covariance in PH domain sequences: physicochemical profile and information theoretical analysis of XLA-causing mutations in the Btk PH domain. Protein Eng Des Sel 2004;17:267–276.
62. Silverman BD. Underlying hydrophobic sequence periodicity of protein tertiary structure. J Biomol Struct Dyn 2005;22:411–423.
63. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27.
64. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 2011;101:2043–2052.
65. Zhao F, Xu J. A position-specific distance-dependent statistical potential for protein structure and functional study. Structure 2012; 20:1118–1126.
66. Tian L, Wu A, Cao Y, Dong X, Hu Y, Jiang T. NCACO-score: an effective main-chain dependent scoring function for structure modeling. BMC Bioinformatics 2011;12:208.
67. Gniewek P, Leelananda SP, Kolinski A, Jernigan RL, Kloczkowski A. Multibody coarse-grained potentials for native structure recognition and quality assessment of protein models. Proteins 2011;79:1923–1929.
68. 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.
69. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinformatics 2010;11:128.
70. 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;256:623–644.
71. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
72. Gong X, Wang P, Yang F, Chang S, Liu B, He H, Cao L, Xu X, Li C, Chen W, Wang C. Protein-protein docking with binding site patch prediction and network-based terms enhanced combinatorial scoring. Proteins 2010;78:3150–3155.
73. Yahyanejad M, Burge CB, Kardar M. Untangling influences of hydrophobicity on protein sequences and structures. Proteins 2006; 62:1101–1106.
74. Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized born solvent model. Proteins 2002;48:404–422.
75. Lee MR, Kollman PA. Free-energy calculations highlight differences in accuracy between X-ray and NMR structures and add value to protein structure prediction. Structure 2001;9:905–916.