2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering

Empirical probability functions derived from dihedral angles for protein structure prediction

Qiwen Dong[1,2], Xin Geng[3], Shuigeng Zhou[1,2,*] and Jihong Guan[3]
[1]Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China.
[2]School of Computer Science, Fudan University, Shanghai, China.
[3]Dept. of Computer Science and Technology, Tongji University, Shanghai, China.

***Abstract*— The development and evaluation of functions for protein energetics is an important part of current research aiming at understanding protein structures and functions. Knowledge-based mean force potentials are derived from statistical analysis of interacting groups in experimentally determined protein structures. Current knowledge-based mean force potentials are based on the inverse Boltzmann's law, which calculate the ratio of the observed probability with respect to the probability of the reference state. In this study, a general probability framework is presented with the aim to develop novel energy scores. A class of empirical probability functions is derived by decomposing the joint probability of backbone dihedral angles and amino acid sequences. The neighboring interactions are modeled by conditional probabilities. Such probability functions are based on the strict probability theory and some suitable suppositions for convenience of computation. Experiments are performed on several well-constructed decoy sets and the results show that the empirical probability functions presented here outperform previous statistical potentials based on dihedral angles. Such probability functions will be helpful for protein structure prediction, model quality evaluation, transcription factors identification and other challenging problems in computational biology.**

***Index Terms*— knowledge-based potential, statistical potential, joint probability, conditional probability**

I. Introduction

The prediction of the three-dimensional structure of a protein from the amino acid sequence is one of the most important tasks in computational biology. A solution to this enigma is ever more necessary because of the huge increase in completely sequenced genomes [1], while the proteins with known structures are very limited [2]. Theoretical prediction methods are one possible way to fill in this gap. The basic energetic model commonly used to solve this problem is based on the Anfinsen's hypothesis [3], which states that for a given physiological set of conditions the native structure of a protein corresponds to the global Gibbs free energy minimum. A potential that can discriminate between the native and miss-folded structures is crucial for any protein structure prediction protocol to be fully successful.

Because of the intrinsic complexity of protein structures, only moderate progress has been achieved over the last few decades [4]. As an alternative way, one may take a simplified description for a protein structure, in which a protein structure is represented by the backbone dihedral angles and the side-chain dihedral angles while the bond lengths and bond angles are taken as the standard values. Regular patterns in dihedral angles are indicative of the protein's secondary structure such as α-helices and β-sheets. A way to visualize the allowed values of the dihedral angles and their populations is to generate Ramachandran plots, which show the correlations between the Φ, Ψ dihedral angles.

Many researchers have tried to predict the dihedral angles from the amino acid sequence with the ultimate goal of three-dimensional structure prediction. Zimmermann and Hansmann [5] used the support vector machine to predict the dihedral states of residues from sequences. Kuang *et al.* [6] adopted machine learning approaches for protein backbone dihedral angle prediction and provided more useful local structural information than the conventional secondary structure predictions. De Brevern *et al.* [7] proposed a 16-letter alphabet generated by a self-organizing map based on a dihedral angle similarity measure. The prediction accuracy of local three-dimensional structure has been steadily increased by taking sequence information and secondary structure information into consideration [8]. The chemical shifts information was used to improve the prediction accurate of dihedral angles [9].

The dihedral angles of proteins can be used to develop knowledge-based mean force potentials. The single-body residue-level dihedral potential [10] can achieve comparable or even better results in comparison with the two-body [11] or atom-level potentials [12]. The propensity of amino acid to the side-chain χ_1 angle has been incorporated to improve the performance of knowledge-based mean-force potentials [13, 14]. The normalized dihedral potential was used to evaluate the model qualities [15]. The higher-order Φ-Ψ pairs score was presented to evaluate the structural quality of protein models [16]. Shortle [17] discussed the difference between propensities and probabilities, and drew the conclusion that Boltzmann hypothesis may only be applicable for the calculation of statistical potentials after the starting conformation has been specified.

The Ramachandran plots has been extensively investigated [18–20]. When the residue is embedded in the polypeptide chain, its states may be correlated with those of the neighboring residues (local correlations) along the chain and those of distant residues along the chain (long-range correlations). The Flory isolated residue pair hypothesis [21] assumes that the conformation of one residue is independent of its neighbors' conformations. However, many studies has shown that such hypothesis is not valid [22]. The Ψ angle probabilities

*Correspondence author, E-mail: sgzhou@fudan.edu.cn

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.55

estimated from a database were shown to be context sensitive and position dependent [23]. The correlation between ω and Ψ dihedral angles has been analyzed by Esposito *et al.* [24]. The relationship between amino acid sequences and dihedral angles has been widely explored [25–27].

In this study, a class of empirical probability functions is derived by decomposing the joint probability of backbone dihedral angles and amino acid sequence. The neighboring interactions are modeled by conditional probabilities. Testing on several well-constructed decoy sets shows that the empirical probability functions presented here outperform previous statistical potentials based on dihedral angles. We also demonstrate that the probability functions are equivalent to the statistical potentials which are deduced from the inverse Boltzmann's law [28].

II. Materials and Methods

A. Datasets

The parameters are collected from the PDB25 dataset [29], which is a subset of the PDB database [2]. There is less than 25% sequence identity between any two proteins and each protein has a resolution better than 2.5 Å. The structures with missing atoms and chain breaks are excluded. The resulting dataset contains 2238 chains.

To evaluate the usefulness of the statistical potentials, large sets of protein structure models are needed. Several datasets including the HR dataset [30], the LKF dataset [31], the Decoys 'R'Us set [32] are used here to evaluate the probability functions. They are briefly described as follows.

The HR dataset is a high resolution decoy set developed by Rajgaria *et al.* [30], which contains 1400 proteins. Each protein has 500-1600 decoy structures. The decoy generation procedure ensures that the decoy structures preserve information about the distances within the hydrophobic core, so all decoy structures in this dataset has high quality with RMSD (Root Mean Square Deviation) less than 8 Å. A number of stringent filters are performed on the dataset. The proteins that miss some of amino acids are removed. The decoy structures that contain incomplete atoms especially main-chain atoms are also discarded. The resulting dataset contain 1137 proteins and 500 decoy structures are randomly selected for each protein.

The LKF dataset is generated by Loose *et al.* [31] using the program of DYANA [33], which takes as input the sequence of a protein, along with information about its secondary structure that gives bounds for the distances and torsion angles between atoms. Several operations are conducted to ensure the quality of the decoy set. The Decoys of 185 proteins are downloaded from the authors' website and each of them has about 200 decoy structures. 23 of the 185 proteins containing incomplete atom are removed from the decoy set. Totally, there are 161 proteins and 32047 decoy structures.

The Decoys 'R'Us dataset [32] contains several well-constructed decoy subsets and has been widely used to evaluate various potentials [34–37]. Three decoy subsets including 4stat_reduced [38], lattice-ssfit [39, 40] and semfold [41] are selected. These sets were generated by various researchers using different techniques and each set contains a varying number of decoy conformations. Because this decoy set contains multiple decoy subsets generated by different methods, it can give a comprehensive evaluation of the potentials.

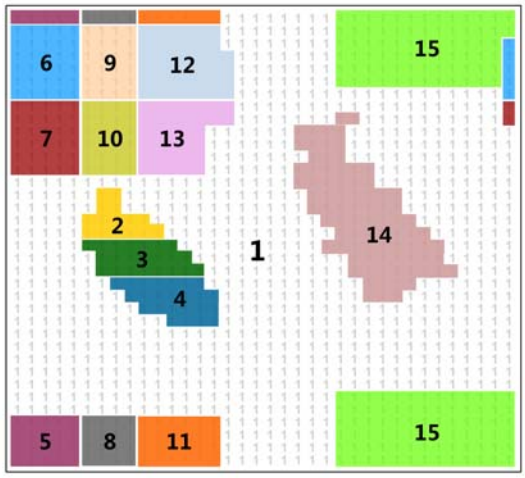

Fig. 1. Discrete representation of backbone dihedral angles. The Ramachandran plot is partitioned into 10 degree by 10 degree grids and 15 classes are defined.

The statistical information of the datasets is listed in the Table I.

B. Protein dihedral angle conformational states

The dihedral angles need to be discretized so that the three-dimensional structures of proteins can be converted into sequences of one-dimensional structural alphabets. The partition of the Ramachandran plot presented by Shortle *et al.* [13] is adopted here. First the Ramachandran plot is partition with 10° interval, which results in 1296 grids for each residue. Then these grids are clustered to 15 conformational states representing the main structure types such as helix, sheet, coil, etc. These 15 conformational states make up of a structural alphabet and shown in Fig.1. The side-chain χ_1 angle is divided into three states, centering at 180°, 60°, -60°, corresponding to trans, gauche+, gauche- states respectively.

C. The probability formalism

1) The problem description by probability theory: The aim of this study is to select the native conformation from a set of incorrect conformations. Let $A = (a_1, a_2, ..., a_n)$ be a amino acid sequence with length of n, and $SD = \{S^1, S^2, ..., S^m\}$ be a set of conformations containing both the native state and the incorrect conformations for the sequence A. For a particular structure $S \in SD$, let $S = (S_1, S_2, ..., S_n)$ be the corresponding structural alphabet sequence. The above mentioned problem becomes to select the optimal structure $S, S \in SD$, which has the best match to sequence A:

$$\Gamma(A, SD) \stackrel{def}{=} \underset{S \in SD}{argmax} P(S|A) \quad (1)$$

TABLE I
SUMMARY OF THE DECOY SETS

Dataset	Number of proteins	Lengths	Number of decoy structures per protein	RMSD(Å)[a]
HR	1137	41-200	500	0.435-7.995 (2.590±0.801)
LKF	161	32-147	52-200	1.463-43.405 (10.061±2.979)
4state_reduced	7	54-75	629-685	0.805-9.391 (5.209±1.687)
lattice_ssfit	7	55-98	1994-2000	4.742-15.608(9.888±1.253)
semfold	6	56-95	10000-21080	2.982-15.065 (10.357±1.665)

[a] RMSD is the Ca Root Mean Square Deviation. Given in the bracket are the average and the standard deviation.

By Bayes' theorem, it can be written as:

$$\begin{aligned}\Gamma(A,SD) &= \underset{S\in SD}{argmax}\frac{P(A,S)}{P(A)} = \underset{S\in SD}{argmax}P(A,S) \\ &= \underset{S\in SD}{argmax} \quad P(a_1,s_1,...,a_n,s_n)\end{aligned} \tag{2}$$

Since all the conformations have the same sequence A, the probability $P(A)$ is const and dropped in the above equation.

Next, the joint probability $P(A,S)$ need to be broken into "bite-size" pieces about which we can collect statistics. This can be done in two ways. The first is like this:

$$\begin{aligned}P(a_1,s_1,...,a_n,s_n) =& P(a_1)P(s_1|a_1) \\ & P(a_2|a_1,s_1)P(s_2|a_{1,2},s_1)... \\ & P(a_n|a_{1,n-1},s_{1,n-1})P(s_n|a_{1,n},s_{1,n-1}) \\ =& P(a_1)P(s_1|a_1) \\ & \prod_{i=2}^{n}P(a_i|a_{1,i-1},s_{1,i-1})P(s_i|a_{1,i},s_{1,i-1}) \\ =& \prod_{i=1}^{n}P(a_i|a_{1,i-1},s_{1,i-1})P(s_i|a_{1,i},s_{1,i-1})\end{aligned} \tag{3}$$

where the sequence like $a_1,a_2,...,a_i$ is simplified as $a_{1,i}$ During the derivation,the equation has been simplified by suitable definition of the terms like $a_{1,0}$ and their probabilities. The equation 3 is derived by first breaking out $P(a_1)$ from the joint probability $P(a_1,s_1,...,a_n,s_n)$. In a similar way, the probability $P(s_1)$ can be first broken out and the following equation can be obtained:

$$P(a_1,s_1,...,a_n,s_n) = \prod_{i=1}^{n}P(s_i|a_{1,i-1},s_{1,i-1})P(a_i|a_{1,i-1},s_{1,i}) \tag{4}$$

So far, no assumption has been made about the probability, so the equation 3 and 4 are correct in theory. In practice, the conditional probability cannot be calculated directly since it has too many conditions. Some suitable hypotheses should be made so that the conditional probability can be empirically evaluated. In following, the Shortle statistical potential is first introduced and then by suitable hypotheses, some calculable probabilities are presented.

2) The Shortle statistical potential: The Shortle statistical potential [13] can be expressed as:

$$P = \prod_{i=1}^{n}\frac{P(s_i|a_i)}{P(s_i)} \tag{5}$$

Here we deduce the above equation by probability theory. The Shortle statistical potential tries to maximize the probability $P(A|S)$. By Bayes' theorem, it can be written as:

$$P(A|S) = \frac{P(A,S)}{P(S)} = \frac{P(a_1,s_1,...a_n,s_n)}{P(s_1,...s_n)} \tag{6}$$

Next, the following two hypotheses are used:

Hypothesis 1: The occurrence of amino acid is independent of the structure S.

$$P(a_i|a_{1,i-1},s_{1,i-1}) = P(a_i|a_{1,i-1}) \tag{7}$$

Hypothesis 2: The occurrence of each conformational state s_i is only dependent on the current amino acid a_i and independent of any neighboring amino acid and any conformational states on the sequence.

$$P(s_i|a_{1,i},s_{1,i-1}) = P(s_i|a_i) \tag{8}$$

Applying the two hypotheses, the equation 6 can be calculated as:

$$P(A|S) = \prod_{i=1}^{n}\frac{P(s_i|a_i)}{P(s_i)} \tag{9}$$

in which the probability of amino acid $P(a_i)$ is ignored since it is const for all conformations. The conditional probability can be counted on a dataset by the maximum likelihood estimation:

$$P(s_i|a_i) = \frac{C(s_i,a_i)}{C(a_i)} \tag{10}$$

where $C(s_i,a_i)$ is the number of times amino acid a_i has the conformational state s_i and $C(a_i)$ is the total number of occurrence for amino acid a_i. This calculation means is denoted as Probability_Shortle. By taking negative logarithm on both sides of equation 9, the Probability_Shortle is the same as the statistical potentials derived by the inverse Boltzmann law, where the reference state is taken as the average probability of conformational state s_i over all amino acids.

3) The simplest model: The Shortle potential is obtained by maximizing the probability $P(A|S)$. However, for the task of selecting native structures, it may be suitable to calculate the probability $P(S|A)$. Again by applying the above two hypotheses to equation 3, the following probability is obtained:

$$P(S|A) = \prod_{i=1}^{n}P(s_i|a_i) \tag{11}$$

This calculation means is denoted as Probability_I. It has a very simple interpretation. For each amino acid a_i, the native

structure will adopt the most common conformational states s_i. By taking negative logarithm, the Probability_I is also the same as the statistical potentials derived by the inverse Boltzmann law, where no reference state is taken.

4) The standard model: The above two calculation means do not take any context into consideration. In protein structures, the conformational states may be influenced by other conformational states or other amino acids. In what follows, other models are developed which take the neighboring interaction into consideration.

Hypothesis 3: The occurrence of each conformational state s_i is dependent on the current amino acid a_i, the anterior amino acid a_{i-1} and its conformational state s_{i-1}.

$$P(s_i|a_{1,i}, s_{1,i-1}) = P(s_i|a_{i-1}, a_i, s_{i-1}) \tag{12}$$

Substituting hypothesis 1 and hypothesis 3 into equation 3, the following equation can be obtained:

$$P(S|A) = \prod_{i=1}^{n} p(s_i|a_{i-1}, a_i, s_{i-1}) \tag{13}$$

where the const probability for $P(a_i)$ is dropped. This calculation means is denoted as Probability_II.

The hypothesis 1 may not be valid for proteins, it can be revised as:

Hypothesis 4: The occurrence of each amino acid is dependent on the the anterior amino acid a_{i-1} and its conformational state s_{i-1}:

$$P(a_i|a_{1,i-1}, s_{1,i-1}) = P(a_i|a_{i-1}, s_{i-1}) \tag{14}$$

Substituting hypothesis 3 and hypothesis 4 into equation 3 or 4, we get:

$$P(S|A) = \prod_{i=1}^{n} p(a_i|a_{i-1}, s_{i-1}) p(s_i|a_{i-1}, a_i, s_{i-1}) \tag{15}$$

This calculation means is denoted as Probability_III.

5) The multi-conditions model: When the conditions are sufficient, they will give accurate description about the probability of the conformational states or the amino acids. However, too many conditions will lead to a large number of parameters to be estimated, the conditional probability will be suffered from the data sparse problem. To overcome this problem, the following hypothesis is made, in which the probability with multiple conditions can be modeled by the weight sum of each individual probability.

Hypothesis 5: The occurrence of each conformational state is dependent on the anterior amino acid a_{i-1}, the current amino acid a_i, the posterior amino acid a_{i+1}, the anterior state s_{i-1}, the posterior state s_{i+1}, the anterior side-chain sc_{i-1}, the current side-chain sc_i and the posterior side-chain sc_{i+1}. This probability with multiple conditions can be calculated by the weight sum of each individual probability.

$$\begin{aligned} &P(s_i|multi-conditions) = \\ &w_1 P(s_i|a_{i-1}) + w_2 P(s_i|a_i) + w_3 P(s_i|a_{i+1}) \\ &+ w_4 P(s_i|s_{i-1}) + w_5 P(s_i|s_{i+1}) \\ &+ w_6 P(s_i|sc_{i-1}) + w_7 P(s_i|sc_i) + w_8 P(s_i|sc_{i+1}) \end{aligned} \tag{16}$$

where

$$\sum_{j=1}^{8} w_i = 1 \tag{17}$$

Substituting hypothesis 1 and hypothesis 5 into equation 3 or 4, we get:

$$P(S|A) = \prod_{i=1}^{n} P(s_i|multi-conditions) \tag{18}$$

This calculation means is denoted as Probability_IV.

D. Performance metrics

The potential functions are evaluated using four criteria as follows:

1) The fraction of the proteins of which the native structure is ranked as one on a decoy set (Accuracy).

2) The rank of the native structure in the list of decoys sorted by the probability (Rank).

3) The Z-score of the native structure in the decoys set, defined as

$$\text{Z-Score} = \frac{|< E^{decoy} > - E^{native}|}{\sqrt{< (E^{decoy})^2 > - < E^{decoy} >^2}} \tag{19}$$

where $<>$ denotes the average over all decoy structures, and E^{native} is the probability of the native structure. Z-score is a measure of the bias toward the native structure.

4) The Pearson correlation coefficient (CC) between probability(x) and RMSD(y), defined as:

$$\text{CC} = \frac{\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{M}(x_i - \overline{x})^2][\sum_{i=1}^{M}(y_i - \overline{y})^2]}} \tag{20}$$

III. Results and discussions

The conditional probabilities are evaluated on the PDB25 dataset and tested on several well-constructed decoy sets. The Probability_Shortle is used as the baseline. All the computation means are converted into logarithm form to avoid overflow, so the product of multiple probabilities are replaced by the summation of the logarithms of the corresponding probabilities.

A. Performance on the HR dataset

The performance on the HR dataset is listed in Table II. The Probability_Shortle gets the lowest performance with accuracy only 64.2%. As mentioned above, such means calculates the probability $P(A|S)$ in order to obtain the similar reference state with the statistical potentials derived from the inverse Boltzmann's law. The simplest model (Probability_I) gives a relatively improved results and obtains an accuracy of 67.2%. The standard models (Probability_II and Probability_III) do not reach the expected results with accuracy only 65.3% and 67.5% respectively. This may be caused by the data sparse problem. The PDB25 dataset contains 2328 chains and 509, 083 amino acids. The number of conditional probability used by Probability_II is 90, 000 ($15 \times 15 \times 20 \times 20$). The calculation of Probability_III needs additional 6, 000 probabilities ($15 \times 20 \times 20$). There are only about 6 samples to count

TABLE II

COMPARATIVE PERFORMANCE ON THE HR DATASET

Probabilites	Accuracy	Rank	Z	CC
Probability_Shortle	0.642	27.424	3.500	0.015
Probability_I	0.672	34.426	4.281	0.058
Probability_II	0.653	45.1	5.022	0.048
Probability_III	0.675	41.647	5.258	0.038
Probability_IV	**0.865**	**12.945**	**6.480**	**0.151**

TABLE III

COMPARATIVE PERFORMANCE ON THE LKF DATASET

Probabilites	Accuracy	Rank	Z	CC
Probability_Shortle	0.689	8.075	3.448	0.042
Probability_I	0.839	9.988	6.457	**0.223**
Probability_II	0.839	7.758	6.395	0.198
Probability_III	0.839	7.665	**6.533**	0.201
Probability_IV	**0.845**	**6.727**	4.970	0.171

one conditional probability, so it is not sufficient to give an accurate estimation. Increasing the size of the training set will be temporarily solve such problem, however it is important to ensure that the dataset is non-redundant.

The Probability_IV adopts another strategy to void the data sparse problem, in which the probability with multiple conditions can be modeled by the weight sum of each individual probability. The conformational state may be influenced by many conditions. The information theory is used to determine whether the conformational state is correlated with the condition. The entropy of the current conformational state s_i is first calculated by

$$H(s_i) = -\sum p(s_i) \ln p(s_i) \tag{21}$$

and then the conditional entropy is calculated as follows:

$$H(s_i|C) = -\sum_{s_i} \sum_{C} p(s_i, C) \ln p(s_i|C) \tag{22}$$

When the conditional entropy is decreased, it means that the conformational state s_i is correlated with the condition C. Eight conditions are selected. They are the anterior amino acid a_{i-1}, the current amino acid a_i, the posterior amino acid a_{i+1}, the anterior state s_{i-1}, the posterior state s_{i+1}, the anterior side-chain sc_{i-1}, the current side-chain sc_i and the posterior side-chain sc_{i+1}. The HR dataset is randomly split into the training set and the test set with the split ratio close to 1:1. The entropies and the weights are calculated and optimized on the training set and then the optimal parameters are used to test the Probability_IV on the test set.

The entropy of the current conformational state s_i is 2.025 bits. The maximum entropy of the conformational state s_i is 2.708 bits, so the conformational state s_i is not randomly distributed. The conditional entropies of the eight conditions are 2.002, 1.914, 1.970, 1.667, 1.664, 2.009, 1.925, 2.003 bits respectively. All the conditional entropies are more or less decreased, which indicates that all the conditions can influence the conformational state s_i. The genetic algorithm is then used to optimize weights with accuracy as the objective function. The optimal weights are 0.0449, 0.2121, 0.0697, 0.0521, 0.2402, 0.2156, 0.1612, 0.0040 respectively. The test result of Probabilit_IV is given at the last row of Table II. As can be seen, the Probability_IV gets an accuracy of 86.5%, which is a significant improvement in comparison with other calculation means. At the same time, a low rank and a high Z-Score are obtained.

TABLE IV

COMPARATIVE PERFORMANCE ON THE DECOYS 'R'US DATASET

Probabilites	4stat_reduced	lattice-sfit	semfold
Probability_Shortle	5/7 (3.245)[a]	4/7 (3.433)	0/6 (1.448)
Probability_I	**7/7** (4.096)	5/7 (5.258)	1/6 (2.001)
Probability_II	**7/7** (4.980)	**7/7**(7.772)	3/6 (3.324)
Probability_III	**7/7 (5.042)**	**7/7 (8.096)**	3/6 (3.610)
Probability_IV	**7/7** (3.504)	6/7 (6.115)	**4/6 (4.063)**

[a] Given are the number of correctly identified proteins in the total number of proteins and the Z-Scores (in bracket)

B. Performance on the LKF dataset

The test results on the LKF dataset are listed in Table III. The Probability_Shortle still gets the lowest performance with accuracy of 68.9%. The simplest model and the standard model achieve the same accuracy of 83.9% but different Ranks and Z-Scores. The weights used by Probability_IV do not optimized on this dataset, which take the same values as used by the HR dataset. This gives an open test on the LKF dataset. The performance of Probability_IV is better than that of other calculation means, which shows the robustness of Probability_IV.

C. Performance on the Decoys 'R'Us dataset

Three subsets are selected to test the performances of different calculation means and the test results are given in Table IV. The weights used by Probability_IV do not re-optimized, so the test on the Decoys 'R'Us dataset is still an open test. Testing on the 4stat_reduced subset seems to be an easy task. Except for Probability_Shortle, all other probabilities get perfect discrimination of the native conformations from the decoy structures. The standard model also gets perfect discrimination on the lattice-ssfit subset. On the semfold subsets, the Probability_IV achieves the best performance.

IV. CONCLUSION

In this study, a simplified representation of protein three-dimensional structure is adopted, from which protein structure can be converted into a one-dimensional sequence. A general probability framework is presented to model the relationships between amino acid sequences and their structural alphabet sequences. A class of empirical probability functions is derived by decomposing the joint probability of backbone dihedral angles and amino acid sequences. Such probability functions are based on strict probability theory and some suitable suppositions for convenience of computation. Experimental results

show that the empirical probability functions presented here outperform previous statistical potentials based on dihedral angles. Future work will aim at employing the smoothing algorithms to solve the data-sparse problem and exploring other novel and effective probability functions.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (NSFC) (No. 60873040 and No. 60873070), Shanghai Leading Academic Discipline Project (No. B114) and China Postdoctoral Science Foundation (No. 20080440572). Jihong Guan was also supported by the Program for New Century Excellent Talents in University of China (NCET-06-0376).

REFERENCES

[1] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The universal protein resource (uniprot): an expanding universe of protein information," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D187–91, 2006.

[2] A. Kouranov, L. Xie, J. de la Cruz, L. Chen, J. Westbrook, P. E. Bourne, and H. M. Berman, "The rcsb pdb information portal for structural genomics," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D302–305, 2006.

[3] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 96, pp. 223–30, 1973.

[4] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, and A. Tramontano, "Critical assessment of methods of protein structure prediction-round vii," *Proteins*, vol. 69 Suppl 8, pp. 3–9, 2007.

[5] O. Zimmermann and U. H. Hansmann, "Support vector machines for prediction of dihedral angle regions," *Bioinformatics*, vol. 22, no. 24, pp. 3009–15, 2006.

[6] R. Kuang, C. S. Leslie, and A. S. Yang, "Protein backbone angle prediction with machine learning approaches," *Bioinformatics*, vol. 20, no. 10, pp. 1612–21, 2004.

[7] A. G. de Brevern, C. Etchebest, and S. Hazout, "Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks," *Proteins*, vol. 41, no. 3, pp. 271–87, 2000.

[8] C. Etchebest, C. Benros, S. Hazout, and A. G. de Brevern, "A structural alphabet for local protein structures: Improved prediction methods," *Proteins*, vol. 59, no. 4, pp. 810–827, 2005.

[9] S. Neal, M. Berjanskii, H. Zhang, and D. S. Wishart, "Accurate prediction of protein torsion angles using chemical shifts and sequence homology," *Magn Reson Chem*, vol. 44 Spec No, pp. S158–67, 2006.

[10] H. Zhou and Y. Zhou, "Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition," *Proteins*, vol. 55, no. 4, pp. 1005–13, 2004.

[11] C. Zhang, S. Liu, H. Zhou, and Y. Zhou, "An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state," *Protein Sci*, vol. 13, no. 2, pp. 400–11, 2004.

[12] J. Qiu and R. Elber, "Atomically detailed potentials to recognize native and approximate protein structures," *Proteins*, vol. 61, no. 1, pp. 44–55, 2005.

[13] D. Shortle, "Composites of local structure propensities: evidence for local encoding of long-range structure," *Protein Sci*, vol. 11, no. 1, pp. 18–26, 2002.

[14] Q. Fang and D. Shortle, "A consistent set of statistical potentials for quantifying local side-chain and backbone interactions," *Proteins*, vol. 60, no. 1, pp. 90–6, 2005.

[15] S. C. Tosatto and R. Battistutta, "Tap score: torsion angle propensity normalization applied to local protein structure evaluation," *BMC Bioinformatics*, vol. 8, p. 155, 2007.

[16] G. E. Sims and S. H. Kim, "A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring," *Proc Natl Acad Sci U S A*, vol. 103, no. 12, pp. 4428–32, 2006.

[17] D. Shortle, "Propensities, probabilities, and the boltzmann hypothesis," *Protein Sci*, vol. 12, no. 6, pp. 1298–302, 2003.

[18] G. J. Kleywegt and T. A. Jones, "Phi/psi-chology: Ramachandran revisited," *Structure*, vol. 4, no. 12, pp. 1395–400, 1996.

[19] D. Pal and P. Chakrabarti, "On residues in the disallowed region of the ramachandran map," *Biopolymers*, vol. 63, no. 3, pp. 195–206, 2002.

[20] B. K. Ho, A. Thomas, and R. Brasseur, "Revisiting the ramachandran plot: hard-sphere repulsion, electrostatics, and h-bonding in the alpha-helix," *Protein Sci*, vol. 12, no. 11, pp. 2508–22, 2003.

[21] P. J. Flory, *Statistical Mechanics of Chain Molecules*. New York.: Wiley, 1969.

[22] R. V. Pappu, R. Srinivasan, and G. D. Rose, "The flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding," *Proc Natl Acad Sci U S A*, vol. 97, no. 23, pp. 12565–70, 2000.

[23] L. Ormeci, A. Gursoy, G. Tunca, and B. Erman, "Computational basis of knowledge-based conformational probabilities derived from local- and long-range interactions in proteins," *Proteins*, vol. 66, no. 1, pp. 29–40, 2006.

[24] L. Esposito, A. De Simone, A. Zagari, and L. Vitagliano, "Correlation between omega and psi dihedral angles in protein structures," *J Mol Biol*, vol. 347, no. 3, pp. 483–7, 2005.

[25] A. Pertsemlidis, J. Zelinka, r. Fondon, J. W., R. K. Henderson, and Z. Otwinowski, "Bayesian statistical studies of the ramachandran distribution," *Stat Appl Genet Mol Biol*, vol. 4, p. Article35, 2005.

[26] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick, "Statistical coil model of the unfolded state: resolving the reconciliation problem," *Proc Natl Acad Sci U S A*, vol. 102, no. 37, pp. 13099–104, 2005.

[27] D. B. Dahl, Z. Bohannan, Q. Mo, M. Vannucci, and J. Tsai, "Assessing side-chain perturbations of the protein

backbone: a knowledge-based classification of residue ramachandran space," *J Mol Biol*, vol. 378, no. 3, pp. 749–58, 2008.

[28] M. J. Sippl, "Boltzmann's principle, knowledge-based mean fields and protein folding. an approach to the computational determination of protein structures," *J Comput Aided Mol Des*, vol. 7, no. 4, pp. 473–501, 1993.

[29] Q. Dong, X. Wang, L. Lin, and Y. Wang, "Analysis and prediction of protein local structure based on structure alphabets," *Proteins*, vol. 72, no. 1, pp. 163–72, 2008.

[30] R. Rajgaria, S. R. McAllister, and C. A. Floudas, "A novel high resolution calpha–calpha distance dependent force field based on a high quality decoy set," *Proteins*, vol. 65, no. 3, pp. 726–41, 2006.

[31] C. Loose, J. L. Klepeis, and C. A. Floudas, "A new pairwise folding potential based on improved decoy generation and side-chain packing," *Proteins*, vol. 54, no. 2, pp. 303–14, 2004.

[32] R. Samudrala and M. Levitt, "Decoys 'r' us: a database of incorrect conformations to improve protein structure prediction," *Protein Sci*, vol. 9, no. 7, pp. 1399–401, 2000.

[33] P. Guntert, C. Mumenthaler, and K. Wuthrich, "Torsion angle dynamics for nmr structure calculation with the new program dyana," *J Mol Biol*, vol. 273, no. 1, pp. 283–98, 1997.

[34] M. Lu, A. D. Dousis, and J. Ma, "Opus-psp: an orientation-dependent statistical all-atom potential derived from side-chain packing," *J Mol Biol*, vol. 376, no. 1, pp. 288–301, 2008.

[35] R. Rajgaria, S. R. McAllister, and C. A. Floudas, "Distance dependent centroid to centroid force fields using high resolution decoys," *Proteins*, vol. 70, no. 3, pp. 950–70, 2008.

[36] P. Benkert, S. C. Tosatto, and D. Schomburg, "Qmean: A comprehensive scoring function for model quality assessment," *Proteins*, vol. 71, no. 1, pp. 261–77, 2008.

[37] Y. Wu, M. Lu, M. Chen, J. Li, and J. Ma, "Opus-ca: a knowledge-based potential function requiring only calpha positions," *Protein Sci*, vol. 16, no. 7, pp. 1449–63, 2007.

[38] B. Park and M. Levitt, "Energy functions that discriminate x-ray and near native folds from well-constructed decoys," *J Mol Biol*, vol. 258, no. 2, pp. 367–92, 1996.

[39] R. Samudrala, Y. Xia, M. Levitt, and E. S. Huang, "A combined approach for ab initio construction of low resolution protein tertiary structures from sequence," *Pac Symp Biocomput*, pp. 505–16, 1999.

[40] Y. Xia, E. S. Huang, M. Levitt, and R. Samudrala, "Ab initio construction of protein tertiary structures using a hierarchical approach," *J Mol Biol*, vol. 300, no. 1, pp. 171–85, 2000.

[41] R. Samudrala and M. Levitt, "A comprehensive analysis of 40 blind protein structure predictions," *BMC Struct Biol*, vol. 2, p. 3, 2002.