POTENTIAL FUNCTION OF SIMPLIFIED PROTEIN MODELS FOR DISCRIMINATING NATIVE PROTEINS FROM DECOYS: COMBINING CONTACT INTERACTION AND LOCAL SEQUENCE-DEPENDENT GEOMETRY

Jinfeng Zhang[1], Rong Chen[1,2,3], Jie Liang[1]

[1]Department of Bioengineering, University of Illinois at Chicago, Chicago,IL
[2]Department of Information & Decision Science, University of Illinois at Chicago, Chicago,IL
[3]Department of Business Statistics & Econometrics, Peking University, Beijing, P.R. China

Abstract—**An effective potential function is critical for protein structure prediction and folding simulation. For simplified models of proteins where coordinates of only C_α atoms need to be specified, an accurate potential function is important. Such a simplified model is essential for efficient search of conformational space. In this work, we present a formulation of potential function for simplified representations of protein structures. It is based on the combination of descriptors derived from residue-residue contact and sequence-dependent local geometry. The optimal weight coefficients for contact and local geometry is obtained through optimization by maximizing margins among native and decoy structures. The latter are generated by chain growth and by gapless threading. The performance of the potential function in blind test of discriminating native protein structures from decoys is evaluated using several benchmark decoy sets. This potential function have comparable or better performance than several residue-based potential functions that require in addition coordinates of side chain centers or coordinates of all side chain atoms.**

Key words—**Decoy discrimination, potential function, protein structure prediction, simplified protein models.**

1. INTRODUCTION

Studies of protein folding are often based on the thermodynamic hypothesis, which postulates that the native state of a protein is the state of lowest free energy under physiological conditions. Based on this assumption, a potential function where the native protein has the lowest energy is essential for protein structure prediction, folding simulation, and protein design.

There are two steps to construct such a potential function. The first step is to define a proper representation of protein structures, which is usually based on a set of numerical descriptors characterizing the structure and sequence of the

protein. The second step is to decide on a functional form, which takes the descriptor vector and maps it to a real valued energy or score for the particular structure. Frequently, potential function $H(.)$ takes the form of $H(f(s, a)) = H(c) = w \cdot c$, where the structure s of a protein and its amino acid sequence a is mapped to a numerical d-vector c by the representation f, such that $f : (s, a) \mapsto c \in \mathbb{R}^d$. The potential function $H(c) \in \mathbb{R}$ then maps the vector c to a real valued energy [1]. Protein representations strongly influence the effectiveness of potential function. The most successful potential functions rely on detailed all-atom representation of native protein structures with tens or hundreds thousand of parameters. Most residue-level potential functions with significant less parameters still need atom-level information. However, simplified representation of proteins are essential for structure prediction and folding simulation. Since not all heavy atoms are explicitly represented, sampling of conformational space is far more efficient. A challenging open problem is whether potential function based on a simplified protein representation can have perfect discrimination of native or near native conformations from decoy conformations. If so, it is important to identify the minimal representation through which such discrimination can be achieved.

In this work, we present a formulation of potential function designed for simplified n-state representations of protein structures at residue level. In addition to contact interactions, we encode sequence-dependent local geometric information. These are combined to form a potential function called *Contact and Local Geometric Potential* (CLGP). Our paper is organized as following: first we introduce the simplified n-state representation of protein structures. We then describe the scoring function, which is a weighted linear combination of contact and geometric parameters. We also discuss the optimization method to derive the parameters of the potential function. Finally, we show the performance of CLGP on both Park-Levitt decoy set [2] and decoys generated by gapless threading.
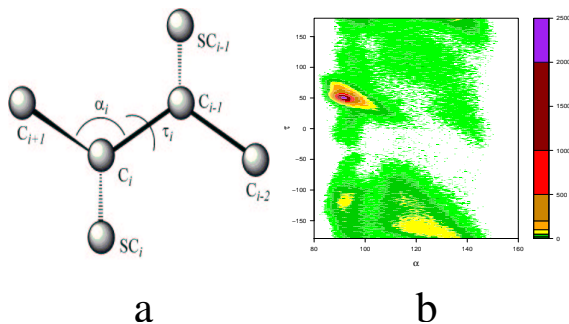
Fig. 1. a. Discrete state model. α_i and τ_i are shown for residue i. Bond angle α_i at position i is formed by $C_{\alpha,i-1}$, $C_{\alpha,i}$, and $C_{\alpha,i+1}$. Torsion angle τ_i is the dihedral angle of the two planes formed by atoms (C_{i-2}, C_{i-1}, C_i) and (C_{i-1}, C_i, C_{i+1}); b. Distribution of α and τ angles.

Fig. 2. a. Average RMSD of the best discrete state models fitted to PDB structures of 348 proteins from 3 to 10 state; b. Hierarchical clustering of amino acids by local sequence geometry propensity.

2. MODELS AND METHODS

REPRESENTATION OF PROTEIN STRUCTURES. Our representation of protein structures is an off-lattice discrete state model [3]. Specifically, we represent all backbone atoms by the C_α atom, and all side chain atoms by one additional atom attached to the main chain C_α atoms (Figure 1a). There are totally 20 different atom types (1 C_α and 19 side chain atoms). The backbone structure is described by the bond angle α and torsion angle τ at each C_α position (Figure 1a). The overall three dimensional structure of a protein of length N is completely determined by the set of the bond angles and torsion angles $\{(\alpha_i, \tau_i)\}$, $1 < i < N$, where i represents the i-th position of the backbone. The distance between each C_α atom and its side chain atom is predefined for each residue type.

In discrete state model, except residues at the two ends each main chain position can take n discrete pairs of (α, τ) angles, which is the states at this position. To determine the optimal number of states and the exact values of the (α, τ) angles for the discrete representation, we examine the distribution of α and τ angles calculated from 980 non-homologous proteins from PDBSELECT. Analogous to the Ramachandran plot, the distribution of (α, τ) also has densely and sparsely populated regions, which corresponds to different secondary structure types (Fig 1b). And the distribution differs for different amino acids. To obtain exact values for each discrete states, we use k-mean clustering to group α and τ values observed in native proteins into k from 3 to 10 clusters for each amino acid residue type. The cluster centers are then taken as the discrete states.

The discrete state representation is associated with discrete conformational space that is different from the continuous space of PDB structures. To represent PDB structure in the simplified discrete space, we need to map an PDB struc-
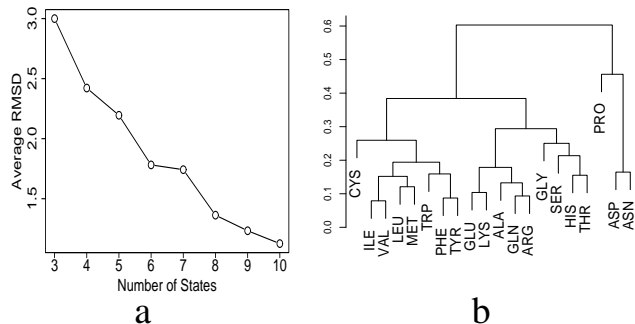
ture to a structure in the discrete space, with the requirement that these two structures are most similar by some criteria. In this study we consider two such criteria: global structural similarity and local structural similarity. To generate a discrete structures s_d that are globally similar to a PDB structure s, we use a heuristic "build-up" algorithm [4], which globally fits the discrete state model $s_{\min}$ to the PDB structure s, namely, $s_{\min} = \arg_{s_d} \min [\text{RMSD}(s, s_d)]$. We obtain discrete best-fit structure for 348 proteins, with length ranging from 40 to more than 1,000. The average RMSD to the PDB structure of the best-fit structures is 2.4 Å for 4-state model, 1.8 Å for 6-state, and 1.1 Å for 10-state. Figure 2a shows the average RMSD for models with 3 to 10 states. The relative high quality of these fitted discrete state structures indicates that a model with four to six states is sufficient to generate near native structures (NNS) that have small RMSD to experimental structures. In this study, we use 4-state model for all amino acid residues. To generate discrete state model with local similarity to PDB structures, each residue is simply assigned a discrete state most similar to its local (α, τ) angle in the PDB structure.

CLUSTERING AMINO ACIDS BY FIRST ORDER STATE TRANSITION PROPENSITY. We reduce the alphabet of twenty amino acids to a simpler alphabet based on their local geometric properties. Simplifying the residue alphabet by geometric properties helps to improve characterization of the local relationship between sequence and backbone structure, and alleviate the problem of inadequate data for deriving empirical potential functions.

A protein structure can be represented uniquely by a sequence of (a, x) pairs, where a is amino acid residue type and x is the discrete state. For four discrete state and 20 amino acid types, the total number of possible descriptors at one residue position is $4 \times 20 = 80$. For simplicity, we use $s \in [1...80]$ to represent the state (discrete state

and amino acid type) a residue may take. We define first order state transition probability as: $p_{s_1,s_2} = p[s_2|s_1] = p[(a_2,x_2)|(a_1,x_1)]$. First order state transition geometric score can then be defined as: $s_g = \ln \frac{p_{s_1,s_2,obs}}{p_{s_1,s_2,exp}}$, where $p_{s_1,s_2,obs}$ is the observed probability of transition from state s_1 to state s_2, and $p_{s_1,s_2,exp}$ is the expected transition probability from state s_1 to state s_2 for a random distribution, which is the product of the relative frequency of a_1, x_1, a_2, and x_2. We cluster twenty amino acids using their first order state transition score. Specifically, we define the distance between two amino acids as the distance between the corresponding row vectors in the state transition matrix (Figure 2b).

This clustering approach based on local sequence-structure relationship provides a useful method to group amino acids and to obtain a simplified protein alphabet. For this study, we take a simplified alphabet of 5 letters: {C,I,V,L,M,W,F,Y}, {A,E,K,Q,R}, {G,S,H,T}, {P}, and {D,N} as our simplified alphabet. Each residue position now has $4 \times 5 = 20$ possible descriptors: there are 4 discrete (α, τ) states and 5 types of amino acids.

The representation of a protein structure is now a vector that includes both contact interaction (with 210 different types of contacts) and local geometric score (with $20 \times 20 = 400$ parameters for each of the 4×5 geometric descriptors for two consecutive residues).

COMBINING CONTACT AND LOCAL GEOMETRIC DESCRIPTORS. The form of the scoring function in this study is a weighted linear combination of all descriptors: $H(c) = \boldsymbol{w} \cdot \boldsymbol{c}$, where $\boldsymbol{c}$ is the descriptor vector, $\boldsymbol{w}$ is the weight vector. We obtain weight vector $\boldsymbol{w}$ using optimization method. For a linear scoring functions, the basic requirement based on Anfinsen experiment is:

$$\boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0, \quad \text{where } b \geq 0,$$

where $\boldsymbol{c}_N$ and $\boldsymbol{c}_D$ are descriptor vectors for native and decoys structures of a specific protein, respectively, and b is the energy gap required between a native and decoy structure. Each pair of native descriptor vector and decoy descriptor vector provides one inequality constraint. Together many such constraints define a convex polyhedron $\mathcal{P}$ where feasible solutions $\boldsymbol{w}$ vectors are located. For nonempty $\mathcal{P}$, there could be infinite number of choices of $\boldsymbol{w}$, all with perfect discrimination. The optimal weight vector $\boldsymbol{w}$ can be found by solving the following quadratic programming problem:

$$\text{Minimize} \qquad \tfrac{1}{2}||\boldsymbol{w}||^2 \qquad (1)$$
$$\text{subject to} \quad \boldsymbol{w} \cdot (\boldsymbol{c}_N - \boldsymbol{c}_D) + b < 0 \text{ for all } N \text{ and } D. \ (2)$$

The solution maximizes the distance $b/||\boldsymbol{w}||$ of the plane $(\boldsymbol{w}, b)$ to the origin. Here, we use a support vector machine to find such an optimal $\boldsymbol{w}$.

Table I. The number of mis-classifications using CLGP and other residue based potential functions. C_α: Computation of potential function need only C_α atoms. SC: Computation of potential function needs side chain center. AT: Computation of potential function needs all-atom representation.

Potential Functions	Complexity	Mis-classified Proteins
CLGP	C_α	4/328
TE13	SC	7/194
BV	AT	2/194
MJ	SC	85/194

We take 980 non-homologous proteins from the PDBS-ELECT database. Among these, 652 proteins are randomly chosen as training set, and the remaining 328 proteins are used as testing set. We first use gapless threading to generate >10 millions of decoys to train and test our contact and local geometric potential function (CLGP). This is then complemented by training using explicit decoys obtained from [5].

3. RESULTS

PERFORMANCE ON GAPLESS THREADING DECOYS. The performance of the potential function on gapless threading decoys is listed in Table I. Among 328 test proteins, only 4 proteins are misclassified. The accuracy of CLGP is near 99%. We compare CLGP with several other residue based potential functions, including those developed by Tobi & Elber *et.al.* (TE13) [6], Miyazawa & Jernigan (MJ) [7], and Bastolla & Vendruscolo *et. al.* (BV) [8]. Although these potential functions are residue based potential function, they still require all-atom representation since they either need to calculate the side chain centers or need to compute explicit atom-atom contacts. CLGP is the only potential function applicable to representations with only C_α atoms. It has better performance than other residue based potential and is comparable to that of all-atom potential.

PERFORMANCE ON OTHER DECOY SETS. Potential function trained using gapless threading decoys usually do not works well for more sophisticated decoys generated by other methods. To further train the CLGP potential function, we include additional decoy set generated by Loose *et. al.* [5], which contains $> 80,000$ decoy structures generated by energy minimization protocols for a set of 829 proteins. The new CLGP potential function is tested using the Park-Levitt decoy set create in [2]. The performance of our potential function on several decoy sets is shown on table II. We also compared our results with three other residue-based potential function, namely, TE13, LHL (Li, Hu, & Liang [9]), and MJ. Performance of CLGP in general is better or comparable to other potential functions. It performs better than other potential functions on LMDS decoy set.

Table II. Performance of CLGP for ParkLevitt decoy set.

Decoy sets	CLGP	TE13	LHL	MJ
Complexity	C_α	SC	AT	SC
A) 4state [3]				
1ctf	1	1	1	1
1r69	1	1	1	1
1sn3	2	6	1	2
2cro	2	1	1	1
3icb	1	–	5	–
4pti	2	7	1	3
4rxn	3	16	51	1
B) lmds [10]				
1b0n-B	1	–	2	–
1bba	436	–	217	–
1ctf	1	1	1	1
1fc2	83	14	500	501
1dtk	1	5	2	13
1igd	1	2	9	1
1shf-A	3	1	17	11
2cro	1	1	1	1
2ovo	4	1	3	2
4pti	1	–	9	–
C) lattice_ssfit [11, 12]				
1beo	1	–	1	–
1ctf	1	1	1	1
1dkt-A	1	2	1	32
1fca	7	36	40	5
1nkl	1	1	1	1
1trl-A	56	1	5	4
1pgb	1	1	1	1
4icb	1	–	1	–

4. SUMMARY AND CONCLUSION

In this study, we have developed an effective potential function for simplified representation of protein structures. By k-mean clustering, we obtained accurate discrete states and demonstrated the accuracy of model generated by these states using a modified build-up algorithm. We then simplified amino acid alphabet using their local sequence-structure relationship. The first order state transition propensity is constructed for this purpose. This potential function combines both contact interaction and local sequence-structure relationship. Contact interaction correlates well with more global topological information, while local sequence-structure relationship contains local geometric information. We find native structures can be well stabilized against decoy structures. The performance of this potential function is better than or comparable to other potential functions but it requires significantly less detailed description. We expect this potential function to be useful in protein structure prediction and fold simulation of simplified protein models.

5. REFERENCES

[1] C. Hu, X. Li, and J. Liang. On design of optimal non-linear kernel scoring function for protein folding and protein design. *www.arxiv.org, cond-mat/0302002*, Jan 31 2003.

[2] R. Samudrala and Levitt M. Decoys 'r' us: a database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9:1399–1401, 2000.

[3] B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *J. Mol. Biol.*, 258:367–392, 1996.

[4] B.H. Park and M. Levitt. The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, 249:493–507, 1995.

[5] C. Loose, J.L. Klepeis, and C.A. Floudas. A new pairwise folding potential based on improved decoy generation and side-chain packing. *Proteins*, 54:303–314, 2004.

[6] D. Tobi, G. Shafran, N. Linial, and R. Elber. On the design and analysis of protein folding potentials. *Proteins*, 40:71–85, 2000.

[7] S. Miyazawa and R.L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, 256:623–644, 1996.

[8] U. Bastolla, J. Farwer, E.W. Knapp, and M. Vendruscolo. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, 44:79–96, 2001.

[9] X. Li, C. Hu, and J. Liang. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins.*, 53:792–805, 2003.

[10] M Levitt. Molecular dynamics of native protein: I. computer simulation of trajectories. *J. Mol. Biol.*, 168:595–620, 1983.

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

[12] Y. Xia and M. Levitt. Extractin knowledge-based energy functions from protein structures by error rate minimization: Comparison of methods using lattice model. *J. Chem. Phys.*, 113:9318–9330, 2000.