Published in final edited form as:

Res Comput Mol Biol. 2009 ; 5541: 59–73. doi:10.1007/978-3-642-02008-7_5.

A Probabilistic Graphical Model for Ab Initio Folding

Feng Zhao¹, Jian Peng¹, Joe DeBartolo², Karl F. Freed³, Tobin R. Sosnick², and Jinbo Xu^{1,*}

Feng Zhao: lfzhao@tti-c.org; Jian Peng: pengjian@tti-c.org; Joe DeBartolo: bartolo@uchicago.edu; Karl F. Freed: freed@uchicago.edu; Tobin R. Sosnick: trsosnic@uchicago.edu; Jinbo Xu: j3xu@tti-c.org

¹ Toyota Technological Institute at Chicago, Chicago, IL 60637

² Department of Biochemistry and Molecular Biology, the University of Chicago, Chicago, IL 60637

³ Department of Chemistry, the University of Chicago, Chicago, IL 60637

Abstract

Despite significant progress in recent years, *ab initio* folding is still one of the most challenging problems in structural biology. This paper presents a probabilistic graphical model for *ab initio* folding, which employs Conditional Random Fields (CRFs) and directional statistics to model the relationship between the primary sequence of a protein and its three-dimensional structure. Different from the widely-used fragment assembly method and the lattice model for protein folding, our graphical model can explore protein conformations in a continuous space according to their probability. The probability of a protein conformation reflects its stability and is estimated from PSI-BLAST sequence profile and predicted secondary structure. Experimental results indicate that this new method compares favorably with the fragment assembly method and the lattice model.

Keywords

protein structure prediction; *ab initio* folding; conditional random fields (CRFs); directional statistics; fragment assembly; lattice model

1 Introduction

Various genome sequencing projects have generated millions of proteins with unknown structures. To fully understand the biological functions of these proteins, the knowledge of their three-dimensional structures is essential. Many computational methods have been developed to predict the structure of a protein from its primary sequence. These methods can be roughly classified into two categories: *ab initio* folding and comparative modeling. Although *ab initio* folding is not as accurate as comparative modeling, *ab initio* folding is still being actively studied for many purposes. The excellent performance of several groups (e.g., Zhang's I-TASSER [1], Baker's Robetta [2] and Skolnick's TASSER [3]) in recent CASP (Critical Assessment of Structure Prediction) events [4,5] indicates that by combining comparative modeling and *ab initio* folding, it is possible to improve the accuracy of a model built from template. Despite significant progress in recent years, *ab initio* folding is still one of the most challenging problems in structural biology.

*The first two authors contribute equally to this paper. Correspondence should be addressed to Dr. Jinbo Xu.

**Author contributions: J.X. designed and performed research and wrote the paper; J.P. performed research; F.Z. performed research and analyzed data; J.D., K.F., and T.S. helped with energy function.

Ab initio folding based on fragment assembly [6,7,8,9,10,11,12,13] and the lattice model [14,15,16] has been extensively studied. These two popular methods and their combination for ab initio protein structure prediction have achieved great success in CASP competitions [17,4,18]. For example, the widely-used fragment assembly program Robetta [19] is one of the most accurate ab initio folding programs. The Zhang-Server [1], which combines the lattice model, fragment assembly and threading-generated distance restraints, has outperformed all the other structure prediction servers in both CASP7 and CASP8. Although these methods demonstrate exciting results, several important issues have yet been addressed. First, due to the limited number of experimental protein structures in the Protein Data Bank (PDB), it is still very difficult to have a library of even moderate-sized fragments that can cover all the possible local conformations of a protein, especially in loop regions. Second, the conformation space defined by a fragment library is discrete in nature. The predicted structural model is not sampled from a continuous space. This discrete nature may restrict the search space and cause loss of prediction accuracy. Fragment-HMM [20], a close variant of Robetta, can sample from a continuous space, but still has the coverage problem. The lattice model used in the TOUCHSTONE programs [15,16] does not have the coverage problem, but it samples protein conformations from a three-dimensional lattice with finite resolution. That is, the conformation space defined by a lattice model is also discrete. More importantly, the sampled conformations may not have a protein-like local structure because the TOUCHSTONE programs do not sample a conformation based upon the primary sequence of a protein. Instead, the TOUCHSTONE programs use a few short-range statistical potentials in its energy function to guide the formation of protein-like local structure.

In addition to the fragment assembly method and the lattice model, there are also a few methods that attempt to sample protein conformations from a continuous space according to the probability of a conformation. The probability of a conformation approximately reflects its stability and is estimated from sequence information. In [21], Feldman and Hogue developed a program FOLDTRAJ, which implements a probabilistic all-atom protein conformation sampling algorithm. Tested on three small proteins 1VII, 1ENH, and 1PMC, FOLDTRAJ can obtain the best structural models with RMSD from native being 3.95, 5.12, and 5.95Å, respectively, out of 100,000 decoys for each protein. However, neither sequence profile nor the local conformation-dependency between two adjacent residues is used in FOLDTRAJ to estimate the probability of a conformation. Therefore, FOLDTRAJ cannot generate models with quality comparable with Robetta. Recently, Hamelryck *et al* have developed an FB5-HMM model [22] and a Torus-HMM model [23] for protein conformation sampling in a continuous space. The HMM models not only capture the relationship between backbone angles and their corresponding primary sequence and predicted secondary structure, but also consider the angle-dependency between two adjacent residues. They demonstrated that their Torus-HMM model can generate local conformations as accurately as the fragment assembly method in Robetta [23]. However, these HMM models do not consider conformation-dependency among more than two residues. It is also very difficult for these HMM models to make use of enriched sequence information such as PSI-BLAST sequence profile or threading-generated restraints to further improve sampling efficiency. Furthermore, it has not been studied if these HMM models can be applied to true ab initio folding. Recently, we have proposed a protein conformation sampling algorithm based on the first-order conditional random fields model [24] and directional statistics. The CRF model is a generalization of the HMM model and is much more expressive. Various sequence and structure features can be easily incorporated to the CRF model so that the probability of a conformation can be more accurately estimated. Our experimental results indicate that using the first-order CRF model, we can sample conformations with better quality than the FB5-HMM model [24]. All these studies have demonstrated that it is promising to probabilistically sample protein conformations from a continuous space.

However, there is no ab initio folding program based upon probabilistic sampling of a continuous space that performs as well as the fragment assembly method and the lattice model.

This paper studies if probabilistic conformation sampling in a continuous space can be used to ab initio folding. In particular, we propose a second-order CRF model for ab initio protein structure prediction. The second-order model captures the conformation-dependency among three adjacent residues instead of only between two adjacent residues. In addition, the second-order model also considers the dependency of the local conformation transition on sequence information. The second-order model has millions of model parameters while the first-order model has only one hundred-thousand model parameters. Therefore, the second-order model is much more expressive in describing the complex protein sequence-structure relationship. Using this new CRF model and directional statistics, we can explore a continuous conformation space very efficiently in a probabilistic way. The probability of a conformation is estimated from PSI-BLAST sequence profile and PISPRED-predicted secondary structure. Furthermore, we test the structure prediction capability of this new graphical model by guiding conformation search using a simple energy function consisting of only three items: DOPE [25] (a distance-dependent pairwise statistical potential), KMBHbond hydrogen bonding energy [26] and ESP (a simplified solvent potential). Experimental results indicate that although using a simple energy function, this new ab initio folding method compares favorably with the fragment assembly program Robetta [19] and the lattice model program TOUCHSTONE II [16].

2 Methods

2.1 Continuous Representation of Protein Conformations

It is time-consuming to evaluate a full-atom energy function, but a residue-level energy function usually is not as accurate as an atom-level energy function. In this paper, we use a simplified and continuous representation of a protein model. In particular, we only consider the main chain and C_β atoms in folding simulation.

C_α -trace Representation—Since the virtual bond length between two adjacent C_α atoms can be approximated as a constant (i.e., $3.8\text{\AA}$)¹, we can represent the C_α -trace of a protein using a set of pseudo backbone angles (θ , τ) [27]. Given a residue at position i , its corresponding θ is defined as the pseudo bond angle formed by the C_α atoms at positions $i-1$, i and $i+1$; τ is a pseudo dihedral angle around virtual bond between $i-1$ and i and can be calculated from the coordinates of the C_α atoms at positions $i-2$, $i-1$, i and $i+1$. Given the coordinates of the C_α atoms at positions $i-2$, $i-1$, and i , the coordinates of the C_α atom at position $i+1$ can be calculated from (θ , τ) at position i . Therefore, given the first three C_α positions and $N-2$ pairs of (θ , τ), we can build the C_α trace of a protein with N residues. The relative positions of the first three C_α atoms are determined by the θ angle at the second residue.

Distribution of Bond Angles—The preferred conformations of an amino acid in the protein backbone can be described as a probabilistic distribution of the θ and τ angles. Each (θ , τ) corresponds to a unit vector in the three-dimensional space (i.e., a point on a unit sphere surface). We can use the 5-parameter Fisher-Bingham (FB5) distribution [28,22] to model the probability distributions over unit vectors. FB5 is the analogue on the unit sphere of the bivariate normal distribution with an unconstrained covariance matrix. The (θ , τ)-space is clustered into one hundred groups, each of which is described by an FB5

¹One rare exception is when the second residue is *cis* proline, the virtual bond length is approximately $3.2\text{\AA}$.

distribution. We calculate the (θ, τ) distribution for each group from a set of approximately 3000 proteins with high-resolution X-ray structures using KentEstimator [22]. Any two proteins in this protein set share no more than 25% sequence identity. Please refer to Section 3 of [24] for a detailed description on how these distribution parameters are calculated. Therefore, given a distribution of (θ, τ) at one residue, we can easily sample a pair of real-valued (θ, τ) angles.

Building Backbone Atoms—Using (θ, τ) representation, only the coordinates of the C_α atoms can be built. To use an atom-level energy function, we also need to build the coordinates of other atoms. Given a C_α trace, there are many methods that can build the coordinates for the main chain and C_β atoms [29,30,31]. For protein folding simulation, we want a method that is both accurate and efficient. We choose to use a method similar to BBQ [30]. The original BBQ method can only build coordinates for the backbone N , C , and O atoms. We extend the method to build coordinates for the C_β atom. Experimental results (data not shown in this paper) indicate that RMSD of this method is approximately 0.5Å supposing the native C_α -trace is available. This level of accuracy is good enough for our folding simulation.

To employ the KMB hydrogen-bonding energy [26] for β -containing proteins, we also need to build the backbone hydrogen atoms. We use a quick and dirty method to build coordinates for the hydrogen atom HN [32]. Let N_i denote the position of the main chain N atom in the same residue as the HN atom. Let $N_i C_{i-1}$ denote the normalized bond vector from the N atom to the C atom in the previous residue. Let $N_i C_\alpha$ denote the normalized bond vector from the N atom to the C_α atom in the same residue. Then the position of the

hydrogen atom HN can be estimated by $N_i - \frac{N_i C_{i-1} + N_i C_\alpha}{|N_i C_{i-1} + N_i C_\alpha|}$. The average RMSD of this method is approximately 0.2Å (data not shown) supposing the native coordinates of other main chain atoms are available.

2.2 A Second-Order CRF Model for Protein Sequence-Structure Relationship

In [24], we have described a first-order CRF model for protein conformation sampling. In this paper, we extend our first-order CRF model to a second-order model to more accurately capture sequence-structure relationship. Experimental results (see Section 3.1) indicate that the second-order model is much more effective in conformation sampling than the first-order model. Please refer to [24] for a detailed description of the first-order CRF model. Here we only briefly explain our extension to the second-order CRF model.

In the context of conditional random fields, the primary sequence (or PSI-BLAST sequence profile) and predicted secondary structure are viewed as observations; the backbone pseudo angles and their distributions are treated as hidden states or labels. Let X denote the PSIPRED-predicted secondary structure likelihood scores. It is a matrix with $3 \times N$ elements (where N is the number of residues in a protein), each of which is the predicted likelihood of one secondary structure type at a specific position. Let X_i denote the predicted likelihood of three secondary structure types at position i . X_i is a vector of three elements and $X_i(x)$ is the predicted likelihood of secondary structure type x at position i . Let M denote the PSI-BLAST sequence profile, which is a position-specific frequency matrix containing $20 \times N$ entries. Each element in this matrix is the occurring frequency of one amino acid at a given position. Let M_i , a vector of 20 elements, denote the sequence profile at position i . Let $M_i(aa)$ denote the occurring frequency of amino acid aa at position i . Let $H = \{h_1, h_2, \dots, h_{100}\}$ denote the set of one hundred backbone angle states, each of which represents an FB5 distribution. Let $S = \{s_1, s_2, \dots, s_N\}$ ($s_i \in H$) be a sequence of labels (i.e., FB5 distribution) corresponding to the observations X and M .

As shown in Figure 1, we use a second-order CRF model to capture the complex relationship between the state sequence S and the observations X and M . Our CRF model defines the conditional probability of S given M and X as follows.

$$P_{\theta}(S|M, X) = \exp\left(\sum_{i=1}^N F(S, M, X, i)\right) / Z(M, X) \quad (1)$$

where $\theta = (\lambda_1, \lambda_2, \dots, \lambda_p)$ is the model parameter and $Z(M, X) = \sum_s \exp\left(\sum_{i=1}^N F(S, M, X, i)\right)$ is a normalization factor summing over all the possible label sequences for the given observations. $F(S, M, X, i)$ consists of two edge features and two label features at position i . It is given by

$$F(S, M, X, i) = e_1(s_{i-1}, s_i) + e_2(s_{i-1}, s_i, s_{i+1}) + \sum_{j=i-\omega}^{i+\omega} v_1(s_i, M_j, X_j) + \sum_{j=i-\omega}^{i+\omega} v_2(s_{i-1}, s_i, M_j, X_j) \quad (2)$$

Meanwhile, $e_1(s_{i-1}, s_i)$ and $e_2(s_{i-1}, s_i, s_{i+1})$ are the first-order and second-order edge feature functions, respectively. and $v_1(s_i, M_j, X_j)$ and $v_2(s_{i-1}, s_i, M_j, X_j)$ are two label feature functions.

The two edge feature functions are given by

$$e_1(s_{i-1}, s_i) = \lambda(h', h'') [s_{i-1} == h'] [s_i == h''] \quad (3)$$

$$e_2(s_{i-1}, s_i, s_{i+1}) = \lambda(h', h'', h''') [s_{i-1} == h'] [s_i == h''] [s_{i+1} == h'''] \quad (4)$$

In the above equations, the indicator function $[s_i == h]$ is equal to 1 if the state at position i is $h \in H$, otherwise 0. The edge feature functions are independent of the observations. They capture the conformation-dependency of two or three adjacent residues. The model parameters (i.e., λ) in Eqs. 3 and 4 are identified by the states at two or three adjacent residues, respectively.

The two label feature functions are given by

$$v_1(s_i, M_j, X_j) = \sum_s \sum_{aa} \lambda(j-i, s, aa, h) X_j(s) M_j(aa) [s_i == h] + \sum_s \lambda(j-i, s, h) X_j(s) [s_i == h] + \sum_{aa} \lambda(j-i, aa, h) M_j(aa) [s_i == h] \quad (5)$$

$$v_2(s_{i-1}, s_i, M_j, X_j) = \sum_s \sum_{aa} \lambda(j-i, s, aa, h', h'') X_j(s) M_j(aa) [s_{i-1} == h'] [s_i == h''] + \sum_s \lambda(j-i, s, h', h'') X_j(s) [s_{i-1} == h'] [s_i == h''] + \sum_{aa} \lambda(j-i, aa, h', h'') M_j(aa) [s_{i-1} == h'] [s_i == h''] \quad (6)$$

Label feature functions model the dependency of backbone angles on protein sequence and secondary structure. Eqs. 5 and 6 define the first-order and second-order label feature functions, respectively. These two equations indicate that not only the state (i.e., angle) itself

but also the state transition depend on the sequence profile and secondary structure. The label feature functions also model the interactions between secondary structure and primary sequence, as shown in the first items of the right hand side of Equations 5 and 6. According to the third and fourth items in the right hand side of Eq. 2, the state (or state transition) at one position depends on the sequence profile and secondary structure information in a window of width $2\omega+1$. In our implementation, we set the half window width ω to 4. The model parameters for label features are identified by one or two states, secondary structure type, amino acid identity and the relative position of the observations.

The second-order CRF model has millions of features, each of which has a model parameter to be trained. We train the model parameters by maximizing $P_{\theta}(S|M, X)$ on a set of 3000 proteins chosen by the PISCES server[33]. Please refer to [24] for a detailed description of how the CRF model can be trained and how the model parameters are chosen using cross-validation. The second-order CRF model is much more expressive than the first-order one and can describe the protein sequence-structure relationship more accurately. We revised the FlexCRFs program [34] to train our CRF model and it takes approximately 24 hours to train a single model on a cluster of 150 CPUs.

Once the CRF model is trained, we can efficiently estimate the probability of a protein conformation, which can be used to sample a protein conformation or resample the local conformation of a protein segment. First we probabilistically sample labels (i.e., angle distribution) from the CRF model and then sample real-valued angles from the labels. Please refer to [24] for more details. The conformation sampling algorithm in [24] is based on the first-order model, but it can be easily extended to the second-order model.

2.3 Energy Function

The energy function we used for protein folding simulation consists of three items: DOPE, KMBhbond and ESP. DOPE is a full-atom, distance-dependent pairwise statistical potential designed by Shen and Sali [25]. DOPE performs as well or better than many other statistical potentials and force fields in differentiating a native structure from decoys [25,35]. The statistical potential in DOPE distinguishes the amino acid identity and atomic identity of two interacting particles. In our folding simulation, we only build coordinates for main chain and C_{β} atoms, so only the statistical potentials related to main-chain and C_{β} atoms are used to calculate the energy of a conformation. We denote this revised DOPE as DOPE- C_{β} . According to [35], DOPE- C_{β} is highly correlated with the full-atom DOPE. DOPE- C_{β} also performs favorably in applications to intra-basin protein folding [36].

KMBhbond is a statistical potential for hydrogen bonding developed by Baker's group [26]. It depends on the distance between the geometric centers of the N-H bond vector and the C=O bond vector, the bond angle between the N-H bond vector and the hydrogen bond, the bond angle between the C=O bond vector and the hydrogen bond, and the dihedral angle about the acceptor-acceptor base bond. The three angles describe the relative orientation of the bond vectors in the hydrogen bond.

ESP is an approximation to the Ooi-Scheraga solvent-accessible surface area (SASA) potential [37]. Since our conformation representation does not contain side-chain atoms, which are necessary for the calculation of the solvent-accessible surface area potential, we employ a simple ESP that assigns each residue with an environmental energy score. ESP is a function of the protein size and the number of C_{α} atoms contained within an $8.5\text{\AA}$ sphere centered on the residue's C_{α} atom [38]. Explicitly, the ESP statistical potential has the form given by

$$ESP(aa, n) = -\ln \frac{P(n|R, aa)}{p(n|R)} \quad (7)$$

where n is the number of C_α atoms in an $8.5\text{\AA}$ sphere centered on the residue's C_α atom, R is the radius of gyration of the protein, aa is the amino acid identity of the residue, $p(n|R)$ is the number of C_α atoms in an $8.5\text{\AA}$ sphere for a given protein radius regardless of amino acid identity, and $P(n|R, aa)$ is the number of C_α atoms in an $8.5\text{\AA}$ sphere for a given protein radius and amino acid identity. We calculate $ESP(aa, n)$ from a set of 3000 non-redundant experimental structures chosen by the PISCES server [33]. Each protein in this set has resolution at least $2.0\text{\AA}$, R factor no bigger than 0.25 and at least 30 residues. Any two proteins in this set share no more than 30% sequence identity.

The weight factors combining these three energy items are trained on the proteins in Table 1 using grid search in a progressive way. First, we fix the weight factor of DOPE to 1 and determine the weight factor for ESP by minimizing the average RMSDs of generated decoys. Then we fix the weight factors of both DOPE and ESP and determine the weight factor for KMBhbond using the same way.

2.4 Energy Minimization

We employ a simulated annealing (SA) algorithm to minimize the energy function for a given protein. The SA routine is based on the algorithm described by Aarts and Korst [39]. We start with sampling an initial conformation and then search for a better one by minimizing the energy function. Given a conformation, we propose a new conformation by resampling the local conformation of a small segment using the CRF model (see [24]). The new conformation is rejected if there are serious steric clashes among atoms², otherwise we calculate its energy. If the new conformation has an energy value smaller than the previous one, then we accept this conformation, otherwise we accept it by a probability $e^{-\frac{\Delta E}{t}}$ where ΔE is the energy increment and t is the annealing temperature.

The initial annealing temperature is chosen so that at the beginning of the annealing process an energy increase is accepted with a given probability p_0 ($=0.8$). The initial temperature t_0 is determined by

$$t_0 = -\Delta E / \ln(p_0) \quad (8)$$

Where ΔE is the average energy increase. Given a protein, we conduct a series of trial conformation samplings and accept all the generated conformations. Then we estimate ΔE by calculating the average energy increase observed in our trial samplings. During the folding simulation process, we decrease the annealing temperature gradually using an exponential cooling schedule. The temperature is updated as follows.

$$t_{k+1} = \alpha t_k \quad (9)$$

Where α is set to 0.9.

At each annealing temperature, the number of sampled conformations is set to $100 \times (1 + N/100)$ where N is the number of residues in the protein. This number is set to achieve thermal equilibrium. The termination of the SA process is triggered when any of the following two

²There is a steric clash if the distance between two C_α atoms is less than $4\text{\AA}$.

conditions is satisfied: 1) either the temperature is low enough such that almost no energy increase is accepted and the annealing process is trapped at a local minima; 2) or the number of conformations generated in a single simulation process reaches a threshold (say 10,000).

3 Results

3.1 Comparison with the First-Order Model

First, we compare our second-order CRF model with the first-order model described in [24] to see how much improvement we can achieve by considering the interdependency among three adjacent residues. In this comparison, we guide conformation search using only compactness and self-avoiding constraints but not energy function. We generated 20,000 decoys for each test protein using these two CRF models and then calculated the average RMSDs of the top 1%, 2%, 5% and 10% decoys, respectively. Table 1 shows the quality of the decoys generated by the first-order and second-order models on a variety of test proteins. In terms of the best decoys, the second-order model is better on 13 out of 22 test proteins and worse on seven proteins. The best decoys may be generated by chance, so they cannot be reliably used to evaluate the performance of two CRF models. We further examine the performance of these two CRF models by comparing the average RMSDs of their top decoys. The second-order model outperforms the first-order model on almost all the test proteins except 1aa2 and 4icb. Both CRF models have similar performance on 1aa2 maybe because that 1aa2 has a large conformation search space so that neither CRF model can search the space efficiently. The reason that the second-order model performs worse on 4icb is because there is a *cis* proline in this protein, and the length of the virtual C_{α} -bond ending at this proline is approximately 3.2Å instead of our assumption 3.8Å. Therefore, the more accurately can the model predict the backbone angles, the more the decoys deviate from the native. This problem will be addressed later since we can predict if a residue is a *cis* proline or not with accuracy 92%.

3.2 Comparison with TOUCHSTONE II

By combining the energy functions described in Section 2.3 and our second-order CRF model, we build a program, denoted as CRFFolder, for ab initio protein structure prediction. We compare CRFFolder with TOUCHSTONE II, a representative lattice-model-based ab initio protein structure prediction program developed by Skolnick *et al.* TOUCHSTONE II is an excellent ab initio folding program and its two derivatives TASSER [3] and I-TASSER [1] perform very well in both CASP7 and CASP8. We do not compare CRFFolder with the two derivatives because both TASSER and I-TASSER use threading-generated constraints to guide conformation search while CRFFolder does not. Due to the limitations of computational power, we tested CRFFolder on only 15 test proteins, which were also tested by TOUCHSTONE II. These test proteins have very different secondary structures and protein sizes ranging from 47 to 157. We generated approximately 3000 decoys for each alpha protein, 7000 decoys for each alpha-beta protein, and 10,000 decoys for each beta protein. By contrast, TOUCHSTONE II used a complex energy function consisting of 21 items and generated 24,000 decoys for each test protein [16]. As shown in Table 2, CRFFolder performs much better than TOUCHSTONE II on all the alpha proteins except one. CRFFolder also has comparable performance on beta and alpha-beta proteins. Looks like that on larger test proteins, CRFFolder is slightly worse than TOUCHSTONE II. This may be because that the replica exchange Monte Carlo algorithm used by TOUCHSTONE II for energy minimization is better than the simulated annealing algorithm used in CRFFolder. Note that since two programs use very different clustering methods, it is not easy to compare these two programs fairly. TOUCHSTONE II used a program SCAR [16] to do decoy clustering while we use MaxCluster³. For the purpose of comparison, we also show the

RMSD of the best decoys and the average RMSDs of the top 1% and 2% decoys generated by CRFFolder.

3.3 Comparison with Robetta in CASP8

In this section, we compare the performance of our method CRFFolder with Baker's Robetta on some hard CASP8 targets. CASP8 was held during the summer of 2008 and these hard targets have no good homologs in the PDB. Note that our second-order CRF model was trained before CASP8 started and the energy function was gradually improved during CASP8. We use the hard targets on which both Robetta and CRFFolder did *ab initio* folding. Robetta is a well-known fragment assembly method for *ab initio* protein structure prediction. The top five models generated by Robetta *ab initio method* for each hard target are available at the Robetta web site.⁴ Using CRFFolder, we generated 7000 decoys for each target and then chose top five models. Note that the top models chosen by CRFFolder are not exactly same as our CASP8 submissions since we also submitted threading-generated models for some hard targets.

Table 3 compares CRFFolder and Robetta in terms of the quality of the first-ranked structure models. The model quality is evaluated by a program TM-score [40], which generates a number between 0 and 1 to indicate the quality of a structure model. Roughly speaking, the higher the TM-score, the better the model quality. As shown in this table, on average CRFFolder is better than Robetta by approximately 0.024 (i.e., 8%). On a variety of proteins such as T0467, T0468, T0480, T0495_2, T0496_1, T0496_2 and T0510_3, CRFFolder is much better than Robetta. By contrast, Robetta is also much better than CRFFolder on some beta-containing proteins such as T0466, T0482, and T0513_1. In particular, Robetta did very well on T0513_1, an alpha-beta protein with two beta sheets. This may indicate that Robetta is better in sampling conformations for large beta sheets or has a better hydrogen-bonding energy item for the formation of large beta sheets. Please note that the TM-scores of T0397_1, T0495_2, T0496_1, T0496_2 and T0513_1 are different from those on Zhang's CASP8 assessment page because of different domain definitions. The domain definition of T0510_3 is from Zhang's CASP8 assessment page⁵ while the domain definitions of the others are from Robetta's CASP8 web site.

4 Conclusion

This paper presented a probabilistic graphical model for *ab initio* protein structure prediction. By using the second-order CRF model and directional statistics, we can accurately describe the protein sequence-structure relationship and search a continuous protein conformation space very efficiently. Although still at its infancy stage and using a simple energy function, our method can do *ab initio* protein structure prediction as well as two well-developed *ab initio* folding programs TOUCHSTONE-II and Robetta. These two programs have been developed for many years and have well-tuned and sophisticated energy functions. Our method is much better than TOUCHSTONE II on alpha proteins and has similar accuracy on beta-containing proteins with TOUCHSTONE II. Our method is also better than Robetta on quite a few CASP8 test proteins but worse than Robetta on some beta-containing proteins. To improve the performance of our method on beta-containing proteins, we will further improve our conformation sampling algorithm on beta regions and develop a better hydrogen-bonding energy item for the formation of beta sheets.

³<http://www.sbg.bio.ic.ac.uk/~maxcluster/>

⁴<http://robetta.bakerlab.org/queue.jsp?UserName=casp8&rpp=100>.

⁵<http://zhang.bioinformatics.ku.edu/casp8/index.html>

Acknowledgments

This work is supported by the internal research funding of TTI-C and NIH research grant. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca), the Open Science Grid Engagement VO and the University of Chicago Computation Institute. The authors are also grateful to Dr. Ian Foster, Dr. John McGee and Mats Rynge for their help with computational resources.

References

1. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology*. 2007; 5:17+. [PubMed: 17488521]
2. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of National Academy Sciences*. 2006; 103(14):5361–5366.
3. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. *Proceedings of National Academy Sciences, USA*. 2005; 102(4):1029–1034.
4. Moulton J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round 6. *Proteins: Structure, Function and Bioinformatics*. 2005; 61(suppl. 7):3–7.
5. Moulton J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction–Round VII. *Proteins: Structure, Function, and Bioinformatics*. 2007; 69(S8):3–9.
6. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. *EMBO Journal*. 1986; 5:819–823. [PubMed: 3709525]
7. Claessens M, van Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with parts from known protein structures. *Protein Engineering*. 1989; 2(5):335–345. [PubMed: 2928296]
8. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Structure, Function and Genetics*. 1989; 5(4):355–373.
9. Simon I, Glasser L, Scheraga HA. Calculation of Protein Conformation as an Assembly of Stable Overlapping Segments: Application to Bovine Pancreatic Trypsin Inhibitor. *Proceedings of National Academy Sciences, USA*. 1991; 88(9):3661–3665.
10. Levitt M. Accurate modeling of protein conformation by automatic segment matching. *Journal of Molecular Biology*. 1992; 226(2):507–533. [PubMed: 1640463]
11. Sippl M. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*. 1993; 17:355–362.
12. Wendoloski JJ, Salemme FR. PROBIT: a statistical approach to modeling proteins from partial coordinate data using substructure libraries. *Journal of Molecular Graphics*. 1992; 10(2):124–126. [PubMed: 1637750]
13. Bowie JU, Eisenberg D. An Evolutionary Approach to Folding Small α -Helical Proteins that Uses Sequence Information and an Empirical Guiding Fitness Function. *Proceedings of National Academy Sciences, USA*. 1994; 91(10):4436–4440.
14. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. *Journal of Molecular Biology*. 2000; 300(1):171–185. [PubMed: 10864507]
15. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences*. 2001; 98(18):10125–10130.
16. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical Journal*. 2003; 85(2):1145–1164. [PubMed: 12885659]
17. Moulton J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP)–round V. *Proteins: Structure, Function, and Genetics*. 2003; 53(S6):334–339.
18. Moulton J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structure Biology*. Jun.2005

19. 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. *Journal of Molecular Biology*. 1997; 268(1):209–225. [PubMed: 9149153]
20. Li SC, Bu D, Xu J, Li M. Fragment-hmm: A new approach to protein structure prediction. *Protein Science*. Aug.2008 ps.036442.108+
21. Feldman HJ, Hogue CWV. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins: Structure, Function, and Genetics*. 2002; 46(1):8–23.
22. Hamelryck T, Kent JTT, Krogh A. Sampling Realistic Protein Conformations Using Local Structural Bias. *PLoS Comput Biology*. Sep.2006 2(9)
23. Boomsma W, Mardia KV, Taylor CC, Ferkinghoff-Borg J, Krogh A, Hamelryck T. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*. 2008; 105(26):8932–8937.
24. Zhao F, Li S, Sterner BW, Xu J. Discriminative learning for protein conformation sampling. *Proteins: Structure, Function, and Bioinformatics*. 2008; 73(1):228–240.
25. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Science*. 2006; 15(11):2507–2524. [PubMed: 17075131]
26. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of National Academy Sciences*. 2004; 101(18):6946–6951.
27. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*. 1976; 104:59–107. [PubMed: 957439]
28. Kent JT. The Fisher-Bingham Distribution on the Sphere. *Journal of Royal Statistical Society*. 1982; 44:71–80.
29. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *Journal of Molecular Biology*. 1991; 218(1):183–194. [PubMed: 2002501]
30. Gront D, Kmiecik S, Kolinski A. Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *Journal of Computational Chemistry*. 2007; 28(9):1593–1597. [PubMed: 17342707]
31. Maupetit J, Gautier R, Tufféry P. SABBAC: online structural alphabet-based protein backbone reconstruction from alpha-carbon trace. *Nucleic Acids Research*. Jul.2006 34(Web-server)
32. Branden, CI.; Tooze, J. *Introduction to Protein Structure*. 2nd. Garland Publishing; 1999.
33. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. *Bioinformatics*. 2003; 19(12): 1589–1591. [PubMed: 12912846]
34. Phan XH, Nguyen LM, Nguyen CT. FlexCRFs: Flexible Conditional Random Field Toolkit. 2005<http://flexcrfs.sourceforge.net>
35. Fitzgerald JE, Jha AK, Colubri A, Sosnick TR, Freed KF. Reduced Cbeta statistical potentials can outperform all-atom potentials in decoy identification. *Protein Science*. 2007; 16(10):2123–2139. [PubMed: 17893359]
36. Colubri A, Jha AK, Shen MY, Sali A, Berry RS, Sosnick TR, Freed KF. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. *Journal of Molecular Biology*. 2006; 363(4):835–857. [PubMed: 16982067]
37. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible Surface Areas as a Measure of the Thermodynamic Parameters of Hydration of Peptides. *Proceedings of the National Academy of Sciences*. 1987; 84(10):3086–3090.
38. Fernández A, Sosnick TR, Colubri A. Dynamics of hydrogen bond desolvation in protein folding. *Journal of molecular biology*. 2002; 321(4):659–675. [PubMed: 12206781]
39. Aarts, E.; Korst, J. *Simulated Annealing and Boltzmann Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing*. Wiley; Chichester: 1991.
40. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*. 2005; 33(7):2302–2309. [PubMed: 15849316]

Fig. 1. The second-order CRF model used to describe protein sequence-structure relationship. Each S_i represents the angle distribution at position i .

Table 1

Decoy quality comparison between the first-order and second-order CRF conformation samplers. Columns 1-3 list the PDB code, protein size and the type of the test proteins. Columns “best” list the RMSDs of the best decoys; the other columns list the average RMSDs of the top decoys. “O-1” and “O-2” denote the first-order and the second-order CRF models, respectively. In total 20,000 decoys are generated for each protein without using energy function.

PDB code	Length	Class	best		1%		2%		5%		10%	
			O-1	O-2	O-1	O-2	O-1	O-2	O-1	O-2	O-1	O-2
1aa2	108	α	7.3	7.3	9.3	9.5	9.8	9.9	10.4	10.4	10.9	10.8
1beo	98	α	6.4	5.8	8.4	8.1	8.8	8.6	9.5	9.3	10.1	9.9
1ctfA	68	$\alpha\beta$	3.7	3.7	5.1	4.6	5.4	4.9	5.9	5.3	6.5	5.7
1dktA	72	β	6.1	5.1	7.6	6.4	8.0	6.7	8.5	7.2	9.0	7.7
1enhA	54	α	2.3	2.2	3.1	2.6	3.3	2.6	3.7	2.8	4.1	2.9
1fc2C	43	α	1.9	2.3	2.7	2.6	2.8	2.7	3.1	2.9	3.4	3.0
1fca	55	β	4.9	5.0	6.4	6.2	6.7	6.5	7.3	6.9	7.7	7.2
1fgp	67	β	7.4	5.9	8.9	7.8	9.2	8.1	9.6	8.6	10.0	8.9
1jer	110	β	9.6	10.2	11.6	11.5	11.9	11.8	12.4	12.3	12.9	12.7
1nkl	78	α	3.6	3.1	4.7	3.8	5.0	3.9	5.4	4.3	5.8	4.6
1pgb	56	$\alpha\beta$	3.1	2.6	4.1	3.6	4.3	3.8	4.6	4.0	4.9	4.2
1sro	76	β	6.2	5.4	7.8	7.2	8.2	7.6	8.8	8.2	9.3	8.8
1trlA	62	α	3.5	3.7	4.4	4.5	4.6	4.6	4.9	4.8	5.2	5.0
2croA	65	α	2.8	2.6	3.6	3.2	3.8	3.4	4.2	3.6	4.6	3.9
2gb1A	56	β	2.9	2.0	4.0	3.5	4.2	3.6	4.6	3.9	4.9	4.1
4icbA	76	α	4.6	4.4	5.9	6.7	6.2	7.1	6.8	7.7	7.4	8.2
T052	98	β	7.6	8.4	10.6	10.1	11.0	10.5	11.6	11.1	12.1	11.6
T056	114	α	7.8	7.6	9.8	9.6	10.3	9.9	10.9	10.5	11.5	11.2
T059	71	β	6.3	6.2	8.5	8.0	8.8	8.2	9.3	8.7	9.6	9.0
T061	76	α	5.3	6.0	7.0	7.1	7.3	7.3	7.8	7.6	8.2	7.9
T064	103	α	7.2	7.5	9.4	9.0	9.9	9.5	10.7	10.4	11.4	11.1
T074	98	α	4.9	4.2	7.3	6.6	7.7	7.0	8.4	7.6	9.0	8.1
average RMSD			5.25	5.05	6.8	6.4	7.2	6.7	7.7	7.2	8.1	7.6

Table 2

Decoy quality comparison between CRFFolder and TOUCHSTONE II. Columns 1-3 list the PDB code, length and the type of the test proteins. The two “Best Cluster” columns list the RMSDs of the representative decoys of the best clusters. In these two columns, the first number in parentheses denotes the rank of the best cluster and the second number is the total number of clusters. Column “best” lists the RMSDs of the best decoys. Columns “1%” and “2%” list the average RMSDs of the top 1% and 2% decoys, respectively. The results of TOUCHSTONE II are from [16].

PDB code	Length	Class	TOUCHSTONE II		CRFFolder		
			Best Cluster	Best Cluster	Best	1%	2%
1bw6A	56	α	4.79(2/3)	3.82(3/3)	2.75	3.38	3.54
1lea	72	α	5.69(5/5)	4.10(5/7)	3.41	4.19	4.48
2af8	86	α	11.07(5/6)	8.9(12/19)	7.07	8.53	8.97
256bA	106	α	3.61(2/3)	2.75(6/11)	2.50	3.45	3.70
1sra	151	α	10.71(3/12)	13.95(17/25)	10.82	13.76	14.24
1gpt	47	$\alpha\beta$	6.30(1/25)	5.55(42/67)	4.34	5.20	5.47
1kp6A	79	$\alpha\beta$	10.01(8/14)	7.99(1/7)	6.29	7.51	7.81
1poh	85	$\alpha\beta$	9.10(5/9)	8.84(5/10)	7.49	8.70	9.04
1npsA	88	$\alpha\beta$	6.89(33/34)	9.91(41/57)	7.87	9.19	9.66
1ttdA	100	$\alpha\beta$	8.96(7/13)	9.22(10/13)	6.51	9.51	9.94
1nmsi	66	β	7.72(19/28)	7.77(12/15)	6.24	7.55	7.89
1hoe	74	β	9.39(5/13)	9.87(16/35)	7.96	10.00	10.37
1ezgA	82	β	11.03(40/44)	10.42(42/66)	9.66	10.35	10.62
1sfp	111	β	7.48(2/18)	11.07(5/11)	9.32	11.09	11.59
1b2pA	119	β	12.52(31/56)	10.01(18/25)	8.76	10.89	11.32

Table 3

Decoy quality comparison between CRFFolder and Robetta. The “Robetta” column lists TM-scores of the first-ranked decoys generated by Baker's Robetta. The “CRFFolder” column lists TM-scores of the first-ranked decoys generated by CRFFolder.

Target ID	Length	Class	Robetta	CRFFolder
T0397_1	70	$\alpha\beta$	0.25	0.258
T0460	111	$\alpha\beta$	0.262	0.308
T0465	157	$\alpha\beta$	0.243	0.253
T0466	128	β	0.326	0.217
T0467	97	β	0.303	0.364
T0468	109	$\alpha\beta$	0.253	0.308
T0476	108	$\alpha\beta$	0.279	0.250
T0480	55	β	0.208	0.307
T0482	120	$\alpha\beta$	0.352	0.223
T0484	62	α	0.253	0.249
T0495_2	65	$\alpha\beta$	0.312	0.436
T0496_1	110	$\alpha\beta$	0.235	0.293
T0496_2	68	α	0.291	0.500
T0510_3	43	$\alpha\beta$	0.147	0.352
T0513_1	77	$\alpha\beta$	0.581	0.367
T0514	145	$\alpha\beta$	0.283	0.277
Average			0.286	0.310
Sum			4.578	4.960