Fragment-HMM: A New Approach To Protein Structure Prediction

Shuai Cheng Li*, Dongbo Bu†, Jinbo Xu‡, and Ming Li‡

*David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1, and †Toyota Technological Institute at Chicago, Chicago, IL, USA, 60637
Submitted to Proceedings of the National Academy of Sciences of the United States of America

Protein structure prediction has been a heuristic science. From homology modeling, threading, to Monte Carlo fragment assembly, decoy clustering, selection, refinement, and consensus, there is no unified model or theory governing the complete process. We believe the protein structure prediction problem should be solved by a simple computational model that encompasses all of above paradigms. We propose such a theory, integrating ideas from fragment assembly, hidden Markov model sampling, and Ramachandran basins. In particular, we design a position-specific hidden Markov model, based on a fragment library, to model the protein folding process. The new framework naturally repeats its process to converge to a final target, conglomerating fragment assembly, clustering, target selection, and potentially refinement, threading and consensus, all in one process. Our initial implementation, FALCON, of this theory converges to within 6Å of the native structures for 100% decoys on all 6 standard benchmark proteins used in ROSETTA [26]. In comparison, only 14% to 94% of the ROSETTA final decoys reached 6Å for these proteins. The qualities of the best decoys and the final decoys FALCON converges to are also better than the best decoys and the final decoys selected by ROSETTA, respectively.

Introduction

We wish to find a unified and the simplest model for folding a protein in silico. That is, we are not interested in trying PSI-BLAST for easy targets, threading by RAPTOR [30] for harder targets, fragment assembly by ROSETTA [26] for ab initio targets, or consensus for everything. We are also not interested in using different methods for different steps, such as Monte Carlo fragment assembly, clustering, selecting, refinement. Nature does not do this. It does not fit with the Occam’s razor principle [2, 19].

Nature prefers simplicity. We wish to find one theory, one model, as simple as possible, that goes from an input sequence to the final structure. This theory should embody homology modeling, threading, fragment assembly (all stages of it), loop modeling, refinement, and even consensus. This theory must be simple, robust, and effective.

This paper presents our initial efforts in building such a theory, and our preliminary implementation of this theory, FALCON, together with clear cut experimental results.

Some ideas of our work come from three lines of research: fragment assembly, hidden Markov model sampling, and Ramachandran basins.

The most successful approach for ab initio structure prediction is to use short structural fragments to model local interactions among the amino acids of a segment, and utilize the non-local interactions to arrange these short structural fragments to form native-like structures [26]. Despite of the importance of non-local interactions in directing the search to discover the native-like protein structures, the relationship between local structures and the interactions among amino acids within a local structure remain active issues of research. An accurate prediction of the local structural bias for a sequence segment is critically important to protein structure prediction.

According to the Levinthal paradox [17], the number of possible conformations of a protein chain is exponential in the protein sequence length due to the large degrees of freedom of the unfolded polypeptide chain. As a consequence, a brute force enumeration of all possible conformations for a given sequence is both computationally and physically infeasible. However, the local structural bias information restricts the possible conformations of each sequence segment, and therefore narrows down the conformation space of the whole polypeptide chain significantly.

Structural motif is a straightforward description of local structural bias. The idea of structural motif can be traced back to [22], in which a protein fold is modeled as an assembly of smaller building blocks from the regular secondary structure elements. In the past years, considerable works [6, 14, 11, 5, 8, 24, 20] have been conducted to define local structural motifs and analyze their structural characterization and sequence preferences. The sequence preferences can be used to predict structural motif for new sequences. Another approach is to search for recurrent sequence patterns first, and then study the structural motif shared by these recurrent sequence patterns. This approach can identify new structural motifs since the important structural properties need not be specified in advance. HMMSTR [7], a hidden Markov model on structural motif space, attempted to describe the overlaps of structural motifs, and the transition probability between motifs. HMMSTR can be considered as a probabilistic version of structural motif library.

Structural motifs serve as the foundation to obtain better predictions. For example, ROSETTA [26] selected 9-mer structural fragments from known protein structures as building blocks, while TASSER [32, 31] generated fragments of various lengths from threading results. Despite of the promising progresses of fragment assembly methods, the structural motif strategy still suffers from its inherent discrete nature. That is, the structural motif library is discrete, while the conformation space of a protein is continuous. Therefore, it is...
impossible to cover the whole conformation space by a limited number of structural motifs. This drawback limits the accuracy of protein structure prediction [15, 12].

An alternative way to describe local structural bias is Ramachandran basin [24]. A Ramachandran basin refers to a specific region of Ramachandran plot imposed by local interactions among amino acids. Ramachandran basin provides a convenient way to present the preference of a specific torsion angle. Sosnick and Freed et al [9] employed the Ramachandran basin technique to investigate the levels of representation required to predict protein structure. Specifically, they tested the ability to recover the native structure from a given Ramachandran basin assignment for each amino acid. In this method, the Ramachandran plot is divided into five pre-determined Ramachandran basins. By decomposing the Ramachandran plot into four or more basins, Shortle [25] calculated the propensities of amino acids mapped to each basin. Shortle argued that these propensities are results of local side-chain-backbone interactions and may restrict the denatured conformation ensemble to a relatively small subset of native-like conformations. Gong et al [12] also investigated the protein structure reconstructing problem from coarse-grained estimation of the native torsion angle. The only difference of these works lies at the definition of Ramachandran basin: Gong et al partitioned both $\phi$ and $\psi$ angle intervals into six ranges, each range of 60°; thus the Ramachandran map is partitioned into 36 basins uniformly. Both of these two studies demonstrate that the knowledge of torsion angles helps the reconstruction of small-sized proteins. However, these two works partition the Ramachandran plot into basins in random manners without statistical explanations describing the torsion angle distributions of each basin. Furthermore, to give each residue a coarse Ramachandran basin assignment, the native structure should be known in advance, which makes these two frameworks infeasible for real-life protein structure prediction.

As the third precursor to our work, Hamelryck et al [13] proposed to apply FB5, a directional distribution, to parameterize the local structural bias. Using this tool, they investigated the local bias in $(\theta, \tau)$ space rather than $(\phi, \psi)$ space. In this method, the local structural bias for each amino acid is trained via a hidden Markov model called FB5-HMM. Instead of HMM, Xu et al [33] proposed CRFSampler, a protein structure sampling framework based on another probabilistic graph model Conditional Random Fields [4]. CRFSampler [33] also employs the FB5 to model backbone angle preferences but with improved results over that of [13]. The successes of these two methods suggest the advantage of continuous torsion angle distributions over discrete structural motifs: using the torsion angle distribution technique, it is possible to generate conformations with local structures not occurring in the structural fragment library. In addition, experimental results demonstrate that the derived local biases can help to generate native-like conformations, and support the view that relatively few conformations are compatible with the local structural biases.

Despite of the promising advancements, the FB5-HMM type of approach has three serious disadvantages. First, FB5-HMM reports the optimal number of local biases as 75 by training on a large set of representative protein structures. In other words, the $(\theta, \tau)$ map which is used to model the local bias is partitioned into 75 basins. This partition scheme implies that a protein sequence of length $n$ has a conformation space of size $O(75^n)$, which is astronomically larger than the estimation of $O(1.6^n)$ by [27]. In addition, it is challenging to select suitable models from these 75 local biases for a particular residue. Second, the derived distributions are general, while a residue may have its specific preference of the distributions and therefore none of the 75 local biases suits it. Third, FB5-HMM is incapable of capturing the relationships among the residues of a local segment. To capture the dependencies among the adjacent residues, we can use a higher order HMM. However, the model complexity increases exponentially with the HMM order. For example, if we want to use an HMM of order nine, the number of nodes is around 75$^9$. Thus the model becomes infeasible and we lack training data for such a model. As a consequence, though equipped with an elegant statistical model, FB5-HMM and CRFSampler have much lower prediction accuracies compared to ROSETTA [13, 33].

The New Paradigm

We propose a simple and unified paradigm for protein structure prediction. The plan is to probabilistically sample protein structure conformations compatible with local structural biases for a given protein. The architecture of the model is as below.

1. For residue $i$, Cosine models [21] are used to describe the local bias of its torsion angle pair $(\phi_i, \psi_i)$.

2. A position specific hidden Markov model (HMM) is used to capture the dependencies among local biases of adjacent residues, based on carefully selected fragments [26, 20]¹. This HMM is referred to as Fragment-HMM.

3. The Fragment-HMM is used to sample a sequence of torsion angle pairs for the given protein sequence. An energy function is used to evaluate the generated decoys, and to direct the sampling process to the better decoys.

4. The generated decoys are fed back to produce more accurate estimations of local structural biases, a more accurate Fragment-HMM and thus, better decoys. This step is executed iteratively to increase the quality of the final decoys, until convergence.

This model differs from existing works as follows.

- Our Fragment-HMM model combines the very successful fragment assembly method [26] and the elegant FB5-HMM idea [13]. Rather than using the fragments as building blocks of the protein structure, we use fragments to produce local bias information. We use the directionality of the model local biases, and use HMM to explore the dependency among the adjacent residues. Unlike FB5-HMM, our Fragment-HMM is position specific as the hidden node sets for each position are mutually disjoint.

- Our Fragment-HMM naturally enables the Step 4 to re-sample decoys. Immediately, the readers would observe that this applies to obtaining fragments from a known

¹ In this paper, in order to obtain strictly fair comparisons, we did not use the fragments of [20]. We uniformly used the ROSETTA fragments.
structure. Thus this naturally enables homology modeling, threading, refinement, loop modeling, and consensus, unifying all these approaches under one roof.

- Step 4 is similar to that of primal and dual optimization process. The primal goal is to minimize the energy which is done by discriminating decoys with an energy function; and the dual process is done via sampling our Fragment-HMM to improve the estimation of torsion angles. Step 4 differs from the traditional fragment assembly methods that end with a population of decoys: some good and some bad. Our model does not stop here, but it iterates until convergence. Thus there are no other things, such as clustering, used in our approach. Note that Step 4 is also different from I-TASSER. I-TASSER is actually a two-step model, and the purpose of its second step is to remove the steric clashes of cluster centroids and refine the topology [31].

- The search space is narrowed down step by step. Monte Carlo is a popular technique for fragment-assembly-based protein structure prediction. However, Monte Carlo suffers from its low efficiency since it does not explore the characteristics of the search space. Specifically, for a protein of length \( n \), the search space size is \( O(200^n) \) if each sequence segment has 200 candidate structural motifs. This search space is unchanged in the whole Monte Carlo search process. In contrast, our Fragment-HMM narrows the search space after each iteration as the local structural biases are estimated more and more accurately.

We have implemented this theory in FALCON, Fragment-HMM approximating local bias and consensus. The first concern obviously is if FALCON will actually converge to some native-like decoys. We take all 6 proteins from the ROSETTA benchmark data used in [26]. FALCON converges 100% to within 6Å for all six proteins after only four iterations.

In order to make further comparisons to evaluate our model. We remove step 4, the iteration step, from FALCON. Then FALCON still generates significantly better results than ROSETTA for these 6 proteins. FALCON improves both the percentages of good decoys and the RMSD of the best decoys for five proteins out of six. For three of them, the percentages of good decoys are improved by more than 80%. Similarly FALCON, without step 4, generates better results than FBS-HMM as well as CRFSampler, by far.

These results suggest that succinct, accurate and flexible descriptions of local biases can significantly improve the quality of protein structure prediction.

### Methods

#### Parameterizing Backbones

A protein backbone of \( n \) residues consists of a sequence of atoms: \( N_1, Ca_1, C_1, \ldots, N_n, Ca_n, C_n \). The backbone conformation is fully specified by bond lengths, bond angles and torsion angles. Three types of torsion angles are defined: \( \phi_i, \psi_i \) and \( \omega_i \) for each residue \( i \) except for the \( N\)- and \( C\)-termini. The angle \( \omega_i \) is restricted to close to 180° or 0°. Bond lengths and bond angles are nearly constants. Therefore, we can parameterize a protein backbone with torsion angle pairs, i.e., the backbone of a protein of length \( n \) can be approximately reconstructed from \( n \) torsion angle pairs \( (\phi_1, \psi_1), \ldots, (\phi_n, \psi_n) \), assuming \( \phi_1 \) and \( \psi_1 \) are defined for notation simplicity.

#### Representing Local Bias of Torsion Angle Pairs

The local structural bias for residue \( i \) is represented by the joint distribution of its torsion angle pair \( (\phi_i, \psi_i) \). We use the Cosine model, a bivariate von Mises distribution over angular or directional space [21, 28]. The probability density function of Cosine model is specified by five parameters \( \kappa_1, \kappa_2, \kappa_3, \mu \) and \( \nu \):

\[
f(\phi, \psi) = c(\kappa_1, \kappa_2, \kappa_3)e^{\kappa_1 \cos(\phi - \mu) + \kappa_2 \cos(\psi - \nu) + \kappa_3 \cos(\phi - \psi + \nu)}
\]

where \( \mu \) is the mean value of \( \phi \), \( \nu \) is the mean value of \( \psi \), and \( c(\kappa_1, \kappa_2, \kappa_3) \) is a normalization constant:

\[
c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \left\{ I_0(\kappa_1)I_0(\kappa_2)I_0(\kappa_3) + 2 \sum_{p=1}^{\infty} I_p(\kappa_1)I_p(\kappa_2)I_p(\kappa_3) \right\}
\]

in which \( I_r(\kappa) \) is the modified Bessel function of the first kind and order \( r \) [1].

An alternative bivariate circular distribution is the Sine model [28]. Mardia et al [21] argued that Cosine model outperforms Sine model due to its ability to fit more closely a larger set of distributions.

Given a set of torsion angle pairs \( A = \{(\phi, \psi)\} \), we utilize a set of \( M \) Cosine models to parameterize these data. The Cosine models are combined into a mixture model, which can be formulated as:

\[
F(\phi, \psi) = \sum_{j=1}^{M} w_j f_j(\phi, \psi)
\]  \[1\]

where \( f_j, 1 \leq j \leq M \), denotes a Cosine model with parameters \( \theta_j = (\kappa_{j1}, \kappa_{j2}, \kappa_{j3}, \mu^j, \nu^j) \), \( w_j \) is the weight of model \( j \) with \( \sum_j w_j = 1 \). We employ an expectation-maximization (EM) algorithm to derive the most likely estimation of the parameters of the mixture model [21].

The number of Cosine models to fit these data is unknown in advance. It is vital to choose a suitable \( M \). Here we use Rissanen’s minimum description length (MDL) principle [19, 18] to determine the best value for \( M \), i.e., we choose \( M \) to minimize the following equation:

\[
\text{MDL}(A) = -2 \ln L(A, M) + 5 \ln(|A|)
\]  \[2\]

where \( L \) is the likelihood that \( M \) mixture models explain \( A \), and \( 5 \) is the number of parameters in each model.

#### Fragment-HMM: A Position Specific Hidden Markov Model

We use an HMM to capture the local dependencies among the adjacent residues. Unlike FBS-HMM, our HMM is position-specific, i.e., each residue is associated with a specific subset of hidden nodes and the subset of hidden nodes for all the residues are mutually disjoint.
Model Topology

An HMM is a directed graph, where the vertices denote the hidden nodes, and directed edges are used to capture transition and emission probabilities. For each residue \( i \), we obtain a set of possible hidden nodes, denoted as \( H_i \). Given two adjacent hidden node sets \( H_i \) and \( H_{i+1} \), a directed edge \((h, h')\) is created for each pair of hidden nodes \( h \in H_i \) and \( h' \in H_{i+1} \).

Each possible hidden node (denoted as \( h \)) has two types of emissions: a secondary structure type (denoted as \( S \)), and a torsion angle pair \( T = (\phi, \psi) \).

For the \( i \)-th amino acid, our position specific HMM describes the following joint probability:

\[
Pr.(S, T) = \sum_{h \in H_i} Pr.(T|S, h)Pr.(S|h)Pr.(h)
\]

where \( S \) is the secondary structure type for the \( i \)-th amino acid, and \( T \) is its torsion angle pair.

Fig. 1 shows an example of Fragment-HMM for five residues. Each residue is associated with a hidden node subset. As illustrated by Fig. 1, the hidden node subset \( H_i \) for residue one has two hidden nodes while \( H_2 \) for residue two has three possible hidden nodes. Each hidden node is associated with its own Cosine model.

Creating Hidden Nodes

Our HMM is both position and sequence specific. We do not assume we have training data available for the target sequences to be predicted. Therefore, the classical Baum-Welch [3] method cannot be applied to estimate the parameters. Here, we build the hidden nodes and estimate parameters with a position specific fragment library, hence the name Fragment-HMM.

The construction process has three steps. First, we parse the target sequence into segments with a sliding window of length \( \ell \) and step size one. Totally there are \( n - \ell + 1 \) segments. We index these sequence segments by \( 1, 2, ..., n - \ell + 1 \). For sequence segment \( i \), we can predict a subset of structural fragments via ROSETTA or FRazor [20]. Since the fragments via FRazor are shown to be better than ROSETTA’s fragments [20], in all experiments in this paper, we use only ROSETTA’s fragments so that the comparisons are fair. A structural fragment for segment \( i \) consists of a predicted torsion angle pair and secondary structure type for each residue from \( i \) to \( i+\ell-1 \).

The set of structural fragments is denoted as \( \mathcal{F} \). Second, we retrieve all predicted torsion angle pairs for residue \( i \) from fragments in \( \mathcal{F} \), and use the EM method to generate a set of Cosine models. Lastly, for each Cosine model, we create a hidden node.

Denote the Cosine model specified by hidden node \( h \) as \( m_h \). The density of \( m_h \) with parameters \((\phi, \psi)\) is written as \( f_h.(\phi, \psi) \).

Estimating Transition Probabilities

We describe the parameter estimation method without considering the secondary structures here for clarity. The secondary structure information is easily integrated into our framework.

The transition probabilities are estimated with fragment library \( \mathcal{F} \). Given a fragment \( q \in \mathcal{F} \) and a hidden node \( h \in H_i \), we first define the probability that \( h \) emits the torsion angle pair predicted by \( q \) for residue \( i \) as:

1. If \( q \) contains a predicted torsion angle pair \((\phi, \psi)\) for residue \( i \), we use the value of probability density function \( f_h \) with parameters \((\phi, \psi)\).
2. Otherwise, we define the probability to be 0.

We denote the above probability as \( g_h(q) \).

The joint probability \( Pr.(h' \in H_{i+1}, h \in H_i|q) \) for edge \((h, h')\) given fragment \( q \) is specified as: if a structural fragment \( q \) does not contain predicted torsion angles for both residue \( i \) and residue \( i+1 \), we let \( Pr.(h \in H_i, h' \in H_{i+1}|q) = 0 \); otherwise, we define \( Pr.(h \in H_i, h' \in H_{i+1}|q) \) as:

\[
Pr.(h \in H_i, h' \in H_{i+1}|q) = \frac{g_h(q)g_{h'}(q)}{\sum_{h'' \in H_i, h''' \in H_{i+1}} g_h(q)g_{h'}(q)} \tag{3}
\]

The above probability is normalized to ensure:

\[
\sum_{h \in H_{i+1}, h' \in H_i} Pr.(h \in H_i, h' \in H_{i+1}|q) = 1 \tag{4}
\]

Then, the joint probability \( Pr.(h \in H_i, h' \in H_{i+1}) \) can be calculated by:

\[
Pr.(h \in H_i, h' \in H_{i+1}) = \sum_{q \in \mathcal{F}} Pr.(h \in H_i, h' \in H_{i+1}|q)Pr.(q) \tag{5}
\]

where \( Pr.(q) \) can be estimated as the inverse of the number of the fragments in \( \mathcal{F} \) which contain predicted torsion angle pairs for both residue \( i \) and residue \( i+1 \).

Now, we are ready to compute the transition probability \( Pr.(h' \in H_{i+1}|h \in H_i) \) by:

\[
\frac{Pr.(h' \in H_{i+1}, h \in H_i)}{\sum_{h' \in H_{i+1}} Pr.(h \in H_i, h' \in H_{i+1})} \tag{6}
\]

The distribution of hidden nodes \( h \in H_i, 1 \leq i \leq n-1 \) is expressed by:

\[
Pr.(h \in H_i) = \sum_{h' \in H_{i+1}} Pr.(h \in H_i, h' \in H_{i+1}) \tag{7}
\]

We have specified a Fragment-HMM of order one for simplicity. A Fragment-HMM with higher order can be defined accordingly. We used order-7 and order-2 Fragment-HMM’s in FALCON.
**Sampling Protein Structure Conformation.** Given the position specific Fragment-HMM, to sample a backbone conformation is straightforward. Our procedure is simplified from the approach in [13] and it consists of two steps.

- **Sampling hidden nodes:** We first sample a sequence of hidden nodes, one for each residue. To sample a sequence of hidden nodes, we start by picking up a hidden node from set $H_1$: a node $h$ is picked according to the probability $Pr.(h), h \in H_1$. Given a hidden node $h$ for residue $i$, we sample a hidden node $h'$ for residue $i+1$, according to the transition probability: $Pr.(h' \in H_{i+1}|h \in H_i)$.

- **Sampling torsion angle pairs:** Then we sample a sequence of torsion angle pairs, one pair per residue according to the Cosine model specified by the respective hidden node. A backbone is constructed according to these torsion angles with ideal bond lengths and bond angles. Coupling the angle sampling process, we also sample a sequence of secondary structure types, which is useful for the energy function to evaluate the sampled structure.

We need to re-construct part of the structure. This is achieved through re-sampling a segment of the local backbone. That is, we re-sample the torsion angle pairs for residues from $i$ to $j$, rather than a whole sequence of torsion angle pairs, by the above two-step strategy.

**Conformation Optimization.** An energy function with non-local information helps to generate native-like conformations. We use the energy function of ROSETTA 2.1.0 (released Sept. 2006).

Initially, we sample a whole sequence of angle pairs, and construct a new 3D backbone structure from these angles. Then we re-sample a subsequence of torsion angle pairs for a given backbone structure and rebuild a new 3D backbone structure. If the new structure has an energy better than or equal to that of the previous structure, the new structure is accepted. Otherwise it is accepted with a certain probability by the Metropolis criteria. The process is repeated till the energy is converged, or the maximum number of iterations is reached.

**Iteratively Improving the Fragment-HMM.** An energy function directs the generation of a set of decoys. Based on these decoys, some infeasible Cosine models may be pruned, and some new Cosine models are to be formed. Then these new Cosine models can be integrated to build a more accurate Fragment-HMM with refined transition and emission probabilities. In turn, accurate Cosine models and HMM will result in better structures.

Therefore, we take a set of generated decoys as a position specific fragment library and use them as input to re-generate Cosine models and a new Fragment-HMM. This way, we generate a set of better decoys. Iterating, more and more accurate Fragment-HMM's and Cosine models are to be obtained providing that the energy function biases to the native-like structures.

**Data Set.** We use the six proteins that were used in previous studies [26, 16, 13]. They are displayed in Table 1: Protein A (code 1FC2), Homeodomain (code 1ENH), Protein G (code 2GB1), Cro repressor (code 2CRO), Protein L7/L12 (code 1CTF) and Calbindin (code 4ICB).

The position specific fragment library for each protein is obtained from ROSETTA. The ROSETTA’s fragment generation code is obtained from the recently-released ROSETTA package (version 2.1.0). Its structural fragments are selected from a set of 1020 protein chains, which are included in ROSETTA’s fragment generation module. We use ROSETTA’s energy function and its default settings.

| Table 1. | The number of Cosine models per residue. Column 2 is length. Column 3 is the number of $\alpha$-helices and $\beta$-strand of the target protein. Column 4-7 are numbers of residues with 1,2,3,4 Cosine models, respectively. Column 8 is the average number of Cosine models per residue. |
|----------|-------------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Name, PDB code | L | $\alpha, \beta$ | 1 | 2 | 3 | 4 | Ave. |
| Protein A, 1FC2 | 43 | 2.0 | 12 | 25 | 3 | 2 | 1.66 |
| Homeodomain, 1ENH | 54 | 2.0 | 24 | 24 | 6 | 0 | 1.21 |
| Protein G, 2GB1 | 56 | 1.4 | 28 | 21 | 7 | 0 | 1.63 |
| Cro repressor, 2CRO | 65 | 5.0 | 52 | 12 | 1 | 0 | 1.22 |
| Protein L7/L12, 1CTF | 68 | 3.3 | 50 | 14 | 3 | 1 | 1.34 |
| Calbindin, 4ICB | 76 | 4.0 | 47 | 23 | 3 | 3 | 1.50 |

**Torsion Angle Distributions.** We derived the most likely torsion angle distributions for each residue of the six proteins. A typical and concrete example for residues 13, 18, and 38 of protein 1FC2 is plotted in Fig. 2. Fig. 2a) contains two Cosine models, which are centered at $(-1.39, -0.18)$ and $(-1.90, 2.39)$. This indicates that this residue has two possible local bias preference: one corresponds to a $\beta$ secondary structure type, and the other corresponds to an $\alpha$ secondary structure type. As shown in Fig. 2(b), the local bias of residue 18 displays a dominant preference that can be described by a Cosine model centered at $(-1.64, 2.31)$. Note that there is only one distribution in Fig. 2(b) since the domains of $\phi$ and $\psi$ are circular. Residue 38 has four preferred Cosine models, which are centered at $(-1.12, -0.65), (-1.69, 2.27), (0.91, 0.71)$ and $(-1.83, 0.30)$. These four models have different weights $w$: the model centered at $(-1.12, -0.65)$ is the dominant one while the other models are relatively minor.

Fig. 2 demonstrates that the number and parameters of Cosine models are varying with residues. This is one of the major differences between our Fragment-HMM and FB5-HMM: in FB5-HMM (and CRFSampler) the number of distributions and the corresponding parameters are uniform for all the residues.

**Results**

FALCON is implemented with C++, on Linux.
We employ a mixture model to represent the local biases for each residue as the local biases of some residues cannot be parameterized perfectly by a single model. Table 1 lists the average number of Cosine models per residue for the six proteins generated from the fragment library. These proteins average 1.66 Cosine models per residue. Most residues have no more than two Cosine models. This observation confirms the fact that for a protein relatively few conformations are compatible with the local biases of all residues. Interestingly, the number of possible conformation clusters, $C$, is estimated to be $C = 1.6^n$ [27], which is consistent with the estimations given in [10, 29], coinciding with our study here. There are several different estimations: according to Levinthal paradox [17], the conformational space has a size at least $C = 3^n$ [34]; the number of possible conformation clusters can be estimated to be $C = 4^n$ via decomposing the Ramachandran map into 4 basins [23, 25]; and Hamelryck et al [13] assigned 75 possi-bile states for every amino acid by decomposing the ($\theta, \tau$) plane, which implies a conformation space of size $C = 75^n$. Compared with these estimations, our estimation is drastically smaller. This observation suggests that: (i) local structural biases can be accurately described and captured; and (ii) the conformation space to be searched is greatly reduced, and thus it may be possible to sample a native-like structure from a space where the conformations are compatible with the derived local biases.

**Local Bias Representation: Fragment-HMM versus Structural Fragments.** Local structural biases can be described by structural fragments or Fragment-HMM with Cosine models. We now investigate which approach is better in representing local bias for generating decoys. To answer this question, we compare FALCON, without Step 4, and ROSETTA (version 2.1.0) in terms of the percentage of good decoys (below 6Å to the native structure) and RMSD values of the best decoys. To do a fair evaluation, we used ROSETTA’s energy function for both programs. The input fragment libraries are generated by ROSETTA and are identical for both programs. The only difference between FALCON and ROSETTA, for this experiment, is the conformation generation strategy: in ROSETTA, local structures are copied directly from the predicted structural fragments; while in FALCON, the torsion angles are sampled from the derived Cosine models in the Fragment-HMM.

In this experiment, 1000 decoys are generated for each protein, by each of ROSETTA and FALCON. 6Å is used as the cutoff value for good decoys. The same criteria is used in [13]. Since all the decoys for both ROSETTA and FALCON are generated independently, the percentage of good decoys is expected not to fluctuate too much when more decoys are generated.

Observing Table 2, FALCON generates significantly more good decoys than ROSETTA. FALCON improves the percentage of good decoys for 1FC2, 2GB1, 2CRO, 1CTF and 4ICB, five out of six proteins. Especially for 2GB1, 1CTF and 4ICB, the improvements are from 53.7%, 14.3%, 19.9% to 93.4%, 25.6% and 46.3%, respectively. The quality of the best decoys for these five proteins 1FC2, 2GB1, 2CRO, 1CTF and 4ICB are improved, too. We obtain a structure that has RMSD only 0.557Å to the native structure for 1CTF as displayed in Fig. 3.

The percentage of good decoys and the quality of the best decoy for 1ENH decreased slightly. Overall, the results suggest that using Fragment-HMM with Cosine models to parameterize the local biases helps to generate better and significantly more native-like decoys.
different conditions, for the same 6 proteins in Table 2 and in that order, FB5-HMM [13] reports the best decoy accuracies to be 2.6Å, 3.8Å, 5.9Å, 4.1Å, 4.1Å, 4.5Å, respectively, and good decoy percentages (<6Å): 17.1, 12.1, 0.001, 1.09, 0.35, 0.38, respectively, from 100,000 decoys. The position-specific local biases obtained from ROSETTA’s fragment library and the higher order HMM, which is infeasible for FB5-HMM, have given the Fragment-HMM an decisive advantage over FB5-HMM.

**FALCON: Zero In On the Native Structure.** We use Cosine model to describe the local bias, and use energy function to capture global interactions. Energy function directs the search to discover native-like structures. Hence, energy function may help to reshape the local biases. This conjecture has led to Step 4 of FALCON: feed back the decoys as new positive-specific fragments, and iterate. The results are pleasing and illustrative.

Six iterations are executed for each protein, and 1000 decoys are generated at each iteration per protein. The first iteration takes as input the position specific fragment libraries from ROSETTA. The (i + 1)-iteration takes the set of decoys generated by the i-th iteration as input.

Table 3 displays the RMSD values of the decoys over the iterations for protein 2CRO. We observe that the RMSDs are converging. Initially, the percentage of decoys with RMSD larger or equal to 7Å is 15.7%. After two iterations, no decoys have a RMSD larger than 7Å. After 5 iterations, the RMSD values of 94.9% of the decoys converge to the range 3Å to 4Å. Also, we notice that both the best and the worst decoys disappear over the iterations. However the best decoys diminish far slower than the worst decoys. The decoy RMSD distributions for other proteins exhibit similar trends.

**Table 3.** RMSD distribution over iterations for protein 2CRO. Col. 2-7: Percentages of decoys with RMSD values in the corresponding intervals.

<table>
<thead>
<tr>
<th>#Iterations</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 3)</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[3, 4)</td>
<td>22.8</td>
<td>47.2</td>
<td>75.3</td>
<td>87.9</td>
<td>94.7</td>
<td>94.9</td>
<td></td>
</tr>
<tr>
<td>[4, 5)</td>
<td>41.5</td>
<td>45.4</td>
<td>24.5</td>
<td>12.0</td>
<td>5.3</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>[5, 6)</td>
<td>11.4</td>
<td>4.7</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>[6, 7)</td>
<td>8.5</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>[7, ∞)</td>
<td>15.7</td>
<td>1.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 shows the evolution of torsion angle pair distributions for residue 41 of protein 2CROA. Initially, the torsion angle pairs acquired from structural fragments display two clusters: one lies at β-strand area; and the other lies at α-helix area. Let’s call them β-strand distribution and α-helix distributions, respectively. The β-strand distribution is incorrect since this residue has a coil structure, and the α-helix is also inaccurate due to its large deviation and the small bias of its center. Interestingly, the subsequent iterations tend to correct the torsion angle distributions step by step.

---

**Table 2.** Decoy quality of ROSETTA and FALCON. Column 2-3: RMSD of the best decoy (Å) and percentage of the good decoys (RMSD < 6Å) for ROSETTA. Column 4-5: corresponding values for FALCON.

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>ROSETTA</th>
<th>FALCON</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Best</td>
<td>&lt;6.0Å(%)</td>
</tr>
<tr>
<td>Protein A, 1FC2</td>
<td>2.82</td>
<td>80.2</td>
</tr>
<tr>
<td>Homeodomain, 1ENH</td>
<td>1.52</td>
<td>94.4</td>
</tr>
<tr>
<td>Protein G, 2GB1</td>
<td>2.21</td>
<td>53.7</td>
</tr>
<tr>
<td>Cro repressor, 2CRO</td>
<td>2.56</td>
<td>70.4</td>
</tr>
<tr>
<td>Protein L7/L12, 1CTF</td>
<td>1.44</td>
<td>14.3</td>
</tr>
<tr>
<td>Calbindin, 4ICB</td>
<td>3.87</td>
<td>19.9</td>
</tr>
</tbody>
</table>

While one of the roots of our Fragment-HMM was FB5-HMM [13], Fragment-HMM has shed its skins to evolve into a significantly stronger model. FB5-HMM describes local biases by using 75 basins in the (θ, τ) plane, while FALCON uses only 1.6 Ramachandran basins on average. Under admittedly
At the second iteration, the initial two clusters are diminishing, and a new cluster centered at \((-1.82, -0.07)\) emerges. At the third iteration, the \(\beta\)-strand distribution disappears completely and the new cluster becomes dominate. Next the \(\alpha\)-helix distribution disappears at the fourth iteration. In the fifth and sixth iterations, the new cluster becomes denser and denser, which means the local bias for this residue is estimated more and more accurately. Finally, we obtain a distribution centered at \((-1.86, -0.13)\) after six iterations. Notice that there is a small gap between the center of this distribution and the native torsion angle pair \((-1.44, -0.63)\). This gap is inevitable since we adopt the standard bond lengths and standard bond angles in our structure generating model [15].

We use the minimum description length principle to determine the fitness of the \(\text{Cosine}\) models to the data. The shorter the description is, the more accurately the \(\text{Cosine}\) models are expected to fit the local biases data.

### Percentage of Good Decoys

Table 5 displays the percentage of good decoys. The percentage of good decoys increased steadily with iterations. All 6 proteins reach 100% of good decoys after 4 iterations. In particular, the percentage of good decoys for 1CTF and 4ICB is boosted up to 100% from 25.6% and 46.3%, respectively.

### Quality of the Finally Reported Decoys

The Fragment-HMM is not only used to sample decoys but also it is used to rank a given decoy. We rank each decoy from the fifth iteration according to the probability that the Fragment-HMM generates the decoy, and output the decoy with the highest probability as FALCON’s prediction. Table 4 displays the comparison of FALCON with ROSETTA. ROSETTA’s results are obtained by the clustering program of ROSETTA’s package with the default configuration.

As illustrated in Table 4, for five of the six benchmark proteins 1FC2, 1ENH, 2CRO, 1CTF and 4ICB, FALCON’s final prediction are better than that of ROSETTA, under the RMSD metric. For protein 2GB1, ROSETTA reports better prediction than FALCON.

### Table 5. Percentage of good decoys with RMSD below 6Å after each iteration.

<table>
<thead>
<tr>
<th>Target Protein</th>
<th># Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Protein A, 1FC2</td>
<td>94.3 98.5 100 100 100 100</td>
</tr>
<tr>
<td>Homeodomain, 1ENH</td>
<td>92.8 95.0 96.9 100 100 100</td>
</tr>
<tr>
<td>Protein G, 2GB1</td>
<td>93.4 96.4 100 100 100 100</td>
</tr>
<tr>
<td>Cro repressor, 2CRO</td>
<td>75.8 97.3 100 100 100 100</td>
</tr>
<tr>
<td>Protein L7/L12, 1CTF</td>
<td>25.6 68.8 97.0 100 100 100</td>
</tr>
<tr>
<td>Calbindin, 4ICB</td>
<td>46.3 90.5 99.3 100 100 100</td>
</tr>
</tbody>
</table>
Table 4. Quality of the finally reported decoys of ROSETTA and FALCON. Column 2-3: RMSD (Å) of the finally chosen decoys of ROSETTA and FALCON.

<table>
<thead>
<tr>
<th>Target Protein</th>
<th>ROSETTA</th>
<th>FALCON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein A, IFC2</td>
<td>3.660</td>
<td>3.652</td>
</tr>
<tr>
<td>Homedomain, 1ENH</td>
<td>2.717</td>
<td>2.464</td>
</tr>
<tr>
<td>Protein G, 2GB1</td>
<td>2.755</td>
<td>3.323</td>
</tr>
<tr>
<td>Cro repressor, 2CRO</td>
<td>3.997</td>
<td>3.477</td>
</tr>
<tr>
<td>Protein L7/L12, 1CTF</td>
<td>8.327</td>
<td>3.035</td>
</tr>
<tr>
<td>Calbindin, 4ICB</td>
<td>4.866</td>
<td>4.770</td>
</tr>
</tbody>
</table>

Conclusion

Based on the belief that the simplest theories and models are the better ones, we set out to look for a simple, clean, and unified theoretical model with which we can compute the protein structures.

We have presented our initial efforts of building such a theory, based on several previous ideas including the fragment assembly method and hidden Markov model sampling. Our new Fragment-HMM overcomes the difficulties of stiff structural fragments in sequence assembly approach, and the high dimensionality problem in the simple HMM approach. We have implemented our theory and produced clear-cut results.

With the iteration technique that is enabled by our Fragment-HMM, we have unified the procedures of fragment assembly, clustering, and final decoy selection.

Ideally, the quality of decoys should converge to its native structure over iterations. However, we notice that for example the RMSD values of the decoys for 2CRO converge to 3Å-4Å. We believe that the main reason is due to the lack of an accurate energy function at the backbone level to direct the search process, and an all-atom energy function for a refinement process. Another reason might be that we have used the idealized bond angle and bond length. This problem can be modeled into our model as well, allowing the sampling program to sample different bond angles and bond lengths as well.

We have alluded that our new model conveniently embodies other approaches such as homology modeling, threading, loop modeling, refinement, and consensus. We will also integrate our own FRazor fragment library [20] to improve accuracy. These projects are underway.

We thank David Baker and his ROSETTA group for developing and allowing us using the ROSETTA program. We also thank Dr. Yang Zhang for helpful discussions. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET::www.sharcnet.ca). This work has been supported by an NSERC grant OGP0046506, and the Canada Research Chair program.