Journal of Bioinformatics and Computational Biology
Vol. 9, No. 3 (2011) 383–398

DOI: 10.1142/S0219720011005574

Imperial College Press
www.icpress.co.uk

IN SEARCH OF THE PROTEIN NATIVE STATE WITH A PROBABILISTIC SAMPLING APPROACH

BRIAN OLSON*, KEVIN MOLLOY* and AMARDA SHEHU*,†,‡

*Department of Computer Science
George Mason University 4400 University Drive
Fairfax, VA 22030, USA

†Department of Bioinformatics and Computational Biology
George Mason University 4400 University Drive
Fairfax, VA 22030, USA
‡ashehu@gmu.edu

Received 31 January 2011
Revised 7 April 2011
Accepted 11 April 2011

The three-dimensional structure of a protein is a key determinant of its biological function. Given the cost and time required to acquire this structure through experimental means, computational models are necessary to complement wet-lab efforts. Many computational techniques exist for navigating the high-dimensional protein conformational search space, which is explored for low-energy conformations that comprise a protein's native states. This work proposes two strategies to enhance the sampling of conformations near the native state. An enhanced fragment library with greater structural diversity is used to expand the search space in the context of fragment-based assembly. To manage the increased complexity of the search space, only a representative subset of the sampled conformations is retained to further guide the search towards the native state. Our results make the case that these two strategies greatly enhance the sampling of the conformational space near the native state. A detailed comparative analysis shows that our approach performs as well as state-of-the-art *ab initio* structure prediction protocols.

Keywords: Protein native state; conformational ensemble; probabilistic search; tree-based projection-guided exploration; fragment library.

1. Introduction

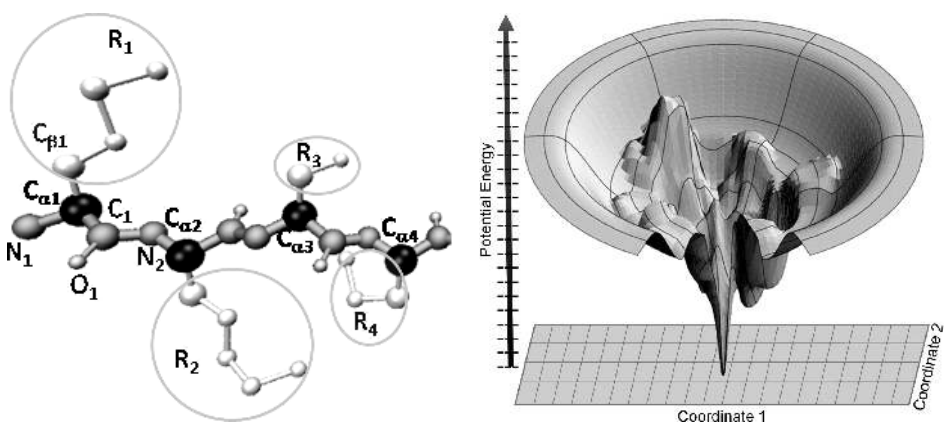
Protein molecules play a central role in biochemical processes in the cell and in various diseases. The spatial arrangement of a protein's atoms, referred to as a structure or conformation, is a key determinant of a protein's biological function. A protein molecule assumes specific conformations under physiologic (native) conditions to fit and interact with other molecules. The great number of novel protein

‡Corresponding author.

sequences with no known structures and the time and cost associated with resolving structures in the wet lab call for computational methods to complement wet-lab efforts.

The Anfinsen experiments showed that the amino-acid sequence governs the folding of a protein chain into a "biologically active conformation" under a "normal physiological milieu".[1] Research also shows that proteins are not rigid. The biologically active state is an ensemble of (native) conformations.[2–4] Probing this ensemble when employing only knowledge of the amino-acid sequence of a protein at hand continues to challenge computational structural biology.[5] Computing native conformations, however, is crucial in associating structural and functional information with novel protein sequences, engineering novel proteins, predicting protein stability, and modeling protein–ligand or protein–protein interactions.[6–8]

A protein chain consists of smaller building blocks, amino acids. Amino acids connect their backbone atoms to form a backbone chain, as shown in Fig. 1(a), with side-chain atoms dangling off the backbone. Tracking the various conformations of a protein chain involves exploring a vast conformational space of many dimensions. Many degrees of freedom (DOFs) are needed to represent a protein chain. One can reduce the level of detail through coarse-grained representations, such as backbone-only models, which track only conformations of the backbone. Once a native backbone conformation is found, computational techniques can be used to find physically relevant placements of the side chains.[9,10]



(a) (b)

Fig. 1. (a) Backbone atoms in the chain of four amino acids are labeled N (gray), C_α (black), C (gray), and O (silver). A peptide bond N_i-C_{i+1} links two amino acids together (i proceeds from N- to C-terminus, which refer to backbone N and C atoms not involved in peptide bonds). Atoms in white are labeled R for residue. There are 20 distinct residues or side chains in natural proteins. (b) We cross-sect energy landscape (grid on z axis) and projection of conformational space (grid on xy axis, two coordinates shown for visualization).

A chain of n amino acids contains $2n$ backbone dihedral angles that can be modified to obtain backbone conformations. The conformational space of interest is further narrowed when focusing on native conformations. These conformations are associated with the lowest energies in a funnel-like energy surface.[3] The totality of atomic interactions in a conformation results in a potential energy that determines the probability of that conformation to be populated under native conditions.[3]

The energy surface is rich in local minima, some of which are introduced by the current empirical energy functions that measure potential energy. By reducing the number of atoms modeled, coarse-grained representations and the energy functions that operate on them are more computationally appealing than all-atom functions but may introduce more inaccuracies. It is generally accepted, however, that modern functions do not significantly hamper a powerful conformational search.[11] A powerful search algorithm needs to populate a sufficient number of energy minima in order to probe the native state without spending impractical resources on irrelevant regions of the search space. These relevant regions are not known *a priori*.

We have recently proposed a probabilistic search algorithm that essentially addresses the question of where to devote exploration time.[12] The algorithm gathers information about regions of the conformational space and energy surface it explores. Discretizations of the explored conformational space and energy surface are employed to further guide the search in the conformational space.

The algorithm grows a search tree in conformational space, reconciling two goals: (i) expanding towards low-energy conformations while (ii) not oversampling geometrically similar conformations. The first goal guides the tree deep in the energy surface. The second goal grows the tree wide in conformational space. Both the energy surface and the explored conformational space are discretized in order to balance these two goals [see Fig. 1(b)]. The employment of discretization layers is inspired by sampling-based motion-planning work that balance exploration between coverage and progress toward the goal.[13–21]

In this paper, we propose two strategies to enhance the sampling of the conformational space near the native state within reasonable resources. First, an enhanced library of structurally diverse fragment configurations is used to assemble low-energy conformations and increase the complexity of the search space. Increasing the complexity appears counterintuitive to efforts to expedite search. The discretizations employed in our algorithm, however, allow implementing a second strategy to address complexity. Only a representative subset of the sampled conformations is maintained and employed to further guide the search for native conformations, effectively reducing the granularity of the conformational ensemble maintained in the search tree. Results show that these strategies enhance the sampling of the conformational space near the native state. This work is promising for large-scale proteomics applications, where the focus is on quickly probing the native state and then refining selected conformations in detailed biophysical studies.

The rest of this paper is organized as follows. A brief summary of related work is provided in Sec. 1.1. Our method is described in Sec. 2. Results follow in Sec. 3. The article concludes with a discussion in Sec. 4.

1.1. *Related work*

Where should a search algorithm devote its time? Regions that lead to the solution space are not known *a priori*, since stochastic search of a high-dimensional space affords only a local view. An effective search algorithm needs to strike the right balance between populating a large number of distinct low-energy regions and focusing further resources to regions likely to lead to the energy basin corresponding to the native state. Early work introduced the idea of a two-stage hierarchical exploration that searches the whole conformational space first and then narrows the search in a later stage to smaller regions with low energy and distinct geometry.[22]

Since the success of locating the energy basin in the second stage depends on the regions populated by the first stage, the emerging state-of-the-art template is to sample a large number of low-energy conformations in the first stage and build a broad map of the energy landscape.[6,23–27] Conformational clustering is then conducted to reveal distinct minima that constitute good starting points from which expensive (in finer detail) local searches in the second stage can reach the basin. Since the local searches employed in the second stage are computationally expensive, it is important that the first stage reveal few distinct local minima worth exploring in greater detail. It still takes weeks on multiple CPUs to obtain a large number of low-energy conformations potentially relevant for the native state.[6,23,24,26,27]

The first stage of the search and the analysis over the conformations are often independent of each other. As a result, computed conformations cannot be ensured to be geometrically distinct. Incorporating geometric diversity during the exploration is nontrivial, in part because it remains difficult to find meaningful conformational (reaction) coordinates on which to measure geometric diversity. Popular measures like least Root-Mean-Squared-Deviation (lRMSD) and radius of gyration (Rg) are confined to the analysis because they can mask away important differences. Specifically, work in Ref. 23 has shown that important minima can be missed even when employing Rg to select distinct conformations obtained at a current temperature to initiate MC trajectories at the next temperature in a Simulated Annealing MC search. Significant work in biophysics is devoted to finding effective reaction coordinates for proteins (cf. to Ref. 11).

Our recently proposed search algorithm incorporates analysis over explored regions of the conformational space and where they map in the energy surface in order to adaptively direct computational resources.[12,28] The analysis is carried out over discretizations of the explored space in order to guide the search over the continuous conformational space (a brief summary of the algorithm is provided in Sec. 2). As our methods and results shows, sampling of the conformational space near the native state is further enhanced if one increases the complexity of the

space while reducing the size of the conformational ensemble maintained in the search tree.

2. Methods

We first summarize the main steps of the algorithm proposed in Refs.12 and 28 (shown below). Given a protein sequence α, the goal is to obtain an ensemble Ω_α, where the lowest-energy backbone-only conformations are sufficiently close to the native state that they can be further refined to recover this state in all-atom detail.

Input: α, amino-acid sequence
Output: ensemble Ω_α of conformations

1: $C_{\text{init}} \leftarrow$ extended coarse-grained conf from α
2: AddConf(C_{init}, Layer$_E$, Layer$_{\text{Proj}}$)
3: **while** Time AND $|\Omega_\alpha|$ do not exceed limits **do**
4: $\quad \ell \leftarrow$ SelectEnergyLevel(Layer$_E$)
5: $\quad$ cell $\leftarrow$ SelectGeomCell(ℓ.Layer$_{\text{Proj}}$.cells)
6: $\quad C \leftarrow$ SelectConf(cell.confs)
7: $\quad C_{\text{new}} \leftarrow$ ExpandConf(C)
8: $\quad$ AddConf(C_{new}, Layer$_E$, Layer$_{\text{Proj}}$)
9: $\quad \Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$
10: **end while**

An explicit one-dimensional (1D) grid is defined over interval $[E_{\min}, E_{\max}]$, where $E_{\min}$ is the minimum energy over computed conformations, and $E_{\max}$ is the energy of the extended conformation. Energy levels ℓ are generated every δE units, which is set to a small 2 kcal/mol, so that the average energy $E_{\text{avg}}(\ell)$ over conformations in a level $\ell \in \text{Layer}_E$ captures the distribution of energies in ℓ well. This discretization is used to bias the selection towards conformations in lower energy levels through the quadratic weight function $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell)$. A level ℓ is selected with probability $w(\ell)/\sum_{\ell' \in \text{Layer}_E} w(\ell')$.

An implicit three-dimensional (3D) grid is associated with ℓ based on a uniform discretization of geometric coordinates. Three coordinates that capture extrema in a 3D structure are adapted from the ultrafast shape recognition (USR) features proposed in Ref. 29. A second weight function selects cells with fewer conformations as in $1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$, where `nsel` records how often a cell is selected, and `nconfs` is the number of conformations that project to the cell. Once a cell is chosen, the actual conformation selected for expansion is obtained at random over those in the cell.

A new conformation C_{new} that expands the tree (and grows the conformational ensemble Ω_α) from a selected C conformation is sampled through a Metropolis Monte Carlo technique that employs fragment-based assembly. The backbone dihedral angles of a selected fragment of three amino acids (trimer) in C are exchanged

with angles from a library of trimer configurations built from a nonredundant subset of known protein native structures. A total of $n-2$ (n amino acids in the chain) exchanges are evaluated according to the Metropolis criterion to obtain C_{new}.

Applications on different protein sequences reveal that the ensemble Ω_α of low-energy backbone conformations sampled for a sequence in a few CPU hours contains many conformations similar to the known native structure.[12] Comparisons with a Monte Carlo trajectory show the algorithm has a higher sampling capability.[12,28] However, detailed inspection of how the algorithm navigates the conformational space near the native state reveals that the ability to add low-energy conformations diminishes significantly with time. It becomes more difficult to find new low-energy conformations in underexplored regions of the conformational space. Moreover, the multitude of conformations retained in Ω_α imposes restrictions on execution time.

We propose two strategies to sample more low-energy conformations near the native state. An enhanced fragment library with greater structural diversity is used to assemble low-energy conformations and sample more conformations near the native state. To efficiently handle the ensuing vast conformational space, only a representative subset of the sampled conformations is maintained and employed to further guide the tree in conformational space. We now detail these strategies.

2.1. *Enhancing the trimer configuration library*

Fragment-based assembly has been incorporated into most state-of-the-art folding algorithms.[6,24–26,30,31] The diversity of the fragment library influences the quality of the assembled conformations.[31] The domain of the conformational search space is primarily determined by the fragment library. To provide the exploration a greater domain in which to search for native conformations, we propose an enhanced fragment library that essentially adds complexity to the conformational space.

The original fragment library (OFL) used in our recent work[12,28] contains trimer configurations, organized by trimer amino-acid sequence. A subset of nonredundant protein structures is extracted through the PISCES server[32] from the Protein Data Bank (PDB).[33] The subset contains only proteins that have $\leq 40\%$ sequence similarity, ≤ 2.5 Å resolution and R-factor ≤ 0.2. The 40% cutoff reduces the topologies that are overpopulated by similar protein sequences in the PDB. The remaining 6,000 protein chains are split into all overlapping trimers. The configurations, backbone dihedral angles, of these trimers are recorded in a fragment library indexed by trimer amino-acid sequences.

When a conformation is selected for expansion, each of the $n-2$ Monte Carlo moves propose to replace a trimer configuration with a configuration extracted from the fragment library. In the OFL, the candidate configurations are only those with matching amino-acid sequences to the trimer configuration chosen for replacement. Focusing only on trimer configurations with the same amino-acid sequence does not allow considering configurations that, while slightly different in sequence, may allow assembling novel conformations that meet the Metropolis criterion. Analysis of

protein structures reveals that proteins have similar native structures with as little as 15% sequence identity.[34] Excluding trimer configurations because their amino-acid sequence is not identical to that of the trimer selected for replacement restricts the conformational search space. This may prevent sampling novel conformations potentially relevant for the native state of the given protein sequence.

We propose to expand the conformational space available to our algorithm with an enhanced fragment library (EFL). Local features predicted from the given sequence α are used to design a structurally diverse high-quality library of configurations. The candidate trimer configurations in EFL are dependent on α. We refer to a specific library instance designed from a given α as EFL_α. Our construction of EFL_α biases toward trimer configurations that share features with those predicted from α. Essentially, EFL_α, whose construction is detailed below, allows selecting configurations that have *similar* (not necessarily identical) sequences to a trimer configuration selected for replacement. While placing a more diverse set of configurations at the disposal of the expansion routine in the algorithm, EFL_α does not contain more configurations than OFL. The configurations are limited to those that share secondary structure annotations with the annotation predicted on α.

EFL_α is constructed as follows. A multiple-sequence alignment (MSA) lists proteins that have similar sequences to the given α. PSI-BLAST[35] is then employed to analyze the MSA and yield for each position i in α a list of amino acids that can replace the amino acid at position i. The resulting position-specific profile for α reveals what alternative trimer sequences can be considered as similar to a trimer from position i to $i+2$. The configurations of these trimers, extracted from a nonredundant database of protein structures as detailed above, can be added as candidate configurations to those extracted for the trimer sequence from i to $i+2$. A filtering step improves the quality of the resulting configurations. Only configurations with the same secondary structure (as present in the known protein structures from which the trimer configurations are extracted) as that predicted for α with PSI-PRED[36] are added as candidate configurations for a trimer. Considering configurations of similar sequences but identical secondary structures has become very popular in *ab initio* structure prediction methods that employ fragment-based assembly.[6]

The resulting EFL_α represents (combinatorially) a conformational space that is both larger and more likely to share local structural motifs with the native structure of the sequence α. Results in Sec. 3 show that our algorithm is able to take advantage of this more complex conformational space to discover more conformations relevant for the native state than when employing the original fragment library.

2.2. *Reducing the granularity of the conformational ensemble* Ω_α

One of the benefits of employing trimer configurations to assemble conformations is that hundreds of thousands of conformations can be sampled this way in less than a day on one CPU. Maintaining all these conformations in the ensemble Ω_α introduces

both a practical memory limitation and unnecessary difficulty in selecting a conformation for expansion. Our recent work limits the exploration to three hours on one CPU in order to limit the size of the conformational ensemble.[12,28] Limiting the size of the conformational ensemble, however, limits the explorative power of the algorithm. Moreover, the enhanced fragment library increases the size of the conformational space to be sampled. In order to explore this broader space while not limiting the sampling capability of the algorithm, we change the purpose of the conformational ensemble Ω_α. Instead of maintaining every sampled conformation in Ω_α, the ensemble now maintains only a carefully selected subset of the sampled conformations through which to represent the explored conformational space.

By essentially reducing the granularity of Ω_α, the linear relationship between running time and memory requirements is removed. Each C_{new} generated is first evaluated for geometric novelty before being added to Ω_α. Clustering by lRMSD is computationally prohibitive to be performed after every sampled conformation C_{new}. Instead, we propose a less costly but effective strategy, which reduces the size of Ω_α by a factor of two or more (see Fig. 3 in Sec. 3). The strategy adds minimal computation overhead and does not impact the ability of the algorithm to sample low-energy conformations near the native state.

The granularity reduction exploits a feature of the energetic and geometric projection layers that is also exploited in the selection process: two conformations that lie in the same energy level ℓ and projection cell r will be geometrically similar (for some similarity threshold τ). Analysis shows that for the chosen granularity of 30 geometric cells per dimension (in the geometric projection grid) the value of τ is less than 1 Å lRMSD. If two conformations share the same ℓ and r, their similarity is determined using lRMSD. If the lRMSD is below a chosen τ (1 Å in our experiments), then only one of the conformations is retained; either the existing conformation is replaced or the new conformation is discarded with equal probability.

2.3. *Implementation details*

The algorithm is implemented in C++ and runs single-threaded on an AMD 2.66 GHz Dual-Core Opteron and all executions are run for 48 CPU hours. This runtime provides ample time to sample different combinations of fragment configurations and reduces the role of stochastic variations. The similarity threshold τ is set to 1 Å. All other parameters are as in previous work.[12,28]

3. Results

We apply the proposed strategies to enhance the sampling of the native state of two sets of target proteins listed in Tables 1 and 2. Section 3.1 compares the effectiveness of the original and enhanced fragment libraries, and Sec. 3.2 compares our method (using the enhanced fragment library) to published results from two established *ab initio* structure prediction methods. Section 3.3 demonstrates the degree to which

Table 1. PDB ID, fold, and length in amino acids are shown for each of the six proteins. PDB ID refers to a unique identifier associated with an experimentally-resolved native structure deposited for a protein in the PDB. The minimum lRMSD to the native structure is shown for both the original and enhanced fragment libraries. The final column compares the results from our method to results obtained by the state-of-the-art Rosetta structure prediction program.

				Minimum lRMSD (Å)		
Protein	PDB ID	Length	Fold	Original	Enhanced	Rosetta
wwD	1i6c	26	β	4.52	3.47	2.90
hbd2	1fd4	41	α/β	5.34	5.84	6.17
L20	1gyz	60	α	5.11	3.66	3.68
GB1	1gb1	60	α/β	6.89	6.31	2.67
Calbindin D_{9k}	4icb	76	α	5.76	4.70	2.73
pB119L	3gwl	106	α	10.32	8.30	9.13

Table 2. The minimum lRMSD to the native structure and the secondary structure Q3 score is given for each of the eleven proteins. Our results (Shehu) are presented along with those published by the Sosnick and Baker research groups.

			Q3 Score (%)			Minimum lRMSD (Å)		
PDB ID	Length	Fold	Shehu	Sosnick	Baker	Shehu	Sosnick	Baker
1ail	70	α/β	84	76	64	3.6	5.4	6.0
1aoy	78	α/β	73	82	89	4.7	5.7	5.7
1c8cA	64	α/β	66	86	67	7.4	3.7	5.0
1cc5	76	α	73	92	86	5.6	6.5	6.2
1dtdB	61	α/β	59	71	69	7.3	6.5	5.7
1fwp	69	α/β	64	70	68	5.9	8.1	7.3
1hz6A	67	α/β	70	80	87	6.2	3.8	3.4
1isuA	62	α/β	64	82	89	6.5	6.5	6.9
1sap	66	α/β	42	85	65	6.2	4.6	6.6
1wapA	68	β	62	80	68	7.9	8.0	7.7
2ezk	93	α	90	80	85	5.8	5.5	6.6

granularity reduction compresses the conformational ensemble Ω_α. Finally, Sec. 3.4 showcases the results of all-atom refinement on selected proteins.

3.1. *Effectiveness of the enhanced fragment library*

Table 1 lists the six targeted protein sequences selected to compare the effectiveness of the enhanced fragment library to that of the original fragment library: Pin1 Trp-Trp ww domain (wwD), human β-defensin 2 (hbd2), bacterial ribosomal protein (L20), immunoglobulin binding domain of streptococcal protein G (GB1), calbindin D_{9k}, and the African Swine Fever Virus pB119L protein. The proteins vary in length (number of amino-acids) and in known native topologies.

The ensemble Ω_α contains low-energy coarse-grained conformations that are good candidates for all-atom energetic refinement. In Table 1 we report the lowest lRMSD between the conformations in Ω_α and the known native structure for each

protein, calculated with both the original and the enhanced fragment library. Lower lRMSDs are obtained when employing the enhanced fragment library, which allows the search algorithm to assemble conformations that are closer in lRMSD to the native state than the original fragment library does.

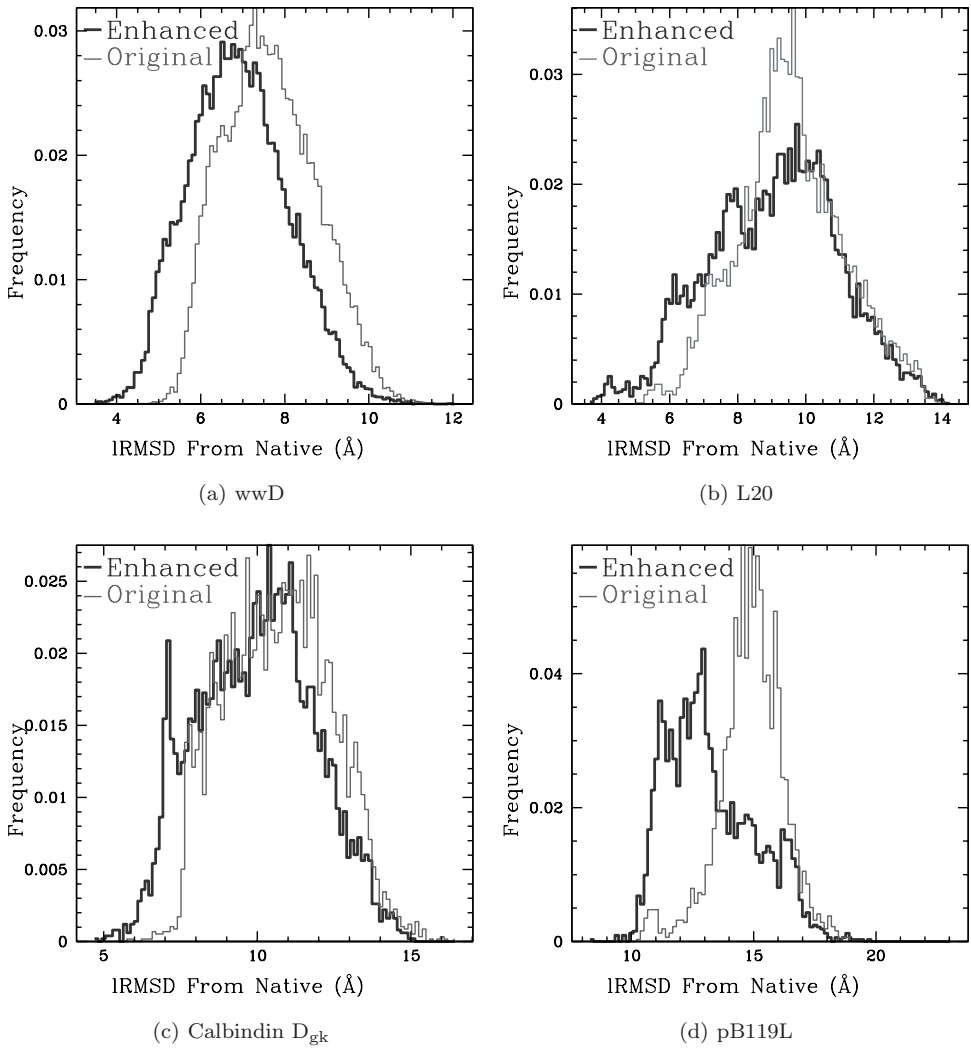
Table 1 also shows the lowest lRMSD obtained for each protein when employing the state-of-the-art Rosetta structure-prediction method.[6] To perform a relevant comparison between our algorithm and Rosetta, only the coarse-grained structure prediction component of Rosetta is employed. This component is initiated from each of the six target sequences and allowed to run for the same amount of time, 48 CPU h. Comparison of the lowest lRMSDs obtained with Rosetta to those obtained with our method (when employing the enhanced fragment library) shows that Rosetta significantly outperforms our method by more than 2 Å on only one protein, GB1. Our method obtains better results on three of the target proteins. Rosetta's performance on GB1 may be due to the coarse-grained energy function and the modulation of temperature during its coarse-grained search. Rosetta employs a fragment library similar to the enhanced library used in our method. However, Rosetta uses both 9-mer and trimer fragment lengths during its search. Our discussion in Sec. 4 lists a more accurate energy function and incorporation of temperature modulation as interesting directions for future research. In addition, the use of variable length fragments is an active area of research in our lab.

The enhanced fragment library, coupled with the reduction of the ensemble Ω_α, allows the search to enhance its sampling of the native state. Figure 2 shows histograms of lRMSDs of conformations in Ω_α from the known native structure, superimposing the histograms obtained when employing both the enhanced and original fragment library. These histograms are shown for only a few selected proteins (the entire list can be found in our recent work[37]). These histograms show that the enhanced fragment library allows the search algorithm to increase the number of computed conformations with lower lRMSD to the known native structure. The increase is significant for wwD, L20, calbindin, and pB119L; pB119L is a long protein chain used here to test the upper limits of our algorithm, with neither library yielding conformations below 8 Å lRMSD from the native structure.

The histogram representation in Fig. 2 is useful, because local maxima in the histograms correspond to potential clusters of conformations that can be detected with simple clustering techniques. The ensembles obtained with the enhanced fragment library for each protein contain more of these maxima at low lRMSDs. A technique interested in selecting a few conformations would obtain more native-like conformations if the enhanced fragment library is employed.

3.2. *Comparison to state-of-the-art methods*

Table 2 compares our results on eleven medium length target proteins to results published by the Sosnick[26] and Baker[38] research groups. Our algorithm samples conformations closer to the native structure for four out of the eleven target proteins



(a) wwD (b) L20

(c) Calbindin D_{gk} (d) pB119L

Fig. 2. (a)–(d) show the percentage of conformations in the ensemble Ω_α for a given lRMSD from the native structure. Data obtained with the enhanced fragment library are shown with a thick dark line and those obtained with the original fragment library are shown with a thin light line.

evaluated (1ail, 1aoy, 1cc5, and 1dtdB). In three cases (1isuA, 1wapA, and 2ezk), our results are in between those of the other groups, with no significant difference between out method and the best performing method. The four cases in which our method performs worse than the both of the other methods (1c8cA, 1dtdB, 1hz6A, and 1sap) are explained by inferior secondary structure prediction, shown by the Q3 scores in columns 4–6 (The Q3 score measures the percentage of amino acids in correctly predicted secondary structures). The enhanced fragment library heavily biases our method towards the secondary structure predicted during library

generation. In the case of 1c8cA, 1dtdB, 1hz6A, and 1sap the Q3 score produced by our method is at least 10% lower than the best performing method.

3.3. *Reduction of the ensemble Ω_α*

Reducing the granularity of the Ω_α ensemble significantly reduces the number of conformations retained in memory. The rate of memory consumption is now directly related to the algorithm's ability to discover geometrically novel conformations with similar energies. In practice, this enhancement allows exploration of the conformational space for days as opposed to hours. Figure 3 illustrates the relationship between runtime and memory requirement for the algorithm on wwD (similar results are observed for all other tested systems). Our ongoing research attempts to further quantify the diversity of Ω_α and develop novel geometric projection methods to enhance clustering of similar conformations.

3.4. *All-atom refinement*

Conformations sampled in Secs. 3.1 and 3.2 are coarse-grained models representing only the backbone of a protein structure. These coarse-grained models are typically further refined in all-atom detail for use in *ab initio* structure prediction and other biophysical studies. Here we present proof of concept all-atom refinements for six selected targets from Secs. 3.2 and 3.1. The refinement is carried out with the protocol available in the Rosetta software package,[6] which adds side-chain atoms to the backbone structure and performs a short Metropolis Monte Carlo energy minimization on the resulting all-atom conformation. Figure 4 shows the lRMSD of the refined structures from the known native structure. Figures 4(a)–4(c) show

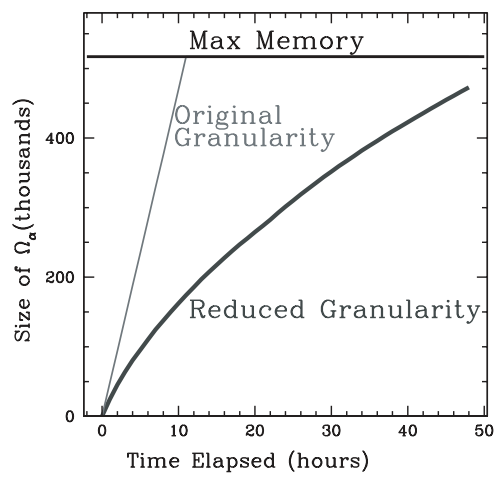
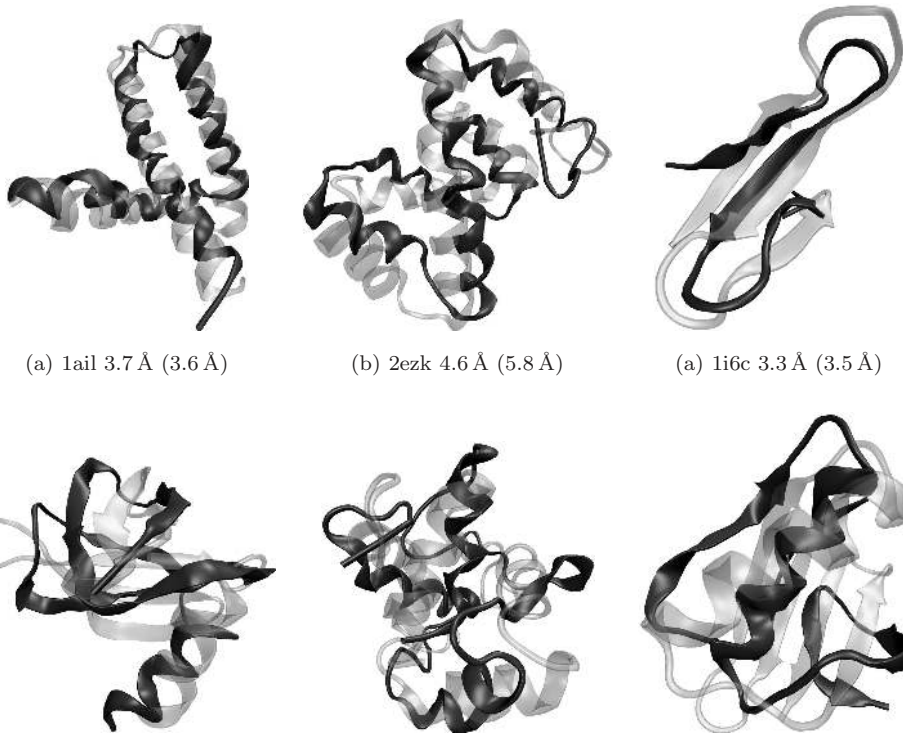


Fig. 3. Granularity reduction lowers the rate of growth of Ω_α (light line versus dark line). Black line shows maximum Ω_α size stored in 16 GB of memory.



(a) 1ail 3.7 Å (3.6 Å) (b) 2ezk 4.6 Å (5.8 Å) (a) 1i6c 3.3 Å (3.5 Å)

(d) 1c8c Å 7.1 Å (7.4 Å) (e) 1cc5 5.7 Å (5.6 Å) (f) gbd1 5.5 Å (6.3 Å)

Fig. 4. The conformation obtained from an all-atom refinement on the lowest lRMSD conformation obtained by our enhanced method is shaded dark and superimposed over the known native structure in light gray. The refined lRMSD to native is shown for each structure. The lRMSDs prior to the refinement are shown in parentheses.

examples where our method was able to sample conformations near the protein native state on a variety of lengths and fold topologies. The targets in Figs. 4(d)–4(f) represent areas for further improvement in our methods, as discussed in Sec. 4. In particular, we expect improved secondary structure prediction during fragment library generation to enhance performance of our algorithm on mixed α/β fold topologies.

4. Discussion

This paper investigates the effect of increasing the complexity of the conformational search space while decreasing the sample size required to represent it on a probabilistic search algorithm. We propose a more structurally diverse fragment library to provide our search algorithm with a larger conformational space. To efficiently handle the vast search space, we reduce the granularity of the conformational ensemble that the algorithm maintains to represent the space it has explored. Our results

show that these two strategies allow the search algorithm to enhance the sampling of conformations relevant for the native state.

Our recently introduced search algorithm[12,28] makes use of discretizations over projection layers of the energy surface and conformational space to guide its search toward diverse low-energy conformations. The algorithm is a first step toward rapidly computing coarse-grained native conformations from amino-acid sequence alone. The strategies proposed here address the need to enhance the sampling capability of the algorithm.

Our results show that the proposed strategies confer the algorithm with the ability to conduct longer, more detailed explorations. Results obtained with these strategies compare favorably against that of established state-of-the-art methods.[26,38] Refinement of all-atom models shows a high degree of accuracy for small to medium length α topologies and a promising first step for more complex structures.

The enhanced sampling capability shown in this work will allow the investigation of new selection-related weight functions, novel projection coordinates, and coarser representations of complex high-dimensional conformational spaces. Furthermore, state-of-the-art coarse-grained energy functions and a temperature modulation scheme will be pursued so that our methods may be applied to larger protein systems with more challenging native topologies. Secondary structure prediction methodologies may also be enhanced with an iterative improvement approach to improve the sampling bias in the fragment library.

References

1. Anfinsen CB, Principles that govern the folding of protein chains, *Science* **181**(4096):223–230, 1973.
2. Frauenfelder H, Sligar SG, Wolynes PG, The energy landscapes and motion on proteins, *Science* **254**(5038):1598–1603, 1991.
3. Dill KA, Chan HS, From levinthal to pathways to funnels, *Nat Struct Biol* **4**(1):10–19, 1997.
4. Huang YPJ, Montellione GT, Structural biology: Proteins flex to function, *Nature* **438**(7064):36–37, 2005.
5. Dill KA, Ozkan B, Shell MS, Weikl TR, The protein folding problem, *Annu Rev Biophys* **37**:289–316, 2008.
6. Bradley P, Misura KMS, Baker D, Toward high-resolution de novo structure prediction for small proteins, *Science* **309**(5742):1868–1871, 2005.
7. Yin S, Ding F, Dokholyan NV, Eris: An automated estimator of protein stability, *Nat Methods* **4**(6):466–467, 2007.
8. Kortemme T, Baker D, Computational design of protein-protein interactions, *Curr Opinion Struct Biol* **8**(1):91–97, 2004.
9. Canutescu AA, Shelenkov AA, Dunbrack Jr RL. A graph-theory algorithm for rapid protein side chain prediction, *Protein Sci* **12**(9):2001–2014, 2003.
10. Heath AP, Kavraki LE, Clementi C. From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes, *Proteins: Struct Funct Bioinf* **68**(3):646–661, 2007.
11. Clementi C, Coarse-grained models of protein folding: Toy-models or predictive tools? *Curr Opin Struct Biol* **18**:10–15, 2008.

12. Shehu A, An *ab initio* tree-based exploration to enhance sampling of low-energy protein conformations, in Trinkle J, Matsuoka Y, Castellanos JA, *Robotics: Science and Systems V*, pp. 241–248, Seattle, WA, USA, 2009.
13. Stilman M, Kuffner JJ, Planning among movable obstacles with artificial constraints, *Int J Robot Res* **12**(12):1295–1307, 2008.
14. Sánchez G, Latombe JC, On delaying collision checking in PRM planning: Application to multi-robot coordination, *Int J Robot Res* **21**(1):5–26, 2002.
15. Plaku E, Kavraki L, Vardi M, Discrete search leading continuous exploration for kinodynamic motion planning, in *Robotics: Science and Systems*, Atlanta, GA, USA, 2007.
16. Yang Y, Brock O, Efficient motion planning based on disassembly, in *Robotics: Science and Systems*, pp. 97–104, Cambridge, MA, 2005.
17. Ladd AM, Kavraki LE, Motion planning in the presence of drift, underactuation and discrete system changes, in *Robotics: Science and Systems*, pp. 233–241, Boston, MA, 2005.
18. Kurniawati H, Hsu D, Workspace-based connectivity oracle: An adaptive sampling strategy for PRM planning, in *WAFR*, Volume 47 of *Springer Tracts in Advanced Robotics*, pp. 35–51, New York, NY, 2006.
19. Rodriguez S, Thomas S, Pearce R, Amato NM, RESAMPL: A region-sensitive adaptive motion planner, in *WAFR*, Volume 47 of *Springer Tracts in Advanced Robotics*, pp. 285–300, 2006.
20. Choset H, *et al.*, *Principles of Robot Motion: Theory, Algorithms, and Implementations*, MIT Press, Cambridge, MA, 1st edn., 2005.
21. Jur P, van den Berg, Overmars MH, Using workspace information as a guide to non-uniform sampling in probabilistic roadmap planners, *Int J Robot Res* **24**(12):1055–1071, 2005.
22. Lee J, Scheraga HA, Rackovsky S, New optimization method for conformational energy calculations on polypeptides: Conformational space annealing, *J Comput Chem* **18**(9):1222–1232, 1997.
23. Shehu A, Kavraki LE, Clementi C, Multiscale characterization of protein conformational ensembles, *Proteins: Struct Funct Bioinf* **76**(4):837–851, 2009.
24. Brunette TJ, Brock O, Guiding conformation space search with an all-atom energy potential, *Proteins: Struct Funct Bioinf* **73**(4):958–972, 2009.
25. Bonneau R, Baker D, De novo prediction of three-dimensional structures for major protein families, *J Mol Biol* **322**(1):65–78, 2002.
26. DeBartolo J, Colubri A, Jha AK, Fitzgerald JE, Freed KF, Sosnick TR, Mimicking the folding pathway to improve homology-free protein structure prediction, *Proc Natl Acad Sci USA*, **106**(10):3734–3739, 2009.
27. Shehu A, Kavraki LE, Clementi C, Unfolding the fold of cyclic cysteine-rich peptides, *Protein Sci* **17**(3):482–493, 2008.
28. Shehu A, Olson B, Guiding the search for native-like protein conformations with an ab-initio tree-based exploration, *Int J Robot Res* **29**(8):1106–11227, 2010.
29. Ballester PJ, Richards G, Ultrafast shape recognition to search compound databases for similar molecular shapes, *J Comput Chem* **28**(10):1711–1723, 2007.
30. Haspel N, Tsai CJ, Wolfson H, Nussinov R, Reducing the computational complexity of protein folding via fragment folding and assembly, *Protein Sci* **12**(6):1177–1187, 2003.
31. Kolodny R, Koehl P, Guibas L, Levitt M, Small libraries of protein fragments model native protein structures accurately, *J Mol Biol* **323**(2):297–307, 2002.
32. Wang G, Dunbrack Jr RL, Pisces: A protein sequence culling server, *Bioinformatics* **19**(12):1589–1591, 2003.

33. Berman HM, Westbrook J, Feng Z, Gilliand G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The protein data bank, *Nucl Acids Res* **28**(1):235–242, 2000.
34. Fersht AR, *Structure and Mechanism in Protein Science, A Guide to Enzyme Catalysis and Protein Folding*, Freeman WH Co., New York, NY, 3rd edn., 631 pp, 1999.
35. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucl Acids Res* **25**(17):3389–33402, 1997.
36. Jones DT, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol* **292**(2):195–202, 1999.
37. Olson B, Molloy K, Shehu A, Enhancing sampling of the conformational space near the protein native state, in *BIONETICS: Intl Conf on Bio-Inspired Models of Network, Information, and Computing Systems*, Boston, MA, December 2010.
38. Meiler J, Baker D, Coupled prediction of protein secondary and tertiary structure, *Proc Nat Acad Sci USA* **100**(21):12105–12110, 2003.

Brian Olson received his B.S.E. in Computer Science from Princeton University in 2005. He is currently pursuing his Ph.D. in Computer Science at George Mason University. His research interests include high-dimensional search and optimization, evolutionary algorithms, and clustering analysis. His work applies these interests to problems in computational biology. Brian Olson is a member of the ACM.

Kevin Molloy is pursuing his M.S. in Computer Science at George Mason University where he also received his B.S. in Computer Science in 1998. He is currently working as an independent consultant specializing in capacity planning/performance management for very large database systems. His research interests include computational biology, analytical performance modeling, and parallel computation. He is a member of the ACM.

Amarda Shehu is an Assistant Professor in the Department of Computer Science at George Mason University. She holds affiliated appointments in the Department of Bioinformatics and Computational Biology and the Bioengineering Program at George Mason University. She earned her Ph.D. in Computer Science at Rice University in Houston, TX, in 2008, where she was also an NIH fellow of the Nanobiology Training Program of the Gulf Coast Consortia. Her research encompasses probabilistic search frameworks, evolutionary algorithms, and machine learning for problems in computational biology and biophysics. Dr. Shehu is a member of the IEEE and ACM.