Enhancing Sampling of the Conformational Space Near the Protein Native State

Brian Olson[1], Kevin Molloy[1], and Amarda Shehu[1,2]

[1] Department of Computer Science
[2] Department of Bioinformatics and Computational Biology
George Mason University, Fairfax, VA, 22030
amarda@gmu.edu

Abstract. A protein molecule assumes specific conformations under native conditions to fit and interact with other molecules. Due to the role that three-dimensional structure plays in protein function, significant efforts are devoted to elucidating native conformations. Many search algorithms are proposed to navigate the high-dimensional protein conformational space and its underlying energy surface in search of low-energy conformations that comprise the native state. In this work, we identify two strategies to enhance the sampling of native conformations. We show that employing an enhanced fragment library with greater structural diversity to assemble low-energy conformations allows sampling more native conformations. To efficiently handle the ensuing vast conformational space, only a representative subset of the sampled conformations are maintained and employed to further guide the search for native conformations. Our results show that these two strategies greatly enhance the sampling of the conformational space near the native state.

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

1 Introduction

The genomic revolution has resulted in millions of protein sequences for which little functional information is available [24]. Due to the central role that protein molecules play in biochemical processes in the cell, knowledge of the biological function of a protein molecule promises to advance our understanding of the living cell and various diseases. The spatial arrangement of a protein's atoms, interchangeably referred to as a structure or conformation, determines to a great extent biological function. A protein molecule assumes specific conformations under physiologic (native) conditions to fit and interact with other molecules.

Due to the role that structure plays in the biological function of a protein, significant efforts are devoted to elucidating native structures. The Protein Structure Initiative has pushed experimental efforts and yielded native structures of many proteins [27]. The great number of novel protein 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.

J. Suzuki and T. Nakano (Eds.): BIONETICS 2010, LNICST 87, pp. 249–263, 2012.

The Anfinsen experiments have shown that the amino-acid sequence governs the folding of a protein chain into a "biologically-active conformation" under a "normal physiological milieu" [2]. Anfinsen posited that, if one were to understand how the amino-acid sequence determines the biologically-active or native conformation, one could find such a conformation in silico. Research shows that proteins are not rigid and that the biologically-active state is an ensemble of (native) conformations [15, 12, 18]. Probing this ensemble when employing only knowledge of the amino-acid sequence of a protein at hand continues to challenge structural biology and has been proved NP-hard [13].

A protein chain consists of smaller building blocks, amino acids, each of which contains many atoms. 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 of each amino acid. 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 representational detail through coarse-grained representations, such as backbone-only representations, 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 [8, 17].

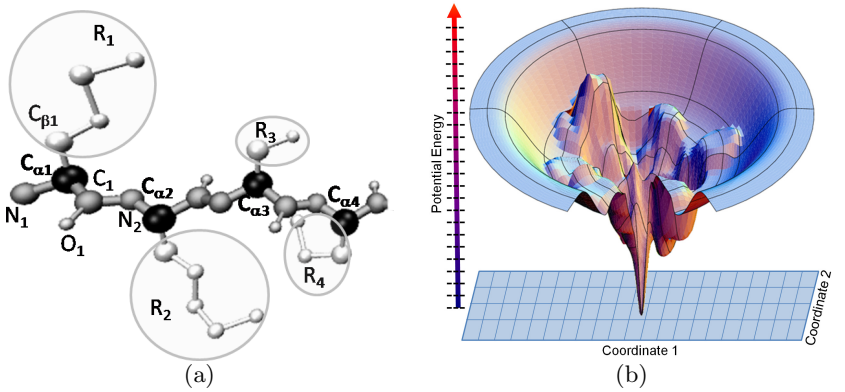

Fig. 1. (a) A chain of four amino acids is shown. Backbone atoms are labeled N (gray), C_α (black), C (gray), and O (silver). A peptide bond between N_i and C_{i+1} links two amino acids together (i proceeds from N- to C-terminus, which refer to backbone N and C atoms not involved in peptide bonds). Atoms in white are labeled R for residue. There are 20 distinct residues or side chains in natural proteins. Side chain atoms dangle off the backbone. (b) We cross-sect energy landscape (grid on z axis) and projection of conformational space (grid on xy axis, 2 coordinates shown for visualization).

Many coarse-grained representations have been proposed [10]. Even when focusing on the backbone, many dofs remain. A protein chain of n amino acids poses $2n$ backbone dihedral angles that can be modified to obtain backbone conformations. The conformational space of interest is narrowed when focusing on native conformations. These conformations are associated with the lowest energies in a funnel-like energy surface that underlies the protein conformational

space [12]. The totality of atomic interactions in a protein conformation results in a potential energy that is directly related to the probability of that conformation to be populated under native conditions [12].

The search for the low-energy native conformations is challenging, because the energy surface is rich in local minima. Some of the minima may be introduced by coarse-grained energy functions designed to operate on coarse-grained conformations. 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. All available energy functions are empirical. However, it is generally accepted that modern functions do not significantly hamper a powerful conformational search [10].

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. Without any a priori information, it is not possible to know what a sufficient number of minima is or where the relevant regions are. The only knowledge is that native conformations are associated with low energies. This is the main reason why it remains challenging for search algorithms to obtain native conformations. Computing these 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, 39, 21].

We have recently proposed a probabilistic search algorithm that essentially addresses the question of where to devote exploration time [31]. 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 essentially 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. Energies of computed conformations are partitioned into levels through a discretized one-dimensional (1d) grid. The grid helps select conformations associated with lower energy levels more often for expansion. The search keeps track of computed conformations in a low-dimensional projection space, which is discretized to select for expansion low-energy conformations that fall in under-explored regions (see Fig. 1(b)). The employment of discretization layers is inspired by sampling-based motion-planning work that uses decompositions, subdivisions, and projections of the search space to balance the exploration between coverage and progress toward the goal [35, 30, 28, 38, 23, 22, 29, 9, 36].

In this work, we focus on further enhancing the sampling of the conformational space near the native state while employing reasonable resources. This goal is crucial, given the potential inaccuracies inherent in a coarse-grained energy function and the fact that the native state is an ensemble of conformations. We identify two strategies to enhance sampling. We propose to increase the complexity of the conformational space while reducing the granularity of the

conformational ensemble maintained in the search tree. 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 search algorithm, however, allow exploiting the complexity without wasting resources. Moreover, a second strategy is proposed to efficiently handle the vast conformational space that ensues from employing the enhanced fragment library. Only a representative subset of the sampled conformations are maintained and employed to further guide the search for native conformations. Results show the proposed strategies enhance the sampling of the conformational space near the native state. Our work may be 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 section 1.1. Our method is described in section 2. Results follow in section 3. The article concludes with a discussion in section 4.

1.1 Related Work

Where should a search algorithm devotes 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. Ingredients for success were identified most notably in [25, 26]. Work in [25] 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.

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, essentially to build a broad map of the energy landscape [33, 6, 7, 5, 11, 32]. Clustering is then conducted over the conformations 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. In contrast, coarse graining is employed to reduce the computational cost of the first stage. It still takes weeks on multiple CPUs to obtain a large number of low-energy conformations potentially relevant for the native state [33, 6, 7, 11, 32]. 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.

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 non-trivial, 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 [33] 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 [10]).

The search algorithm we have recently proposed [31, 34] incorporates analysis over explored regions of the conformational space and where they map in the protein energy surface in order to adaptively determine where next to devote resources. The analysis is carried out over discretizations of the explored space in order to properly guide the search over the continuous conformational space (a brief summary of the essential ingredients of the algorithm is provided in section 2). As the description of our method and results shows, the sampling of the conformational space near the native state can be 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 [31, 34] (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: $\quad \Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$

An explicit 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 well the distribution of energies in ℓ. This discretization is used to bias the selection towards conformations in lower energy levels through the quadratic weight function $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell) + \epsilon$, where $\epsilon = 2^{-22}$ ensures a non-zero probability of selection for conformations with higher energies. A level ℓ is selected with probability $w(\ell) / \sum_{\ell' \in \text{Layer}_E} w(\ell')$.

An implicit 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 [3]. 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, since conformations in the same cell have similar energies (within δE).

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 non-redundant subset of known protein native structures. A total of $n - 2$ (n amino acids in the chain) exchanges are evaluated and accepted with probability 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 [31]. Comparisons with a Monte Carlo trajectory show the algorithm has a higher sampling capability [31, 34]. 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, further restricting the search.

We propose two strategies to help the exploration find more low-energy conformations near the native state. An enhanced fragment library with greater structural diversity is proposed 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 are maintained and employed to further guide the tree in conformational space. We detail each of these strategies next.

2.1 Enhancing the Trimer Configuration Library

In recent years fragment-based assembly has been incorporated into most state-of-the-art protein conformational search algorithms [16, 6, 7, 5, 11, 20]. The diversity of the fragment library influences the quality of the assembled conformations [20]. Indeed, 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 [31, 34] contains trimer configurations, organized by trimer amino-acid sequence. A subset of nonredundant protein structures is extracted through the PISCES server [37] from the Protein Data Bank (PDB) [4]. The subset contains only proteins that have $\leq$ 40% sequence similarity, $\leq$ 2.5Å resolution and R-factor $\leq$ 0.2. The

40% cutoff reduces the topologies that are over-populated 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 OFL, the candidate configurations are only those with the same amino-acid sequence as the sequence of 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 [14]. Excluding trimer configurations simply because their amino-acid sequence is not identical to that of the trimer configuration 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 employed to design a structurally-diverse high-quality library of configurations. The candidate trimer configurations in EFL are dependent on α, and we refer to a specific library instance designed from a given α as EFL_α. Our construction of EFL_α biases towards 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 containing a more diverse set of configurations at the disposal of the expansion routine in the algorithm, EFL_α does not contain more configurations that 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 [1] 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 [19] 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 (in the number of ways conformations can be assembled with the configurations in the library) a conformational space that is not only larger, but also more likely to share local structural motifs with the native structure of the given sequence α. Results in section 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 [31, 34]. 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 a 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 10 to 100 (see Fig 2 in section 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 actually 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 1Å (using lRMSD). For this value of τ, an arbitrary cutoff of one conformation per ℓ and r would suffice. However, the strategy we employ is not dependent on the chosen granularity of the geometric projection grid. Instead, if two conformations share the same ℓ and r, their similarity is determined using lRMSD. If the lRMSD is below a chosen τ (set at 1Å in our experiments), then only one of the conformations, selected at random, 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 with 4 GB of RAM. All reported times are based on CPU user time. The similarity threshold τ is set to 1Å. All other parameters are as in previous work [31, 34]. Results that compare the enhanced fragment library to the original library are obtained after 48 hours. This gives the algorithm ample time to sample different combinations of fragment configurations in the libraries and reduces the role of stochastic variations in our comparisons.

3 Results

We apply the proposed strategies to enhance the sampling of the native state of the six protein sequences listed in section 3.1. Section 3.2 compares the quality of the enhanced fragment library with the original one. Section 3.3 then shows the degree to which granularity reduction compresses the conformational ensemble Ω_α. Finally, section 3.4 shows how the proposed strategies enhance the sampling of conformations near the native state for each of the chosen protein sequences.

3.1 Target Proteins

Table 1 lists the six targeted protein sequences, 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 are selected to span different sizes (number of amino-acids) and known native topologies.

Table 1. PDB Id, fold, and number of 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.

Protein	wwD	hbd2	L20	GB1	Calbindin D_{9k}	pB119L
PDB Id	1I6C	1FD4	1GYZ	1GB1	4ICB	3GWL
Fold	β	α/β	α	α/β	α	α
Nr. AAs	26	41	60	60	76	106

3.2 Quality of the Enhanced Fragment Library

The quality of the fragment libraries is evaluated using the local-fit score introduced in [20]. The local-fit score measures the degree to which a fragment library fits a given native protein structure not used to construct the library. The chain of a given protein is broken into all its overlapping trimers. The configurations available for each resulting trimer in the library are then scanned to find the configuration closest (in terms of lRMSD) to the configuration of the trimer in the given native structure. The local-fit score associated with a given protein is

the average over the lowest lRMSDs obtained for all trimers defined over the chain of the protein. The local-fit score, referred to as lRMSD_f, is calculated for each of the six proteins listed above and is reported in Table 2. The scores obtained when employing the original fragment library are compared with those obtained when employing the enhanced fragment library.

Table 2. Local-fit lRMSD_f scores and lRMSDs between assembled conformations and known native structures are shown when employing the original fragment library (normal font) and enhanced fragment library (bold) for each of the six target proteins

Protein	wwD	hbd2	L20	GB1	Calbindin D_{9k}	pB119L
lRMSD_f (Å)	0.09 **0.08**	0.02 **0.04**	0.06 **0.06**	0.06 **0.06**	0.03 **0.02**	0.05 **0.03**
lRMSD (Å)	5.36 **5.02**	8.46 **6.21**	9.74 **7.94**	6.39 **9.01**	6.04 **5.78**	25.04 **10.69**

Table 2 shows (row 2) that overall lower local-fit scores are obtained when employing the enhanced fragment library. Similarly, if one assembles the conformation with the lowest lRMSD configurations from each library for each selected trimer in a given protein chain, the enhanced fragment library yields conformations that are closer to the known native structures. Lower lRMSDs from the native structure are reported for most of the proteins in Table 2 (row 3). This is not surprising, since the enhanced fragment library does not limit the search for fragment configurations to those with the same amino-acid sequence as the selected trimer. The high lRMSDs between the assembled conformations and the known native structures, especially for GB1 and pB119L, make the case that suboptimal fragment configurations are needed to assemble an optimal conformation. This further attests to the difficulty of assembling native conformations and the need for non-trivial search methods with powerful sampling capability.

3.3 Reduction of 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 exploring the conformational space for an indefinite period of time. Fig 2 illustrates the relationship between runtime and memory requirement for the algorithm on Calbindin D_{9k} (similar results are observed for all other tested systems).

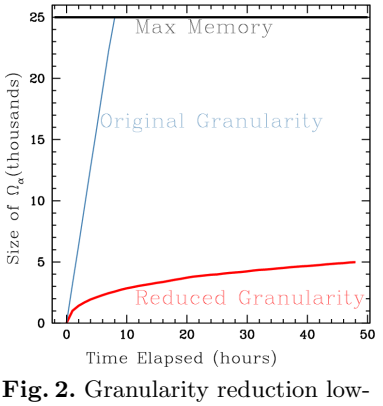

Fig. 2. Granularity reduction lowers the rate of growth of Ω_α (red line vs. blue line). The Black line shows maximum ensemble size stored in a 32-bit machine.

3.4 Effectiveness of Enhanced Fragment Library

The ensemble Ω_α contains low-energy coarse-grained conformations that are good candidates for all-atom energetic refinement. In Table 3 we report the lowest lRMSD between the conformations in Ω_α and the known native structure for each protein. The lowest lRMSDs are compared when employing the original versus the enhanced fragment library. Table 3 shows that lower lRMSDs are obtained when employing the enhanced fragment library. This library allows the search algorithm to assemble conformations that are closer in lRMSD to the native state compared to the original fragment library. Fig. 3 shows the lowest-lRMSD conformation obtained with the enhanced fragment library superimposed over the known native structure for each of the six targeted proteins.

Table 3 also shows the lowest lRMSD obtained on each protein when employing the state-of-the-art Rosetta structure prediction method [6]. To keep the comparisons similar, only the coarse-grained structure prediction component of

Table 3. The minimum lRMSD to the native structure is shown for each of the six target proteins. Data obtained when employing the original library is in normal font. Data obtained with the enhanced fragment library is highlighted in bold. The final row shows the data obtained when employing Rosetta [6].

Protein	wwD	hbd2	L20	GB1	Calbindin D_{9k}	pB119L
min-lRMSD (Å)	4.52 **3.47**	5.34 **5.84**	5.11 **3.66**	6.89 **6.31**	5.76 **4.70**	10.32 **8.30**
Ros-lRMSD (Å)	2.90	6.17	3.68	2.67	2.73	9.13

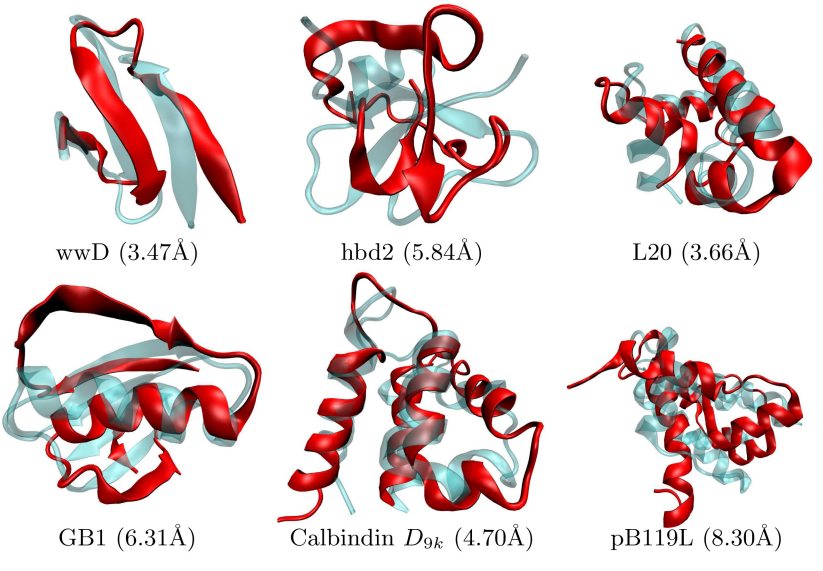

Fig. 3. The lowest lRMSD conformation obtained with the enhanced fragment library is drawn in red and superimposed over the known native structure in transparent blue

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 hours. 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 $\mathring{A}$ on only one protein, GB1. Our method obtains better results on three of the target proteins. The ability of Rosetta to perform better on GB1 may originate from the coarse-grained energy function and the modulation of temperature during its coarse-grained search. Our discussion in section 4 lists a more accurate energy function and incorporation of temperature modulation as interesting directions for future research.

The enhanced fragment library, coupled with the reduction of the conformational ensemble Ω_α, allows the search algorithm to enhance its sampling of the native state. Figure 4 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 ensemble. These histograms show that the enhanced fragment ensemble allows the search algorithm to increase the number of computed conformations with lower lRMSD to the known native structure. This increase is significant for wwD, L20, calbindin, and pB119L. pB119L is longer than the other proteins and is used here to test the upper limits of the search algorithm, with neither library allowing us to obtain conformations below 8Å lRMSD from the native structure.

The histogram representation in Figure 4 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.

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 search algorithm, recently introduced in [31, 34], makes use of discretizations over projection layers of the energy surface and conformational space to guide its search towards diverse low-energy conformations. The algorithm is a first step towards 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 capability to conduct a longer, more detailed exploration and improve its sampling of native conformations.

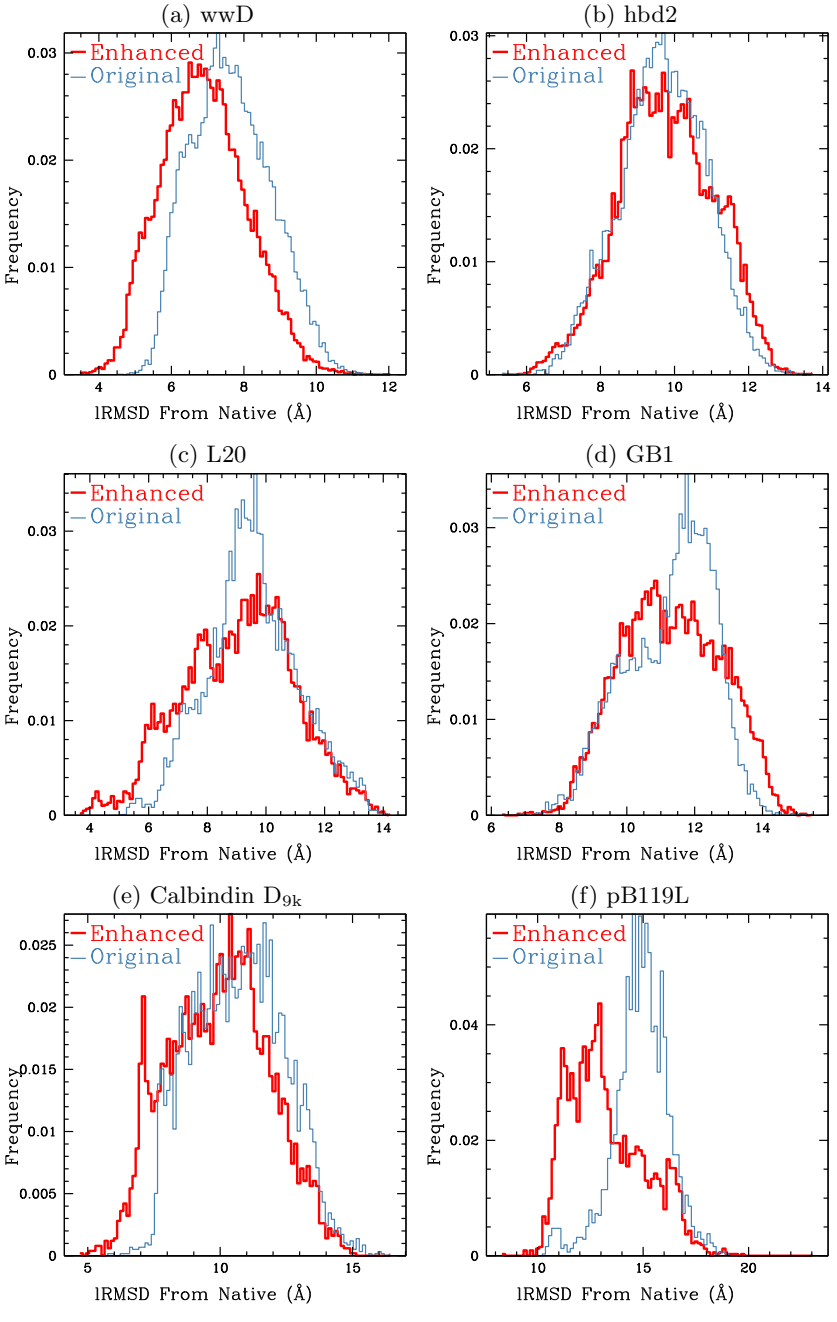

Fig. 4. (a-f) show the percentage of conformations in the ensemble Ω_α for a given lRMSD from the native structure. Data obtained with the enhanced fragment library are shown with a thick red line and those obtained with the original fragment library are shown with a thin blue line.

The work conducted in this paper lays the foundations for further future work. The enhanced sampling capability shown in this work will allow investigating different selection-related weight functions, novel projection coordinates, and coarser representations to further enhance the sampling capability of the algorithm on more complex high-dimensional conformational spaces of larger protein systems with challenging native topologies. Furthermore, state-of-the-art coarse-grained energy functions and a temperature modulation scheme will be pursued to further enhance the sampling capability of the method.

References

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25(17), 3389–33402 (1997)
2. Anfinsen, C.B.: Principles that govern the folding of protein chains. Science 181(4096), 223–230 (1973)
3. Ballester, P.J., Richards, G.: Ultrafast shape recognition to search compound databases for similar molecular shapes. J. Comput. Chem. 28(10), 1711–1723 (2007)
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliand, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucl. Acids Res. 28(1), 235–242 (2000)
5. Bonneau, R., Baker, D.: De novo prediction of three-dimensional structures for major protein families. J. Mol. Biol. 322(1), 65–78 (2002)
6. Bradley, P., Misura, K.M.S., Baker, D.: Toward high-resolution de novo structure prediction for small proteins. Science 309(5742), 1868–1871 (2005)
7. Brunette, T.J., Brock, O.: Guiding conformation space search with an all-atom energy potential. Proteins: Struct. Funct. Bioinf. 73(4), 958–972 (2009)
8. Canutescu, A.A., Shelenkov, A.A., Dunbrack Jr., R.L.: A graph-theory algorithm for rapid protein side chain prediction. Protein Sci. 12(9), 2001–2014 (2003)
9. Choset, H., et al.: Principles of Robot Motion: Theory, Algorithms, and Implementations, 1st edn. MIT Press, Cambridge (2005)
10. Clementi, C.: Coarse-grained models of protein folding: Toy-models or predictive tools? Curr. Opinion Struct. Biol. 18, 10–15 (2008)
11. DeBartolo, J., Colubri, A., Jha, A.K., Fitzgerald, J.E., Freed, K.F., Sosnick, T.R.: Mimicking the folding pathway to improve homology-free protein structure prediction. Proc. Natl. Acad. Sci. USA 106(10), 3734–3739 (2009)
12. Dill, K.A., Chan, H.S.: From levinthal to pathways to funnels. Nat. Struct. Biol. 4(1), 10–19 (1997)
13. Dill, K.A., Ozkan, B., Shell, M.S., Weikl, T.R.: The protein folding problem 37, 289–316 (2008)
14. Fersht, A.R.: Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding, 3rd edn. W.H. Freeman and Co., New York (1999)
15. Frauenfelder, H., Sligar, S.G., Wolynes, P.G.: The energy landscapes and motion on proteins. Science 254(5038), 1598–1603 (1991)
16. Haspel, N., Tsai, C., Wolfson, H., Nussinov, R.: Reducing the computational complexity of protein folding via fragment folding and assembly. Protein Sci. 12(6), 1177–1187 (2003)
17. Heath, A.P., Kavraki, L.E., Clementi, C.: From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes. Proteins: Struct. Funct. Bioinf. 68(3), 646–661 (2007)

18. Huang, Y.J., Montellione, G.T.: Structural biology: Proteins flex to function. Nature 438(7064), 36–37 (2005)
19. Jones, D.T.: Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292(2), 195–202 (1999)
20. 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)
21. Kortemme, T., Baker, D.: Computational design of protein-protein interactions. Curr. Opinion Struct. Biol. 8(1), 91–97 (2004)
22. Kurniawati, H., Hsu, D.: Workspace-based connectivity oracle: An adaptive sampling strategy for PRM planning. In: WAFR. Springer Tracts in Advanced Robotics. vol. 47, pp. 35–51 (2006)
23. Ladd, A.M., Kavraki, L.E.: Motion planning in the presence of drift, underactuation and discrete system changes. In: Robotics: Sci. and Syst., pp. 233–241 (2005)
24. Lee, D., Redfern, O., Orengo, C.: Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. 8(12), 995–1005 (2007)
25. Lee, J., Scheraga, H.A., Rackovsky, S.: New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. J. Comput. Chem. 18(9), 1222–1232 (1997)
26. Lee, J., Scheraga, H.A., Rackovsky, S.: Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46(2), 103–115 (1998)
27. Matthews, B.W.: Protein structure initiative: geeting into gear. Nat. Struct. Biol. & Mol. Biol. 14(6), 459–460 (2007)
28. Plaku, E., Kavraki, L., Vardi, M.: Discrete search leading continuous exploration for kinodynamic motion planning. In: Robotics: Sci. and Syst., Atlanta, GA, USA (2007)
29. Rodriguez, S., Thomas, S., Pearce, R., Amato, N.: RESAMPL: A Region-Sensitive Adaptive Motion Planner. In: WAFR. Springer Tracts in Advanced Robotics, vol. 47, pp. 285–300 (2006)
30. Sánchez, G., Latombe, J.-C.: On delaying collision checking in PRM planning: Application to multi-robot coordination. Int. J. Robot. Res. 21(1), 5–26 (2002)
31. Shehu, A.: An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations, Seattle, WA, USA (2009)
32. Shehu, A., Kavraki, L.E., Clementi, C.: Unfolding the fold of cyclic cysteine-rich peptides. Protein Sci. 17(3), 482–493 (2008)
33. Shehu, A., Kavraki, L.E., Clementi, C.: Multiscale characterization of protein conformational ensembles. Proteins: Struct. Funct. Bioinf. 76(4), 837–851 (2009)
34. 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)
35. Stilman, M., Kuffner, J.J.: Planning among movable obstacles with artificial constraints. Int. J. Robot. Res. 12(12), 1295–1307 (2008)
36. van den Berg, J.P., Overmars, M.H.: Using workspace information as a guide to non-uniform sampling in probabilistic roadmap planners. Int. J. Robot. Res. 24(12), 1055–1071 (2005)
37. Wang, G., Dunbrack, R.L.: Pisces: a protein sequence culling server. Bioinformatics 19(12), 1589–1591 (2003)
38. Yang, Y., Brock, O.: Efficient motion planning based on disassembly. In: Robotics: Sci. and Syst., Cambridge, MA, pp. 97–104 (2005)
39. Yin, S., Ding, F., Dokholyan, N.V.: Eris: an automated estimator of protein stability. Nat. Methods 4(6), 466–467 (2007)