An Ab-initio Tree-based Exploration to Enhance Sampling of Low-energy Protein Conformations

Amarda Shehu
Department of Computer Science, George Mason University, Fairfax, Virginia 22030
Email: {amarda}@cs.gmu.edu

Abstract— **This paper proposes a robotics-inspired method to enhance sampling of native-like protein conformations when employing only amino-acid sequence. Computing such conformations, essential to associate structural and functional information with gene sequences, is challenging due to the high-dimensionality and the rugged energy surface of the protein conformational space. The contribution of this work is a novel two-layered method to enhance the sampling of geometrically-distinct low-energy conformations at a coarse-grained level of detail. The method grows a tree in conformational space reconciling two goals: (i) guiding the tree towards lower energies and (ii) not over-sampling geometrically-similar conformations. Discretizations of the energy surface and a low-dimensional projection space are employed to select more often for expansion low-energy conformations in under-explored regions of the conformational space. The tree is expanded with low-energy conformations through a Metropolis Monte Carlo framework that uses a move set of physical fragment configurations. Testing on sequences of seven small-to-medium structurally-diverse proteins shows that the method rapidly samples native-like conformations in a few hours on a single CPU. Analysis shows that computed conformations are good candidates for further detailed energetic refinements by larger studies in protein engineering and design.**

I. Introduction

A globular protein molecule repeatedly populates its functional (native) state at room temperature after denaturation [1]. Despite this discovery in 1973 by Anfinsen, the problem of computing the conformations that comprise the protein native state from knowledge of amino-acid sequence alone continues to challenge structural biology [2]. Computing native conformations, however, is essential in associating structural and functional information with newly discovered gene sequences, engineering novel proteins, predicting protein stability, and modeling protein-ligand or protein-protein interactions [3]–[5].

Sampling native conformations is inherently difficult due to the vast high-dimensional conformational space available to a protein chain. The high-dimensionality challenge has drawn robotics researchers to adapt and apply algorithms that plan motions for articulated mechanisms with many degrees of freedom (dofs) to the study of protein conformations [6]–[11]. Though these methods often have to be adapted to deal with hundreds of dofs in protein chains (from dozens of dofs in articulated mechanisms), the motion-planning framework has allowed addressing the problem of computing paths from a given initial to a given goal protein conformation [6], [7].

The problem addressed in this work is the discovery of native conformations from knowledge of amino-acid sequence. No information is available on the goal conformations besides the energy landscape view that associates native conformations with lowest energies [12]. Energetic considerations complicate the search for native conformations. Interactions among atoms in a protein, which scale quadratically with the number of modeled atoms, give rise to an energy surface that can currently be probed only with empirical energy functions. A fundamental challenge in probing the native state is efficiently computing native-like conformations associated with the true global minimum (or competing minima) in an approximated energy surface constellated with local minima [12].

Faced with a vast space rich in local minima, some methods narrow the search space relevant for the native state by employing additional information about this state, often obtained from experiment [13]–[15]. Such information, however, is not available for millions of newly discovered protein-encoding gene sequences, nor is it easily obtained for novel sequences proposed in silico [16]. In such cases, ab-initio methods that employ only amino-acid sequence become very important.

In order to explore a vast conformational space, ab-initio methods often proceed in two stages: they first obtain a broad view of the conformational space, usually at a coarse-grained level of detail, to reveal candidate conformations that then undergo further refinement at a second stage [3], [17]. Coarse-grained representations of a protein chain are employed to reduce the number of dofs (from thousands to hundreds on small-to-medium proteins), and as a result, the dimensionality of the ensuing conformational space. The current generation of physically-realistic coarse-grained energy functions allows employing coarse-grained representations of protein chains as a practical alternative to all-atom representations without sacrificing predictive power [18].

Obtaining a broad view of the conformational space is crucial when postponing the computationally demanding all-atom detail and energetic refinement on coarse-grained conformations deemed to be native-like. Sampling a large number of low-energy conformations often entails enhanced sampling methods that build on Molecular Dynamics (MD) or Monte Carlo (MC) (cf. [19]). While MD-based techniques systematically search the conformational space, MC-based techniques often exhibit higher sampling efficiency [19]. Most notable among MC-based ab-initio methods are those that employ fragment assembly. These methods simplify the search space by using a limited move set of physical fragment configurations to assemble conformations [3], [14], [17].

Enhanced sampling on a simplified search space, coupled with realistic energy functions, have allowed fragment assembly methods to achieve high prediction accuracy of the native state, albeit at a time cost. It takes weeks on multiple CPUs to obtain a large number of low-energy conformations potentially relevant for the native state. It also remains difficult to ensure that computed conformations are geometrically-distinct and not representative of only a few regions of the conformational space [17]. Part of the difficulty lies in the inability to define a few conformational (reaction) coordinates on which to define distance measures. Popular measures like least Root-Mean-Squared-Deviation (lRMSD) and radius of gyration (Rg) can mask away differences among conformations [17].

The contribution of this work is a novel two-layered method to enhance the sampling of geometrically-distinct low-energy conformations. The method uses no information about the native state of a protein beyond amino-acid sequence. Employing a coarse-grained representation of a protein chain, the method focuses on efficiently obtaining diverse native-like conformations. The goal of the method is to serve as a filtering step that, in a matter of a few hours on a single CPU, reveals native-like conformations that can be further refined through detailed studies. From now on the method is referred to as FeLTr for Fragment Monte CarLo Tree Exploration.

As in fragment assembly, FeLTr assembles a conformation with configurations of short fragments compiled over a nonredundant database of protein structures. Employing fragments from the database (rather than random) improves the likelihood of assembling a physical conformation. A physically-realistic coarse-grained energy function estimates the energy of assembled conformations. The fragment assembly is implemented in a Metropolis MC framework, where the chain of a current conformation is scanned and new fragment configurations are proposed to replace old ones. Replacements that meet the Metropolis criterion are accepted, resulting in a new conformation. Using a limited move set of fragment configurations and a coarse-grained representation, FeLTr is able to rapidly obtain low-energy native-like conformations. Most importantly, the method obtains diverse native-like conformations through a novel tree-based exploration of the conformational space.

Inspired by tree-based methods in robotics motion planning, FeLTr grows a tree in conformational space, reconciling two goals: (i) expanding the tree towards conformations with lower energies while (ii) not oversampling geometrically-similar conformations. The first goal is warranted due to the fact that the desired native-like conformations are associated with the lowest energies in the energy surface sculpted by an amino-acid sequence. The second goal attempts to obtain a broad view of the conformational space near the native state by not oversampling geometrically-similar conformations.

To achieve the first goal, FeLTr partitions energies of computed conformations into levels through a discretized one-dimensional (1d) grid. The grid is used to select conformations associated with lower energy levels more often for expansion. The second goal is achieved by keeping track of computed conformations in a three-dimensional (3d) projection space using the recently proposed ultrafast shape recognition (USR) features [20]. The projection space is discretized in order to select for expansion low-energy conformations that fall in under-explored regions of the conformational space. After a conformation is selected for expansion, a short Metropolis MC trajectory summarized above is employed to expand the tree with a new low-energy conformation. This two-layered exploration, detailed in section II, is illustrated in Fig. 1.

The employment of the projection space in FeLTr is inspired by recent sampling-based motion-planning methods that make use of decompositions, subdivisions, and projections of the configurational space or the workspace of a robot to balance the exploration between coverage and progress toward the goal [21]–[25]. It is worth mentioning that, while the focus of this work is on protein chains, the projections employed in FeLTr are not tied to protein conformations. Since the projections rely only on geometry, they can be used for any articulated mechanism (manipulators, humanoid, modular robots) (see section IV for more on this point).

FeLTr is tested on sequences of seven proteins of different lengths (20-76 amino acids) – amounting to 40-152 dofs – and native topologies (β, α/β, α). Results show that FeLTr obtains compact low-energy conformations on all proteins. Clustering the lowest-energy conformations reveals that the native state is captured among the computed conformations. These conformations are good candidates for further refinement with detailed all-atom energy functions. The proposed method can serve as an efficient initial filtering step in larger detailed studies focused on extracting structural and functional properties of uncharacterized protein sequences [17], [26].

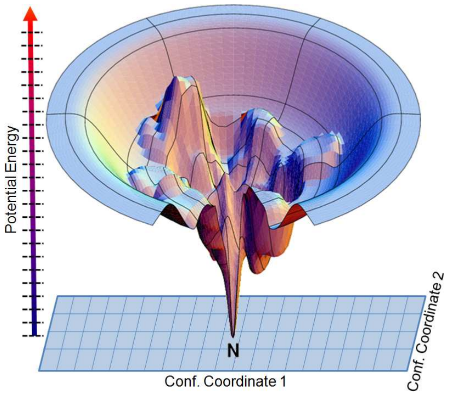


Fig. 1. According to the landscape view of protein folding, the native state, labeled N, is associated with the global minimum of a funnel-like energy surface [12]. The surface shown here is adapted from [12]. The proposed FeLTr method cross-sects this surface by discretizing potential energy values in a 1d grid, illustrated here with the z-axis. The axis is color-coded in a red-to-blue scheme to denote energy levels that reach lower values with the native state. The grid on the xy-plane discretizes the projection of the conformational space onto a few conformational coordinates. The illustration here shows two coordinates for visualization purposes. FeLTr uses three coordinates to describe projected conformations, as detailed in section II.

The following summary of related work further places the proposed FeLTr in context. The method is detailed in section II. Analysis of the conformations computed on the seven sequences chosen for testing is presented in section III. Discussion follows in section IV.

Related Work

Computational methods search conformational space systematically or at random [19]. Systematic searches rely on MD to sample conformations by numerically solving Newton's equations of motion. Since the solution accuracy dictates a small timestep (femtoseconds) between consecutive conformations in an MD trajectory, a broad view of the space demands multiple long trajectories. MC-based searches conduct biased probabilistic walks in conformational space [19].

Since both MD and MC are prone to converge to local minima in the energy surface, many methods build on them to enhance sampling. Such methods include simulated annealing, importance and umbrella sampling, replica exchange (also known as parallel tempering), local elevation, activation relaxation, local energy flattening, jump walking, multicanonical ensemble, conformational flooding, Markov state models, discrete timestep MD, and many more (cf. [19]).

Some of these methods narrow the search space by employing Nuclear Magnetic Resonance (NMR) data such as order parameters, residual couplings, and other NMR data [15], [27]. Such data are often incorporated in a new term in the energy function to construct a pseudo-energy function that biases the search towards conformations that reproduce the experimental data. Other methods employ an experimental structure representing the native state as a semi-rigid template and conduct a geometrically-constrained search around it for additional native-like conformations [13], [28].

Ab-initio methods employ only amino-acid sequence for a protein at hand. Some rely on sophisticated energy functions to guide MD trajectories to native-like conformations [29]. The most successful methods employ fragment assembly [3]. While the fragment length varies among methods, the basic process is to assemble conformations with physical fragment configurations extracted from a nonredundant database of protein structures [3], [14], [17]. Deciding on a suitable fragment length depends on the richness of the database to provide a comprehensive picture of fragment configurations. The current diversity of the Protein Data Bank (PDB) supports a minimum length of three amino acids [17].

While a limited move set of fragment configurations speeds up an MC-based exploration, there is no guarantee that assembled conformations are distinct or capture the native state. A large number of low-energy coarse-grained conformations are often computed to increase the probability that a few are sufficiently close to the native state that they will reach this state upon detailed energetic refinements [3], [17]. Since these refinements, often in all-atom detail, are computationally expensive, they can be conducted only on few conformations. Hence, it is important that the exploration reveal diverse coarse-grained conformations near the native state. It is currently difficult to find reaction coordinates on which to measure diversity [30]. Popular but lacking measures like lRMSD and Rg are not integrated in the exploration but largely confined to analysis over computed conformations [17].

FeLTr integrates coordinates proposed in [20] in its tree-based search. The tree is expanded in conformational space through an MC framework, keeping track of computed conformations in the low-dimensional projection space. MC has been used in sampling-based motion planing to escape local minima [31] and enhance sampling for closed chains with spherical joints [32]. The employment of the projection space in FeLTr is also inspired by motion-planning methods that decompose, subdivide, and project a robot's configurational space or workspace to balance the exploration between coverage and progress toward the goal [21]–[25].

II. FeLTr

In the usual MC framework, the probabilistic walk in the conformational space resumes from the last conformation computed. FeLTr enhances this framework by conducting a tree-based exploration of the conformational space. Given that only amino-acid sequence information is available, the root of the tree is an extended coarse-grained conformation. FeLTr then explores the conformational space iteratively, at each iteration selecting a conformation and then expanding it. While every expansion involves a short Metropolis MC trajectory, the important decision about which conformation to select for expansion depends on (i) the energy levels populated and (ii) the projection space covered by computed conformations. Pseudocode is given in Algo. 1. Sections describing the main steps in FeLTr are referenced at the end of each line.

Algo. 1 FeLTr: Fragment Monte CarLo Tree Exploration

```
Input: $\alpha$, amino-acid sequence
Output: ensemble $\Omega_\alpha$ of conformations
$C_{\text{init}} \leftarrow$ extended coarse-grained conf from $\alpha$          ▷II-A
$\mathrm{G}_E \leftarrow$ explicit 1d energy grid                                  ▷II-B.1
for $\ell \in \mathrm{G}_E$ do
    $\ell.\mathrm{G}_{\text{USR}} \leftarrow$ implicit 3d geom projection grid      ▷II-B.2
AddConf($C_{\text{init}}, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
while Time AND $|\Omega_\alpha|$ do not exceed limits do
    $\ell \leftarrow$ SelectEnergyLevel($\mathrm{G}_E$)                            ▷II-B.1
    cell $\leftarrow$ SelectGeomCell($\ell.\mathrm{G}_{\text{USR}}$.cells)          ▷II-B.2
    $C \leftarrow$ SelectConf(cell.confs)                                          ▷II-B.2
    $C_{\text{new}} \leftarrow$ MC_ExpandConf($C$)                                  ▷II-C
    if $C_{\text{new}} \neq$ NIL then                                  ▷MC succeeded
        AddConf($C_{\text{new}}, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
        $\Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$       ▷add conf to ensemble

AddConf($C, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
▷add $C$ to appropriate energy level in $\mathrm{G}_E$
$E(C) \leftarrow$ CoarseGrainedEnergy($C$)                                       ▷II-C.2
$\ell \leftarrow$ level in $\mathrm{G}_E$ where $E(C)$ falls into
$\ell$.confs $\leftarrow \ell$.confs $\cup \{C\}$
▷add $C$ to appropriate cell in geom grid associated with $\ell$
$P(C) \leftarrow$ USRGeomProj($C$)                                               ▷II-B.2
cell $\leftarrow$ cell in $\ell.\mathrm{G}_{\text{USR}}$ where $P(C)$ falls into
if cell = NIL then                                     ▷cell had not yet been created
    cell $\leftarrow$ new geom projection cell
    $\ell.\mathrm{G}_{\text{USR}}$.cells $\leftarrow \ell.\mathrm{G}_{\text{USR}}$.cells $\cup$ {cell}
cell.confs $\leftarrow$ cell.confs $\cup \{C\}$
```

A. Coarse-grained Representation of a Protein Chain

Only backbone heavy atoms and side-chain C_β atoms are explicitly represented. Employing the idealized geometry model, which fixes bond lengths and angles to idealized (native) values, positions of backbone atoms are computed from ϕ, ψ angles. These angles are set to $120°, -120°$ in an extended conformation (Algo. 1:1). Positions of C_β atoms are determined from the computed backbone as in [33].

B. Selection: Combination of Energy Layers and Low-dimensional Geometric Projections

1) Energy Layers: A 1d grid, G_E (Algo. 1:2), is defined on the segment $[E_{\min}, 0]$. $E_{\min}$ refers to the lowest expected energy on computed conformations (set to -200 kcal/mol on the tested proteins), and 0 refers to the highest energy. Since the Metropolis MC expansion quickly obtains negative-energy conformations (and conformations with nonnegative energies are not relevant because they are highly infeasible), it is not necessary to maintain a grid over nonnegative energies. Energy levels are generated every δE units. δE is set to a small value (2 kcal/mol), so that the average energy $E_{\text{avg}}(\ell)$ over conformations populating a specific energy level $\ell \in \mathrm{G}_E$ captures well the distribution of energies in ℓ.

This discretization is used to bias the selection towards conformations in the lower energy levels. The weight function $w(\ell)$ associated with an energy level $\ell \in \mathrm{G}_E$ is set to $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell) + \epsilon$, where $\epsilon = 1.0/2^{22}$ ensures that conformations with higher energies have a nonzero probability of selection. An energy level ℓ is then selected with probability $w(\ell)/\sum_{\ell' \in \mathrm{G}_E} w(\ell')$ (Algo. 1:7). This quadratic weight function biases the selection towards conformations with lower energies while allowing for some variation. Allowing higher-energy conformations to be selected provides `FeLTr` with the ability to jump over barriers in the energy surface. The Metropolis MC expansion also allows jumping over barriers.

2) Low-dimensional Geometric Projections: Conformations in a chosen energy level are then projected onto a low-dimensional space. Borrowing from the USR features proposed in [20], three projection coordinates are defined on each computed conformation: the mean atomic distance μ^1_{ctd} from the centroid (ctd), the mean atomic distance μ^1_{fct} from the atom farthest from the centroid (fct), and the mean atomic distance μ^1_{ftf} from the atom farthest from fct (ftf). The ctd, fct, and ftf atoms capture well-separated extremes of a conformation. So, the distribution of atomic distances from each extreme point (approximated with the first moment) is likely to yield new geometric information on a conformation.

The three projection coordinates capture overall topologic differences among conformations. As discussed in section IV, the coordinates are not specific to protein conformations but can be applied to any articulated mechanism. The coordinates allow introducing a second layer of discretization in `FeLTr`. The reason for the second layer is that conformations with similar energies may be geometrically different (a notion captured by entropy in statistical mechanics), and `FeLTr` aims to compute geometrically-distinct low-energy conformations.

Conformations in an energy level are partitioned in cells of the projection space to employ coverage in this space as a second criterion for selection. An implicit 3d grid, G_{USR}, is associated with each energy level (Algo.1:3-4), based on a uniform discretization of the projection coordinates. The selection is then biased towards cells with fewer conformations through the weight function $1.0/[(1.0 + \text{nsel}) \cdot \text{nconfs}]$, where nsel records how often a cell is selected, and nconfs is the number of conformations that project to the cell (Algo. 1:8). Similar selection schemes have been advocated in motion-planning literature as a way to increase geometric coverage during exploration [21], [24]. Once a cell is chosen, the actual conformation selected for expansion is obtained at random over those in the cell (Algo. 1:9), since conformations in the same cell have similar energies (within δE).

C. Expansion: Metropolis MC with Fragment Assembly

After a conformation is selected for expansion, its chain of N amino acids is scanned, defining N-2 consecutive fragments of three amino acids referred to as trimers. A conformation can now be updated by replacing configurations of its trimers. A trimer configuration consists of 6 ϕ, ψ angles defined over its backbone. The expansion procedure (Algo. 1:10) iterates N-2 times, at each iteration choosing a trimer at random over the chain. Upon choosing a trimer, a database of trimer configurations, whose construction is detailed in section II-C.1, is then queried with the amino-acid sequence of the trimer.

Of all configurations available for the trimer in the database, one obtained at random is proposed to replace the configuration in the current conformation. The reason for the random rather than consecutive iteration over the trimers in a chain is that an iterative scanning of the chain may result in local minima; that is, configurations are proposed but not accepted. The decision on whether to accept a trimer configuration (Algo. 1:11) is done under the Metropolis criterion, with the coarse-grained energy function defined in section II-C.2.

1) Database of Fragment Configurations: A PDB subset of nonredundant protein structures (as of November 2008) is extracted through the PISCES server [34] to contain proteins that have $\leq 40\%$ sequence similarity, ≤ 2.5 Å resolution and R-factor ≤ 0.2. Proteins studied in this work are removed from the database. The 40% similarity cutoff ensures that topologies that are over-populated by similar protein sequences in the PDB are not over-represented in the database. Around $6,000$ obtained protein chains are split into all possible overlapping trimers. The database maintains a list of configurations populated by each trimer over all extracted chains - a total of more than ten million configurations. No less than 10 configurations are populated for any trimer of each tested sequence.

2) Coarse-grained Energy Function: The function that evaluates the energy of a coarse-grained conformation, recently proposed in [17], is a linear combination of non-local terms (local terms are excluded since conformations are assembled with physical trimer configurations): $E = E_{\text{Lennard-Jones}} + E_{\text{H-Bond}} + E_{\text{contact}} + E_{\text{burial}} + E_{\text{water}} + E_{\text{Rg}}$. The $E_{\text{Lennard-Jones}}$ term is implemented after the 12-6

Lennard-Jones potential in AMBER9 [35], with a modification that allows a soft penetration of van der Waals spheres. The $E_{\mathrm{H-Bond}}$ term allows formation of local and non-local hydrogen bonds. The terms E_{contact}, E_{burial}, and E_{water}, implemented as in [29], allow formation of non-local contacts, a hydrophobic core, and water-mediated interactions.

The E_{Rg} term in this work penalizes a conformation by $(\mathrm{Rg} - \mathrm{Rg}_{\mathrm{PDB}})^2$ if the conformation's Rg value is above the $\mathrm{Rg}_{\mathrm{PDB}}$ value predicted for a chain of same length from proteins in the PDB. The predicted value fits well to the line $2.83 \cdot N^{0.34}$ [14], which is used to compute $\mathrm{Rg}_{\mathrm{PDB}}$ for each sequence of N amino acids. The E_{Rg} term penalizes non-compact conformations, since native-like conformations are compact and with a well-packed hydrophobic core. Moreover, $\mathrm{Rg}_{\mathrm{PDB}}$ and the Rg value of an extended conformation are used to define the boundaries of the projection space.

3) Metropolis Criterion: After replacing a trimer configuration in a selected conformation, the resulting energy is evaluated with the above energy function. The proposed replacement is accepted if it results in a lower energy (Algo. 1:10-11). Otherwise, it is accepted with probability $e^{-\beta \cdot \Delta E}$, where ΔE is the difference in energy after the replacement, and β is a temperature scaling factor. In this work, β is chosen to allow an energy increase of 10 kcal/mol with probability 0.1 so the tree is expanded with conformations that cross energy barriers.

D. Analysis of Computed Conformations

As shown in Algo. 1, conformations computed by FeLTr are gathered in the ensemble Ω_α. The distribution of energies of conformations in Ω_α is analyzed to obtain the average energy $\langle E \rangle$ and standard deviation σE. Let Ω^*_α denote the subensemble of conformations with energies no higher than $\langle E \rangle - \sigma E$. Ω^*_α is clustered with a simple leader-like algorithm [36], using a conservative cluster radius of 2.0 Å. The lowest-energy conformations of each cluster are offered by FeLTr as candidates for further detailed refinement. The results below show that this analysis reveals distinct clusters of native-like conformations that capture the native state.

The purpose of the analysis is to reveal possibly more than one energy minimum. Since exact quantum mechanics calculations cannot be afforded on long chains, empirical energy functions are used instead. These functions (like the one employed in this work) need to rank lower in energy those computed conformations that are more native-like. The lowest energy value reported, however, may not correspond to the most native-like conformation.

III. Experiments and Results

Since the conformational space available to a protein chain is high-dimensional, the ability of a method to reproduce conformations that populate the protein native state provides an important benchmark [26]. FeLTr is applied to seven structurally-diverse protein sequences of varying lengths. Comparing computed conformations with experimentally-available native structures of each protein reveals that FeLTr captures the native state. The results below, which benchmark FeLTr against a Metropolis MC simulation, show that FeLTr consistently obtains lower energies than the MC simulation.

The low-energy conformations obtained by FeLTr are analyzed not only for the presence of native-like conformations, but also for their geometric diversity. The results presented below show that FeLTr populates diverse energy minima significantly better than the MC simulation. While the native state is usually present among the highest-populated minima, other obtained minima contain compact low-energy conformations that differ on content of secondary structure segments or overall 3d arrangement of these segments.

A. Chosen Systems

The seven proteins chosen to test FeLTr, listed in Table III-A, include tryptophan cage (Trp-cage), Pin1 Trp-Trp ww domain (wwD), villin headpiece (hp36), engrailed homeodomain (eHD), bacterial ribosomal protein (L20), immunoglobulin binding domain of streptococcal protein G (GB1), and calbindin $\mathrm{D_{9k}}$. These proteins are chosen because they vary in size (number of amino acids) and ultimately number of dofs, native fold (3d global arrangement of local secondary structure segments), and are actively studied both in silico and in the wet lab due to the importance of their biological functions.

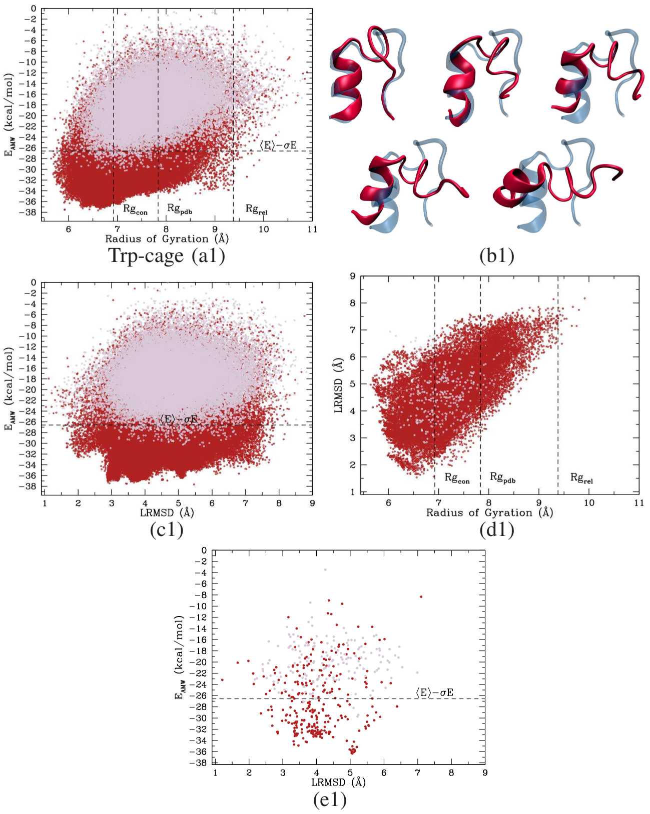
Protein	Trp-cage	wwD	hp36	eHD	L20	GB1	Calbindin $\mathrm{D_{9k}}$
Fold	α	β	α	α	α	α/β	α
Size	20	26	36	54	60	60	76
Dofs	40	52	72	108	120	120	152

FeLTr is implemented in C++, run on an Intel Core2 Duo machine with 4GB RAM and 2.66GHz CPU, and allowed to compute no more than $50,000$ conformations in no more than 3 hours. Limiting the number of conformations ensures that the exploration tree and ensemble Ω_α fit in memory. On small proteins like Trp-cage, wwD, and hp36, this number of conformations is reached in 40 minutes, 1 hour, and 2 hours, respectively. (Time grows quadratically with chain length due to the Lennard-Jones energy term; e.g., computing $50,000$ conformations takes ~ 36 hours on 200 amino-acid chains.)

B. Analyzing the Efficiency of FeLTr

The efficiency of FeLTr is estimated in comparison with a Metropolis MC simulation that is limited by the same time and number of conformations like FeLTr. To keep all other conditions similar, the MC simulation also employs the same fragment assembly to compute conformations in its trajectory and the same coarse-grained representation and energy function to calculate the energy of a computed conformation. This comparison allows to directly probe the effect of the FeLTr tree-based exploration guided by the novel two-layer discretization in sampling native-like conformations.

Fig. 2(a1) plots energies versus Rg values of computed conformations. Significantly lower energies are obtained with FeLTr (shown in red) than the MC simulation (shown in light purple). This result, also revealed when plotting energies versus lRMSDs of computed conformations from an



Trp-cage (a1)

(b1)

(c1)

(d1)

(e1)

Fig. 2. Plots superimpose Trp-cage MC-computed data in light purple over `FeLTr`-computed data in red. (a1) Energies of conformations are plotted vs. Rg values. The horizontal line marks the energetic cutoff. The three vertical lines show different Rg thresholds. (b1) `FeLTr`-computed conformations that meet the energetic criterion are clustered. Lowest-energy conformations of five main clusters (red) are superimposed over the native structure (transparent blue). (c1) Energies are plotted vs. lRMSDs from native. (d1) For conformations that meet the energetic criterion, lRMSDs from native are plotted vs. Rg values. (e1) For conformations that map to same cell of the projection space as the native, energies are plotted vs. lRMSDs from native.

experimentally-available native structure (Fig. 2(c1) for Trp-cage and 3(b2-7) for the other proteins), is not surprising. `FeLTr` guides the tree towards lower energies, whereas the MC simulation resumes from the last conformation generated.

The MC simulation employed for comparison is a powerful probabilistic walk that uses fragment assembly to make large hops in conformational space; that is, the lRMSD between two consecutive conformations can be significant. Figs. 2(c1) and 3(b2-7) show that the MC simulation comes close (in lRMSD) to the native structure of most tested proteins (`FeLTr` comes closer in Figs 2(c1) and 3(b3, b6)). However, `FeLTr` populates more energy minima, revealed when plotting energies of computed conformations versus their Rg values and versus lRMSDs from the native structure. This indicates the geometric projection layer helps to explore different conformational topologies with which to populate low energy levels. It is worth noting that, though Rg and lRMSD are not general conformational coordinates, they are useful to visualize 2d projections of the probed energy surface.

The inability of an MC trajectory to populate diverse energy minima is addressed in recent work by executing numerous trajectories. The trajectories are initiated from different carefully selected conformations in a simulated annealing framework [17]. This exploration, while offering a broad view of the conformational space near the native state, requires 1-3 weeks of computation on 50 CPUs [17]. The selection of conformations from which to initiate MC trajectories is seamlessly integrated in `FeLTr` through the tree-based exploration.

Native structures of many of the proteins presented here have been computed by detailed MD simulations that employ discrete timesteps [26]. Running times of such simulations, while resulting in high-quality native-like conformations, are limited by the dynamics of the proteins considered. While fast folders require ns-long MD simulations, slow folders may require longer than μs-long simulations (more than one week on one CPU). Other work that computes coarse-grained native-like conformations with MC trajectories (of similar length, $\geq 50,000$ consecutive conformations) employs long angle-based (rather than sequence-based) fragments [14].

C. Extracting Native-like Conformations with `FeLTr`

The goal of `FeLTr` is not to obtain single structures with high accuracy but to compute coarse-grained native-like conformations whose accuracy can be later improved through further refinements. Determining what makes a `FeLTr`-computed conformation native-like depends only on energetic considerations. The $\langle E \rangle - \sigma E$ cutoff defines a subensemble Ω^*_α of conformations that can be considered for further refinement. Employing other measures such as Rg or lRMSD from the native structures employs information that is not available from knowledge of amino-acid sequence only. Focusing on Ω^*_α reduces the number of conformations by more than 50%.

Fig. 2(a1) shows that conformations with energies no higher than $\langle E \rangle - \sigma E$ have diverse Rg values. The vertical lines in Fig. 2(a1) mark three Rg thresholds: $\mathrm{Rg}_{\mathrm{rel}}$ - the Rg value of an extended conformation, $\mathrm{Rg}_{\mathrm{PDB}}$ - the Rg value predicted for a chain of same length from the PDB, and $\mathrm{Rg}_{\mathrm{con}}$ - a smaller Rg value proposed in [14] for more compact conformations. Specifically, $\mathrm{Rg}_{\mathrm{con}} = 2.5 \cdot N^{0.34}$, where N is the number of amino acids in a protein sequence. The three vertical lines show that (i) almost all computed conformations are more compact than an extended conformation; (ii) the number of conformations proposed for refinement can be further reduced (down to $0.25|\Omega_\alpha|$) by discarding those with Rg $> \mathrm{Rg}_{\mathrm{PDB}}$; and (iii) the method obtains more compact conformations than rewarded by the coarse-grained energy function.

Clustering Ω^*_α allows offering only the lowest-energy conformations of the top-populated clusters for further refinement. These conformations are superimposed in opaque red over the transparent blue native structure of each considered protein - see Fig. 2(b1) and 3(a2-7). The native structures are obtained from the PDB: PDB id 1l2y for Trp-cage, 1i6c for wwD, 1vii for hp36, 1enh for eHD, 1gyz for L20, 1gb1 for GB1, 4icb for calbindin. With the exception of GB1, the native structure is captured among the top clusters. As Fig. 2(b1) shows, the top two clusters are very similar to the Trp-cage native structure (2-3 Å lRMSD), with some variability in the loop. Plotting

Rg values of conformations in Ω^*_α versus their lRMSDs from the native in Fig. 2(d1) shows a well-separated cluster of conformations. The cluster is around 2.0 Å in lRMSD from the native and around an Rg value of 6.5 Å, which is similar to the Rg value of 6.93 Å of the native structure.

D. The Projection Space Layer Helps Obtain Geometrically-distinct Conformations

Conformations that map to the same cell in the 3d grid over the projection space can still be significantly different (in terms of lRMSD) from one another. Fig. 2(e1), which plots energies versus lRSMDs from the native structure, shows that FeLTr obtains lower energies even for conformations that map to the same cell as the native structure. In addition, these conformations have diverse lRMSDs, up to 7 Å.

Projection coordinates capture overall topology, with fine structural details handled by the energy function. For example, the GB1 conformation representative of the top cluster projects to the same cell as the native structure. The native topology is captured, but the β-sheets are not fully formed. Since β-sheets arise from non-local interactions, they cannot be captured at the fragment level but through an energy function. Improvements in energy functions to capture non-local backbone pairings are the subject of much research [3], [26].

IV. Discussion

The coarse graining in FeLTr is based on the backbone-based theory of protein folding [37]. Other work that also employs coarse graining shows that geometry presculpts the protein energy surface [38]. FeLTr leverages the role of geometry in shaping the protein energy surface by employing both geometry and energy to guide its tree-based exploration.

Since time grows quadratically with the number of atoms (due to the Lennard-Jones term), coarse graining (which reduces the number of atoms) and the focus on computing diverse low-energy conformations make FeLTr particularly effective to handle high-dof chains. Coarser protein representations like C_α traces may extend applicability to longer chains.

Coarse graining can also benefit methods that search high-dimensional spaces of articulated robots. The importance of coarse graining is indeed starting to emerge in sampling-based motion planning. Together with work in [21], which shows benefits in using different layers of granularity (from geometric to kinematic to dynamic), FeLTr also supports the use of reduced models to address high dimensionality.

The projection coordinates employed here are not proposed as general reaction coordinates on which to project the energy surface. Finding such coordinates remains the subject of much research in computational biology [30]. Rather, these coordinates are a first attempt towards integrating a projection space in the exploration of conformational space. Directions for future research include the design of novel projections and parallel implementations to further enhance the exploration.

Since the USR projections rely only on geometry, they are not tied to protein chains but can apply to any articulated mechanism. In particular, sampling-based motion planners like DSLX [21], PDST [24] and [39] that rely on low-dimensional projections can potentially benefit from using USR projections to effectively explore high-dimensional spaces.

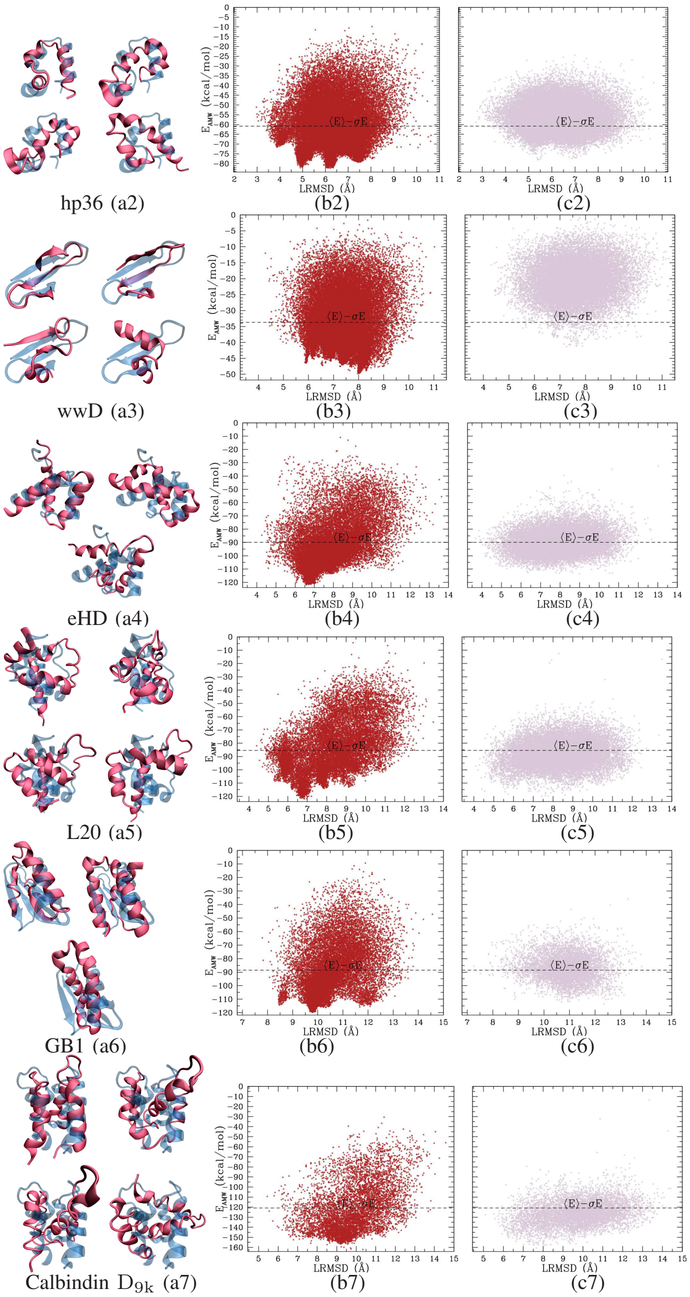


Fig. 3. (a2-7) FeLTr-computed conformations that meet the energetic criterion are clustered. The lowest-energy conformations of the most populated clusters are superimposed in opaque red over the native structure drawn in transparent blue. Energies of conformations are plotted vs. their lRMSDs from the native structure in red for FeLTr-computed conformations in (b2-7) and light purple for MC-computed conformations in (c2-7).

FeLTr also offers an interesting insight on how to generate valid samples for articulated mechanisms. For instance, since random sampling of dofs in manipulation planning often results in self-colliding configurations, the equivalent of a fragment database can be employed to extract good configurations for different fragments. Work in [40] has also proposed the usage of chain fragments in sampling valid configurations.

The $\mathrm{Rg}_{\mathrm{PDB}}$ employed to obtain compact conformations does not capture proteins with diverse functional states. Different values of Rg thresholds, obtained from experiment or defined systematically over a range as in [17], can be employed in future work to extend applications on such proteins.

`FeLTr` makes a first step towards rapidly computing coarse-grained native-like conformations from sequence. Analysis shows the native structure is among computed conformations. The lowest-energy conformations are good candidates for further refinement in all-atom detail in order to associate structural and functional information with novel sequences.

Acknowledgment

The author thanks Erion Plaku, Srinivas Akella, and the anonymous reviewers for comments on this paper.

References

[1] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[2] R. H. Lathrop, "The protein threading problem with sequence amino acid interaction preferences is NP-complete," *Protein Eng*, vol. 7, no. 9, pp. 1059–1068, 1994.

[3] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[4] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nat Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[5] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Curr. Opinion Struct. Biol.*, vol. 8, no. 1, pp. 91–97, 2004.

[6] N. M. Amato, K. A. Dill, and G. Song, "Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures," *J. Comp. Biol.*, vol. 10, no. 3-4, pp. 239–255, 2002.

[7] M. S. Apaydin, D. L. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe, "Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion," *J. Comp. Biol.*, vol. 10, no. 3-4, pp. 257–281, 2003.

[8] J. Cortes, T. Simeon, R. de Angulo, D. Guieysse, M. Remaud-Simeon, and V. Tran, "A path planning approach for computing large-amplitude motions of flexible molecules," *Bioinformatics*, vol. 21, no. S1, pp. 116–125, 2005.

[9] A. Lee, I. Streinu, and O. Brock, "A methodology for efficiently sampling the conformation space of molecular structures," *J. Phys. Biol.*, vol. 2, no. 4, pp. 108–S115, 2005.

[10] K. M. Kim, R. L. Jernigan, and G. S. Chirikjian, "Efficient generation of feasible pathways for protein conformationa transitions," *Biophys. J.*, vol. 83, no. 3, pp. 1620–1630, 2002.

[11] I. Georgiev and B. R. Donald, "Dead-end elimintation with backbone flexibility," *Bioinformatics*, vol. 23, no. 13, pp. 185–194, 2007.

[12] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nat. Struct. Biol.*, vol. 4, no. 1, pp. 10–19, 1997.

[13] A. Shehu, C. Clementi, and L. E. Kavraki, "Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations," *Proteins: Struct. Funct. Bioinf.*, vol. 65, no. 1, pp. 164–179, 2006.

[14] H. Gong, P. J. Fleming, and G. D. Rose, "Building native protein conformations from highly approximate backbone torsion angles," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 45, pp. 16 227–16 232, 2005.

[15] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, "Simultaneous determination of protein structure and dynamics," *Nature*, vol. 433, no. 7022, pp. 128–132, 2005.

[16] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[17] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Funct. Bioinf.*, 2009, in press.

[18] C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?" *Curr. Opinion Struct. Biol.*, vol. 18, pp. 10–15, 2008.

[19] W. F. van Gunsteren and et al., "Biomolecular modeling: Goals, problems, perspectives," *Angew. Chem. Int. Ed. Engl.*, vol. 45, no. 25, pp. 4064–4092, 2006.

[20] P. J. Ballester and G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *J. Comput. Chem.*, vol. 28, no. 10, pp. 1711–1723, 2007.

[21] E. Plaku, L. Kavraki, and M. Vardi, "Discrete search leading continuous exploration for kinodynamic motion planning," in *Robotics: Sci. and Syst.*, Atlanta, GA, USA, 2007.

[22] G. Sánchez and J.-C. Latombe, "On delaying collision checking in PRM planning: Application to multi-robot coordination," *Int. J. Robot. Res.*, vol. 21, no. 1, pp. 5–26, 2002.

[23] Y. Yang and O. Brock, "Efficient motion planning based on disassembly," in *Robotics: Sci. and Syst.*, Cambridge, MA, 2005, pp. 97–104.

[24] A. M. Ladd and L. E. Kavraki, "Motion planning in the presence of drift, underactuation and discrete system changes," in *Robotics: Sci. and Syst.*, Boston, MA, 2005, pp. 233–241.

[25] H. Choset and et al., *Principles of Robot Motion: Theory, Algorithms, and Implementations*, 1st ed. Cambridge, MA: MIT Press, 2005.

[26] F. Ding, D. Tsao, H. Nie, and N. V. Dokholyan, "Ab initio folding of proteins with all-atom discrete molecular dynamics," *Structure*, vol. 16, no. 7, pp. 1010–1018, 2008.

[27] M. Vendruscolo, E. Pacci, C. Dobson, and M. Karplus, "Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange," *J. Am. Chem. Soc.*, vol. 125, no. 51, pp. 15 686–15 687, 2003.

[28] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *J. Phys. Biol.*, vol. 2, no. 4, pp. 127–136, 2005.

[29] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 10, pp. 3352–3357, 2004.

[30] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 26, pp. 9885–9890, 2006.

[31] J. Barraquand and J.-C. Latombe, "Robot motion planning: A distributed representation approach," *Int. J. Robot. Res.*, vol. 10, pp. 628–649, 1991.

[32] L. Han, "Hybrid probabilistic RoadMap-Monte Carlo motion planning for closed chain systems with spherical joints," in *ICRA*, New Orleans, LA, 1994, pp. 920–926.

[33] M. Milik, A. Kolinski, and J. Skolnick, "Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates," *J. Comput. Chem.*, vol. 18, no. 1, pp. 80–85, 1997.

[34] G. Wang and R. L. Dunbrack, "Pisces: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[35] D. A. Case and et al., "Amber 9," University of California, San Francisco, 2006.

[36] A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 55, pp. 628–633, 1987.

[37] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, "A backbone-based theory of protein folding," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 45, pp. 16 623–16 633, 2006.

[38] T. H. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, "Geometry and symmetry presculpt the free-energy landscape of proteins," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 21, pp. 7960–7964, 2007.

[39] H. Kurniawati and D. Hsu, "Workspace-based connectivity oracle: An adaptive sampling strategy for PRM planning," in *WAFR*, ser. Springer Tracts in Advanced Robotics, New York, NY, 2006, vol. 47, pp. 35–51.

[40] L. Han and N. M. Amato, "A kinematics-based probabilistic roadmap method for closed chain systems," in *Algorithmic and Computational Robotics: New Directions*, B. R. Donald, K. M. Lynch, and D. Rus, Eds. MA: AK Peters, 2001, pp. 233–246.