Guiding the Search for Native-like Protein Conformations with an Ab-initio Tree-based Exploration

Amarda Shehu and Brian Olson

The International Journal of Robotics Research 2010 29: 1106 originally published online 6 May 2010
DOI: 10.1177/0278364910371527

Amarda Shehu
Department of Computer Science and
Department of Bioinformatics and Computational Biology
George Mason University,
Fairfax, VA 22030, USA
amarda@gmu.edu

Brian Olson
Department of Computer Science,
George Mason University,
Fairfax, VA 22030, USA

Guiding the Search for Native-like Protein Conformations with an Ab-initio Tree-based Exploration

Abstract

In this paper we propose a robotics-inspired method to enhance sampling of native-like conformations when employing only amino-acid sequence information for a protein at hand. Computing such conformations, essential to associating 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 paper 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 oversampling 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 eight 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.

KEY WORDS—native-like protein conformations; tree-based search; guided exploration; discretization layers; projection space; energy landscape; robotics-inspired; probabilistic sampling

The International Journal of Robotics Research
Vol. 29, No. 8, July 2010, pp. 1106–1127
DOI: 10.1177/0278364910371527

1. Introduction

Protein molecules are central to many biochemical processes in the cell. Modeling proteins and their biologically active state is crucial to our understanding and treatment of disease. It is now widely accepted that the sequence of amino acids in a protein chain determines the spatial arrangements in which the chain is biologically active (Anfinsen 1973). Wet-lab techniques have elucidated these arrangements, also referred to as native conformations, for only a small fraction of millions of available protein sequences (Lee et al. 2007). Obtaining native conformations *in silico* is essential to engineer novel proteins, predict protein stability, model protein interactions, and elucidate the molecular basis of disease (Kortemme and Baker 2004; Bradley et al. 2005; Yin et al. 2007).

Computing native conformations is NP-hard (Hart and Istrail 1997). The protein conformational space is vast and high-dimensional, as a protein chain may contain thousands of atoms. Even simplified (coarse-grained) chain representations may have hundreds of degrees of freedom (dofs). Figure 1 highlights different representations on a short chain. Given the high number of dofs, exhaustive search is impractical, although early work employed it on greatly simplified lattice representations (cf. the review in Clementi (2008)).

Analogies between protein chains and articulated mechanisms, together with the high-dimensionality of the protein conformational space, have long attracted robotics researchers to adapt and apply algorithms that plan motions for articulated mechanisms with many dofs to the study of protein conformations (Apaydin et al. 2001; Amato et al. 2002; Kim et al. 2002; Apaydin et al. 2003; Song and Amato 2004; Cortes et al. 2005; Lee et al. 2005; Chiang et al. 2007; Georgiev and Donald 2007; Kirillova et al. 2008). Although these methods often have to be adapted to deal with hundreds of dofs in protein

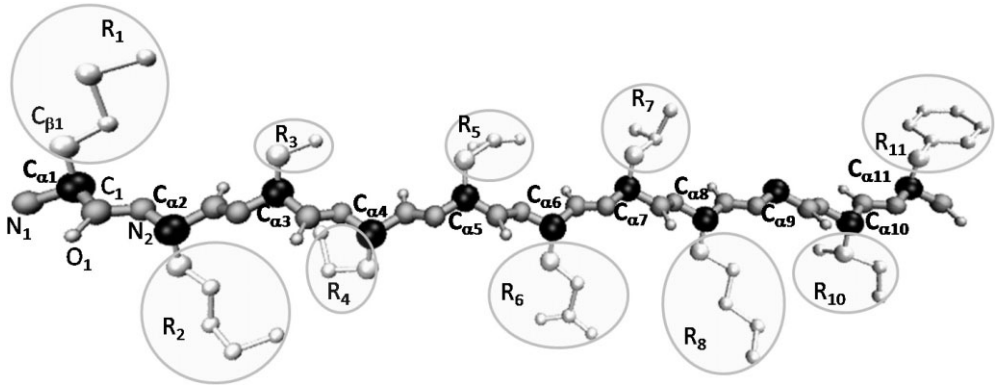

Fig. 1. A protein chain of 11 amino acids is shown in an extended conformation. Atoms are drawn as spheres and bonds as cylinders. Amino acids share a common group of backbone atoms, labeled N, C_α, C, and O. $C_{\alpha,i}$ atoms are in black, N_i and C_i in gray, and O_i in silver (i is amino-acid position). Backbone atoms and bonds that connect them make up the backbone chain. Backbone representations model only this chain. Coarser representations such as C_α traces model only C_α atoms. Atoms in white are labeled R for residue. There are 20 distinct residues or side chains in natural proteins, hence 20 amino-acid types. Glycine at $C_{\alpha,9}$ has no side chain. Side chains connect to C_α through the C_β atom. Extended backbone representations include C_β atoms. All-atom representations model all atoms. Rotational dofs can be associated with N–C_α, C_α–C, and side-chain bonds to generate conformations analogously to configurations for articulated mechanisms with revolute joints.

chains (from dozens of dofs in articulated mechanisms), the motion-planning framework has allowed the problem of computing paths from a given initial to a given goal conformation to be addressed. This problem, known as protein folding, concerns itself with understanding how a protein chain traverses the energy surface that underlies the conformational space to fold from an initial to a goal conformation. The goal is often a known native conformation.

The problem addressed in this paper is the discovery of native conformations from knowledge of amino-acid sequence. Rather than addressing how a protein folds, this problem concerns itself with understanding into what the protein folds. Computing native conformations from a protein's amino-acid sequence is known as the ab-initio structure prediction problem.

This paper makes use of the energy landscape view that places native conformations at the global minimum of the protein energy surface (Dill and Chan 1997). Locating this minimum is not trivial, as the energy surface is rich in local minima. The true surface is probed with empirical energy functions, as quantum mechanics calculations can only be afforded on chains of a few amino acids. Decades of research have resulted in sufficiently accurate all-atom and coarse-grained energy functions (operating on all-atom and coarse-grained representations) (Clementi 2008). These functions may introduce distortions and false local minima, but they do not hamper conformational search as long as the distortions are slight and the reported global minimum corresponds to the known native state. Since atomic interactions scale quadratically with the number of atoms modeled, coarse-grained energy functions are appealing due to their lower computational cost.

This paper employs coarse-grained energy functions and a novel conformational search method to enhance the sampling of native-like conformations. The term native-like applies to conformations that are sufficiently close to the true global minimum so that further local search with an expensive all-atom energy function will steer these conformations to the native state (Bradley et al. 2005). This process is often referred to as energetic refinement. It is not possible to determine that a conformation is native-like in the absence of a known native conformation. Since all that is known of native-like conformations is that they are low energy, a viable strategy is to populate many distinct local minima in order to increase the probability that some of them are sufficiently close and will actually reach the global minimum upon further refinement.

This paper proposes a novel two-layered ab-initio conformational search method to enhance the sampling of geometrically distinct low-energy coarse-grained conformations. The goal of the method is to serve as a filtering step and reveal potentially native-like low-energy conformations that can be refined through detailed studies to obtain the native state. From now on the method is referred to as `FeLTr` for `Fragment` Monte Carlo `Tree` Exploration.

`FeLTr` is tested on eight protein sequences of different lengths (20–76 amino acids), amounting to 40–152 dofs, and native topologies (β, α/β, α). Results show that `FeLTr` captures and populates the native state on all of these proteins. Short proof-of-concept refinement of selected distinct lowest-energy conformations shows they are good starting points for

detailed biophysical studies focused on extracting fine structural and functional properties of novel sequences (Ding et al. 2008; Shehu et al. 2009).

A summary of the main ingredients of FeLTr is given in the following in order to place FeLTr in context of other conformational search methods in Section 1.2. A detailed description of FeLTr follows in Section 2. Analysis of the conformations computed on the protein sequences chosen for testing is presented in Section 3. This section also discusses limitations of the current sequential implementation of FeLTr on longer sequences and highlights directions for future research. The paper concludes in Section 4 with a discussion.

1.1. Main Ingredients of the Conformational Search in FeLTr

FeLTr gathers information about explored regions of the conformational space and the underlying energy landscape to further guide the exploration towards energetically relevant under-explored regions. Gathering information during the search to further advance the search towards promising regions of the solution space is an important topic in artificial intelligence (AI) that is central to searching high-dimensional solution spaces (Russell and Norvig 2002).

Inspired by tree-based methods in robot 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 native-like conformations have the lowest energies. The second goal attempts to enhance sampling 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 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 projection space using the recently proposed ultrafast shape recognition (USR) features (Ballester and Richards 2007). 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 Monte Carlo (MC) trajectory expands the tree with a new low-energy conformation. Figure 2 illustrates this two-layered exploration.

The employment of the projection space in FeLTr is inspired by recent sampling-based motion planners that use decompositions, subdivisions, and projections of the robot configurational space or workspace to balance exploration between coverage and progress toward the goal (Sánchez and Latombe 2002; Choset et al. 2005; Ladd and Kavraki 2005; van den Berg and Overmars 2005; Yang and Brock 2005; Kurniawati and Hsu 2006; Rodriguez et al. 2006; Plaku et al. 2007; Stilman and Kuffner 2008). It is worth mentioning that, while FeLTr focuses on proteins, the employed projections rely only on geometry and can be used for any articulated mechanism (manipulators, humanoid, modular robots); see Section 4 for more on this point.

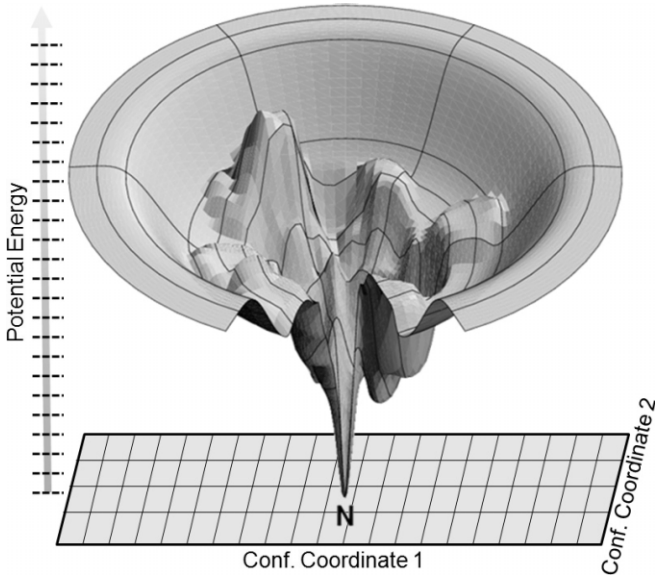

Fig. 2. The native state, labeled N, is associated with the global minimum of the protein energy surface (Dill and Chan 1997). The surface shown here is adapted from Dill and Chan (1997). FeLTr cross-sects this surface by discretizing potential energy values in a one-dimensional grid, illustrated here with the z-axis. The axis changes color from light to dark gray to show the decrease in energy values towards the native state. The grid on the xy-plane discretizes the projection of the conformational space onto a few conformational coordinates. Two coordinates are shown here for visualization purposes. FeLTr uses three coordinates, as detailed in Section 2.

1.2. FeLTr in Context of Conformational Search Methods

Where should conformational search devote exploration time? It is important to focus computational resources on regions that allow the search to progress to the global minimum. Such regions are not known *a priori*, since stochastic search of a high-dimensional space affords only a local view. A successful conformational search needs to strike the right balance between populating a large number of distinct low-energy regions and focusing further resources to low-energy regions likely to lead to the global minimum.

Ingredients of a successful approach were identified most notably by Lee et al. (1997, 1998) and Lee and Scheraga

(1999). The conformation space annealing method of Lee et al. (1997) 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. The success of the second stage to locate the global minimum depends on the regions populated by the first stage.

The emerging template in state-of-the-art conformational search methods is to sample a large number of low-energy conformations in the first stage, essentially aiming to build a broad map of the energy landscape (Bonneau and Baker 2002; Bradley et al. 2005; Shehu et al. 2008; Brunette and Brock 2009; DeBartolo et al. 2009; Shehu et al. 2009). Clustering analysis is then conducted over the conformations to reveal distinct minima that constitute good starting points from which expensive (often all-atom detail) local searches in the second stage can reach the global minimum. In contrast, coarse-grained representations are 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 (Bradley et al. 2005; Shehu et al. 2008; Brunette and Brock 2009; DeBartolo et al. 2009; Shehu et al. 2009). In addition, 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 exploration and the analysis over the conformations are independent of each other. As a result, computed conformations cannot be ensured to be geometrically distinct and not representative of only a few regions in conformational space. 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 such as least root-mean-squared deviation (lRMSD) and radius of gyration (Rg) are often confined to the post stage 1 analysis because they can mask away important conformational differences.

Specifically, the work of Shehu et al. (2009) has shown that important local 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 computational biophysics is devoted to finding effective reaction coordinates for proteins (cf. Clementi (2008)). The proposed `FeLTr` circumvents this difficulty and incorporates geometric diversity in its exploration by employing general geometric-based coordinates not specific to proteins.

`FeLTr` is an alternative approach to the first stage in ab-initio structure prediction. Its goal is to rapidly reveal distinct minima worth exploring in greater detail. The novelty in `FeLTr` is the employment and interplay of two layers, energy and geometry, to guide the exploration to distinct local minima. While `FeLTr` integrates geometric diversity in its tree-based search, the method of (Brunette and Brock 2009) regularly employs expensive all-atom refinement of low-energy regions (deemed "funnels") sampled in the first stage as a way to gauge which regions are more likely to lead to the global minimum upon further refinement. `FeLTr` does not switch between coarse-grained and all-atom representations, a costly technique employed to guide a simulated annealing MC search in our earlier work (Shehu et al. 2009). Instead, `FeLTr` incorporates a geometric projection layer to guide the exploration to under-populated low-energy regions of conformational space.

The geometric projection layer in `FeLTr` directly determines which branches of the search tree will be further extended in conformational space. The USR coordinates of Ballester and Richards (2007) are employed as general geometric coordinates (not specific to proteins) that allow an initial implementation of the idea to integrate geometric diversity in the exploration itself. In order to guide the exploration towards low-energy regions, `FeLTr` employs a second discretization layer over energy values. The weight function that selects low-energy levels for expansion (detailed in Section 2) can be tuned to balance between progress toward the global minimum and coverage of conformational space (afforded by the geometric projection layer). The specific weight function in this paper drives aggressively towards lower energies, which allows capturing the native state for small proteins (with no alternative functional states) in a few CPU hours. The relationship between energy and geometry is highlighted in detail in the results in Section 3.

`FeLTr` can indeed cross energetic barriers due to its employment of a short MC trajectory to expand the tree with a new low-energy conformation. MC has been used in sampling-based motion planing to escape local minima (Barraquand and Latombe 1991) and enhance sampling for closed chains with spherical joints (Han 1994). It has recently been proposed to enhance RRT and extend towards higher-cost regions of the configuration space (Jaillet et al. 2008). While there is no notion of geometric diversity in Jaillet et al. (2008), the combination of RRT, an established tree-based search, with MC, an established local optimization method, is analogous to the combination of a tree-based search with an MC-based expansion in `FeLTr`. The MC component allows `FeLTr` to extend to higher-energy regions and possibly escape local minima while biasing the expansion of the tree from nodes that lie deep in the energy surface.

An important component of `FeLTr` is the employment of fragment-based assembly, where a conformation is assembled with physical fragment configurations extracted from a non-redundant database of structures (Bonneau and Baker 2002; Bradley et al. 2005; Gong et al. 2005; Brunette and Brock 2009; DeBartolo et al. 2009; Shehu et al. 2009). Deciding on a suitable fragment length depends on the richness of the database to provide a comprehensive picture of fragment configurations (Kolodny et al. 2002). The current diversity of the Protein Data Bank (PDB) supports a minimum length of

Algorithm 1 FeLTr: Fragment Monte CarLo Tree Exploration.

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

1: $C_{\text{init}} \leftarrow$ extended coarse-grained conf from α ▷2.1
2: $\mathrm{G}_E \leftarrow$ explicit 1d energy grid ▷2.2.1
3: **for** $\ell \in \mathrm{G}_E$ **do**
4: $\quad \ell.\mathrm{G}_{\text{USR}} \leftarrow$ implicit three-dimensional geom projection grid ▷2.2.2
5: AddConf($C_{\text{init}}, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
6: **while** Time AND $|\Omega_\alpha|$ do not exceed limits **do**
7: $\quad \ell \leftarrow$ SelectEnergyLevel(G_E) ▷2.2.1
8: $\quad$ cell $\leftarrow$ SelectGeomCell($\ell.\mathrm{G}_{\text{USR}}$.cells) ▷2.2.2
9: $\quad C \leftarrow$ SelectConf(cell.confs) ▷2.2.2
10: $\quad C_{\text{new}} \leftarrow$ MC_ExpandConf(C) ▷2.3
11: $\quad$ **if** $C_{\text{new}} \neq$ NIL **then** ▷*MC succeeded*
12: $\quad\quad$ AddConf($C_{\text{new}}, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
13: $\quad\quad \Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$ ▷*add conf to ensemble*

AddConf($C, \mathrm{G}_E, \mathrm{G}_{\text{USR}}$)
▷*add C to appropriate energy level in* G_E
1: $E(C) \leftarrow$ CoarseGrainedEnergy(C) ▷2.3.2
2: $\ell \leftarrow$ level in G_E where $E(C)$ falls into
3: ℓ.confs $\leftarrow \ell$.confs $\cup \{C\}$
▷*add C to appropriate cell in geom grid associated with ℓ*
4: $P(C) \leftarrow$ USRGeomProj(C) ▷2.2.2
5: cell $\leftarrow$ cell in $\ell.\mathrm{G}_{\text{USR}}$ where $P(C)$ falls into
6: **if** cell = NIL **then** ▷*cell had not yet been created*
7: $\quad$ cell $\leftarrow$ new geom projection cell
8: $\quad \ell.\mathrm{G}_{\text{USR}}$.cells $\leftarrow \ell.\mathrm{G}_{\text{USR}}$.cells $\cup$ {cell}
9: cell.confs $\leftarrow$ cell.confs $\cup \{C\}$

three amino acids (Shehu et al. 2009). FeLTr's employment of physical (versus random) fragment configurations improves the likelihood of assembling a physically realistic conformation. The assembly is implemented in a Metropolis MC framework, where new fragment configurations are proposed to replace those in a current conformation. Replacements that meet the Metropolis criterion are accepted, resulting in a new conformation.

2. FeLTr

In the usual MC framework, the probabilistic walk in conformational space resumes from the last conformation computed. FeLTr enhances this framework by conducting a tree-based exploration. 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 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 Algorithm 1. Sections describing the main steps in FeLTr are referenced at the end of each line in the pseudocode.

2.1. Coarse-grained Representation of a Protein Chain

The extended backbone representation is employed, modeling backbone N, C_α, C, O atoms and side-chain C_β atoms. 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 (Algorithm 1:1). Positions of C_β atoms are determined from the backbone as in Milik et al. (1997).

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

2.2.1. Energy Layers

A one-dimensional grid, G_E (Algorithm 1:2), is defined on the segment $[E_{\min}, E_{\max}]$. Here $E_{\min}$ refers to the lowest expected energy on computed conformations, and $E_{\max}$ refers

to the highest energy. Values to these parameters may depend on the coarse-grained energy function employed. The experiments in Section 3 test FeLTr when employing two different coarse-grained energy functions. The main results detailed in Section 3 are obtained with E_{AMW}, a coarse-grained energy function recently proposed by Shehu et al. (2009) and detailed in Section 2.3.2. Section 3.8 compares results obtained with E_{AMW} to those obtained with our adaptation of Rosetta, another well-established energy function in fragment-based assembly literature (Bonneau and Baker 2002). We refer to our adaptation as $E_{\text{Rosetta*}}$ (detailed in Section 2.3.2). When employing E_{AMW}, low-negative-energy values are associated with feasible conformations, whereas $E_{\text{Rosetta*}}$ may associate low-positive-energy values with feasible conformations. To employ the same $[E_{\min}, E_{\max}]$ segment, the range of energy values obtained with $E_{\text{Rosetta*}}$ is properly shifted.

Since the Metropolis MC expansion quickly obtains low-energy conformations, it is not necessary to maintain a grid over energies higher than $E_{\max}$. Energy levels are generated every δE units. Here δE is set to a small value so that the average energy $E_{\text{avg}}(\ell)$ over conformations populating a specific energy level $\ell \in \text{G}_E$ captures well the distribution of energies in ℓ. This discretization is used to bias the selection towards conformations in the lower energy levels through a weight function.

The weight function $w(\ell)$ associated with an energy level $\ell \in \text{G}_E$ is set to $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell) + \epsilon$, where ϵ ensures that conformations with higher energies have a non-zero probability of selection. An energy level ℓ is then selected with probability $w(\ell)/\sum_{\ell' \in \text{G}_E} w(\ell')$ (Algorithm 1:7). This quadratic weight function biases selection towards conformations with lower energies while allowing for some variation. Allowing higher-energy conformations to be selected for expansion provides FeLTr with the ability to jump over barriers in the energy surface. It is worth pointing out that the Metropolis MC expansion additionally allows jumping over energetic barriers.

2.2.2. Low-dimensional Geometric Projections

Conformations in a chosen energy level are projected onto a low-dimensional space. Borrowing from the USR features of Ballester and Richards (2007), 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. In this way, the distribution of atomic distances from each extreme point (approximated with μ^1) is likely to yield new geometric information on a conformation.

The three projection coordinates capture overall topologic differences among conformations. As discussed in Section 4, the coordinates are not specific to proteins but can be applied to any articulated mechanism. The coordinates allow introducing a second layer of discretization. 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 into cells of the projection space to employ coverage in this space as a second criterion for selection. An implicit three-dimensional grid, G_{USR}, is associated with each energy level (Algorithm 1: 3–4), based on a uniform discretization of the projection coordinates. The selection is biased towards cells with fewer conformations through the weight function $1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$, where nsel records how often a cell is selected, and nconfs is the number of conformations that project to the cell (Algorithm 1: 8). Similar selection schemes have been advocated in motion-planning literature as a way to increase geometric coverage during exploration (Sánchez and Latombe 2002; Ladd and Kavraki 2005; Burns and Brock 2007; Plaku et al. 2007). Once a cell is chosen, the actual conformation selected for expansion is obtained at random over those in the cell (Algorithm 1: 9), since conformations in a cell have similar energies (within δE).

2.3. Expansion: Metropolis MC with Fragment-based 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 six ϕ, ψ angles defined over its backbone. The expansion procedure (Algorithm 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 2.3.1, 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 (Algorithm 1: 11) is done under the Metropolis criterion.

2.3.1. Database of Fragment Configurations

A PDB subset of non-redundant protein structures (as of November 2008) is extracted through the PISCES server (Wang

and Dunbrack 2003) to contain proteins that have $\leq 40\%$ sequence similarity, ≤ 2.5 Å resolution and R-factor ≤ 0.2. Proteins studied in this paper are removed from the database. The 40% similarity cutoff ensures that topologies that are overpopulated 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 10 million configurations. No less than 10 configurations are populated for any trimer of each tested sequence.

Kolodny et al. (2002) highlight the relationship between fragment length and quality of the fragment configuration database that is needed to assemble native structures. The longer the fragment, the larger the database needs to be in order to provide the configurational diversity needed to capture novel protein structures. The shorter the fragment, the higher the probability that a fragment configuration database constructed from non-redundant protein structures will recover novel native structures not used in the construction (Kolodny et al. 2002). That is why `FeLTr` employs a database of trimer configurations.

The quality of the database can be quantified. The local-fit score introduced by Kolodny et al. (2002) allows us to measure the degree to which the database fits a testing set of known native protein structures not used to construct the database. The chain of each protein in the testing set is broken into all of its overlapping trimers. For each such trimer, the set of configurations available for it in the database is scanned to find the configuration closest to the configuration in the native structure. Configuration similarity is measured in terms of lRMSD. A local-fit score is then associated with each protein as the average over the lRMSD values obtained for all trimers of a given protein chain. It is worth noting that the implementation of the local-fit score in this paper limits the search for similar configurations of a trimer to configurations of that same trimer sequence in the database (the search in Kolodny et al. (2002) is over the entire database). This modification better reflects the usage of the fragment configuration database in the fragment-based assembly process.

The eight proteins employed in this paper to measure the performance of `FeLTr` are used as the testing set by which to illustrate the quality of the trimer configuration database. Results in Section 3.2 show that the trimer database is diverse enough to allow `FeLTr` obtain native-like conformations.

2.3.2. Coarse-grained Energy Function

The main coarse-grained energy function employed by `FeLTr` has recently been proposed by Shehu et al. (2009). The function is a modification of the `A`ssociative `M`emory Hamiltonian with `W`ater (AMW) employed by Papoian et al. (2004) for ab-initio structure prediction. The function, referred to as E_{AMW}, 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 (Case et al. 2006), with a modification that allows a soft penetration of van der Waals spheres. The $E_{\text{H-Bond}}$ term allows formation of local and non-local hydrogen bonds. The terms E_{contact}, E_{burial}, and E_{water}, implemented as in Papoian et al. (2004), allow formation of non-local contacts, a hydrophobic core, and water-mediated interactions.

The E_{Rg} term in this paper penalizes a conformation by $(\text{Rg} - \text{Rg}_{\text{PDB}})^2$ if the conformation's Rg value is above the Rg_{PDB} value predicted for a chain of the same length from proteins in the PDB. The predicted value fits well to the line $2.83 \cdot N^{0.34}$ (Gong et al. 2005), which is used to compute Rg_{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, Rg_{PDB} and the Rg value of an extended conformation are used to define the boundaries of the projection space.

A second coarse-grained energy function introduced in David Baker's lab and implemented as part of the Rosetta structure prediction package (Bradley et al. 2005) has been configured into `FeLTr` to additionally test the ability of the method to compute native-like conformations with different state-of-the-art coarse-grained energy functions. The Rosetta coarse-grained energy function is also a linear combination of local and non-local terms, some of which are statistically derived from analysis of native structures of proteins in the PDB (cf. Bonneau and Baker (2002)). We add the E_{Rg} term described above and refer to the modification as E_{Rosetta^*}.

2.3.3. 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 (Algorithm 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 paper, 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.

2.4. Analysis of Computed Conformations

As shown in Algorithm 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

Table 1. Fold, Size, and Number of Degrees of Freedom (dofs) are Shown for Each of the Eight Proteins

Protein	Trp-cage	wwD	hp36	hbd2	eHD	L20	GB1	Calbindin D_{9k}
Fold	α	β	α	α/β	α	α	α/β	α
Size	20	26	36	41	54	60	60	76
dofs	40	52	72	82	108	120	120	152

Table 2. $\langle$lRMSD$\rangle_f$ Refers to the Average Least Root Mean-squared Deviation (lRMSD) Obtained Over All Trimers f Defined Over Each Protein Chain; Values are Reported for All Eight Proteins

Protein	Trp-cage	wwD	hp36	hbd2	eHD	L20	GB1	Calbindin D_{9k}
$\langle$lRMSD$\rangle_f$(A)	0.04	0.09	0.07	0.02	0.02	0.06	0.06	0.03

$\langle E\rangle - \sigma E$. Here Ω_α^* is clustered with a simple leader-like algorithm (Jain et al. 1987), 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, whether E_{AMW} or E_{Rosetta^*} are employed as energy functions.

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 two employed in this paper) 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. Such inherent uncertainties in the energy functions warrant the focus on the distinct clusters of conformations with energies no higher than $\langle E\rangle - \sigma E$ rather than on the lowest-energy conformation.

3. Experiments and Results

FeLTr is applied to eight structurally diverse protein sequences of varying lengths listed in Section 3.1. Section 3.2 details the quality of the trimer configuration database as measured through the local-fit score described in Section 2. Implementation details are related in Section 3.3. The rationale for the conducted analysis is laid out in Section 3.4. Detailed analysis of the obtained results follows in Sections 3.5–3.10.

3.1. Chosen Systems

The eight proteins chosen to test FeLTr, listed in Table 1, include tryptophan cage (Trp-cage), Pin1 Trp-Trp ww domain (wwD), villin headpiece (hp36), human β-defensin 2 (hbd2), engrailed homeodomain (eHD), bacterial ribosomal protein (L20), immunoglobulin binding domain of streptococcal protein G (GB1), and calbindin D_{9k}. These proteins are chosen because they vary in size (number of amino acids) and so number of dofs, native fold (three-dimensional global arrangement of local secondary structure segments), and are actively studied *in silico* and wet lab due to the importance of their biological functions.

3.2. Analyzing the Quality of the Trimer Configuration Database

The local-fit score is measured on each of the eight proteins as described in Section 2. Table 2 reports $\langle$lRMSD$\rangle_f$, the average lRMSD obtained over all trimers f of each protein. The results in Table 2 show that the trimer configuration database is diverse enough to obtain on average a low lRMSD configuration for the trimers on each of the eight protein systems.

It is interesting to see what conformations could be constructed by assembling the lowest-lRMSD configurations for each trimer in the protein chain. These conformations are shown in dark gray in Figure 3 superimposed over the native structure of the respective protein shown in transparent light gray. Figure 3 also lists for each protein the lRMSD between the constructed conformation and the native structure. Comparing the constructed conformation to the native structure illustrates that the closest configuration is not always the best choice for a fragment when aiming to assemble a native-like conformation (especially obvious on hbd2, L20, and calbindin D_{9k}). A trivial reason for the high lRMSDs is that lRMSD is known to increase with protein size and is an effective measure for similar structures but less descriptive with increasing dissimilarity. The main reason, however, which is elucidated by the superimposition of the conformations, is that suboptimal fragment configurations are often needed to construct a final optimal conformation. This observation emphasizes the need for non-trivial search methods and realistic energy functions to assemble native-like conformations.

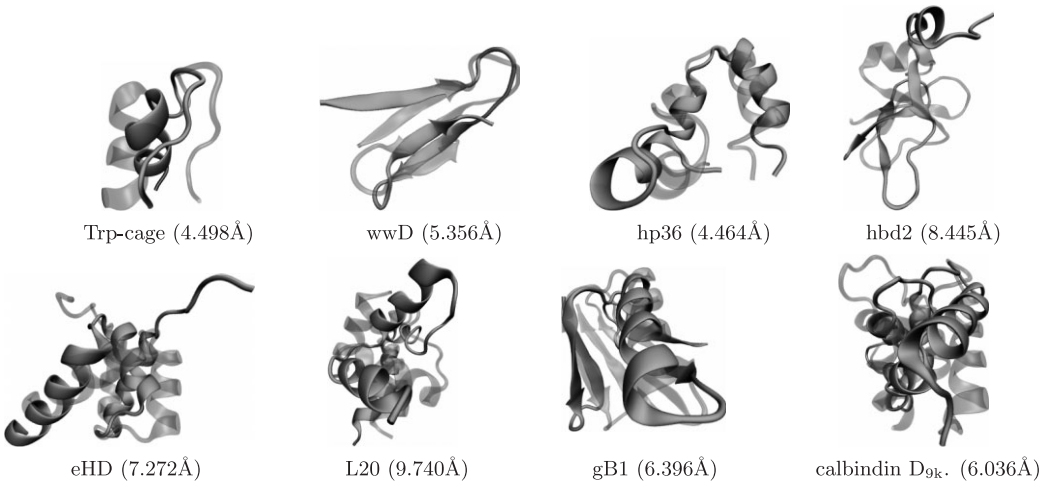

Fig. 3. Conformations constructed for each protein during the calculation of the local-fit score are drawn in dark gray and superimposed over the native structure drawn in transparent light gray. Least root-mean-squared deviations (lRMSDs) between constructed conformations and native structures are in parentheses. The shown native structures are obtained from the Protein Data Bank (PDB) with PDB id 1l2y for Trp-cage, 1i6c for wwD, 1vii for hp36, 1fd4 for hbd2, 1enh for eHD, 1gyz for L20, 1gb1 for GB1, and 4icb for calbindin.

3.3. Implementation Details

`FeLTr` is implemented in C++, run on an Intel Core2 Duo machine with 4 GB RAM and 2.66 GHz CPU. The segment $[E_{\min}, E_{\max}]$ for G_E is set to $[-200, 0]$ kcal/mol. The ϵ parameter in $w(l)$ is set to 2^{-22}. The δE parameter that separates energy levels is 2 kcal/mol.

In the current sequential implementation, `FeLTr` keeps the entire exploration tree in memory. To ensure that the ensemble Ω_α in this tree fits in memory, tests on the sequences considered here are terminated before reaching the memory limitation. On small proteins such as Trp-cage, a larger conformational ensemble can fit in memory than on longer proteins such as calbindin. On the other hand, the limit is reached more quickly on the small proteins, as it is faster to add a conformation to the growing conformational ensemble. As time grows quadratically with chain length due to the Lennard–Jones energy term, it takes longer to evaluate the energy of a generated conformation for a longer sequence. Therefore, the memory limitation is combined with a reasonable bound on the running time. On all proteins tested here, no more than 3 hours are needed to either reach the memory limitation or sufficiently populate the native state. On Trp-cage, wwD, and hp36, 40 minutes, 1 hour, and 2 hours, respectively, are sufficient to generate over 50,000 conformations and reach the memory limit. On longer chains like calbindin, ∼10,000 conformations are generated in 3 hours.

3.4. Rationale and Summary of the Analysis of FeLTr

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 (Ding et al. 2008). Comparing computed conformations with experimentally available native structures of each protein reveals that `FeLTr` captures the native state, regardless of whether E_{AMW} or $E_{\text{Rosetta*}}$ are employed. 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 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 three-dimensional arrangement of these segments.

Comparing `FeLTr` with MC allows directly probing the effect of the novel tree-based exploration guided by the two-layer discretization in sampling native-like conformations. In addition, the role and contribution of each of the discretization layers in `FeLTr` is investigated in detail by expanding the comparison between `FeLTr` and MC with versions of `FeLTr` where either the energy or the geometric projection layers are turned off. Comparison of conformations generated under each setting shows that the combination of both disretization layers allows `FeLTr` to better populate diverse minima.

Two systems, Trp-cage and eHD, are selected to compare the conformations obtained under these four experimental settings. The rest of the results in Sections 3.5–3.7 compare MC with `FeLTr` when employing the E_{AMW} function. Sec-

tion 3.8 compares the results obtained with E_{AMW} with those obtained when employing E_{Rosetta^*}. The comparison shows no significant differences in the ability of FeLTr to capture the native state when employing either E_{AMW} or E_{Rosetta^*}. Section 3.9 subjects conformations representative of the highest-populated low-energy minima obtained for each protein sequence to a short all-atom refinement. The analysis, which compares the conformations before and after refinement to the known native structure, illustrates the practical usage of FeLTr to obtain a few native-like coarse-grained conformations worth refining in further detail. Finally, Section 3.10 applies FeLTr on a longer protein sequence that showcases both current limitations of the method and interesting directions for future research.

3.5. Analyzing the Efficiency of FeLTr

The efficiency of FeLTr is estimated in comparison with a Metropolis MC simulation and with versions of FeLTr where we turn off either the energy layer or the geometric projection layer. Turning off one of the layers in FeLTr affects the selection of conformations for expansion and so allows us to probe the contribution of each layer in the ability of FeLTr to populate diverse energy minima. Comparison with an MC simulation essentially allows us to probe the effect of turning off both layers. The limit on execution time in each experimental setting is kept the same. To keep all other conditions similar, the MC simulation employs the same fragment-based assembly to compute conformations in its trajectory and the same coarse-grained representation and energy function to calculate the energy of a computed conformation.

These four settings are compared with one another on two selected proteins, Trp-cage in Figure 4(a1)–(f1) and eHD in Figure 4(a5)–(f5) (eHD is the fifth system selected for this study). Figure 4(a)–(f) plots energies versus lRMSDs of conformations from a known native structure. Data obtained from the MC simulation are shown in lightest gray in Figure 4(a1), (a5), followed by data obtained from FeLTr with only the energy layer in dark gray in Figure 4(b1), (b5). Data obtained from FeLTr with only the geometric projection layer are shown in darker gray in Figure 4(c1), (c5), followed by data obtained from the full FeLTr (with both layers) in darkest gray in Figure 4(d1), (d5). Figure 4(e1), (e5) plot the minimum energy over the growing exploration tree (the MC trajectory can be regarded as a degenerate tree). Figure 4(f1), (f5) plot the minimum lRMSD from the native over the growing tree.

For both Trp-cage and eHD, the MC simulation does not achieve as low energies as FeLTr or FeLTr with only the energy layer; see Figure 4(a), (b), (d), and (e). Figure 4(a1)–(f1) and (a5)–(f5) show that MC can come close in lRMSD to the native structure. This is also confirmed on most proteins in Figure 6(c2)–(c8), where MC data are compared with FeLTr data (FeLTr comes closer in Figures 4(d1) and 6(b2), (b7)). This is expected, as the MC simulation employed for comparison is a powerful probabilistic walk that uses fragment-based assembly to make large hops in conformational space; that is, the lRMSD between two consecutive conformations can be significant.

Figure 4(b1), (b5) show that FeLTr with only the energy layer yields lower energies than the MC simulation on both Trp-cage and eHD. The energy layer in FeLTr biases the selection of conformations for expansion according to a quadratic weighting scheme. This layer greedily seeks already-populated lower-energy levels for expansion. Such a strategy, on its own, can steer the exploration towards a local minimum far away from the native state. For instance, the data obtained for Trp-cage in Figure 4(b1) show this version of FeLTr trapped in a local energy minimum as far as 6 Å away from the native structure. eHD data in Figure 4(b5) show this strategy heavily populates a local minimum about 1 Å away from the lowest-lRMSD minimum obtained with FeLTr in Figure 4(d5). While the energy layer drives the Trp-cage exploration towards lower energies than those obtained with FeLTr (see Figure 4(a1), (d1), (f1)), FeLTr obtains lower energies on eHD (see Figure 4(a5)–(e5)). The combination of the energy and geometric projection layers allows FeLTr to expand deeper in the more complex energy landscape of eHD.

Figure 4(c1), (c5) show that FeLTr with only the geometric projection layer yields higher energies than the other settings, both on Trp-cage and eHD. This can be seen in Figure 4(e1), (e5). The geometric projection layer fails to drive the exploration towards low-energy conformations or to sufficiently populate local minima near the native state in the given exploration time. While Figure 4(c1), (f1) show that the exploration for Trp-cage can get close in lRMSD to the native structure (Trp-cage is a small protein), Figure 4(c5), (f5) for eHD shows that seeking coverage may waste precious exploration time on vaster conformational spaces with complex energy landscapes.

Figure 4(d1), (d5) shows that FeLTr, where both layers are turned on, populates more minima (and may even obtain lower energies than FeLTr with only the energy layer, as pointed out on eHD). On both Trp-cage and eHD, at least one of the minima is very close in lRMSD to the corresponding native structure. This result is also evident in Figure 4(f1), (f5), which tracks the lowest lRMSD from the native structure over all conformations in the growing exploration tree. The combination of both layers make FeLTr superior to the other three experimental settings. For this reason, the analysis on the rest of the protein systems used for testing focuses on comparing FeLTr with an MC simulation. This analysis shows that significantly lower energies are obtained with FeLTr than the MC simulation (also revealed when comparing MC to FeLTr for the other proteins in Figure 6). This is not surprising, as FeLTr guides the tree towards lower energies, whereas the MC simulation resumes from the last conformation generated.

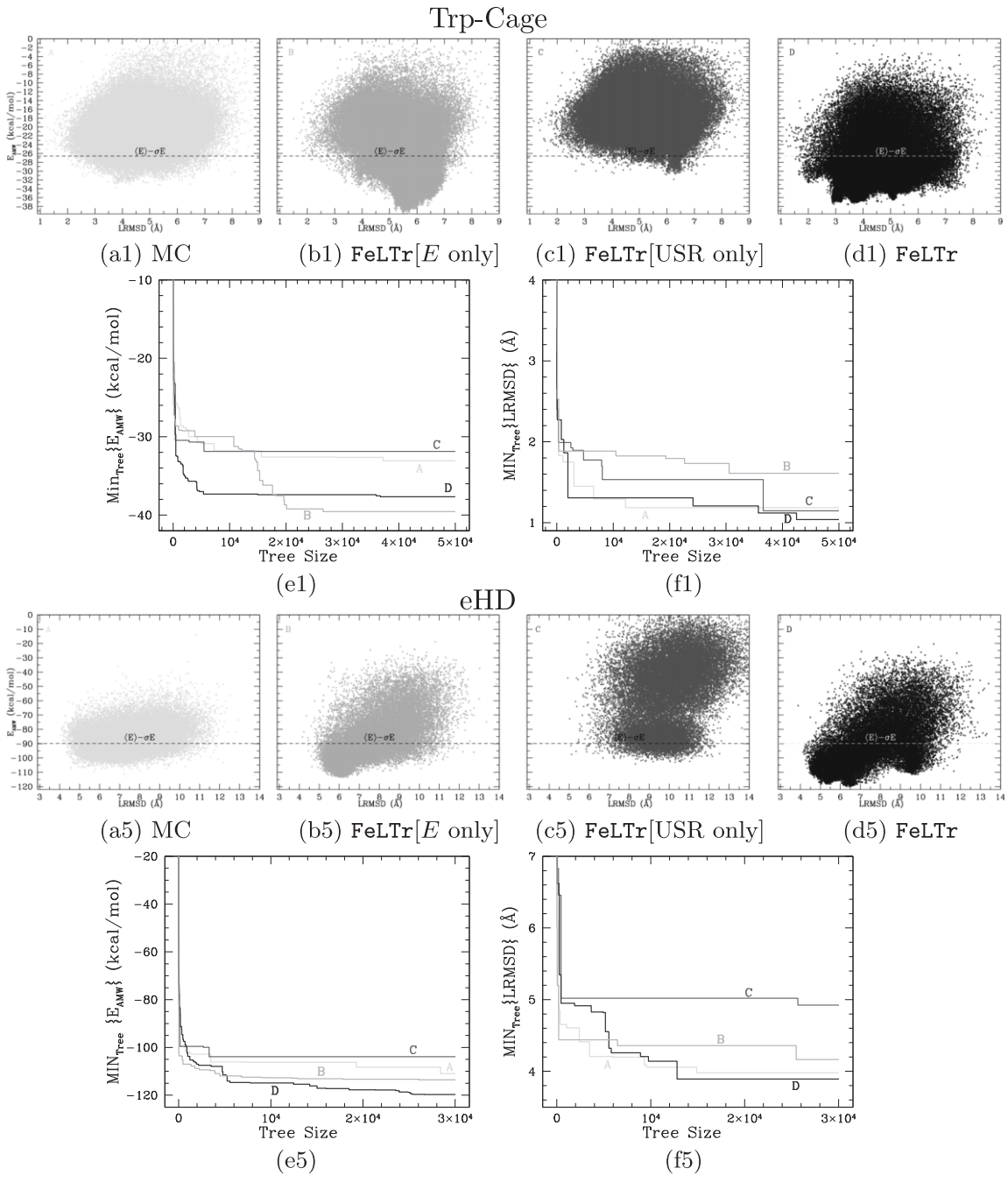

Fig. 4. (a1)–(f1) show data obtained on Trp-cage, and (a5)–(f5) show data obtained on eHD. (a)–(d) Energies of conformations are plotted versus lRMSDs from native; MC in (a1), (a5), FeLTr with only the energy layer turned on in (b1), (b5), FeLTr with only the geometric projection layer turned on in (c1), (c5), and FeLTr (with both layers) in (d1), (d5). The horizontal line marks the energetic cutoff with respect to energies obtained with FeLTr. (e1), (e5) track the minimum lRMSD from native over the growing exploration tree as conformations are computed and added to the tree (the tree in the MC simulation is a trajectory). (f1), (f5) track the minimum energy over the growing tree. Curves are labeled, A for the MC simulation, B for FeLTr with only the energy layer turned on, C for FeLTr with only the geometric projection layer turned on, and D for the full FeLTr, where both layers are turned on.

The geometric projection layer prevents FeLTr from over-populating minima. This is important, as Figure 4(b1) shows the presence of non-native energy minima far away from the native state. FeLTr populates more minima, revealed when plotting energies of computed conformations versus their lRMSDs from the native. The geometric projection layer helps explore different conformational topologies with which to populate low-energy levels. It is worth noting that, although lRMSD is not a general conformational coordinate, it is useful to visualize two-dimensional projections of the probed energy surface.

It is worth emphasizing the inability of a single MC trajectory to populate diverse energy minima. Recent work by Shehu et al. (2009) attempts to address this inability by executing numerous trajectories, each one initiated from different carefully selected conformations in a simulated annealing framework (Shehu et al. 2009). This strategy comes at a computational cost, requiring 1–3 weeks of computation on 50 CPUs (Shehu et al. 2009). In contrast, the selection of conformations from which to initiate MC trajectories is seamlessly integrated in FeLTr through the tree-based exploration.

3.6. 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 the amino-acid sequence only. Focusing on Ω_α^* reduces the number of conformations by more than 50%.

Figure 5(g1), (g5), which plots Rg values of conformations in Ω_α^* versus their lRMSDs from the native (Trp-cage in (g1), eHD in (g5)) , shows that conformations with energies no higher than $\langle E \rangle - \sigma E$ have diverse Rg values. The vertical lines 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 by Gong et al. (2005) 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 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 $\mathrm{Rg} > \mathrm{Rg}_{\mathrm{PDB}}$; and (iii) FeLTr 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 dark gray over the transparent light gray native structure of each considered protein; see Figures 5(h1), (h5) and 6(a2)–(a8). The native structures are obtained from the PDB: PDB id 1l2y for Trp-cage, 1i6c for wwD, 1vii for hp36, 1fd4 for hbd2, 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 Figure 5(h1) shows for Trp-cage, the top two clusters are very similar to the Trp-cage native structure (2–3 Å lRMSD), with some variability in the loop. Figure 5(g1) 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. Similarly, Figure 5(h5) shows the top eHD cluster is very similar (3–4 Å lRMSD) to the native structure. Figure 5(g5) shows that FeLTr obtains more eHD conformations with lower lRMSD from the native and lower Rg values than the MC simulation.

3.7. The Projection Space Layer Helps Obtain Geometrically Distinct Conformations

Conformations that map to the same cell in the projection space can still be significantly different in lRMSD. Figure 5(i1), (i5), which plots energies versus lRSMDs from the native structure (for Trp-cage in (i1) and eHD in (i5)), 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 for Trp-cage and 11 Å for eHD.

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 (see Figure 6(a7)). 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 (Bradley et al. 2005; Ding et al. 2008). Proteins with extended β-sheets such as GB1 and longer sequences (see Section 3.10) highlight challenging topologies for FeLTr and directions for future research.

3.8. FeLTr as a Configurable Platform with Different Coarse-grained Energy Functions

The E_{Rosetta^*} energy function is configured into FeLTr to determine the ability of FeLTr to recover native-like conformations with other state-of-the-art energy functions well established in fragment-based assembly literature. Both MC and FeLTr are employed to generate conformations with the

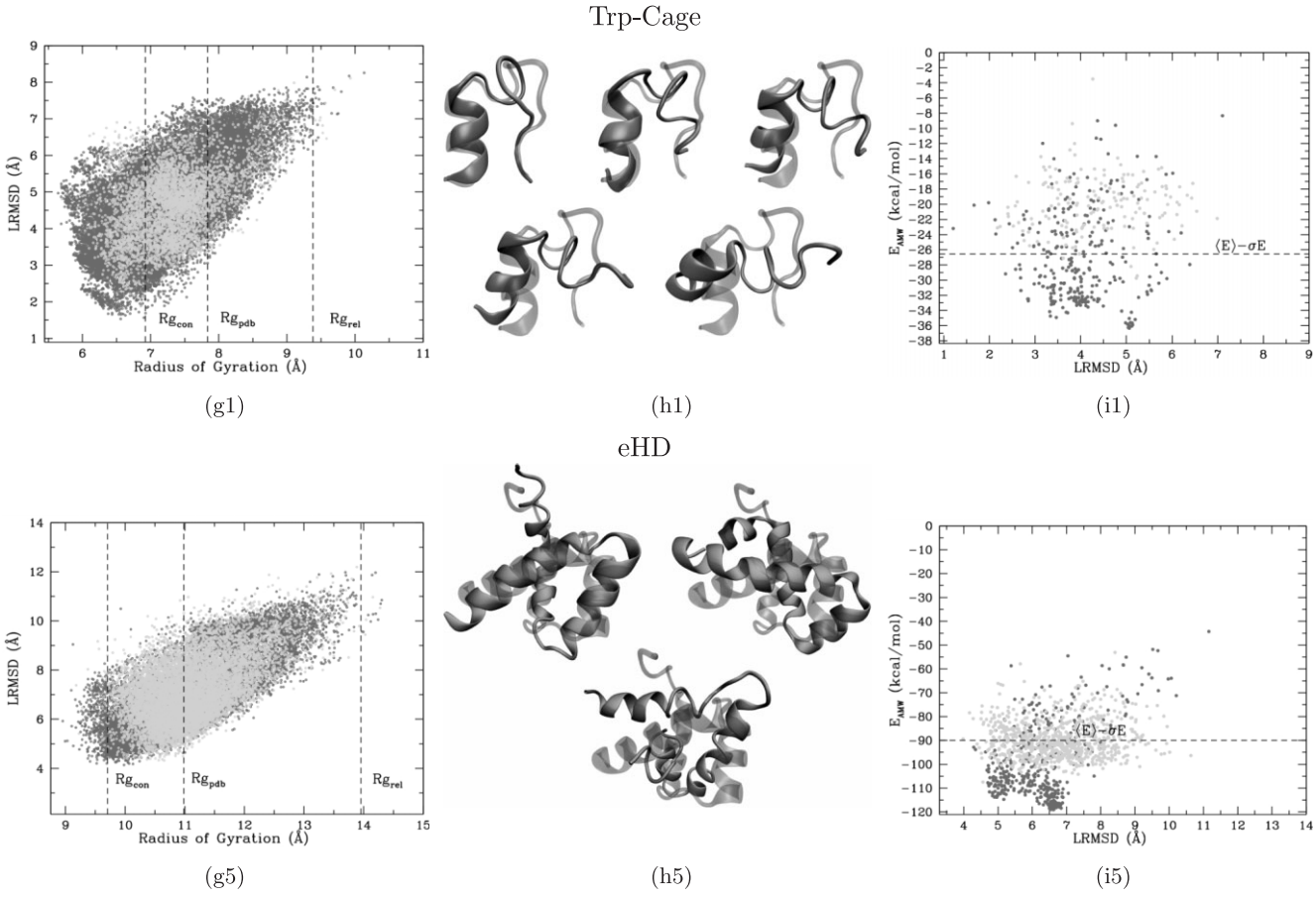

Fig. 5. (g1)–(i1) show data obtained on Trp-cage, and (g5)–(i5) show data obtained on eHD. (g1),(g5) For conformations that meet the energetic criterion, lRMSDs from native are plotted versus Rg values. MC data in light gray are superimposed over `FeLTr` data in dark gray. (h1), (h5) `FeLTr`-computed conformations that meet the energetic criterion are clustered. Lowest-energy conformations of five main clusters (dark gray) are superimposed over the native structure (transparent light gray). (i1), (i5) For conformations that map to the same cell of the projection space as the native, energies are plotted versus lRMSDs from native (MC data over `FeLTr` data).

E_{Rosetta^*} function on the eight protein systems. Figure 7 superimposes energies versus lRMSDs from the native of conformations generated with E_{Rosetta^*} over those generated with E_{AMW}. Data plotted in the left column in Figure 7 are obtained with MC, whereas those in the right are obtained with `FeLTr`.

Figure 7 shows that, overall, the employment of E_{Rosetta^*} does not affect the recovery of the native state among the lowest-energy conformations, whether these conformations are computed with MC or `FeLTr`. It is worth pointing out that, while the lowest-energy conformations obtained with `FeLTr` when employing E_{Rosetta^*} do populate the minima obtained with `FeLTr` when employing E_{AMW}, E_{AMW} seems to allow `FeLTr` populate more and deeper energy minima (see, for instance, results for Trp-cage, hp36, hbd2, eHD, and calbindin). An important observation to make is that, while different realistic coarse-grained energy functions may differ in what energy minima they associate with different regions of conformational space, conformations near the native state will be among the lowest-energy conformations. While differing on the details, both E_{AMW} and E_{Rosetta^*} largely allow `FeLTr` to recover the native state among the lowest-energy conformations.

3.9. Short Refinement of Candidate Conformations in All-atom Detail

Selected conformations obtained by `FeLTr` on each of the eight protein systems are subjected to a short proof-of-concept all-atom energetic refinement in order to showcase how results obtained by `FeLTr` can be employed in detailed biophysical studies or in ab-initio structure prediction. It is worth pointing out that the all-atom energetic refinement stage in ab-initio

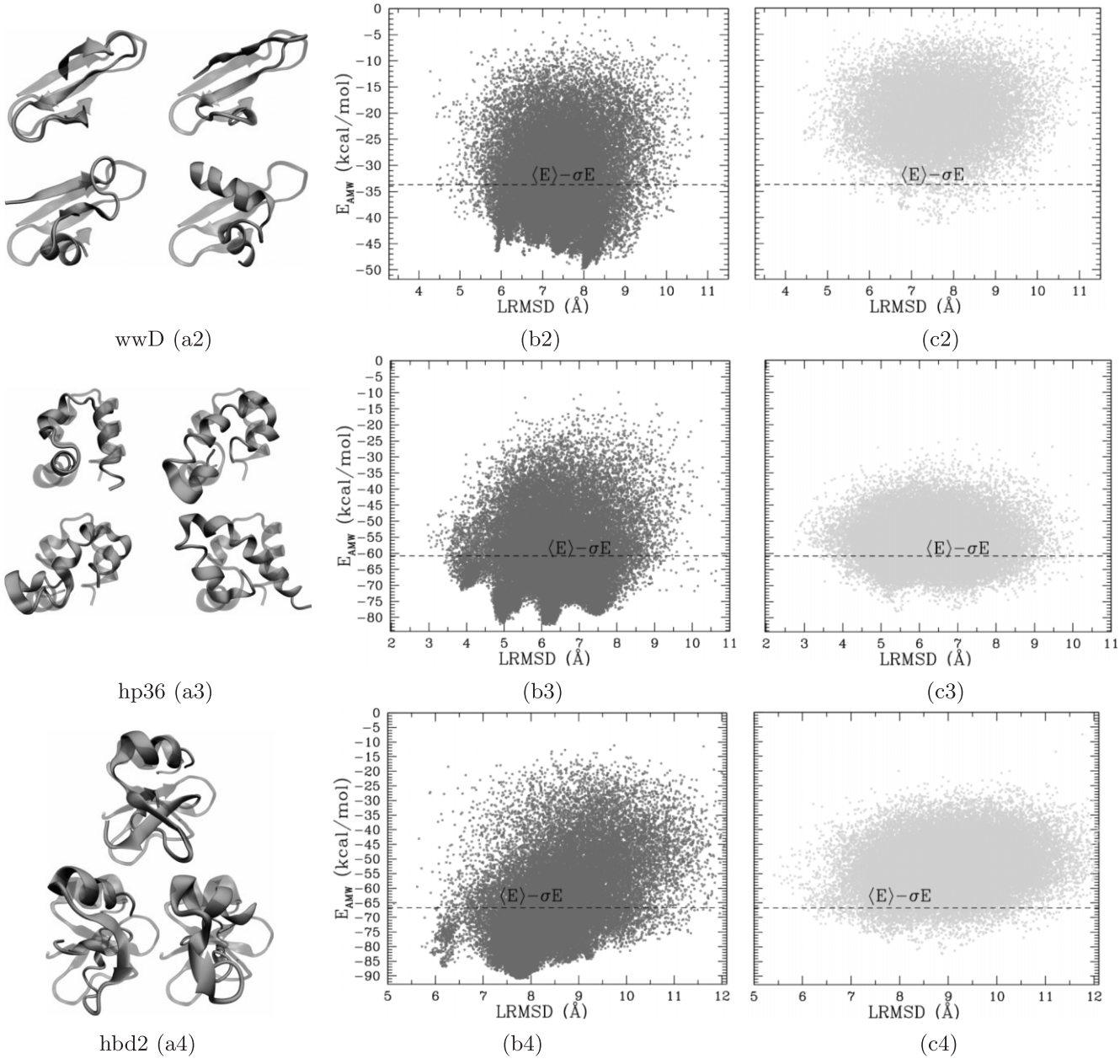

Fig. 6. (a2)–(a8) FeLTr-computed conformations that meet the energetic criterion are clustered. The lowest-energy conformations of the most populated clusters are superimposed in dark gray over the native structure drawn in transparent light gray. Energies of conformations are plotted versus their lRMSDs from the native structure in dark gray for FeLTr-computed conformations in (b2)–(b8) and light gray for MC-computed conformations in (c2)–(c8).

structure prediction is conducted over thousands of coarse-grained conformations chosen for refinement. In contrast, the proof-of-concept refinement showcased here is conducted over only one lowest-energy conformation chosen to represent each of the few top-populated clusters revealed by FeLTr.

We employ the refinement protocol available in the Rosetta package (Bradley et al. 2005). The protocol adds low-energy side-chain configurations to a FeLTr-obtained coarse-grained conformation and minimizes the all-atom Rosetta energy of the resulting all-atom conformation through an MC minimization. The minimization alternates between small random perturbations of the backbone, assignment of low-energy side-chain configurations through side-chain rotamer optimization, and a gradient-based minimization of the energy function in

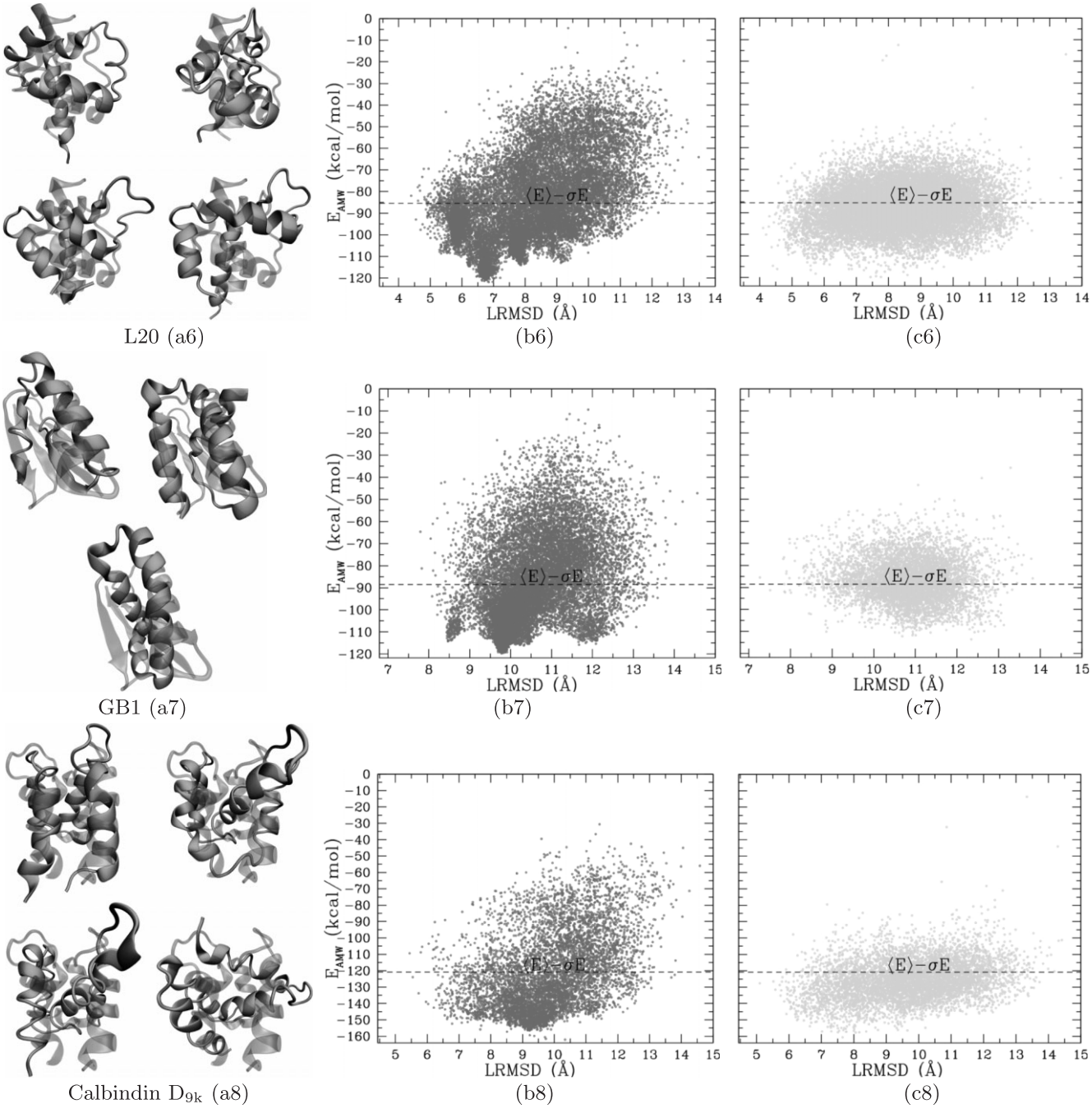

Fig. 6. Continued.

the space of backbone and side-chain dihedral angles. Default parameter values are used for each of the stages of the minimization so that no more than 10 minutes are spent on refining a conformation (the same parameter values that refine a Trp-cage conformation in one minute yield a 10-fold increase in refinement time on Calbindin D_{9k} owing to the quadratic dependence of the Lennard–Jones energy term on protein size). Details on the refinement protocol are available in Bradley et al. (2005).

Table 3 compares `FeLTr` conformations before and after refinement to the native structure for each of the eight protein systems. Column 2 reports the lowest lRMSD value between the native structure and the coarse-grained conformations selected to represent `FeLTr`-obtained minima. The

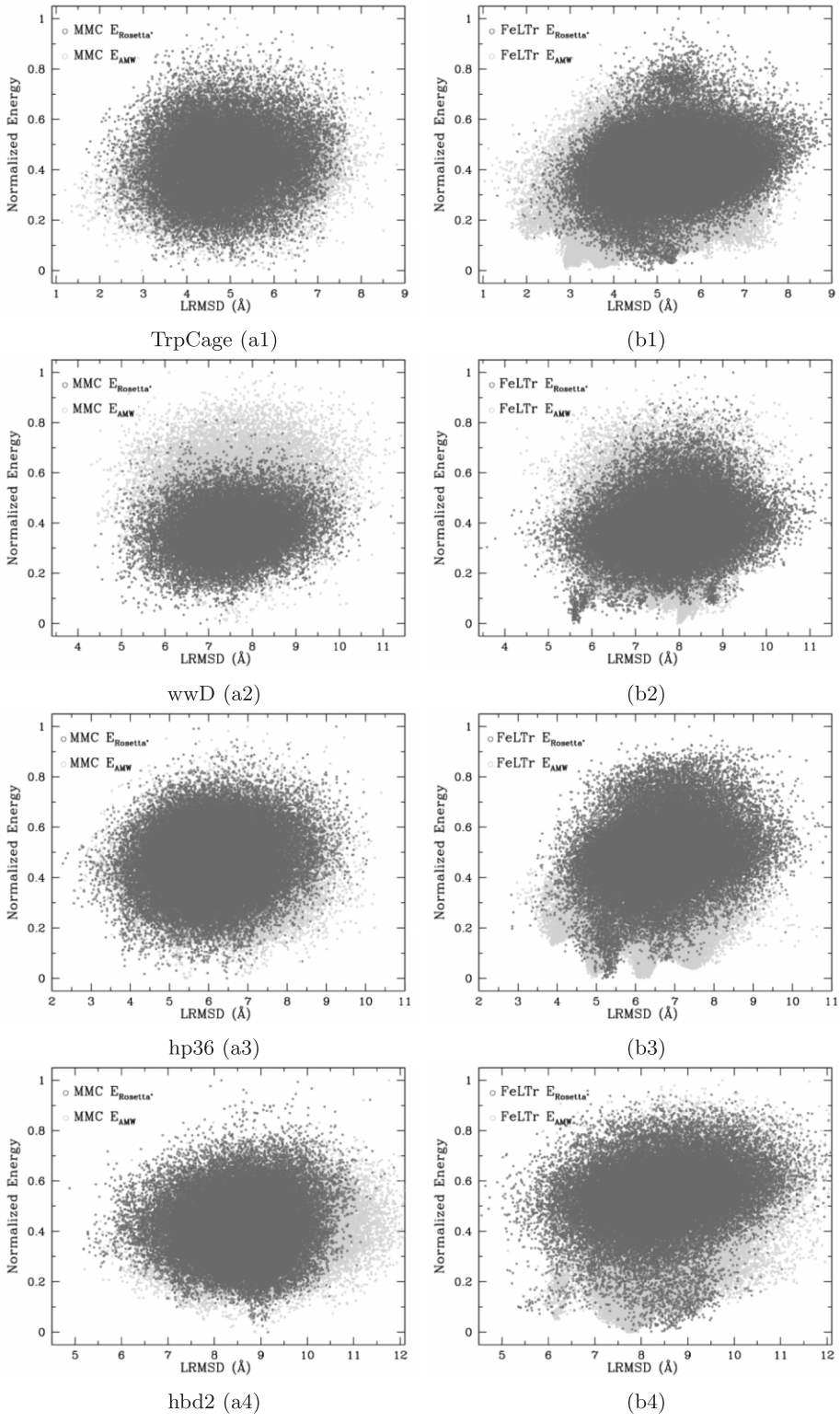

Fig. 7. Data obtained with $E_{\text{Rosetta*}}$ are superimposed in dark gray over those obtained with E_{AMW} in light gray. Energy values are normalized due to the different energy ranges obtained with the two energy functions. Data in (a1)–(a8) are obtained with Metropolis MC (MMC), whereas those in (b1)–(b8) are obtained with `FeLTr`.

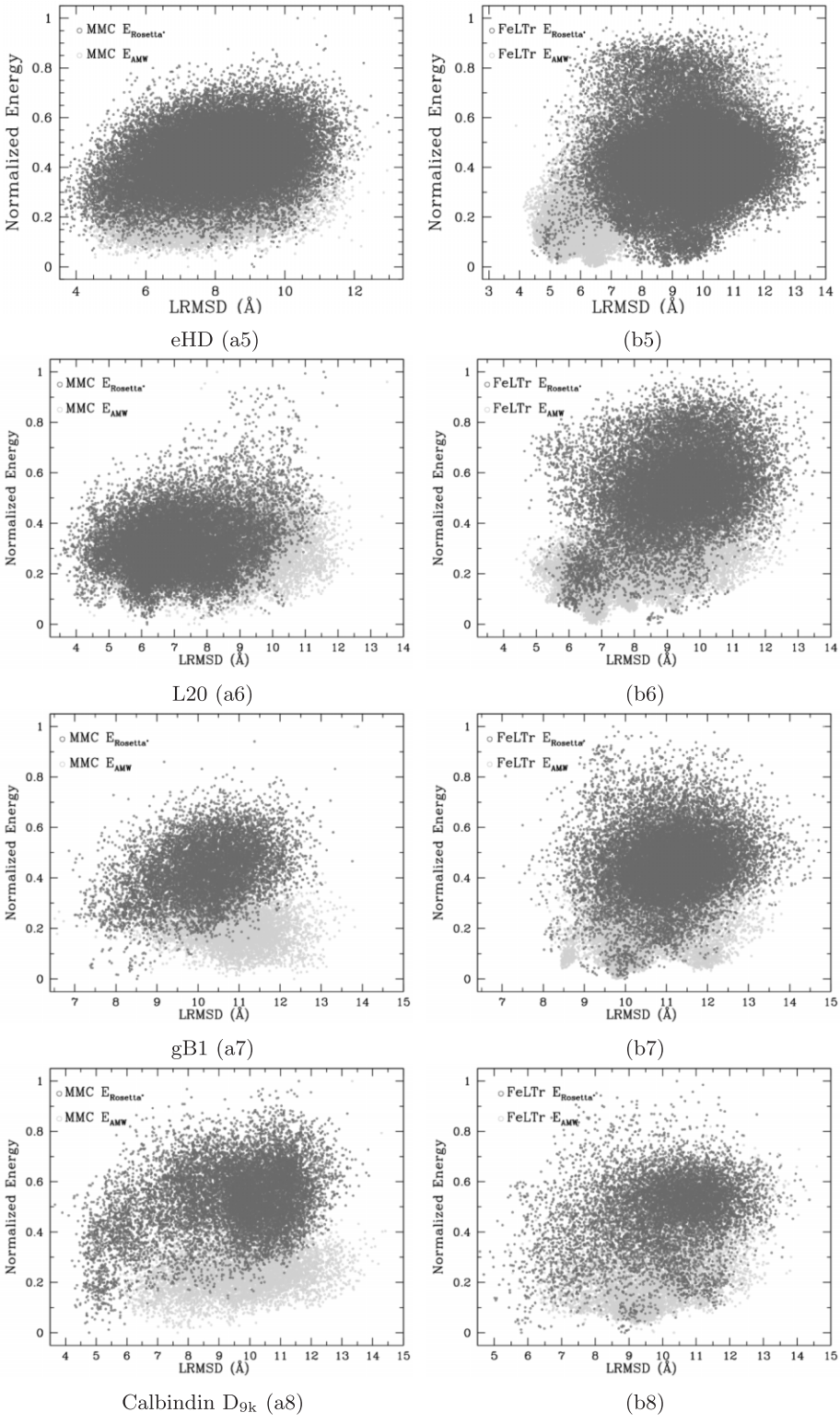

Fig. 7. Continued.

Table 3. Conformations Before and After the Refinement Protocol are Compared with the Corresponding Native Structure for Each Protein; the Last Column Shows the Average Least Root-mean-squared Deviation Obtained when Refining the Native Structure Itself 50 Times

	lRMSD (Å)		GDT_TS (%)		
Protein	Before	After	Before	After	$lRMSD_N$ (Å)
Trp-cage	1.53	2.01	90.00	82.50	0.79
wwD	4.28	3.66	64.42	67.22	1.35
hp36	3.05	2.86	73.61	73.29	2.08
hbd2	4.88	4.76	53.05	55.49	0.61
eHD	4.89	4.75	50.00	51.85	0.84
L20	4.62	4.03	53.33	59.58	1.53
GB1	8.25	8.15	40.63	49.02	1.12
Calbindin D_{9k}	4.62	5.31	41.78	43.42	0.85

coarse-grained conformation with the lowest lRMSD from the native is then subjected to the refinement. To account for the probabilistic feature of the Rosetta refinement protocol, the refinement is carried out on a conformation 50 different times. The lRMSD value reported in column 3 is the average obtained over these runs. In order to directly compare the results in columns 2 and 3, the lRMSD values do not include side-chain atoms. Comparing columns 2 and 3 shows that on most proteins the short refinement brings the backbone of `FeLTr`-obtained conformations closer to the native structure.

The last column in Table 3 subjects the native structure (as obtained from the PDB) of each protein to the all-atom refinement. Again, 50 different runs are carried out, and the average lRMSD between the resulting conformation and the original native structure is reported. These values provide an estimate for the width of the global minimum of each protein. Comparison of these values with the lRMSDs from the native of `FeLTr`-obtained conformations after the refinement shows that in the case of hp36, the refined conformation is within the expected deviation from the native structure. On other proteins, the expected deviations are less than 1.6 Å, suggesting that the narrow global minima may be better populated by further large-scale energetic refinements on selected `FeLTr` conformations.

In addition to comparing lRMSD values between conformations (before and after refinement) and the native structure, Table 3 reports and compares GDT_TS scores. GDT_TS is a similarity score used in the biennial Critical Assessment of Structure Prediction (CASP) competition to compare predictions of different structure prediction groups more robustly than lRMSD (Moult et al. 2009). The GDT_TS score (GDT Total Score), introduced by Zemla (2003), is based on the Global Distance Test (GDT).

GDT_d is the number of atoms of a computed conformation not deviating more than d Å from the atoms of the native structure after optimal alignment. The alignment is optimal in that it maximizes the number of atoms that can fit under the cutoff d. If N is the total number of amino acids, GDT_TS is the mean fraction of amino acids of the native structure not deviating from the predicted conformation after a certain number of optimal alignments with different distance cutoffs. The GDT_TS scores reported in Table 3 are averaged over four optimal alignments:

$$\text{GDT\_TS} = 100 * \frac{\sum_d \text{GDT}_d/\text{N}}{5},$$

for $d \in \{1.0, 2.0, 3.0, 4.0, 5.0\}$. Column 4 shows GDT_TS scores between coarse-grained conformations and the native structure. Scores obtained between the refined conformations and the native structure are shown in column 5. Comparison of these columns reveals that the refinement improves the similarity of computed conformations with the native structure. A more expensive refinement and at a larger scale as part of a detailed study could improve the GDT_TS scores even further.

3.10. FeLTr Sampling Capability on Longer Proteins

The current sequential implementation of `FeLTr` keeps the entire exploration tree in memory. This limits applications on sequences longer than those above, as even a conformational ensemble of a few thousand conformations cannot fit in memory. While directions for future research are laid out in Section 4 to address this limitation, it remains interesting to showcase the results that can be obtained by `FeLTr` on a longer sequence. XF2673, a protein 89 amino-acids long and a target protein (T0464) in CASP 2008, has been selected for this purpose. The obtained results are shown in Figure 8.

Tracking the ability of `FeLTr` to obtain low-lRMSD conformations for XF2673 shows that 24 hours are needed (resulting in a total of 73920 conformations) for the exploration

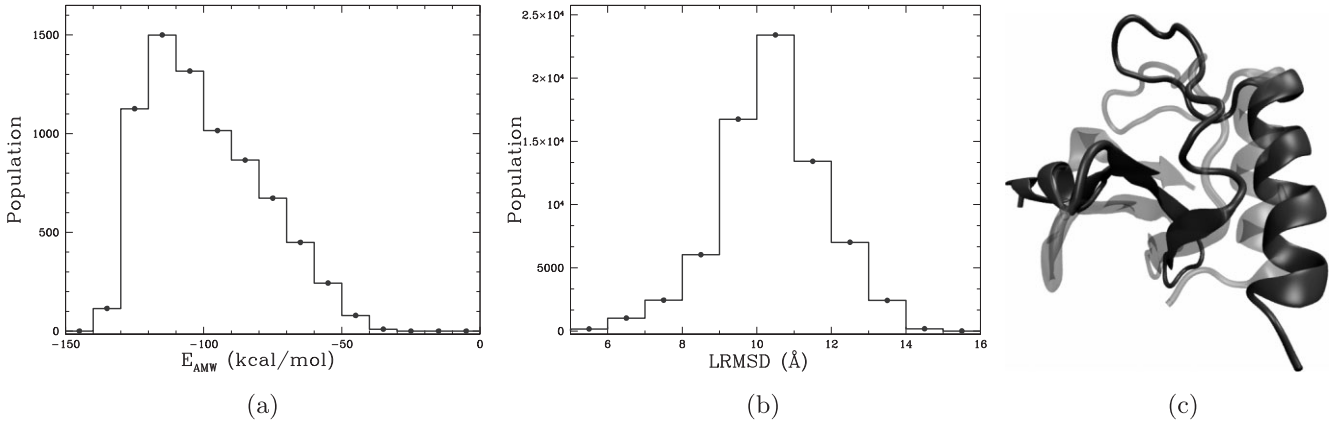

Fig. 8. Histograms show the absolute number of conformations per (a) energy or (b) lRMSD bin. (c) A lowest-energy conformation of the lowest-lRMSD cluster is superimposed (dark gray) over the native structure (transparent light gray).

to obtain a few hundred (432) conformations within 6 Å lRMSD of the native structure. Figure 8 shows the distribution of (a) energies and (b) lRMSDs from the native of the computed conformations. The lowest-energy conformation of the lowest-lRMSD cluster of conformations is shown in Figure 8(c), superimposed in dark gray over the native structure drawn in transparent light gray. The lRMSD value of this conformation from the native structure is 6.4 Å, and its GDT_TS score is 61.89%. By comparison, the three top predictors in CASP 2008 achieved GDT_TS scores of 73.19%, 65.68%, and 64.49% on this protein. These results allow us to conclude that it is possible to extend the sampling capability of `FeLTr` to longer protein chains. Further discussion follows.

4. Discussion

This paper has proposed `FeLTr`, a robotics-inspired tree-based exploration of protein conformational space to enhance the sampling of coarse-grained native-like conformations when employing only amino-acid sequence information for a protein at hand. Two discretization layers of the energy surface and a low-dimensional projection space are employed to guide the tree towards low-energy conformations in under-explored regions of the conformational space. Analysis on a diverse set of proteins suggests that `FeLTr` can serve as a filtering step. In a matter of a few hours on a single CPU, `FeLTr` reveals native-like conformations that are good candidates for further detailed refinements by larger studies in protein engineering and design.

The coarse graining in `FeLTr` is based on the backbone-based theory of protein folding (Rose et al. 2006). Other work that also employs coarse graining shows that geometry presculpts the protein energy surface (Hoang et al. 2007). `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.

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 the work in Plaku et al. (2007), 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. Finding such coordinates remains the subject of much research (Das et al. 2006). Rather, these coordinates are a first attempt towards integrating a projection space in the exploration of the protein conformational space. Since the 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 such as DSLX (Plaku et al. 2007), PDST (Ladd and Kavraki 2005) and the approach of Kurniawati and Hsu (2006) that rely on low-dimensional projections can potentially benefit from using USR projections to effectively explore high-dimensional spaces.

Owing to its employment of a database of physical fragment configurations, `FeLTr` 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. Han and Amato (2001)

have also proposed the usage of chain fragments in sampling valid configurations.

In addition to improving performance on proteins with native topologies that are rich in extended β sheets, extending applicability to longer proteins and/or proteins with diverse functional states are the subject of future research. The design of novel projections to further enhance the exploration is a valid direction for future work. A distributed implementation together with strategies to make the exploration tree sparser will allow the application domain of `FeLTr` to be extended to longer sequences than those considered here. Employing coarser representations such as C_α traces may additionally enhance the sampling capability of `FeLTr` and allow the conformational space of longer ($>$ 300 amino acids) sequences to be explored in more detail. Future work will also address proteins with diverse functional states. Instead of one Rg_{PDB} threshold, different values of Rg thresholds, obtained from experiment or defined systematically over a range as in Shehu et al. (2009), can be employed to extend applications on such proteins.

The interplay between energy and geometry allows `FeLTr` to populate distinct local minima. The weight function that biases the expansion of the search tree from low-energy levels is effective on small proteins with no alternative functional states. While the MC expansion and the geometric projection layer allow `FeLTr` to extend towards higher-energy and less-populated regions of conformational space, different weighting functions are currently under investigation for their ability to extend applicability to proteins with more complex energy landscapes and alternative functional states.

The two-layered tree-based search in `FeLTr` can be employed to populate specific regions of conformational space. Given a coarse-grained representation, energy function, and desired reaction coordinates, the basic search approach in `FeLTr` can be employed to efficiently compute coarse-grained conformations that are of low energy according to the employed energy function and diverse in the projection space of the desired reaction coordinates.

`FeLTr` makes a first step towards rapidly computing coarse-grained native-like conformations from amino-acid sequence. Analysis shows the native structure is among computed conformations. The short proof-of-concept refinements suggest that the lowest-energy conformations obtained by `FeLTr` are good candidates for further refinement in all-atom detail. `FeLTr` can serve as an initial filter in larger studies aimed at extracting detailed structural and functional properties of novel sequences.

Acknowledgments

The authors thank Erion Plaku, Srinivas Akella, Jyh-Ming Lien, Oliver Brock, and anonymous reviewers for constructive feedback, and Adam Zemla for sharing an implementation of the GDT_TS score. This research was partly funded by the Bioengineering Seed Grant of the Volgenau School of Information Technology and Engineering at George Mason University. A preliminary version of this work was published in *Robotics: Science and Systems*, 2009, pp. 241–248. This work substantially improves upon the overall method introduced in the preliminary work and presents new experiments that were not included in the preliminary work.

References

Amato, N. M., Dill, K. A. and Song, G. (2002). Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *Journal of Computational Biology*, **10**: 239–255.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**: 223–230.

Apaydin, M. S., Singh, A. P., Brutlag, D. L. and Latombe, J.-C. (2001). Capturing molecular energy landscapes with probabilistic conformational roadmaps. *Proceedings IEEE International Conference on Robotics and Automation*, Vol. 1, pp. 932–939.

Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D. and Latombe, J.-C. (2003). Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. *Journal of Computational Biology*, **10**: 257–281.

Ballester, P. J. and Richards, G. (2007). Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry*, **28**: 1711–1723.

Barraquand, J. and Latombe, J.-C. (1991). Robot motion planning: a distributed representation approach. *The International Journal of Robotics Research*, **10**: 628–649.

Bonneau, R. and Baker, D. (2002). De novo prediction of three-dimensional structures for major protein families. *Journal of Molecular Biology*, **322**: 65–78.

Bradley, P., Misura, K. M. S. and Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, **309**: 1868–1871.

Brunette, T. J. and Brock, O. (2009). Guiding conformation space search with an all-atom energy potential. *Proteins: Structure, Function and Bioinformatics*, **73**: 958–972.

Burns, B. and Brock, O. (2007). Single-query motion planning with utility-guided random trees. In *ICRA*, pages 3307–3312, Rome, Italy.

Case, D. A., Darden, T. A., Cheatham, T. E. I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S. and Kollman, P. A. (2006). Amber 9.

Chiang, T. H., Apaydin, M. S., Brutlag, D. L., Hsu, D. and Latombe, J.-C. (2007). Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: folding rates and phi-values. *Journal of Computational Biology*, **14**: 578–593.

Choset, H. et al. (2005). *Principles of Robot Motion: Theory, Algorithms, and Implementations*, 1st edition. Cambridge, MA, MIT Press.

Clementi, C. (2008). Coarse-grained models of protein folding: toy-models or predictive tools? *Current Opinion in Structural Biology*, **18**: 10–15.

Cortes, J., Simeon, T., de Angulo, R., Guieysse, D., Remaud-Simeon, M. and Tran, V. (2005). A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, **21**: 116–125.

Das, P., Moll, M., Stamati, H., Kavraki, L. E. and Clementi, C. (2006). Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Science of the U S A*, **103**: 9885–9890.

DeBartolo, J., Colubri, A., Jha, A. K., Fitzgerald, J. E., Freed, K. F. and Sosnick, T. R. (2009). Mimicking the folding pathway to improve homology-free protein structure prediction. *Proceedings of the National Academy of Science of the U S A*, **106**: 3734–3739.

Dill, K. A. and Chan, H. S. (1997). From levinthal to pathways to funnels. *Nature Structural Biology*, **4**: 10–19.

Ding, F., Tsao, D., Nie, H. and Dokholyan, N. V. (2008). Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure*, **16**: 1010–1018.

Georgiev, I. and Donald, B. R. (2007). Dead-end elimintation with backbone flexibility. *Bioinformatics*, **23**: 185–194.

Gong, H., Fleming, P. J., and Rose, G. D. (2005). Building native protein conformations from highly approximate backbone torsion angles. *Proceedings of the National Academy of Science of the U S A*, **102**: 16227–16232.

Han, L. (1994). Hybrid probabilistic RoadMap-Monte Carlo motion planning for closed chain systems with spherical joints. *Proceedings of ICRA*, New Orleans, LA, pp. 920–926.

Han, L. and Amato, N. M. (2001). A kinematics-based probabilistic roadmap method for closed chain systems. *Algorithmic and Computational Robotics: New Directions*, Donald, B. R., Lynch, K. M. and Rus, D. (eds). Wellesley, MA, A.K. Peters.

Hart, W. E. and Istrail, S. (1997). Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. *Journal of Computational Biology*, **4**: 1–22.

Hoang, T. H., Trovato, A., Seno, F., Banavar, J. R. and Maritan, A. (2007). Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Science of the U S A*, **101**: 7960–7964.

Jaillet, L., Cortes, J. and Simeon, T. (2008). Transition-based RRT for path planning in continuous cost spaces. *IEEE/RSJ International Conference on Intelligent Robotic Systems*. Stanford, CA, AAAI, pp. 22–26.

Jain, A. K., Dubes, R. C. and Chen, C. C. (1987). Bootstrap techniques for error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **9**: 628–633.

Kim, K. M., Jernigan, R. L. and Chirikjian, G. S. (2002). Efficient generation of feasible pathways for protein conformationa transitions. *Biophysical Journal*, **83**: 1620–1630.

Kirillova, S., Cortes, J., Stefaniu, A. and Simeon, T. (2008). An nma-guided path planning approach for computing large-amplitude conformational changes in proteins. *Proteins: Structure, Function and Bioinformatics*, **70**: 131–143.

Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology*, **323**: 297–307.

Kortemme, T. and Baker, D. (2004). Computational design of protein–protein interactions. *Current Opinion in Structural Biology*, **8**: 91–97.

Kurniawati, H. and Hsu, D. (2006). Workspace-based connectivity oracle: An adaptive sampling strategy for PRM planning. *Proceedings of WAFR* (*Springer Tracts in Advanced Robotics*, Vol. 47). New York, NY, Springer, pp. 35–51.

Ladd, A. M. and Kavraki, L. E. (2005). Motion planning in the presence of drift, underactuation and discrete system changes. *Robotics: Science and Systems*, Boston, MA, pp. 233–241.

Lee, A., Streinu, I. and Brock, O. (2005). A methodology for efficiently sampling the conformation space of molecular structures. *Journal of Physical Biology*, **2**: 108–S115.

Lee, D., Redfern, O., and Orengo, C. (2007) Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology*, **8**: 995–1005.

Lee, J. and Scheraga, H. A. (1999). Conformational space annealing by parallel computations: Extensive conformational search of Met-enkephalin and of the 20-residue membrane-bound portion of melittin. *International Journal of Quantum Chemistry*, **75**: 255–265.

Lee, J., Scheraga, H. A. and Rackovsky, S. (1997). New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *Journal of Computational Chemistry*, **18**: 1222–1232.

Lee, J., Scheraga, H. A. and Rackovsky, S. (1998). Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. *Biopolymers*, **46**: 103–115.

Milik, M., Kolinski, A. and Skolnick, J. (1997). Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *Journal of Computational Chemistry*, **18**: 80–85.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. and Tramontano, A. (2009). Critical assessment of methods of protein

structure prediction (CASP) round VIII. *Proteins: Structure, Function and Bioinformatics*, **77**: 1–4.

Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z. and Wolynes, P. G. (2004). Water in protein structure prediction. *Proceedings of the National Academy of Science of the U S A*, **101**: 3352–3357.

Plaku, E., Kavraki, L., and Vardi, M. (2007). Discrete search leading continuous exploration for kinodynamic motion planning. In *Robotics: Science and Systems*, Atlanta, GA, pp. 326–333.

Rodriguez, S., Thomas, S., Pearce, R. and Amato, N. (2006). RESAMPL: A Region-Sensitive Adaptive Motion Planner. In *Proceedings of WAFR* (*Springer Tracts in Advanced Robotics*, Vol. 47). New York, NY, Springer, pp. 285–300.

Rose, G. D., Fleming, P. J., Banavar, J. R. and Maritan, A. (2006). A backbone-based theory of protein folding. *Proceedings of the National Academy of Science of the U S A*, **103**: 16623–16633.

Russell, S. and Norvig, P. (2002). *Artificial Intelligence: A Modern Approach*, 2nd edition. New York, NY, Prentice Hall.

Sánchez, G. and Latombe, J.-C. (2002). On delaying collision checking in PRM planning: application to multi-robot coordination. *The International Journal of Robotics Research*, **21**: 5–26.

Shehu, A., Kavraki, L. E. and Clementi, C. (2008). Unfolding the fold of cyclic cysteine-rich peptides. *Protein Science*, **17**: 482–493.

Shehu, A., Kavraki, L. E. and Clementi, C. (2009). Multi-scale characterization of protein conformational ensembles. *Proteins: Structure, Function and Bioinformatics*, **76**; 837–851.

Song, G. and Amato, N. M. (2004). A motion planning approach to folding: from paper craft to protein folding. *IEEE Transactions on Robotics and Automation*, **20**: 60–71.

Stilman, M. and Kuffner, J. J. (2008). Planning among movable obstacles with artificial constraints. *The International Journal of Robotics Research*, **12**: 1295–1307.

van den Berg, J. P. and Overmars, M. H. (2005). Using workspace information as a guide to non-uniform sampling in probabilistic roadmap planners. *The International Journal of Robotics Research*, **24**: 1055–1071.

Wang, G. and Dunbrack, R. L. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, **19**: 1589–1591.

Yang, Y. and Brock, O. (2005). Efficient motion planning based on disassembly. *Robotics: Science and Systems*, Cambridge, MA, pp. 97–104.

Yin, S., Ding, F. and Dokholyan, N. V. (2007). Eris: an automated estimator of protein stability. *Nature Methods*, **4**: 466–467.

Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, **31**: 3370–3374. http://as2ts.llnl.gov/AS2TS/LGA/lga.html.