A Robotics-inspired Method to Sample Conformational Paths Connecting Known Functionally-relevant Structures in Protein Systems

Kevin Molloy[1] and Amarda Shehu[1,2,3*]
[1]*Department of Computer Science*
[2]*Department of Bioinformatics and Computational Biology*
[3]*Department of Bioengineering*
George Mason University
Fairfax, VA, 22030, United States
kmolloy1@gmu.edu, amarda@gmu.edu
**Corresponding Author*

***Abstract*—Characterization of transition trajectories that take a protein between different functional states is an important yet challenging problem in computational biology. Approaches based on Molecular Dynamics can obtain the most detailed and accurate information but at considerable computational cost. To address the cost, sampling-based path planning methods adapted from robotics forego protein dynamics and seek instead conformational paths, operating under the assumption that dynamics can be incorporated later to transform paths to transition trajectories. Existing methods focus either on short peptides or large proteins; on the latter, coarse representations simplify the search space. Here we present a robotics-inspired tree-based method to sample conformational paths that connect known structural states of small- to medium-size proteins. We address the dimensionality of the search space using molecular fragment replacement to efficiently obtain physically-realistic conformations. The method grows a tree in conformational space rooted at a given conformation and biases the growth of the tree to steer it to a given goal conformation. Different bias schemes are investigated for their efficacy. Experiments on proteins up to 214 amino acids long with known functionally-relevant states more than 13Å apart show that the method effectively obtains conformational paths connecting significantly different structural states.**

***Keywords*-protein conformational transitions; diverse functional states; transition trajectories; conformational paths; tree-based stochastic search**

I. Introduction

Protein molecules often modulate their biological function by undergoing conformational changes that allow them to transition between different functional states [1]. Experimental techniques, such as X-ray crystallography, have caught some of these multi-functional proteins in the act. Significantly different X-ray structures are now available for many proteins in the Protein Data Bank [2].

While experimental techniques can resolve structures relevant for function, they are not able to elucidate the detailed transition in terms of intermediate conformations. Computational techniques are in principle able to explore the protein energy surface in search of transition trajectories. However, computing these trajectories is challenging [3]. Transition trajectories can connect structural states more than 100Å apart. This length scale is up to 2 orders of magnitude larger than a typical interatomic distance of 2Å. Transitions can demand even larger μs-ms time scales, which is $6-12$ orders of magnitude larger than typical atomic oscillations of the fs-ps timescale.

Given these scales, it is not surprising to draw analogies between computing transition trajectories and folding, where the starting conformation represents an unfolded rather than an alternative functional state. Due to analogies with folding, typical exploration methods based on Molecular Dynamics (MD) and Monte Carlo [4] can be used to compute transition trajectories. In particular, targeted, steered, and biased MD approaches bias the MD exploration towards the goal [5].

MD-based approaches take into account protein dynamics and so provide detailed information into the time scales associated with conformational changes in a transition trajectory. Given the length and time scales involved, these approaches can be computationally demanding. Long simulation times may be needed for a transition trajectory to go over energy barriers related with transient intermediate states. It is very costly to navigate the protein energy surface in search of transition trajectories with MD-based approaches [5].

Many methods enhance MD sampling to improve efficiency. They often do so by incorporating bias at the expense of obtaining possibly different transition trajectories. Methods include umbrella sampling, replica exchange, local flattening of the energy surface, activation relaxation, conformational flooding, and others [6]–[11]. Efficiency concerns are also addressed through coarse graining and techniques based on normal mode analysis and elastic network modeling [12], [13], [13]–[20]. Some methods focus on deforming a trivial conformational path (obtained, for instance, through morphing) to improve its energy profile. Examples include the nudged elastic band, morphing, zero-temperature string, and finite-temperature string methods [21]–[26].

978-1-4673-2747-3/12/$31.00 ©2012 IEEE

A different class of methods addresses efficiency concerns by adapting sampling-based search algorithms developed for the robot motion-planning problem. The objective is to obtain trajectories that take a robot from a start to a goal configuration. Methods building over the Probabilistic Roadmap (PRM) framework [27], [28] have exploited analogies between robot articulated chains and protein chains to fold small proteins [29], [30]. Methods based on tree-based planners, such as Rapidly-exploring Random Tree (RRT) [31], Expansive Spaces Tree (EST) [32], and Path-directed Subdivision Tree (PDST) [33] have been proposed to model conformational changes and flexibility, predict the native structure, and even compute conformational paths connecting given structural states in proteins [34]–[39].

In particular, the T-RRT [38] and PDST [39] methods have focused on the problem of computing conformational changes connecting two given structures. Due to removal of dynamics, the contribution of these methods is not in producing transition trajectories but rather a sequence of conformations (a conformational path) with a credible energy profile. The working assumption is that credible conformational paths can be locally deformed with techniques that consider dynamics to obtain transition trajectories. While T-RRT has been shown to connect known low-energy states of the dialanine peptide (2 amino acids long) [38], the PDST method has been shown to produce credible information on the order of conformational changes connecting functional states of large proteins ($200-500$ amino acids long) [39]. Both methods control the dimensionality of the conformational space by either focusing on systems with very few amino acids [38] or by employing very coarse-grained representations to limit the number of modeled parameters in large proteins [39].

In this paper we propose a robotics-inspired tree-based probabilistic search method to sample conformational paths connecting two given structural states of small- to medium-size proteins. These systems range in size from a few dozen to a few hundred amino acids (214 amino acids in the largest system studied in this paper). Instead of employing very coarse-grained representations to simplify the search space, the proposed method models all backbone dihedral angles. The size of the search space is controlled, however, through the molecular fragment replacement technique, which allows efficiently obtaining physically-realistic protein conformations by essentially bundling together backbone dihedral angles and sampling physically-realistic moves for them.

The method grows a tree rooted at a given start conformation in iterations. At each iteration, a conformation is selected for expansion. The expansion employs molecular fragment replacement and the Metropolis criterion to bias the tree towards low-energy conformations over time. Due the employment of expansions and discretization layers to make decisions on how and where to grow the tree, the method can be considered an adaptation of EST and grid-based methods in robotics [28].

Since the objective is to steer the tree towards the given goal conformation, we experiment with different bias schemes both in the selection and expansion procedures. A discretization layer organizes conformations in the tree in levels based on their proximity to the goal. A probability distribution is defined through weighting functions over the levels in order to bias the selection to conformations close to the goal. Different weighting functions are analyzed here for trade-off between efficiency in reaching the goal and premature convergence to local optima. A randomized strategy is also proposed that shifts between different bias strengths during selection. Biasing the expansion of the tree is also considered, by only adding to the tree conformations that improve the proximity to the goal over the parent.

Experiments are conducted on systems ranging in size from a few dozen to a few hundred amino acids. Our study in this paper focuses on three well-characterized systems, Trp-Cage, calmodulin (CaM), and adenylate kinase (AdK). The results show that the method is effective in elucidating conformational paths on these systems. Due to the Metropolis criterion and a state-of-the-art energy function, the paths also have credible energy profiles.

A tree elucidates only one path, unless a tolerance region is defined around the goal structure, as we do here. While roadmap-based methods can in principle highlight many paths, they suffer from the steering problem (finding moves to connect neighboring conformations). To an extent, this problem also arises in RRT-based methods. This is one of the reasons we pursue an EST-based method in this work, where the tree pushes out in conformational space by expanding selected conformations. As is common practice [38], [39], multiple executions can be used to obtain different paths. We emphasize that these paths are not transition trajectories. The conformations in them can be considered milestones during deformations of these paths into transition trajectories.

The proposed method makes an important contribution to the problem of computing conformational paths connecting two given states of a protein. Sampling values for individual dihedral angles is not feasible on proteins more than a few amino acids to search the space connecting states sometimes more than 13Å apart. On the other hand, the work described here makes the case that one does not have to resort to very coarse-grained representations to limit the number of modeled parameters. Instead, parameters can be bundled and credible moves, extracted from known low-energy structures of proteins, can be proposed for a series of consecutive angles in order to efficiently obtain physically-realistic intermediate conformations. This paper shows a proof of concept of the proposed approach. As we discuss in section IV, the method proposed here opens up new lines of investigation. The results in section III suggest that work in this direction is very promising to obtain credible conformational paths connecting diverse functional states of a protein. We now proceed in section II with details on the proposed method.

II. METHODS

A. Problem Statement

The input to the method are two PDB-obtained structures (start and goal) corresponding to functional states. The output consists of conformational paths. A path is a series of conformations terminating within some threshold distance of the goal. We use a semi-metric, least Root-Mean-Squared-Deviation (lRMSD), which measures structural dissimilarity between two conformations. lRMSD measures the weighted Euclidean distance between corresponding atoms in two aligned conformations (alignment minimizes distances due to rigid-body transformations). Low values indicate high similarity, but interpretation is difficult for values > 6Å. The energetic difference between conformations in a path is limited through the Metropolis criterion which employs an effective temperature to control the height of energetic barriers that can be crossed by a path (detailed below). Different executions of the method result in different paths, which can then be analyzed through clustering and other techniques to determine, for instance, highly-populated intermediates.

B. Main Algorithmic Components of Proposed Search

The method grows a tree in conformational space, rooted at a given start conformation. The tree grows in iterations, at each iteration selecting a conformation and expanding it. The selection determines where the tree grows. Our selection procedure employs a discretization layer. Conformations in the tree are projected on a 1-dimensional grid discretizing their lRMSDs to the goal. Grid boundaries are set at $[0, D]$, where D is the lRMSD between the start and goal. Conformations with lRMSD higher than D to the goal are mapped to the D level. Levels in the grid are separated by 1Å here.

The grid discretizes the explored conformational space and allows biasing the growth of the tree towards the goal. While lRMSD is an imperfect measure, its employment as a progress coordinate has some precedent in biophysical studies that detect conformational transitions in CaM [12]. Below we discuss additional progress coordinates. Using lRMSD, we investigate different bias schemes, as a strong bias towards selecting low-lRMSD conformations may perform well in a small system or in a particular run due to the probabilistic nature of the method and quickly drive the tree towards the goal. However, a strong bias may also lead to premature convergence to local optima and prevent the tree from approaching the goal. This is the classic depth vs. breadth issue that characterizes greedy exploration.

Different bias schemes can be naturally defined through weighting functions over levels of the grid. A quadratic function, referred to as QUAD, can be defined to associate a weight $w(l) = 1/[1 + l^2] + \epsilon$, with level l in the grid. The function biases the selection towards levels with low lRMSD to the goal, and ϵ is set to a small value to ensure that higher-lRMSD conformations can be selected if the method is given enough time. Another weighting function, LINEAR, defined as $w(l) = 1/[1 + l] + \epsilon$, reduces the bias. UNIFORM removes bias entirely, as in $w(l) = 1/[\#\text{levels}]$. A probability distribution function then associates probability of selection $p(l) = w(l)/[\sum_{levels\, l'} w(l')]$ with a level l. Once a level l is selected with probability $p(l)$, any conformation that maps to it is selected with equal probability for expansion.

We also provide the first steps towards a dynamic combination of different bias schemes. We compare the three basic bias schemes above in COMBINE_{90-10}, which $p = 90\%$ of the time grows the tree with no selection bias (effectively employing UNIFORM), and $1-p = 10\%$ of the time employs QUAD. The value of p can be adaptively set in a reactive scheme to balance between tree depth and breadth, and this is a direction we will investigate in future work.

The expansion procedure employs the molecular fragment replacement technique. A move sampled from a library of physically-realistic moves is proposed to modify the conformation selected for expansion. The modification is accepted according to the Metropolis criterion with probability $e^{-\delta E/K_B T}$. δE is the energy difference between the resulting conformation and its parent, and K_B is the Boltzmann constant. T is an effective temperature that serves as a scaling parameter through which to control the height of an energy barrier crossed by the child conformation. In the study presented here, we employ a fixed temperature that accepts a 10 kcal/mol energy increase with probability 0.1.

Section III shows that this temperature is effective, but connectivity in more complex systems, such as AdK, can benefit from the ability to cross higher-energy barriers. Reactive schemes that change the temperature as needed to make progress, introduced in [38] for the dialanine peptide system, present an interesting direction to further enhance the exploration of the method we propose here.

In addition to the global bias schemes described above in the context of selection, we investigate a local bias in the context of the expansion to grow the tree with conformations that improve proximity to the goal. Essentially, moves are proposed until m conformations are obtained that all meet the Metropolis criterion. The conformation with the lowest lRMSD to the goal is considered for addition to the tree. Two different schemes are analyzed in this paper, one in which the lowest-lRMSD child is added to the tree, and another where the addition is only carried out if the child's lRMSD to the goal is no higher than that of the parent.

A tolerance region of 3-4Å around the goal allows defining a goal region and essentially obtaining many paths from one execution of the method. To some extent, the paths can be redundant, as the tree may waste time sampling similar conformations. An additional discretization layer can be defined over conformations in the tree, using shape- or contact-summarizing features (which we have previously studied for their ability to organize conformations [36], [37], [40]) to represent and project conformations. Here we

investigate one such projection. In addition to the 1d lRMSD grid described above, The 3 USR-based shape-summarizing features employed in [36] are used to associate a 3d grid with the tree. A second weighting function can be defined to bias the tree away from similar conformations (essentially, cells with many conformations and that have been selected many times before are penalized with higher weights). A preliminary investigation of this additional discretization shows that this forces the tree to find diverse paths. In one extreme instance shown in section III, insisting on diversity allows finding alternative paths that come closer to the goal.

The tree-based method summarized above employs a specific representation of a protein chain, a state-of-the-art coarse-grained energy function capable of interfacing with this representation, and the molecular fragment replacement technique. We now describe each in detail.

1) Representation of a Protein Chain: The representation employed here uses only ϕ, ψ backbone dihedral angles (there are $2n$ such angles for a chain of n amino acids). Side chains are sacrificed, as side-chain packing techniques can be used to add all-atom detail if needed. The representation is a version of the idealized geometry model, which fixes bond lengths and angles to idealized (native) values. Forward kinematics allows computing cartesian coordinates of backbone atoms from the ϕ, ψ angles in the representation.

2) Employed Energy Function: The function is a modification of the Associative Memory hamiltonian with Water (AMW) [41] used in structure prediction [36], [40]. AMW sums non-local terms (local terms are kept at ideal values in the idealized geometry model): $E_{\mathrm{AMW}} = E_{\mathrm{Lennard-Jones}} + E_{\mathrm{H-Bond}} + E_{\mathrm{contact}} + E_{\mathrm{burial}} + E_{\mathrm{water}}$. $E_{\mathrm{Lennard-Jones}}$ is implemented after the 12-6 Lennard-Jones potential in AMBER9 [42] but allows a soft penetration of van der Waals spheres. $E_{\mathrm{H-Bond}}$ models hydrogen bonds. E_{contact}, E_{burial}, and E_{water}, allow formation of non-local contacts, a hydrophobic core, and water-mediated interactions. Further details can be found in [41], [43].

3) Molecular Fragment Replacement: The fragment replacement technique has allowed ab-initio structure prediction methods to make great advancements [44]–[47]. The basic idea is that a subset of non-redundant structures are obtained from the PDB, and configurations (ϕ, ψ angles) that can be defined for k consecutive amino acids (fragments) are excised from these structures and stored in a library. We direct the reader to [37] for details on how the library is constructed. Here we use a fragment of length 3 rather than a longer fragment to limit the magnitude of the jump in conformational space resulting from one replacement. A position i in the chain is sampled uniformly at random, and a fragment $[i, i+2]$ is defined. A configuration (6 backbone dihedral angles) is sampled uniformly at random over those available for the fragment in the library. This constitutes a move, accepted according to the Metropolis criterion detailed above. The key advantage is that the move is physically-realistic and increases the probability of obtaining a feasible conformation.

III. Results

A. Implementation Details and Experimental Setup

Experiments are conducted on three systems, Trp-Cage, CaM, and AdK of respective lengths of 20, 144, and 214 amino acids (aa). Ten independent executions of the method are carried out on each system. The termination criterion is $1,000$ conformations for Trp-Cage and $10,000$ conformations for CaM and AdK (Trp-Cage is a small system, and our investigation shows that $1,000$-$2,000$ conformations yield similar overall performance). The time demands of one execution of the method span from $1-2$ minutes on Trp-Cage to 8 hours for CaM and 36 hours for AdK. Energy function evaluations make up 90% of CPU time.

On Trp-Cage we connect an extended conformation to the native structure (PDB id 1l2y). On CaM, we analyze the ability to connect all 6 directed pairs that can be defined over its three functional states. These states are documented under PDB ids 1cfd (apo), 1cll (holo), and 2f3y (collapsed). CaM is an ideal system to study, as it is a key signaling protein in many cellular processes exhibiting a particularly large conformational rearrangement. On AdK, a variety of states have been reported, but we focus here on connecting the apo (PDB id 4ake) to the closed state (PDB id 1ake).

B. Summary Results on Connecting Start to Goal

Table I summarizes performance in terms of the lowest lRMSD obtained to the goal from a tree rooted at a given start structure. Results are averaged over 10 executions, and Table I shows averages (μ) and standard deviations (σ) across bias schemes. `QUAD`, `LINEAR`, `UNIFORM`, and $\texttt{COMBINE}_{90-10}$ described in section II are combined with the local bias in expansion in the results reported in columns $3-6$ in Table I. The local bias is removed in the $\texttt{COMBINE}^{*}{}_{90-10}$ scheme reported in column 7.

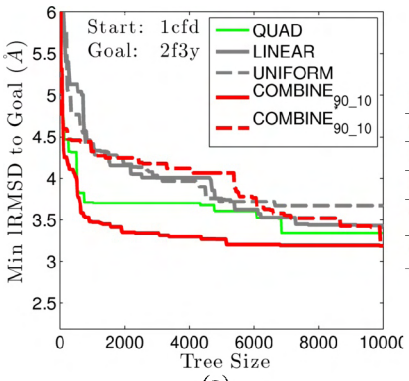
Table I shows that the method effectively approaches the goal within the $3-4$Å tolerance. On Trp-Cage, the average lowest lRMSDs are 1.8-2Å, which is comparable with the 1.5-2.5Å range reported by folding and structure prediction studies [36], [48]. We note that the test on Trp-cage is intended to show the capability of the method on a small system. The paths are not to be interpreted as folding paths, as the fragment configurations are extracted from native protein structures. On CaM, the average lowest lRMSDs range from sub-angstrom to 4Å depending on the pair connected. Some pairs seem more difficult than others. On the 1cfd to 1cll paths, the average lowest lRMSDs are below 4Å, which is in general agreement with the $1.5-5$Å proximity reported by MD- and MC-based biophysical studies [12], [49]. AdK represents a more challenging case for method. Lowest lRMSDs obtained here are 3-5Å , slightly higher than the 2.5Å obtained with very coarse-grained models [39].

Table I: Average (μ) and standard deviations (σ) are reported for the lowest tree lRMSD over 10 executions of the method.

System	Start → Goal	$\mu \pm \sigma$ over lowest lRMSDs (Å)				
		QUAD	LINEAR	UNIFORM	COMBINE$_{90-10}$	COMBINE*$_{90-10}$
Trp-Cage (20 aa)	E → 1l2y (18 Å)	1.86 ± 0.53	2.12 ± 0.57	1.83 ± 0.63	1.87 ± 0.43	2.42 ± 0.43
CaM (140 aa)	1cfd → 1cll (10.7 Å)	3.17 ± 0.25	3.27 ± 0.10	3.49 ± 0.26	3.32 ± 0.12	3.36 ± 0.13
	1cll → 1cfd (10.7 Å)	3.35 ± 0.51	3.56 ± 0.29	3.70 ± 0.23	3.50 ± 0.21	3.49 ± 0.24
	1cfd → 2f3y (9.9 Å)	3.93 ± 0.37	3.93 ± 0.42	3.99 ± 0.24	3.76 ± 0.41	4.01 ± 0.34
	2f3y → 1cfd (9.9 Å)	3.43 ± 0.39	3.65 ± 0.45	3.62 ± 0.13	3.34 ± 0.13	3.57 ± 0.28
	1cll → 2f3y (13.44 Å)	1.91 ± 0.58	1.67 ± 0.49	2.01 ± 0.86	1.68 ± 0.37	1.50 ± 0.2
	2f3y → 1cll (13.44 Å)	0.82 ± 0.30	0.72 ± 0.08	1.02 ± 0.43	0.73 ± 0.10	0.94 ± 0.14
AdK (214 aa)	1ake → 4ake (6.95 Å)	3.91 ± 0.34	4.28 ± 0.36	5.15 ± 0.30	4.19 ± 0.22	4.32 ± 0.41
	4ake → 1ake (6.95 Å)	4.65 ± 0.71	5.32 ± 0.79	5.62 ± 0.37	5.21 ± 0.41	5.09 ± 0.69

C. Comparison of Bias Schemes

Results in Table I suggest that all bias schemes allow approaching the goal. Here we take a closer look at how these schemes lower lRMSD to goal over time. We limit the analysis to CaM and the "best" execution (over 10) that allows a bias scheme to achieve its lowest lRMSD. Fig. 1(a) highlights the expected behavior, showing that QUAD can drive the exploration rapidly towards the goal but may plateau for long periods of time. LINEAR and UNIFORM show similar rate of descent, suggesting that LINEAR has a weak enough bias to allow a broader exploration. COMBINE$_{90-10}$ seems to temper the bias in QUAD through its randomization, plateauing later when fragment replacements do not allow further approaching the goal. COMBINE*$_{90-10}$, which removes the local bias in expansion, shows similar behavior to LINEAR and UNIFORM.



(a)

Bias Scheme	Breadth (Å)
QUAD	29.62
LINEAR	45.73
UNIFORM	53.28
COMBINE$_{90\text{-}10}$	33.17
COMBINE*$_{90\text{-}10}$	40.25

(b)

Figure 1: (a) Minimum lRMSDs to goal are plotted as a function of tree size and compared among bias schemes. (b) Bias schemes are compared in terms of path diversity.

In Fig. 1(b) we analyze path diversity. On the same 1cfd to 2f3y case studied in terms of depth in Fig. 1(a), all paths on the best execution of a bias scheme are collected. A path is defined as a series of conformations that reach the goal with no higher than 4Å here. An estimate of breadth over paths is defined as $b = (\sum_{i=0}^{h} (i+1) \cdot d_i)/h$, where h is the number of nodes on the shortest path, and d_i is the maximum pairwise lRMSD among conformations at level i across all paths (i grows from goal to root). This measure downweights differences in lower levels (closer to goal). Fig. 1(b) shows this estimate across all bias schemes and confirms that diversity is lowest in QUAD and highest in UNIFORM. Taken together, this suggests that COMBINE*$_{90-10}$ balances best between breadth and depth during tree growth.

D. Detailed Analysis

On CaM, the method is able to surpass initial lRMSDs >13.44Å. Sub-angstrom lRMSDs are obtained when the method is setup to approach 1cll from 2f3y; 1−2Å are obtained in the other direction. Connecting the other 4 directed pairs is more difficult; lowest lRMSDs across all bias schemes are 3−4Å. The employment of USR-based discretization (described in section II) to bias away from similar paths seems to improve lowest lRMSDs by 1−2Å in these difficult cases; lowest lRMSDs of 2.0−2.4Å are obtained on connecting 1cfd to 1cll and 1cfd to 2f3y and vice-versa (COMBINE*$_{90-10}$ is used).

Results on CaM are in qualitative agreement with those observed in experiment and simulation [12], [50], [51]. The transition between 1cll and 2f3y is easier than between the other pairs. Though the other pairs have initial lRMSDs that are lower than that between 1cll and 2f3y, the true distance that has to be surpassed is in angular space, which partially explains why the method performs well. Due to its use of molecular fragment replacement, the method is particularly suitable to obtain paths of "angular" rearrangements. Some paths highlighting these rearrangements are shown in Fig. 3.

We note that the use of fragment configurations is justified when functional transitions do not involve unfolding. This is true of many proteins, including CaM and AdK. In particular, wet-lab studies on CaM wildtypes and mutants exclude the possibility that the transition involves a significant population of unfolded or disordered states [51]. These studies also suggest that the transition between 1cfd and 1cll is a complex process with energy barriers rather than a single global transition between two substates. A pseudo-free energy landscape produced by our method is shown in

Fig. 2. All paths from 10 runs obtained with $\texttt{COMBINE}^{*}_{90-10}$ on connecting 1cfd to 1cll and vice-versa are combined. Pseudo-free energies are calculated along the ΔR coordinate (defined as lRMSD$(C, C_{1\text{cfd}})$ - lRMSD$(C, C_{1\text{cll}})$) through the weighted histogram analysis method [52]. The pseudo-free energy landscape in Fig. 2 shows that paths have to cross regions of high free energy, which qualitatively agrees with wet-lab findings in [51]. The shown pseudo-free energies need to be taken with caution. Pseudo-free energy values are affected by potential lack of sampling density and path diversity, issues we discuss in section IV.

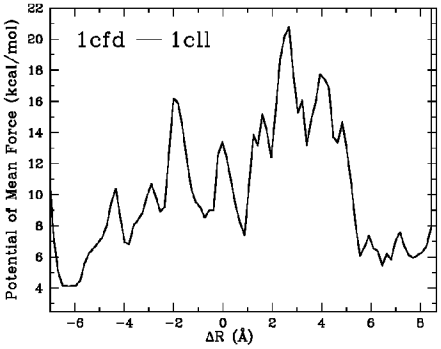


Figure 2: Pseudo-free energies along ΔR are shown for sampled paths connecting 1cfd to 1cll and vice versa.

AdK presents an extremely challenging case for our method, not only due to its size but also due to the presence of a significant energy barrier in the transition [9]. Table I shows that lowest lRMSDs can be above 4Å. Lack of density in sampling makes a pseudo-free energy analysis here premature. Our future work will consider reactive temperature schemes to enhance sampling in AdK, but here we report some interesting results on the presence of a known intermediate structure (PDB id 2rh5) in the paths computed by the method. Table II shows the lowest lRMSD over paths connecting 1ake to 4ake and vice versa to the intermediate structure. These results show that the method is able to capture this intermediate structure, in agreement with findings in other studies [39], [53].

Table II: Lowest lRMSD to 2rh5 is shown.

	lowest lRMSD to 2rh5 (Å)
1ake → 4ake	3.62
4ake → 1ake	1.97

IV. Conclusion

We have presented a robotics-inspired tree-based method to compute paths connecting functional states of a protein. Molecular fragment replacement is employed to grow the tree with physically-realistic conformations. As a result, computed paths are mainly applicable to connecting states that do not involve unfolding. External factors in conformational switching (e.g., binding partners) are not considered under the general assumption that diverse functional states co-exist at equilibrium albeit with different probabilities.

The approach proposed in this paper is a proof of concept that is demonstrated to compute credible conformational paths connecting structural states in small- to medium-size proteins. To the best of our knowledge, this is the first employment of an EST-based method and its combination with the molecular fragment replacement technique. Several directions of research will be pursued. In the long run, MD-based techniques will allow deforming paths into transition trajectories and obtain timescale information. This will necessitate limiting the distance between parent and child conformations, which we will investigate with techniques we have studied in other contexts [54]. Additional path smoothing techniques will be investigated. Immediate lines of research include the pursuit of general reactive schemes for the global bias and temperature.

More permissive temperatures may enhance sampling for more complex systems, such as AdK. A reactive scheme for temperature updates will be pursued in future work (based on work in [38] for dialanine peptide). General reactive schemes will also be employed to adaptively balance the exploration and switch between different biasing schemes.

While we use lRMSD as a progress coordinate here, others based on contact topology and structural profiles will be investigated based on our previous work in ab-initio structure prediction [40]. Additional layers (e.g. based on USR coordinates) will be considered to balance the tree growth between coverage (path diversity) and progress (proximity to goal) and so weaken path correlations. Sampling multiple structurally diverse paths with statistical rigor is an important issue to address in future work. Due to its incorporation of discretization layers, the proposed method provides a natural bed on which to investigate balancing progress and coverage. Obtaining a broad view of possibly diverse paths connecting two states allows identification of highly-populated intermediates and rigorous calculation of free energy barriers for direct quantitative validation with experimental and computational biophysical studies.

Acknowledgment

This work is supported in part by NSF CCF No. 1016995 and NSF IIS CAREER Award No. 1144106.

References

[1] Y. J. Huang and G. T. Montellione, "Structural biology: Proteins flex to function," *Nature*, vol. 438, no. 7064, pp. 36–37, 2005.

[2] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[3] P. Majek, H. Weinstein, and R. Elber, *Pathways of conformational transitions in proteins*. Taylor and Francis group, 2008, ch. 13, pp. 185–203.

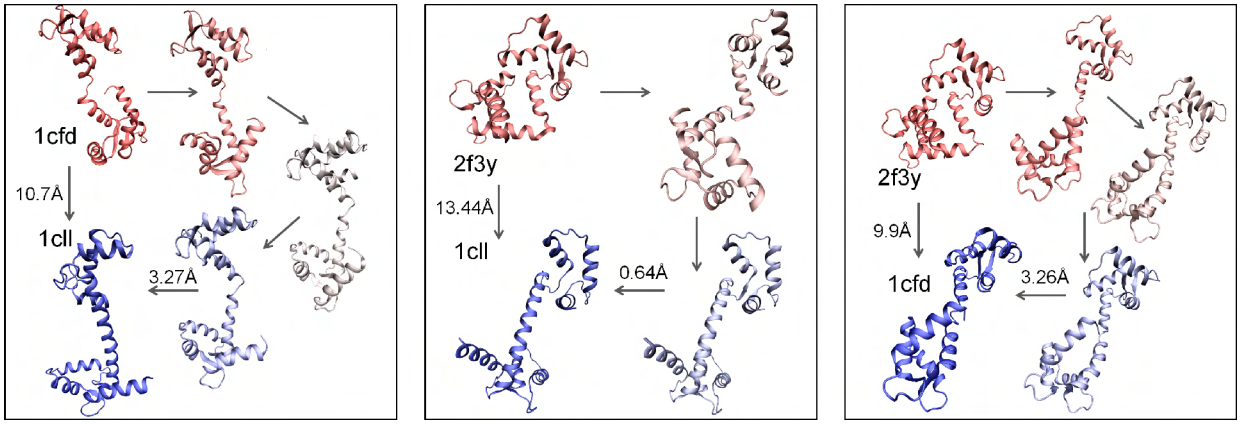


Figure 3: Three paths are highlighted. Start and goal structures are in red and blue, respectively. Selected conformations in the path are drawn in a red-to-blue interpolated scheme.

[4] 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.

[5] H. Huang, E. Ozkirimli, and C. B. Post, "A comparison of three perturbation molecular dynamics methods for modeling conformational transitions," *J. Chem. Theory Comput.*, vol. 5, no. 5, pp. 1301–1314, 2009.

[6] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling," *J. Comput. Phys.*, vol. 23, no. 2, pp. 187–199, 1977.

[7] R. Malek and N. Mousseau, "Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique," *Phys. Rev. E*, vol. 62, no. 6, pp. 7723–7728, 2000.

[8] D. J. Earl and M. W. Deem, "Parallel tempering: Theory, applications, and new perspectives," *Phys. Chem. Chem. Phys.*, vol. 7, pp. 3910–3916, 2005.

[9] K. Arora and C. L. I. Brooks, "Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 47, pp. 18 496–18 501, 2007.

[10] Y. Zhang, D. Kihara, and J. Skolnick, "Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding," *Proteins: Struct. Funct. Bioinf.*, vol. 48, no. 2, pp. 192–201, 2002.

[11] B. G. Schulze, H. Grubmueller, and J. D. Evanseck, "Functional significance of hierarchical tiers in carbonmonoxy myoglobin: conformational substates and transitions studied by conformational flooding simulations," *J. Am. Chem. Soc.*, vol. 122, no. 36, pp. 8700–8711, 2000.

[12] B. W. Zhang, D. Jasnow, and D. M. Zuckermann, "Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin," *Proc. Natl. Acad. Sci. USA*, vol. 104, no. 46, pp. 18 043–18 048, 2007.

[13] 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.

[14] A. d. Schuyler, R. L. Jernigan, P. K. Wasba, B. Ramakrishnan, and G. S. Chirikjian, "Iterative cluster-nma (icnma): A tool for generating conformational transitions in proteins," *Proteins: Struct. Funct. Bioinf.*, vol. 74, no. 3, pp. 760–776, 2009.

[15] W. Zheng and B. Brooks, "Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model," *J. Mol. Biol.*, vol. 346, no. 3, pp. 745–759, 2005.

[16] R. Bahar and A. J. Rader, "Coarse-grained normal mode analysis in structural biology," *Curr. Opinion Struct. Biol.*, vol. 204, no. 5, pp. 1–7, 2005.

[17] N. Kantarci-Carsibasi, T. Haliloglu, and P. Doruker, "Conformational transition pathways explored by monte carlo simulation integrated with collective modes," *Biophys. J.*, vol. 95, no. 12, pp. 5862–5873, 2008.

[18] A. Korkut and W. A. Hendrickson, "Computation of conformational transtions in proteins by virtual atom molecular mechanics as validated in application to adenylate kinase," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 37, pp. 15 673–15 678, 2009.

[19] M. Teknipar and W. Zheng, "Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model," *Proteins: Struct. Funct. Bioinf.*, vol. 78, no. 11, pp. 2469–2481, 2010.

[20] S. Kirillova, J. Cortes, A. Stefaniu, and T. Simeon, "An nma-guided path planning approach for computing large-amplitude conformational changes in proteins," *Proteins: Struct. Funct. Bioinf.*, vol. 70, no. 1, pp. 131–143, 2008.

[21] J. W. Chu, B. L. Trout, and C. L. I. Brooks, "A super-linear minimization scheme for the nudged elastic band method," *J. Chem. Phys.*, vol. 119, pp. 12 708–12 717, 2003.

[22] E. Weinan, W. Ren, and E. Vanden-Eijnden, "Simplified and improved string method for computing the minimum energy paths in barrier-crossing events," *J. Chem. Phys.*, vol. 126, p. 164103, 2007.

[23] L. Maragliano and E. Vanden-Eijnden, "On-the-fly string method for minimum free energy paths calculation," *Chem.*

Phys. Lett., vol. 446, pp. 182–190, 2007.

[24] E. Weinan, W. Ren, and E. Vanden-Eijnden, "Finite temperature string methods for the study of rare events," *J. Phys. Chem.*, vol. 109, pp. 6688–6693, 2005.

[25] W. Ren, E. Vanden-Eijnden, P. Maragakis, and E. Weinan, "Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide," *J. Chem. Phys.*, vol. 123, p. 134109, 2005.

[26] D. R. Weiss and M. Levitt, "Can morphing methods predict intermediate structures?" *J. Mol. Biol.*, vol. 385, no. 2, pp. 665–674, 2009.

[27] L. E. Kavraki, P. Svetska, J.-C. Latombe, and M. Overmars, "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Trans. Robot. Autom.*, vol. 12, no. 4, pp. 566–580, 1996.

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

[29] G. Song and N. M. Amato, "A motion planning approach to folding: From paper craft to protein folding," *IEEE Trans. Robot. Autom.*, vol. 20, no. 1, pp. 60–71, 2004.

[30] T. H. Chiang, M. S. Apaydin, D. L. Brutlag, D. Hsu, and J.-C. Latombe, "Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: folding rates and phi-values," *J. Comp. Biol.*, vol. 14, no. 5, pp. 578–593, 2007.

[31] S. M. LaValle and J. J. Kuffner, "Randomized kinodynamic planning," *Int. J. Robot. Res.*, vol. 20, no. 5, pp. 378–400, 2001.

[32] D. Hsu, R. Kindel, J.-C. Latombe, and S. Rock, "Randomized kinodynamic motion planning with moving obstacles," *Int. J. Robot. Res.*, vol. 21, no. 3, pp. 233–255, 2002.

[33] 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.

[34] 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.

[35] A. Shehu, "An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations," in *Robot: Sci. and Sys.*, Seattle, WA, USA, 2009, pp. 241–248.

[36] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–11 227, 2010.

[37] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *J. Bioinf. and Comp. Biol.*, vol. 9, no. 3, pp. 383–398, 2011.

[38] L. Jaillet, F. J. Corcho, J.-J. Perez, and J. Cortes, "Randomized tree construction algorithm to explore energy landscapes," *J. Comput. Chem.*, vol. 32, no. 16, pp. 3464–3474, 2011.

[39] N. Haspel, M. Moll, M. L. Baker, W. Chiu, and K. L. E., "Tracing conformational changes in proteins," *BMC Struct. Biol.*, vol. 10, no. Suppl1, p. S1, 2010.

[40] B. S. Olson, K. Molloy, S.-F. Hendi, and A. Shehu, "Guiding search in the protein conformational space with structural profiles," *J. Bioinf. and Comput. Biol.*, vol. 10, no. 3, p. 1242005, 2012.

[41] 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.

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

[43] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Funct. Bioinf.*, vol. 76, no. 4, pp. 837–851, 2009.

[44] R. Bonneau and D. Baker, "De novo prediction of three-dimensional structures for major protein families," *J. Mol. Biol.*, vol. 322, no. 1, pp. 65–78, 2002.

[45] 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.

[46] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins: Struct. Funct. Bioinf.*, vol. 73, no. 4, pp. 958–972, 2009.

[47] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 10, pp. 3734–3739, 2009.

[48] 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.

[49] E. Project, R. Friedman, E. Nachliel, and M. Gutman, "A molecular dynamics study of the effect of Ca^{2+} removal on calmodulin structure," *Biophys. J.*, vol. 90, no. 11, pp. 3842–3850, 2006.

[50] B. E. Finn, J. Evenäs, T. Drakenberg, J. P. Waltho, E. Thulin, and S. Forsén, "Calcium-induced structural changes and domain autonomy in calmodulin," *Nat. Struct. Biol.*, vol. 2, no. 9, pp. 777–783, 1995.

[51] J. Evenäs, S. Forsén, A. Malmendal, and M. Akke, "Backbone dynamics and energetics of a calmodulin domain mutant exchanging between closed and open conformations," *J. Mol. Biol.*, vol. 289, no. 3, pp. 603–617, 1999.

[52] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules: I. The method," *J. Comput. Chem.*, vol. 13, no. 8, pp. 1011–1021, 1993.

[53] K. P. Ravindranathan, E. Gallicchio, and R. M. Levy, "Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein," *J. Mol. Biol.*, vol. 353, no. 1, pp. 196–210, 2005.

[54] B. Olson and A. Shehu, "Efficient basin hopping in the protein energy surface," in *IEEE BIBM*, 2012, in press.