A Probabilistic Roadmap-based Method to Model Conformational Switching of a Protein Among Many Functionally-relevant Structures

Kevin Molloy[1] Amarda Shehu[1,2,3*]
[1]Dept. of Computer Science, [2]Dept. of Bioengineering, [3]School of Systems Biology
George Mason University, Fairfax, VA 22030
*amarda@gmu.edu

Abstract

Obtaining a detailed microscopic view of protein transitions among key structural states is central to obtaining a deeper understanding of the relationship between protein dynamics and function. Doing so in the wet laboratory is currently not possible. It is also infeasible to model conformational switching through computational treatments based on Molecular Dynamics, particularly when the objective is expanded to model switching of medium-sized proteins among an arbitrary number of given states. In this paper, we consider this expanded objective and propose a novel probabilistic method to sample conformational paths connecting functionally-relevant structures of a protein. The method achieves this without launching expensive simulations but instead by mapping the connectivity of the conformational space around given thermodynamically-stable and semi-stable structural states. This is achieved through an adaptation of the probabilistic roadmap framework that has been shown successful at planning motions of articulated mechanisms in robotics. Preliminary analysis shows the method is promising and efficient in modeling motions among various states for medium-size proteins.

1 Introduction

Function modulation requires proteins to access and switch between different thermodynamically-stable and semi-stable structures. Sometimes, these structures are very different, by more than 10Å, and switching between them occurs on very long micro-μ second timescales. Yet, elucidating the conformational pathways that a protein uses to switch between potentially many structurally-diverse functionally-relevant states is important to obtain a deeper understanding and characterization of proteins as dynamic molecular machines. Doing so is also a first step towards addressing further interesting questions on how sequence mutations potentially impact conformational switching and thus function in proteinopathies [13].

Conformational switching remains hard to capture in the wet laboratory. Elucidating conformational pathways is also non-trivial in silico. In particular, modeling system dynamics through protocols based on Molecular Dynamics demands significant computational time to witness the system transition from a given start to a given goal structure. This is exacerbated when the computational objective is expanded to mapping the possible motions that a system uses to switch among many different structural states.

In this paper we propose a robotics-inspired computational treatment of the conformational switching problem. In particular, the objective here is to map the connectivity of the protein structure space relevant for function. The setup we consider in this paper is as follows. A certain number of functionally-relevant structures, possibly elucidated in the wet-laboratory or obtained in silico (or through a combination of both) are provided as *landmarks* in the protein structure space. The objective of the method we propose here is to enrich the structure space with more structural detail biased towards providing paths, series of conformations/structures, between any two of the given structural states.

The method we propose in this paper is based on the probabilistic roadmap method (PRM) that is used in the robot motion planning community to compute feasible, collision-free, paths that allow a robot to go from a given start to a given goal state [8]. Adaptations of PRM have been debuted in the computational structural biology community to model unfolding trajectories of small proteins ($<$ 80 amino acids long), where the goal is to see protein systems transition from a given folded to some unfolded state [16–18]. Recent work building on robotics treatments has addressed the conformational switching problem, but it has been limited to providing paths connecting two given structural states. These methods have typically been adaptations of tree-based (as opposed to graph/roadmap-based) robot motion planning methods [1, 5, 6, 9, 10]. There are intrinsic differences among tree-based methods, including important decisions on how to improve exploration capability through a combination of low-level decisions on which degrees of freedom to modu-

late and higher-level ones on how to balance the tree search between progress towards the goal state and coverage of the conformational space. These methods have been predominantly limited to the binary setting, where the goal is to sample paths connecting two given configurations. Work in [7] has expanded this setting to mapping trajectories among many (>2) different states, but the work is limited to very small peptides of no more than 2 amino acids.

In this work we address medium-size proteins of length between 144 and 214 amino acids. At a lower level, important decisions are made on which degrees of freedom to modify and when in order to handle the high dimensionality of the conformational space in these systems. At a higher level, the method is an adaptation of the PRM framework (hence, roadmap-based) so that it can obtain a broad view of the connectivity of the structure space around many given landmark structures and compute paths in a computationally viable manner. Other adaptations revolve around making sure that the conformational paths have credible energetic profiles, the sampling is dense in regions of the space that matter, the connections between nearby conformations are credible, and local energetic barriers are crossed as needed to improve connectivity and path diversity. We report here promising preliminary results, which motivate us to pursue further enhancements of this method to obtain a detailed and comprehensive picture of the thermodynamically-stable and semi-stable structure space (and its connectivity) available to a protein for biological function.

2 Methods

We first describe the setup of the problem considered in this paper and then proceed to relate details on the proposed method.

2.1 Problem Statement

The setup we consider is as follows: We are given l structures (each in a PDB file format) of a protein system, which can be obtained in the wet or dry laboratory. Our assumption is that the input structures are representative of diverse functional states of a given protein sequence, and hence are thermodynamically-stable or semi-stable; they have low energies. We refer to these given structures as *landmarks*. In addition to the landmarks, the method is provided with a-priori constructed libraries of protein fragment configurations (their purpose is detailed below in the description of the method). The method produces a graph or roadmap, whose vertices consist of sampled conformations and the given landmarks. The purpose of the graph is to obtain connectivity and thus reveal, through graph-analysis techniques, conformational pathways connecting any two given landmarks.

The method is comprised of three main stages: exploration — to obtain novel conformations beyond the provided landmarks, roadmap construction — to connect pairs of conformations, and graph analysis — to extract paths connecting the given landmarks. Our description of the method proceeds along these stages.

2.2 Exploration

The objective in this stage is to generate an ensemble of low-energy conformations of the given protein sequence. This ensemble will populate the vertex set V of the roadmap $G = (V, E)$ that we want to construct. V is first initialized with the given landmark structures. This stage in the proposed method is equivalent to the *learning phase* in the PRM framework in robotics, where robot configurations are sampled, discarding those that violate constraints (for instance, due to collisions with obstacles). In our setting, obstacles take the form of high-energy regions of the protein conformational space. While conformational switching in a protein is a stochastic process, paths that allow switching between states in shorter timescales avoid going over high-energy barriers. This justifies our setup and objective in the exploration stage of sampling low-energy protein conformations.

The main challenge in the exploration stage lies in the sampling technique used to increase the likelihood of sampling conformations that satisfy the constraints. In robotics, various biased sampling techniques exist that make use of the location of obstacles [4]. These are not readily portable in our setting. On the other hand, uniform random sampling of the degrees of freedom — dofs (the parameters that we encode to represent a protein conformation) is rather inefficient, as the probability to obtain a high-energy conformation (for instance, due to self collisions) increases with the length of the protein chain.

For this reason, the exploration stage makes use of four key strategies. First, the only dofs considered are the backbone dihedral angles (for a review of protein representations and their implication for sampling, the reader is referred to Ref. [13]). Second, the dimensionality of the conformational space is reduced by bundling together consecutive dofs into fragments and sampling values for them simultaneously. Values are obtained from fragment configuration libraries constructed over known native structures of proteins,

which increases the likelihood that sampled conformations have low energies. Third, new *sampled* conformations are obtained by modifying fragment configurations of conformations selected from the existing vertex set. Fourth, each sampled conformation is evaluated according to the Metropolis criterion, using a state-of-the-art energy function, to ensure that the conformation does not raise energy by more than what is allowed at a given temperature.

In order to improve coverage of the conformational space, we make use of a progress coordinate. We use ΔR, defined in literature as $\text{LRMSD}(\text{start}, \text{C}_\text{i}) - \text{LRMSD}(\text{goal}, \text{C}_\text{i})$ to provide a coordinate to a conformation C_i in the segment from some given start to some given goal conformation. In our setting, we have $k > 2$ landmark (start and goal) conformations. Therefore, we associate $\binom{k}{2}$ coordinates with each sampled conformation, each coordinate tracking the position of a conformation between any 2 of the given landmarks. In order to improve coverage, we maintain $\binom{k}{2}$ 1d grids, each grid covering the range $\pm(\text{LRMSD}(\text{start}, \text{goal}) + 1)$ and having cells 1Å wide. Each conformation is mapped to each of the grids.

Each grid cell is assigned a weight w, defined as $w_c = \frac{1}{(1+nsels)*nconfs}$, where $nsels$ is the number of times cell c has been selected for expansion, and $nconfs$ is the number of conformations mapped to cell c (this weight has been introduced by us in tree-based search for structure prediction in [15]). A probability distribution function can be calculated using the sum of these weights as a normalization constant. This function is then used to bias the selection towards conformations falling in grid cells that are either under-populated or have not been selected many times before. This promotes uniform coverage of the space as defined by the progress coordinate. The selection mechanism is as follows. First, a pair of landmarks are chosen uniformly at random. A cell in the grid tied to the selected pair is then selected according the probability distribution function defined above. Once a cell is selected, any of the vertices in the current roadmap that fall in that cell are selected uniformly a random to obtain a new conformation.

Let us refer to the selected vertex/conformation as C. The procedure we use to sample a new conformation C_new is similar to that used in the Rosetta structure prediction protocol and our own employment of fragment-based conformational sampling for structure prediction and conformational path sampling. Due to limited space, we are not detailing the procedure here (details can be found in [11]). In summary, a fragment library is used to propose that a series of ϕ, ψ angles in C, a fragment, be changed to values of a configuration stored in the library. We use a fragment library of length 9 amino acids in the exploration stage of the method. This allows for rapidly sampling of low-energy conformations. While using fragment libraries increases the likelihood that the sampled C_new will have low energy, there may exist unfavorable long-range interactions. Therefore, the Metropolis criterion is used on the change of energy $\delta E = E(C_\text{new}) - E(C)$, where E is measured on a conformation with the Rosetta coarse-grained `score3` energy function (after removing from this function the compaction term that penalizes non-compact conformations). The Metropolis criterion is probabilistic, $exp(-\delta E/T)$, where T is an effective temperature that determines how much of an energetic increase is allowed over the energy of C in C_new. For a selected C, we make ≤ 20 attempts at obtaining a C_new that satisfies the Metropolis criterion. If the update fails, a new vertex is selected, and the so-described update process begins anew.

Energetic barriers may exist between landmark pairs. Given that each landmark is a stable or semi-stable structure, it may reside in a local minimum of the energy surface. For this reason, we employ a reactive temperature scheme that increases T as needed to cross energy barriers. We associate temperatures T with each grid cell. Initially, all temperatures are set to values that allow accepting a 10kcal/mol energetic increase in the update step with a 0.17 probability. A cell's temperature is adjusted every 25 times the cell is selected. If the last 25 selections did not result in a conformation being accepted (the Metropolis criterion failed), the temperature of the cell is increased. If a new conformation meeting the Metropolis criterion was obtained at least 60% of the time a given cell was selected, the temperature of the cell is decreased. The scheme employed for temperature increases and decreases is based on a simulated annealing proportional cooling schedule. Temperature is increased or decreased by setting it to the previous or next temperature in the schedule. This schedule is detailed and employed in various of our works on structure prediction and conformational sampling [10, 14].

2.3 Roadmap Construction

Now that the vertex set is populated, the objective shifts to providing connectivity and populating the edge set E of the roadmap G. We put an edge between two conformations whose LRMSD is less than some threshold ϵ (set at 1.5Å in our initial implementation here). The vertices in the graph are analyzed, and edges are added for those that meet the criterion. A more involved procedure is then followed to further improve the connectivity of the graph, even if additional conformations need to be sampled to do so.

Local Planner: For each conformation C_i created in the exploration stage, we identify its n nearest neighbors, employing LRMSD as the distance metric (we use n=5 here). For each of the n neighbors C_{ij} (where $1 \leq j \leq n$), we attempt to create a new conformation C_{new} with the following attributes. First, the new conformation must be within ϵ of C_i so we can put an edge between C_i and C_{new}. We also want to maximize the distance from C_i and minimize the distance to C_{ij}. For this reason, 50 candidate C_{new} conformations are sampled (all by updating C_i and passing the Metropolis criterion), selecting the one that optimizes the two distance criteria. This results in a new vertex, C_{new}, and edge $\langle C_i, C_{\text{new}} \rangle$ being added to the graph. The distance from C_{new} to other vertices in the graph are also checked, and an edge is added from C_{new} to any vertex that is within ϵ. If C_{ij} is also within ϵ of C_{new}, we have finally connected C_i to its nearest neighbor C_{ij} through an intermediate conformation C_{new} and can move on to repeating this procedure for C_i and its next nearest neighbor. Else, if no direct edge can be placed between C_{ij} and C_{new}, we repeat the procedure one more time, attempting to extend from C_{ij} to C_{new} with another intermediate conformation.

The sampled conformations in the local planner are obtained using configuration libraries of fragments of length 3 (shorter length means that the modification to a conformation will have a smaller magnitude). Similar to the sampling procedure in the exploration stage, each new conformation is subjected to the same Metropolis criterion, using the same energy function.

Merging of Connected Components: Since the dimensionality of the conformational space is high and the energetic constraints are non-trivial, there is no guarantee that the roadmap will be connected. We now focus on individual connected components (CC). We first relax ϵ to 3 Å and add new edges to G based on this relaxed criterion. This helps identify the most challenging regions that are further than 3Å from any others. We attempt to merge CCs as follows. For each pair of CCs, k pairs of closest conformations (one on each CC) are identified. The local planner described above is then applied on each pair, but the number of candidate C_{new} conformations is increased to 500 in order to increase the likelihood that we can connect the chosen CCs. The sampled conformations and the edges meeting the relaxed 3.0 Å threshold in the local planner are added to the roadmap.

2.4 Path Analysis

Among other internal graph-theoretic analyses that can be conducted on the roadmap, the main interesting one in the context of our application is the evaluation of the presence or not of conformational paths connecting pairs of the provided landmarks and the energetic quality of these path. Path analysis is non-trivial, unless we limit it to the computation of the shortest path. As such, a first set of results we provide below focuses on the shortest path between a pair of landmarks. We associate an overall weight with a path that is the sum of the negative logarithms of probabilities associated with edges in the roadmap (these probabilities make use of $exp(-\delta E)$ and are defined as in the application of a PRM-based framework for protein-ligand binding and protein unfolding in [2]). This allows us to use Dijkstra's algorithm. Provided that there may be more than a path connecting two landmarks in the roadmap, techniques to extract multiple paths, such as k-shortest paths [12], are useful for our analysis in this paper.

3 Results

Implementation Details and Experimental Setup Experiments are conducted on calmodulin (CaM) and adenylate kinase (AdK) of respective lengths 144 and 214 amino acids. On CaM, we are provided with 3 landmark X-ray structures documented under PDB ids 1cfd (apo), 1cll (holo), and 2f3y (collapsed). CaM is a key signaling protein involved in many cellular processes that undergoes large conformational rearrangements. On AdK, we initialize the roadmap with 2 landmark X-ray structures, the apo (PDB id 4ake) and the closed states (PDB id 1ake). AdK is another important protein catalyzing the interconversion of adenine nucleotides, and thus playing a critical role in cellular energy homeostasis. On each protein, the exploration stage of the method is run until $2,500$ vertices are added to the roadmap. In total, the runtime is 24−36 hours on one CPU. All our experiments are conducted on 2.66 GHz Opteron processor with 24 GB of memory.

3.1 Comparing to a Tree-based Search

The roadmap-based method we propose here provides more connectivity than a tree-based method and can answer multiple queries. In our prior work on a tree-based method for conformational switching, we grew a tree to find paths from a given start structure to a given start structure [10]. While the roadmap-based method we provide here has a more general setup (the number of landmark structures can exceed 2), we demonstrate that even in the strict context of computing conformational paths between two given structures, the roadmap-based method provides more information. We highlight this on the AdK system.

Table 1: Comparing tree-based method in [10] to roadmap-based method proposed here.

Adk: 4ake → 1ake	Tree (EST) [10]	This Work
Nr. Samples	10,000	11,200
Run Time	8 hrs.	24 hrs
Nr. Paths	> 200	> 200
Max Step (Å)	3.9	2.9

Table 1 compares the methods side by side in terms of overall statistics, such as number of conformations in the search structure (whether that is a tree or a graph/roadmap), time, number of paths, and path quality. We note that the amount of time spent (row 3) by the roadmap-based method is higher, due to mainly the computation of nearest neighbors in the roadmap construction stage. The number of paths found by the roadmap-based method is higher, as expected (see row 4), as a roadmap provides a more global view than the local view afforded in tree-based methods. In addition, the path quality is better (row 5). The maximum LRMSD between two consecutive conformations in a path is almost 1Å lower in the roadmap-based method than in the tree-based method in our prior work.

3.2 Analysis on AdK

Using the path analysis described in Methods, we were able extract hundreds of distinct pathways (with no loops) connecting the two given landmark structures of AdK (1AKE and 4AKE) in each direction. We show two such representative paths here, one from 1AKE to 4AKE and another one from 4AKE to 1AKE. The paths are shown on a 2D embedding of the conformational space sampled by the method for AdK. The two projection coordinates used for the embedding are the potential energy of each conformation and the conformation's ΔR coordinate. The embedding is discretized into small cells so that a density of state can be associated with each cell (to show the number of conformations per cell/state). The density of state is used to color-code the embedding, with darker shades showing where the method focused its sampling.

Fig. 1 shows that both paths go through high-energy regions, which is qualitatively in agreement with previous work by us and others [3, 10]. For reference, additional known structures of AdK are shown on the embedding, with ΔR values calculated from their PDB atomic coordinates (ids are shown in the legend). We note that the potential energies of these structures are higher than the landmarks, as expected (no refinement of these structures is carried out prior to this analysis). These structures are known intermediates in the conformational switching between the closed and open structures of AdK. The embedding in Fig 1 shows that the intermediate structures span the sampled conformational space and the conformational switching captured by our method for AdK.

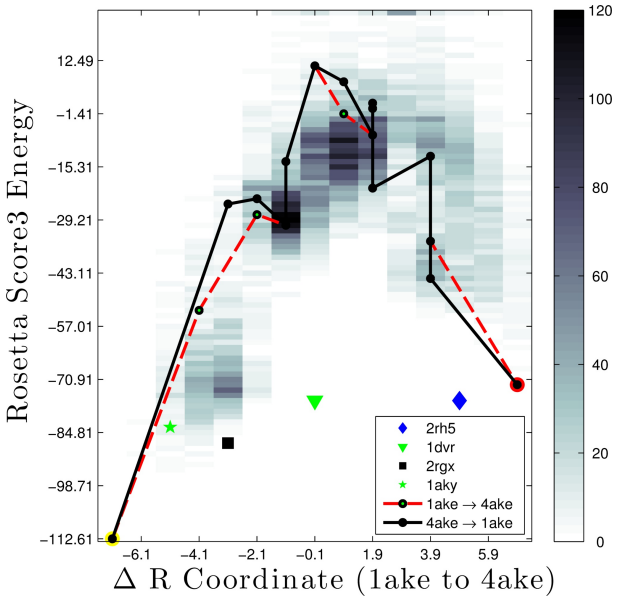

Figure 1: Two paths, one in each direction, are shown over a 2D embedding of the sampled conformational space. Known intermediate structures are indicated. Cells are color-coded by density of state.

3.3 Analysis on CaM

We apply our method to the 3 landmark states of CaM. When incorporating multiple landmarks, it becomes increasing difficult to establish good connectivity. Figure 2 illustrates this challenge during the merging of CCs. Initially the method rapidly reducing the number of CCs, but this becomes increasing difficult. Improving this portion of the local planner is a clear direction for future work, and may warrant incorporating tree-based local planners based on our previous work [10].

4 Summary

This paper has presented a novel method for discovering conformational pathways between multiple functional states in a protein system. The adaptation of the probabilistic roadmap method from robotics allows the method to map the connectivity of the conformational subspace of a protein around given structural states. The results we have presented in this preliminary investigation are promising, warranting further research in this direction.

Acknowledgements

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

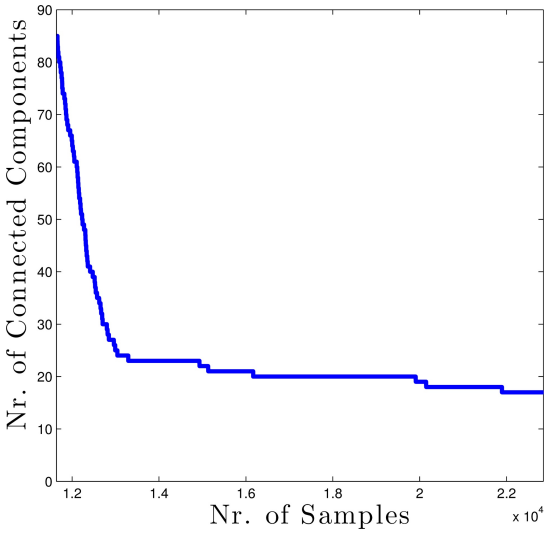

Figure 2: This plot shows the progress during the merging of connected components phase.

a Jeffress Trust Program in Interdisciplinary Research Award.

References

[1] Ibrahim Al-Bluwi, Marc Vaisset, Thierry Siméon, and Juan Cortés. Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC Structural Biology*, 13(Suppl 1):S8, 2013.

[2] M. S. Apaydin, D. L. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe. Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. *J. Comp. Biol.*, 10(3-4):257–281, 2003.

[3] Oliver Beckstein, Elizabeth J. Denning, Juan R. Perilla, and Thomas B. Woolf. Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open-closed transitions. *J. Mol. Biol.*, 394(1):160–176, 2009.

[4] H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. MIT Press, Cambridge, MA, 1st edition, 2005.

[5] 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*, 21(S1):116–125, 2005.

[6] N. Haspel, M. Moll, M. L. Baker, W. Chiu, and Kavraki. L. E. Tracing conformational changes in proteins. *BMC Struct. Biol.*, 10(Suppl1):S1, 2010.

[7] L. Jaillet, F. J. Corcho, J.-J. Perez, and J. Cortes. Randomized tree construction algorithm to explore energy landscapes. *J. Comput. Chem.*, 32(16):3464–3474, 2011.

[8] 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.*, 12(4):566–580, 1996.

[9] 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.*, 70(1):131–143, 2008.

[10] K. Molloy and A. Shehu. Elucidating the ensemble of functionally-relevant transitions in protein systems with a robotics-inspired method. *BMC Struct Biol*, 13(Suppl 1):S8, 2013.

[11] B. Olson, K. Molloy, and A. Shehu. In search of the protein native state with a probabilistic sampling approach. *J. Bioinf. and Comp. Biol.*, 9(3):383–398, 2011.

[12] M. Pascoal and E. Martins. A new implementation of Yens ranking loopless paths algorithm. *4OR Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, 2003.

[13] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and A. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2013.

[14] A. Shehu, L. E. Kavraki, and C. Clementi. Multi-scale characterization of protein conformational ensembles. *Proteins: Struct. Funct. Bioinf.*, 76(4):837–851, 2009.

[15] A. Shehu and B. Olson. Guiding the search for native-like protein conformations with an ab-initio tree-based exploration. *Int. J. Robot. Res.*, 29(8):1106–11227, 2010.

[16] G. Song and N. M. Amato. A motion planning approach to folding: from paper craft to protein folding. *IEEE Trans. Robot. Autom.*, 20(1):60–71, 2004.

[17] L. Tapia, S. Thomas, and N. Amato. A motion planning approach to studying molecular motions. *Communications in Information Systems*, 10(1):53–68, 2010.

[18] S. Thomas, G. Song, and N. Amato. Protein folding by motion planning. *Physical Biology*, (2):S148–S155, 2005.