Tracing Conformational Changes in Proteins Represented at a Coarse Level

Nurit Haspel*

Department of Computer Science, University of Massachusets Boston, Boston MA 02125 USA
nurit.haspel@umb.edu

Abstract. Many large protein complexes undergo extensive conformational changes as part of their functionality. Tracing these changes is important for understanding the way these proteins function. It is not always possible to obtain a high resolution structure for very large complexes. Electron cryo-microscopy (Cryo-EM) enables the representation of large macromolecular structures at a medium resolution level (4–8Å). Traditional conformational search methods cannot be applied to medium resolution data where structural information may be partial or missing. Additionally, simulating large scale conformational changes in proteins require a massive amount of computational efforts. We apply a search method from robotics to structural information obtained from medium resolution Cryo-EM maps, modeled to approximate backbone trace level. The pathways obtained by this method can be useful in understanding protein motions, providing reliable results for the medium resolution data. To provide a baseline validation for our method, we tested it on Adenylate Kinase and Cyanovirin. To test the data on actual cryo-EM determined structures, we simulated the conformational opening of the GroEL single ring complex. We show that we can produce low energy conformational pathways which correspond to known structural data. The method presented here is a promising step towards exploring the conformational motion of even larger complexes.

1 Introduction

Proteins are flexible molecules that undergo conformational changes as part of their interactions with other proteins or drug molecules [1]. Changes in torsional angles may induce localized changes or large scale domain motions. Tracing these changes is crucial for understanding the way these proteins perform their function. Only a relatively small number of proteins have their structure determined experimentally at atomic resolution. In cases where the high resolution structure cannot be determined, as in the case of very large complexes, electron cryo-microscopy (Cryo-EM) enables the structural determination of large macromolecular complexes at sub-nanometer resolution (4–8Å) [2, 3]. At this resolution the backbone or sidechain positions cannot always be determined, and typically only partial information exists about the location and connectivity of the secondary structure elements. More recently, with the advance of Cryo-EM determination methods and data processing tools, an increasing number of medium resolution data can be traced to a backbone level with high accuracy. An example of a medium

* Part of the work was done when the author was at the Department of Computer Science at Rice University, Houston, TX 77005 USA.

J. Suzuki and T. Nakano (Eds.): BIONETICS 2010, LNICST 87, pp. 343–356, 2012.

resolution structure can be seen in Figure 1(a), where a backbone trace of the GroEL chaperonin monomer is modeled, based on Cryo-EM image. Several recent studies perform local refinement and fitting of high resolution structures to Cryo-EM models [4, 5], but to the best of our knowledge, no algorithm exists for tracing and analyzing the large scale conformational changes undergone by complexes represented at medium resolution. Hence, a gap of knowledge exists. Existing physics-based computational methods that trace and simulate conformational changes in proteins where full atomic resolution is available include Molecular Dynamics (MD) [6], Monte Carlo (MC) [7] and their variants. They normally require full atomic representation of the molecule, and most available force fields are not suitable to deal with coarse grained representation, especially when modeled from sub-nanometer resolution images. In the past, several efficient conformational search algorithms have been developed. Some use a coarse representation of the protein molecule [8, 9] and employ various efficient search methods such as Normal Mode Analysis (NMA) [4, 10], elastic network modeling (ENM) [11, 12, 13], or morphing [14]. ENM- and NMA-based methods are especially useful in sampling local motions, and it is not always clear whether large scale conformational sampling methods can be sampled correctly using normal mode analysis, and the normal modes have to be re-calculated for the sampling to be correct. Sampling based motion planning methods have been successfully applied towards an efficient exploration of the conformational space of macromolecules. Motion planning is an area in robotics concerned with finding a pathway for robot-like objects in constrained environments [15, 16]. When applied to biological problems, the protein is represented as an articulated body with the degrees of freedom in all or selected torsional angles. The physical constraints are implicitly encoded in a penalty function which approximates the potential energy of the molecule. The conformational space of the protein is explored so that high energy regions are avoided and feasible conformational pathways are obtained more efficiently than with traditional simulation methods. Among the many applications of motion planning to biology are the characterization of near-native protein conformational ensembles [17, 18, 19], the study of conformational flexibility in proteins [20, 21], protein folding and binding simulation [22, 23], modeling protein loops [24, 20], simulation of RNA folding kinetics [25, 26] and recently the elucidation of conformational pathways in proteins, subject to pre-specified constraints [27, 28].

Many of those methods are successful in sampling the conformational landscape of proteins but are often biased by the protein native conformation and some of them require additional, problem specific information.

In this work we present an efficient motion-planning based methodology to perform conformational search on complex macromolecular structures represented at medium resolution extracted from Cryo-EM imaging. The molecule is mapped into a reduced representation using a small number of parameters that represents its degrees of freedom. This allows for larger complexes to be explored efficiently. A similar method has been applied recently to higher resolution structures [28]. The results of the previous research encouraged us to take the methodology one step further and explore lower resolution structures with high degree of uncertainty. To the best of our knowledge, there are no large scale conformational sampling methods that accept lower than backbone

resolution and this is the main strength of this method and the main difference from our previous work at [28].

Problem Statement. Given two conformational states of a molecule, denoted by start and goal, and represented as two medium resolution Cryo-EM maps, our goal is to find a set of affine transformations that, when applied successively to the degrees of freedom of the start conformation, the start conformation will be brought within a tolerance range of the goal conformation under a defined distance metric. Furthermore, each intermediate conformation along the pathway must be feasible under a given scoring function, which approximates its potential energy. The degrees of freedom of the structures lie in the flexible parts connecting rigid structural elements. In this work we assume that secondary structure elements do not change significantly during domain motions and that the flexible parts are the loops connecting them. While this assumption is true in many cases, there are cases where secondary structure elements melt or change. In these cases, and if backbone or higher resolution modeling of these parts is available, it is possible to incorporate a more detailed modeling of the flexible parts into the general framework of the algorithm without limiting the proposed procedure. The flexible parts can be modeled using existing methods such as elastic network modeling [11]. We also assume that approximate backbone trace can be made using structural modeling tools and Cryo-EM data analysis methods [29, 30]. Extension to cases where only partial information is given about the location of secondary structures is a subject of on-going and future work. Figure 1 shows an illustration of the conformational transition of the GroEL monomer from the closed structure (Figure 1(a)) to the opened structure (GroEL-GroES-ADP7) (Figure 1(b)). Figure 1(c-d) show the single-ring 7 member complex. The closed conformation is a backbone trace model generated from a mid-resolution cryo-EM image.

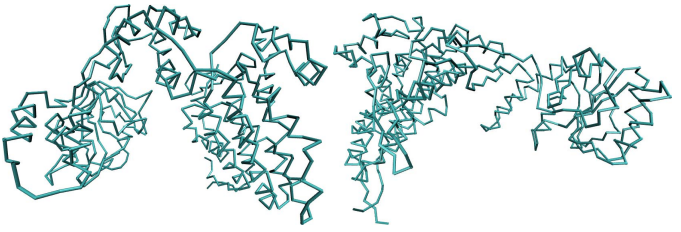
It should be emphasized that the algorithm does not always produce the same conformational pathway, but rather a possible pathway. By repeating the procedure several times we produce a set of feasible pathways, thus limiting the huge search space to a manageable number of possibilities which can later be refined and filtered using information about the tested systems.

2 Methods

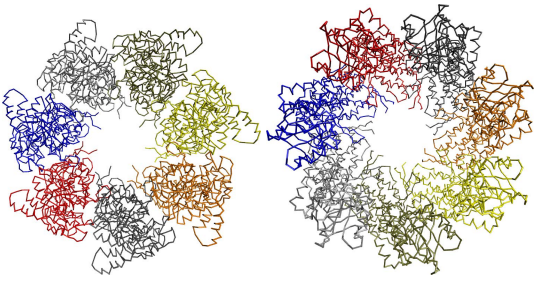
2.1 Data Representation

This part is the main difference from our previous work [28]. The data representation here is based upon a medium resolution Cryo-EM map rather than on a detailed atomic representation. Given a Cryo-EM map which contains electron density data, the data is encoded into a compact representation using the EMAN software package [31], developed to process Cryo-EM maps. A multi-level representation of the data allows us to conveniently manipulate different parts of the structure at will. The different levels of representation include:

1. Pseudo atoms [29], which are feature points of increased density, where most likely atoms are found. Notice that due to the resolution pseudo atoms do not correspond to exact atom locations.



(a) GroEL monomer (b) GroEL-GroES-ADP7



(c) GroEL complex (d) GroEL-GroES-ADP7 complex

Fig. 1. Cryo-EM maps generated for GroEL monomer using EMAN [31]. (a) The GroEL monomer (Created from cryo-EM model). (b) The GroEL-GroES-ADP7 monomer taken from PDB structure 1SX4, chain A. (c) The GroEL complex (built as a symmetric complex of the monomer modeled in (a).). (d) The GroEL-GroES-ADP7 complex - taken from PDB structure 1SX4.

2. A skeleton [32], which captures the topology of the structure and helps detecting helix and sheet regions and their connectivity.
3. Secondary structures, which are assigned using either prior knowledge or the SSE-Hunter and SSEBuilder tools [29]. These tools assign secondary structure elements to a given Cryo-EM map using local topology information obtained from the pseudo atoms and connectivity information obtained from the skeleton.

Based on this data, the location of secondary structure elements and often even partial or full backbone and side chain information can be obtained with high degree of accuracy [33]. The high-level data structure that represents a conformation is a graph $G = (V,E)$ such that each secondary structure element is a node $v \in V$ in the graph. Two secondary structure elements v_1 and v_2 are connected by an edge $e \in E$ if there is at least one pair of adjacent amino acids r_1, r_2, such that $r_1 \in v_1$ and $r_2 \in v_2$. The backbone angles in r_1, r_2, and a small number of sequentially adjacent residues form the degrees of freedom of the protein. In other words, the protein motions consist of bond rotations in these residues while the remaining angles stay fixed.

Based on the graph we construct a spanning tree $T = (V,X)$ where X is a subset of E using a greedy approach. The root of the tree is specified as the structure that is expected to move the least during the search as determined by aligning the start and goal structures and measuring the least RMSD between corresponding secondary structure

elements. Each one of the root's neighbors forms a child node in the tree, and at each stage the selected node and its adjacent edges are removed from the graph. The process repeats iteratively until all the secondary structure elements are represented in the tree. There may be more than one correct topology to the structure. We picked the topology that follows the order of amino acids, which seems to give the best results. It should be noted that it is not always possible when no information about the exact location of the amino acids is available when the resolution is too low. In some cases we may know that the positions of certain secondary structure elements is likely to stay fixed. This allows us to speed up the search for a feasible pathway by restricting motions to the remaining secondary structure elements. Let $K \subseteq V$ be the set of secondary structures that is free to move. This set is used below in the definition of a distance metric for our representation.

It should be emphasized that this representation can be used even if there is no detailed structural information such as backbone location or even if only partial information about secondary structures breaking down, since the conformation representation assumes only secondary structure knowledge.

2.2 Distance between Structures

The search method requires a distance measure to estimate the progress in the conformational search. In the case of proteins and protein complexes the distance measure is not trivial to define due to the complexity of protein structures and the high dimensionality of the search space. Finding a good distance measure between two molecules is an active area of research [34]. This issue becomes especially challenging when the proteins are represented at a coarse resolution and traditional distance measures such as RMSD may not be accurate due to the approximate location of the α carbon representation that may cause inaccuracies to accumulate. In order to measure the distance between structures we use a method we developed previously and gave good result in a previous work [28]. Since it only requires knowledge of the location of secondary structure elements it is especially suitable for coarse grained molecular representations.

The distance measure is defined in terms of the relative positions between secondary structure elements. We compute for a conformation C a feature vector:

$$v_C = \langle score(C^1), score(C^2), \ldots, score(C^k) \rangle \tag{1}$$

where the components of the vector are scores calculated for the K secondary structure elements of the conformation, based on their positioning with respect to one another. The distance between two conformations, C_1 and C_2 is defined as the Euclidean distance between their feature vectors, i.e., $\|v_{C_1} - v_{C_2}\|^2$. By definition, when C_2 is the goal structure, the *score* of C_1 is the magnitude of its vector representation. Therefore, the lower the score for a given conformation, the more similar it is to the goal structure.

2.3 Penalty Function

There are several potential functions that are suitable for C-α representation [8]. These potential functions take into account the hydrophobicity of the amino acids, their

interactions with the solvent and with one another. While coarse-grained energy models are an approximation of the protein potential energy and are often biased towards a folded state, they give good results given their simplicity. They allow for an efficient exploration of the protein conformational space where more detailed structural representations, which require a more accurate potential function, consume a vast amount of resources. We use the energy function developed by Brown and Head-Gordon [35].This function has been shown to give good results while using only one bead per amino acid. The energy function classifies the 20 amino acids into three categories - Hydrophobic (H), Polar (P) and Neutral (N). The potential energy is given by the following equation:

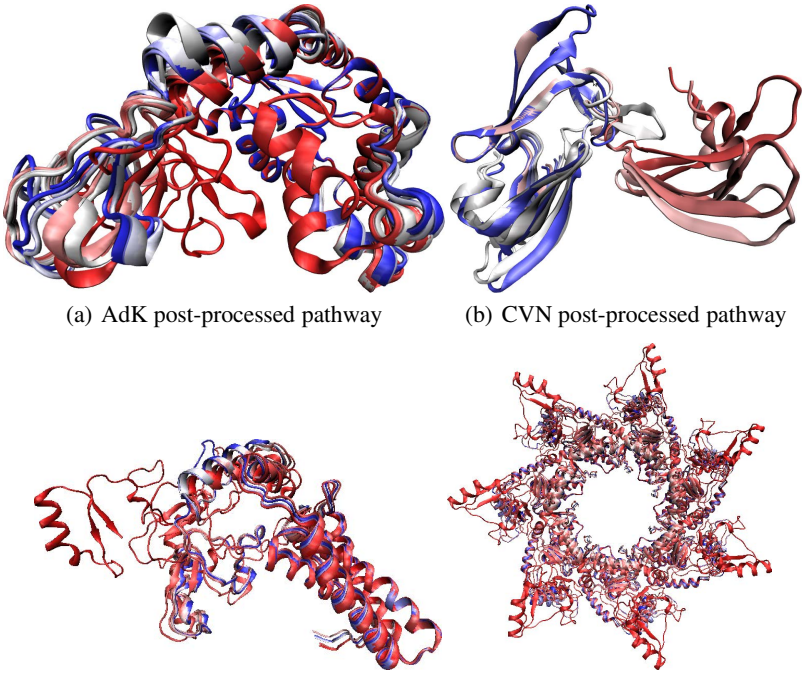
$$E_{\text{total}} = \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} [A(1+\cos\phi) + B(1-\cos\phi) \\ +C(1+\cos 3\phi) + D(1+\cos[\phi + \frac{\pi}{4}])] + \sum_{i,j \geq i+3} 4\varepsilon_H S_1 [\frac{\sigma}{r_{ij}^{12}} - \frac{\sigma}{r_{ij}^{6}}] \tag{2}$$

Where $r_{i,j}$ is the distance between amino acids and the constants A, B, C, and D, S_1 and S_2 are determined by the secondary structure type and the amino acid type. See [35] for more details.

When no backbone trace is available, none of the above coarse grained potential functions can be used. Our structural representation and search algorithm are suitable for sub-nanometer resolution and do not assume knowledge about the locations of specific amino acids. However, there are only a few papers regarding coarser than backbone representation [36]. The use of these very coarse models in the context of our model, where no normal mode analysis is performed, should be tested carefully, since they may admit non-biological motions as well. However, it can be used as a filtering tool that greatly reduces the number of possible pathways. In cases where a more detailed structural model exists, a more realistic energy function can be used for filtering and refinement. This area is the subject of on-going and future research.

2.4 Search Methodology

The search is performed using a sampling-based motion planning algorithm. Motion planning algorithms have been applied extensively in the past to solve biological problems due to the analogy between protein chains and robotic articulated mechanisms [22, 23]. The search methodology applied in this paper is based on the Path-Directed Subdivision Tree (PDST) planner [37]. We chose this algorithm because of its good performance with articulated systems with complex dynamics moving in physically constrained environments. It has shown good results in our previous work [28]. We adapted the algorithm to model protein motions. In our adaptation, the planner iteratively constructs a tree of conformational pathways as the search progresses. The input to the algorithm consists of the start and end conformations of a molecule, represented as sets of articulated secondary structures as discussed in Data Representation above. The root of the search tree is a "pathway" of length 0 consisting only of the starting structure. At every iteration a previously generated pathway is selected for propagation using a deterministic scoring scheme described below. From a random conformation



(a) AdK post-processed pathway

(b) CVN post-processed pathway

(c) Groel Monomeric post-processed pathway

(d) GroEL complex post-processed pathway

Fig. 2. Illustration of the results for AdK, CVN and GroEL: (a) An example of a conformational pathway obtained for AdK after resolution enhancement and basic energy minimization. (b) An example of a conformational pathway for CVN. (c-d) An example of a conformational path of GroEL. The monomeric path is shown in (c) for clarity, and the entire complex is illustrated in (d). The conformation colors are interpolated on the red (start) to blue (goal) scale.

along that pathway, a new pathway is propagated by applying a small random rotation to the ϕ or ψ backbone dihedral angle of a residue that resides on a loop connecting two randomly chosen secondary structure elements. A molecular motion is sampled by applying the rotation until a high energy conformation is reached. The coarse grained energy function described above is used to determine when a high energy conformation is encountered. A high energy conformation is defined as being more than 100 kcal/mol above the starting energy. The relatively high threshold aims to compensate for inaccuracies in the structure due to the low resolution. The algorithm maintains a subdivision of the low-dimensional projection of the conformational space (described in Distance Between Structures above) into cells, such that no sample spans more than one cell in the subdivision. The goal of the subdivision is to guarantee coverage of the search space [37]. After a sample is selected for propagation, the cell containing that sample is subdivided into two cells. The algorithm keeps track of how many samples are contained in each cell to estimate how dense the sampling is in different areas of the space. It maintains a scoring scheme that gives selection preference to samples residing

in large, empty cells, thus pushing the exploration towards unvisited areas in the conformational space. Probabilistic completeness is obtained via a scoring scheme that favors the selection of samples contained in larger cells and leads to unexplored areas of the search space. The sample scores are updated in a way that guarantees that every sample in the tree will eventually be selected for propagation and avoids over-sampling of parts of the space. Previous studies in path-directed motion planning algorithms [28, 38] showed that employing a biasing scheme in a small percentage of the iterations greatly improves the performance of the planner. We employed biasing at 10% of the iterations. During these iterations the scoring scheme described above is ignored and a sample is chosen out of a pool of conformations closest to the goal conformation, which gives the planner a better chance to successfully terminate the search. We found that the biasing improves the performance of the algorithm. Our top-level algorithm runs PDST iteratively. Each iteration runs until a generated conformation is closer to the goal conformation than a pre-specified intermediate distance threshold, where the distance threshold is determined by the distance measure described above. We found that a threshold of 0.8–0.9 of the distance between the start and goal conformations is usually sufficient to achieve good results and running further does not benefit the results much. The iterative runs of the PDST planner help reduce memory use and improve performance, as also shown in [39].

3 Results and Discussion

We ran the algorithm on three test cases, Adenylate Kinase (AdK), cyanovirin (CVN) and the GroEL single-ring complex. In this study we seek to provide a proof of concept and some real test-cases. AdK and CVN undergo extensive conformational transitions, they are well studied and have an abundance of data for testing and comparison. To provide test for real backbone-resolution data we tested the algorithm on the GroEL 7-member ring complex. A CryoEM model is available only for the closed form, and the open form was taken from the PDB, using accession code 1SX4. The PDB structures were respresented as C-α traces. Figure 2 shows an illustration of the AdK, CVN and GroEL examples.

We ran the algorithm 100 times per protein on the UMass Boston Supercomputing Cluster, where each machine runs at 2.2 Ghz and has 4 GB RAM. For comparison purposes, we produced conformational pathways using a random walk Monte Carlo like algorithm [7] with the same resolution, representation and penalty function described in this work. The random walk algorithm is similar to the one used for comparison in our previous work [28]. Using the same representation, similarity score and potential function described in our algorithm, the random walk algorithm differs from the common use of Monte Carlo in protein conformational search. Rather than optimizing the energy, it optimizes the above mentioned similarity score in order to simulate a conformational pathway from the start to the goal conformation. The energy, while not optimized, is used to filter out non-feasible conformations. The random walk implementation uses the Metropolis criterion for the selection of steps. At each iteration a random conformational pathway is generated from the current conformation by applying a small random transformation to one of the randomly chosen degrees of freedom

connecting secondary structure elements, in a similar way to the one used to generate new conformations described in the Search Methodology subsection above. If a step brings the similarity score of the generated conformation closer to the goal it will be accepted. Otherwise it is accepted with a probability proportional to $e^{\Delta S}$ where ΔS is the difference in the similarity score of the current step and the previous step. In practice, this criterion accepts all "good" steps while allowing a very small fraction of "bad" steps.

No comparison was made to any other conformational search method since random walk is very easy to implement, but no other method exists that models conformational changes in low resolution level and therefore such a comparison would be meaningless. In order to compare the performance of the two methods by an objective standard, each was run for a fixed amount of time and the least RMSD (lRMSD) of the closest conformation to the goal at that given time step was measured. Generally, lRMSD is not available for medium resolution structures. However, it was used for the sake of this initial baseline validation due to the fact that the full resolution structures are available. lRMSD was measured after 10 and 20 minutes for AdK. In the case of GroEL, which is a longer running example, measurements were taken after 15 and 30 minutes. For CVN, which was a shorter example, measurements were taken after 5 and 10 minutes. Table 3 summarizes the average lRMSD results over 80 test runs, where the top and bottom 10% outliers were removed.

Adenylate Kinase (AdK). AdK is a monomeric phosphotransferase enzyme that catalyzes reversible transfer of a phosphoryl group from ATP to AMP. AdK contains 214 amino acids and assumes an "open" conformation in the unligated structure and a bound, "closed" conformation. The lRMSD between the two structures is 6.95Å. Supposedly, during the transition from the "open" to "closed" form, the largest conformational change occurs in the LID (residues 118–167) and NMP (residues 30–67) domains with the rest of the protein – the CORE domain being relatively rigid. We modeled the closed state to open state motion using the C-α traces of PDB codes 1AKE and 4AKE for the closed and open states respectively. Our model contains 7 secondary structure elements where most of the CORE domain was modeled as one large segment and was considered fixed, since it does not undergo a large-scale motion. Figure 2(a) shows the overlapped resulting structure with the goal structure. The final set of transformations was applied to the backbone traces (which are known in this case) to generate the figure. The C-α RMSD from the goal structure is 1.622Å. As seen in table 3 the resulting average lRMSD was 1.93Å in the end of the runs. Random walk performed slightly worse compared to our planner with an average lRMSD of 2.26Å .

Cyanovirin-N (CVN). CVN is an anti-viral fusion inhibitor protein that binds to viral sugars, and is trialed for preventing sexual transmission of HIV. It comprises two repeat domains of 30% sequence identity which undergo swapping [40]. We simulated the unpacking of the repeat domains of a single chain from the intertwined monomeric conformation to an extended domain-swapped conformation. The swapped conformations deviate by approximately 16Å. CVN contains 101 amino acids and our model contains 6 rigid elements. The flexible rotation axis resides mainly between residues 48–55. The distance measure threshold for successful termination of the algorithm was

Performance statistics for the AdK, CVN and GroEL complex examples. The average ±(standard deviation) lRMSD data were taken over 80 runs where the top and bottom 10% outliers were removed from the original set of 100 runs.

	AdK	**AdK RW†**	**CVN**	**CVN RW†**	**GroEL**	**GroEL RW†**
Initial lRMSD (Å)	6.95	6.95	16.01	16.01	14.64	14.64
#Residues	214	214	101	101	525‡	525‡
lRMSD at first measurement (Å)	2.49±0.41	2.621±0.54	4.85±1.55	4.46±1.84	7.06±0.38	7.07±0.44
lRMSD at second measurement (Å)	2.36±0.38	2.54±0.53	3.57±0.95	4.42±1.67	6.62±0.43	6.95±0.37
Final lRMSD (Å)	1.93±0.16	2.32±0.49	2.29±0.25	3.20±1.2	6.26±0.43	6.83±0.34

† Random walk. See Results section for details.

a normalized distance of 0.91 from the goal conformation. Figure 2(b) shows an example of a pathway from the start to the end conformation. The C_α RMSD from the goal structure is 1.67Å. As seen in Table 3, our algorithm significantly outperformed random walk with an average lRMSD of about 2.29Å comparing to 3.2Å for random walk. Many of our runs got as low as 1.5Å from the final conformation. The average run time was approximately 29 minutes.

GroEL Complex. The GroEL protein belongs to the chaperonin family and is found in a large number of bacteria [41]. It is required for the correct folding of many proteins. GroEL requires the lid-like cochaperonin protein complex GroES. Binding of substrate protein, in addition to binding of ATP, induces an extensive conformational change that allows association of the binary complex with GroES. We modeled the epical domain movement from the GroEL-GroES-ADP7 complex (modeled from chain A of PDB code 1SX4) to the GroEL monomer (modeled using a Cα trace extracted from Cryo-EM data, analogous to the closed GroEL monomer) The monomer contains 525 amino acids, and our model contains 8 secondary structure elements where most of the equatorial domain, whose structure does not change significantly, was modeled as one large segment and was considered fixed. The initial lRMSD between the Cα atoms of the two complexes is 14Å. The lRMSD from the goal structure was measured with respect to the Cα trace of the Cryo-EM model. Figure 2(c-d) shows an example of a pathway from the start to the end conformation. Table 3 shows that our method significantly outperforms random walk. The average lRMSD between the resulting structures and the goal structure was 6.2Å after 30 minutes compared to approximately 6.8Å for MC. The pathways obviously indicate a closing motion of the complex, but lRMSD was used here mainly for compatibility with the other examples. It is not a very good indicator of the quality of the pathway in this case, since the Cryo-EM model contains inaccuracies that makes it somewhat different than the atomic structure of GroEL which was used to model the open structure. In a previous work [28] we used PDB code 1SS8, which is close in structure to the Cryo-EM model used in this work. We then observed an lRMSD of less than 5Å, with some runs reaching as close as 2.5Å from the goal structure. When comparison between the two structures is made less accurate, other distance measurements should probably be considered [34].

Analysis of the Results. In order to provide initial validation for our results, we tested whether our algorithm produces biologically reasonable, low energy pathways when using an all-atom force field. Such an analysis was done in earlier works [27, 28],

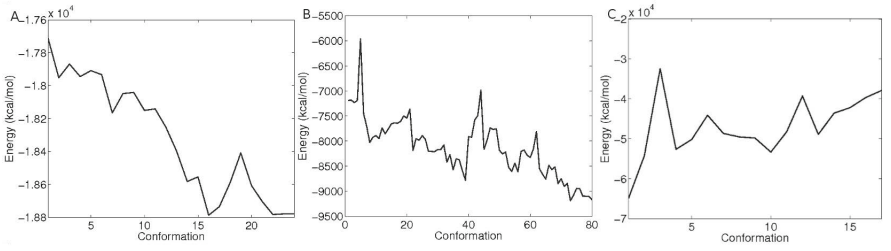


Fig. 3. Energetic profiles of the resulting pathways (a) A potential energy plot for ADK, CVN and GroEL slightly minimized conformational pathways. Notice the different potential energy scale.

where the authors used a similar method to show that their conformational search was reasonable. For this purpose we chose targets for which high resolution structures are available. Side chain information was completed for the resulting pathways using the algorithm described in [42]. The resulting full-atomic structures were minimized for 500 Steepest Descent steps using the AMBER energy minimization package [6]. The minimization was done for a relatively small number of steps and was restrained in order to resolve initial clashes but not cause large conformational changes to the structures. The purpose of this test is not to provide a fully minimized pathway, but to show that the algorithm produces pathways with reasonable conformations whose clashes can be resolved within a small number of minimization steps. Figure 3 shows the potential energy plot of a sampled conformational pathway for the three test cases. As seen, even with a small number of energy minimization steps all the intermediate structures exhibit low potential energies, below -20000 kcal/mol for AdK, below -7000 kcal/mol for CVN and around -40000 kcal/mol for GroEL, as measured by AMBER.

4 Conclusions

We present a novel method for exploring large scale conformational changes in proteins represented at medium resolution using Cryo-EM data and image processing techniques. The search methodology is based on robot motion planning, and it strikes a balance between an efficient coverage of the conformational space and fast exploration towards the goal structure. The input molecule is modeled using a coarse representation and a relatively simple potential function to guide the search. This representation does not require atomic details and thus makes the computation tractable and especially useful in cases where a detailed structural model is not available and facilitates dealing with partial or missing data. We tested our algorithm on the following well studied proteins: Adenylate Kinase and Cyanovirin. Additionally, we ran the algorithm on actual backbone-resolution models obtained from cryo-EM data, the GroEL complex (where one state was obtained from cryo-EM and the other state from the PDB). We show that our method performs significantly better than random walk by producing low energy pathways with resulting structures closer to the goal structure. We believe this is an important step towards a larger scale modeling of more complex biological systems such

as virus capsid shells. Additionally, since there is an abundance of cryo-EM models whose resolution do not allow backbone trace modeling, this method can be the basis to simulating conformational changes in even lower than backbone resolution models. The algorithmic framework is similar, and the main difference is selecting a potential function that handles lower-than-backbone resolution. This is an area of on-going work.

Acknowledgements. I thank Dr. Steve Ludtke for providing part of the data and the training on Cryo-EM processing software. I also thank Dr. Tao Ju for useful ideas and Dr. Erion Plaku and Ioan Şucan for providing code which was incorporated in this work. Special thanks to Dr. Lydia Kavraki, Dr. Mark Moll, Dr. Matt Baker and Dr. Wah Chiu for their collaboration and valuable contribution to this work. Computational experiments were conducted on the supercomputing facilities in the College of Science and Mathematics at UMass Boston.

References

1. Perutz, M.F.: Mechanisms of cooperativity and allosteric regulation in proteins. Quart. Rev. Biophys. 22, 139–236 (1989)
2. Schmid, M.F., Sherman, M.B., Matsudaira, P., Chiu, W.: Structure of the acrosomal bundle. Nature 431, 104–107 (2004)
3. Jiang, W., Li, Z., Zhang, Z., Baker, M., Prevelige Jr., P.E., Chiu, W.: Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. Nature Structural Biology 10(2), 131–135 (2003)
4. Schroeder, G., Brunger, A.T., Levitt, M.: Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure 15, 1630–1641 (2007)
5. Lasker, K., Dror, O., Shatsky, M., Nussinov, R., Wolfson, H.J.: EMatch: discovery of high resolution structural homologues of protein domains in intermediate resolution cryo-EM maps. IEEE/ACM Trans. Comput. Biol. Bioinform. 4(1), 28–39 (2007)
6. Case, D.A., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz Jr., K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.: The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668–1688 (2005)
7. Kirkpatrick, S., Gelatt Jr., C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
8. Head-Gordon, T., Brown, S.: Minimalist models for protein folding and design. Curr. Opin. Struct. Biol. 13(2), 160–167 (2003)
9. Whitford, P.C., Miyashita, O., Levy, Y., Onucic, J.N.: Conformational transitions of adenylate kinase: Switching by cracking. Journal of Molecular Biology 366(5), 1661–1671 (2007)
10. Schuyler, A., Jernigan, R., Qasba, P., Ramakrishnan, B., Chirikjian, G.: Iterative cluster-nma: A tool for generating conformational transitions in proteins. Proteins 74, 760–776 (2009)
11. Zheng, W., Brooks, B.: Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. J. Mol. Biol. 346(3), 745–759 (2005)
12. Temiz, N., Meirovitch, E., Bahar, I.: Escherichia coli adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)n-nmr relaxation data. Proteins 57, 468–480 (2004)
13. Gohlke, H., Thorpe, M.: A natural coarse graining for simulating large biomolecular motion. Biophysical Journal 9, 2115–2120 (2006)

14. Weiss, D., Levitt, M.: Can morphing methods predict intermediate structures? J. Mol. Biol. 385, 665–674 (2009)
15. Choset, H., Lynch, K.M., Hutchinson, S., Kantor, G., Burgard, W., Kavraki, L.E., Thrun, S.: Principles of Robot Motion: Theory, Algorithms, and Implementations. MIT Press (2005)
16. Kavraki, L.E., Švestka, P., Latombe, J.-C., Overmars, M.H.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation 12, 566–580 (1996)
17. Shehu, A., Kavraki, L., Clementi, C.: On the characterization of protein native state ensembles. Biophysical Journal 92(5), 1503–1511 (2007)
18. Shehu, A., Kavraki, L., Clementi, C.: Multiscale characterization of protein conformational ensembles. Proteins: Structure, Function and Bioinformatics 76(4), 837–851 (2009)
19. Haspel, N., Geisbrech, B., Lambris, J., Kavraki, L.E.: Multi-scale characterization of the energy landscape of proteins with application to the c3d/efb-c complex. Proteins: Structure, Function and Bioinformatics 78(4), 1004–1014 (2010)
20. Cortés, J., Siméon, T., de Angulo, V.R., Guieysse, D., Remauld-Siméon, M., Tran, V.: A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 21(suppl. 1), i116–i125 (2005)
21. Shehu, A., Clementi, C., Kavraki, L.E.: Sampling conformation space to model equilibrium fluctuations in proteins. Algorithmica 48, 303–327 (2007)
22. Thomas, S., Tang, X., Tapia, L., Amato, N.M.: Simulating protein motions with rigidity analysis. J. Comp. Biol. 14(6), 839–855 (2007)
23. Chiang, T.H., Apaydin, M.S., Brutlag, D.L., Hsu, D., Latombe, J.-C.: Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics. J. Comp. Biol. 14(5), 578–593 (2007)
24. Yao, P., Dhanik, A., Marz, N., Propper, R., Kou, C., Liu, G., van den Bedem, H., Latombe, J.-C., Halperin-Landsberg, I., Altman, R.B.: Efficient algorithms to explore conformation spaces of flexible protein loops. IEEE/ACM Trans. Comput. Biol. Bioinform. 5(4), 534–545 (2008)
25. Tang, X., Thomas, S., Tapia, L., Amato, N.M.: Tools for simulating and analyzing rna folding kinetics. In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), San Francisco, CA, USA, pp. 268–282 (April 2007)
26. Shehu, A., Clementi, C., Kavraki, L.E.: Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. Proteins: Structure, Function and Bioinformatics 65, 164–179 (2006)
27. Raveh, B., Enosh, A., Furman-Schueler, O., Halperin, D.: Rapid sampling of molecular motions with prior information constraints. Plos Comp. Biol. (2009) (in press)
28. Haspel, N., Moll, M., Baker, M., Chiu, W., Kavraki, L.E.: Tracing conformational changes in proteins. BMC Structural Biology (2010) (in press)
29. Baker, M.L., Ju, T., Chiu, W.: Identification of secondary structure elements in intermediate resolution density maps. Structure 15, 7–19 (2007)
30. Abeysinghe, S.S., Ju, T., Baker, M., Chiu, W.: Shape modeling and matching in identifying protein structure from low-resolution images. In: ACM Symposium on Solid and Physical Modeling, pp. 223–232 (2007)
31. Ludtke, S.J., Baldwin, P.R., Chiu, W.: EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. 128, 82–97 (1999)
32. Ju, T., Baker, M.L., Chiu, W.: Computing a family of skeletons of volumetric models for shape description. Computer Aided Design 39(5), 352–360 (2007)
33. Zhang, J., Baker, M., Schroeder, G., Douglas, N., Reissman, S., Jakana, J., Dougherty, M., Fu, C., Levitt, M., Ludtke, S., Frydman, J., Chiu, W.: Mechanism of folding chamber closure in a group ii chaperonin. Nature 463, 379–383 (2010)

34. Ballester, P.J., Richards, W.G.: Ultrafast shape recognition to search compound databases for similar molecular shapes. J. Comput. Chem. 28(10), 1711–1723 (2007)
35. Brown, S., Fawzi, N., Head-Gordon, T.: Coarse grained sequences for protein folding and design. Proc. Nat. Acad. USA 100, 10 712–10 717 (2003)
36. Doruker, P., Jernigan, R., Bahar, I.: Dynamics of large proteins through hierarchical levels of coarse-grained structures. J. Comput. Chem. 23(1), 119–127 (2002)
37. Ladd, A.M.: Motion planning for physical simulation. Ph.D. dissertation, Dept. of Computer Science, Rice University, Houston, TX (Dec. 2006)
38. Şucan, I.A., Kruse, J.F., Yim, M., Kavraki, L.E.: Reconfiguration for modular robots using kinodynamic motion planning. In: ASME Dynamic Systems and Control Conference. Michigan, Ann Arbor (2008)
39. Tsianos, K., Kavraki, L.E.: Replanning: A powerful planning strategy for hard kinodynamic problems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France, pp. 1667–1672 (September 2008)
40. Botos, I., O'Keefe, B., Shenoy, S., Cartner, L., Ratner, D., et al.: Structures of the complexes of a potent anti-hiv protein cyanovirin-n and high mannose oligosaccharides. J. Biol. Chem. 277, 34336–34342 (2002)
41. Zeilstra-Ryalls, J., Fayet, O., Georgopolous, C.: The universally conserved GroE (Hsp60) chaperonins. Annu. Rev. Microbiol. 45, 301–325 (1991)
42. Heath, A.P., Kavraki, L.E., Clementi, C.: From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. Proteins: Structure, Function and Bioinformatics 68(3), 646–661 (2007)