Improving Protein Conformational Sampling by Using Guiding Projections

Anastasia Novinskaya, Didier Devaurs, Mark Moll, and Lydia E. Kavraki
Department of Computer Science
Rice University
Houston, Texas, USA 77005
Email: {an17, devaurs, mmoll, kavraki}@rice.edu

***Abstract*—Sampling-based motion planning algorithms from the field of robotics have been very successful in exploring the conformational space of proteins. However, studying the flexibility of large proteins with hundreds or thousands of Degrees of Freedom (DoFs) remains a big challenge. Large proteins are also highly-constrained systems, which makes them more challenging for standard robotic approaches.**

So-called "expansive" motion planning algorithms were specifically developed to address highly-dimensional and highly-constrained problems. Many such planners employ a low-dimensional projection to estimate exploration coverage and direct their search based on this information. We believe that such a projection plays an essential role in the success of these planners.

This paper shows how the low-dimensional projection used by expansive planners can be tailored with respect to a given molecular system to enhance the process of conformational sampling. We introduce a methodology to generate an expert projection using any available information about a given protein. We evaluate this methodology on several conformational search problems involving proteins with hundreds of DoFs. Our experiments demonstrate that incorporating expert knowledge into the projection can significantly benefit the exploration process.

I. Introduction

Proteins are involved in almost every process within living organisms. A protein's function is known to be defined by its structural conformation and the way it changes. Often a protein's activity is modulated/characterized by its ability to switch among several stable conformations. Understanding how a protein shifts between these states is essential for treating or preventing diseases related to the protein's dysfunction [1]. However, the shift happens so rapidly that it is extremely hard to monitor it experimentally. For this reason, a common approach to gain knowledge of how proteins move is to model this process computationally. There exist several classes of algorithms for simulating protein motion. They vary from highly physically precise and computationally expensive simulation techniques, such as Molecular Dynamics (MD) [2], to methods producing fast but rather approximate analysis of protein motions, such as Normal Mode Analysis (NMA) [3] and Elastic Network Models (ENM) [4].

Our work on modeling protein flexibility involves sampling-based motion planning algorithms that have been adapted from the field of robotics. These methods fill the gap between the two classes of approaches mentioned above: they represent a trade-off between physical accuracy and computational cost. As a result, these methods have been very efficient at producing representative large-scale protein motions [5], [6]. Sampling-based algorithms explore the conformational space of a protein by randomly sampling it (usually using a special heuristic) and constructing a graph where each node represents a feasible low-energy protein conformation (or state), and each edge represents a possible low-energy local transition between two states. The computed graph describes the topology of a protein's energy landscape and the connectivity of its low-energy areas. This graph can be used to find possible large-scale transitions between two given protein conformations.

All current computational methods for modeling protein flexibility, including sampling-based techniques, suffer from the curse of dimensionality: their complexity grows exponentially with the number of dimensions. Moreover, large proteins are also highly-constrained systems, which increases simulations complexity even further. Identifying representative motions for even middle-sized proteins (with hundreds of residues) remains an active field of research. In contrast to other methods, sampling-based techniques offer various ways to mitigate the above-mentioned problems.

The work presented in this paper involves a specific kind of motion planners that use low-dimensional projections to overcome the curse of dimensionality [7], [8]. These planners constitute a large part of the group of so-called "expansive" planners, that grow their conformational graph by iteratively applying an expansion procedure. The planners on which we focus use linear projections to assess how they cover the conformational space of a protein with the samples they produce. Based on this information, these planners direct their search towards unexplored regions of the conformational space.

Even though large proteins have thousands of Degrees of Freedom (DoFs), the extensive analysis of protein conformations generated by various methods (such as MD [9], X-Ray Crystallography [10], or Normal Mode Analysis [11]) has shown that the majority of their residues move in a correlated fashion. As a result, protein motions can usually be characterized by just a few collective DoFs [12], [13]. Therefore, projections that are aligned with the low-dimensional manifold of protein motion can represent a good approximation of the high-dimensional conformational space of a protein.

The contribution of this paper is an assessment of the role

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

of a projection on the process of conformational exploration. We introduce a new methodology to construct effective low-dimensional projections using simple biological knowledge available for a given protein. We demonstrate that such "expert" projections can improve the process of conformational search for expansive planners. We have applied such projections to two different kinds of conformational search problems: (1) finding a feasible low-energy transition between two given protein states, and (2) exploring the conformational space starting from a given protein state. The expert projections show improvements in algorithm runtime, in the case of the first problem, and in space coverage, in the case of the second problem.

The rest of the paper is organized as follows: the next section presents some related work as well as the context of our work. The framework we use for protein modeling is described in Section 3; the section includes further explanations on the role of projections in protein conformational sampling. In Section 4, we report the results of our experiments; they involve three middle-sized protein systems (having at least one hundred residues): Cyanovirin-N, Calmodulin, and Ribose-binding protein. Finally, Section 5 concludes the paper and presents some of our future work.

II. Related work

A. Sampling-based Path Planning Methods

Sampling-based methods have been very effective for the fast computation of representative motions of molecular systems [5], [6]. A broad range of approaches exploit sampling-based techniques to address various biological problems, such as exploring energy landscapes [14], modeling protein folding pathways [15], analyzing protein loops [16], or modeling large-scale transitions in a protein structure [17].

The sampling-based methods explore the conformation space (i.e., the space of all possible combinations of values that the system's DoFs can take) of a system and build a graph connecting the feasible conformations. At each step, a sampling-based algorithm samples a conformation. Then it performs a validity check for the chosen sample. In protein modeling, this means eliminating high-energy protein conformations. If the sampled state satisfies all the constraints of the problem, it is added to the graph as a new node, otherwise it is discarded. Finally, the valid states are connected into a graph structure by adding edges between the nearest configurations. Edges also often undergo a validity test: only the edges that satisfy the system's constraints are added to the graph. The constructed graph represents the topology of conformation space; the nodes represent the low-energy clash-free conformations of the protein, and the edges represent feasible local transitions between the corresponding conformations.

B. Using Projections to Guide Conformational Sampling

Despite the capability of sampling-based methods to generate large-scale protein motions much faster than physics-based simulations, they still suffer from the curse of dimensionality. Middle-sized and large proteins require hundreds or thousands of variables to encode a conformation. Moreover, because large proteins often represent highly-constrained systems, they can only move in a very limited fashion. These issues represent a significant challenge for sampling-based approaches as their complexity grows exponentially with the dimensionality of the system as well as with the decrease in volume of the space of low-energy conformations.

In this work we use a group of expansive planners that were developed to specifically tackle high-dimensional and highly-constrained problems [7], [8]. Expansive planners iteratively grow a tree of feasible protein conformations by choosing a state which is already in the tree (and therefore has low energy), and slightly perturbing some DoFs of that state to generate a new, child conformation. These planners employ a low-dimensional projection to store statistics of the exploration progress. To identify a promising parent state for expansion, these planners use the coverage estimate provided by the projection.

Different approaches have been exploited in the context of sampling-based methods in general and "expansive" algorithms in particular to overcome the curse of dimensionality. One of the common techniques is to identify flexible and rigid parts of a protein on-the-fly and bias exploration towards the flexible regions. The framework employed in the current work has the functionality of identifying the flexible protein regions automatically based on a protein's secondary structure. An alternative approach for rigidity analysis based on the pebble game computations [18] is presented in the works of Amato's group (see, for example, [19]), and Streinu's group [20]. The rigidity analysis technique has been mainly used to bias the local search (choosing the local protein motion) [19], [21]. In the current work we focus on enhancing the global search (choosing the conformation from the tree for further expansion) by estimating the effect of a low-dimensional projection on the overall exploration process.

Many recent approaches use the expansive planners to explore protein conformational landscapes [22]–[24]. They suggest a variety of algorithms for computing low-dimensional projections: including simplistic 1D projections based on lRMSD towards a goal structure (or some milestone) [24]; slightly more advanced 3D projections computed from average interatomic distances to the given points of the structure [25], [22]; and the quite intricate 1D projections generated from the contact matrix with usage of hashing algorithms [23]. Often, mentioned projections are also combined with 1D projection layer based on the energy of the structure [23]. All of these methods have some biological intuition to support them. However, only one of them, [23], provides some analysis of how the suggested structural profiles enhance the conformational search.

In many cases, low-dimensional projections for sampling-based planners are chosen randomly. Prior work [26] studies the influence of such projections in the context of robotic systems with at most a few dozen of DoFs. That work demonstrates that some projections enhance sampling-based planners more than others, even for systems with moderate dimensionality. Because the conformational space grows exponentially with the number of dimensions, the importance of proper guidance increases significantly for high-dimensional systems. This gives us a reason to believe that for high-dimensional systems the difference between "successful" and "unsuccessful" projections is even more drastic. In [26] the authors find that the projection showing the best performance usually belongs to the group of randomly generated projections. However, the described conclusions cannot be applied to protein modeling without additional investigation. Proteins represent significantly larger systems than the ones considered in the above-mentioned paper. When the dimensionality of a system increases, chances of constructing a "good" low-dimensional projection randomly diminish greatly. Furthermore, the user-defined projections in the analyzed paper are built under the assumption that a projection is independent from the environment of the system. In the case of proteins, the environment is encoded by their energy landscape: it defines which parts of the protein are mostly rigid and which parts could change their shape and participate in large-scale conformational transitions. This information provides essential insight for enhancing conformational search. If an expert projection is tailored for the efficient exploration of a particular protein's energy landscape, the same projection will not benefit the investigation of another protein.

In our work, we evaluate the performance of expert-defined projections that incorporate any available information about a protein's flexibility. We demonstrate that our expert projections accelerate and enhance the exploration of the conformational space compared to the traditionally-used, randomly-generated projections. However, the question of how to generate a "good" projection automatically still remains open and is a subject for future work (see Section V).

III. Methods

A. Structured Intuitive Move Selector

Our work is based on a framework for exploring the conformational space of proteins using a sampling-based motion planning approach called Structured Intuitive Move Selector (SIMS) [27]. The main purpose of SIMS is to explore the space of low-energy conformations of a protein. For this, SIMS employs an advanced expansive sampling-based planner, and defines its main propagation procedures in terms of known protein *moves* (biophysically plausible perturbations of a protein's structure).

Protein model: SIMS encodes a protein's conformation by the vector of its backbone dihedral angles. The angles between the bonds and their lengths are considered to be constant because they change insignificantly in comparison to the variation of the torsional angles. Side chains are optimized on-the-fly by the state-of-the-art Rosetta library [28], [29]. Such model has been shown to be a good enough approximation of a protein [30] and it drastically reduces the number of considered DoFs. Taking into account the planarity of the peptide bond we restrict ω to be $180°$. Thus, for each residue of the studied protein we keep only its (ϕ, ψ) values. This model induces 2N DoFs for a protein with N residues.

Fragments: Not all residues are equally important for a large-scale transition of a protein. Often only very few flexible parts of a protein are actively involved in its motion. SIMS is designed to allow for the prioritization of the most "active" parts of a protein: the algorithm gives more computational time to the exploration of these flexible regions. For that purpose, SIMS represents a protein as a set of flexible fragments, which can be defined by an expert or automatically (from a protein's secondary structure). Each fragment consists of one or several subsets of a protein's residues. Depending on which parts of the molecule are known to be the most "active" in the studied motion, the fragments are assigned probabilities to be chosen during the sampling procedure (for details on assigning probabilities of fragment selection see [27]).

SIMS algorithm: SIMS samples the state space and grows a tree of low-energy conformations, where each edge represents a possible transition from the parent state to the child state. To increase the chances of sampling a low-energy conformation, SIMS grows its tree by randomly choosing a conformation which already belongs to the tree and trying to expand from it by slightly perturbing some of its DoFs.

The search algorithm takes start and goal states, the maximum allowed energy, the minimal distance (resolution step) as an input. There are several main steps of this algorithm (pseudo-code is presented in Algorithm 1): 1. use a planner to identify a possible parent state to expand from (line 4); 2. slightly perturb the chosen state in a specific way (propagation step; line 5); 3. compute the energy of the newly produced conformation (line 6); 4. if the energy is below a user-defined threshold, the conformation is accepted and the tree is updated accordingly, otherwise the state is discarded (lines 6-9).

Algorithm 1 Search (startState, goalState, minDist, E_{max})

```
1:  addToTree(startState)
2:  lastState $\leftarrow$ startState
3:  while distance(lastState, goalState) $>$ minDist do
4:      $parentState$ $\leftarrow$ SampleParent()
5:      $currentState$ $\leftarrow$ Propagate(parentState)
6:      if $Energy(currentState) < E_{max}$ then
7:          addToTree(currentState)
8:          lastState $\leftarrow$ currentState
9:      end if
10: end while
11: return $Tree$
```

At each propagation step (pseudo-code is presented in Algorithm 2), SIMS samples some fragment with a user-defined probability and slightly perturbs the residues of that fragment. To perturb conformations in a biologically feasible way, the framework involves the most common protein moves, such as loop motion, rigid body motion (fix one end of a loop and move the other end), energy minimization, and random perturbation. All mentioned moves (except energy minimization) are applied at the fragment level (i.e., the move affects only the residues of the chosen fragment). To implement these moves, as well as for fast and accurate energy computations, SIMS uses the Rosetta library.

Algorithm 2 Propagate ($state$)

```
$fragment \leftarrow$ SampleFragment()
$move \leftarrow$SampleMove()
$newState \leftarrow$ APPLY($state, fragment, move$)
return $newState$
```

B. Projections as Sampling Guides

In the first step, Algorithm 1 randomly samples a promising parent state for the expansion of the tree towards unexplored areas of the conformational space. This step is essential for the overall success of the algorithm. To enhance the overall exploration of a protein's conformational space we need some lever to softly bias the search procedure out of the well-sampled areas. A low-dimensional projection becomes such a lever for expansive planners. Employing a low-dimensional projection to keep track of the exploration progress allows the planner to successfully model high-dimensional and highly-constrained systems.

The procedure of projecting a conformation is performed by multiplying its initial vector by a projection matrix. The conformations of a protein with N residues have $2N$ variables: (ϕ_i, ψ_i) for each residue i. Before applying the projection, we first transform the conformation vector into a vector of sines and cosines: $(\phi_i, \psi_i) \rightarrow (\sin(\phi_i), \cos(\phi_i), \sin(\psi_i), \cos(\psi_i))$. This step transfers angular data to Euclidean space and allows reasoning about projected conformation points in terms of Euclidean distances. Finally, the produced $4N$-dimensional vector is projected into a k-dimensional subspace by multiplying it by the projection matrix of size $k \times 4N$.

The projection space is discretized into a grid of equal-sized cells. This way, a projected point falls into some cell of the k -dimensional grid. The planner keeps track of the number of conformations projected on each cell of this grid. The algorithm prioritizes cells based on the density of coverage in different parts of the projection grid. At each iteration, the planner chooses the highest-priority cell and randomly picks a state from this cell.

C. Construction of "good" and "bad" projections

The intuition behind the technique of approximating a high-dimensional space with a low-dimensional projection is inspired by the Johnson-Lindenstrauss theorem [31]. This theorem states that distances between points in the initial n-dimensional space can be estimated with $(1 + \epsilon)$ distortion by the distances between the corresponding points embedded into a $\log(n/\epsilon^2)$-dimensional subspace. However, in the case of protein modeling, the dimensionality of the employed projection is usually much less than $\log(n)$. In general, the computational cost of maintaining a projection as well as the required memory resources grow exponentially with the number of dimensions. Therefore, in most cases the projection has no more than 10 dimensions; most often just 2 or 3. In many cases, such dimensionality reduction represents a reasonable approximation of the initial molecular system. In the particular case of protein modeling (as opposed to modeling a robotic articulated chain), there is an additional factor justifying the usage of a low-dimensional projection: even though proteins represent extremely high-dimensional systems, only very few of their "effective" DoFs are involved in large-scale motions [12], [13]. Therefore, the projection constructed from the few vectors corresponding to these flexible parts of the protein could represent a good approximation of the initial high-dimensional conformational space.

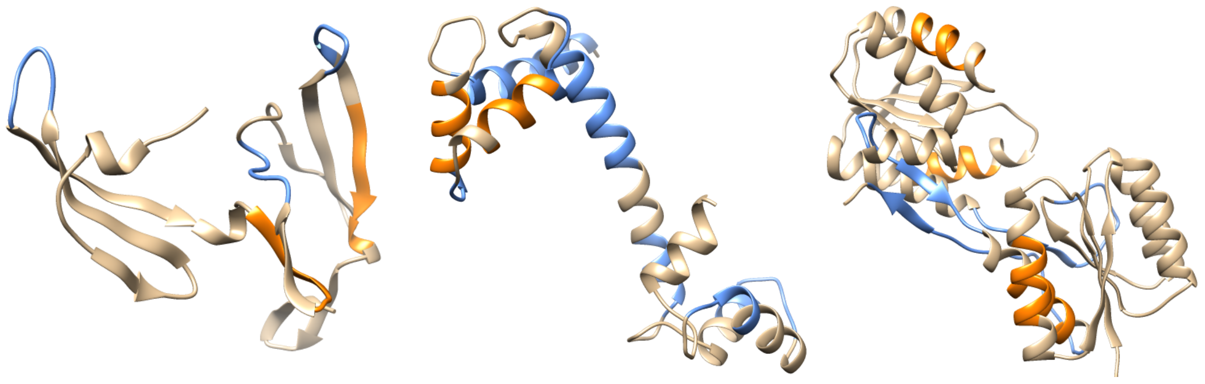
In this paper we propose a methodology to choose the rows of a projection matrix based on simple biological intuitions about the studied protein. In the next section we demonstrate that a projection designed this way represents a good approximation of the initial high-dimensional system.

To construct an expert projection, we first identify the main flexible regions of the considered protein. Second, we try to predict how correlated these regions are. If some parts of a protein move mostly in a correlated way, we will encode them together into one of the projection's dimensions (instead of putting them into different dimensions of the projection matrix). Getting such biological insights about the studied protein involves the use of information available in the literature, analysis of available datasets of conformations, as well as visual inspection of the protein's secondary structure. After identifying which regions should be present in the projection matrix and how they should be coupled, we are ready to build the matrix. In each row we assign non-zero values only to the variables encoding regions that should be coupled. The rows are then normalized. This construction process ensures the orthonormality of the produced matrix (which is important to preserve relative distances in the projection space).

By construction, an expert projection differs from a random projection in the way that it employs only some predefined groups of residues; a random projection uses *all* residues but with different weights. Thus, to ensure the comprehensive analysis of the generated expert projections, we also build "misguided" projections. A misguided projection has the same nature as an expert one: in each row it encodes only some groups of residues. In contrast to an expert projection, the residues of a misguided projection are chosen in some protein's parts that are anticipated to be mostly rigid. With such construction, the subspace of a misguided projection should be mostly orthogonal to the subspace of the protein motion. Because of that, a misguided projection cannot approximate the conformation space well and is not expected to enhance the exploration process.

IV. Experiments

The main goal of our experiments is to investigate whether we can improve the process of conformational sampling by defining a "good" low-dimensional projection. We design expert projections that take into account biological insights about a given protein. More specifically, we use the main



(a) Cyanovirin-N (CVN) (b) Calmodulin (CaM) (c) Ribose-Binding Protein (RBP)

Fig. 1. The three proteins involved in our experiments. Blue and orange areas indicate the residues involved in the expert and misguided projections, respectively.

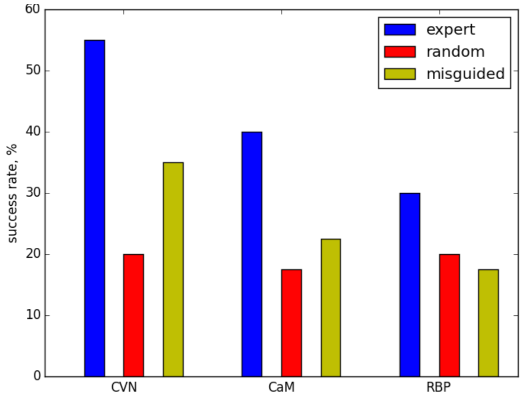


Fig. 2. Success rates associated to the three projection types: percentage of runs (among 40) that successfully found a feasible transition between start and goal states using a particular type of projection (expert, random or misguided) for CVN, CaM, and RBP within a 24 hour time limit.

"active" residues of the protein. We compare the average performance of the planner and the quality of its exploration when it uses an expert projection, as opposed to a random projection or a misguided projection.

For our experiments, we chose three well-studied protein systems with two known stable states: Cyanovirin-N, Calmodulin, and Ribose-binding protein. For each protein, we carried out a series of conformational searches to find possible low-energy transitions between these stable states (see Section IV-B). This experiment allowed us to evaluate the influence of the various projections on the success rate of the planner and on its average runtime. To assess their influence on the search-space coverage achieved by the planner, we also performed a series of conformational searches involving a single stable state (see Section IV-C). All experiments were performed using the SIMS framework [27] which internally uses the KPIECE planner [8].

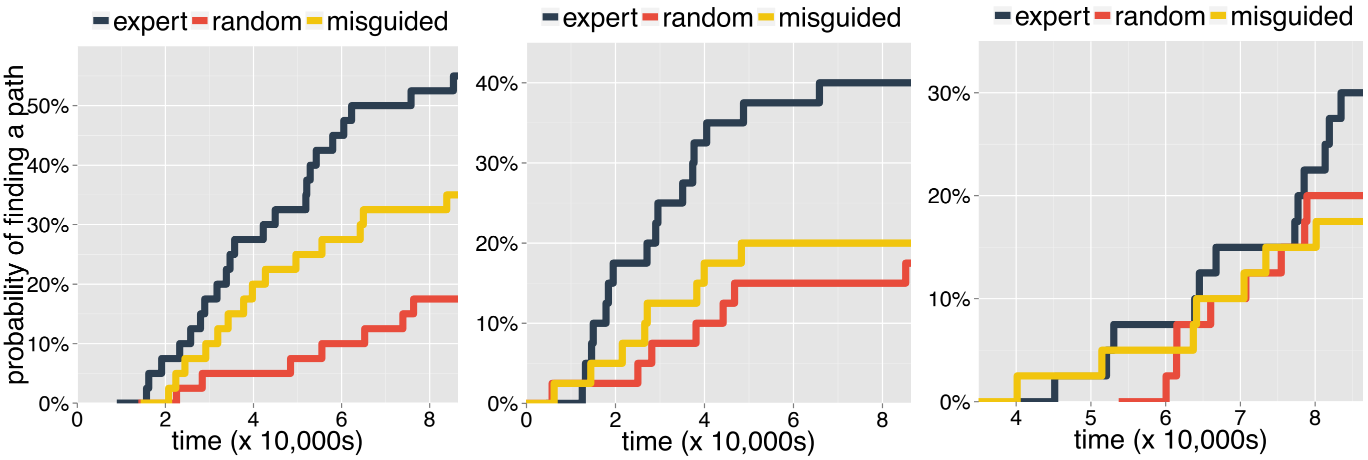
A. Studied Proteins

Cyanovirin-N: Cyanovirin-N (CVN) [32] is a bacterial protein with 101 residues, which corresponds to 202 DoFs in our framework. It demonstrates an antiviral activity towards several viruses including the human immunodeficiency virus (HIV). CVN is known to exist in two stable states, which can be found together in solution. To switch between these states, CVN goes through a domain swapping process, which involves a large-scale motion (the RMSD distance between the start state, PDB 2EZM, and the goal state, PDB 1L5E, is 17Å) via the correlated activity of three separate loop regions: residues 24-28, residues 50-55, and residues 75-80.

Based on this knowledge, we generated a 3-dimensional expert projection matrix. Each row of this matrix encodes one of the mentioned loop regions by setting only the elements corresponding to this loop's residues to non-zero values (represented by the blue regions in Fig. 1a). The misguided projection also has 3 dimensions: the first two dimensions encode residues 40-45 and residues 83-88, respectively (represented by the orange regions in Fig. 1a), and the last dimension encompasses all the other residues. Residues 40-45 and 83-88 correspond to the middle parts of beta-strands, which are likely to be inactive during the transition because of hydrogen bonds.

Calmodulin: Calcium-loaded Calmodulin (CaM) [33] is a middle-sized protein consisting of 144 residues (encoded by 288 DoFs in our framework). CaM is a calcium-binding protein involved in interactions between calcium ions and various target proteins. CaM exists in an open state (PDB 1CLL), and a closed state (PDB 1PRW) that are far apart from each other: the distance is about 16Å. The transition is known to happen mainly through unfolding of the middle part of the central helix.

Based on this information, we constructed a 2-dimensional expert projection matrix. The first dimension contains the active residues of the central hinge (residues 67-80) [33]. The second dimension encodes regions of remaining active loops and some alpha helices involved in the transition (residues 5-20, 35-41, 52-57, 87-93, 107-116, 126-129). The misguided projection is generated from the residues of the alpha helices that are not involved in the main motion: both, the first and the



(a) CVN

(b) CaM

(c) RBP

Fig. 3. Probability of finding a solution path as a function of time for each of the three projection types, for CVN, CaM, and RBP. Dark blue color is associated with an expert projection; red color - with a random projection; and yellow color - with a misguided projection.

second, rows contain residues 30-35, 47-52 (but with different signs in half of the values to ensure orthonormality of the matrix).

Ribose-binding protein: Our last system, Ribose-binding protein (RBP) [34], [35], is larger than CVN and CaM: it has 271 residues, which induces 542 DoFs in our model representation. RBP consists of two domains connected by three loop regions which form a hinge. The open conformation (PDB 2DRI) and the closed conformation (PDB 1URP) of this protein are only 4Å apart, but the transition between them requires a correlated motion of the three loop regions in the main hinge.

For this system we created a 2-dimensional expert projection encoding the three loop regions of the hinge connecting the two domains. The first row contains two loop regions (residues 91-104 and 226-237). The second row corresponds to the third loop region (residues 253-269). Such choice of a placement of the flexible parts in the projection is made because the third loop region belongs to the very tail of the protein, and thus has more freedom for motions, whereas the first two loops are more constrained to move in a correlated way. The misguided projection is constructed from the residues of several alpha helices as follows: the first row contains residues 19-26 and 241-248; the second row contains residues 140-147 and 168-175.

B. Performance Improvement

For each protein and each type of projection we performed 40 runs of a conformational search aimed at finding a feasible transition between a given pair of start and goal conformations. Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24 hour time limit.

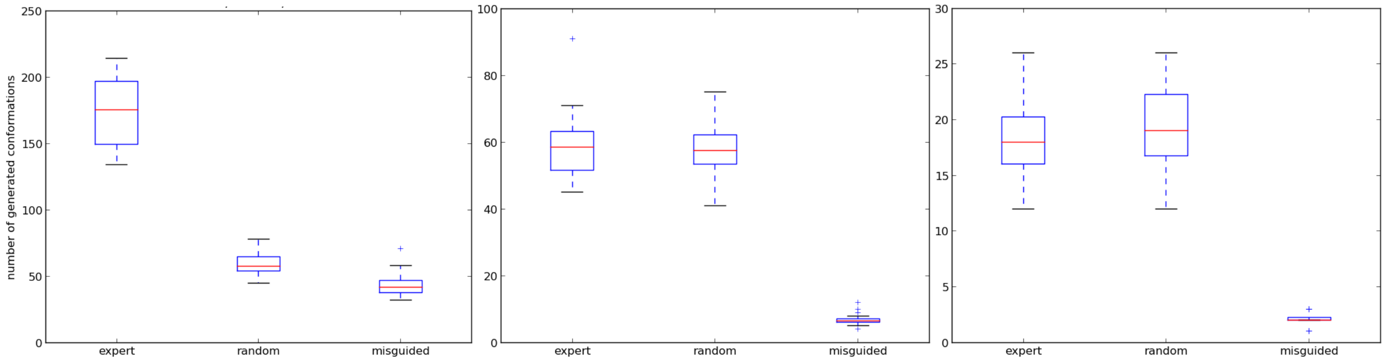
For each protein, we compared the success rates of the planners involving the expert, random, or misguided projections, respectively (see Fig. 2). We define the success rate of a projection as the percentage of runs (among 40) that successfully find a feasible transition using that projection within a 24 hour time limit. For CVN, the planner with the expert projection was 2.8 times more likely to find a solution than the planner with a random projection, and 1.4 times more likely than the planner with the misguided projection. Similar results were obtained for CaM: the expert projection was successful 2.3 times more often than the random projection and 1.8 times more often than the misguided projection. For RBP, the expert projection was 1.5 times more successful than the random one, and 1.7 times than the misguided one. Therefore, despite the differences between the three conformational search problems, the expert projections demonstrate a consistent improvement over the random and misguided projections, in terms of success rate.

Fig. 3 shows the probabilities of finding a solution path as a function of time for each of the three projection types. The success probabilities corresponding to the final time step, 86400 seconds (24 hours), in Fig. 3 are the values of the overall success rates presented in Fig. 2. Fig. 3 illustrates that, even though, the expert projection does not demonstrate a significant improvement for RBP, it still performs as well or better than the random projection. Moreover, for the other studied proteins, the expert projection consistently has a higher probability to find a solution in a given time. As computational time is a very limited resource, especially for modeling large proteins, the usage of the expert projection can significantly benefit the simulations.

C. Exploration Coverage Improvement

Another way to quantify the influence of the projection on the process of conformational search is to quantitatively assess the amount of explored projection space. We are interested in increasing the volume of the explored *projection* space, because this translates into enlarging the volume of the explored *conformational* space.

The projection space can be represented as a grid divided into cells. The number of non-empty cells in the grid serves



(a) CVN (b) CaM (c) RBP

Fig. 4. Average number of projection cells explored during the conformational search with each type of projection within 24 hours for CVN, CaM, and RBP.

as a measure of the coverage of this space. It is important to note that a small number of explored projection cells does not necessarily indicate a bad exploration of the conformational space. On the other hand, a large number of non-empty cells is an indicator of good conformational space coverage.

To compare the projection space coverage produced by the planners using different types of projection, we performed a second experiment involving another kind of conformational search. The purpose of this search is an extensive exploration of the conformational space of a protein starting from a given state (in other words, no goal state is involved). This way, we are not exploring only the protein flexibility inherent to a single transition, but the overall flexibility of the protein. In this experiment, we performed 40 runs of this conformational search for each protein and each type of projection. Each experiment was held on a single thread of quad core 2.4 GHz Intel Xeon (Nahalem) CPUs with a 24 hour time limit. All runs for CaM and RBP generated a similar number of conformations (on average 2500 for CaM; 1500 for RBP). For CVN, runs with the expert and random projections generated twice as many conformations as runs with the misguided projection (on average, about 6500–7000 for the expert and random projections, and 3200 for the misguided one).

Fig. 4 illustrates the average number of projection cells explored during the conformational search with each type of projection. For CVN, the searches with the expert projection discovered more than twice as many projection cells as the searches with the random or misguided projections. For CaM and RBP, runs with the expert projections explored a volume of projection space similar to what runs with the random projections did, but much greater than what runs with the misguided projections did: the expert projection generated 12 times more cells than the misguided projection did for CaM protein, and 6 times more cells for RBP.

Even though the expert projections described in Section IV-A does not incorporate all flexible parts of the studied proteins (they employ only the flexible parts that are anticipated to be involved in the transition between the start and goal states), the exploration coverage they produce is at least as good as the one produced by a random projection, and sometimes better, see Fig. 4. The poor exploration coverage produced by the misguided projections was expected. The misguided projections were designed specifically to focus mostly on rigid parts of a given protein. As a result, the generated low-energy conformations of the studied protein are likely to be distributed along the directions that are mostly orthogonal to its projection space. In this set of experiments, a random projection demonstrates relatively good performance, especially for flexible proteins. In the case of flexible proteins, there is a higher chance to randomly generate a projection that induces good exploration of the projection space because almost any combination of residues could be involved in *some* motion.

V. Conclusions and future work

In this paper, we have investigated the problem of improving the exploration of the conformational space of a protein. The framework we use for protein conformational sampling is based on robotics-inspired expansive path planning algorithms. These algorithms use a low-dimensional projection to guide their search in the high-dimensional conformational space. Even though the definition of this projection is essential to ensure good performance of the planning algorithms, little work had been devoted to this problem.

Our contribution consists of proposing a methodology to define "good" projections that accelerate the conformational search and improve the exploration coverage. Using the biological knowledge available for a given protein, it is possible to define a so-called "expert" projection that can efficiently guide the search through the high-dimensional conformational space of this protein. We have evaluated the use of such expert projections for three middle-sized proteins. We have shown that our expert projections perform consistently better than randomly-defined or poorly-defined (so-called "misguided") projections. Our results show that using the expert projection increases the success rate of the planning algorithm at finding a transition pathway between two conformations, and improves computational runtime. Furthermore, using an expert projection allows the planning algorithm to produce a better coverage of the conformational space.

As part of our future work, we want to compare these ”expert” projections with projections generated automatically by various methods, such as NMA, Principal Component Analysis (PCA), or graph-theory-based rigidity analysis. We are also planning to analyze the influence of the dimensionality of the projections. Our future research will concentrate on generating good projections automatically. A possible idea could be to use the decomposition of a protein into fragments corresponding to its flexible regions. A version of such decomposition is already incorporated into the SIMS framework (see Section III-A), and can be performed automatically (based on the secondary structure of a protein) or by an expert. The fragments represent the parts of a protein which are the most flexible and most likely to be involved in a transition. As such, they naturally define the regions that should be included in a “good” guiding projection.

Improving protein conformational sampling by defining a successful projection can open new horizons for studies of proteins by enabling modeling of larger proteins, such as viruses with several thousands of residues.

Acknowledgment

This work has been supported in part by NSF ABI 0960612, NSF CCF 1423304, and Rice University Funds. Experiments were run on equipment that is supported in part by the Data Analysis and Visualization Cyberinfrastructure funded by NSF under grant OCI-0959097, and on equipment that is supported by the Cyberinfrastructure for Computational Research funded by NSF under Grant CNS-0821727.

References

[1] H. Carlson, “Protein flexibility is an important component of structure-based drug discovery,” *Curr. Pharm. Design.*, vol. 8, no. 17, pp. 1571–1578, 2002.

[2] S. Adcock and J. McCammon, “Molecular dynamics: survey of methods for simulating the activity of proteins,” *Chem. Rev.*, vol. 106, no. 5, pp. 1589–1615, 2006.

[3] D. Case, “Normal mode analysis of protein dynamics,” *Curr. Opin. Struc. Biol.*, vol. 4, no. 2, pp. 285–290, 1994.

[4] E. Fuglebakk, N. Reuter, and K. Hinsen, “Evaluation of protein elastic network models based on an analysis of collective motions,” *J. Chem. Theory Comput.*, vol. 9, no. 12, pp. 5618–5628, 2013.

[5] I. Al-Bluwi, T. Siméon, and J. Cortés, “Motion planning algorithms for molecular simulations: A survey,” *Comput. Sci. Rev.*, vol. 6, no. 4, pp. 125–143, 2012.

[6] B. Gipson, D. Hsu, L. E. Kavraki, and J.-C. Latombe, “Computational models of proteins kinematics and dynamics: Beyond simulation,” *Annu. Rev. Anal. Chem.*, vol. 5, no. 1, pp. 273–291, 2012.

[7] D. Hsu, J.-C. Latombe, and R. Motwani, “Path Planning in Expansive Configuration Spaces,” *Int. J. Comput. Geom. Ap.*, vol. 9, no. 4-5, pp. 495–512, 1999.

[8] I. A. Şucan and L. E. Kavraki, “Kinodynamic motion planning by Interior-Exterior Cell Exploration,” in *Algorithmic Foundation of Robotics VIII*, G. Chirikjian, H. Choset, M. Morales, and T. Murphey, Eds. Springer Berlin Heidelberg, 2010, vol. 57, pp. 449–464.

[9] A. E. García, “Large-amplitude nonlinear motions in proteins,” *Phys. Rev. Lett.*, vol. 68, pp. 2696–2699, 1992.

[10] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. Phillips, “Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements,” *Proteins.*, vol. 22, no. 4, pp. 311–321, 1995.

[11] L. Skjaerven, S. M. Hollup, and N. Reuter, “Normal mode analysis for proteins,” *J. Mol. Struc-theochem.*, vol. 898, no. 1-3, pp. 42–48, 2009.

[12] M. L. Teodoro, G. N. Phillips Jr., and L. E. Kavraki, “Singular value decomposition of protein conformational motions: Application to HIV-1 protease,” in *Currents in Computational Molecular Biology*. Universal Academy Press Inc., 2000, pp. 198–199.

[13] A. Amadei, A. Linssen, B. De Groot, and H. Berendsen, “Essential degrees of freedom of proteins,” *Molecular Engineering*, vol. 5, no. 1-3, pp. 71–79, 1995.

[14] D. Devaurs, A. Shehu, T. Siméon, and J. Cortés, “Sampling-based methods for a full characterization of energy landscapes of small peptides,” in *IEEE BIBM 2014*, pp. 37–44.

[15] S. Thomas, G. Song, and N. M. Amato, “Protein folding by motion planning,” *Phys. Biol.*, vol. 2, no. 4, p. S148, 2005.

[16] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran, “Geometric algorithms for the conformational analysis of long protein loops,” *J. Comput. Chem.*, vol. 25, no. 7, pp. 956–967, 2004.

[17] B. Raveh, A. Enosh, O. Schueler-Furman, and D. Halperin, “Rapid sampling of molecular motions with prior information constraints,” *PLoS Comput. Biol.*, vol. 5, no. 2, p. e1000295, 2009.

[18] D. J. Jacobs and M. F. Thorpe, “Generic rigidity percolation: The pebble game,” *Phys. Rev. Lett.*, vol. 75, pp. 4051–4054, 1995.

[19] S. Thomas, X. Tang, L. Tapia, and N. M. Amato, “Simulating protein motions with rigidity analysis,” *J. Comput. Biol.*, vol. 14, no. 6, pp. 839–855, 2007.

[20] N. Fox, F. Jagodzinski, Y. Li, and I. Streinu, “Kinari-web: a server for protein rigidity analysis,” *Nucleic Acids Research*, vol. 39, no. suppl 2, pp. W177–W183, 2011.

[21] D. Luo and N. Haspel, “Multi-resolution rigidity-based sampling of protein conformational paths,” *BCB*, pp. 786–792, 2013.

[22] A. Shehu and B. Olson, “Guiding the search for native-like protein conformations with an ab-initio tree-based exploration,” *The International Journal of Robotics Research*, vol. 29, no. 8, pp. 1106–1127, 2010.

[23] B. Olson, K. Molloy, S. F. Hendi, and A. Shehu, “Guiding probabilistic search of the protein conformational space with structural profiles,” *J. Bioinform Comput Biol.*, vol. 10, no. 03, p. 1242005, 2012.

[24] K. Molloy and A. Shehu, “Interleaving global and local search for protein motion computation,” in *Bioinformatics Research and Applications*, ser. Lecture Notes in Computer Science, R. Harrison, Y. Li, and I. Mndoiu, Eds. Springer International Publishing, 2015, vol. 9096, pp. 175–186.

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

[26] I. A. Şucan and L. E. Kavraki, “On the performance of random linear projections for sampling-based motion planning,” in *IEEE/RSJ*, 2009, pp. 2434–2439.

[27] B. Gipson, M. Moll, and L. E. Kavraki, “SIMS: A hybrid method for rapid conformational analysis,” *PLoS ONE*, vol. 8, no. 7, p. e68826, 2013.

[28] R. Das and D. Baker, “Macromolecular modeling with Rosetta,” *Annu. Rev. Biochem.*, vol. 77, no. 1, pp. 363–382, 2008.

[29] K. W. Kaufmann, G. H. Lemmon, S. L. Deluca, J. H. Sheehan, and J. Meiler, “Practically useful: what the Rosetta protein modeling suite can do for you,” *Biochemistry.*, vol. 49, no. 14, pp. 2987–98, 2010.

[30] M. Levitt, “A simplified representation of protein conformations for rapid simulation of protein folding.” *J. Mol. Biol.*, vol. 104, no. 1, pp. 59–107, 1976.

[31] J. L. William Johnson, “Extensions of lipschitz mappings into a hilbert space,” *Contemporaty Mathematics*, vol. 26, pp. 189–206, 1984.

[32] I. Botos, B. R. O’Keefe, S. R. Shenoy, L. K. Cartner, D. M. Ratner, P. H. Seeberger, M. R. Boyd, and A. Wlodawer, “Structures of the complexes of a potent anti-HIV protein Cyanovirin-N and high mannose oligosaccharides.” *J. Biol. Chem.*, vol. 277, no. 37, pp. 34 336–42, 2002.

[33] N. J. Anthis, M. Doucleff, and G. M. Clore, “Transient, sparsely populated compact states of Apo and calcium-loaded Calmodulin probed by paramagnetic relaxation enhancement: Interplay of conformational selection and induced fit,” *J. Am. Chem. Soc.*, vol. 133, no. 46, pp. 18 966–18 974, 2011.

[34] A. J. Björkman, R. A. Binnie, H. Zhang, L. B. Cole, M. A. Hermodson, and S. L. Mowbray, “Probing protein-protein interactions. The ribose-binding protein in bacterial transport and chemotaxis.” *J. Biol. Chem.*, vol. 269, no. 48, pp. 30 206–11, 1994.

[35] A. J. Björkman and S. L. Mowbray, “Multiple open forms of ribose-binding protein trace the path of its conformational change.” *J. Mol. Biol.*, vol. 279, no. 3, pp. 651–64, 1998.