Computing Transition Paths in Multiple-Basin Proteins with a Probabilistic Roadmap Algorithm Guided by Structure Data

Tatiana Maximova[1], Erion Plaku[2,*], and Amarda Shehu[1,3,4*]
[1]*Department of Computer Science, George Mason University,*
[2]*Department of Computer Science, The Catholic University of America,*
[3]*Department of Bioengineering,* [4]*School of Systems Biology, George Mason University*
tmaximov@gmu.edu, plaku@cua.edu, amarda@gmu.edu
∗*Corresponding Authors*

***Abstract*—Proteins are macromolecules in perpetual motion, switching between structural states to modulate their function. A detailed characterization of the precise yet complex relationship between protein structure, dynamics, and function requires elucidating transitions between functionally-relevant states. Doing so challenges both wet and dry laboratories, as protein dynamics involves disparate temporal scales. In this paper we present a novel, sampling-based algorithm to compute transition paths. The algorithm exploits two seminal ideas. First, it leverages known structures to initialize its search and define a reduced conformation space for rapid sampling. This is key to address the insufficient sampling issue suffered by sampling-based algorithms. Second, the algorithm embeds samples in a nearest-neighbor graph where transition paths can be efficiently computed via queries. The algorithm adapts the probabilistic roadmap framework that is popular in robot motion planning. In addition to efficiently computing lowest-cost paths between any given structures, the algorithm allows investigating hypotheses regarding the order of known structures in a transition event. This novel contribution is likely to open up new venues of research. Detailed analysis is presented on multiple-basin proteins of relevance to human disease. Multiscaling and the AMBER ff12SB force field are used to obtain energetically-credible paths at atomistic detail.**

I. Introduction

While it is now known that protein dynamics is complex [1], it is exquisitely exploited for participation in various molecular recognition events in the cell [2]. In many proteins, the energy landscape is rich in broad and deep minima, also known as basins, in which a perpetually-fluctuating protein dwells long enough to participate in molecular recognition events [3]. Such basins correspond to thermodynamically-stable and meta-stable structural states, and proteins switch/transition between these states to modulate their biological function [4]. Elucidating such transitions is key not only to a detailed characterization of protein function, but also to drug and sensor design, and other protein engineering applications [5], [6].

Elucidating transitions of a protein between stable and meta-stable states is challenging in the wet laboratory. Protein dynamics involves disparate temporal scales; while typical atomic oscillations occur in the femto-pico second scale, transitions between stable and meta-stable states may occur in the micro-milli second scale, as a protein needs to gain enough kinetic energy to cross the energy barrier typically separating basins corresponding to stable and meta-stable states in the energy landscape. The main issue with catching a protein in the act and seeing the series of structures it uses in a transition event in the wet laboratory is due to the fact that the protein may spend a very long time diffusing in a (broad and deep) basin before making a sudden jump to another basin. While single-molecule wet-laboratory techniques have made great strides in revealing transitions [7], in principle, wet-laboratory techniques cannot obtain a complete picture of protein dynamics, as dwell times at successive structural states in a transition may be too short to be detected in the wet laboratory.

Neither wet- nor dry-laboratory techniques can on their own span all spatial and temporal scales in protein dynamics [8]. The presence of disparate temporal scales challenges methods that simulate dynamics (known as Molecular Dynamics – MD – methods) by iteratively solving Newton's equation of motion on a finely discretized time scale [9]. Other methods that instead navigate the energy landscape via biased random walks (known as Monte Carlo – MC methods) have to address the multiple minima issue; protein energy landscapes are rich in both shallow and deep minima (manifesting in the disparate temporal scales). Sampling-based methods, while in principle promise a higher exploration capability, have to rapidly get out of local minima so that the random walks reach desired states within practical computational budgets [10]. The presence of local minima often confines sampling-based methods to specific regions of the search space, resulting in insufficient sampling.

In this paper, we propose a novel, sampling-based algorithm that circumvents insufficient sampling by leveraging known, experimental structures of a protein to restrict sampling in a space of a reasonable number of dimensions and on regions of relevance for transition events. Known

structures are used both to define a reduced (conformation) variable space and to initialize an iterative sampling process. While the algorithm samples in a reduced space, it employs multiscaling, lifting conformations/samples in a higher-dimensional, structure space and then improving them with the AMBER ff12SB force field to obtain energetically-credible transition paths at an atomistic level of detail.

The proposed algorithm is not confined to computing one path between only a pair of given structures from one run (which is what the majority of related methods do) but is able to compute various paths between any pair of known stable and meta-stable structural states of a protein from one run within a practical computational budget (a few hours to a few days on one CPU for medium-size proteins up to 166 amino acids). The ability to compute various paths is due to the fact that the algorithm adapts the well-known probabilistic road map (PRM) framework that is a cornerstone of algorithmic robot motion planning. From now on, we will refer to the algorithm as `SoPriM` for Structure-guided Roadmap-based Protein Transition Modeling.

An additional contribution of `SoPriM` is the computation of tours that explore hypotheses regarding the position of known structures in a transition event. This new feature allows comparing lowest-cost paths to lowest-cost tours that go through a user-specified set of experimental structures, and then categorizing experimental structures as on- or off-pathway intermediates. The cost associated with a path measures the amount of work, in the thermodynamic sense, needed for a transition. In this way, the lowest-cost path is that of minimum work and can credibly represent a transition path. The additional computation of tours allows obtaining more paths of possibly higher costs but with differences that can be surpassed via thermal fluctuations at room temperature. Obtaining an ensemble of paths allows addressing the stochastic nature of protein transitions.

We first place the proposed `SoPriM` algorithm in the context of related work before describing it in section II. Detailed analysis is presented in section III on three multiple-basin proteins of relevance to human biology and disease. Conclusions and future prospects follow in section IV.

Related Work

Analogies between protein and robot motions have been exploited to model protein dynamics with robotics-inspired methods. A detailed review of these methods is provided in [10]. In summary, one of the main challenges is to determine an effective set of variables that define the conformation (search) space of interest in modeling a desired transition (a conformation is an instantiation over the variables selected to represent a protein structure). Individual dihedral angles are used to model unfolding in protein chains no longer than 100 amino acids, whereas system-specific insight and structure analysis are used to bundle individual variables together in fragments or reveal fewer, collective variables to capture larger transitions.

Robotics-inspired methods are tree-based or (PRM-) roadmap-based. Tree-based methods grow a tree in conformation space from a start to a goal conformation representing the structures desired to be bridged by a transition. The growth of the tree is biased so the goal conformation can be reached in reasonable time. As a result, tree-based methods are efficient but limited in their sampling and cannot be employed to reveal multiple paths between any two conformations. They are known as single-query methods, as they can only answer one start-to-goal query at a time. Even running them multiple times is not desirable, as the biasing results in high path correlations.

Roadmap-based methods support multiple queries, as they embed sampled conformations in a nearest-neighbor graph/roadmap. Several challenges exist with broadening their scope beyond unfolding of small proteins. Focusing sampling to regions of interest is difficult with no a priori information. A preliminary adaptation for modeling transitions in proteins of any size employs a diverse set of perturbation operators under a probabilistic scheme to sample conformations and several geometric and energetic constraints to restrict samples near given functionally-relevant structures. This adaptation has a high computational demand of up to several hundred hours on a CPU for a few paths [11].

The roadmap-based algorithm proposed here leverages experimental structures of a protein to define the conformation space of relevance for sampling. While years ago the reliance on experimental structures would be a limitation, nowadays over hundred of thousand structures exist in the Protein Data Bank (PDB) [12]. For proteins of importance to human disease and biology, significant resources in wet laboratories have resulted in diverse stable and meta-stable structures of wildtype and variant sequences. The proposed algorithm exploits such structures to expedite sampling.

Another challenge with adapting roadmap-based algorithms for protein transitions relates to edge realization. When edges connect conformations far away, a local planner is needed to reveal intermediate conformations. This is in effect another transition modeling instance and can tax computational resources. The proposed `SoPriM` algorithm addresses this challenge in the way it samples conformations; moreover, nearest neighbors in the roadmap pass a distance constraint so that an edge represents a motion expected to occur within thermal fluctuations.

II. Methods

As a roadmap-based algorithm, `SoPriM` consists of three stages: conformation sampling, roadmap building, and roadmap querying, as shown in Alg. 1. The result of the sampling stage is an ensemble of conformations, denoted by $\mathcal{C}$, that provides a discrete representation of the conformation space expected to be of relevance for the transition event.

In roadmap building, a graph $\mathcal{R} = (\mathcal{C}, \mathcal{E})$ is constructed by connecting each conformation $c \in \mathcal{C}$ to several of its nearest neighbors. In roadmap querying, costs associated with roadmap edges are used to obtain a set of lowest-cost paths that connect the given start and goal conformations by going over all the possible subsets of a set of specified structures, referred to as landmarks and denoted by $\mathcal{L}$. The rest of the section describes each stage in more detail.

A. Conformation Sampling

The input to SoPriM is a set Ω of experimental structures of a protein, collected and curated as described in section III under *Data Preparation*. These structures are projected to the space of considered variables to obtain conformations with which to seed the growing ensemble $\mathcal{C}$. While uniform sampling has worked well when applying roadmap algorithms to robot motion-planning problems [13], it is impractical when dealing with high-dimensional conformation spaces, since the sampled conformations are highly likely to have high energies. To effectively populate the roadmap, SoPriM relies on a low-dimensional conformation space on which iterative application of a selection and a perturbation operator result in new samples/conformations that satisfy geometric and energetic constraints. The selection operator selects a conformation from the current $\mathcal{C}$ ensemble. The perturbation operator modifies the selected conformation to yield a new one, which is subjected to a local improvement operator before being added to the $\mathcal{C}$ ensemble. This process is repeated until at least a user-specified minimum number of conformations have been sampled. As described later in the section, roadmap sampling is interleaved with roadmap building until the start, goal, and landmark conformations are connected, i.e., belong to the same graph component in the roadmap $\mathcal{R} = (\mathcal{C}, \mathcal{E})$, or a maximum number of conformations have been obtained.

1) Defining the Conformation Space for Sampling: SoPriM leverages the set Ω of known structures of a protein. These structures are stripped down to their CA atoms and are subjected to Principal Component Analysis (PCA) [14] in order to reveal collective variables (principal components – PCs) over which to define the conformation space for sampling. This is motivated by prior work on evolutionary algorithms that employ PCA to find basins in energy landscapes of proteins [15], [16]. PCA and other linear dimensionality reduction techniques are shown to be effective for many multiple-basin proteins of relevance to human biology and disease [15].

The CA traces of the experimental structures are first aligned to a reference CA trace (arbitrarily set to the first one) using the optimal superimposition process employed when identifying least root-mean-squared-deviation (lRMSD) between two structures [17]. The purpose for the alignment is so that PCA does not capture trivial structural variations due to rigid-body motions. An average trace is then computed and subtracted from all traces so that a centered matrix of structural variations can be defined. The matrix is subjected to the *dgesvd* routine in LAPACK [18] to obtain a singular value decomposition $X = U\Sigma V^T$. Rows of the U matrix contain the new axes (PCs), rotated to identify the axes of highest variance. The variance of the data along each axis is given by its corresponding eigenvalue, which can be calculated by squaring the singular values contained in the diagonal of the Σ matrix.

Ordering the PCs by the variance they capture allows identifying a few (if PCA has been effective) that cumulatively capture a desired total variance. In this paper and related employments of PCA, a 90% cutoff is used. When PCA is effective, this cutoff can be reached by a number of PCs that is a significant reduction over the original dimensionality of the space. For instance, for all the proteins considered here, the original dimensionality is over 300 (number of x, y, z coordinates of CA atoms), whereas no more than 25 PCs

Algorithm 1 SoPriM

Input: Ω: initial ensemble of conformations
$c_s, c_g \in \Omega, \mathcal{L} \subset \Omega$: start, goal, and landmark conformations
n_{min}, n_{max}: min/max number of conformations in roadmap
n_{add}: number of conformations to add to roadmap at each stage
k, r: number and range for nearest neighbors
$\mathcal{G}$: 2D-grid, i.e., $min_{x,y}, max_{x,y}$, number of rows and columns
$-\delta_{min}, \delta_{max}$: min/max perturbation step
Output: a set of paths $\mathcal{P} = \{path_{\mathcal{S}} : \mathcal{S} \subseteq \mathcal{L}\}$ over the roadmap $\mathcal{R} = (\mathcal{C}, \mathcal{E})$ where $path_{\mathcal{S}}$ is the lowest-cost path in $\mathcal{R}$ that starts at c_s, ends at c_g, and reaches each conformation in $\mathcal{S}$

define $\rho(c_i, c_j) = ||\text{PCPROJECTION}(c_i) - \text{PCPROJECTION}(c_j)||_2$
define $\text{COST}(c_i, c_j) = \max\{\text{SCORE}(c_j) - \text{SCORE}(c_i), 0\}$

$\mathcal{R} = (\mathcal{C}, \mathcal{E}) \leftarrow (\emptyset, \emptyset)$; $\Gamma \leftarrow \emptyset$; $\mathcal{P} \leftarrow \emptyset$; $n \leftarrow n_{min}$; $i \leftarrow 1$
for each $c \in \Omega$ **do** ADDCONFORMATION$(\mathcal{R}, \Gamma, c)$
repeat
  while $|\mathcal{C}| < n$ **do**
    $\gamma \leftarrow$ SELECTGRIDCELL(Γ)
    $c \leftarrow$ SELECTCONFORMATION(γ)
    $c_{new} \leftarrow$ GENERATESUCCESSOR$(c, \text{RAND}(\delta_{min}, \delta_{max}))$
    UPDATESTATISTICS(γ, c, c_{new})
    if $c_{new} \neq$ `null` **then** ADDCONFORMATION$(\mathcal{R}, \Gamma, c_{new})$
  $n \leftarrow \min\{|\mathcal{C}| + n_{add}, n_{max}\}$
  while $i \leq |\mathcal{C}|$ **do**
    $c \leftarrow i$-th conformation in $\mathcal{C}$; $i \leftarrow i + 1$
    neighs $\leftarrow$ NEARESTNEIGHBORS$(\mathcal{R}, \rho, c, k, r)$
    for $c' \in$ *neighs* **do** $\mathcal{E} \leftarrow \mathcal{E} \cup \{(c, c'), (c', c)\}$
until CONNECTED$(\mathcal{R}, c_s, c_g, \mathcal{L}) =$ `true` **or** $|\mathcal{C}| > n_{max}$
for each $\mathcal{S} \subseteq \mathcal{L}$ **do**
  $path_{\mathcal{S}} \leftarrow$ SHORTESTPATH$(\mathcal{R}, \text{COST}, c_s, c_g, \mathcal{S})$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \{path_{\mathcal{S}}\}$
return $\mathcal{P}$

local procedure ADDCONFORMATION$(\mathcal{R}, \Gamma, c_{new})$
$\mathcal{C} \leftarrow \mathcal{C} \cup \{c_{new}\}$
$\langle p_1 p_2 \ldots p_d \rangle \leftarrow$ PCPROJECTION(c_{new})
$\gamma \leftarrow$ LOCATEGRIDCELL(p_1, p_2)
if $\gamma \notin \Gamma$ **then** $\Gamma \leftarrow \Gamma \cup \{\gamma\}$
INSERT(γ, c_{new})

are needed to preserve 90% of the original data variance. This effectively results in a reduced search space, where the conformations are points with coordinates on the top-selected PCs (axes).

2) Generating a Successor via Perturbation and Improvement: A new conformation, c_{new}, is obtained from a conformation $c \in \mathcal{C}$ via perturbation and local improvement (Alg. 1:7). Given c as a point in the space of the top d PCs, the perturbation operator computes c_{new} as $c + v$, where $v = \langle v_1 \dots v_d \rangle$ specifies displacements along each PC. The displacement v_1 along PC_1 is sampled uniformly at random inside a given interval $[\delta_{min}, \delta_{max}]$. In order to ensure that displacements are proportionate with the variations captured by each PC, every other displacement is computed as $v_i = v_1 \lambda_i / \lambda_1$, where λ_i is the eigenvalue of PC_i.

After the perturbation, c_{new} is subjected to a local improvement operator so a potential energy can be associated with it. First, c_{new} is lifted to an all-atom structure where the CA trace is obtained by adding c_{new} to the reference trace, the backbone is obtained via the BBQ program [19], and side chains are packed via the SCWRL4 program [20]. The resulting all-atom structure is then subjected to a standard, AMBER-recommended minimization protocol [21]. The protocol uses the ff12SB force field and *sander* to conduct 50 steps of steepest descent followed by 50 steps of conjugate gradient descent ($maxcyc = 100, ncyc = 50$). Nonbonded interactions beyond 10Å are cut off. The implicit, generalized Born solvation model is used ($igb = 1$).

The resulting structure corresponds to a local minimum in the all-atom energy surface. If the potential energy is above 0kcal/mol, the minimization is considered to have failed and `null` is returned by GENERATESUCCESSOR. Since the minimization can change CA coordinates, the all-atom structure is projected back onto the PCs to obtain final coordinates for c_{new}. This is key to controlling the accumulation of structural errors expected from iterative-based sampling. The experimental structures are subjected to the same minimization protocol to resolves unfavorable interactions often present in X-ray and NMR models.

3) Conformation Selection: Conformation selection is key to controlling sampling in conformation space. A two-dimensional, implicit grid $\mathcal{G}$ is imposed over the top two PCs (which capture $> 50\%$ of the original dynamics/variations for all proteins here) to bias sampling so the roadmap can cover the reduced conformation space. The grid $\mathcal{G}$ is also used to promote the generation of low-energy conformations. More specifically, each grid cell $\gamma \in \mathcal{G}$ keeps track of the conformations in $\mathcal{C}$ that map to it. Note that $c \in \mathcal{C}$ maps to γ if the point defined by the coordinates associated with PC_1 and PC_2 is inside γ. Moreover, a weight $w(\gamma)$ is maintained for each grid cell $\gamma \in \mathcal{G}$ as

$$w(\gamma) = \frac{e^{-minE(\gamma)\cdot\alpha}}{(nrConfs(\gamma) \cdot nrSel(\gamma) \cdot nrFailures(\gamma))^2}, \quad (1)$$

where $minE(\gamma)$, $nrConfs(\gamma)$, $nrSel(\gamma)$, and $nrFailures(\gamma)$ denote the minimum potential energy over conformations that map to γ, the number of conformations in γ, the number of times γ has been selected (Alg. 1:5), and the number of times GENERATESUCCESSOR has failed to generate a successor when using a conformation mapped to γ (Alg. 1:9, when $c_{new} = \texttt{null}$), respectively. The probability of selecting γ is then defined according to its weight as

$$prob(\gamma) = \frac{w(\gamma)}{\sum_{\gamma' \in \Gamma} w(\gamma')}, \quad (2)$$

where $\Gamma = \{\gamma' : \gamma' \in \mathcal{G} \text{ and } nrConfs(\gamma') > 0\}$ keeps track of all the non-emtpy grid cells.

In this way, SELECTGRIDCELL (Alg. 1:5) discourages cells that lead to failures, encourages cells that protrude deep in the energy landscape, and rejects cells that have been selected many times and have too many conformations in them already so as to penalize oversampling in the same region of conformation space. The α parameter is a user-defined constant to tune the importance of selecting based on energy versus the other statistics.

Once a cell γ is selected, a similar weighting function and probability distribution is used over the conformations that map to γ in order to select a conformation (Alg. 1:6) for the GENERATESUCCESSOR function , i.e.,

$$w(c) = \frac{e^{-E(c)\cdot\alpha}}{(nrSel(c) \cdot nrFailures(c))^2}, prob(c) = \frac{w(c)}{\sum_{c' \in \gamma} w(c')}. \quad (3)$$

B. Roadmap Building

To capture the connectivity of the conformation space, each $c \in \mathcal{C}$ is connected to several of its nearest neighbors according to the Euclidean distance ρ in the space of d PCs (Alg. 1:11–14). Specifically, NEARESTNEIGHBORS$(\mathcal{R}, \rho, c, k, r)$ returns at most k nearest neighbors whose distance from c is also $\leq r$. Correlation analysis between lRMSD and ρ yields a reasonable value for r (data not shown). The idea behind imposing a r constraint is to restrict edges to thermal fluctuations.

The latter stage of roadmap querying depends on the start c_s, goal c_g, and landmark structures $\mathcal{L}$ belonging to the same graph component in the roadmap $\mathcal{R}$. Therefore, the sampling and roadmap building proceed iteratively. Once a minimum number n_{min} of conformations have been sampled, conformations are then sampled in sets of n_{add}, checking for the presence of a connected component after each such set has been added to the growing ensemble $\mathcal{C}$. Hard cases where no connected component can be obtained are identified by stopping the computation when a maximum number of conformations n_{max} have been sampled. Note that the parameter n_{min} effectively gives a burn-in phase to the algorithm. If the algorithm checks every n_{add} conformations without this burn-in phase, possibly very distant neighbors can be joined through edges (when no distance criterion is added to the nearest-neighbor computation).

C. Roadmap Querying

The input to the query consists of the start c_s, goal c_g, and a set $\mathcal{L}$ of other experimental structures serving as possible intermediate structures in the sought transition event. These structures, referred to as landmarks, are part of the ensemble Ω initializing the sampling stage, so they are already in the roadmap. The objective of roadmap querying is to compute a set of paths $\mathcal{P} = \{path_{\mathcal{S}} : \mathcal{S} \subseteq \mathcal{L}\}$ where $path_{\mathcal{S}}$ is the lowest-cost path in the roadmap that starts at c_s, ends at c_g, and reaches each conformation in $\mathcal{S}$ (Alg. 1:16–18).

The cost of a roadmap edge $(c, c') \in \mathcal{E}$ is defined as

$$\text{COST}(c, c') = \max\{E(c') - E(c), 0\}, \tag{4}$$

where $E(c)$ stands for the ff12SB energy of the reconstructed structure corresponding to c (roadmap edges are directed). This definition only records uphill energetic variations which measure the amount of energy that the protein needs to accumulate through thermal vibrations to move from c to c'. The cost of a path, which is the sum of the costs of its edges, represents the total amount of energy needed for a transition event to occur. This definition of edge weight implements the concept of mechanical work, which has been shown to assess the quality of a path and thus the relevance of a lowest-cost path as a representative of the transition event better than the integral cost along the path [22].

In order to effectively obtain $\mathcal{P} = \{path_{\mathcal{S}} : \mathcal{S} \subseteq \mathcal{L}\}$, Dijkstra's algorithm is used to compute the lowest-cost path, denoted by $path(c, c')$, for every pair (c, c') where $c, c' \in \{c_s, c_g\} \cup \mathcal{L}$. Given $\mathcal{S} = \{s_1, \ldots, s_\ell\}$, $path_{\mathcal{S}}$ is computed by considering all the possible orderings of the configurations $s_1, \ldots, s_\ell$. In fact, let $s_{\pi_1}, \ldots, s_{\pi_\ell}$ denote a permutation of $s_1, \ldots, s_\ell$. Let $path(\langle s_{\pi_1} \ldots s_{\pi_\ell} \rangle)$ denote the lowest-cost path in $\mathcal{R}$ that starts at c_s, reaches $s_{\pi_1}, \ldots, s_{\pi_\ell}$ in order, and ends at c_g. Such path is obtained by concatenating $path(c_s, s_{\pi_1}), path(s_{\pi_1}, s_{\pi_2}), \ldots path(s_{\pi_{\ell-1}}, s_{\pi_\ell}), path(s_{\pi_\ell}, c_g)$. Thus, $path_{\mathcal{S}}$ corresponds to the lowest cost $path(\langle s_{\pi_1} \ldots s_{\pi_\ell} \rangle)$ over all possible permutations of $s_1, \ldots, s_\ell$. As described in the next section, the set $\mathcal{P} = \{path_{\mathcal{S}} : \mathcal{S} \subseteq \mathcal{L}\}$ is analyzed to identify experimental structures that serve as intermediates in a transition event.

D. Implementation Details

`SoPriM` is implemented in C/C++ and run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University. Compute nodes used for testing are Intel Xeon E5-2670 CPU with 2.6GHz base processing speed and 3.5TB of RAM. Different parameter values are investigated for the proteins considered here, resulting in running times from 3 to 48 hours on one CPU. `SoPriM` is run 5 times on each parameter setting in order to account for the stochastic nature of the algorithm.

Parameter Values: n_{min} is set at $3,000$, n_{max} is set at $5,000$ and n_{add} is set at 50; Different parameter values are investigated, with k ranging in $[10, 30]$, r corresponding to lRMSD of $1-3$Å, depending on the magnitude of the transition and protein size. δ_{min} is set to $-\delta_{max}$, and δ_{max} is varied in the set $\{0.25, 0.5, 1.0, 2.0, 3.0\}$.

III. Results

A. Test Cases and Data Preparation

Performance is evaluated on 3 proteins of importance to human biology and disease, (the catalytic domain of) H-Ras, the superoxide dismutase [Cu-Zn] (SOD1), and Calmodulin (CaM). H-Ras is 166 amino acids long, mediates signaling pathways controlling cell proliferation, growth and development, and switches between two structural states to regulate its activity. SOD1 is 150 amino acids long with mutations linked to Amyotrophic lateral sclerosis (ALS). CaM is 144 amino acids long and has been captured in diverse bound and unbound states in the wet laboratory.

X-ray and NMR structures collected for each of these proteins are restricted to those of sequences with no more than 3 mutations over the wildtype sequence. Structures with missing internal regions are discarded. This results in 86 structures for H-Ras, 186 for SOD1, and 697 structures for CaM. PCA is applied to each of these three datasets, and a cumulative variance of 90% is reached at 10, 25, and 10 PCs for H-Ras, SOD1, and CaM, respectively. In the interest of space, the cumulative variance profile is not shown here; the reader is referred to prior work by us on evolutionary algorithms in PC projection spaces [15].

B. Experimental Setup

To take into account variations in results from different runs of the algorithm (with the same parameters or different parameters), the results from all the runs are collected and analyzed. The paths obtained from one run are the lowest-cost path between the start and goal and the lowest-cost tours (which go over all possible non-empty subsets of the specified landmarks). The lowest-cost path over all runs is recorded, and its cost is used as a baseline. A threshold corresponding to 1.5kcal/mol per residue is then applied to extract other paths and tours (from all runs) with costs no more than the threshold above the baseline. These are visualized in projections over the top two PCs. For all the proteins studied here, the top two PCs capture at least 50% of the cumulative variance; as such, projections of samples and paths on the top two PCs can be used to draw observations.

C. Summary Analysis of Transition Paths for SOD1

Known and generated SOD1 structures and computed lowest-costs paths are shown in the PC1-PC2 embedding in Fig. 1(a). Known structures (empty cyan circles) organize in two distinct clusters/basins. The start (PDB id 4FF9, chain A) and goal (4B3E, chain A) structures are selected from the different basins to evaluate whether `SoPriM` can compute inter-basins transitions and highlight participating structures. Four landmarks are considered, drawn in orange (PDB ids shown). Generated samples connect the basins. Samples are

color-coded based on the cost of the paths in which they participate (the darker the color, the lower the cost of the path in which a conformation participates). For SOD1, all the lowest-cost paths and tours from different runs that meet the energetic threshold are identical, going through two of the four landmarks, as shown in Fig. 1(a). Tours that are forced to make use of other subsets of the four landmarks have higher costs (not shown); one can conclude that inter-basins transitions go through a high-energy barrier, and the lowest-cost path over the barrier is mediated by two distinct structures (chains A and C in PDB entry 2NNX).

Fig. 1(a) also shows the structures in the lowest-cost path (superimposed and drawn in a red-to-blue color scheme, with the start in red and the goal in blue. Residues that bind copper are in green, and those that bind zinc are in yellow. Fig. 1(a) shows that backbone fluctuations in the transition are small and predominantly in regions not directly involved in metal binding. PDB id 2NNX corresponds to the structure a disease-associated double mutant that diminishes copper binding and dimer stability; such a structure has not been captured in the wet laboratory for the wildtype. The results here suggest that this double-mutant structure may be meta-stable in the wildtype and possibly mediates the inter-basins transition. The mutations stabilize this structure and possibly slow the transition, affecting SOD1 function. This result on SOD1 points to specific structures that can be further investigated in the wet laboratory to better understand function modulation in the wildtype and mutants of SOD1.

D. Summary Analysis of Transition Paths for H-Ras

Results are shown for H-Ras in Fig. 1(b). The start (PDB id 1QRA) and goal (4Q21) structures represent the known On and Off known states/basins. Various landmarks are employed, shown in Fig. 1(b). Generated samples connect the On and Off basins and mediate the transition (a representative lowest-cost path is drawn, together with the structures involved). Many slight, but energetically-similar variations to the reported path are observed from different runs (data not shown), pointing to the existence of low-energy structures bridging the On-to-Off transition that have yet to be reported in the wet laboratory. Other paths that go through structures with PDB entry 1Q21 and nearby have higher cost. The results on H-Ras suggest a possible transition that can be probed in the wet laboratory. The structural fluctuations occurring in this transition are predominantly on the SI and SII regions involved in GTP and GDP binding.

E. Detailed Analysis of Transition Paths for CaM

Different settings are investigated for CaM to observe its transitions from an apo (unbound) state to different closed, peptide-binding states. The apo state is represented by the structure with PDB id 1CLL and used as start. Two different structures are used as goals, the ones with PDB id 2F3Y and 1NWD. One landmark is specified (the calcium-binding structure with PDB id 1CFD) to investigate the hypotheses that apo-to-closed transitions go through the calcium-binding state, where the internal helix connecting the N- and C-terminal domains partially unfolds to possibly accommodate further collapse of the domains.

Fig. 1(c1) shows all paths and tours that meet the set energetic criterion; projections of all known structures of CaM are also shown (generated samples are not drawn for clarity). The two paths with the lowest cost among all are drawn in black, and they correspond to the transition from 1CLL to 2F3Y (total cost of about 0.2kcal/mol per residue). The lowest-cost path obtained for the transition from 1CLL to 1NWD has a slightly higher cost of 0.8kcal/mol per residue (drawn in gray in Fig. 1(c1)). The lowest-cost paths and tours obtained from different runs of the algorithm that lie below 3.5kcal/mol per residue are drawn in gray, with lighter gray indicating higher cost per residue.

The information provided in Fig. 1(c1) is summarized in a schematic in Fig. 1(c2), which shows the PDB ids of the experimental structures that participate in the obtained paths. Fig. 1(c2) shows that the transitions from 1CLL to the closed, peptide-binding states may not make use of 1CFD; in fact, tours forced to go through 1CFD have a higher cost, and lower cost is obtained if the paths go through structures in the NMR ensemble with PDB id 2KOE. In fact, Fig. 1(c2) shows that the structures in this ensemble are key to the CaM transition from its apo to its peptide-binding states.

The right panel in Fig. 1(c2) shows the successive structures corresponding to the two lowest-cost paths (that do not make use of 1CFD) to the closed states. The succession of structures shows that the domain collapse, re-arrangement, and partial unfolding of the helix linker are gradual and correlated, as captured in the various structures in the NMR ensemble with PDB id 2K0E. This ensemble has been contributed to the PDB by work in [23]. The 2K0E ensemble represents the structure and dynamics of calmodulin (CaM) in the calcium-bound state (Ca(2+)-CaM) and in the state bound to myosin light chain kinase (CaM-MLCK). Analysis in [23] shows that correlated motions within the Ca(2+)-CaM state direct the structural fluctuations toward complex-like substates. This is in great agreement with the results obtained by `SoPriM` for CaM. `SoPriM` elucidates that the lowest-cost path for the transition between the apo and peptide-binding states of CaM does not make use of the Ca(2+)-bound structure reported under PDB id 1CFD but instead of other Ca(2+)-bound structures and MLCK-bound structures captured in the wet laboratory under PDB id 2K0E. While work in [23] was restricted to MLCK binding, the results obtained for CaM here suggest that the same mechanism observed in [23] prepares CaM for binding to other peptides (the C-terminal Domain of Petunia Glutamate Decarboxylase in 1NWD and the IQ domain in 2F3Y). Taken together, the results obtained here for CaM point to a general mechanism for its apo-to-closed/complexed

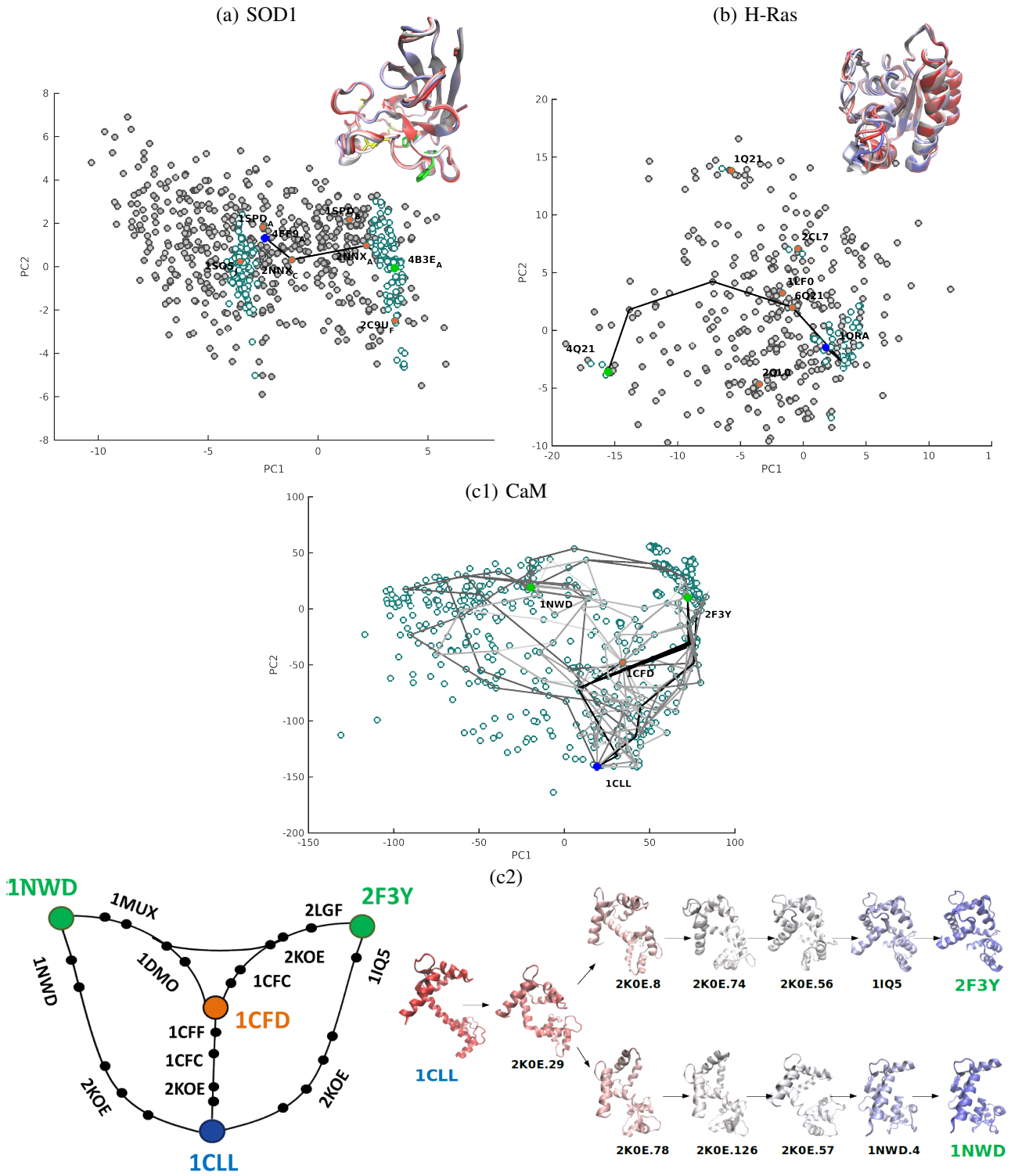

Figure 1: Top panel: Known and generated structures and lowest-cost path are shown in the PC1-PC2 embedding for SOD1 in (a) and H-Ras in (b). The start and goal are in blue and green, and landmarks are in orange. Darker-colored samples participate in lower-cost paths. Structures in the lowest-cost path are drawn, superimposed over one another. For H-Ras, new, generated conformations are needed to obtain transitions. Bottom panel: (c1) Lowest-cost path obtained for CaM is in black, and other paths are in shades of gray, with lighter shades indicating higher costs (only projections of experimental structures are drawn for clarity). (c2) Schematic summarizes lowest-cost paths obtained by `SoPriM`, showing PDB ids of known structures participating in the transitions. Successive structures in the lowest-cost paths found for the 1CLL to 2F3Y and 1CLL to 1NWD transitions are also shown. Numbers indicate model number within an NMR entry.

transition dynamics, where correlated motions within the calcium-bound state direct the fluctuations and population shift to the peptide-bound states. This result illustrates the capability of `SoPriM` to both confirm wet-laboratory work and make new discoveries.

IV. Conclusion

The definition of edge weight here is based on mechanical work. While comparison of different criteria is beyond the scope of this paper, future work will consider other criteria, such as those based on minimum resistance [24]. Other lines of investigation concern dimensionality reduction. While non-linear techniques exist, they do not allow direct sampling. While beyond the scope of this paper, several strategies can be employed to remedy this limitation. Techniques such as NMA can also be used to soften the reliance on diverse experimental structures. Future work will also focus on more detailed analysis of computed paths, as well as additional applications of the algorithm on more proteins and different variant sequences of a given protein. The latter setting will allow comparing transitions between wildtype and variants to formulate structure-based hypotheses regarding the functional impact of sequence mutations.

Acknowledgment

This work is supported in part by NSF CCF No. 1440581, NSF IIS CAREER Award No. 1144106, and the Thomas F. and Kate Miller Jeffress Memorial Trust Award. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu).

References

[1] K. Jenzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, pp. 964–972, 2007.

[2] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chem Biol*, vol. 5, no. 11, pp. 789–96, 2009.

[3] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 32, pp. 11 844–11 849, 2006.

[4] J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," *PLoS Comp Biol*, vol. 5, no. 8, p. e1000480, 2009.

[5] C. F. Wong and M. J. A., "Protein simulation and drug design," *Adv. Protein Chem.*, vol. 66, no. 1, pp. 87–121, 2003.

[6] M. Merkx, M. V. Golynskiy, L. H. Lindenburg, and J. L. Vinkenborg, "Rational design of FRET sensor proteins based on mutually exclusive domain interactions," *Biochem Soc Trans*, vol. 41, no. 5, pp. 128–134, 2013.

[7] H. M. Lee, K. S. M., H. M. Kim, and Y. D. Suh, "Single-molecule surface-enhanced Raman spectroscopy: a perspective on the current status," *Phys Chem Chem Phys*, vol. 15, pp. 5276–5287, 2013.

[8] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Veláquez-Muriel, and A. Sali, "The structural dynamics of macromolecular processes," *Curr Opin Cell Biol*, vol. 21, no. 1, pp. 97–108, 2009.

[9] R. E. Amaro and M. Bansai, "Editorial overview: Theory and simulation: Tools for solving the insolvable," *Curr. Opinion Struct. Biol.*, vol. 25, pp. 4–5, 2014.

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

[11] K. Molloy and A. Shehu, "Interleaving global and local search for protein motion computation," in *LNCS: Bioinformatics Research and Applications*, R. Harrison, Y. Li, and I. Mandoiu, Eds., vol. 9096. Norfolk, VA: Springer International Publishing, 2015, pp. 175–186.

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

[13] 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, 2005.

[14] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, 1973.

[15] R. Clausen and A. Shehu, "A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes," *J Comp Biol*, 2015, in press.

[16] R. Clausen, E. Sapin, K. A. De Jong, and A. Shehu, "Mapping multiple minima in protein energy landscapes with evolutionary algorithms," in *Genet Evol Comput Conf (GECCO)*. New York, NY, USA: ACM, July 2015, pp. 923–927.

[17] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallogr. A.*, vol. 26, no. 6, pp. 656–657, 1972.

[18] E. Anderson *et al.*, "LAPACK: A portable linear algebra library for high-performance computers," in *ACM/IEEE Conf on Supercomputing*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, pp. 2–11.

[19] D. Gront, S. Kmiecik, and A. Kolinski, "Backbone building from quadrilaterals: a fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates," *J. Comput. Chem.*, vol. 28, no. 29, pp. 1593–1597, 2007.

[20] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack, "Improved prediction of protein side-chain conformations with SCWRL4," *Proteins: Struct. Funct. Bioinf.*, vol. 77, no. 4, pp. 778–795, 2009.

[21] D. Case *et al.*, "Amber 14," University of California, San Francisco, 2015.

[22] L. Jaillet, J. Cortés, and T. Siméon, "Sampling-based path planning on configuration-cost costmaps," *IEEE Trans Robot*, vol. 26, no. 4, pp. 635–646, 2010.

[23] J. Gsponer, J. Christodoulou, A. Cavalli, J. M. Bui, B. Richter, C. M. Dobson, and M. Vendruscolo, "A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction," *Structure*, vol. 16, no. 5, pp. 736–746, 2008.

[24] S. Huo and J. Straub, "The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature," *J Chem Phys*, vol. 107, no. 13, pp. 5000–5006, 1997.