Structure-guided Protein Transition Modeling with a Probabilistic Roadmap Algorithm

Tatiana Maximova, Erion Plaku, and Amarda Shehu, *Members, IEEE*

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 main 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 experimentally-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 ff14SB force field are used to obtain energetically-credible paths at atomistic detail.

Index Terms—Protein structure, transitions, energy landscape, basins, robotics-inspired search, motion computation.

1 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 due to thermal energy 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 landscape. While single-molecule wet-laboratory techniques have made great strides in revealing transitions [7], in principle, wet-laboratory techniques cannot obtain a complete picture, 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 Molecular Dynamics (MD) methods that simulate dynamics 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 have to address the multiple minima issue; protein energy landscapes are rich in both shallow and deep minima (manifesting in the disparate temporal scales). These sampling-based methods, also known as Monte Carlo (MC) methods, while in principle promising of a higher exploration capability, have to rapidly escape local minima so paths reach desired states within limited computational budgets [10]. The presence of local minima often confines sampling-based algorithms to specific regions of the search space, resulting in insufficient sampling.

Here we propose a novel, sampling-based algorithm that addresses the issue of insufficient sampling by leveraging experimentally-determined structures of a protein to restrict sampling in a space of a reasonable number of dimensions and on regions of relevance for transition events. These structures are used both to define a reduced (conformation) search space and to initialize an iterative sampling process. While the algorithm samples in a reduced space, it operates at different scales, as it lifts conformations/samples in a higher-dimensional, structure space and then improving them with the AMBER ff14SB force field to obtain energetically-credible 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

• *T. Maximova and A.Shehu are with the Department of Computer Science, George Mason University, Fairfax, VA, 22030. E. Plaku is with the Department of Electrical Engineering and Computer Science at the Catholic University of America, Washington, DC 20064.*
E-mail: amarda@gmu.edu

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 [11]. 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 experimentally-known structures in a transition event. This new feature allows comparing lowest-cost paths to paths that go through a user-specified set of experimental structures, and thus 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 the 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 different 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.

A proof-of-concept demonstration of the promise of exploiting structural information in a roadmap-based algorithm and extending the path analysis to include tours has been recently presented in [12]. Here we investigate different algorithmic decisions in SoPriM and analyze their impact on the exploration capability, as well as present a more comprehensive analysis of computed landscapes and transitions in comparison with wet-laboratory findings.

This paper proceeds as follows. The proposed SoPriM algorithm is placed in the context of related work in Section 2, before it is described in detail in Section 3. Detailed analysis of the algorithm and its application on several proteins of relevance for human biology and disease are presented in Section 4. The paper concludes in Section 5.

2 Related Work

There is now a rich literature of sampling-based algorithms that leverage robot motion planning to model structural transitions in biomolecules [10], [13]. These algorithms exploit analogies between molecular and robot motions to model molecular dynamics. For instance, direct analogies between molecular bonds and robot links and molecular atoms and robot joints allow employing and adapting techniques that perform fast forward and inverse kinematics for kinematic linkages to molecular kinematics [14], [15], [16], [17], [18], [19]. The problems of robot motion planning and protein transition modeling (or, more generally, molecular motion modeling), are similar. In robot motion planning, the objective is to compute paths from a start to a goal state while satisfying constraints due to the obstacles and the underlying robot dynamics [20], [21]. In transition modeling, the objective is to compute paths from a start to a given structure while satisfying constraints due to the physics-based, energetic interactions among atoms in the molecule.

Sampling-based algorithm that exploit the robot-protein motion planning analogy exploit the observation that transitions of a dynamical system, whether mechanical or biological, between two given states can be modeled via discrete, kinetic models. These models embed computed states of the system in graph-like structures amenable to rapid, shortest, or lowest-cost path queries. Algorithms that embed computed states in a tree are referred to as tree-based, and those that embed samples in a nearest-neighbor graph are referred to as roadmap-based.

An outstanding challenge for both tree- and roadmap-based algorithms involves how to focus limited computational resources to computing transition-relevant states with no a priori information on such states. This challenge concerns both the selection of an effective set of variables that define the search space of interest and the employment of such variables in representation-aware variation or perturbation operators to efficiently sample regions relevant for the sought transition. These two issues are often grouped into one and known as the sampling issue, as they ultimately relate to the ability of a sampling-based algorithm to rapidly obtain a discrete, sample-based representation of the variable space that then allows finding transition-relevant paths connecting the start and goal states.

Variable selection directly determines the dimensionality and complexity of the search space. A popular choice is for the selected variables to be all or a subset of the backbone ϕ, ψ dihedral angles that can be defined over covalently-linked atoms in a protein molecule [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. Dihedral angles are more appealing over Cartesian coordinates, as there are about $3N/7$ dihedral angles in a protein of N atoms [35], and about $2n$ dihedral angles in a protein of n amino acids. Moreover, consecutive backbone dihedral angles can be bundled together into k-long fragments or bundles that capture variable dependencies. This technique, known as molecular fragment replacement, has been incorporated in robotics-inspired algorithms for modeling equilibrium protein structure and dynamics [32], [33], [34], [36].

Other variable reduction or variable prioritization strategies employ system-specific insight, structure analysis, or statistical analysis. For instance, rigidity-based analysis, which detects least-constrained regions in a given structure, and suggests a prioritization scheme for modification of dihedral angles, has been successfully incorporated in robotics-inspired algorithms for modeling the dynamics of proteins as long as 100 amino acids [26]. Application of robotics-inspired algorithms on proteins of more than 200 amino acids often relies on a direct comparison of the start and goal structures to identify the differently-valued dihedral angles as variables [37], [38]. Fragment replacement has allowed tree-based algorithms to address an important range of proteins between $100-200$ amino acids [34], [36]. Other variable reduction strategies rely on statistical analysis to identify collective variables that capture collective motions of atoms in Cartesian space [12], [39], [40], [41], [42], [43]. For instance, Normal Mode Analysis (NMA) [44], a popular technique in computational structural biology [39], [45], [46], [47], is often incorporated in robotics-inspired algorithms [40], [41], [42], [43]. NMA extracts collective motions from one structure at a time. The SoPriM algorithm

proposed here uses a multivariable statistical technique that extracts collective variables from a set of known structures.

The choice of variables is key to the design of effective perturbation or variation operators for generating samples or conformations that satisfy a set of desired geometric and/or energetic constraints; from now on, we refer to an instantiation over the selected variables as a conformation. Conformations obtained uniformly at random have very low probability of being relevant for a sought transition due to collisions among atoms. Biased sampling techniques can be used to remedy this issue, as pioneered by Amato and colleagues [23], [24]. Since it is generally hard to know a priori which operators will be effective, recent work employs diverse operators on diverse sets of variables during sampling [36], [37], [48]. Work in [36] implements a probabilistic scheme that selects among a rich menu of operators making use of angular or Cartesian variables.

Operators that generate new conformations by incremental modifications of existing conformations are more likely to yield conformations relevant for a transition. Recent algorithms generate conformations in neighborhoods of existing "parent" conformations [32], [34], [49] (hence, the term perturbation or variation operator). Perturbation operators that perturb a selected conformation to obtain a new one tend to preserve some good structural features in the new conformation while introducing enough change to explore new regions of the variable space [50]. The SoPriM algorithm proposed here does not sample conformations over the variable space uniformly at random but via perturbation operators that modify selected conformations in the space of collective variables.

In the context of perturbation operators, (parent) selection schemes are critical to control sampling. Traditionally, such selection schemes have been confined to tree-based algorithms, which grow a tree in conformation space in iterations, at every iteration selecting a parent for perturbation, and then perturbing it via some operator to obtain a new child conformation to add to the tree [32], [34]. In this paper, we inject selection schemes in a roadmap-based algorithm. The proposed SoPriM algorithm selects at every iteration a parent conformation for perturbation, and then perturbs it in the space of collective variables. The resulting conformation, if it is passes an energetic threshold, is added to the growing set of sampled conformations and the roadmap that embeds all sampled conformations.

Most robotics-inspired algorithms for modeling biomolecular transitions do not exploit existing structural information about a protein beyond the start and goal structures. In this paper, SoPriM is initialized with an ensemble of experimentally-known structures, which are exploited both to define the reduced space of collective variables and initialize the roadmap vertex set. While years ago the reliance on experimental structures would be a limitation, nowadays over hundreds of thousands structures exist in the Protein Data Bank (PDB) [51]. Moreover, 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 (WT) and variant sequences.

Considering many experimentally-available structures allows SoPriM to focus sampling to the conformation space of interest for the sought transition. These structures (collected from different variants of a protein) represent, per the conformational selection principle, possible semi-stable states in the space of a given protein sequence. By populating the conformation space first around these structures and then gradually further away, SoPriM is likely to sample regions that are relevant for a transition. It is worth noting that the idea of utilizing more than the start and goal structures has been recently proposed in [36], but only SoPriM gradually expands in conformation space from a diverse set of experimentally-known structures (in [36], several geometric and energetic constraints are employed to restrict sampled conformations near experimental structures).

In robotics-inspired algorithms, once conformations are generated, the tree or roadmap is queried for a least-cost path. Unlike trees, where the sampling is specifically biased to connect the start and goal and only one path can be obtained, roadmap-based algorithms support multiple queries; in principle, the roadmap/graph allows extracting paths for different start and goal structures. Typically, in the context of modeling the transition between a given start and goal structure, most algorithms focus on the least-cost path. However, protein transitions are stochastic processes, and a set of energetically-similar paths may provide a broader view of the transition tube of likely paths. Recent work recognizes this issue by providing the K lowest-cost paths, or by treating the roadmap as a Markov state model over which average statistics over paths can be computed [52]. In this paper, SoPriM is executed several times in order to get different lowest-cost paths. In addition, more paths are obtained via the notion of tours, lowest-cost paths that contain subsets of given experimental structures. Collecting these paths and focusing on those within an energetic threshold of the lowest cost one provides a picture of the possible different, energetically-similar transition routes.

Finally, a challenge unique to 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.

3 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 relevant 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 structures 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.

3.1 Conformation Sampling

The input to SoPriM is a set Ω of experimental structures of a protein, collected and curated as described in section 4. 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 [11], it is impractical when dealing with high-dimensional conformation spaces, since the sampled conformations are 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 ensemble $\mathcal{C}$. The perturbation operator modifies the selected conformation to yield a new one, which is subjected to a local improvement operator before being added to $\mathcal{C}$. 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.

3.1.1 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) [53] to reveal collective variables (principal components – PCs) over which to define the conformation space for sampling. This is motivated by our prior work on evolutionary algorithms that employ PCA to find basins in energy landscapes of proteins [54], [55]. PCA and other linear dimensionality reduction techniques are shown effective for many multiple-basin proteins important to human biology and disease [54], [56].

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 [57]. 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 [58] 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 the eigenvalues, 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 are needed to preserve 90% of the original data variance. This effectively results in a reduced search space, where conformations are points with coordinates on each of the top-selected 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 nr. of conformations in roadmap
n_{add}: nr. of conformations to add to roadmap at each stage
k, r: nr. and range for nearest neighbors
$\mathcal{G}$: 2D-grid, i.e., $min_{x,y}, max_{x,y}$, nr. of rows/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 $\text{COST}(c_i, c_j) = \max\{\text{SCORE}(c_j) - \text{SCORE}(c_i), 0\}$

1: $\mathcal{R} = (\mathcal{C}, \mathcal{E}) \leftarrow (\emptyset, \emptyset)$; $\Gamma \leftarrow \emptyset$; $\mathcal{P} \leftarrow \emptyset$; $n \leftarrow n_{min}$; $i \leftarrow 1$
2: **for each** $c \in \Omega$ **do** ADDCONFORMATION$(\mathcal{R}, \Gamma, c)$
3: **repeat**
4: **while** $|\mathcal{C}| < n$ **do**
5: $\gamma \leftarrow$ SELECTGRIDCELL(Γ)
6: $c \leftarrow$ SELECTCONFORMATION(γ)
7: $c_{new} \leftarrow$ GENERATESUCCESSOR$(c, \text{RAND}(\delta_{min}, \delta_{max}))$
8: UPDATESTATISTICS(γ, c, c_{new})
9: **if** $c_{new} \neq$ null **then** ADDCONFORMATION$(\mathcal{R}, \Gamma, c_{new})$
10: $n \leftarrow \min\{|\mathcal{C}| + n_{add}, n_{max}\}$
11: **while** $i \leq |\mathcal{C}|$ **do**
12: $c \leftarrow i$-th conformation in $\mathcal{C}$; $i \leftarrow i + 1$
13: *neighs* $\leftarrow$ NEARESTNEIGHBORS$(\mathcal{R}, \rho, c, k, r)$
14: **for** $c' \in$ *neighs* **do** $\mathcal{E} \leftarrow \mathcal{E} \cup \{(c, c'), (c', c)\}$
15: **until** $\text{CONNECTED}(\mathcal{R}, c_s, c_g, \mathcal{L}) =$ true **or** $|\mathcal{C}| > n_{max}$
16: **for each** $\mathcal{S} \subseteq \mathcal{L}$ **do**
17: $path_{\mathcal{S}} \leftarrow$ SHORTESTPATH$(\mathcal{R}, \text{COST}, c_s, c_g, \mathcal{S})$
18: $\mathcal{P} \leftarrow \mathcal{P} \cup \{path_{\mathcal{S}}\}$
19: **return** $\mathcal{P}$

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

3.1.2 Perturbation and Improvement Operator: 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 $[\delta_{min}, \delta_{max}]$. 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 [59], and side chains are packed via the SCWRL4 program [60]. The resulting all-atom structure is then subjected to a standard, AMBER-recommended minimization protocol [61]. The protocol uses the Amber *ff14SB* 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 cutoff. The implicit, generalized Born solvation model is used ($igb = 1$), and energies associated with conformations include the solvation term.

The conformation now 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 and others arising when threading structures reported for a variant onto the WT sequence.

3.1.3 Selection Operator: Conformation Selection

Conformation selection is key to controlling sampling in conformation space. A two-dimensional 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. 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. The set $\Gamma = \{\gamma : \gamma \in \mathcal{G} \text{ and } nrConfs(\gamma) > 0\}$ denotes all the non-empty cells, where $nrConfs(\gamma)$ indicates the number of conformations in $\mathcal{C}$ that map to γ. Moreover, a weight $w(\gamma)$ is maintained for each non-empty grid cell $\gamma \in \Gamma$. The specific formula employed for w directly impacts the exploration capability of `SoPriM`, and here we investigate two different ways of defining w that correspond to two different implementations of the selection operator.

The first, intuitive definition for w is one where all non-empty cells have equal probability of being selected, i.e., $w(\gamma) = 1/|\Gamma|$. A second definition biases the selection by including various statistics gathered and updated during the course of `SoPriM`'s execution, i.e.,

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

where $minE(\gamma)$, $nrSel(\gamma)$, and $nrFailures(\gamma)$ denote the minimum potential energy over conformations that map to 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 (Alg. 1:5) is then defined according to its weight as

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

The weight formulation in Equation 1 allows SELECTGRIDCELL (Alg. 1:5) to discourage cells that lead to failures, encourage cells that protrude deep in the energy landscape, and reject 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 weighting function and probability distribution can be used over the conformations in γ in order to select a conformation (Alg. 1:6) for the GENERATESUCCESSOR function. An intuitive choice is to grant equal weight to each conformation in γ, i.e., $w(c) = 1/nrConfs(\gamma)$. Alternatively, the selection can be biased, i.e.,

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

In section 4, we compare the uniformly at random selections of γ and c to the biased selections that results from defining $w(\gamma)$ and $w(c)$ as in Equations 1 and 3.

3.2 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 root-mean-squared deviation and Euclidean distance in the space of PCs yields a reasonable value for r (data not shown). The idea 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 (in the absence of a distance criterion in nearest-neighbor computations).

3.3 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

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

where $E(c)$ stands for the ff14SB energy of the reconstructed structure corresponding to c (roadmap edges are directed). This definition only records uphill energetic variations; the latter represent 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 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 [62].

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 permutations of $s_1, \ldots, s_\ell$. For a permutation $s_{\pi_1}, \ldots, s_{\pi_\ell}$, 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 $path(\langle s_{\pi_1} \ldots s_{\pi_\ell} \rangle)$ with the lowest cost over all permutations of $s_1, \ldots, s_\ell$. As described in section 4, the set $\mathcal{P}$ is analyzed to identify experimental structures that serve as intermediates in a transition event.

3.4 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 $k \in \{10, 20, 30, 40, 50\}$, and r corresponding to lRMSDs of $\{0.25, 0.5, 1.0, 2.0, 3.0\}$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 $\{0.5, 1.0, 2.0\}$.

4 Results

4.1 Test Cases and Data Preparation:

Performance is evaluated on 3 proteins of importance to human biology and disease, the catalytic domain of uncomplexed H-Ras, the superoxide dismutase [Cu-Zn] (SOD1), and Calmodulin (CaM). The lengths of these proteins varies from 144 to 166 amino acids. 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. The cumulative variance profiles, not shown here, can be found in [54], where evolutionary algorithms operate over PC projection spaces to probe multiple-basin energy surfaces.

4.2 Experimental Setup

Two separate analyses are conducted. The first analysis, presented in Section 4.3, investigates the impact of the selection operator (detailed in Section 3.1.3) on conformation ensemble obtained by SoPriM. The second analysis, presented in Section 4.4, is conducted on applications of SoPriM on selected proteins.

4.3 Analysis of Impact of Selection Mechanism

Energetic and structural features of the growing conformation ensemble $\mathcal{C}$ are tracked over the iterations in SoPriM. In the analysis below, the displacement magnitude δ_{max} is set to 1; Appendix A analyzes the impact of the displacement magnitude and its interplay with the selection operator.

The growing ensemble $\mathcal{C}_i$ is evaluated at regular snapshots $i \in \{0, w, 2w, \ldots\}$, where w is the number of iterations SoPriM is allowed to execute before the ensemble is re-evaluated. Recall that $\mathcal{C}_0$ consists of (energetically-refined) conformations obtained by threading the experimentally-known structures onto the sequence of interest, as described in Section 3. Two settings are considered, one where SoPriM uses the uniformly at random selection operator, and another where SoPriM uses the biased selection operator.

Two separate evaluations of $\mathcal{C}_i$ are conducted. In the first, conformations in $\mathcal{C}_i$ are shown in the PC1-PC2 embedding (effectively using only the first two coordinates of each conformation), and the projections are color-coded based on the AMBER ff14SB energy values of the all-atom structures corresponding to conformations in $\mathcal{C}_i$. The result of this projection of the probed energy surface over the top two collective variables (PCs) is what is often referred to as an energy landscape; for all proteins studied here the top two PCs capture at least 50% of the cumulative variance, and in turn projections on the top two PCs can be used to drawn observations. By visualizing the energy landscape as $\mathcal{C}$ grows, one obtains a dynamic view of how the selection operator steers the exploration in SoPriM. The baseline at $i = 0$ allows seeing how the emerging landscape features are influenced, if at all, by the initialization. Comparing landscapes obtained at the point in time, but by the two different selection operators, allows drawing qualitative observations on the impact of the two implementations on the exploration capability of SoPriM. The second evaluation concerns the structural diversity of $\mathcal{C}_i$. This is measured via the distribution of pairwise CA lRMSDs among the structures corresponding to the ensemble $\mathcal{C}_i$.

These two evaluations are related on a selected protein, H-Ras. Fig. 1 juxtaposes the energy landscape corresponding to $\mathcal{C}_0$ (the experimental structures) to three landscapes corresponding to $i = 500$, $i = 1,500$ and $i = 3,000$ iterations, shown on the left panel for the uniformly at random selection operator and on the right for the biased selection operator.

Comparison of the right panel to the left panel in Fig. 1 shows that the biased selection scheme affords higher exploration capability to SoPriM. New regions are populated more rapidly. In contrast, the uniformly at random selection scheme results in SoPriM exploring regions nearby the energy basins populated by the experimentally-known structures. While all conformations in $\mathcal{C}_i$ have equal probability of being selected, in terms of organization by basins, the basins with higher population will be selected more often for enrichment by the uniformly at random selection operator than basins with lower population. As can be seen in the slow growth of the landscape in Figures 1(b1),(c1),(d1), significant execution time has to pass for the uniformly at random operator to allow SoPriM to expand away from the existing basins, as new, under-populated regions have lower probability of selection than well-populated regions.

Fig. 2 shows the structural diversity of $\mathcal{C}_i$ by plotting the pairwise CA lRMSD distributions for H-Ras at the same 4 intervals, juxtaposing SoPriM with the uniformly at random selection operator on the left panel with SoPriM with the biased selection operator on the right panel. The distributions obtained at later iterations are superimposed over those obtained at earlier iterations.

Fig. 2(a) shows that the uniformly at random selection operator does not readily change the distribution in $\mathcal{C}_0$; the operator causes SoPriM to spend further time populating the already well-populated bins at $i = 0$. In contrast, Fig. 2(b) shows that the distribution obtained by SoPriM when using the biased selection operator quickly fills out new regions of the conformation space; in particular, the population of bins around lRMSD $1.2 - 1.5$Ågrows. This region is about halfway between the On and Off active states of H-Ras, and corresponds to an energy barrier in the landscape (as found by SoPriM and visible in Fig. 1(c2), (d2)). Comparison of Fig. 2(a) to Fig. 2(b) shows that the biased selection operator allows SoPriM to quickly populate this region and more readily explore regions different from those populated by the initial structures.

Taken together, the above analysis suggests that the biased selection operator is more effective at exploration than the uniformly at random operator. However, a uniformly at random selection operator affords more exploitation capability to SoPriM; that is, drilling down in a basin. The latter behavior can also be accomplished by interleaving a more exploration-driven selection operator, such as the biased selection one used here, with a prudent perturbation operator (demonstrated via additional analysis in Appendix A). The analysis suggests that the biased selection operator combined with different displacement magnitudes in the perturbation operator provides a good balance between exploration and exploitation. For this reason, the rest of our analysis on the three protein systems is conducted over conformation ensembles that combine those obtained by 5 independent runs; as suggested by the above analysis, all runs use the biased selection operator, but three different displacement magnitudes are considered; $\delta_{max} \in \{1.0, 2.0, 3.0\}$. Once conformations are obtained, lowest-cost paths and tours are computed at different values of nearest neighbors k and range r. Specifically, k ranges in $[10, 20, 30, 40, 50]$. For H-Ras and SOD1, r ranges in $\{0.25, 0.5\}$Å. For CaM, where the maximum pairwise lRMSD between the experimentally-known structures is 21Å, r ranges in $\{1, 2, 3\}$Å. The lowest-cost paths and tours obtained by the different runs are collected, and only those with costs no higher than a threshold of 3.5kcal/[mol $\cdot$ residue] of the lowest cost obtained are retained for further analysis. Such paths and tours are visualized on the PC1-PC2 energy landscapes.

4.4 Summary Analysis of Landscapes and Paths Obtained by SoPriM

The analysis on the three selected proteins is conducted over ensembles obtained by SoPriM with the Amber ff14SB force field. In a proof-of-concept demonstration of SoPriM in [12], we utilized the ff12SB force field. Comparative analysis in Appendix B shows that the two force fields lead SoPriM to obtain highly similar landscapes.

4.4.1 Landscape and Low-cost Paths Obtained for SOD1

SOD1 is a 150 amino-acid long enzyme critical in the detoxification of superoxide radicals in the body. SOD1 misfolding and aggregation have been associated with the development of late-onset neurodegenerative diseases such as Parkinson's, Alzheimer's, and Amyotrophic Lateral Sclerosis (ALS). Mutations in SOD1 have been linked to familial ALS. SOD1 WT is a dimer held together by disulfide bridges. Oxidation of SOD1 involves binding of Cu and Zn ions to each subunit. Cu- and Zn-deficient states of SOD1 are destabilized and prone to monomerization. Different states of metallation of SOD1 in monomeric forms have been the focus of several studies, as these structures tend to be pathogenic. At normal conditions, SOD1 functionality is satisfied by the Cu-Zn binding regulation mechanism. Wet-laboratory evidence suggests that in the absence of free Cu ions, the Zn alone is able to provide the Zn-Zn regulation, maintaining SOD1 functionality [63].

We use SoPriM to study the transformations between WT apo- and fully-metallated WT SOD1 structures to better understand Cu-Zn and Zn-Zn regulation. SoPriM is applied to the SOD1 functional region that is implicated in three biological processes, SOD1 oxidation regulated through the binding Cu and Zn ions, glutathionylation, and phosphorylation. The latter two impact SOD1 activity through interactions with other proteins [64].

The preliminary investigation in [12] applies SoPriM to the region consisting of amino acids at positions $1-150$. SoPriM-obtained conformations are projected onto PC1 and PC2 and color-coded by Amber ff14SB energy values to visualize the energy landscape. The landscape shown in Fig. 3(a) contains two basins separated by an energy barrier; the known structures project to the basins (black dots). This organization is due to structural changes upon phosphorylation and has been also reported by prior work applying an evolutionary algorithm to probe multiple-basin energy surfaces [54]. Analysis of SoPriM-obtained lowest-cost paths and tours in [12] shows that structures reported

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2586044, IEEE/ACM Transactions on Computational Biology and Bioinformatics

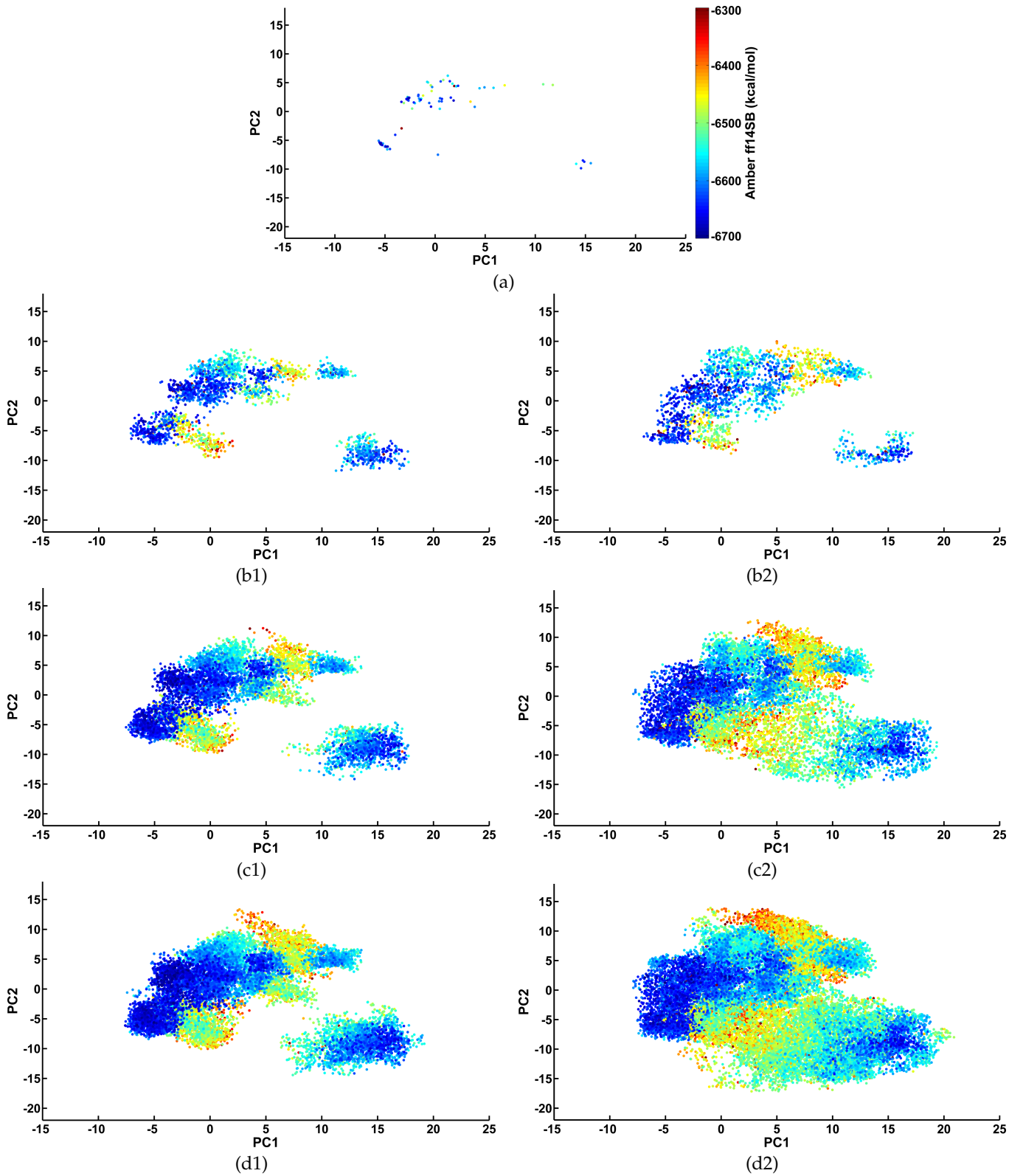
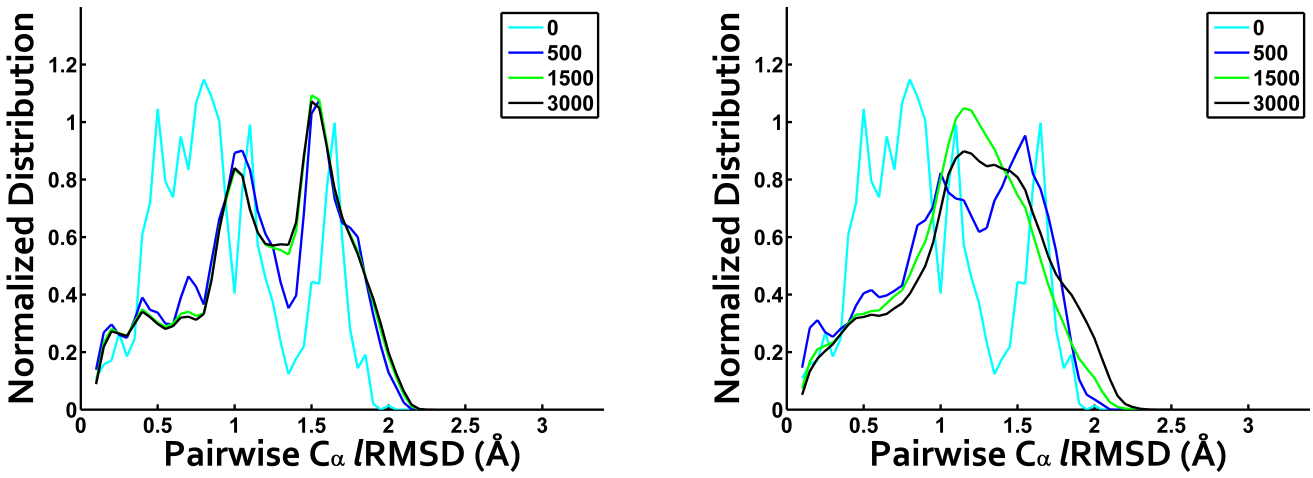


Fig. 1. The growing ensemble $\mathcal{C}_i$ is projected onto the top two PCs and color-coded by AMBER ff14SB energy values to obtain a dynamic view of the SoPriM-generated H-Ras energy landscape. (a) shows the landscape probed immediately after initialization, when only experimentally-known structures of H-Ras are present in the ensemble. (b1) and (b2) juxtapose the landscapes probed after 500 iterations when using (b1) the uniformly at random versus (b2) the biased selection operator. Similarly, (c1) and (c2) juxtapose the landscapes probed after $1,500$ iterations, and (d1) and (d2) provide the juxtaposition after $3,000$ iterations by the uniformly at random versus the biased selection operator, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2586044, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Uniformly at random selection (b) Biased selection

Fig. 2. The distribution of pairwise CA lRMSDs among conformations obtained after every $i \in \{0, 500, 1, 500, 3, 000\}$ is shown ($i = 0$ corresponds to the distribution over experimentally-known structures that initialize the ensemble $\mathcal{C}_i$) to track the structural diversity of the growing ensemble. (a) does so over when SoPriM uses the uniformly at random selection operator and (b) shows the results when SoPriM uses the biased selection operator. In both settings, the displacement magnitude $\delta_{max} = 1$.

in PDB id 2NNX, orange dots in Fig. 3(a), mediate the inter-basins transition corresponding to phosphorylation. This PDB entry is reported for a disease-associated double mutant that diminishes Cu binding and dimer stability; SoPriM finds these structures to be semi-stable for the WT and mediate the transition.

Phosphorylation-induced motions of N-terminal residues in SOD1 are large and dominate PC1. To study other, more subtle motions possibly related to the oxidation mechanism in SOD1, the PCA analysis is repeated here on residues $3-150$, and SoPriM is reapplied. The new landscape is shown in Fig. 3(b), which shows that, when the phosphorylated region is removed, known structures organize in a single basin. As a result, a greater variety of energetically-similar paths are involved in the oxidation process. Fig. 3(b) relates this by showing the region of the landscape involved in low-cost paths from the same start structure, WT apo-SOD1 form (PDB id 1HL4) to four different goal structures (PDB ids 1HL5, 1SPD, 2C9S, and 2C9V) corresponding to different metallated forms of SOD1.

SoPriM-obtained paths from apo- to metallated Cu-Zn binding structures range in energetic costs from 0.05 to 1.5 kcal/mol, with an average of 0.45 kcal/mol. Paths from the apo to the Zn-Zn binding structures have higher costs in the range 0.6 to 3.4 kcal/mol, with an average of 0.65. These quantitative results suggest that, even though the Zn-Zn state appears to be as stable as the Cu-Zn metallated state [65], the formation of Cu-Zn structures may be more preferable from the apo state.

An additional observation can be drawn regarding the role of SOD1 structures reported under PDB id 2NNX. In addition to a mediating role for phosphorylation, these structure seem to mediate the transitions in the oxidation process, as well. Based on the location in the transitions obtained for SOD1 and drawn in Fig. 3(b), further stabilization of this structure upon mutations may slow the transition, thus affecting SOD1 function. Taken together, these results suggest that, though SOD1 is a challenging protein to study, specific structures, such as the highlighted double mutant, can be further investigated to better understand function modulation in SOD1 WT and variants.

4.4.2 *Landscape and Low-cost Paths Obtained for H-Ras*

H-Ras is 166 amino acids long, mediates signaling pathways controlling cell proliferation and growth and switches between two states, On and Off, to regulate its activity. The switch is a slow process, and at normal conditions is accelerated by the binding of On H-Ras to the G-protein activating (GAP) and Off H-Ras to the guanine nucleotide exchange factor (GEF) [66]. Most wet-laboratory studies aim to find alternatives to the GAP- and GEF-regulated mechanism to accelerate the On-Off switch in the case of mutations in H-Ras.

The landscape and low-cost On-Off paths obtained by SoPriM on WT H-Ras are drawn in Fig. 4. All conformations obtained are projected onto the top two PCs and color-coded by their Amber ff14SB energy values on order to visualize the energy landscape reconstructed by SoPriM. The experimentally-known structures used by SoPriM are shown by drawing their projections in black. Specific structures crucial to our analysis are drawn in different colors. The start structure (projection is drawn blue) used by SoPriM for path queries is the WT On structure with PDB id 1QRA. The goal structure (projection drawn in green) is the WT Off structure under PDB id 4Q21. Structures used by SoPriM in tour queries are highlighted by coloring their projections in orange. Various annotations on the sides of the landscape provide information on the PDB ids and known biophysical characteristics and activity of these structures.

Fig. 4 shows multiple basins that contain different groups of known structures. We note that many of these structures are only observed on variants of H-Ras and not the WT. Threading them onto the WT and minimizing them as described in Section 3 provides possible semi-stable structures for the WT. The location of their projections on low-energy regions discovered by SoPriM is a confirmation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2586044, IEEE/ACM Transactions on Computational Biology and Bioinformatics

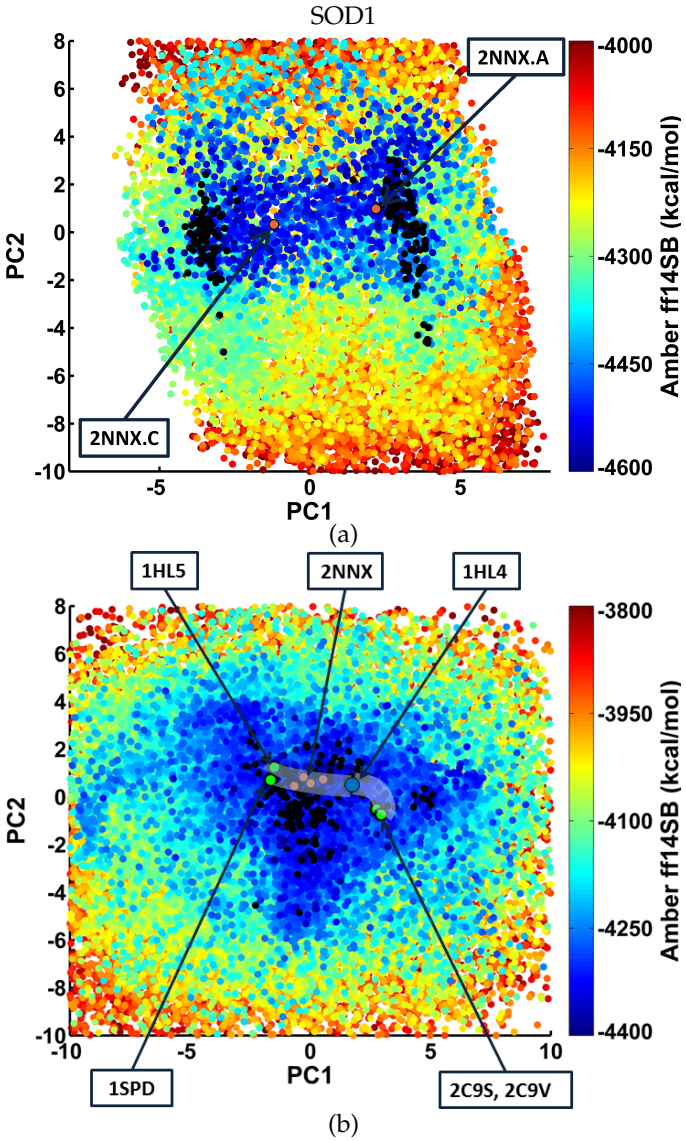


Fig. 3. Conformations obtained by SoPriM on the (a) $1-150$ amino-acid region and then on the (b) $3-150$ amino-acid region of SOD1 are projected onto the top two PCs and color-coded by AMBER ff14SB energy values. Projections of known structures are drawn in black. Projections of structures under PDB id 2NNX are in orange. (b) Lowest-cost paths and tours that meet the energetic criterion are summarized via transition tubes. Different settings are considered, using the same start structure (in blue) but four different goal structures (in green).

of the principle of conformational selection; that is, stable structures of variants are semi-stable for the WT. Three distinct low-energy regions are observed, separated by energy barriers. The widest, low-energy region appears on the top left of the landscape and corresponds to the On state (based on the location of the projection of the On structure under PDB id 1QRA). While containing many deep local minima, this region is highly rugged. This has implications for the On-Off regulation, as observed when comparing low-cost paths and tours between the On and Off states (detailed later). The second widest appears on the bottom right and corresponds to the Off state (based on the location of the projection of the Off structure under PDB id 4Q21).

Several known WT structures project onto the region corresponding to the On state in Fig. 4. These include not only the GTP-bound structure under PDB id 1QRA and the canonical structure under PDB id 1CTQ, but also the GTP GAP-bound structure under PBD id 1WQ1 and the WT SOS-bound structure under PDB id 1NVW. The co-location of the latter two structures supports wet-laboratory evidence that both GAP and SOS effectors use the same mechanism of On-Off activation [67], [68]. Additional insights can be drawn. In particular, the landmark structures that SoPriM uses in tour calculations have been selected to be structures with severe, oncogenic mutations (at G12 and Q61) and structures employed in MD simulations [69], [70], [71]. These structures are organized in different sub-groups and project onto different regions within the On basin, based on whether they have effectors or not. For instance, oncogenic structures without effectors are closer to the canonical and GTP-bound On structures, whereas oncogenic structures with effectors, such as Raf, PI3K kinase, and GSP, are further away.

Recent experiments report the presence of an allosteric switch in H-Ras [72] (marked by light green circles and annotations in Fig. 4); it is speculated that this allosteric switch may mediate the On-Off switch in H-Ras [72], [73]. These recently-found structures are classified as either having an On or an Off shift, based on structural features [73]. Inspection of the projections of these structures on the SoPriM-obtained landscape shows that the projections fall in two distinct groups, in agreement with the On and Off shifts. The On-shifted structures are closer to the On structures, whereas the Off-shifted ones are further away.

Low-cost paths and tours (that meet the energetic criterion) are now analyzed. The paths are shown in detail in Appendix C. In the interest of clarity, Fig. 4 summarizes them via transition tubes, essentially grouping the observed paths based on regions they navigate. The widths of the tubes correspond to path diversity. As can be seen in Fig. 4, SoPriM obtains three conformation switching scenarios/routes for the On to Off transition. The widest transition tube makes use of the allosteric switch structures. This route has greater energetic diversity (paths have costs ≥ 2.5kcal/mol) but is also marked by higher path diversity. MD studies, which confirm the stability of allosteric binding [74], [75], provide complementary evidence of the possible role of the allosteric switch in the On-Off regulation in H-Ras, as obtained here by SoPriM.

The other two routes obtained by SoPriM have less path diversity but lower costs, as low as 1.3kcal/mol. These paths are harder to find (and result only when $k = 50$), which is due to high ruggedness in the vicinity of 1LF0 (a stringent energetic threshold in our preliminary investigation in [12] provided evidence of these two routes only). These routes are interesting, as they have also been proposed in previous MD simulation studies of the On to Off transition [69], [71]. Taken together, these results suggest that the On to Off transition in H-Ras may make use of distinct routes and thus presents an opportunity for drug-induced On-Off regulation in oncogenic variants.

4.4.3 *Summary Analysis of Low-cost Paths of CaM*

CaM is a 144 amino-acid long enzyme that has been captured in diverse bound and unbound states in the wet laboratory. Different settings are investigated to observe CaM transitions from an open state to closed/peptide-binding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2586044, IEEE/ACM Transactions on Computational Biology and Bioinformatics

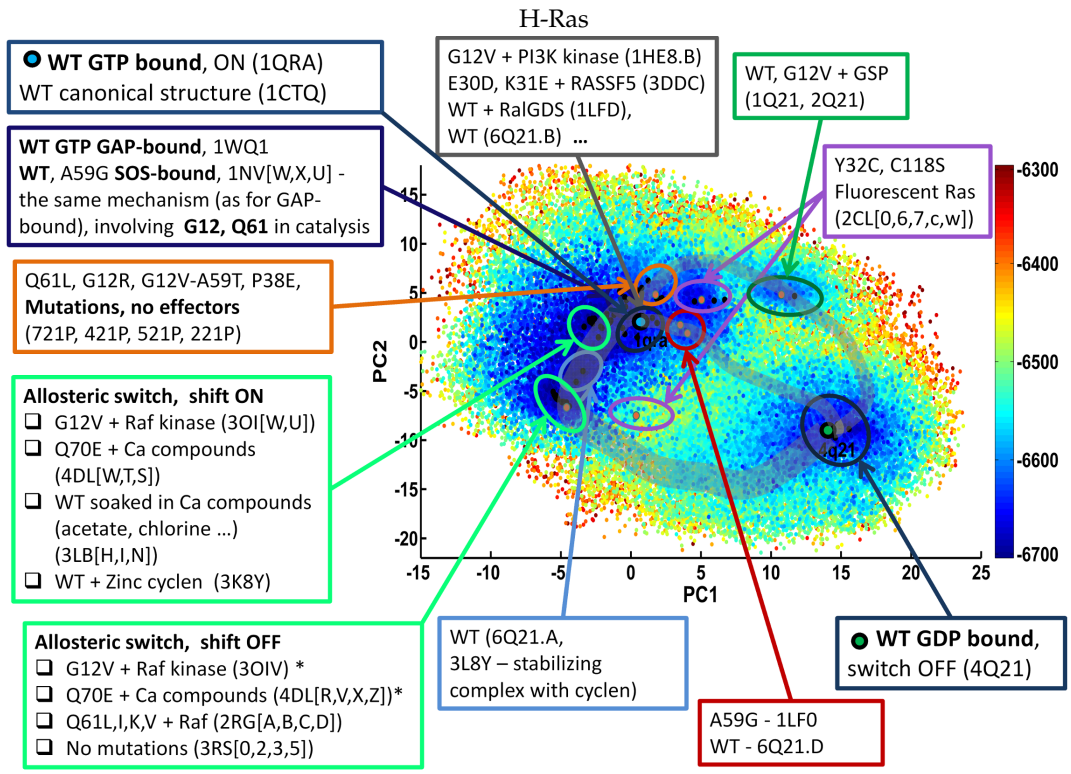


Fig. 4. Conformations obtained by SoPriM are projected onto the top two PCs and color-coded by AMBER ff14 energy values to obtain a view of the SoPriM-generated H-Ras energy landscape. The landscape is enriched with information on where experimentally-known structures project in the embedding. PDB ids are shown, together with brief function information. The start and goal are in blue and green, and landmarks are in orange. The regions of the space navigated by energetically-similar lowest-cost paths and tours are shown via transition tubes. The width of the tubes relates to the diversity of paths in them (detailed paths can be found in Appendix C).

states. The open state is represented by the structure with PDB id 1CLL and used as start. Two different structures are employed as goals in different runs of the algorithm, the ones with PDB id 2F3Y and 1NWD (these bind to different peptides and show different degrees of collapse of the N- and C- domains in CaM). One landmark is specified, the structure with PDB id 1CFD, to investigate the hypotheses that open-to-close transitions in CaM go through the apo (calcium-free) state, where the internal helix connecting the N- and C-terminal domains is partially unfolded to possibly accommodate further collapse of the domains.

Many low-cost paths and tours are related in an earlier investigation of SoPriM [12], where it is observed that paths that go through 1CFD have higher cost. These results are synthesized in a schematic in Fig. 5(d), which shows that the transitions from 1CLL to the closed state 1NWD may not make use of 1CFD (tours obtained by SoPriM forced to go through 1CFD have a higher cost of about 3.5kcal/mol per residue). PDB ids of known structures participating in the different routes to each of the closed states are shown. Fig. 5(b) also 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, as captured in various structures in the NMR ensemble with PDB id 2K0E. This ensemble has been contributed to the PDB by work in [76].

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 [76] 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 open and peptide-binding states of CaM does not make use of the apo/calcium-free structure reported under PDB id 1CFD but instead makes use of Ca(2+)-bound structures and MLCK-bound structures captured in the wet-laboratory under PDB id 2K0E. While work in [76] was restricted to MLCK binding, the results obtained for CaM here suggest that the same mechanism observed in [76] 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 here point to a general mechanism for the open-to-closed/complexed dynamics of CaM, 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.

5 Conclusion

The definition of edge weight here employs the concept of mechanical work. Future work will consider additional concepts, such as minimum resistance. Another direction will consider adaptive selection operators that switch between exploration and exploitation during execution.

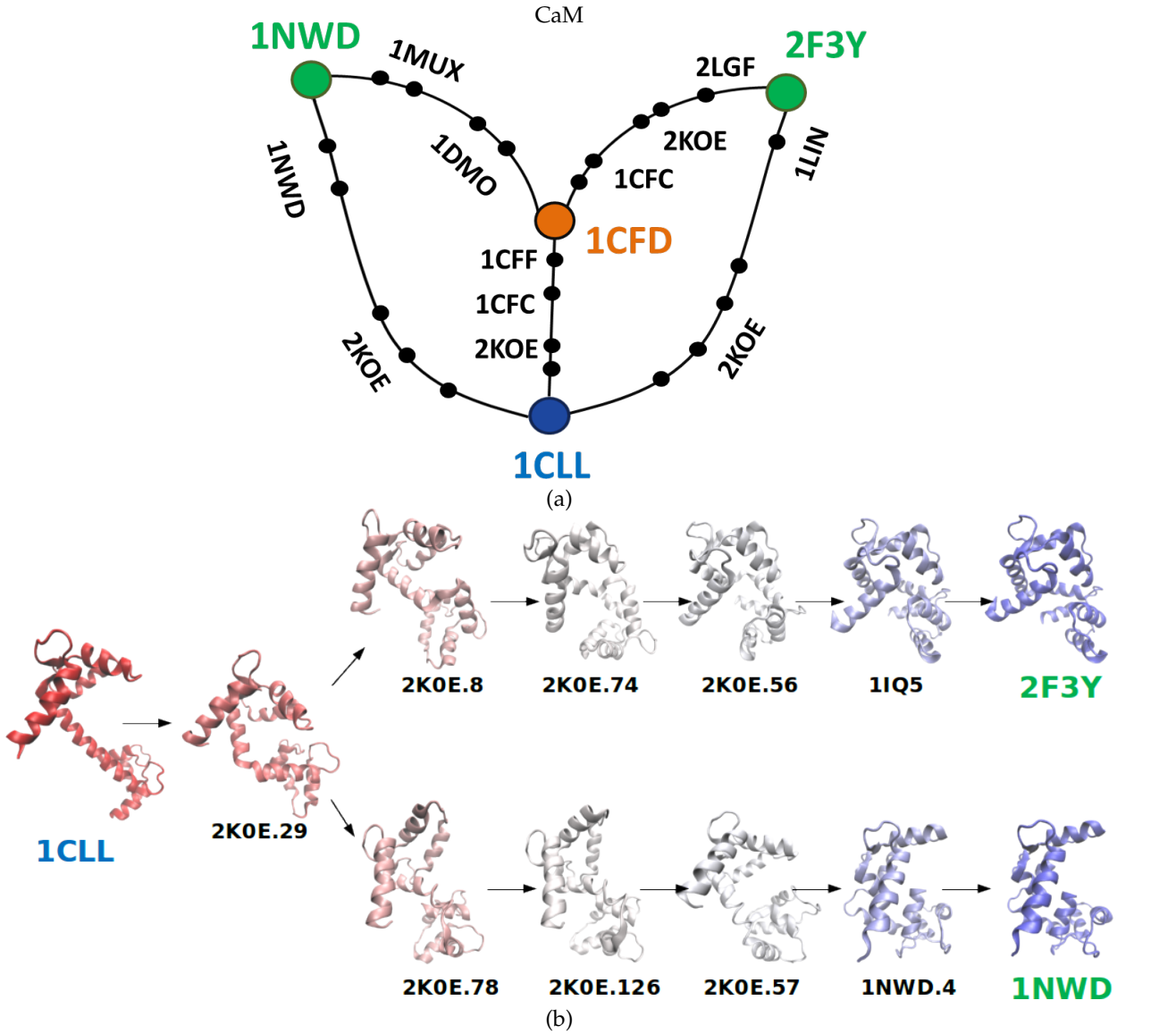


Fig. 5. (a) Schematic summarizes lowest-cost paths obtained by SoPriM, showing PDB ids of known structures participating in the transitions. (b) Successive structures in the lowest-cost paths found for the 1CLL to 2F3Y and 1CLL to 1NWD transitions are also shown. Numbers indicate model number within an NMR entry.

Other lines of further investigation concern multivariate statistical analysis techniques of existing experimental structures. Here we employ a linear technique that allows directly sampling in the reduced space. Non-linear techniques will be considered in the future. The additional demand of direct sampling in the reduced space will have be addressed. While beyond the scope of this paper, several local strategies can be employed in this regard. Techniques, such as NMA, can also be employed to soften the reliance on a set of experimental structures.

Future work will investigate applications on variants of a protein. This setting will allow comparing landscapes and paths to understand the role of structure and energetics in the impact of sequence mutations on misfunction.

Acknowledgment

This work is supported in part by NSF SI2 No. 1440581 and NSF IIS CAREER Award No. 1144106. 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] 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*, 1st ed. Cambridge, MA: MIT Press, 2005.

[12] T. Maximova, E. Plaku, and A. Shehu, "Computing transition paths in multiple-basin proteins with a probabilistic roadmap algorithm guided by structure data," in *IEEE Intl Conf Bioinf and Biomed (BIBM)*, 2015, pp. 35–42.

[13] B. Gipson, D. Hsu, L. E. Kavraki, and J. Latombe, "Computational models of protein kinematics and dynamics: Beyond simulation," *Annu. Rev. Anal. Chem.*, vol. 5, pp. 273–291, 2012.

[14] D. Manocha, Y. Zhu, and W. Wright, "Conformational analysis of molecular chains using nano-kinematics," *Comput. Appl. Biosci.*, vol. 11, no. 1, pp. 71–86, 1995.

[15] M. Zhang and L. E. Kavraki, "A new method for fast and accurate derivation of molecular conformations," *Chem. Inf. Comput. Sci.*, vol. 42, no. 1, pp. 64–70, 2002.

[16] G. S. Chirikjian, "General methods for computing hyper-redundant manipulator inverse kinematics," in *Proc IEEE/RSJ Int Conf Intell Robot Sys (IROS)*, vol. 2. Yokohama, Japan: IEEE, 1993, pp. 1067–1073.

[17] D. Manocha and J. Canny, "Efficient inverse kinematics for general 6r manipulator," *IEEE Trans. Robot. Autom.*, vol. 10, no. 5, pp. 648–657, 1994.

[18] M. Zhang and L. E. Kavraki, "Finding solutions of the inverse kinematics problem in computer-aided drug design," in *Currents in Computational Molecular Biology*, L. Florea, B. Walenz, and S. Hannenhalli, Eds., no. TR02-385. Washington, DC: ACM Press, 2002, pp. 214–215.

[19] R. Kolodny, L. Guibas, M. Levitt, and P. Koehl, "Inverse kinematics in biology: the protein loop closure problem," *Int. J. Robot. Res.*, vol. 24, no. 2-3, pp. 151–163, 2005.

[20] 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.

[21] E. Plaku, "Region-guided and sampling-based tree search for motion planning with dynamics," *IEEE Transactions on Robotics*, vol. 31, pp. 723–735, 2015.

[22] L. Han and N. M. Amato, "A kinematics-based probabilistic roadmap method for closed chain systems," in *Algorithmic and Computational Robotics: New Directions*, B. R. Donald, K. M. Lynch, and D. Rus, Eds. MA: AK Peters, 2001, pp. 233–246.

[23] N. M. Amato, K. A. Dill, and G. Song, "Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures," *J. Comp. Biol.*, vol. 10, no. 3-4, pp. 239–255, 2002.

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

[25] S. Thomas, G. Song, and N. Amato, "Protein folding by motion planning," *Physical Biology*, no. 2, pp. S148–S155, 2005.

[26] 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.

[27] L. Jaillet, J. Cortés, and T. Siméon, "Transition-based rrt for path planning in continuous cost spaces," in *IEEE/RSJ Int. Conf. Intel. Rob. Sys.* Stanford, CA: AAAI, 2008, pp. 22–26.

[28] X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. Amato, "Simulating rna folding kinetics on approximated energy landscapes," *J. Mol. Biol.*, vol. 381, no. 4, pp. 1055–1067, 2008.

[29] L. Tapia, X. Tang, S. Thomas, and N. Amato, "Kinetics analysis methods for approximate folding landscapes," *Bioinformatics*, vol. 23, pp. i539–i548, 2007.

[30] L. Tapia, S. Thomas, and N. Amato, "A motion planning approach to studying molecular motions," *Communications in Information Systems*, vol. 10, no. 1, pp. 53–68, 2010.

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

[32] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–1127, 2010.

[33] K. Molloy, S. Saleh, and A. Shehu, "Probabilistic search and energy guidance for biased decoy sampling in ab-initio protein structure prediction," *IEEE/ACM Trans Bioinf and Comp Biol*, vol. 10, no. 5, pp. 1162–1175, 2013.

[34] K. Molloy and A. Shehu, "Elucidating the ensemble of functionally-relevant transitions in protein systems with a robotics-inspired method," *BMC Struct Biol*, vol. 13, no. Suppl 1, p. S8, 2013.

[35] R. Abayagan, M. Totrov, and D. Kuznetsov, "ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation," *J. Comput. Chem.*, vol. 15, no. 5, pp. 488–506, 1994.

[36] K. Molloy and A. Shehu, "A general, adaptive, roadmap-based algorithm for protein motion computation," *IEEE Trans NanoBioScience*, 2015, in press.

[37] B. Raveh, A. Enosh, O. Schueler-Furman, and D. Halperin, "Rapid sampling of molecular motions with prior information constraints." *PLoS Computational Biology*, vol. 5, no. 2, 2009.

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

[39] M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, "Elastic models of conformational transitions in macromolecules," *J Mol Graph Model*, vol. 21, no. 2, pp. 151–160, 2002.

[40] K. M. Kim, R. L. Jernigan, and G. S. Chirikjian, "Efficient generation of feasible pathways for protein conformational transitions," *Biophys. J.*, vol. 83, no. 3, pp. 1620–1630, 2002.

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

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

[43] I. Al-Bluwi, M. Vaisset, T. Siméon, and J. Cortés, "Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods," *BMC Struct Biol*, vol. 13, no. S2, p. Suppl 1, 2013.

[44] Q. Ciu and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, 1st ed. CRC Press, 2005.

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

[46] W. Zheng, B. R. Brooks, and G. Hummer, "Protein conformational transitions explored by mixed elastic network models," *Proteins: Struct Funct Bioinf*, vol. 69, no. 1, pp. 43–57, 2007.

[47] A. Das, M. Gur, M. H. Cheng, S. Jo, I. Bahar, and B. Roux, "Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model," *PLoS Comput Biol*, vol. 10, no. 4, p. e1003521, 2014.

[48] 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.

[49] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu, "Principles and overview of sampling methods for modeling macromolecular structure and dynamics," *PLoS Comput Biol*, 2015, in press.

[50] B. Olson, I. Hashmi, K. Molloy, and A. Shehu, "Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules," *Advances in AI J*, vol. 2012, no. 674832, 2012.

[51] 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.

[52] K. Molloy, R. Clausen, and A. Shehu, "A stochastic roadmap method to model protein structural transitions," *Robotica*, 2015, in press.

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

[54] R. Clausen and A. Shehu, "A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes," *J Comp Biol*, vol. 22, no. 9, pp. 844–860, 2015.

[55] 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.

[56] R. Pandit and A. Shehu, "A principled comparative analysis of dimensionality reduction techniques on protein structure decoy data," in *Intl Conf on Bioinf and Comp Biol (BICoB)*, Las Vegas, NV, 2016.

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

[58] 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.

[59] 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.

[60] 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.

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

[62] 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.

[63] Y. Sheng, M. Chattopadhyay, J. Whitelegge, and J. S. Valentine, "SOD1 aggregation and ALS: role of metallation states and disulfide status," *Curr Top Med Chem*, vol. 12, no. 22, pp. 2560–2572, 2012.

[64] K. Wilcox, L. Zhou, J. Jordon, Y. Huang, Y. Yu, R. Redler, X. Chen, M. Caplow, and N. Dokholyan, "Modifications of superoxide dismutase (sod1) in human erythrocytes a possible role in amyotrophic lateral sclerosis," *J Biol Chem*, vol. 284, no. 20, pp. 13 940–13 947, 2009.

[65] R. Strange, S. Antonyuk, M. Hough, P. Doucette, J. Valentine, and S. Hasnain, "Variable metallation of human superoxide dismutase: atomic resolution crystal structures of Cu, Zn and as-isolated wild-type enzymes," *J Mol Biol*, vol. 356, no. 5, pp. 1152–1162, 2006.

[66] K. Scheffzek, M. R. Ahmadian, W. Kabsch, L. Wiesmüller, A. Lautwein, F. Schmitz, and A. Wittinghofer, "The Ras-RasGAPcomplex: structural basis for GTPase activation and its loss in oncogenic Ras mutants," *Science*, vol. 277, no. 5324, pp. 333–339, 1997.

[67] G. Buhrman, G. Wink, and C. Mattos, "Transformation efficiency of RasQ61 mutants linked to structural features of the switch regions in the presence of raf," *Structure*, vol. 15, no. 12, pp. 1618–1629, 2007.

[68] S. M. Margarit, H. Sondermann, B. E. Hall, B. Nagar, A. Hoelz, M. Pirruccello, D. Bar-Sagi, and J. Kuriyan, "Structural evidence for feedback activation by Ras-GTP of the Ras-specific nucleotide exchange factor SOS," *Cell*, vol. 112, no. 5, pp. 685–695, 2003.

[69] B. J. Grant, A. A. Gorfe, and J. A. McCammon, "Ras conformational switching: simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics," *PLoS Comput Biol*, vol. 5, no. 3, p. e1000325, 2009.

[70] A. A. Gorfe, B. J. Grant, and J. A. McCammon, "Mapping the nucleotide and isoform-dependent structural and dynamical features of Ras proteins," *Structure*, vol. 16, no. 6, pp. 885–896, 2008.

[71] S. Lukman, B. J. Grant, A. A. Gorfe, G. H. Grant, and J. A. McCammon, "The distinct conformational dynamics of K-Ras and H-Ras A59G," *PLoS Comput Biol*, vol. 6, no. 9, p. e1000922, 2010.

[72] G. Buhrman, O. Casey, B. Zerbe, B. M. Kearney, R. Napoleon, E. A. Kovrigina, S. Vajda, D. Kozakov, E. L. Kovrigin, C. Mattos *et al.*, "Analysis of binding site hot spots on the surface of Ras GTPase," *J Mol Biol*, vol. 413, no. 4, pp. 773–789, 2011.

[73] G. Buhrman, V. S. Kumar, M. Cirit, J. M. Haugh, and C. Mattos, "Allosteric modulation of ras-gtp is linked to signal transduction through raf kinase," *J Biol Chem*, vol. 286, no. 5, pp. 3323–3331, 2011.

[74] M. McCarthy, P. Prakash, and A. A. Gorfe, "Computational allosteric ligand binding site identification on Ras proteins," *Acta Biochim et Biophys Sinica*, p. gmv100, 2015.

[75] B. J. Grant, S. Lukman, H. J. Hocker, J. Sayyah, J. H. Brown, J. A. McCammon, and A. A. Gorfe, "Novel allosteric sites on Ras for lead generation," *PLoS One*, vol. 6, no. 10, p. e25711, 2011.

[76] 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.

Tatiana Maximova is a postdoctoral fellow in the department of Computer Science at George Mason University. She obtained her Ph.D. from Odessa National Polytechnic University and was a researcher at the Ben Gurion University in Israel. Her research interests include computational structural biology, protein biophysics, protein structure prediction, force-fields development, cooperative energy terms, and sampling-based algorithms with a focus on transition modeling.

Erion Plaku is an Assistant Professor in the Department of Electrical Engineering and Computer Science at The Catholic University of America in Washington, DC. He earned his Ph.D. from Rice University, Houston, TX. He was a postdoctoral fellow in the Laboratory for Computational Sensing and Robotics at The Johns Hopkins University, Baltimore, MD. His research focuses on robot motion planning for ground, underwater, and aerial vehicles. Plaku's expertise is in developing probabilistic search algorithms over high-dimensional hybrid spaces consisting of discrete and continuous components.

Amarda Shehu is an Associate Professor in the Department of Computer Science at George Mason University. She earned her Ph.D. from Rice University, Houston, TX. Shehu's research contributions are in computational structural biology, biophysics, and bioinformatics with a focus on issues concerning the relationship between biomolecular function, sequence, equilibrium structures and dynamics. Shehu's research is supported by an NSF CAREER award and various NSF programs, including Intelligent Information Systems, Computing Core Foundations, and Software Infrastructure. Shehu is a member of the IEEE and ACM.