A Stochastic Roadmap Method to Model Protein Structural Transitions

Kevin Molloy[†], Rudy Clausen[†], and Amarda Shehu[†,‡,§*]

[†]Department of Computer Science, George Mason University, Fairfax, VA, USA 22030

[‡]Department of Bioengineering, George Mason University, Fairfax, VA, USA 22030

[§]School of Systems Biology, George Mason University, Fairfax, VA, USA 20110

SUMMARY

Evidence is emerging that the role of protein structure in disease needs to be rethought. Sequence mutations in proteins are often found to affect the rate at which a protein switches between structures. Modeling structural transitions in wildtype and variant proteins is central to understanding the molecular basis of disease. This paper investigates an efficient algorithmic realization of the stochastic roadmap simulation framework to model structural transitions in wildtype and variants of proteins implicated in human disorders. Our results indicate that the algorithm is able to extract useful information on the impact of mutations on protein structure and function.

1 Introduction

The increasingly accepted view of proteins as inherently dynamic systems[1] is raising questions on the role of protein structure in diseases that are proteinopathies. The simplified view of proteins assuming a unique structure to carry out their biological activity[2] allows understanding some protein conformational diseases.[3] In these, the protein is unable to assume its designated function-carrying structure due to internal perturbations (sequence mutations) or external ones in the environment (cellular stress). However, increasing evidence is emerging on enzymes and other proteins making use of a menu of thermodynamically stable or semi-stable structures to modulate their function and participate in numerous complex chemical processes in the cell.[4] Both experiment and com-

*Corresponding Author. Email: amarda@gmu.edu

putation have shown that many proteins switch between such structures, undergoing productive structural displacements of less than an angstrom (Å) or on the order of a few angstroms.[5]

In light of this dynamic view of proteins, it is unclear how mutations that cause or participate in disease affect structure, which is the intermediate link in the relationship between protein sequence and function. There is evidence that some of the most complex human diseases, including cancer, do not arise from the inability of a protein variant to occupy a specific structure but rather from changes to the rate at which the variant transitions between thermodynamically-stable or semi-stable structures.[6] Due to such evidence and the inability of wet-laboratory techniques to elucidate structural transitions in great structural detail, it is central to explore computational techniques to model such transitions and extract information related to kinetics, such as transition rates. While typically kinetic data can be extracted with methods based on Molecular Dynamics (MD), such methods tend to be prohibitively computationally expensive.[7] Many MD simulations need to be launched to sample an ensemble of trajectories over which to calculate any statistics of interest.

In this paper, we propose a computationally-efficient method that is not based on MD but instead builds upon the stochastic roadmap simulation (SRS) framework as detailed by Latombe and colleagues.[8] The proposed method does not obtain structures of a given protein sequence in an MD setting but instead makes use of an efficient stochastic optimization algorithm published in Ref.[9] to sample structures that are local minima of an employed energy function. The method relies on the SRS formalization[8] to build a stochastic roadmap, but rather than doing so over all sampled structures, the method organizes structures into structural states before imposing connectivity information over them. Effectively, the constructed roadmap over states is lazy, as probabilistic edges are added between two nearby structural states to indicate the estimated feasibility of a protein making the structural transition between the states.

By building a stochastic roadmap over structural states, the method is then able to conduct a rich analysis over the roadmap that is not limited to querying the roadmap for the lowest-cost path. Instead, the analogy between the stochastic roadmap and a Markov state model (MSM), facilitated by the employment of structural states, allows extracting interesting statistics, such as the expected

number of edges in computed transition paths between any two states of interest. Such statistics, while not direct measurements of transition rates due to the lack of timescale information from non-MD methods, allow conducting comparisons between wildtype (WT) and variant (mutated) sequences of a protein. Here we employ such statistics to obtain a structural explanation for the role of specific mutations on function in two proteins implicated in human disorders. We investigate the human Superoxide dismutase 1 (SOD1) enzyme, whose sequence mutations have been linked to familial ALS,[10] and the H-RAS protein, whose sequence mutations have been implicated in many human cancers.[11]

In addition to conducting queries and extracting statistics over the lazy roadmap, we pursue here a path smoothing algorithm in order to provide a more detailed structural transition associated with a low-cost path obtained from the roadmap. We adapt for these purposes the conjugate peak refinement (CPR) algorithm originally proposed in.[12] In addition, we investigate the generality of our conclusions on the impact of sequence mutations on structure and energetics when employing two different energy functions, one a representative of hybrid energy functions and another a representative of physics-based functions.

The rest of this paper is organized as follows. In Section 2 we provide a treatment of related work. Details on the proposed method are provided in Section 3. Our experimental analysis is laid out in Section **??**. The paper concludes with a brief discussion in Section 4.

2 Related Work

2.1 Roadmap-based Methods for Modeling Protein Dynamics

Seminal work by Latombe and colleagues proposed the employment of the Probabilistic RoadMap (PRM) framework to plan robot motions over a nearest-neighbor graph of feasible robot configurations.[13] The PRM framework was then adapted to model the binding of a flexible small ligand onto a rigid protein receptor.[14] Many adaptations then followed, notably by Amato and colleagues. In Ref.,[15] the PRM framework was applied for the first time to study folding of a short polypeptide

chain of 10 alanine residues from an extended to an α-helical structure. Notably, the notion of probabilistic edges was introduced in this early work, with the probability of an edge estimated as a function of the Boltzman-related factor $e^{-\Delta E}/K_B T$, with ΔE denoting the energetic difference between two structures, K_B the Boltzmann constant, and T the chosen simulation temperature.

Other adaptations of PRM focused on modeling folding and unfolding events of small proteins as a means of studying and ordering important folding events, such as secondary structure formation.[16] Later adaptations extended applicability to longer protein chains of around 60 amino acids[17] and then 110 amino acids[18] as well as improved computational time demands by focusing sampling over a subset of degrees of freedom selected via rigidity analysis.[19] Other work extended applicability to RNA and additionally extracted kinetics-related measurements, such as relative folding rates.[20–22] A detailed review of PRM-based methods for protein folding is presented in Ref.[23] Adaptations of PRM for studying molecular motions beyond folding of small proteins have proven challenging, particularly due to the difficulty of providing efficient sampling of the relevant structure space.[24] Tree-based variations to model molecular motions beyond protein folding have also been proposed over the years. Such methods are less amenable to calculation of statistics and do not easily allow for large-scale analysis compared to PRM-based adaptations. While a summary of tree-based methods is beyond the scope of this work, a detailed review can be found in Ref.[25]

2.1.1 The Stochastic Roadmap Framework

While seminal ideas on the probabilistic roadmap for protein dynamics had been previously introduced, most notably in the Amato lab,[15,16] it was work in Ref.[8] that formalized the notion of the stochastic roadmap through the SRS framework. In a roadmap, quantitative information is associated with edges in the form of weights. Typically, such weights or costs capture the extent to which the motion of a robot between the edge-connected vertices satisfies underlying constraints. In a roadmap where vertices correspond to structures of a protein under investigation, these costs can measure energetic differences and can be used to determine, for instance, the minimum-cost path connecting two structures of interest. Seminal work by Amato and others[15,16] had already introduced the notion of probabilistic edges, realizing that edges connecting two vertices represent

transitions between two neighboring structures in the structure/conformation space of a protein. In Ref.,[8] this idea was further developed, realizing that if structures were obtained via an MD simulation at some temperature T, then the probability of an edge representing a transition from a vertex u to a vertex v could be measured via the Boltzmann-related Metropolis criterion $e^{-(E(v)-E(u))/(K_B \cdot T)}$. This realization allowed Apaydin, Latombe, and colleagues to see the clear connection between the stochastic (probabilistic) roadmap of structures and an MSM, with vertices seen as states of the MSM and edges between vertices in the roadmap as transitions between states in the MSM.

This analogy between the stochastic roadmap and an MSM also allowed seeing beyond the calculation of minimum-cost paths. the analogy allowed treating the stochastic roadmap as a generative model; hence, the naming of the SRS as a framework to simulate trajectories. Monte Carlo (MC) trajectories could be obtained by simply generating sequences of states from the stochastic roadmap. The correct estimation of transition probabilities was key not only to making the analogy between the stochastic roadmap and an MSM but also to calculating kinetic measurements without launching a single MD simulation.[26]

While it would not be an exaggeration to claim that the work in Ref.[8] was a paradigm shift, the impact of this work was largely confined to the community of robotics researchers working on modeling protein dynamics with robotics-inspired methods. One possible reason was that further formalization of statistics that could be extracted from an MSM was not provided. There was also no detailed discussion of how one could reliably organize structures of the roadmap into structural states in the MSM. Other key questions remained unanswered, including how to associate credible state-state transition probabilities when structures of the roadmap are not obtained via MD simulation but via sampling-based methods that operate without a notion of a physical temperature.

The issue on extracting further information from MSMs that organize MD simulation data would be explicitly addressed in the computational biology community by Pande, Amaro, Noe, Fischer, and others.[27–32] The issue on how to extend the SRS-MSM analogy to settings where structures are not obtained via MD simulations largely remains answered. One of the contributions of the work we present here is to show how one can do so for specific protein systems where dense structural

ensembles are obtained via effective non-MD, sampling-based methods.

Below we provide a review of methods originating in the computational biology community for modeling protein dynamics for the purpose of computing transition paths between structural states relevant for function. This allows placing the proposed SRS-based method here and its contributions in a broader context.

2.2 MD and MSM-MD Methods for Modeling Protein Dynamics

Traditionally, protein dynamics is simulated via MD. MD trajectories are initiated from a structure of interest and conducted until either a time limit, a specified convergence criterion, or the goal structure of interest has been reached; the second setting is expedited when the MD simulation is biased to reach the goal structure within a more reasonable time budget. The biasing can be achieved via different ways. One way is to modify the energy potential to include a penalty term along a coordinate measuring the progress of the trajectory from the start to the goal structure. The issue with biased or steered MD simulations is that they modify the energy surface and can yield a transition trajectory that does not correspond to the true one. For instance, the application of biased MD to capture transitions of Ras between its active and inactive states yielded unrealistic, high-energy structures.[33] Work in[34] introduces a different strategy to bias an MD simulation; specifically, many MD trajectories are launched from a given start structure, but only relevant, productive MD trajectories are further grown; productive movements towards the goal structure are identified by measuring the progress of a specific MD trajectory via RMSD to the given goal structure. Other strategies to expedite an MD simulation elevate a deep basin corresponding to a stable structural state so as to allow the simulation to cross an otherwise very high energy barrier. This strategy is known as the accelerated MD method[35] and has been applied to simulate transitions of Ras between its active and inactive states.[7] While more realistic than biased MD, accelerated MD has also been observed to occasionally get stuck in specific states, failing to report a transition.

An interesting complementary direction is not to rely on long MD trajectories but rather on short, off-equilibrium MD trajectories, which can be simulated in parallel on large-scale, high-performance

computing platforms. The structures from the various, short MD simulations are collected, clustered to identify structural states, and then embedded in an MSM. Transitions are trivially determined. If a structure u obtained at time step t in a particular MD trajectory is followed by a structure v at time step $t+dt$ in the same trajectory, then a directed transition is noted from the state to which u maps to the state to which v maps. The transition counts are tallied up and normalized to associate probabilities with state-state transitions. By now, there are two widely-used software by MD researchers to embed structures in MSMs and analyze transitions. One is EMMA,[36] and the other is MSMBuilder.[29]

The availability of MSM software has allowed the proposal of MD-MSM methods for extracting reliable kinetic information off MD trajectories simulating protein dynamics. MD-MSM methods are widely employed to approximate the underlying folding dynamics of proteins,[27–31] as well as model the kinetics of protein-ligand binding.[32] Detailed reviews of MD-MSM methods for modeling protein dynamics can be found in Refs.[37,38]

MD-MSM methods can have a heavy computational footprint. In general, many MD trajectories are needed to sample enough of the structure space so the MSM is ergodic and captures the relevant dynamics of the system under investigation.[31] In contrast, non-MD methods promise more reasonable computational budgets at the expense of some detail. We review these methods below.

2.3 Morphing and Chain-of-States Methods for Modeling Protein Dynamics

Non-detailed, mechanistic approaches focus on extracting functional modes.[39–42] For instance, geometric morphing methods use the linear interpolation of each atom to construct a path between two structures for which a transition is sought.[43,44] Trajectories based on linear interpolation do not necessarily represent actual transition paths.[45] In response, several, non-linear morphing methods have been developed that provide non-linear interpolations between the start and goal structures. Work based on elastic, plastic network models (ENM, PNM), and their variants falls in this category.[46–55] Non-linear morphing methods rely on the assumption that macromolecules can be treated as deformable elastic bodies, and the interatomic potential function can be represented by

harmonic-type models (ENMs, PNMs, and variants). These methods employ normal mode analysis (NMA) of such models to obtain principal motions of a macromolecule about a local minimum. Since, typically ENMs involve only a single energy minimum and are not immediately applicable to model transitions, mixed ENMs (MENMs)[50,51] and other, related, ENM-based models have been developed.[49,52–55]

Some non-linear morphing methods based on ENMs, MENMs, and PNMs compute transitions that are minimum-energy paths (MEP) in the energy landscape.[49] This is achieved by employing the CPR algorithm,[12] which represents one of the earliest chain-of-states methods for modeling transition paths.

Chain-of-states methods, such as the nudged elastic band (NEB) and string methods, rely on the assumption that a transition trajectory can be encoded as a series/chain of structures (also referred to as a string of images).[56–66] In these methods a string of images is created between the given start and goal structures, and the images are relaxed to the transition trajectory.

While NEB-based methods assume the energy landscape is smooth, string methods do not. However, there are two drawbacks in string methods.

First, their computational cost is high due to the multiple gradient calculations that need to be performed on images located far away from the transition state. Methods are proposed to address this issue, most notably by the Head-Gordon lab.[67–70] In these methods, two strings are grown independently, one from the start and another from the goal structure, until the strings meet.

A second drawback of string methods is their assumption that the flux associated with transition paths is very likely to be concentrated inside one or a few thin (reaction) tubes. This may not be reasonable, particularly for complex macromolecules. To overcome such a limitation, string methods and other chain-of-states methods are combined with enhanced sampling algorithms.[71]

The work presented here can be considered to rely on a similar combination. An algorithm with enhanced sampling capability is used to obtain a broad view of the structure space relevant for the transition. Embedding structures in a connectivity structure (the roadmap) exposes several paths. The CPR, a chain-of-state method, is then used to locally deform these paths to transition

trajectories.

2.4 Contributions of this Work

By comparison to MD-MSM methods, realizations of SRS promise to be more efficient, as in principle they allow the structures embedded in the roadmap to be obtained from non-MD methods. Drawing on the analogy between a stochastic roadmap and an MSM then promises to efficiently extract kinetics-related measurements that can be valuable for the purpose of comparative analysis. For instance, while no time scale information and thus no transition rates can be extracted when structures are obtained via non-MD methods, other related, average statistics can be estimated. We employ one such statistic in this paper, the expected number of state-state transitions (edges) connecting a start to a goal structural state. Comparison of this statistic between the WT and a variant of a protein can provide valuable information on how a sequence mutation impacts the transition of a protein between a reactant and a product state.

It is not exactly clear how to obtain an MSM, which is a kinetic model, from structures obtained via non-kinetic methods following an optimization process to sample local minima of a protein's energy landscape. In this paper, we provide the first steps in this direction. A non-MD algorithm recently developed by us is particularly effective at obtaining dense, sample-based representations of energy landscapes of small-to-medium proteins. Making use of such an algorithm, the work proposed in this paper essentially addresses three key questions: (i) how to map sampled structures to structural states that become states of the MSM; (ii) how to determine which states transition to which states; and (iii) how to associate credible probabilities with designated transitions.

We address the first question via clustering. Essentially, similar structures are grouped together in a cluster. Clusters become states of the MSM. Clustering is also the way structures obtained via MD simulations are grouped to identify states of an MSM in MD-MSM related work.

We address the second question by exploiting the roadmap formalization. States of the MSM can also be viewed as vertices of a roadmap. A state is connected to its k nearest neighbors in the roadmap. Directed edges in the roadmap are then state-state transitions in the MSM.

We address the third question by borrowing from the formalization of transition probabilities in the SRS framework.[8] However, as there is no notion of a physical temperature, we propose a reasonable protocol to define an effective temperature to associate Metropolis-based transition probabilities with state-state transitions. This also requires associating an energy with a state based on energies of the structures mapped to that state.

The roadmap formalization allows obtaining a minimum-cost path as a credible representative of a transition trajectory. The MSM formalization allows obtaining interesting average statistics over all paths connecting two states of interest (protein dynamics is inherently stochastic). One such statistic, the expected number of edges (direct state-state transitions), is particularly interesting. In lieu of actual information on time scales and thus transition rates, comparison of the expected number of edges in a transition between the WT and a variant sequence of a protein provides similar information. We exploit such information here to propose mechanisms by which known mutations impact function in two proteins of central importance to human biology and health.

The work presented here essentially shows how to embed structures obtained via non-MD methods in an MSM. This work is useful and can be seen as representing a meta approach to modeling transitions in proteins. On one end, there are detailed, MD simulations. On the other end are the harmonic-based and chain-of-states approaches focusing on extracting functional modes. The presented work sits in the middle. While different decisions can be made on each of the algorithmic components that allow building an MSM over structures obtained from non-MD methods, the proposed work shows one way to do so and can be regarded as a proof of concept.

3 Methods

The proposed method follows the SRS framework. Briefly, it proceeds in three stages. The first stage samples structures in the search space of interest and is described in Section 3.1. The second organizes these structures into structural states as described in Section 3.2. The third embeds a roadmap over the states and is described in Section 3.3. Once the roadmap is constructed, analysis is conducted on it, as described in Section 3.4. Our analysis consists of path query and calculation

of interesting statistics based upon the casting of the roadmap as an MSM. Given that no local planners are used in the construction of the roadmap, a path smoothing technique is designed and described in Section 3.5 to provide more structural and energetic detail behind queried paths. We now proceed to relate details.

3.1 Stage I: Sampling

One of the key challenges with adaptations of roadmap-based methods for molecular structure and motion computation lies in the sampling stage. Sampling needs to be dense and focus on the relevant regions of the structure space. In this paper, we employ an evolutionary algorithm (EA) to obtain a dense ensemble of structures representing local energy minima in the structure space of interest. We provide an indication of the density of the ensemble in Section **??**. Though a detailed description of this EA is beyond the scope of this paper, we provide here a brief summary, focusing on its salient algorithmic ingredients.

EAs are investigated in detail in our lab in diverse protein modeling scenarios, including *de novo* structure prediction[72,73] and protein-protein docking.[74,75] The EA we employ here has been recently proposed[9] to further populate the structure space of a protein for which many experimental structures already exist in the Protein Data Bank (PDB).[76]

Briefly, the EA leverages the abundance of experimentally-available structures to define the structure space of interest in a lower-dimensional embedding. The algorithm relies on the principle of conformational selection, also known as population shift, which allows understanding that structures caught as stable in the wet laboratory for different sequences of a protein are all populated, albeit with different probabilities, by a given sequence. Hence, all experimentally-available structures of a protein, whether for the WT or variant sequences, are present and possibly represent stable and meta-stable states of a given sequence. This is precious, albeit incomplete, information on the structure space of a protein sequence under investigation. The objective of the EA is to exploit this information to further populate the structure space of a protein sequence.

The EA obtains a lower-dimensional embedding of the structure space of a given protein via Prin-

cipal Component Analysis of CA-traces (using only CA atoms to represent structures[†]) of all available structures for a protein. Note that when stripped down to CA atoms, structures are not specific anymore to any given sequence of the protein under investigation. While PCA is generally not guaranteed to be effective, the EA only proceeds if at least 50% of the variance can be captured with the top two principal components (PCs). This is the case with the protein system we have chosen to investigate in this paper. The EA directly searches in the low-dimensional PC map of m dimensions, ensuring that m PCs are sufficient to capture 90% of the variance in the original structure data.

Starting with an initial population of p structures built on the experimentally-available ones, reproductive operators are used to generate child structures (in a CA trace representation) from parents in the PC map, using sampled perturbations along the available PCs. A multiscaling reconstruction procedure maps a child structure to an all-atom structure representative of a local energy minimum. It is this procedure that makes structures generated by the EA specific to a protein sequence under investigation. In summary, the procedure first reconstructs a backbone from the CA trace of a child, adds side chains, and then minimizes the entire resulting all-atom structure using the Rosetta *relax* protocol (keeping backbone heavy atoms fixed). This procedure ensures that structures obtained by the EA are minima of the all-atom Rosetta *score12* energy function[77] of a given protein sequence under investigation. The resulting minima structures compete with neighboring parents based on their energies, and p winners become parents of the next generation. This proceeds for a certain number g of generations.

It is worth noting that searching in a PC-based embedding and making use of multiscaling have been previously analyzed in detail in the context of a robotics-inspired (tree-based) search algorithm,[78] and these components are integrated in the recently proposed EA[9] we employ in the sampling stage here. The ensemble Ω of structures fed to stage II of the SRS-based method in this paper consists of all the populations of local minima obtained by the EA across all its g generations for a protein sequence at hand.

[†]a CA atom is the principal, backbone carbon atom in an amino acid

3.2 Stage II: Organizing Structures into Structural States

The ensemble Ω potentially contains many structures that are geometrically similar to one another. Therefore, in this stage, the structures in Ω are grouped into structural states both to remove redundancy and to allow constructing a roadmap over these states that can then be treated and analyzed as a Markov state model. We employ a simple unsupervised clustering algorithm, leader clustering,[79] to efficiently group structures into states. That is, a structural state is a cluster.

The leader clustering algorithm has the benefit of not having to specify the number of clusters/states a priori. Its results are dependent on the order in which the data is processed. In this paper, we use a sorted order, ordering first all the structures in the Ω ensemble by their Rosetta energies. This ordering allows the first structure mapped to a new cluster to be the lowest-energy structure over all others that will be mapped to that same cluster. The algorithm proceeds in the sorted order, mapping a structure to one of the existing clusters if its distance to the cluster representative is below a specified cluster radius. Otherwise, a new cluster is created with the unmapped structure as its representative. The algorithm proceeds until all structures have been processed, resulting in a list of $C_1, \ldots, C_l$ clusters/states. The decision on what distance function to use is important. Here we employ least Root Mean Squared Deviation (lRMSD), which is a popular dissimilarity measure to compare protein structures.[80] We do so over only CA atoms of a structure; that is, we use CA lRMSD. We experiment with different values of cluster radii, as presented in Section **??**.

3.3 Stage III: Roadmap Construction

Roadmap construction proceeds over the identified clusters. The roadmap is encoded as a weighted directed graph $G = (V, E)$. A vertex $v \in V$ is created for each of the clusters identified in stage II; that is, vertices encode states over the sampled structure space. Edges are added to the roadmap as follows. Each vertex is connected to up to k_{nn} of its nearest neighbors that are within an ϵ_{nn} CA lRMSD of v. Since vertices correspond to structural states/clusters, the lRMSD comparison is conducted between the cluster representatives. When a vertex u is deemed to be a neighbor of v

that passes the k_{nn} and ϵ_{nn} criteria, two edges are added to the roadmap, (u, v) and (v, u). To improve the connectivity of the roadmap, a final pass across all connected components is performed, adding an edge when the two components can be merged (subject to the same ϵ_{nn} CA lRMSD constraint). Edges are weighted based on the energetic difference between the states the vertices they connect encode. For a directed edge (u, v), its weight P_{uv} measures the probability of a direct transition from u to v. We assign edge weights following closely the original formulation of the SRS in Ref.,[8] per the following equations:

$$P_{uv} = \begin{cases} (1/|N_u|) \cdot e^{-\Delta E_{uv}/\alpha} & \text{if } \Delta E_{uv} > 0 \\ 1/|N_u| & \text{otherwise} \end{cases}$$

$$P_{uu} = 1 - \sum_{u \neq v} P_{uv}$$

For each vertex v, $|N_v|$ represents the number of outgoing edges from v excluding the edge back to itself. The $e^{-\Delta E_{uv}/\alpha}$ factor is a Boltzmann-related factor that mimics the Metropolis criterion for accepting the energetic transition from state u to state v. Note that $\Delta E_{uv} = E(v) - E(u)$, where $E(u)$ and $E(v)$ are the energies of the structures corresponding to vertices u and v, respectively. There are two important decisions that need to be made. First, since vertices here encode structural states, how is the energy of a state measured? Second, how is a reasonable value for the α parameter estimated? Our specific choices for these two design decisions are important adaptations of the original SRS formulation on weighting directed transitions.

3.3.1 Energy of a State

Theoretically, if the states correspond to energetic states, one should measure $E(u)$ as the free energy F of the state u. This can be estimated, in theory, as $F(u) = \langle E \rangle_u - \alpha \cdot ln(|C_u|)$, where $\langle E \rangle_u$ captures the average energy over all structures in the state u, and $|C_u|$ measures the number of structures in u. However, in practice, an accurate estimate requires a theoretically-sound definition of a state. By employing clustering to define states as clusters, as we do in this work, one cannot guarantee energetic homogeneity in addition to structural homogeneity of a deemed cluster/state.

The distinction is important, as a rigorous calculation of pseudo-free energy requires that the average energy $\langle E \rangle_u$ of a state u be descriptive of the distribution of energies of the structures mapped to u. Clustering based on structural similarity cannot make such guarantees. Potential alternative groupings of structures may include finer clustering by considering energetic similarity in addition to structural similarity. However, this introduces even more parameters that need to be set to determine such similarity and our own work indicates that the results are not more accurate than associating with each cluster the lowest-energy over structures mapped onto it. In this paper, we associate with a cluster/state the energy of the cluster representative, which is the lowest-energy structure in a cluster due to the energy-sorted order in which structures are processed in the clustering algorithm described above.

3.3.2 Weighting of Edges in the Roadmap:

Effective Temperature as a Scaling Parameter

The parameter α replaces the $K_B \cdot T$ term that scales the energetic difference between two states in Ref.[8] This is necessary, as the EA employed to obtain structures here is not an MD-based algorithm and thus does not make use of any physical temperature. Typically, in MD simulations, the simulation is conducted at a chosen temperature. When studying transitions, equilibrium (room) temperature is specified. In the non-MD setting in this paper, there is no physical meaning for temperature. In addition, employment of a physical temperature assumes an energy function with units of kcal/mol. Instead, one of the functions employed in this paper combines physics-based terms with knowledge-based ones. In particular, the Rosetta *score12* energy function used by the EA is a hybrid function in *Rosetta Energy Units*. As a result, the $K_B \cdot T$ term needs to be adapted if a Metropolis-like criterion is to be employed to determine the probability of a transition between two edge-connected states/vertices.

So, we rewrite here the Boltzmann-related probability as $e^{(-\delta E/\alpha)}$, where α is an energy-scaling parameter in energy units. Determining the value for α is an important decision. A high value results in high transition probabilities even for large δE values; the transition crosses large energetic

barriers. A low value makes it unlikely that the transition will cross large energetic barriers. Since α directly determines the magnitude of the energetic barrier crossed by a transition, this parameter can be seen as a knob to scale an energy barrier between two states; hence, the designation as a scaling parameter.

A decision on a meaningful value for α is critical to the accuracy of any analysis on lowest-cost paths or average statistics. In other works, an initial value for α is set arbitrarily and then updated in a reactive scheme to allow MC simulations and other robotics-inspired exploration to balance between crossing energy barriers and drilling down energy basins.[81,82] In this paper, we propose a novel protocol to assign a reasonable value to α. The protocol is based on analysis on what energetic barriers a system can easily jump as an indirect way of associating an effective (note, not physical) temperature parameter.

The analysis is based on statistical mechanics. We measure the energetic variance over structures that the Rosetta *score12* energy function reports to be in the same energy basin. The assumption is made that a system should readily exchange/diffuse between structures in a basin within thermal vibrations. We restrict the analysis over the distribution of structures obtained by the Rosetta *relax* protocol when minimizing the same crystal structure many times to map the depth of a basin. The *relax* protocol is based on simulated annealing, so different structures can be obtained, thus providing a view of the basin where the Rosetta energy function maps a given crystal structure. We conduct this analysis various times, over different crystal structures and observe the energetic variance in δE_{basin}; the basin depth δE_{basin} is recorded as the difference between the maximum and minimum energies obtained with the *relax* protocol. We note that such analysis is protein-dependent and needs to be conducted separately on each protein systems studied in order to find a reasonable value for α for each system.

Based on a statistical mechanics treatment, structures in the same basin should exchange into one another with high probability. Let us refer to this probability as a target probability t_{prob}. Rather than directly setting values for α, we derive α based on a user-defined value for t_{prob}. Given t_{prob}, α is derived by solving the equation $e^{-\delta E_{\text{basin}}/\alpha} = t_{\text{prob}}$; so, $\alpha = -\delta E_{\text{basin}}/ln(t_{\text{prob}})$. This formulation

supports $\alpha \to \infty$ as $t_{\text{prob}} \to 1$, which means that a probability of acceptance of 1, which allows exchange between any two structures regardless of their energetic difference, corresponds to a very high temperature; in contrast, for a given δE_{basin}, lower exchange probabilities give smaller negative values for the denominator, which result in smaller values for α (analogous to lower temperatures). We note that the actual value for α is also dependent on the energy function employed and requires that a target probability be specified a priori, but the process outlined here is general.

3.3.3 Construction of Lazy Stochastic Roadmap

Each edge in the stochastic roadmap G now encodes a potential transition between two structural states. In this work, we employ a "lazy" strategy that avoids the computation of these transitions and instead focuses on the global connectivity. This has some similarities to the Lazy PRM.[83] We note, however, that foregoing a local planner is made possible here because of the stringent criterion of structural proximity ϵ_{nn} when considering connecting two vertices via an edge. This in itself exploits the dense structural sampling afforded by the EA employed in stage I.

We note that, by construction, G consists of a set of strongly connected components (SCCs); when $\epsilon_{nn} = \infty$, G consists of a single SCC. As demonstrated in Ref.,[8] a random walk in G can be interpreted as a discretized version of a Monte Carlo trajectory. More importantly, various analyses can be conducted over the roadmap to obtain path-ensemble averages without launching Monte Carlo simulations, as the roadmap encodes multiple such trajectories.

3.4 Roadmap Analysis

Treating the constructed stochastic roadmap as a graph allows using path search algorithms to obtain paths connecting structural states of interest. Treating the roadmap as a Markov state model allows using transition state theory to obtain measurements approximating kinetic quantities of interest.

3.4.1 Querying the Roadmap

As demonstrated in the original proposal of the PRM method in Ref.,[13] the roadmap can be queried given two states of interest. Dijkstra's algorithm can be used to obtain a shortest path. Here, edges are weighted by the probabilities of transition. The negative logarithms of these probabilities can be employed to obtain a minimum-cost path. In addition to such a path, more information can be obtained by analyzing not just one path but several. Yen's K-shortest paths algorithm[84] can be employed for this purpose.

3.4.2 Treating the Roadmap as a Markov State Model

The roadmap G can be treated as a Markov state model encoding the stochastic behavior of the system being studied. In this paper, we use the roadmap to model the structural transitions between functionally-relevant states of a protein and understand how these transitions are affected by sequence mutations. For this purpose, the roadmap G is analyzed to determine the expected number of edges across all transition paths allowing a protein to switch from one structural state to another.

Recall that structural states are vertices in the vertex set V in our roadmap G. For each vertex $v_i \in V$, one can utilize first-step analysis theory to measure the expected number of edges t_i from vertex v_i to some specific vertex of interest. As demonstrated in Ref.,[8] random walks need not be performed to obtain such a measure, as a closed-form solution can be computed via a linear solver. The formulation of t_i is recursive. Let us generalize and state that the goal is to measure the expected number of edges from some vertex v_i to a set of vertices $v_j \in A$, where A is a subset of V that does not include v_i (A is in an SCC). Then, provided that v_i and A are in the same SCC:

$$t_i = 1 + \sum_{v_j \in A} P_{ij} \cdot 0 + \sum_{v_j \notin A} P_{ij} \cdot t_j \quad \forall\, v_i \notin A$$

This results in a system of equations that is the same order as the number of vertices in the SCC containing the functionally-relevant states of interest in the roadmap. Since clustering of structures into structural states reduces the number of vertices in the roadmap, an exact solver (as opposed

to a slow-converging iterative solver) can be afforded, and that is what we employ in this paper to solve the linear system above algebraically and obtain t_i for all the vertices simultaneously.

In this paper, we are specifically interested in measuring the expected number of edges across all present transition paths from a given structural state to another given structural state. Such states may be critical to the ability of a protein to function normally. By repeating the sampling, clustering, roadmap construction, and analysis on different sequence variants, we are then able to compare the expected number of edges in transitions between two states of interest in the WT versus disease-participating variants of a protein of interest.

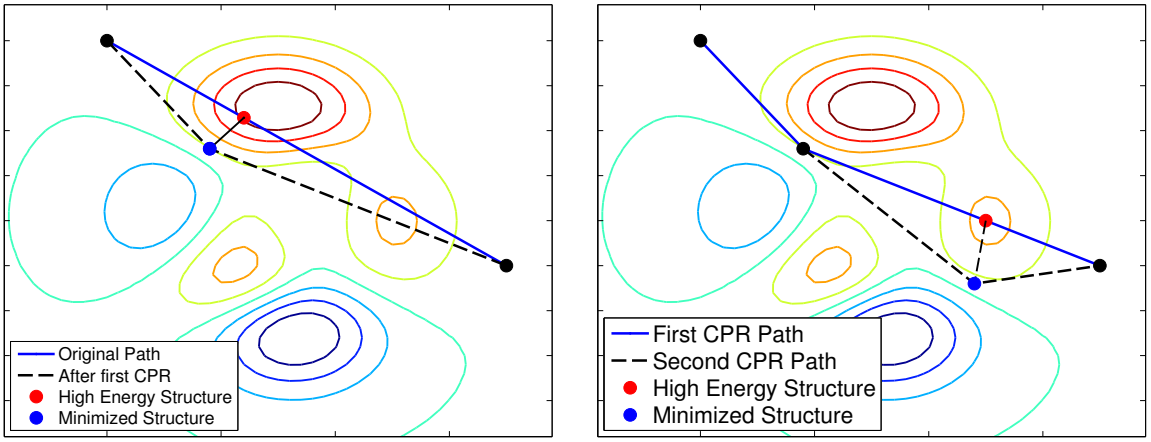
3.5 Path Smoothing and Analysis

The path smoothing procedure we employ is an adaptation of the conjugate peak refinement (CPR) algorithm originally introduced in Ref.[12] Briefly, CPR produces a series of intermediate conformations to approximate a potential reaction path between two given (start and goal) conformations p and r. The initial guess of the path is a straight line interpolation between p and r. We ensure in our adaptation that the distance between two consecutive conformations in the straight line interpolation does not exceed a parameter l. The original CPR algorithm is illustrated in Fig. 1.

In the original CPR algorithm, the highest-energy conformation is identified among the ones resulting from the interpolation. Instead, in our adaption, we identify the conformation x_q that represents the largest jump in energy between two adjacent conformations in the current straight line path. This conformation is then minimized to obtain x_q^*. This results in two path segments, $[p\ x_q^*]$ and $[x_q^*\ r]$. The process now repeats to identify (and then minimize) the conformation representing the largest energetic jump over the existing path segments continues until the series of resulting conformations from this procedure are energetically feasible.

Energetic feasibility is evaluated as follows. A short Markov chain is constructed, where states correspond to the conformations (existing structures and new ones produced by CPR). Adjacent states in the chain are connected and probabilities are assigned utilizing the Metropolis criterion described earlier. In this way, the CPR chain of structures can be regarded as a Markov chain, and a

Fig. 1: A cartoon example of the CPR algorithm. The energy surface is projected on two hypothetical collective variables. The left panel shows the initial interpolated path in blue, with the highest energy conformation shown in red. This conformation undergoes an energy minimization, resulting in the blue point. A new path is now constructed via the blue point. The right panel illustrates the next iteration of the algorithm.



(a) (b)

random walk can be performed over the chain (in this work, of length 1000). If the end of the path is reached at any point during this walk, the path is deemed energetically acceptable. Otherwise, application of CPR continues in order to identify more intermediate conformations as described above.

In its original form, CPR requires an energy function that is continuous and for which the first derivative can thus be defined. As an alternative, we employ Rosetta's *relax* protocol, which performs a simulated-annealing minimization after adding side chains to backbone-resolution conformations. In our employment of the Rosetta *relax* protocol, we constrain the movement of backbone atoms so the minimized conformation x_q^* lies nearby the one prior to minimization x_q. The pseudocode of the algorithm is shown in Algorithm 1.

We make use of CPR as follows. A path returned from a query on the stochastic roadmap is realized by having the endpoints of each of its edges supplied as input to CPR. Since CPR is an interpolation-based algorithm, we set its termination criterion to be a resolution of ℓÅ , where ℓ is the minimum distance between two consecutive structures produced by CPR.

We note that CPR is related to an interpolation-based path planner. However, the interpolation is

Algorithm 1 The Conjugate Peak Refinement algorithm[12]

Input:
Function states C_s, C_t
ϵ, interpolation interval
Output: Path $C_s, C_1, C_2, ..., C_n, C_t$

```
P ← InterpolateInitialPath(ε)
while Time AND EnergyOK = FALSE do
    H ← SelectHighDeltaEnergy(P)
    H_Min ← Miniimize(H)
    P ← SegmentPath(P,H,H_Min,ε)
end while
```

not over the straight line connecting two structures, as described above. We only employ CPR not to realize each edge in the roadmap but instead to provide more energetic and structural detail with a low-cost path extracted from the lazy stochastic roadmap. Moreover, we elect to deem it a path smoothing algorithm in order to retain analogies with path smoothing algorithms in algorithmic robot motion planning, where such algorithms are often used to improve the satisfaction of present constraints in a computed path. CPR maps a high-energy intermediate structure into a local minimum. As a result, intermediate motions better satisfy the present energetic constraints.

3.6 Implementation Details

The method is implemented in C++. The EA in the sampling stage runs for g=100 generations, with p=500 structures in a population. Thus, the ensemble of structures Ω fed to the clustering stage contains $50,500$ structures. It takes 48 days of CPU time on a single 2.66 GHz Opteron processor with 24 GB of memory to obtain this ensemble. Various cluster radii are investigated in the clustering stage. For the results shown in this paper, a cluster radius of 0.35Å is used to provide a compromise between a reduction in the number of clusters and structural homogeneity within a cluster. The clustering stage takes approximately 7 hours of CPU time. Parallelizing reduces the run time to just under 45 minutes on a 64 core AMD Opteron processor with 542 GB of memory. This same hardware is used to perform the roadmap construction and analysis, which each execute in approximately 60 minutes. While various values for k_{nn} and e_{nn} are investigated for how they affect connectivity, the results related here are obtained with k_{nn}=30 for SOD1 and k_{nn}=20 for Ras.

The value for ϵ_{nn} is 0.65Å for both systems of study. We note that the selection of this value for ϵ_{nn} is slightly less than twice the cluster radius. In determining a reasonable value for the effective temperature α per the process described above, we err here on the conservative side and set tprob to 0.25. In the CPR-based path smoothing algorithm, we set l to be 0.3Å.

4 Conclusion

This paper has proposed an efficient realization of the SRS framework to model structural transitions in dynamic proteins involved in proteinopathies. Central to the ability of the method to capture the connectivity of the thermodynamically-available structure space of a protein in this paper is the employment of a powerful evolutionary algorithm in the sampling stage. Organization of sampled structures into structural states and lazy evaluation of edges in the roadmap are not only critical to computational efficiency but also important to treat the resulting stochastic roadmap as an MSM. The latter allows employing calculations based on transition state theory to estimate transition rates between structural states of interest.

Given that the method does not employ MD, no timescale information can be extracted from the roadmap. However, important statistics, such as expected number of transitions between two states provide indirect estimates of kinetics-related measurements, such as transition rates. A higher expected number of transitions relates to a lower transition rate, whereas a low expected number of transitions relates to a higher transition rate. Under such relationship, while statistics obtained from analysis of the roadmap are not meaningful as absolute measurements, they are precious in a comparative setting, where the goal is to compare transition rates between WT and variant sequences of a protein to elucidate the impact of mutations on protein function.

In particular, application of the proposed method in a comparative setting to model and compare transitions in the WT and selected variants of two important proteins, SOD1 and Ras, shows promising results. The method is able to provide a structural basis for how mutations affect protein function. Analysis of results obtained when employing different energy functions indicates that reliable conclusions can be reached that are not dependent on a specific energy functions.

The work presented here constitutes a first step into obtaining a better understanding of the role of structure in the complex relationship between protein sequence and function in the healthy and diseased cell. Unraveling the mechanisms of oncogenic mutations is stated as a critical element in genetics-based decision-making in cancer treatments.[85] Mechanistic insight into the conformational behavior of molecules in isolation and complex may provide a foundation for the structural basis of cancer decisions. Such insight may be valuable, for instance, for exposing novel allosteric sites for lead generation.[86] Elucidating transient structures may facilitate the design of small ligands to stabilize these structures and so decrease the pool of activated proteins through a population shift mechanism.[5]

Future work will continue to investigate the algorithmic richness of the SRS framework in order to improve both accuracy and efficiency in protein structure modeling for the purpose of unraveling the role of protein structure and energetics in proteinopathies. Particular directions to investigate include extending applicability to longer protein systems with potentially larger structural transitions. The measurement of pseudo-free energies is also worth investigating, but this will require investigation of the definition of a state to ensure its usefulness for extracting reliable statistics.

Finally, the balance between spending computational resources to obtain a global albeit coarse view of state connectivity vs. a local but detailed view afforded by often expensive local planners is a worthy issue for investigation. Directions in algorithmic robotics that shift the focus from obtaining an ensemble paths a priori to path sampling on demand, as in fuzzy PRM, may prove useful in this direction. A preliminary investigation of such approaches is available in Ref.,[87] but further work is needed to exploit the analogies with MSMs that have proven so useful in this paper in extracting summary statistics for structural transitions in proteins.

Acknowledgement

Many of these experiments 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). Funding for this work is provided in part by the National Science Foundation (Grant Nos. 1421001,

1440581, and CAREER Award No. 1144106) and the Thomas F. and Kate Miller Jeffress Memorial Trust Award.

References

1. K. Jenzler-Wildman and D. Kern. Dynamic personalities of proteins. *Nature*, 450:964–972, 2007.

2. C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

3. C. Soto. Protein misfolding and neurodegeneration. *JAMA Neurology*, 65(2):184–189, 2008.

4. D. D. Boehr, D. McElheny, J. Dyson, and P. E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313(5793):1638–1642, 2006.

5. D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol*, 5(11):789–96, 2009.

6. A. Fernández-Medarde and E. Santos. Ras in cancer and developmental diseases. *Genes Cancer*, 2(3):344–358, 2011.

7. B. J. Grant, A. A. Gorfe, and J. A. McCammon. Ras conformational switching: Simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS Comp Biol*, 5(3):e1000325, 2009.

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

9. R. Clausen and A. Shehu. A multiscale hybrid evolutionary algorithm to obtain sample-based representations of multi-basin protein energy landscapes. In *ACM Conf on Bioinf and Comp Biol (BCB)*, pages 269–278, Newport Beach, CA, September 2014.

10. R. A. Conwit. Preventing familial ALS: a clinical trial may be feasible but is an efficacy trial warranted? *J Neurol Sci*, 251(1-2):1–2, 2006.

11. Antoine E. Karnoub and Robert A. Weinberg. Ras oncogenes: split personalities. *Nature Reviews Molecular Cell Biology*, 9:517–531, 2008.

12. S. Fischer and M. Karplus. Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chemical Physics Letters*, 194(3):252–261, June 1992.

13. L. E. Kavraki, P. Svetska, J.-C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Autom.*, 12(4):566–580, 1996.

14. A. P. Singh, J.-C. Latombe, and D. L. Brutlag. A motion planning approach to flexible ligand binding. In R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes, and R. Zimmer, editors, *Proc Int Conf Intell Sys Mol Biol (ISMB)*, volume 7, pages 252–261, Heidelberg, Germany, 1999. AAAI.

15. G. Song and N. M. Amato. A motion-planning approach to folding: From paper craft to protein folding. Technical Report TR00-001, Department of Computer Science, Texas A & M University, January 2000.

16. 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.*, 10(3-4):239–255, 2002.

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

18. S. Thomas, G. Song, and N. M. Amato. Protein folding by motion planning. *J. Phys. Biol.*, 2(4):148, 2005.

19. S. Thomas, X. Tang, L. Tapia, and N. M. Amato. Simulating protein motions with rigidity analysis. *J. Comput. Biol.*, 14(6):839–855, 2007.

20. L. Tapia, X. Tang, S. Thomas, and N. Amato. Kinetics analysis methods for approximate folding landscapes. *Bioinformatics*, 23:i539i548, 2007.

21. X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. Amato. Simulating rna folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, 381(4):1055–1067, 2008.

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

23. M. Moll, D. Schwartz, and L. E. Kavraki. Roadmap methods for protein folding. *Methods Mol. Biol.*, 413:219–239, 2008.

24. K. Molloy and A. Shehu. A probabilistic roadmap-based method to model conformational switching of a protein among many functionally-relevant structures. In *Intl Conf on Bioinf and Comp Biol (BICoB)*, Las Vegas, NV, 2014.

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

26. T. H. Chiang, D. Hsu, and Latombe. J. C. Markov dynamic models for long-timescale protein motion. *Bioinformatics*, 26(12):269–277, 2010.

27. N. Singhal, C. D. Snow, and V. S. Pande. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121(1):415–425, 2004.

28. K. A. Beauchamp, D. L. Ensign, R. Das, and V. S. Pande. Quantitative comparison of villin headpiece subdomain simulations and triplettriplet energy transfer experiments. *Proc. Natl. Acad. Sci. USA*, 108(31):12734–12739, 2011.

29. K. A. Beauchamp, G. R. Bowman, T. J. Lane, L. Maibaum, I. S. Haque, and V. S. Pande. MSMBuilder2: Modeling conformational dynamics at the picosecond to millisecond scale. *J Chem Theory Comput*, 7(10):3412–3419, 2011.

30. F. Noé, C. Schutte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA*, 106(45):19011–19016, 2009.

31. Robert D Malmstrom, Christopher T Lee, Adam T Van Wart, and Rommie E Amaro. Application of molecular-dynamics based markov state models to functional proteins. *Journal of chemical theory and computation*, 10(7):2648–2657, 2014.

32. M. Held and F. Noé. Calculating kinetics and pathways of protein-ligand association. *Eur J Cell Biol*, 91(4):357–364, 2012.

33. J. Ma and M. Karplus. Molecular switch in signal transduction: reaction paths of the conformational changes in ras p21. *Proc. Natl. Acad. Sci. USA*, 94(22):11905–11910, 1997.

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

35. D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys*, 120(24):11919–11929, 2004.

36. M Senne, B. Trendelkamp-Schroer, A. S. Mey, C. Schütte, and F. Noé. EMMA: A software package for markov model building and analysis. *J Chem Theory Comput*, 8(7):2223–2238, 2012.

37. G. R. Bowman, V. S. Pande, and F. Noé. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer, Heidelberg, Germany, 2014.

38. B. Gipson, D. Hsu, L. E. Kavraki, and J.-C. Latombe. Computational models of protein kinematics and dynamics: Beyond simulation. *Annu. Rev. Anal. Chem.*, 5:273–291, 2012.

39. J. S. Hub and B. L. de Groot. Detection of functional modes in protein dynamics. *PLoS Comp Biol*, 5(8):e1000480, 2009.

40. Q. Ciu and I. Bahar. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. CRC Press, 1st edition, 2005.

41. R. Bahar and A. J. Rader. Coarse-grained normal mode analysis in structural biology. *Curr. Opinion Struct. Biol.*, 204(5):1–7, 2005.

42. I. Bahar, T. R. Lezon, L. W. Yang, and E. Eyal. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys*, 39:23–42, 2010.

43. W. G. Krebs and M. Gerstein. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res*, 28(8):1665–1675, 2000.

44. Nathaniel Echols, Duncan Milburn, and Mark Gerstein. Molmovdb: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res*, 31(1):478–482, 2003.

45. Dahlia R Weiss and Patrice Koehl. Morphing methods to visualize coarse-grained protein dynamics. In *Protein Dynamics*, pages 271–282. Springer, 2014.

46. M. K. Kim, G. S. Chirikjian, and R. L. Jernigan. Elastic models of conformational transitions in macromolecules. *J Mol Graph Model*, 21(2):151–160, 2002.

47. K. M. Kim, R. L. Jernigan, and G. S. Chirikjian. Efficient generation of feasible pathways for protein conformational transitions. *Biophys. J.*, 83(3):1620–1630, 2002.

48. 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.*, 74(3):760–776, 2009.

49. P. Maragakis and M. Karplus. Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J. Mol. Biol.*, 352(4):807–822, 2005.

50. Protein conformational transitions explored by mixed elastic network models. Protein conformational transitions explored by mixed elastic network models. *Proteins: Struct. Funct. Bioinf.*, 69(1):43–57, 2007.

51. F. Zhu and G. Hummber. Gating transition of pentameric ligand-gated ion channels. *Biophys J*, 97(9):2456–2463, 2009.

52. O. Miayshita, J. N. Onuchic, and P. G. Wolynes. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. USA*, 100(22):12570–12575, 2003.

53. O. Miayshita, P. G. Wolynes, and J. N. Onuchic. Simple energy landscape model for the kinetics of functional transitions in proteins. *J Phys Chem B*, 5(1959-1969):109, 2005.

54. J. W. Chu and G. A. Voth. Coarse-grained free energy functions for studying protein conformational changes: a double-well network model. *Biophys J*, 93(11):3860–3871, 2007.

55. J. Franklin, P. Koehl, S. Doniach, and M. Delarue. MinActionPath: maximum likelihood trajectory for large-scale structural transitions in a coarse-grained locally harmonic energy landscape. *Nucleic Acids Res*, 35(Web Server issue):W477–W482, 2007.

56. R. Elber and M. Karplus. A method for determining reaction paths in large molecules: Application to myoglobin. *Chem Phys Lett*, 139(5):375–380, 1987.

57. G. Henkelmann and H. Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J Chem Phys*, 113:9978–9985, 2000.

58. E. Weinan, W. Ren, and E. Vanden-Eijnden. String method for the study of rare events. *Phys Rev B*, 66:052301, 2002.

59. J. W. Chu, B. L. Trout, and C. L. III Brooks. A super-linear minimization scheme for the nudged elastic band method. *J. Chem. Phys.*, 119(24):12708–12717, 2003.

60. Matthias U Bohner, Johannes Zeman, Jens Smiatek, Axel Arnold, and Johannes Kästner. Nudged-elastic band used to find reaction coordinates based on the free energy. *J Chem Phys*, 140(7):074109, 2014.

61. R. Olender and R. Elber. Yet another look at the steepest descent path. *J Mol Struct THEOCHEM*, 398-399:63–71, 1997.

62. L. Maragliano, A. Fiser, E. J. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.*, 125:024106, 2006.

63. Luca Maragliano and Eric Vanden-Eijnden. On-the-fly string method for minimum free energy paths calculation. *Chem Phys Lett*, 446(1):182–190, 2007.

64. Weiqing Ren and Eric Vanden-Eijnden. Finite temperature string method for the study of rare events. *J Phys Chem B*, 109(14):6688–6693, 2005.

65. Weiqing Ren, Eric Vanden-Eijnden, Paul Maragakis, and E Weinan. Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide. *J Chem Phys*, 123(13):134109, 2005.

66. E. Weinan, W. Ren, and E. Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, 126:164103, 2007.

67. A. Goodrow, A. T. Bell, and M. Head-Gordon. Development and application of a hybrid method involving interpolation and ab initio calculations for the determination of transition states. *J Chem Phys*, 129(17):174109/1–174109/12, 2008.

68. A. Goodrow, A. T. Bell, and M. Head-Gordon. Transition state-finding strategies for use with the growing string method. *J Chem Phys*, 130(24):244108/1–244108/14, 2009.

69. A. Goodrow, A. T. Bell, and M. Head-Gordon. A strategy for obtaining a more accurate transition state estimate using the growing string method. *Chem Phys Lett*, 484(4-6):392–398, 2010.

70. A. Behn, P. M. Zimmerman, A. T. Bell, and M. Head-Gordon. Efficient exploration of reaction paths via a freezing string method. *J Chem Phys*, 135(22):224108–224116, 2011.

71. Wenxun Gan, Sichun Yang, and Benoît Roux. Atomistic view of the conformational activation of src kinase using the string method with swarms-of-trajectories. *Biophys J*, 97(4):L8–L10, 2009.

72. B. Olson and A. Shehu. Efficient basin hopping in the protein energy surface. In *IEEE Intl Conf on Bioinf and Biomed*, Philadelphia, PA, October 2012. 119-124.

73. B. Olson, K. A. De Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In *Conf on Genetic and Evolutionary Computation (GECCO)*, New York, NY, 2013. ACM.

74. I. Hashmi and A. Shehu. Informatics-driven protein-protein docking. In *ACM Conf on Bioinf and Comp Biol Workshops (BCBW)*, pages 772–779, Washington, D. C., September 2013.

75. 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*, 2012(674832), 2012.

76. H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, 10(12):980–980, 2003.

77. Kristian W. Kaufmann, Gordon H. Lemmon, Samuel L. DeLuca, Jonathan H. Sheehan, and Jens Meiler. Practically useful: What the rosetta protein modeling suite can do for you. *Biochemistry*, 49(14):2987–2998, 2010.

78. R. Clausen and A. Shehu. Exploring the structure space of wildtype ras guided by experimental data. In *ACM Conf on Bioinf and Comp Biol Workshops (BCBW)*, pages 757–764, Washington, D. C., September 2013.

79. J.A. Hartigan. *Clustering Algorithms*. John Wiley and Sons, New York, 1975.

80. A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.*, 26(6):656–657, 1972.

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

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

83. R. Bohlin and Lydia E. Kavraki. Path planning using lazy prm. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 1, pages 521–528, San Fransisco, CA, April 2000. IEEE Press, IEEE Press.

84. M. Pascoal and E. Martins. A new implementation of Yen's ranking loopless algorith. *Quart J of the Belgian, French and Italian Oper Res Soc*, 1(2):121–133, 2003.

85. R. Nussinov, H. Jang, and C.-J. Tsai. The structural basis for cancer treatment decisions. *Oncotarget*, 5(17):7285–7302, 2014.

86. 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*, 6(10):e25711, 2011.

87. K. Molloy. *Probabilistic Algorithms for Modeling Protein Structure and Dynamics*. PhD thesis, George Mason University, 2015.