Robotica (2016) volume 34, pp. 1705–1733. © Cambridge University Press 2015
doi:10.1017/S0263574715001058

A stochastic roadmap method to model protein structural transitions

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

†*Department of Computer Science, George Mason University, Fairfax, 22030 VA, USA*
‡*Department of Bioengineering, George Mason University, Fairfax, 22030 VA, USA*
§*School of Systems Biology, George Mason University, Fairfax, 20110 VA, USA*

(Accepted November 13, 2015. First published online: December 11, 2015)

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.

KEYWORDS: Protein dynamics; Structural transitions; Conformational switching; SOD1 and Ras variants; Human disorders; Stochastic roadmap simulation; Markov state model; First-step analysis; Path smoothing; Energy landscape.

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 computation 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

* Corresponding author. E-mail: amarda@gmu.edu
E-mails: kmolloy1@gmu.edu, rclausen@gmu.edu

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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 Refs. [9] and [10] 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,[11] and the H-RAS protein, whose sequence mutations have been implicated in many human cancers.[12]

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 Ref. [13]. 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 4. The paper concludes with a brief discussion in Section 5.

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.[14] The PRM framework was then adapted to model the binding of a flexible small ligand onto a rigid protein receptor.[15] Many adaptations then followed, notably by Amato and colleagues. In Ref. [16], 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.[17] Later adaptations extended applicability to longer protein chains of around 60 amino acids[18] and then 110 amino acids[19] as well as improved computational time demands by focusing sampling over a subset of degrees of freedom selected via rigidity analysis.[20] Other work extended applicability to RNA and additionally extracted kinetics-related measurements, such as relative folding rates.[21–23] A detailed review of PRM-based methods for protein folding is presented in Ref. [24]. 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.[25,26] 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. [27].

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

2.1.1. The stochastic roadmap framework. While seminal ideas on the PRM for protein dynamics had been previously introduced, most notably in the Amato lab,[16,17] 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[16,17] 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.[28]

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.[29–34] 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.[35] Work in Ref. [36] 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

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

the progress of a specific MD trajectory via Root Mean Squared Deviation (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[37] 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 $t + \mathrm{d}t$ 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,[38] and the other is MSMBuilder.[31]

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,[29–33] as well as model the kinetics of protein-ligand binding.[34] Detailed reviews of MD–MSM methods for modeling protein dynamics can be found in Refs. [39, 40].

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.[33] 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.[41–44] 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.[45,46] Trajectories based on linear interpolation do not necessarily represent actual transition paths.[47] 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.[48–57] 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)[52,53] and other, related, ENM-based models have been developed.[51,54–57]

Some non-linear morphing methods based on ENMs, MENMs, and PNMs compute transitions that are minimum-energy paths (MEP) in the energy landscape.[51] This is achieved by employing the CPR algorithm,[13] 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).[58–68] 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.[69–72] In these methods, two strings are grown independently, one from the start and another from the goal structure, until the strings meet.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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.[73]

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.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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 4. 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[74–76] and protein–protein docking.[77–79] The EA we employ here has been recently proposed[9,10] to further populate the structure space of a protein for which many experimental structures already exist in the Protein Data Bank (PDB).[80] A detailed analysis of the energy landscapes obtained from the algorithm on various H-Ras sequences has also been published recently.[81]

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 metastable 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 Principal Component Analysis of CA-traces (using only CA atoms to represent structures[(1)]) 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[82] 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.

[(1)] A CA atom is the principal, backbone carbon atom in an amino acid.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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,[83] and these components are integrated in the recently proposed EA[9,10] 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.

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 an MSM. We employ a simple unsupervised clustering algorithm, leader clustering,[84] 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.[85] 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 4.

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:

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

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

For each vertex v, $|N_v|$ represents the number of outgoing edges from v excluding the edge back to itself. The $\mathrm{e}^{-\Delta E_{uv}/\alpha}$ 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.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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.[86,87] 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.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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.[88] 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 an MC trajectory. More importantly, various analyses can be conducted over the roadmap to obtain path-ensemble averages without launching MC 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 MSM 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. [14], 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[89] can be employed for this purpose.

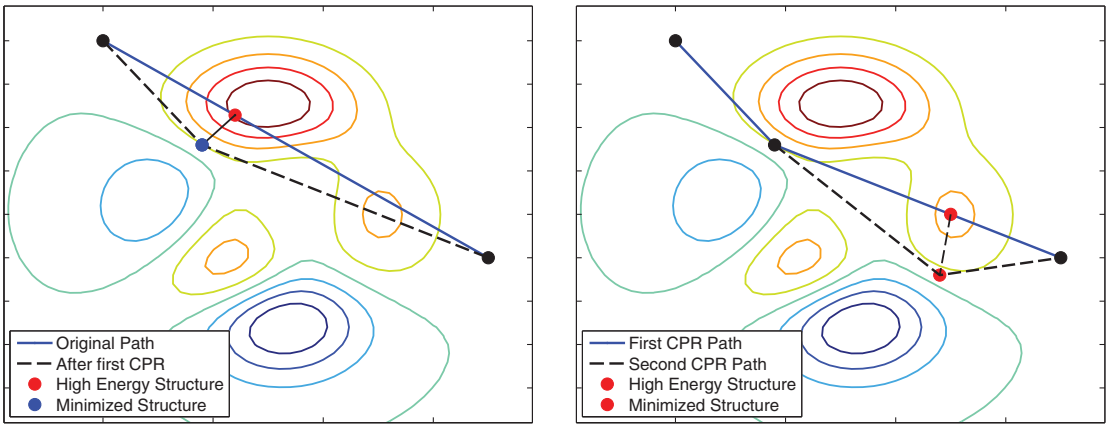
3.4.2. Treating the roadmap as a Markov state model. The roadmap G can be treated as an MSM encoding the stochastic behavior of the system being studied. In this paper, we use the roadmap to model the structural transitions between functionally relevant states of a protein and understand how these transitions 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.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press



(a) (b)

Fig. 1. A cartoon example of the CPR algorithm. The energy surface is projected on two hypothetical collective variables. (a) 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. (b) The right panel illustrates the next iteration of the algorithm.

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 CPR algorithm originally introduced in Ref. [13]. 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 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 pseudo-code of the algorithm is shown in Algorithm 1.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

Algorithm 1 The Conjugate Peak Refinement algorithm[13]

Input:
Function states C_s, C_t
ϵ, interpolation interval
Output: Path $C_s, C_1, C_2, \ldots, C_n, C_t$
1: $P \leftarrow$ INTERPOLATEINITIALPATH(ϵ)
2: **while** TIME AND ENERGYOK = FALSE **do**
3: $H \leftarrow$ SELECTHIGHDELTAENERGY(P)
4: $H_{\text{Min}} \leftarrow$ MINIIMIZE(H)
5: $P \leftarrow$ SEGMENTPATH(P,H,H_{MIN},ϵ)
6: **end while**

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 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 0.35 Å 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 h of CPU time. Parallelizing reduces the run time to just under 45 min 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 min. 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 0.65 Å 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 0.3 Å.

4. Results

4.1. Experimental setup

Here, we present results on the application of the proposed algorithm on the WT and a variant sequence of two enzymes, SOD1 and Ras. SOD1 is 150 amino acids long, whereas Ras is a 166 amino acids long.

Both enzymes are important due to their implication in severe human disorders. Mutations in SOD1 are implicated in familial ALS,[11] though the full function of this enzyme remains unknown. In particular, no simple characterization exists in terms of conformational switching between documented structural states, as experimental studies have shown great functional diversity. For

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press



Fig. 2. SOD1 crystal structures under PDB entries 1pu0 and 4b3e are superimposed over each other to highlight the subtle structural differences. The structure under PDB entry 1pu0 is colored red, whereas that under PDB entry 4b3e is colored blue. The Cu and Zn metals that bind to SOD1 are also drawn. Structures are rendered with the NewCartoon representation through the Visual Molecular Dynamics (VMD) software.[93]

instance, mutations not linked to ALS have been found to trigger aggregation, whereas the impact of other mutations on structure and energetics remains unknown.

In this paper, we take a first step and study differences between the WT and the US-dominant ALS-causing variant, A4V, of SOD1. In the A4V variant, the alanine (A) residue at position 4 is mutated and replaced with a valine (V). The WT sequence of SOD1 can be obtained from UniProt[90] by typing the keyword "superoxidase dismutase" or "SOD1" in the text box. The query with each of these keywords additionally returns all known variants of SOD1.

We elect to study here the ability of these two sequences, the WT and the A4V variant, to navigate opposite sides of the SOD1 energy landscape. An additional reason for selecting SOD1 as a system of study in this paper is that the EA in Refs. [9] and [10] is very effective at providing a dense sample-based representation of the SOD1 WT and A4V energy landscapes that are employed here.

A PCA-based analysis in Refs. [9, 10] of all known crystal structures populated by the SOD1 WT and variant sequences has demonstrated that known structures populate both sides of the energy landscape (corresponding to positive and negative values of projection coordinates over the first PC). For this reason, we study transitions between two crystal structures, each residing on the left and right side of the two-dimensional PCA-based embedding of the SOD1 energy landscape. The PDB ids of these structures are 1pu0[91] and 4b3e[92] (the WT sequence can also be directly obtained from 4b3e, and these structures are shown superimposed over each other in Fig. 2. The CA lRMSD between these structures is 0.70 Å (all-atom lRMSD is 1.13 Å). Figure 2 shows that the main-chain structural differences are not localized.

Unlike SOD1, Ras (we will refer to H-Ras as Ras from now on) is better studied in the wet and dry laboratories. This is partly due to the fact that Ras regulates cell proliferation, and some of its variants that deregulate activity are involved in over 25% of human cancers.[12,94] The native activity of Ras is to switch between an ON/active (GTP-bound, reactant) and an OFF/inactive (GDP-bound, product) state. These two states have been characterized in the wet laboratory and can be found under structures with PDB ids 1qra[95] and 4q21,[96] respectively. Both structures are reported for the WT, and the WT sequence for Ras can be obtained from either entry). We show these structures side by side in Fig. 3. The CA lRMSD between these structures is 1.47 Å (all-atom lRMSD is 2.26 Å), but changes are largely localized on two loop regions, switch I and switch II (which our previous analysis of PCA

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

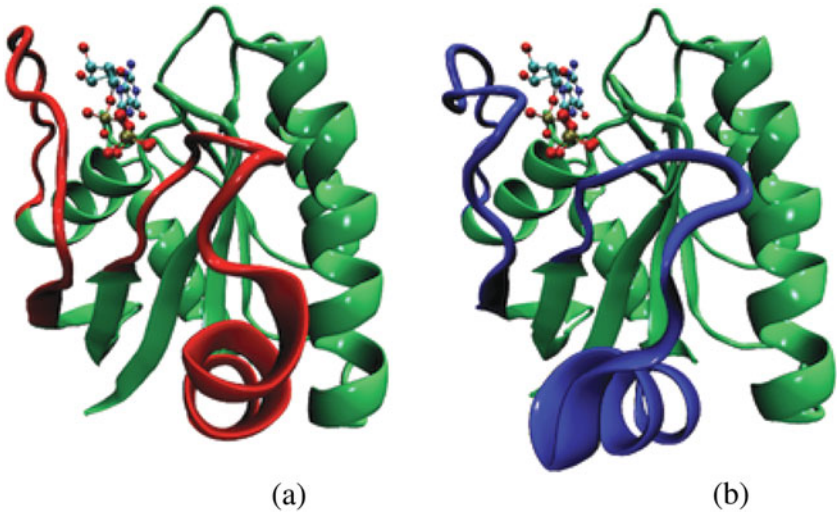


Fig. 3. (a) A representative of the ON (GTP-bound) state of Ras (PDB id: 1qra). (b) A representative of the OFF (GDP-bound) state (PDB id: 4q21). The reactant (GTP) and product (GDP) are shown, as well. (a), (b) The two loop regions that undergo a structural change in the ON to OFF transition and (reverse) are shown color-coded in red (left) and blue (right). The loop region closer to the drawn ligand is known as the switch I (SI) region, and that further from the ligand is known as the switch II (SII) region. Q61 is on SI and not directly involved in ligand binding; however, wet-laboratory studies suggest an allosteric role for this residue in Ras catalysis.[98] Structures are rendered with the NewCartoon representation through the VMD software.[93]

for Ras shows PCA is able to capture with a few PCs[83]). How variations in the Ras sequence affect its capacity for switching between states is the focus of much research and is the reason we apply our SRS-based method to study the impact of a well-characterized oncogenic mutation, Q61L.[94,97,98] In the Q16L variant, the glutamine (Q) residue at position 61 is mutated and replaced with a leucine (L). Mutations of residue Q61 in Ras are shown to impair activity and are prominent in human cancers; however, while this residue is known to contribute to catalysis,[98] the role of its mutations on the relationship between structure and activity is not well understood.

The reduced space over which the sampling stage operates is obtained via PCA on WT and variant structures extracted for Ras and SOD1 from the PDB (details on the data collection step can be found in Refs. [9] and [10]). The SRS-based algorithm is run twice, once on the WT sequence and once on the disease-participating variant of the protein under investigation. It is important to note that, while the PCs are the same in each setting (for WT or variant sequence of a protein under investigation), the EA algorithm obtains different structural ensembles, as the initial structures are threaded onto the sequence of study, and thus mapped by the reconstruction procedure to minima of different sequence-dependent energy surfaces. Thus, the results of the SRS-based algorithm are dependent on the sequence used and can be used to draw comparisons between the WT and variants to determine how sequence mutations affect transitioning between the ON and OFF states.

The collection, threading, and reconstruction procedures are described in detail in the context of the EA proposed in Refs. [9] and [10] that is used here to obtain a structural ensemble for a protein sequence of interest. For the sake of completeness, we briefly summarize these procedures here.

4.1.1. Collection. Experimentally available structures in the PDB are collected by querying with the name of the protein under investigation. Structures with missing internal regions and amino-acid sequences differing from the WT sequence by more than three amino acids are discarded. This results in 86 structures for Ras and 186 for SOD1.

4.1.2. Threading and reconstruction. Since these structures may be reported for different variants, they are first threaded onto the sequence of interest. Cartesian coordinates of common atoms are retained; differences are localized to no more than three side chains. Coordinates of side-chain atoms that are different in the sequence of interest are rebuilt via the Rosetta *relax* protocol. This protocol essentially allows bridging scales, as it can pack side chains and via a simulated annealing procedure

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

map a threaded structure onto a nearby local minimum in the all-atom Rosetta *score12* surface of the protein sequence of interest.

The EA uses these reconstructed minima structures of a protein sequence of interest to seed the initial population. More structures are generated during the execution of the EA, from the rest of the generations. As further described in Section 3, the structures obtained from all generations of the EA are collected in the Ω ensemble and clustered by the proposed SRS-based method to identify structural states. Given the objective to identify transition paths and collect statistics over paths connecting two structural states of interest, the first task is to identify these states from representative structures.

Specifically, for Ras, the structure in the PDB entry 1qra is considered a representative of the ON state, whereas 4q21 is a representative of the OFF state (such information is available in the description of the structures in these entries). The structures under these PDB entries have been caught in the wet laboratory for the WT, but other variants of Ras also populate the ON and OFF structural states. These two PDB-obtained structures are each minimized 500 times with the Rosetta *relax* protocol (the protocol is stochastic), and the resulting structures are added to the Ω ensemble. This is done for two reasons; first, to map these structures to local minima in the Rosetta all-atom (*score12*) energy surface and in the process also resolve any unfavorable interactions and suboptimally placed hydrogen or other side-chain atoms that may exist in the PDB entries; second, to seed the roadmap with structures comprising the states of interest (ON and OFF for Ras). These states are identified by clustering the structures resulting from the independent applications of the *relax* protocol on the PDB-obtained structures.

So, for Ras, the cluster containing the most minimized structures resulting from 500 independent applications of the *relax* protocol on the structure under PDB entry 1qra is labeled the ON state. The cluster containing the most minimized structures resulting from 500 independent applications of the *relax* protocol on the structure under PDB entry 4q21 is labeled the OFF state. A similar process is followed to define the states of interest for SOD1 from the structures found under PDB entries 1pu0 and 4b3e.

4.2. Analysis of application on SOD1

We first provide details onto the selection of parameter values and then analyze the transitions between the states of interest in the WT and ALS variant studied here for SOD1.

4.2.1. Justification of parameter values for application of method on SOD1. Figure 4 shows the number of clusters as a function of increasing cluster radii, from 0.25 to 0.40 Å. There is no significant reduction in the number of clusters. This indicates that the conformations (corresponding to local minima) sampled for SOD1 by the EA-based algorithm employed by our method in the sampling stage are spread out in conformational space. We proceed with the same cluster radius of 0.35 Å used for Ras to favor structural homogeneity of a state over a decrease in the number of states.

We refer to the cluster/state containing most of the minimized (*relaxed*) all-atom conformations of structure under PDB id 1pu0 as state A, and the state containing most of the minimized all-atom conformations of structure under PDB id 4b3e as state B. The CA lRMSD between the leaders (structure representatives) of the A and B states in WT SOD1 is 0.77 Å (all-atom lRMSD is 1.32 Å). In Fig. 5, we show the distribution of distances of all clusters from states A and B in the SOD1 WT. The distribution for the A4V variant is similar (data not shown). Inter-cluster distances are measured through the average lRMSD between the members of each cluster.

The distribution of edge lengths in the roadmap constructed for the SOD1 WT is shown in Fig. 6 (results are similar for the A4V variant, data not shown). Since an edge connects two clusters in the roadmap, the length of an edge is the distance between the two clusters it connects, measured as described above.

4.2.2. Roadmap analysis on SOD1 WT and A4V variant. We apply the analysis techniques discussed in Section 3.4 to the roadmaps created for the WT and A4V SOD1 sequences. The minimum-cost paths between the A and B states for each sequence are computed and analyzed. Column 3 in Table I shows the total cost of each of these paths. Comparison of these values shows that the B → A structural transition is slightly more costly than the A → B one for both the WT and A4V. However, both transitions have slightly higher costs in the A4V variant.

The detailed energetic profiles of the minimum-cost paths for the A → B and B → A transitions are shown for the WT and A4V variant in Fig. 7. The Rosetta all-atom (score12) energy is shown for

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

Table I. The cost of the minimum-cost paths are shown for the structural transitions between the A and B states in both the WT and A4V variant of SOD1.

Sequence	Transition	Minimum cost
WT	A → B	5.81
	B → A	6.78
A4V	A → B	6.38
	B → A	7.49

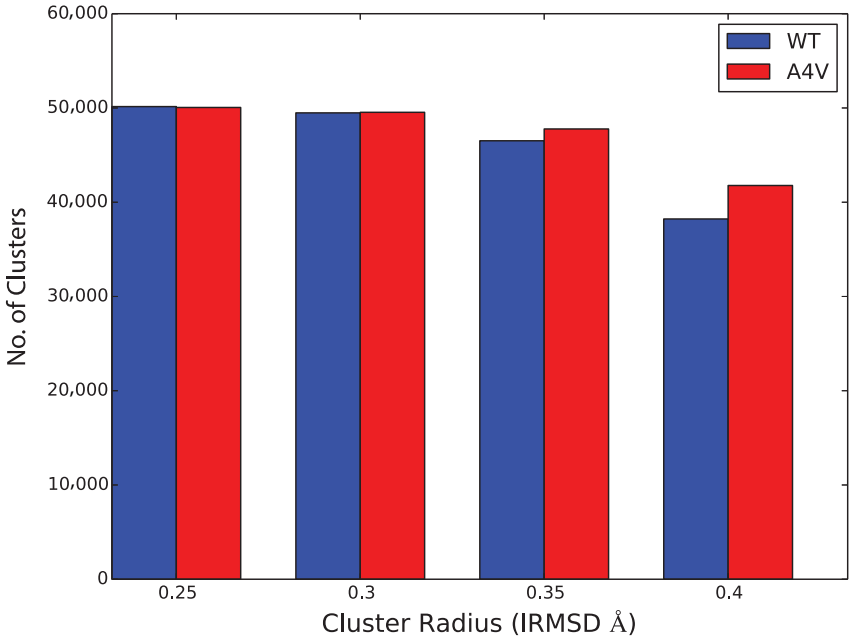


Fig. 4. The number of clusters resulting from different values selected for the cluster radius is shown for both the SOD1 WT and A4V variant.

each vertex in these paths, but the path lengths are normalized to allow a direct comparison between paths of different lengths. Figure 7 shows that the A4V mutation affects the energy barrier that SOD1 has to cross both in the A → B and the B → A transitions.

The results shown in Fig. 7 suggest that the A4V mutation causes the transition to become slightly more costly but does not affect the stability of the A and B states. However, the rise is no more than 10 Rosetta Energy Units, which is comparable to the in-basin variations observed upon relaxation of native structures. At present, there are no wet-laboratory or computational studies on the transitions of the SOD1 WT and A4V variant, but wet-laboratory studies have shown that the A4V variants retains full enzymatic activity as the WT.[99] This in agreement with the slight effect that the mutation has on the transition barrier. Wet-laboratory studies have also reported that the A4V variant is prone to aggregation.[100] While this may be due to slight changes in the transition barrier, the results here cannot be employed to obtain a detailed mechanistic insight into the process of aggregation, as they are obtained over the computed landscape of monomeric SOD1. While the results presented here do not suggest significant changes between the WT and A4V variant, it is possible that entropic contributions may stabilize aggregated species in the variant. Further investigation of the WT and other variants of SOD1 in both wet and dry laboratories are needed to obtain a more comprehensive view of SOD1 misfunction.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

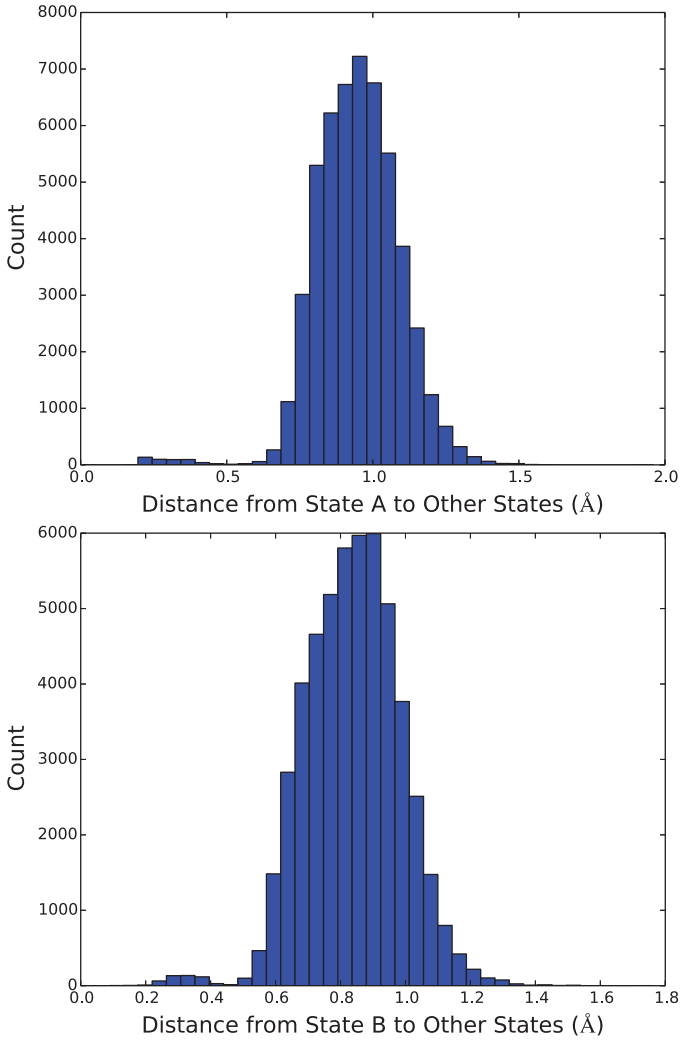


Fig. 5. The histogram shows the distribution of distances of all clusters/states from the cluster/state designated as the A and B states for the SOD1 WT sequence. Distance between two clusters C_i and C_j is measured as the average lRMSD between members of C_i and members of C_j.

4.3. Analysis of application on Ras

As for SOD1, we first justify the selection of parameter values. We then provide a detailed analysis of the impact of the Q61L mutation on the energy landscape and transitions between the ON and OFF states in Ras.

4.3.1. Justification of parameter values for application of method on Ras. Figure 8 shows the reduction in the number of clusters as a function of increasing the cluster radius. A linear dropoff is observed. Specifically, the number of clusters goes down from 46, 193 to 37, 250, 26, 461, and then 17, 547 when the cluster radius varies in {0.25, 0.3, 0.35, 0.4} Å for Ras. The selected radius of 0.35 Å provides good reduction in the number of clusters/states while allowing a state to be structurally homogeneous.

The identification of the ON and OFF states among the clusters is described in Section 3. The CA lRMSD between the leaders (structure representatives) of the ON and OFF states in WT Ras is 1.40 Å(all-atom lRMSD is 2.38 Å). In Fig. 9, we show the distribution of distances of all clusters from the designated ON and OFF clusters/states for the WT. The distribution for the variant is similar (data not shown). Inter-cluster distances are measured through the average lRMSD between the members of each cluster.

The distribution of edge lengths in the roadmap constructed for the Ras WT is shown in Fig. 10 (results are similar for the Q61L variant, data not shown). Since an edge connects two clusters in

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

Table II. The minimum-cost paths and the expected number of edges are shown for the structural transitions between the ON and OFF states in both the WT and Q61L variants in Ras.

Sequence	Transition	Minimum cost	Expected No. of edges
WT	OFF → ON	12.9	3.4×10^{8}
	ON → OFF	16.5	3.9×10^{10}
Q61L	OFF → ON	20.9	1.9×10^{12}
	ON → OFF	24.3	3.8×10^{14}

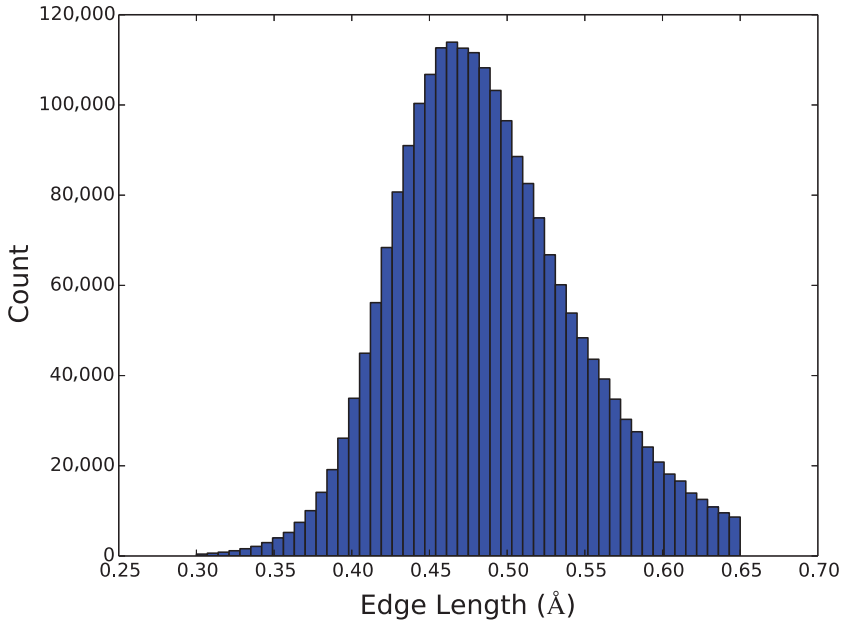


Fig. 6. The histogram shows the distribution of edge lengths in the roadmap constructed for the SOD1 WT. The length of an edge is the distance between the two clusters it connects.

the roadmap, the length of an edge is the distance between the two clusters it connects, measured as described above.

4.3.2. Roadmap analysis on Ras WT and Q61L variant. We apply the analysis techniques described in Section 3.4 to the roadmaps created for the WT and Q61L Ras sequences. The minimum-cost paths between the ON and OFF states for each sequence are computed, analyzed, and the total cost of each of these paths are shown in column 3 in Table II. Total cost of a path is computed as $\sum_{e=(u,v)} -\log(\exp(-\frac{(E(v)-E(u)}{\alpha})$, where u and v are the two states connected by a directed edge, and α is the effective temperature parameter calculated as described in Section 3. Comparison of these values shows that the ON → OFF structural transition is more costly than the OFF → ON one for both the WT and Q61L (the minimum costs are higher by 3.4–3.6 units in each case). However, both transitions have higher cost in the Q61L variant, indicating a significant change of the energy landscape upon this mutation.

We now visualize the minimum-cost paths for each of these transitions in the WT and Q61L variant by drawing them over the Rosetta energy landscape of each sequence. The landscape is obtained by projecting the energy surface over the top two PCs for ease of visualization. The two-dimensional PC1-PC2 map is discretized, and grid cells are color-coded according to the average Rosetta score12 energy over conformations mapped to that cell. The top panel of Fig. 11 shows the minimum-cost ON → OFF and OFF → ON paths over the Rosetta landscape for the WT. The bottom panel of Fig. 11 does so for the Q61L variant.

The wider and deeper basin on the right of the Ras WT Rosetta landscape corresponds to the ON state, whereas the narrower basin on the left corresponds to the OFF state. Comparison of the

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

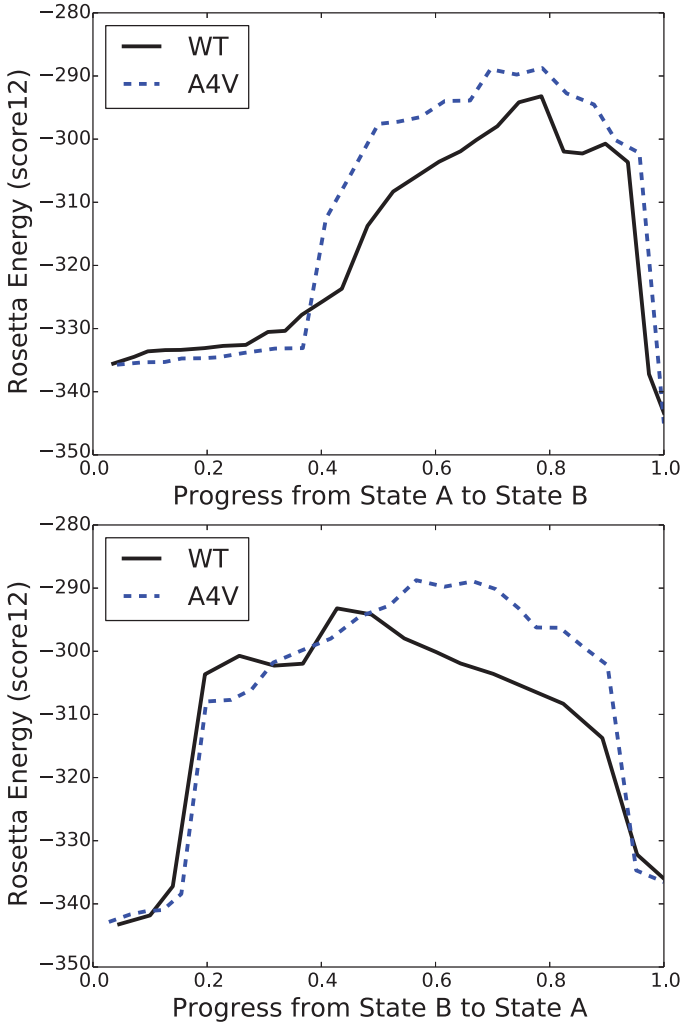


Fig. 7. The energetic profiles of the minimum-cost paths (top panel: A to B and bottom panel B to A) are shown for the WT and A4V variant in SOD1.

minimum-cost ON → OFF and OFF → ON paths over the WT landscape shown in the top panel of Fig. 11 reveals that the paths are largely different in how they exit and enter the two states. The higher-energy region in the middle of the landscape constrains both paths. The higher energetic cost of the ON → OFF path reported in column 3 for the WT in Table II is largely a result of the difficulty of the path to exit the deeper and wider ON basin. Differences between the WT and Q61L variant are made clear by the landscapes. The bottom panel of Fig. 11 shows that the Rosetta landscape for the Q61L variant is markedly different from that of the WT. The basin corresponding to the ON state is shallower, and that corresponding to the OFF state is narrower. The landscape is now much more energetically unfavorable, with the middle region presenting a higher energy barrier over that seen for the WT. This is the reason that the energetic costs of the minimum-cost paths shown for the Q61L variant in column 3 in Table II are higher than those of the minimum-cost paths obtained for the WT. Again, within Q61L, the differences between the minimum-cost ON → OFF and OFF → ON states can be largely explained by the higher difficulty to get out of as opposed to get into the ON state, as the rest of the landscape similarly constrains both paths.

To provide insight into the generality of the conclusions that can be reached on Ras, the analysis is repeated by reconstructing the roadmap over energies obtained with a physics-based force field. Each conformation is rescored using the all-atom ff12SB Amber force field,[101] and edge weights are recalculated. The analysis is repeated, and the minimum-cost paths now obtained over the new

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

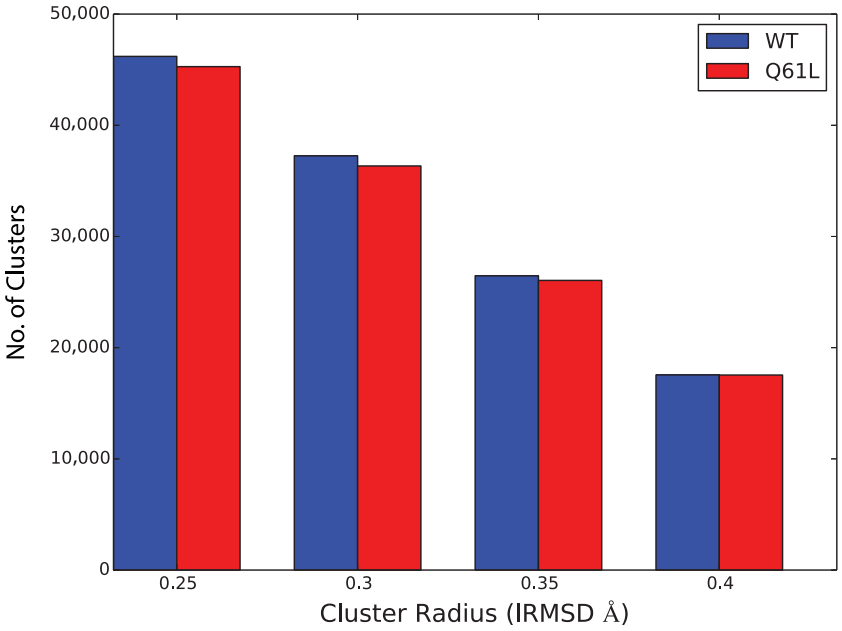


Fig. 8. The number of clusters resulting from different values selected for the cluster radius is shown for both the Ras WT and Q61L variant.

roadmaps for the WT and Q61L variant are shown over the AMBER landscapes for these two sequences in Fig. 12.

Largely, the same observations made over the Rosetta landscapes apply over the Amber landscapes. On the WT sequence, the paths have to go over a similar energy barrier. On the Q61L variant, the AMBER landscape has a lower barrier than the Rosetta landscape between the two states, but the OFF basin has also become much shallower. The same conclusions can be reached, that the minimum-cost paths between the ON and OFF states in the Q61L variant are more costly than the minimum-cost paths in the WT.

To facilitate direct comparison between the WT and Q61L variant, the detailed energetic profiles of the minimum-cost paths for the ON $\rightarrow$ OFF and OFF $\rightarrow$ ON transitions (when using Rosetta score12) are shown for the WT and Q61L variant in Fig. 13. The Rosetta all-atom energy is shown for each vertex in these paths, but the path lengths are normalized to allow a direct comparison between the two Ras sequences.

Figure 13 shows that the Q61L mutation magnifies the energy barrier that Ras has to cross both in the ON $\rightarrow$ OFF and the OFF $\rightarrow$ ON structural transitions. The Q61L mutation does not affect the stability of the ON and OFF structural states, since the potential energies of the corresponding states remain largely the same between the WT and variant.

Figure 14 illustrates the effect of the CPR-based path smoothing algorithm. The algorithm is applied to the minimum-cost paths obtained from the Ras WT (top panel) and Q61L (bottom panel) roadmaps for the ON $\rightarrow$ OFF transition. The smoothed paths are superimposed over the original ones and drawn over the PC1–PC2 embedding of the sampled conformational space for each Ras sequence. The superimposition shows that the path smoothing does not significantly alter the original paths in the conformational space, at least in the PC1–PC2 embedding that captures more than 50% of the structural variance in Ras (similar observations are made for the Q61L variant—data not shown).

The small differences are largely a result of the constraints imposed on our path smoothing algorithm to seek a nearby path of lower energetic cost. As described in Section 3, this is obtained by greatly restricting backbone movements during the energetic minimizations in the Rosetta *relax* protocol (the option of minimizing backbone movements is employed in the protocol).

Figure 15 illustrates the effect of the CPR-based path smoothing algorithm on the energetic profile of the minimum-cost paths. As can be observed, the algorithm discovers that there are specific edges in the original paths that go over high-energy hills where constrained minimization does not significantly steer away towards lower-energy valleys. While an argument can be made that in some cases that

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

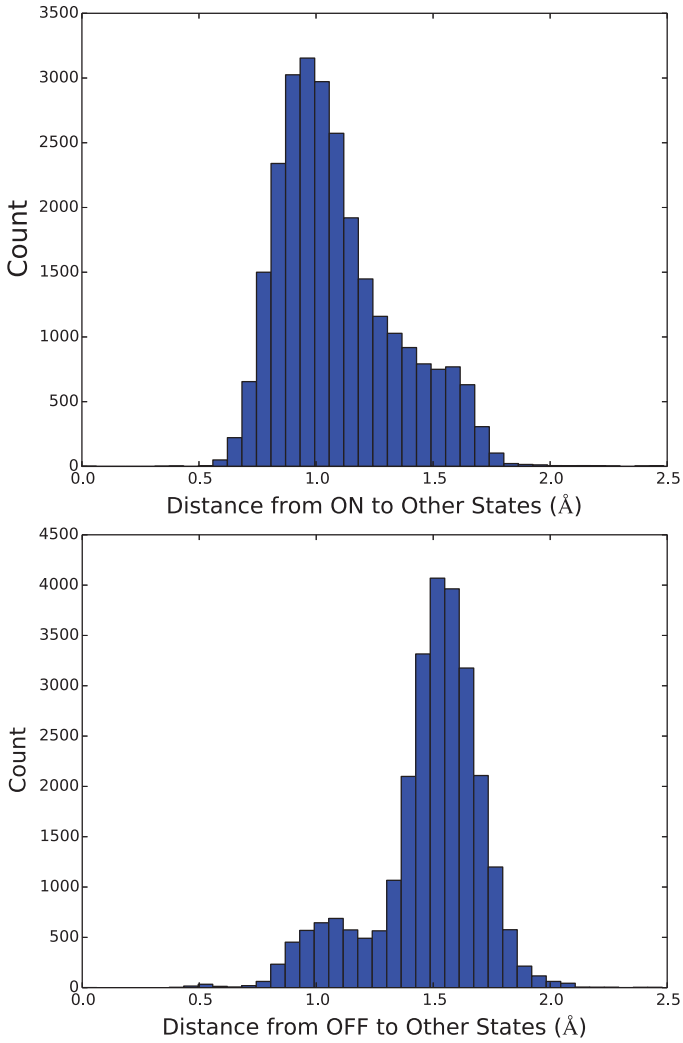


Fig. 9. The histogram shows the distribution of distances of all clusters/states from the cluster/state designated as the ON and OFF states for the Ras WT sequence. Distance between a cluster C_i and cluster C_j is measured as the average lRMSD between members of C_i and members of C_j.

allowing indiscriminate backbone movements may be beneficial, the goal here is not to rely on the path smoothing algorithm to discover largely different paths but instead to seek local improvements of paths by searching in a restricted neighborhood of the conformational space.

The first-step analysis described in Section 3 is applied to measure and compare the expected number of edges in the transitions between the ON and OFF states. These results are related in column 4 in Table II. Comparison of these results for the WT sequence shows that the expected number of edges to allow switching from the ON to the OFF state is two orders of magnitude higher than that from the OFF to the ON state. This also holds for the Q61L variant, though switching from ON to OFF and vice versa becomes more difficult in the variant than in the WT.

4.3.3. Comparison with published literature. Taken altogether, these results suggest that a careful realization of the SRS framework may allow a more detailed understanding of the role of sequence mutations in misfunction. In particular, the application of our method on the WT and Q61L variant of Ras, yields results that support the hypothesis that the Q61L mutation does not remove the ON and OFF basins from the Ras energy landscape but instead slows down the switching of Ras between these states by raising the energy barrier of the transition.

In particular, the results shown in Fig. 13 are in qualitative agreement with other published studies; direct, quantitative comparisons with transition rates cannot be made at this stage, as the work here is the first to provide detailed energy landscapes of H-Ras WT and the Q61L variant. However,

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

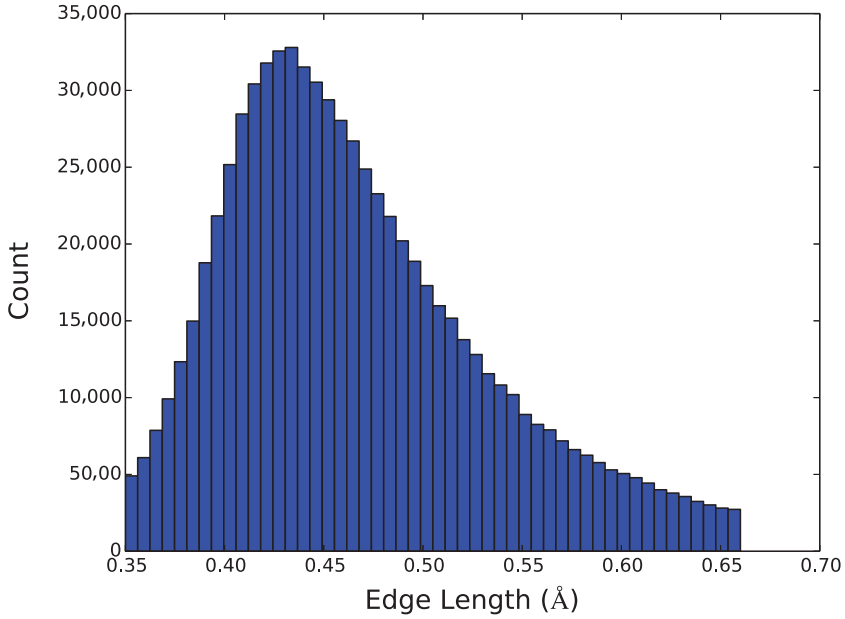


Fig. 10. The histogram shows the distribution of edge lengths in the roadmap constructed for the Ras WT. The length of an edge is the distance between the two clusters it connects.

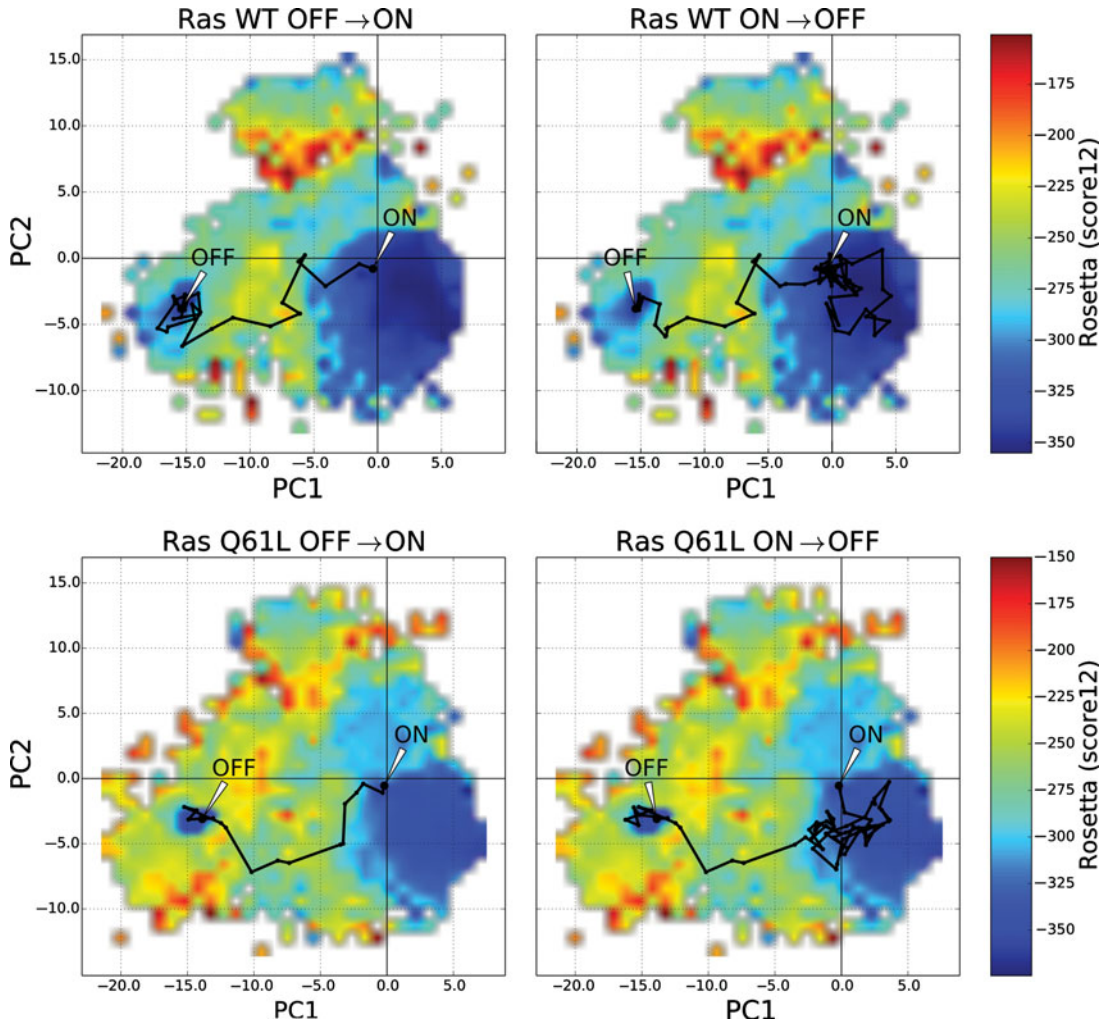


Fig. 11. The minimum-cost paths are shown for the WT (top panel) and Q61L variant (bottom panel) between the ON and OF states over the Rosetta landscapes.

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

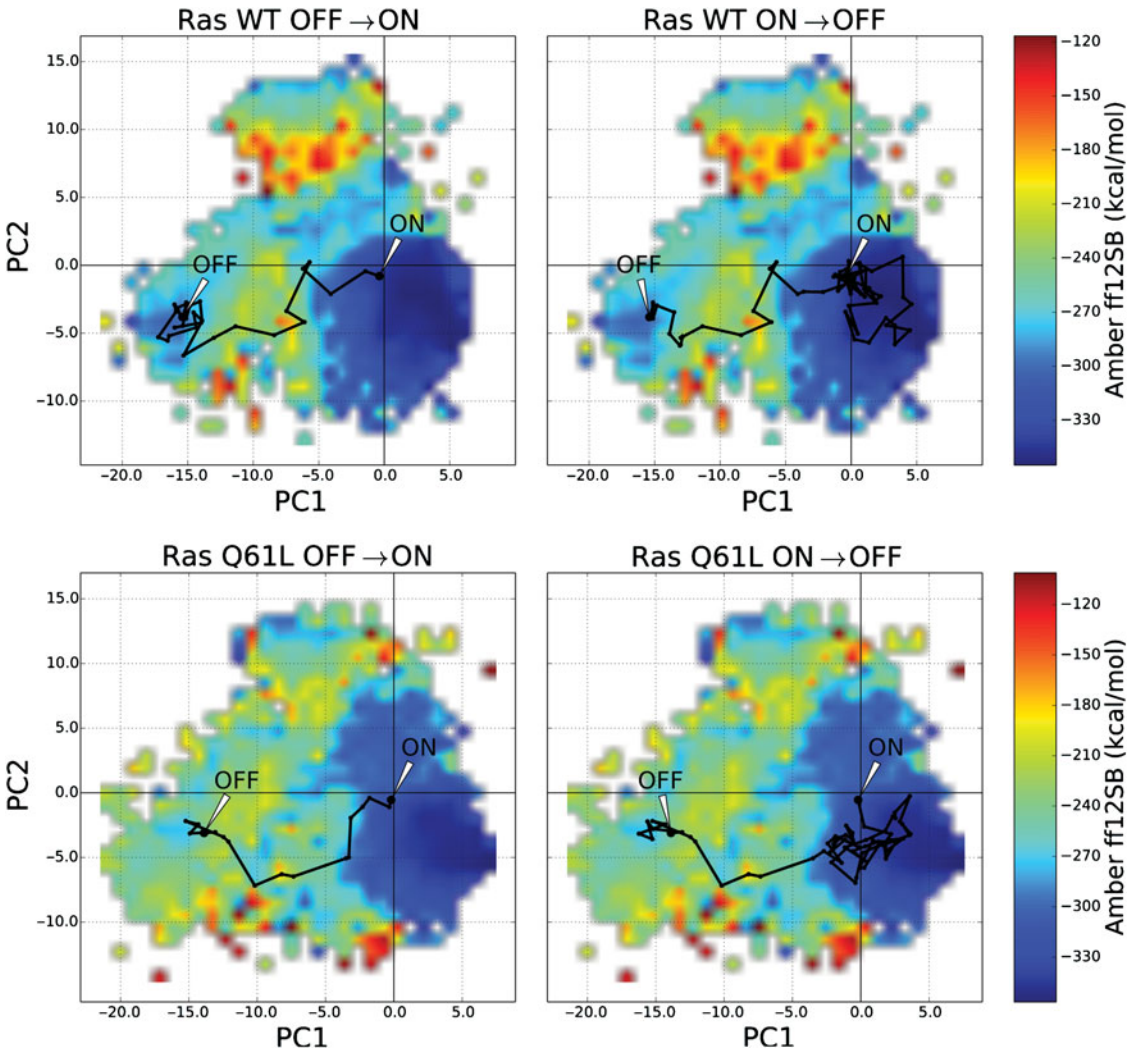


Fig. 12. The minimum-cost paths are shown for the WT (top panel) and the Q61L variant (bottom panel) between the OFF and ON states over the Amber landscapes.

wet-laboratory investigations of the detailed catalysis of the Q61L Ras variant has shown that, while the variant has binding affinity similar to that of the WT,[102] the variant cannot form the transition state mimic[103] and, as a result, its hydrolysis rate is not enhanced by the interaction.[104] These wet-laboratory investigations suggest that the Q61L variant has largely well-defined and possibly similarly stable ON and OFF states to that of the WT; however, the inability to form the transition state mimic suggests higher energetic barriers in the transition between these states than in the WT. The energy profiles in Fig. 13 and the higher costs and expected number of edges in Table II support this hypothesis on the effect of the Q61L in raising the energetic barrier in the transition.

Figure 13 suggests that the energy barrier in the WT is between 50 and 100 units higher than the average energy of the ON and OFF states. A direct comparison with published experimental rates, which only exist for the WT is difficult for two reasons. First, the rate of the ON→OFF transition is not resolved kinetically.[105] Second, considerable entropic contributions are expected to influence this rate. Wet-laboratory studies provide a lower bound for the transition rate on the order of 104 s1,[105] which means that the free energy barrier along the transition cannot exceed 23 kcal/mol. Scaling down the 50–100 units here to the observed 6–7 units in-basin variation corresponding to the expected 0.6 kcal/mol thermal vibrations within a basin suggests that the energy barrier obtained here cannot exceed 10 kcal/mol. The barriers estimated from biased MD simulations[35,106] vary from 60 kcal/mol[35] to 500 kcal/mol,[106] which are considerably higher than what is found here. It is questionable, however, whether biased MD simulations provide information on the true transition pathway.[7] In contrast,

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

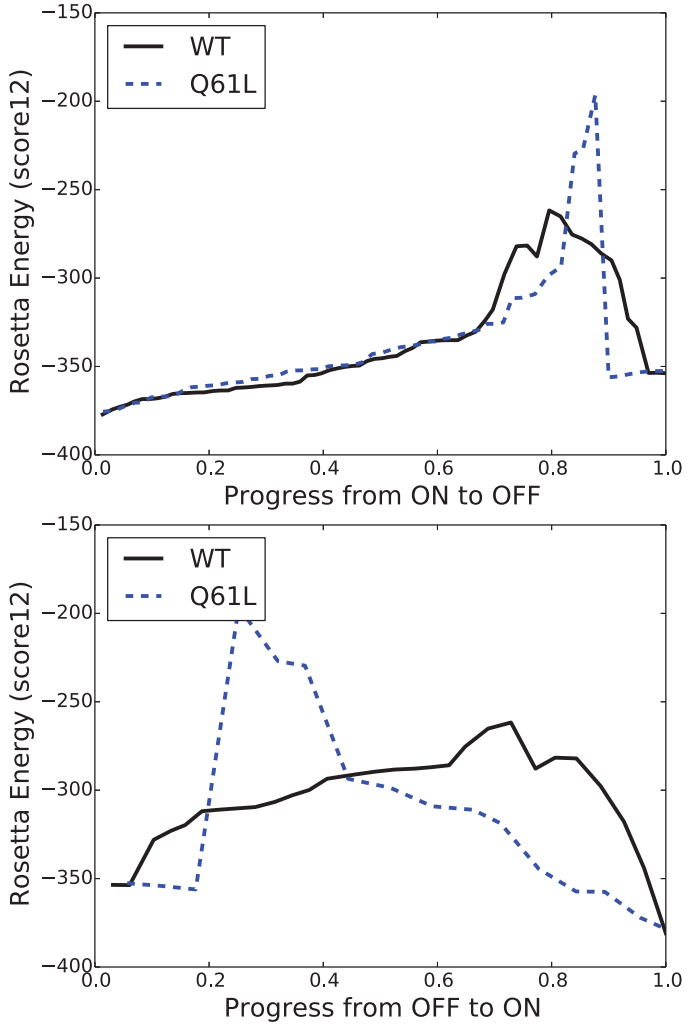


Fig. 13. The energetic profiles of the minimum-cost paths (top panel: ON to OFF and bottom panel OFF to ON) are shown for the WT and Q61L variant.

the barrier estimated from application of a chain-of-states approach[107] is much lower, about 24.3 kcal/mol. However, this rate was found highly sensitive to the solvation model and the non-bonded cutoff employed in energetic evaluations. In light of these, comparisons between WT and variant, as provided in this paper, are more informative than absolute values. The analysis provided in this paper may suggest further wet-laboratory investigations in this direction to obtain rates for Ras variants to facilitate such comparisons.

5. 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 EA 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

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

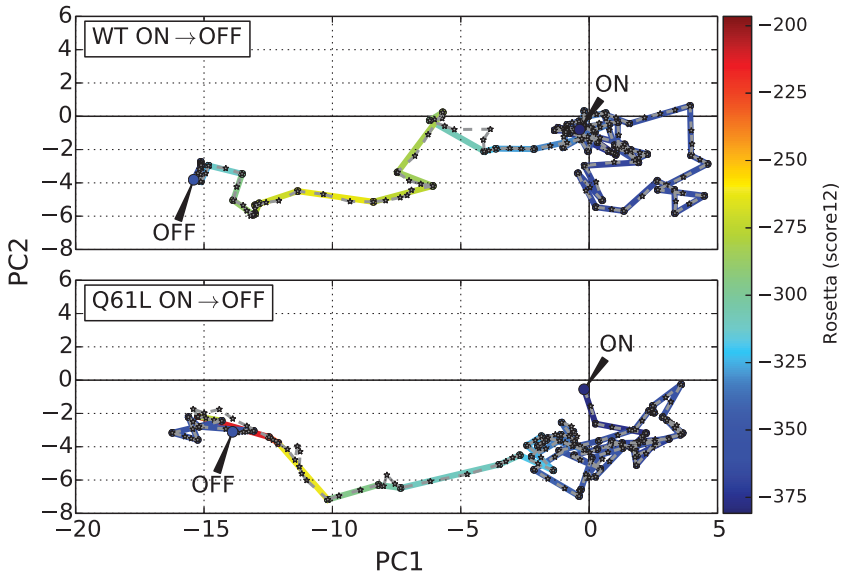


Fig. 14. The minimum-cost paths for the ON → OFF transition obtained for the RAS WT (top panel) and Q61L variant (bottom panel) are shown over the PC1–PC2 embedding, with edges color-coded according to energies. The paths shown dashed and in light gray are the result of applying the CPR-based path smoothing algorithm to the original minimum-cost paths. Structures in the roadmap that participate in the minimum-cost paths are shown as black dots. New conformations computed by the CPR-based path smoothing algorithm are shown as black stars.

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.[108] 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.[109] Elucidating transient structures may also 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 versus 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

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

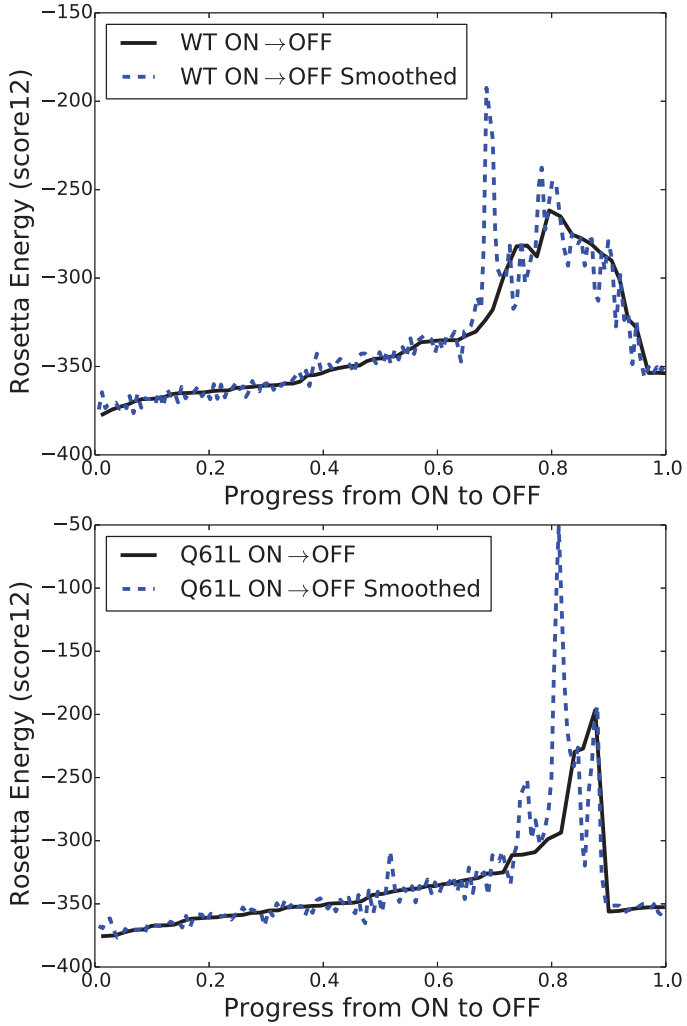


Fig. 15. The energetic profile of the minimum-cost path is compared to that of the smoothed path for the ON → transition in the Ras WT (top panel) and Q61L (bottom panel) roadmap. The original path is drawn in black. When the path smoothing algorithm is applied, the resulting path is drawn in dashed blue.

direction. A preliminary investigation of such approaches is available in Ref. [110], 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).

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

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 data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes," *J. Comp. Biol.* **22**(9), 844–860 (2015).
10. R. Clausen and A. Shehu, "A Multiscale Hybrid Evolutionary Algorithm to Obtain Sample-Based Representations of Multi-Basin Protein Energy Landscapes," *ACM Conference on Bioinformatics, Computational Biology (BCB)*, Newport Beach, CA (Sep. 2014) pp. 269–278.
11. 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).
12. A. E. Karnoub and R. A. Weinberg, "Ras oncogenes: Split personalities," *Nature Rev. Mol. Cell Biol.* **9**, 517–531 (2008).
13. 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," *Chem. Phys. Lett.* **194**(3), 252–261 (1992).
14. 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).
15. A. P. Singh, J.-C. Latombe and D. L. Brutlag, "A Motion Planning Approach to Flexible Ligand Binding," **In**: *Proceedigs of the International Conference on Intelligent Systems for Molecular Biology (ISMB)* vol. 7, (R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes and R. Zimmer, eds.) (AAAI, Heidelberg, Germany, 1999) pp. 252–261.
16. 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, Jan. 2000.
17. 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).
18. 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).
19. S. Thomas, G. Song and N. M. Amato, "Protein folding by motion planning," *J. Phys. Biol.* **2**(4), 148 (2005).
20. S. Thomas, X. Tang, L. Tapia and N. M. Amato, "Simulating protein motions with rigidity analysis," *J. Comput. Biol.* **14**(6), 839–855 (2007).
21. L. Tapia, X. Tang, S. Thomas and N. Amato, "Kinetics analysis methods for approximate folding landscapes," *Bioinformatics* **23**, i539–i548 (2007).
22. 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).
23. L. Tapia, S. Thomas and N. Amato, "A motion planning approach to studying molecular motions," *Commun. Inform. Syst.* **10**(1), 53–68 (2010).
24. M. Moll, D. Schwartz and L. E. Kavraki, "Roadmap methods for protein folding," *Methods Mol. Biol.* **413**, 219–239 (2008).
25. K. Molloy and A. Shehu, "A Probabilistic Roadmap-Based Method to Model Conformational Switching of a Protein Among Many Functionally-Relevant Structures," *International Conference on Bioinformatics & Computational Biology (BICoB)*, Las Vegas, NV (2014) pp. 1–6.
26. K. Molloy and A. Shehu, "Interleaving Global and Local Search for Protein Motion Computation. **In**: *LNCS: Bioinformatics Research and Applications,* vol. 9096, (R. Harrison, Y. Li and I. Mandoiu, eds.) (Springer International Publishing, Norfolk, VA, 2015) pp. 175–186.
27. 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).
28. T. H. Chiang, D. Hsu and Latombe. J. C, "Markov dynamic models for long-timescale protein motion," *Bioinformatics* **26**(12), 269–277 (2010).
29. 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).
30. 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).
31. 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).

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

32. 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).
33. R. D. Malmstrom, C. T. Lee, A. T. Van Wart and R. E. Amaro, "Application of molecular-dynamics based markov state models to functional proteins," *J. Chem. Theory Comput.* **10**(7), 2648–2657 (2014).
34. M. Held and F. Noé, "Calculating kinetics and pathways of protein-ligand association," *Eur. J. Cell Biol.* **91**(4), 357–364 (2012).
35. 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).
36. O. Beckstein, E. J. Denning, J. R. Perilla and T. 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).
37. 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).
38. 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).
39. 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).
40. 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).
41. J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," *PLoS Comp. Biol.* **5**(8), e1000480 (2009).
42. Q. Ciu and I. Bahar, *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*, 1st ed. (CRC Press, 2005).
43. R. Bahar and A. J. Rader, "Coarse-grained normal mode analysis in structural biology," *Curr. Opin. Struct. Biol.* **204**(5), 1–7 (2005).
44. 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).
45. 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).
46. N. Echols, D. Milburn and M. Gerstein, "Molmovdb: Analysis and visualization of conformational change and structural flexibility," *Nucleic Acids Res.* **31**(1), 478–482 (2003).
47. D. R. Weiss and P. Koehl, "Morphing methods to visualize coarse-grained protein dynamics," **In**: *Protein Dynamics*, vol. 1084, (Dennis R. Livesayed) (Humana Press, 2014) pp. 271–282.
48. 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).
49. 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).
50. 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).
51. 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).
52. W. Zheng, B. R. Brooks, G. Hummer, Protein conformational transitions explored by mixed elastic network models. *Proteins: Struct. Funct. Bioinf.* **69**(1), 43–57 (2007).
53. F. Zhu and G. Hummber, "Gating transition of pentameric ligand-gated ion channels," *Biophys. J.* **97**(9), 2456–2463 (2009).
54. 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).
55. 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).
56. 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).
57. 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).
58. 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).
59. 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).
60. E. Weinan, W. Ren and E. Vanden-Eijnden, "String method for the study of rare events," *Phys. Rev. B* **66**, 052301 (2002).
61. 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).
62. M. U. Bohner, J. Zeman, J. Smiatek, A. Arnold and J. Kästner, "Nudged-elastic band used to find reaction coordinates based on the free energy," *J. Chem. Phys.* **140**(7), 074109 (2014).

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

63. R. Olender and R. Elber, "Yet another look at the steepest descent path," *J. Mol. Struct. THEOCHEM* **398–399**, 63–71 (1997).
64. 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).
65. L. Maragliano and E. Vanden-Eijnden, "On-the-fly string method for minimum free energy paths calculation," *Chem. Phys. Lett.* **446**(1), 182–190 (2007).
66. W. Ren and E. Vanden-Eijnden, "Finite temperature string method for the study of rare events," *J. Phys. Chem. B* **109**(14), 6688–6693 (2005).
67. W. Ren, E. Vanden-Eijnden, P. 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).
68. 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).
69. 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).
70. 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).
71. 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).
72. 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).
73. W. Gan, S. Yang and B. 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).
74. B. Olson and A. Shehu, "Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface," *ACM Conference on Bioinformatics, Computational Biology (BCB)*, Washington, D. C. (Sep. 2013) pp. 430–439.
75. B. Olson and A. Shehu, "Efficient Basin Hopping in the Protein Energy Surface," *IEEE International Conference on Bioinformatics and Biomedical*, Philadelphia, PA (Oct. 2012) pp. 119–124.
76. B. Olson, K. A. De Jong and A. Shehu, "Off-Lattice Protein Structure Prediction with Homologous Crossover," *Conference on Genetic and Evolutionary Computation (GECCO)*, ACM, New York, NY (2013) pp. 287–294.
77. I. Hashmi and A. Shehu, "idDock+: Integrating machine learning in probabilistic search for protein-protein docking," *J. Comp. Biol.* **22**(9), 1–18 (2015).
78. I. Hashmi and A. Shehu, "Informatics-Driven Protein-Protein Docking," *ACM Conference on Bioinformatics, Computational Biology Workshops (BCBW)*, Washington, D. C. (Sep. 2013) pp. 772–779.
79. B. Olson, I. Hashmi, K. Molloy and A. Shehu, "Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules," *Adv. AI J.* **2012**(674832) (2012).
80. H. M. Berman, K. Henrick and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.* **10**(12), 980–980 (2003).
81. R. Clausen, B. Ma, R. Nussinov and A. Shehu, "Mapping the conformation space of wildtype and mutant h-ras with a memetic, cellular, and multiscale evolutionary algorithm," *PLoS Comput. Biol.* **11**(9), e1004470 (2015).
82. K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan and J. Meiler, "Practically useful: What the rosetta protein modeling suite can do for you," *Biochemistry* **49**(14), 2987–2998 (2010).
83. R. Clausen and A. Shehu, "Exploring the Structure Space of Wildtype Ras Guided by Experimental Data. *ACM Conference on Bioinformatics, Computational Biology Workshops (BCBW)*, Washington, D. C. (Sep. 2013) pp. 757–764.
84. J. A. Hartigan, *Clustering Algorithms* (John Wiley and Sons, New York 1975).
85. A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallogr. A.* **26**(6), 656–657 (1972).
86. 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).
87. 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.* **13**(Suppl 1), S8 (2013).
88. R. Bohlin and L. E. Kavraki, "Path Planning using Lazy PRM," *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 1 (San Fransisco, CA, IEEE Press, Apr. 2000) pp. 521–528.
89. M. Pascoal and E. Martins, "A New Implementation of Yen's Ranking Loopless Algorith," *Quart. J. Belg. French Ital. Oper. Res. Soc.* **1**(2), 121–133 (2003).
90. M. Magrane and the UniProt consortium, "UniProt knowledgebase: A hub of integrated protein data," *Database* **2011**(bar009), 1–13 (2011).
91. M. DiDonato, L. Craig, M. E. Huff, M. M. Thayer, R. M. Cardoso, C. J. Kassmann, T. P. Lo, C. K. Bruns, E. T. Powers, J. W. Kelly, E. D. Getzoff and J. A. Tainer, "ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization," *J. Mol. Biol.* **332**(3), 601–615 (2003).

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press

92. R. W. Strange, M. A. Hough, S. V. Antonyuk and S. S. Hasnain, "Structural evidence for a copper-bound carbonate intermediate in the peroxidase and dismutase activities of superoxide dismutase," *PLoS One* **7**(9), e44811 (2012).
93. W. Humphrey, A. Dalke and K. Schulten, "VMD - Visual Molecular Dynamics," *J. Mol. Graph. Model.* **14**(1), 33–38 (1996). http://www.ks.uiuc.edu/Research/vmd/.
94. I. A. Prior, P. D. Lewis and C. Mattos, "A comprehensive survey of Ras mutations in cancer," *Cancer Res.* **72**(10), 2457–2467 (2012).
95. A. J. Scheidig, C. Burmester and R. S. Goody, "The pre-hydrolysis state of p21(ras) in complex with GTP: New insights into the role of water molecules in the GTP hydrolysis reaction of ras-like proteins," *Structure* **7**(11), 1311–1324 (1999).
96. M. V. Milburn, L. Tong, A. M. deVos, A. Brünger, Z. Yamaizumi, S. Nishimura and S. H. Kim, "Molecular switch for signal transduction: Structural differences between active and inactive forms of protooncogenic ras proteins," *Science* **247**(4945), 939–945 (1990).
97. 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* **15**(12), 1618–1629 (2007).
98. G. Buhrman, G. Holzapfel, S. Fetics and C. Mattos, "Allosteric modulation of Ras positions Q61 for a direct role in catalysis," *Proc. Natl. Acad. Sci. U.S.A.* **107**(11), 4931–4936 (2010).
99. T. Ratovitski, L. B. Corson, J. Strain, P. Wong, D. W. Cleveland, V. C. Culotta and D. R. Borchelt, "Variation in the biochemical/biophysical properties of mutant superoxide dismutase 1 enzymes and the rate of disease progression in familial amyotrophic lateral sclerosis kindreds," *Hum. Mol. Genet.* **8**(8), 1451–1460 (1999).
100. G. S. Wright, S. V. Antonyuk, N. M. Kershaw, R. W. Strange and S. Hasnain, "Ligand binding and aggregation of pathogenic SOD1," *Nat. Commun.* **4**, 1758 (2013).
101. D. A. Case, T. A. Darden, T. E. III Cheatham, C. L. Simmerling, J. Wang R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Götz, I. Kolossváry, K.F.Wong, F. Paesani, J. Vanicek, R.M.Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman. AMBER 12, University of California, San Francisco (2012).
102. U. S. Vogel, R. A. Dixon, M. D. Schaber, R. E. Diehl, M. S. Marshall, E. M. Scolnick, et al, "Cloning of bovine gap and its interaction with oncogenic Ras p21," *Nature* **335**(6185), 90–93 (1988).
103. L. Gremer, B. Gilsbach, M. R. Ahmadian and A. Wittinghofer, "Fluoride complexes of oncogenic Ras mutants to study the Ras-RasGap interaction," *Biol. Chem*. **389**(9), 1163–1171 (2008).
104. J. B. Gibbs, M. D. Schaber, W. J. Allard, I. S. Sigal and E. M. Scolnick, "Purification of Ras GTPase activating protein from bovine brain," *Proc. Natl. Acad. Sci. USA* **85**(14), 5026–5030 (1988).
105. S. E. Neal, J. F. Eccleston and M. R. Webb, "Hydrolysis of GTP by p21NRAS, the NRAS protooncogene product, is accompanied by a conformational change in the wild-type protein: Use of a single fluorescent probe at the catalytic site," *Proc. Natl. Acad. Sci. USA* **87**(9), 3562–3565 (1990).
106. J. F. Diaz, B. Wroblowski and Y. Engelborghs, "Molecular dynamics simulation of the solution structures of Ha-ras-p21 GDP and GTP complexes: Flexibility, possible hinges, and levers of the conformational transition," *Biochemistry* **34**(37), 12038–12047 (Sep. 1995).
107. F. Noé, F. Ille, J. C. Smith and S. Fischer, "Automated computation of low-energy pathways for complex rearrangements in proteins: Application to the conformational switch of Ras p21," *Proteins: Struct. Funct. Bioinf.* **59**(3), 534–544 (1990).
108. R. Nussinov, H. Jang and C.-J. Tsai, "The structural basis for cancer treatment decisions," *Oncotarget* **5**(17), 7285–7302 (2014).
109. 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).
110. K. Molloy, Probabilistic Algorithms for Modeling Protein Structure and Dynamics *Ph.D. Thesis* (Fairfax, VA: George Mason University, 2015).

https://doi.org/10.1017/S0263574715001058 Published online by Cambridge University Press