

Statistical Analysis of Computed Energy Landscapes to Understand Dysfunction in Pathogenic Protein Variants

Wanli Qiao*
Dept of Statistics
George Mason University

Tatiana Maximova
Dept of Computer Science
George Mason University

Erion Plaku
Dept of Electrical Engineering and Computer Science
The Catholic University of America

Amarda Shehu†
Dept of Computer Science
George Mason University

ABSTRACT

The energy landscape underscores the inherent nature of proteins as dynamic systems interconverting between structures with varying energies. The protein energy landscape contains much of the information needed to characterize protein equilibrium dynamics and relate it to function. It is now possible to reconstruct energy landscapes of medium-size proteins with sufficient prior structure data. These developments turn the focus to tools for analysis and comparison of energy landscapes as a means of formulating hypotheses on the impact of sequence mutations on (dys)function via altered landscape features. We present such a method here and provide a detailed evaluation of its capabilities on an enzyme central to human biology. The work presented here opens up an interesting avenue into automated analysis and summarization of landscapes that yields itself to machine learning approaches at the energy landscape level.

CCS CONCEPTS

•Applied computing → Molecular structural biology;

KEYWORDS

protein energy landscape, dysfunction, mutation, basins, barriers, saddle points, landscape analysis

ACM Reference format:

Wanli Qiao, Tatiana Maximova, Erion Plaku, and Amarda Shehu. 2017. Statistical Analysis of Computed Energy Landscapes to Understand Dysfunction in Pathogenic Protein Variants. In *Proceedings of ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.*, 6 pages. DOI: <http://dx.doi.org/10.1145/3107411.310749>

1 INTRODUCTION

Proteins switch between different shapes/structures to interact with different molecular partners [15]. These motions, if visualized,

*Corresponding Author: wqiao@gmu.edu

†Corresponding Author: amarda@gmu.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM-BCB'17, August 20–23, 2017, Boston, MA, USA.

© 2017 ACM. ISBN 978-1-4503-4722-8/17/08...\$15.00

DOI: <http://dx.doi.org/10.1145/3107411.310749>

correspond to hops in a high-dimensional energy landscape that organizes structures by their potential energies. An energy landscape can reveal basins and barriers, with basins corresponding to thermodynamically-stable and semi-stable structural states, and barriers corresponding to higher-energy regions that slow down basin-to-basin diffusions and regulate the structural rearrangements that allow a protein to interact with different molecules [6].

Obtaining detailed representations of biomolecular energy landscapes is central to elucidating the structural dynamics that regulates the function or dysfunction of a biomolecule [15]. Some of the most complex human disorders, including cancer, are driven by DNA mutations that percolate to protein dysfunction [18]. While it is known that mutations percolate to dysfunction by changing the energy landscape and in turn the structural dynamics of a protein, quantifying changes to the landscape and dynamics of a protein in response to a mutation remains elusive [11]. Due to the disparate spatio-temporal scales involved, currently, neither wet- nor dry-laboratory techniques can fully reconstruct energy landscapes [21].

Recent algorithmic efforts circumvent the issue of disparate spatio-temporal scales by exploiting information available in resolved structures of healthy and diseased forms (pathogenic variants) of a protein [4]. Several stochastic optimization algorithms have been proposed and shown to reconstruct energy landscapes of various medium-size proteins with reasonable computational budgets (no more than a few CPU days) [3, 5, 12, 13, 22].

These algorithms, whether employing concepts from evolutionary search strategies (thus operating under the umbrella of evolutionary computation) or concepts adapted from robot motion planning (thus often referred to as robotics-inspired algorithms), effectively sample new structures (ranging anywhere from 50K to 1M structures depending on the energy function employed) of a protein sequence under consideration in a carefully-selected, low-dimensional variable space; the latter is extracted via Principal Component Analysis (PCA) of known structures. Some of these (robotics-inspired) algorithms integrate the latest AMBER energy functions, yielding detailed, sample-based representations of AMBER energy landscapes (containing thousands of new structures) that allow understanding the role of stable and semi-stable states and transitions between them in (dys)function [12, 13].

As a result of these recent algorithmic developments, it is now possible to reconstruct detailed energy landscapes of medium-size proteins with sufficient prior structure data. These developments allow us to turn our focus from computing to mining protein energy landscapes. Specifically, given the ability to compute landscapes

of variants of a protein effectively on demand and to do this in a comprehensive fashion, for the WT and many pathogenic variants, the goal is to develop computational tools for analysis and comparison of energy landscapes. The latter is necessary to quantify and summarize energy landscape features altered by mutations and then formulate thermodynamics-based hypotheses on the impact of sequence mutations on (dys)function.

In this paper, we present such a landscape analysis method. The method analyzes computed samples (representing structures of a protein sequence) and automatically identifies basins and ridges in the landscape comprised by the samples. Visualization of detected basins and ridges allows qualitatively comparing landscapes of healthy and pathogenic variants of a protein. More importantly, the method exposes the deepest point in a basin, which, alongside with saddle points allow us to define quantitative landscape descriptors. The latter, as we demonstrate here, open the way for landscape mining techniques to summarize landscapes, compare them, and even relate landscape features to wet-laboratory measurements of biological mechanisms impacted by mutations.

In Section 1.1, we provide a better context for the landscape analysis method proposed here via a background and summary of related work. The method is described in detail in Section 2, and evaluated on variants of H-Ras, a cell-growth regulating enzyme central to human biology and health. Our evaluation in Section 3 draws several mechanisms of interest by which oncogenic and non-oncogenic mutations alter features of the energy landscape and in turn impact the wildtype structural dynamics. An interesting avenue is also related on how automated landscape analysis opens the way for machine learning approaches to operate over energy landscapes. The paper concludes in Section 4 with a summary and description of future directions of work.

1.1 Related Work

Let $e(x)$ be an energy function defined on the domain $D \subseteq \mathbb{R}^d$ ($d \geq 1$). Denote $e_{\min} = \min_{x \in D} e(x)$ and $e_{\max} = \max_{x \in D} e(x)$. For $c \in [e_{\min}, e_{\max}]$, let the *lower-level set* $L(c) = \{x \in D : e(x) \leq c\}$. $L(c)$ may contain components with separate boundaries. The concept of a *basin*, though not explicitly defined in current literature, can be related to each component, but basins may have hierarchical structure (a smaller basin can be a subset of a larger basin). This structure ought to be revealed by a landscape analysis method but only emerges when iteratively decreasing the energy level c . As energy decreases, a basin shrinks until it disappears or gets split into more components with separate boundaries; i.e., more basins. The method we propose detects these scenarios.

Other concepts relate to *barriers* or *ridges*, which are collections of local maxima on paths between basins; *saddles* are critical points on these paths. Statistical analyses of ridges exist in literature [16, 19]. The identification of basins is strongly connected to ridges, but here we focus on basins, leaving ridge detection to future research.

Problems related to landscape analysis arise in other disciplines. The vast amount of cosmological data exhibits complex network-like spatial structures, consisting of clusters, filaments, walls, and voids. Persistent diagrams are employed to extract topological summarization of cosmological data [23]. Here, however, the focus is

on geometric features of spatial data (in particular, basins). A related method, proposed in [1], analyzes a nearest-neighbor graph of structures, seeking critical points in the graph in a discrete manner. Basins are represented by local minima. In contrast, our approach analyzes the smoothed energy landscape. Therefore, the method we propose is robust against both the ruggedness of the energy landscape and the non-uniform sampling density that is often expected from stochastic optimization methods. Additionally, the basins identified by our method are sub-regions of the landscape and have clear boundaries that facilitate visualization.

2 METHODS

The SoPriM algorithm [12] and its faster version, SoPriMp [10], leverage known structures to obtain an ensemble of structures that provide a discrete, sample-based representation of the energy landscape of a protein sequence of interest. The algorithms feasibly provide detailed representations of landscapes of many medium-size proteins [13]. Both algorithms do not directly operate in the structure space of a protein, so as to circumvent the dimensionality issue, but instead compute samples that represent structures in a low-dimensional variable space (of principal components – PCs). Fast transformations between the variable space and the all-atom structure space allow the algorithms to obtain low-energy all-atom structures. We refer the reader to work in [10] for more details. Here we employ SoPriMp (due to its higher exploration capability [10]) to provide the proposed landscape analysis method with samples (points in the variable space) corresponding to computed all-atom structures with associated Amber ff14SB energies.

The proposed method explicitly reconstructs the landscape by finding basins and saddles. Here we limit the application to automated analysis of 3D samples, using as observations the PC1, PC2 coordinates of the SoPriMp-computed samples and the Amber ff14SB energy values of the structures corresponding to the samples. In principle, all the variables employed by SoPriMp can be used (all selected principal components, which typically range from 10 to 25), but the number of variables determines the dimensionality of the space; increasing dimensionality adds to the computational cost associated with the analysis. It is worth noting that cases where SoPriMp can be employed to obtain sample-based representations of energy landscapes, the top two PCs capture more than 50% of the structural variance, and the top three capture well above 70% of the variance. In this paper, we demonstrate that a coarse and fast analysis can still be very valuable to extract features of a landscape. The emphasis on reasonable computational costs is due to our objective to apply this algorithm in a comparative analysis setting, where we screen dozens or more variants of a protein in search of energetic features that help us summarize, categorize, and elucidate the impact of mutations on dysfunction.

Finding Basins and Saddles

We refer to the method as Basin.Finder, though the method finds basins and saddles separating basins. The method first finds the α -convex hull [17, 20] of the PC1-PC2 sample locations and then defines a 2D grid over the samples in the hull; the distance between adjacent grid points is a parameter δ_1 . The collection of the resulting grid points is denoted as S_{\max} . The energy of each grid point is

estimated via the Nadaraya-Watson (NW) kernel regression [14]. At each grid point x , the energy estimate is the weighted average of the observed sample energies in a small neighborhood (bandwidth h) around x . We note that kernel regression is a smoothing technique that is central in spatial data analysis (the amount of smoothing is controlled by h). In particular, energy landscapes reconstructed with all-atom energy functions, such as Amber ff14SB here, are overly rugged [11]. Kernel regression is a mechanism to reduce the ruggedness and also address non-uniform density of samples.

Let the minimum and maximum (estimated) energy on S_{\max} be c_{\max} and c_{\min} , respectively. The method then iteratively decreases the energy level c , starting from c_{\max} by a small step δ_2 , detecting when basins split, and storing detected basins and basin-separating saddle points in a list Ω for further visualization and quantitative analysis. A recursive implementation of the proposed Basin_Finder method is shown in pseudocode in Algo 1. The initial arguments are S_{\max} , c_{\max} , and the list Ω , which is initially empty.

On any given collection S of grid points, the method calculates the (possibly more than 1) k boundaries of S using the α -convex hull (lines 3-4). If S is one component, the method proceeds to identify possible hierarchical structure emerging at lower-energy levels, decreasing c by δ_2 (line 6) and updating the grid locations that meet the energetic threshold (line 7) before calling itself recursively (lines 12-13). If instead S contains $k > 1$ components with separate boundaries, a basin splitting event has been detected (lines 8-9). The newly identified basins $S_1 \dots S_k$ (line 8) are added to the list Ω (line 11). Possible saddle points emerging at the boundaries of neighboring basins are also identified (line 10) and added to the list Ω . The minimum distance between vertices on polygonal boundaries of different basins S_i, S_j is computed, and if this distance is not above a threshold d_{th} (S_i, S_j are deemed neighboring), the middle point of the minimum-length line is estimated to be the saddle point saddle_{ij} . The method terminates when no more grid locations are left (lines 1-2).

Algo. 1 Basin_Finder(S, c, Ω)

```

1: if  $S = \emptyset$  then
2:   return
3:  $\partial S \leftarrow bd(S)$  //boundary of  $S$  using the  $\alpha$ -convex hull
4:  $k \leftarrow nb(\partial S)$  //number of separated boundaries in  $\partial S$ 
   //no basin splitting detected
5: if  $k = 1$  then
6:    $c \leftarrow c - \delta_2$  //lower energy
7:    $S_k \leftarrow$  subset of  $S$  with energy  $\leq c$  //update lower-level set
   //basin splitting detected
8: else
9:    $\{S_1, \dots, S_k\}$  //subsets of  $S$  with separate boundaries
   //are new basins
10:   $\text{saddle}_{i,j} \leftarrow$  saddle point between neighboring basins  $S_i, S_j$ 
11:   $\Omega \leftarrow \Omega \cup \{S_1 \dots S_k\} \cup \{\text{saddle}_{ij}\}$ 
12: for  $1 \leq i \leq k$  do
13:   Basin_Finder( $S_i, c, \Omega$ ) //seek further hierarchical structure

```

For clarity, the pseudocode in Algo 1 does not show some modifications we make to lower computational costs. Since the basins

in which we are most interested are in the low-energy regions of the landscape, the dense boundaries in the high-energy regions can be both distracting and consuming of computational time. So, the method is adjusted via a cutoff point c_0 . If $c \leq c_{\min} + c_0$, the method is run without any adjustment. If $c > c_{\min} + c_0$, we make use of a threshold n_0 , such that, if the number of grid points within any basin S_i is less than n_0 , we do not dig further into S_i for any hierarchical structure in it.

The value of δ_2 also has an impact on computational costs. We balance between keeping computational costs reasonable and capturing the exact moment when a basin splits. Instead of always lowering the energy level by δ_2 (line 6), we first perform a jump test by decreasing the energy by $m\delta_2$ for some moderately large m . If the resulting lower level set at the level $c - m\delta_2$ splits into smaller basins, it means we have missed the important moment of splitting; so, we revert back to $c - \delta_2$. Otherwise, we conclude that no interesting events occur between the current energy level c and $c - m\delta_2$. We then directly move down to the level $c - m\delta_2$ and continue with the jump test until the test fails. Again, jumps are only applied for $c > c_{\min} + c_0$, because the computational costs decrease with $c \leq c_{\min} + c_0$ (in general, computation time correlates negatively with c). It is still possible that some very small and shallow basins can quickly appear and then disappear in a jump between c and $c - m\delta_2$. In this sense, the parameter m has a similar effect as n_0 in terms of ignoring small basins. This approach recognizes that it is best to apportion computational costs and thus keep all the details in the region with energy level $c \leq c_{\min} + c_0$.

Implementation Details

Basin_Finder is implemented in R and uses $\delta_1 = 0.1$, $\delta_2 = 0.3$, $m = 50$, $c_0 = 200$, $n_0 = 100$, $\alpha = 0.15$ (parameter for α -convex hull), and $d_{\text{th}} = 20 \cdot \alpha$. Different values of h are analyzed (see Section 3). Gaussian kernel is used in the NW regression. On each of the test cases related in Section 3, the method takes anywhere from 12-24 CPU hours to complete on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University. Compute nodes used for testing are Intel Xeon E5-2670 CPU with 2.6GHz base processing speed and 3.5TB of RAM.

3 RESULTS

First we relate 2D landscapes obtained via the method described in Section 2. We do so for the H-Ras WT and several diseased variants and draw larger observations into what mechanisms the different mutations employ to alter the energy landscape with implications for dysfunction. Second, we extract quantitative descriptors of basins and basin-separating saddles detected automatically by the method proposed in Section 2 and showcase how such descriptors can be correlated to wet-laboratory measurements that allow us to relate specific characteristics of mutation-altered landscapes to biological mechanisms in which H-Ras participates.

3.1 Visualization and Qualitative Comparison of 2D Landscapes

All the landscapes reconstructed by our method on variants of H-Ras add specific energetic features to the schematic shown in Fig. 1. Fig. 1 summarizes the presence of two large basins, one

corresponding to the GTP-activated state (also known as On), and the other to the GDP-activated state (also known as Off). A saddle point in the schematic denotes the presence of a barrier separating the two basins. The GTP-activated basin is larger, and contains many structures reported by different wet laboratories, such as those with PDB ids 1QRA, 1CTQ, 3L8Y, 2RGD, 3K8Z, and more. The latter three structures represent allosteric states of H-Ras, denoted as Reactive (R) and Tardy (T). Interconversion between the R- and T-states is reported in [8] and suggested to be more important than the On to Off interconversion for dysfunction in oncogenic variants. Visual comparison of landscapes reconstructed (with contour lines denoting detected basins) shows that changes from mutations impact the size of basins, the appearance or disappearance of basins within basins, elevation of existing barriers, and/or separation/split of basins.

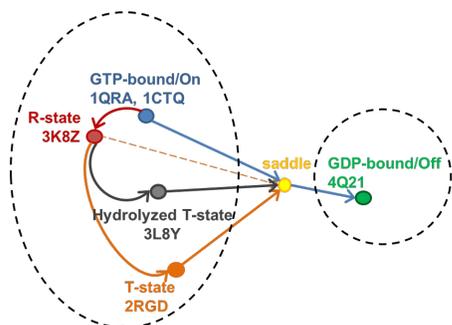


Figure 1: Known and putative Ras state interconversions.

Reconstructed landscapes of representative variants are shown in Fig. 2. We first draw attention to the impact of the bandwidth value by showing the landscape of H-Ras WT at two different such values (top panel of Fig. 2). Lower bandwidth values provide more detail, detecting narrow basins, as well as basins within basins. Higher bandwidth values may miss some of these details; for instance, the GTP-activated structures of H-Ras, under PDB ids 1QRA and 1CTQ occupy their own narrow basin (within the larger On basin), which is smoothed away at the higher bandwidth value (see top panel of Fig. 2). The hydrolyzed T-state (represented by the structure with PDB id 3L8Y) is within the larger, GTP-activated basin but appears to be just outside the basin at the higher bandwidth value (due to basin narrowing by the kernel regression). The R- and T-states (PDB ids 3K8Z, 2RGD, 3L8Y) reside within one basin, validating work in [8] that reports interconversions of these structures. In contrast, the interconversion of the R- and T-states into the GTP-activated forms needs external energy, as correctly reproduced in our work in the form of an energy barrier. However, the hydrolyzed T-state appears to be just outside the basin at the higher bandwidth value (due to basin narrowing by the kernel regression). We have determined that bandwidth values ≥ 0.8 remove too many important energetic features, whereas the high level of detail obtained when using values ≤ 0.5 makes visualization difficult. For this reason, all 2D landscapes shown and visually compared in the rest of the analysis are those obtained with a bandwidth value of 0.7.

In Fig. 2 we also show landscapes of select oncogenic (middle panel) and non-oncogenic but syndrome-causing (bottom panel) variants. Comparison of landscapes (using the WT landscape as

baseline) allows making the following observations (also holding for other variants not shown here). In oncogenic variants, the barrier between the On and Off basins typically rises (Q61L illustrates this mechanism). The Off basin shrinks or disappears (see Q61L and F28L). The On basin splits, separating the R- and T-states (see Q61L). We note that the rigidification of H-Ras (due to barrier elevation or shrinkage or disappearance of the Off basin) is validated by prior work [3]. A novel feature that prior work does not capture is the separation of the R- and T-states (due to basin splitting) in oncogenic variants. This, together with wet-laboratory work suggesting that the R-to-T interconversion in H-Ras is central to function suggests an interesting mechanism via which oncogenic mutations percolate to dysfunction, by essentially disrupting the allosteric switch. On syndrome-causing variants, changes over the WT are less drastic (see bottom panel of Fig. 2) but degeneracy emerges. The Off basin leaks into other regions (K5N and Q22R are illustrative examples), or both the Off and On basins degenerate, even merging with one another and spilling over much of the landscape (see Q22R). The degeneracy suggests an interesting mechanism for dysfunction via delay of the On-to-Off diffusion (by, for instance, internal diffusions within the larger, degenerate Off basin).

3.2 Landscape Mining

We now compare landscape descriptors with biochemical parameters of several catalytic activities of Ras measured in the wet laboratory and reported in [2, 7]. These include GTP activation, GAP sensitivity, (MEK, ERK) activation of the RAF-kinase pathway and AKT activation of the PI3K-kinase pathway, GTP/GDP dissociation, GEF activity of SOS1, intrinsic GTP hydrolysis, GAP-regulated hydrolysis, and RAF1-RBD binding affinity. We number these as P0 through P9 (so, P2-P4 for the three kinase pathways). In [2, 7], these parameters are reported for the WT, 2 oncogenic variants (G12V and F28L) and 12 non-oncogenic but syndrome-causing variants (K5N, V14I, Q22E, Q22R, P34L, P34R, T58I, G60R, Y71H, K147E, E153V, F156L). We curate and normalize these parameters (not shown) to allow the following correlation-based analysis.

All 15 variants characterized in [2, 7] are subjected to SoPriMp and the method presented here. We extract descriptors related to spatial and energetic distances of known states of H-Ras (these states are related in Fig. 1) to the saddle point separating the On and Off basins. Each state can be mapped to the global minimum of the basin containing it, or alternatively, the location and energy of the known wet-lab structure representing it; we relate results with the latter, as results are similar. For instance, the spatial distance $d(\text{On}, \text{Saddle})$ is measured via Euclidean distance in PC1-PC2 space between the canonical structure representing this basin (PDB id 1QRA) and the saddle point detected by the proposed landscape analysis method. The energetic distance $dE(\text{On}, \text{Saddle})$ measures the height of the barrier ($dE(\text{Saddle}) - E(\text{On})$). So, each landscape is summarized with these 10 descriptors. Each descriptor across 15 variants is compared with each biochemical parameter reported for these variants and Pearson correlations are measured.

Table 1 lists comparisons resulting in correlations ≥ 0.5 . These yield many insights. We highlight a few. Work in [2, 7] reveals that intrinsic hydrolysis (named P7 by us) is higher in the variants over the WT. Table 1 shows that this occurs due to elevated barriers of all

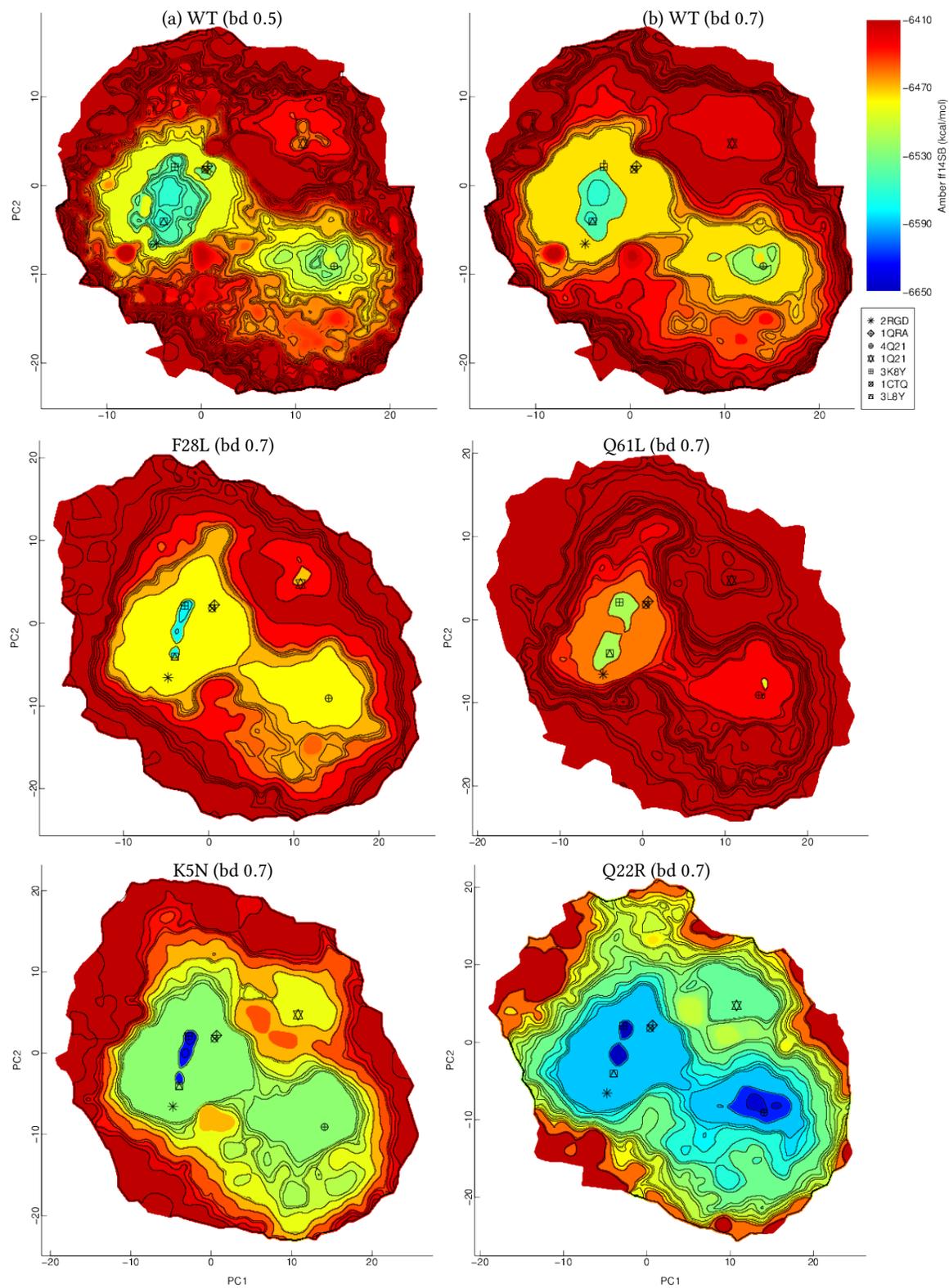


Figure 2: Top panel: landscape of WT at two bandwidth values. Landscapes of selected oncogenic (middle panel) and syndrome-causing (bottom panel) variants are also shown. The color coding-scheme is based on Amber ff14SB energy values estimated for every grid point as described in Section 2. Symbols that annotate projections of select experimentally-known structures are also shown.

states (positive correlations), movement of the Off state away from the saddle point (positive correlation), and movement of all other states towards saddle point (negative correlations). This suggests that equilibrium diffusions from the various states within the On basin to the Off basin directly relate to intrinsic hydrolysis; thus, this activity is perturbed in pathogenic variants by changing landscape features. GAP-catalyzed (as opposed to intrinsic) hydrolysis (P8) is another activity impacted by mutations. Table 1 shows correlations of 0.50 and 0.53 between P8 and Off-to-saddle and hydrolyzed T-to-saddle barrier height variations, which suggests a specific role of these states in GAP-catalyzed hydrolysis. A prior study relating FoldX energies (FoldX is a protein design algorithm) of specific structures to biochemical parameters in [2, 7] could only obtain two correlations, 0.65 for intrinsic hydrolysis and 0.43 for GAP-activated hydrolysis [9]. The highest correlations we obtain are -0.85 and 0.58 , respectively. In addition, Table 1 shows that spatial and energetic distances of states from the On-to-Off saddle point correlate well with parameters measuring GTP activation (P0) and ERK activation of the RAF-kinase pathway (P3). These results show that the On and R-states are important for activation of this pathway, and suggest that the increased barrier heights between the GTP-activated states and the saddle point delay activation and so increase the amount of unbound GTP in pathogenic variants.

Table 1: Measured landscape descriptors and biochemical parameters (reported in [2, 7]) with correlations ≥ 0.5 . T* indicates hydrolyzed T-state.

State	d(State, Saddle)	dE(State, Saddle)
On	P7(-0.84), P3(0.53)	-
Off	P7(0.83)	P7(0.58), P0(0.54), P8(0.50)
T-	P7(-0.79)	P0(0.62)
R-	P7(-0.85), P3(0.51)	P7(0.61), P0(0.51)
T*	P7(-0.82)	P7(0.62), P0(0.54), P8(0.53)

4 CONCLUSION

This paper evaluates a new line of research on mining energy landscapes of protein variants as a means of elucidating how mutations associated with various disorders alter the landscape. This is now possible due to a method that obtains detailed sample-based representations of energy landscapes of medium-size proteins with experimentally-resolved structures and a novel method, described here, that can reconstruct landscapes thus facilitating visualization, as well as automatically extract basins and saddles from them.

Visual comparison of reconstructed landscapes of pathogenic H-Ras variants validates prior dry- and wet-lab work and reveals novel mechanisms via which mutations percolate to dysfunction. The quantitative analysis suggests that a simple correlation-based investigation can reveal insights into which structures and inter-conversions can be related to specific Ras activities. Many more quantitative descriptors can be extracted from the list of basins and saddles that the proposed method extracts from a reconstructed landscape. This opens the way for learning from computed landscapes and macroscopic observations made in the wet laboratory. Altogether, the results suggest the approach of an exciting stage

where one can compute and then mine landscapes of protein variants to learn *in-silico* models of how mutations impact function, as well as elucidate the role of specific structures and structural rearrangements in key biological activities.

5 ACKNOWLEDGMENTS

This work is supported in part by NSF CCF No. 1421001 and NSF IIS CAREER Award No. 1144106.

REFERENCES

- [1] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. 2015. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. of Computational Chemistry* 36, 16 (2015), 1213–1231.
- [2] I. C. Cirstea et al. 2013. Diverging gain-of-function mechanisms of two novel KRAS mutations associated with Noonan and cardio-facio-cutaneous syndromes. *Human Molecular Genetics* 22, 2 (2013), 262–270.
- [3] R. Clausen, B. Ma, R. Nussinov, and A. Shehu. 2015. Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Comput Biol* 11, 9 (2015), e1004470.
- [4] R. Clausen and A. Shehu. 2014. A Multiscale Hybrid Evolutionary Algorithm to Obtain Sample-based Representations of Multi-basin Protein Energy Landscapes. In *Conf on Bioinf and Comp Biol (BCB)*. ACM, Newport Beach, CA, 269–278.
- [5] R. Clausen and A. Shehu. 2015. A Data-driven Evolutionary Algorithm for Mapping Multi-basin Protein Energy Landscapes. *J Comp Biol* 22, 9 (2015), 844–860.
- [6] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. 1991. The energy landscapes and motion on proteins. *Science* 254, 5038 (1991), 1598–1603.
- [7] L. Gremer et al. 2011. Germline KRAS mutations cause aberrant biochemical and physical properties leading to developmental disorders. *Human Mutation* 32, 1 (2011), 33–43.
- [8] C. W. Johnson and C. Mattos. 2013. The allosteric switch and conformational states in Ras GTPase affected by small molecules. *Enzymes* 33, Pt. A (2013), 41–67.
- [9] C. Kier and C. Serrano. 2014. Structure-energy-based predictions and network modelling of RASopathy and cancer missense mutations. *Mol Syst Biol* 10, 5 (2014), 727–.
- [10] T. Maximova, D. Carr, E. Plaku, and A. Shehu. 2016. Sample-based Models of Protein Structural Transitions. In *Conf Bioinf and Comp Biol (BCB)*. ACM, Seattle, WA, 128–137.
- [11] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput Biol* 12, 4 (2016), e1004619.
- [12] T. Maximova, E. Plaku, and A. Shehu. 2015. Computing Transition Paths in Multiple-Basin Proteins with a Probabilistic Roadmap Algorithm Guided by Structure Data. In *Intl Conf on Bioinf and Biomed (BIBM)*. IEEE, Washington, D.C., 35–42.
- [13] T. Maximova, E. Plaku, and A. Shehu. 2016. Structure-guided Protein Transition Modeling with a Probabilistic Roadmap Algorithm. *IEEE/ACM Trans Comput Biol & Bioinform* 13, 5 (2016), 1–14.
- [14] E. A. Nadaraya. 1964. On estimating regression. *Theory of probability and its applications* 9, 1 (1964), 141–142.
- [15] R. Nussinov and P. G. Wolynes. 2014. A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys Chem Chem Phys* 16, 14 (2014), 6321–6322.
- [16] U. Ozertem and D. Erdogmus. 2011. Locally defined principal curves and surfaces. *Journal of Machine Learning Research* 12 (2011), 1249–1286.
- [17] B. Pateiro-Lopez. 2008. *Set estimation under convexity type restrictions*. Ph.D. Dissertation. Universidad de Santiago de Compostela.
- [18] I. A. Prior, P. D. Lewis, and C. Mattos. 2012. A comprehensive survey of Ras mutations in cancer. *Cancer Res* 72, 10 (2012), 2457–2467.
- [19] W. Qiao and W. Polonik. 2016. Theoretical analysis of nonparametric filament estimation. *Annals of Statistics* 44, 3 (2016), 1269–1297.
- [20] H. Rodriguez-Casal. 2007. *Set estimation under convexity type assumptions. Annales de l'I.H.P.- Probabilites & Statistiques* 43 (2007), 763–774.
- [21] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Velázquez-Muriel, and A. Sali. 2009. The structural dynamics of macromolecular processes. *Curr Opin Cell Biol* 21, 1 (2009), 97–108.
- [22] E. Sapin, D. B. Carr, K. A. De Jong, and A. Shehu. 2016. Computing energy landscape maps and structural excursions of proteins. *BMC Genomics* 17, Suppl 4 (2016), 456.
- [23] R. van de Weygaert et al. 2011. Alpha, betti and the megaparsec universe: on the topology of the cosmic Web. *Transactions on Computational Science XIV* 6970 (2011), 60–101.