Path-based Guidance of an Evolutionary Algorithm in Mapping a Fitness Landscape and its Connectivity

Emmanuel Sapin
Dept of Computer Science
George Mason University
Fairfax, VA 22030
esapin@gmu.edu

Kenneth De Jong
Dept of Computer Science
George Mason University
Fairfax, VA 22030
kdejong@gmu.edu

Amarda Shehu*
Dept of Computer Science
George Mason University
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

Understanding function regulation in proteins that switch between different structural states at equilibrium requires both finding the basins that correspond to such states and computing the sequence of intermediate structures employed (i.e., the path taken) in basin-to-basin switching. Recent work suggests that evolutionary strategies can be used to map protein energy landscapes effectively. Further work has shown that the constructed maps can be additionally equipped with connectivity information to help identify basin-switching paths. Here we highlight a potential issue when the problems of mapping and path finding are considered separately. We conduct a simple, proof-of-principle study that demonstrates the ability of an EA to allow extracting better paths from an EA-built map when the EA is supplied with the right information. The study is conducted on two key, multi-state proteins of importance to human biology and disease. The results presented here suggest that further research efforts to guide an EA with path-based information are warranted and feasible.

Keywords

mapping; protein energy landscape; evolutionary algorithm; protein modeling; computational structural biology.

1. INTRODUCTION

Studying protein energy landscapes is gaining renewed attention in computational structural biology. Mapping energy landscapes is now seen key to understanding a wide range of phenomena, including the structure-function relationship in protein molecules [4, 12, 3, 11]. Moreover, steady algorithmic and hardware advances are increasingly making it feasible to study landscapes of chains that are longer than a few amino acids [6].

*Corresponding Author

GECCO'16 Companion, July 20-24, 2016, Denver, CO, USA

DOI: http://dx.doi.org/10.1145/2908961.2931718

Mapping protein energy landscapes is particularly important on intrinsically-dynamic proteins that switch between different stable and semi-stable structural states at equilibrium for function regulation. Recent work demonstrates the effectiveness of evolutionary algorithms (EA) in mapping the complex, multi-modal landscapes of such proteins with practical computational budgets on modest computational resources. In particular, work in [2, 1, 9, 10] presents steady improvements to an underlying EA framework to move beyond the classic optimization setting, where the goal is to locate the global minimum of a fitness/energy landscape, to a mapping setting in which the goal is to construct a map of a protein energy landscape that retains important energetic features such as deep and broad basins corresponding to the thermodynamically-stable and semi-stable structural states of a protein.

The top panel of Figure 1 summarizes the methodological contributions in recent published work. In particular, the hall of fame mechanism is employed in [9] to serve as a map that is dynamically updated to contain individuals corresponding to non-redundant local minima in the landscape. These efforts have shown that the combination of domain-specific insight and EAs with carefully defined initialization, variation, improvement, and selection operators allows building maps that reproduce known basins of proteins characterized extensively in wet laboratories (typically due to central roles in human biology and disease), as well as exposes novel interesting structural states not captured in wet laboratories [1, 10]. The mapped landscapes provide an opportunity to identify paths that demonstrate how a protein hops between successive structures to switch between two known structural states. In [8], as summarized in the bottom panel of Figure 1, a new method is proposed that embeds individuals stored in the hall of fame into a nearest-neighbor graph, and then uses the graph to identify low-cost paths connecting two structures of interest.

This line of work has led to the following observation. An EA designed to find the basins of an energy landscape will apportion computational resources to exploitation of lower-energy regions in expense of further exploration of high-energy regions. While this is a rational decision to do when the goal is to map the local minima in an energy landscape, this decision may impact the ability to find detailed paths connecting two structures of interest. We illustrate this observation on one particular protein, the superoxide dismutase [Cu-Zn] (SOD1), which is a central protein with mutations implicated in familial Amyotrophic lateral sclerosis (ALS). Due to its role in human disease and a clearly

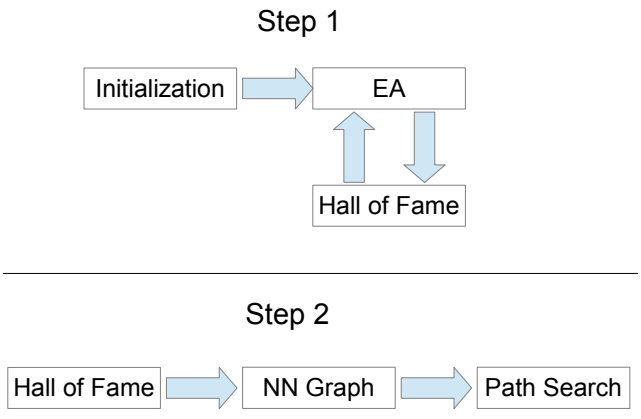

Figure 1: An EA can be used to construct a map of a protein energy landscape via the hall of fame mechanism. This can be considered as step 1 in an approach that later employs the map to identify low-cost paths between two structures of interest. In [8], the hall of fame is embedded in a nearest-neighbor graph that is explored for low-cost paths connecting two given structures.

bimodal landscape, this protein has been employed as a test case for various evolutionary strategies for exploration of protein energy landscapes [2, 9].

Figure 2 shows the energy landscape of SOD1 (wildtype – WT – form) projected on the top two principal components (PCs); One of the evolutionary strategies central to the ability to explore the high-dimensional protein energy landscapes is the employment of Principal Component Analysis (PCA) on a set of known wet-laboratory structures of a protein to define the underlying variable space (space of PCs) for the EA. These known structures are also employed to fill in the initial population.

The interested reader is referred to work in [2] for details on the data collection and PCA procedure. Figure 2 essentially shows the projection of individuals in the hall of fame constructed by the EA onto the top two PCs; the projections are color-coded according to the fitness, which is measured via the Rosetta all-atom energy function `score12` [5]. Two basins can be clearly seen in the two-dimensional projection of the map. The majority of the known wet-laboratory structures (drawn as black dots) fall in either of the basins; a few, corresponding to variants (mutated versions) of the protein, fall in between the basins, which means that these structures, while stable for variants, are higher-energy for the WT sequence.

Figure 2 shows that the majority of individuals in the map fall on or near the basins. Fewer individuals populate the high-energy region between the basins, as an EA designed to find basins has to balance between exploration and exploitation in order to apportion the resources on promising, low-energy regions of the variable space. Figure 2 shows what happens when the map is then queried for a path connecting two structures, one in each of the visible basins. The path shown is the lowest-cost path computed via Dijkstra's shortest-path algorithm, after embedding individuals in the map in a nearest-neighbor graph. The graph is directed, and edges of the graph are weighted, with weights recording only energetic increases (0 REUs otherwise).

The edges of the path shown in Figure 2 can be as long as 2.72 Å; this means that the structures connected by an edge in the path can be as far away as 2.72 Å, when the distance between them is measured via least root-mean-squared deviation [7] over their CA atoms. This distance for an edge is too high for a path to be considered realistic in terms of thermal fluctuations of a protein at equilibrium. One can refer to such a path as a low-resolution approximation of the true structural excursion.

Connecting via an edge two structures that are not near in the landscape is the equivalent of tunneling through a possible mountain; the latter, however, is not present in the map due to the focus on exploitation of basins. As a consequence, under sparse sampling of energetic barriers that connect two distinct basins, reported paths may underestimate the true cost.

In this paper, we investigate this issue further in a proof-of-principle setting. Specifically, we pursue the hypothesis that supplying the EA with information on the location of regions of interest for connecting paths indeed improves sampling of these regions and allows finding more detailed, finer-resolution paths that are better approximations of the true structural excursion. Specifically, here we supply this information to the initialization operator and demonstrate that indeed, more detailed and accurate paths are obtained. We provide evaluation on two extreme test cases, SOD1, where regions needed for connectivity are severely under-sampled, and H-Ras (a protein of importance in human cancer), where the EA provides better coverage of such regions. The results presented here suggest future research on directing the EA towards regions of interest for basin-basin excursions.

2. METHODS

We investigate the following setup. The EA proposed in [9] is employed to construct a map of a protein's energy landscape. The EA runs for a fixed budget of fitness evaluations (details can be found in [9]). When the EA terminates (step 1 in Figure 1), the map is saved, and step 2 in Figure 1 starts; the map is recast as a nearest-neighbor graph; that is, the individuals in it are connected to k nearest neighbors. The Euclidean distance in variable space is used to determine the distance between two neighbors. The graph is directed, and edge weights record only energetic increases. Dijkstra's algorithm is then used over the graph to obtain the lowest-cost path connecting two structures of interest. Further low-cost paths can be obtained if the individuals in the path are removed from the map, and Dijkstra is run again. The process can be repeated until no paths can be obtained anymore, effectively providing a set of low-cost paths connecting two structures of interest with individuals in the constructed map. Details on parameters and implementation of the various procedures are presented in [8].

As shown in Figure 2 for SOD1, the obtained path may be forced to connect two structures that, while nearest-neighbors in the variable space, are far apart in structure space. So, the following setting (summarized in Figure 3) is pursued here. The EA is run again, though with a much smaller computational budget. In the first EA, the initial population consists of individuals corresponding to wet-laboratory structures of a protein and individuals sampled at random in the variable space. In the second/following EA,

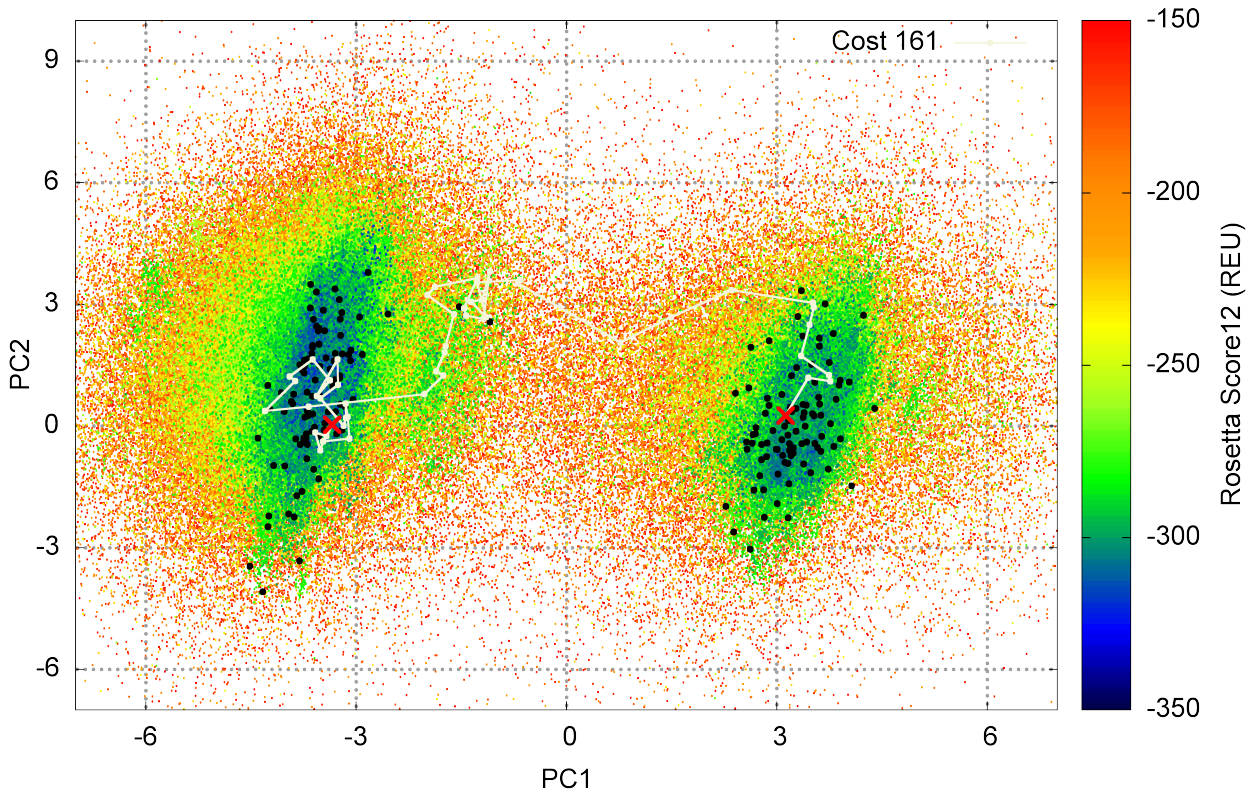

Figure 2: Visualization in 2D of the map and a lowest-Cost path computed for SOD1. PC1 and PC2 refer to the top two PCs. The red-to-blue color code scheme follows high-to-red Rosetta `score12` energy values (measured in Rosetta Energy Units – REUs). Black dots show projections of wet-laboratory structures of both SOD1 WT and variants. A lowest-cost path connecting two structures, one in each of the visible basins, is also shown. The maximum edge length does not exceed 2.72Å. The cost of the path is in REUs.

the initial population contains only individuals corresponding to wet-laboratory structures of a protein and individuals in the lowest-cost path found by the above methodology. Note that the second EA adds individuals to the map. The methodology summarized above on querying a map for a lowest-cost path is repeated on the final map obtained after the second EA terminates. Both the new map and low-cost paths are compared to the old map and old low-cost paths on each of the proteins studied here. Results follow.

3. RESULTS

The methodology described in Section 2 is applied to SOD1 and H-Ras. The (first) EA summarized in Section 2 is applied to each protein using a budget of $1,000,000$ fitness evaluations. This corresponds to roughly 13 days on 16 CPUs for SOD1 and 8 days on 16 CPUs for H-Ras (the CPUs are employed in an embarrassing parallelization scheme to improve (the fitness of) structures corresponding to individuals with the computationally-demanding Rosetta *relax* protocol). The second EA is run for a more modest computational budget of $200,000$ fitness evaluations.

Results on SOD1 are related first. The map obtained by the first EA and the lowest-cost path connecting two specific structures residing in each of the discovered basins are shown in Figure 2. It is worth noting that no more paths are obtained if the individuals in the lowest-cost path

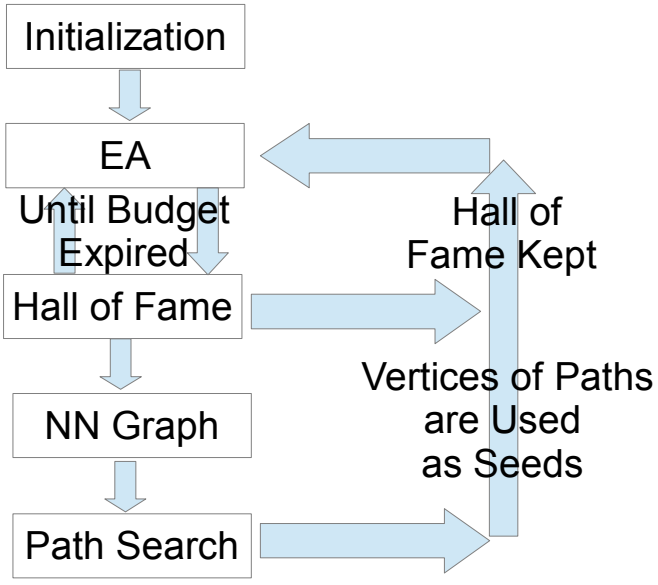

Figure 3: Summary of the new approach investigated here.

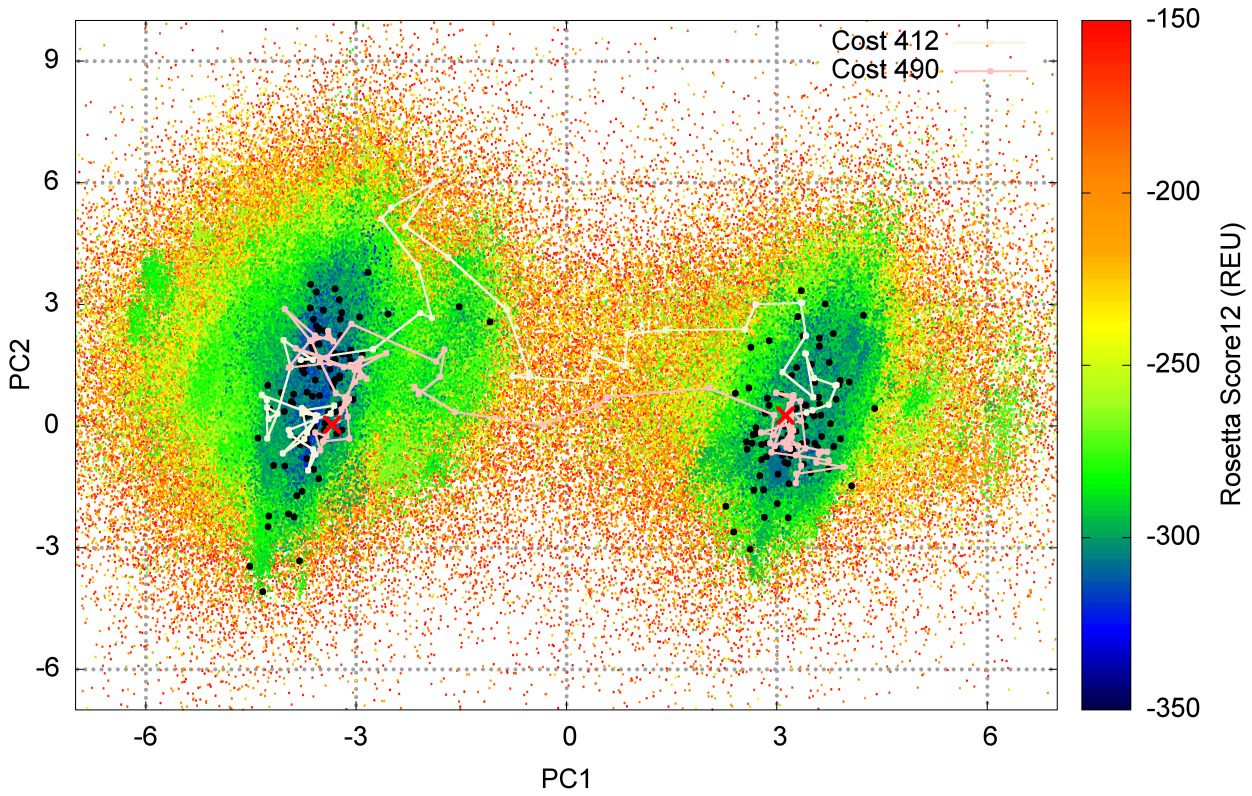

Figure 4: Visualization in 2D of the map and low-cost paths computed for SOD1, using (score-based) color-coded projections on the top two PCs. Projections of wet-laboratory structures are shown as black dots. The top panel shows the map and paths after the first iteration. The bottom panel shows the new map after the second iteration (path-guided EA), and the new low-cost paths obtained after querying the new map. The maximum edge length does not exceed 2.17Å.

are removed from the map and Dijkstra is invoked again. The map obtained after the second is shown in Figure 4. Querying the map for low-cost paths until the input start and goal structures are disconnected (after repeated removal of individuals in the discovered paths from the map) results in the paths shown in Figure 4.

Several observations can be made by comparing Figure 4 to Figure 2. First, the second has managed to add more low-energy individuals to the map, both in the two main basins and, more importantly, in intermediate regions. The latter has allowed finding two paths as opposed to one path. Both paths are finer; the maximum edge length is 2.17Å, an improvement of about 0.6Å over the first run. The costs have increased, confirming one of the possibilities raised that connecting two structures far away in structure space may underestimate the true energetic cost of a structural excursion. In fact, the shorter edge lengths here allow the query process to more faithfully follow the ruggedness of the energy landscape. Taken altogether, this study on SOD1 suggests that exploration in an EA can be influenced (via the initial population mechanism) by information on which regions are needed for connectivity to then allow more accurate modeling of structural excursions.

While the results on SOD1 show an extreme case of a landscape with a prominent energetic barrier that is not sufficiently sampled by an EA. The results on H-Ras are shown in Figure 5. The top panel shows the results of running the first EA and then querying for low-cost paths connecting individuals closest to two wet-laboratory structures representing the active/on and inactive/off structural states of H-Ras. The maximum edge length in the paths is 0.115Å, showing that the EA has achieved very good sampling of the structure space. The bottom panel in Figure 5 shows the new map and the new low-cost paths obtained after following up with the second, path-guided EA. The new paths obtained, shown in the bottom panel, are not more detailed than the older ones, though on average their energetic cost is lower. This suggests that no further exploration can be obtained, and the new individuals in the map improve slightly the average energetic cost but not the resolution of the paths.

4. CONCLUSION

The work presented here expands upon a novel direction of research on EAs designed for exploring protein energy landscapes. While a body of recent work illustrates how EAs can be used to efficiently map the complex energy landscapes of proteins, here we draw attention to an issue that presents itself when the ultimate objective of the maps is to use them to identify the structural paths taken by dynamic proteins during basin-to-basin switches. We present a simple approach that guides the EA exploration not only

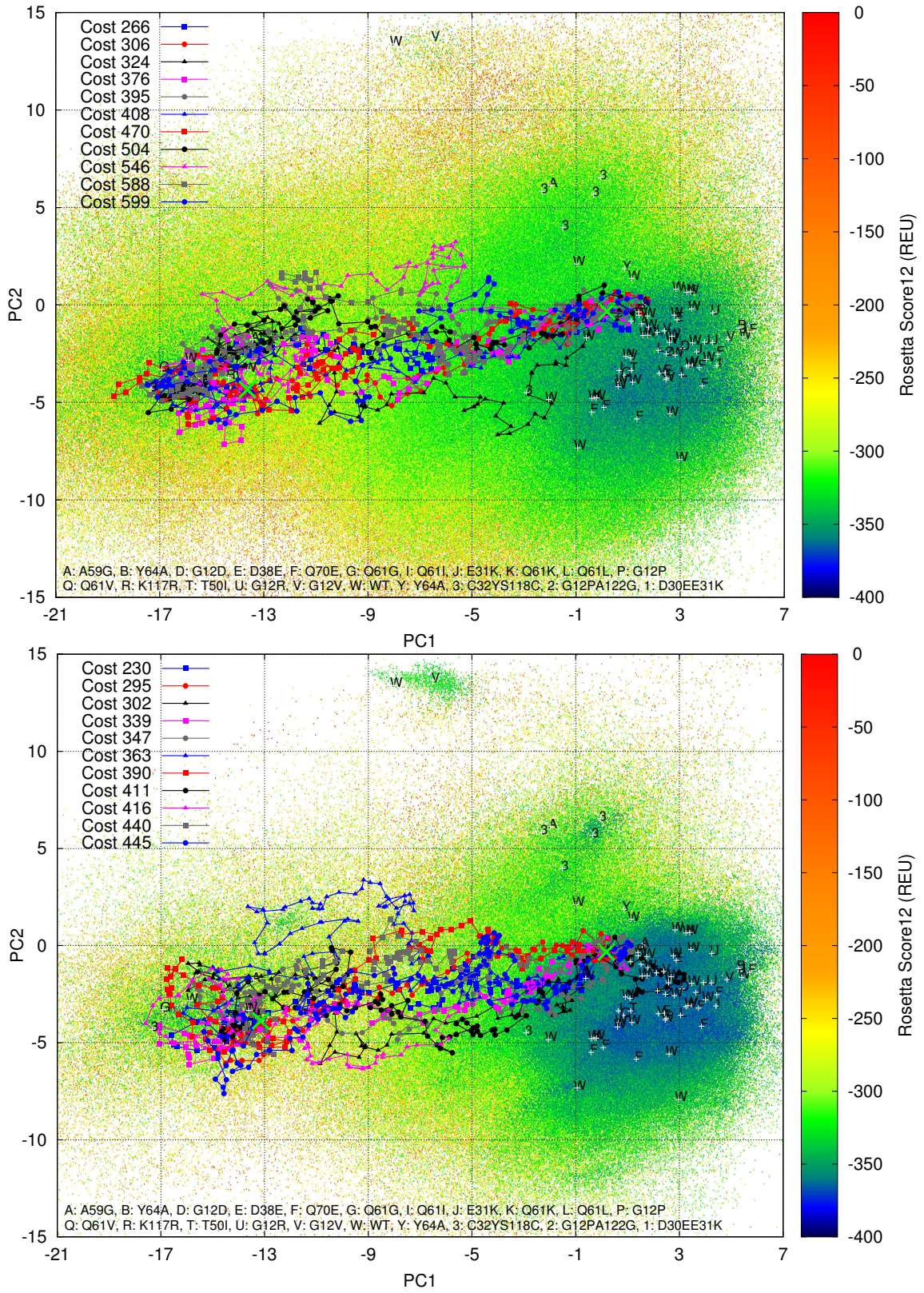

Figure 5: Visualization in 2D of the map and low-cost paths computed for H-Ras WT, using (score-based) color-coded projections on the top two PCs. The top panel shows the map and paths after the first iteration. The bottom panel shows the new map after the second iteration (path-guided EA), and the new low-cost paths obtained after querying the new map. The locations of projections of wet-laboratory structures of H-Ras are indicated by annotations of whether the structures correspond to the WT or variants.

towards regions likely to contain basins but also the regions that connect basins. We provide a proof-of-principle demonstration of the ability of this approach to improve the quality of the paths produced from EA-built maps of energy landscapes. The results presented here on the proteins, SOD1 and H-Ras, demonstrate the promise of the presented approach. In the preliminary investigation conducted in this paper, the exploration in the EA is guided via an initial population mechanism. Further research will consider other avenues as well.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] 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.

[2] 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.

[3] P. G. Debenedetti and F. H. Stillinger. Supercooled liquids and the glass transition. *Nature*, 410(6825):259–267, 2001.

[4] J. P. K. Doye. The network topology of a potential energy landscape: A static scale-free network. *Phys Rev Lett*, 88(23):238701, 2002.

[5] 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.

[6] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput Biol*, 2016. in press.

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

[8] E. Sapin, D. B. Carr, K. A. De Jong, and A. Shehu. Computing energy landscape maps and structural excursions of proteins. *BMC Genomics*, 2016. under review.

[9] E. Sapin, K. A. De Jong, and A. Shehu. Evolutionary search strategies for efficient sample-based representations of multiple-basin protein energy landscapes. In *IEEE Intl Conf Bioinf and Biomed (BIBM)*, 2015.

[10] E. Sapin, K. A. De Jong, and A. Shehu. Randomized search for uncovering basins in protein energy landscapes. *IEEE/ACM Trans Bioinf and Comp Biol*, 2016. under review.

[11] L. C. Smeeton, J. D. Farrell, M. T. Oakley, D. J. Wales, and R. L. Johnston. Structures and energy landscapes of hydrated sulfate clusters. *J Chem Theory Comput*, 11(5):2377âĂŞ2384, 2015.

[12] D. J. Wales, M. A. Miller, and T. R. Walsh. Archetypal energy landscapes. *Nature*, 394(6695):758–760, 1998.