Evolutionary Search Strategies for Efficient Sample-based Representations of Multiple-basin Protein Energy Landscapes

Emmanuel Sapin[1], Kenneth A De Jong[1,2], and Amarda Shehu[1,3,4*]
[1]*Department of Computer Science,* [2]*Krasnow Institute of Neuroscience,*
[3]*Department of Bioengineering,* [4]*School of Systems Biology,*
George Mason University, Fairfax, VA, 22030, United States
esapin@gmu.edu, kdejong@gmu.edu, amarda@gmu.edu
[*]*Corresponding Author*

***Abstract*—Protein function is the result of a complex yet precise relationship between protein structure and dynamics. The ability of a protein to assume different structural states is key to biomolecular recognition and function modulation. Protein modeling research is driven by the need to complement experimental techniques in obtaining a comprehensive and detailed characterization of protein equilibrium dynamics. This is a non-trivial task, as it requires mapping the structure space (and underlying energy landscape) available to a protein under physiological conditions. Existing algorithms invariably adopt a stochastic optimization approach to explore the non-linear and multimodal protein energy landscapes. At the present, such algorithms suffer from limited sampling, particularly in high-dimensional and non-linear variable spaces rich in local minima. In this paper, we equip a recently published evolutionary algorithm with novel evolutionary search strategies to enhance the sampling capability for mapping multi-basin protein energy landscapes. We investigate initialization strategies to delay premature convergence and techniques to maintain and update on-the-fly a sample-based representation that serves as a map of the energy landscape. Applications on three proteins central to human disease show that the novel strategies are effective at locating basins in complex energy landscapes with a practical computational budget.**

I. Introduction

Experimental, theoretical, and computational studies have shown that protein function is the result of a complex yet precise relationship between protein structure and dynamics [1]–[3]. The ability of a protein to switch between different structural states at equilibrium is key to biomolecular recognition and function modulation [4].

This dynamic property of proteins as macromolecules in perpetual motion [2] makes it hard to catch them in the act in specific states in the wet laboratory. While advances have been made by single-molecule techniques [5]–[8], wet-laboratory techniques are in principle limited by the disparate temporal scales in protein dynamics. For instance, dwell times at successive structural states accessed by a protein at equilibrium may be too short to be detected in the wet laboratory.

While the design of computational methods to characterize equilibrium protein structure and dynamics is driven by the need to aid wet-laboratory methods, neither computational nor wet-laboratory methods can on their own span all the spatial and temporal scales in protein dynamics [9]. In principle, a complete and detailed account of protein equilibrium dynamics requires a comprehensive characterization of both the protein structure space and the underlying energy landscape that governs accessibility of structures and transitions between them at equilibrium.

Protein structure spaces are vast and high-dimensional, and their underlying energy landscapes are nonlinear and multimodal. Many algorithms have been proposed over the years to characterize protein structure spaces and energy landscapes [10]. The majority suffer from the well-known issue of limited sampling. Faced with high-dimensional, complex variable spaces, existing algorithms can only obtain sample-based representations of protein energy landscapes; invariably, they approach this task under the umbrella of stochastic optimization, as the thermodynamically-stable and semi-stable structural states at equilibrium correspond to basins in the landscape [11].

Of particular interest are algorithms that forego simulation of protein dynamics (not based on the Molecular Dynamics – MD – framework). Algorithms that navigate the structure space of a protein by adapting the classic Monte Carlo (MC) framework have been shown to have higher exploration capability than MD-based ones [12]. A rather under-explored subclass of algorithms that approach stochastic optimization under the umbrella of evolutionary computation are evolutionary algorithms (EAs), which have been shown to be better at escaping local minima in protein energy landscapes than MC-based algorithms [13].

While EAs have shown promise in the context of *de novo* protein structure prediction [14], they remain under-explored for the more general problem of mapping multiple-basin protein energy landscapes. Given the wealth of different algorithmic knobs that can be tuned in an EA to obtain different behaviors [15], this paper investigates and injects novel evolutionary search strategies in an EA originally proposed in [16]; the result is a powerful algorithm capable of mapping multiple-basin protein energy landscapes within

a practical computational budget.

The evolutionary search strategies proposed here address the well-known issue of premature convergence of EAs in multimodal landscapes [15]. While the original EA proposed in [16] shows promise at locating more than one basin, the algorithm is formulated for the classic problem of finding the global minimum (deepest basin) and relies on a decentralized selection mechanism to delay convergence.

In this paper, we address two main issues that arise when the focus switches from locating the global minimum to obtaining a map of the different basins in protein's multiple-basin energy landscape. First, we investigate the role of *initialization* and propose a novel mechanism to obtain a diverse, yet energetically-relevant (fit) initial population. Second, we investigate techniques to maintain and update on-the-fly a sample-based representation of the energy landscape through the concept of a *hall of fame*. This is a list of all the best protein structures that are discovered by the EA. This list is updated during the evolutionary search process using energetic and geometric constrains. This results in a sparse representation (effectively, a map) of the energy landscape. Finally, we improve the computational time demands of the resulting EA via parallelization so that the algorithm can be applied to longer protein chains within a practical computational budget.

The EA presented in this paper does not operate in *de novo* settings (using only the sequence of a protein). Existing *de novo* algorithms do not address the problem of mapping a protein's energy landscape but restrict their focus to the structure prediction setting, where the goal is to rapidly find the global minimum; their best performance, even in this setting, is typically on proteins no longer than 150 amino acids [17]. Approaching the problem of mapping the energy landscape of a protein using only its amino-acid sequence would be an exceptionally computationally-demanding task, unless *a priori* information is available regarding where in the variable space the algorithm should focus.

To obtain such information, the EA described here leverages known, wet-laboratory structures deposited for a protein in the Protein Data Bank [18] and, more importantly, uses a principal component analysis (PCA) of these structures to define a reduced space of collective variables from which to draw samples during exploration. We have demonstrated in recent work that such PCA transformations open the way to the application and adaptation of various existing evolutionary search techniques for protein structure modeling [19]. In this context, this paper proposes various novel techniques that exploit this transformation to map multiple-basin protein energy landscapes in a computationally efficient manner while preventing the search from getting stuck in the regions populated by the experimental structures.

The main EA template for readers and the algorithm on which this work builds is summarized in section II, focusing on the novel algorithmic contributions made in this paper. In section III analysis of application of this approach on three proteins central to human disease shows that the proposed novel strategies are effective at mapping and locating basins in complex energy landscapes. The paper concludes with a summary and future prospects in section IV.

II. Methods

Given a protein, a map of its energy landscape is constructed during the course of the algorithm. The EA template used here has been recently proposed and applied to locate the global minimum of a protein energy landscape [16]. In this paper, this template has been equipped with new algorithmic components to extend its applicability from a strict optimization setting to the more challenging, mapping setting. Before describing the new components, we summarize the underlying EA template for readers.

A. EA in a Reduced Space of Collective Variables

The fundamental assumption and driving hypothesis here is that known, wet-laboratory structures of different sequence variants of a protein represent stable and semi-stable structural states of the wildtype sequence. This hypothesis is indeed based on the principle of conformational selection [20], under which perturbations such as sequence mutations to a protein do not change its structure space but rather the probabilities with which a given sequence is expected to populate its various structural states; in other words, even a structure detected for a variant is expected to be assumed by the wildtype (and vice-versa) but possibly with a different probability at equilibrium.

Known structures are collected and exploited in order to learn the structure space. The structures are fed to a dimensionality reduction technique in order to detect principal modes of motion that can be employed as collective variables to define a low-dimensional variable space for exploration. The process is detailed in prior work [16], [21]. In summary, a linear dimensionality reduction technique, PCA [22], extracts collective variables known as principal components (PCs). These variables are new axes, obtained from rotation of the original space so as to maximize variance along the axes. Ordering the axes by the variance of the data when projected onto them allows extracting a subset m that is typically much less than the original dimensionality of the data if PCA is indeed effective.

We have shown elsewhere that for many proteins with distinct basins in their landscapes, this is indeed the case; with the top two PCs one captures more than 50% of the variance (which means they can be employed for data visualization) and anywhere between 10-25 PCs allow capturing more than 90% of the variance [16]. The variation operator is described in more detail in prior work [21]. The latter is a reduction by more than ten-fold, as the original structures are of proteins with more than 100 amino acids; stripping them down to their CAs prior to PCA exposes more than

300 Cartesian coordinates on which PCA operates to reveal few PCs that still capture more than 90% of the variance.

The variance-ordered PCs are used as variables through which to represent a structure; we refer to an instantiation over this representation as a *conformation*; a conformation is a point in an m-dimensional space whose axes are the top PCs revealed by the PCA on a given protein. The EA operates in this space, evolving a population of conformations over generations towards low-energy conformations. Variation is introduced in each population through the following variation operator: a vector is defined in the PC space, with elements being magnitudes of movement along each of the m PCs. The magnitude of the movement along the top PC (that captures the most variance) is sampled uniformly at random in the segment [-s, s], where s is a user parameter. The magnitudes of the movements along the other PCs respect their variance relative to the variance captured by the first/top PC.

Once the variation operator yields N offspring from a population of N parent conformations, the offspring are subjected to a local improvement operator that maps them to local minima of an all-atom energy surface. Since the offspring are not yet structures, they are subjected to a multiscale procedure, which first constructs a backbone given CA atoms (these are trivially obtained from a conformation), and then employs the side-chain packing and minimization routine in the *relax* protocol in Rosetta to obtain a local minimum structure for an offspring. The resulting structure is projected back onto the PCs to replace the offspring (its coordinates may have slightly changed during minimization), and the all-atom Rosetta score (score12) is recorded and associated with the offspring conformation. The replacement of the offspring with the result of the local improvement operator makes this EA a Lamarckian EA.

A selection operator selects N new conformations off the combined pool of N parents and N offspring conformations. In [16], to delay premature convergence, an offspring only competes with parent conformations in a neighborhood of a user-defined size. The neighborhood captures the notion of structural similarity, so that offspring only replace structurally-similar parents if they lower energy. This results in preserving diversity longer in a population. Structural similarity is determined efficiently by embedding conformations in an explicit two-dimensional (2d) grid over the top two PCs. Cell width is also a user-defined parameter.

B. Initialization Mechanism

Proper initialization is key to exploration. In [16], the initial population is seeded with the collected wet-laboratory structures (projected to obtain conformations in the reduced space) and then filled with more conformations to reach a desired population size. The latter are obtained by subjecting the conformations corresponding to wet-laboratory structures to the variation and improvement operators.

We investigate two additional mechanisms to enhance the exploration capability in the mapping EA. The first does not employ the wet-laboratory structures but instead seeds the initial population with conformations sampled at random in the reduced space. The second employs the wet-laboratory structures but then reaches the desired population size with conformations sampled at random. To limit the bias to specific regions of the conformation space populated by the wet-laboratory structures, the size of a population is significantly increased over prior work, from 500 to $2,000$. This size allows enriching the initial population with more conformations so as to increase the exploration ability of the EA. In section III we analyze these three alternative initialization mechanisms, tracking the average energy and average structural diversity of a population over the generations. Structural diversity is measured through the Euclidean distance function in the space of PCs.

The inclusion of conformations sampled at random is key to diversify the initial population and allow the EA to explore new regions in the conformation space. In the same direction, the selection operator is modified here so that if an offspring conformation does not have any parent conformations in its neighborhood, it survives; in prior work, such an offspring competed with all parents. Another modification concerns the case when an offspring is killed; the parent of the offspring is then replaced with a conformation sampled at random. The intuition behind modification is as follows: the inability to generate a lower-energy offspring is indicative of the parent being at a local minimum, which is expected to be in the hall of fame. Replacing it with a conformation sampled at random prevents the algorithm from being stuck at the local minima already present in a population.

C. A Discrete, Sample-based Map of Protein Energy Landscapes: Hall of Fame

A large population is critical to capture a possibly large set of local minima in a rugged energy surfaces. However, combined with a large number of generations, this results in hundreds of thousands of conformations. Maintaining these conformations in memory is not an effective strategy for obtaining a discrete representation of the energy surface that can serve as a map. Here we employ the mechanism of hall of fame, which is an evolutionary strategy to equip an EA with memory. We employ the hall of fame as a dynamically-updated map of the energy surface. The algorithm to construct and update the hall of fame is shown in pseudo-code in Algorithm 1 where score(C) is the Rosetta *score12* of a conformation C and distance(c, C) is the Manhattan distance between two conformations c and C.

The hall of fame is a set of conformations that are updated at each generation. The conformations in the population at each generation are considered for inclusion in the hall of fame. The decision is made on a per-conformation basis. First, if the score of a conformation is non-negative, the

conformation is not added to the hall of fame (line 1 in Algorithm 1). Otherwise, the conformation is compared to structurally-similar conformations in the hall of fame (lines 4-8). Again, structural similarity is determined fast, measuring Manhattan distance in the space of top (m) PCs (analysis shows similar results are obtained if using the more expensive Euclidean distance). If the distance between two conformations is no higher than an *a priori* determined threshold (set to half of the minimum pairwise Manhattan distances over the wet-laboratory structures), the conformations are considered similar. If a conformation considered for inclusion in the hall of fame is not similar to any other conformation in the hall of fame, it is then added to the hall of fame, as it represents a new local minimum (line 10). Otherwise, it is compared to the conformations to which it is similar in terms of energies. If its energy is higher, it is not included in the hall of fame (line 8). Otherwise, it is included, and all similar conformations that have higher energy are removed from the hall of fame (line 6). In this way, the hall of fame is a set of distinct local minima, separated by at least the defined threshold in the space of PCs.

Algorithm 1 Hall of fame algorithm

```
1:  if score(C) < 0 then
2:    flag C_candidate_to_hof ← 1
3:    for all c in the hall of fame do
4:      if distance(c, C) < threshold_distance then
5:        if score(C) < score(c) then
6:          hof ← hof \ {c};
7:        else
8:          C_candidate_to_hof ← 0
9:    if C_candidate_to_hof = 1 then
10:     hof ← {C} ∪ hof;
```

D. Implementation Details

The algorithm is run for 500 generations, with a population size of 2000. It is implemented in C/C++ and run on a 16 core red hat Linux box with 3.2GhZ HT Xeon CPU and 8GB RAM. Offspring improvements are carried out in a parallel setting. This results in significant time savings, as shown in Fig. 1 for one of the longest proteins on which the algorithm is tested, and a total running time of 10-20 CPU hours for the three proteins used as test cases. Analysis of variance of various performance measurements employs 3 independent runs of the algorithm.

Parameter Values: The step size used in the variation operator is 1. The cell width in the 2d grid used in the selection operator is 1. The neighborhood model used is C1, effectively restricting structural comparisons only to parents that fall on same grid cell as the offspring under consideration. All parameter values in the *relax* protocol used in the local improvement operator are kept as recommended; to improve the likelihood that the bottom of a local minimum is reached, the number of iterations in the protocol is set to 5. These parameter values have been determined based on their effect in the convergence rate and structural diversity of populations over generations and time (data not shown).

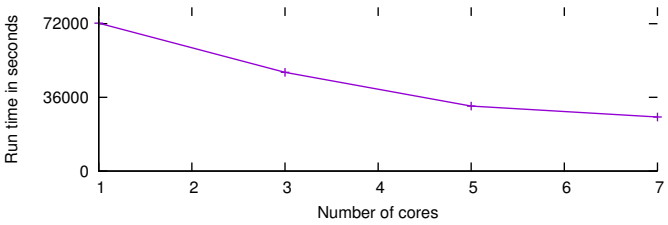

Figure 1: Run time as a function of the number of nodes used to distribute offspring improvements.

III. Results

A. Test Cases and Data Preparation

Performance is evaluated on 3 proteins of importance to human disease and with a significant number of known structures in the PDB [18]. These are H-Ras, SOD1, and HIV-I protease. H-Ras is a 166 amino-acid long protein that mediates signaling pathways that control cell proliferation, growth and development. H-Ras employs conformational switching between two distinct structural states to regulate its biological activity [23]. Sequence mutations are implicated in various human cancers [24]. SOD1 is a 150 amino-acid long protein whose mutations have been linked to familial Amyotrophic lateral sclerosis (ALS) [25]. HIV-I protease is a protein central to the replication of the HIV-I virus [26]. In its native form, the protein is a dimer, containing two identical polypeptide chains, each of 99 amino acids. Here we only consider one polypeptide chain, as the structure space accessed by it in isolation is also the one populated in the complexed form (when bound to an identical subunit), though with possibly different population probabilities, per the conformation selection principle [20].

Due to the implication of these proteins in various critical human diseases, ample structural data of their wildtype and mutated (variant) sequences exist in the PDB. Only X-ray structures are collected for H-Ras and HIV-I protease, whereas NMR structures are additionally included for SOD1 to enrich its dataset. The wildtype sequence of each of these proteins is obtained from UniProt [27]. Structures obtained from the PDB whose sequence changes by more than 3 amino acid from the wildtype sequence are discarded. Structures with missing internal amino acids are also discarded. Remaining structures are cropped at the termini, if necessary, so their lengths match the length of the wildtype. This protocol results in 86 wet-laboratory structures collected for H-Ras, 186 for SOD1, and 254 structures for HIV-I protease.

As described in section II, PCA is applied to each of these three datasets. A cumulative variance of 90% is reached at

$m = 10$, $m = 25$, and $m = 10$ PCs for H-Ras, SOD1, and HIV-I Protease, respectively. In the interest of space, the cumulative variance profiles are not shown here but have been presented in prior work on analysis of the PC spaces for each of these proteins [16].

B. Experimental Setup

The analysis presented here focuses on tracking the quality of the hall of fame and consists of three parts. First, a convergence rate analysis is conducted, which tracks the average fitness of the hall of fame over generations. Second, the structural diversity among individuals in the hall of fame is also tracked across generations, measured as the average Euclidean distance in the variable space (of PCs). The variance in these measurements across 3 independent runs of the algorithm is also presented. These two analyses are presented for H-Ras, as a representative system that highlights the role of the initialization mechanism. Third, the hall of fame obtained by the algorithm for each of the three proteins is visualized on two-dimensional energy maps and analyzed in detail, making comparisons where available with wet-laboratory knowledge on these proteins.

C. Convergence Rate Analysis

Fig. 2 tracks the average fitness of the hall of fame over generations. The three different initialization mechanisms are compared. Data are not drawn after generation 18, as convergence has been reached. The 99% confidence intervals are shown to visualize the variance in results due to the stochasticity of the algorithm (over 3 independent runs). Fig. 2 shows that seeding the initial population with random conformations converges more slowly than the other two settings. The reason for this is that it takes longer for the algorithm to reconstruct fit conformations, as the probability that a randomly-packed polypeptide chain will have low energy is very low (protein chains are highly constrained). The setting where the wet-laboratory structures are combined with conformations drawn at random reaches convergence faster but not as fast as the mechanism where the initial population consists of wet-laboratory structures and conformations obtained via variations of these structures.

D. Structural Diversity Analysis

Fig. 3 tracks the structural diversity in the hall of fame through the average Euclidean distance in the space of m PCs. The three initialization mechanisms are compared, and 99% confidence intervals are shown, calculated over 3 independent runs. Fig. 3 shows that seeding the initial population with random conformations results in higher diversity that is also preserved longer. The mechanism where conformations are drawn via variations of the wet-laboratory structures has the lowest diversity and loses it fast.

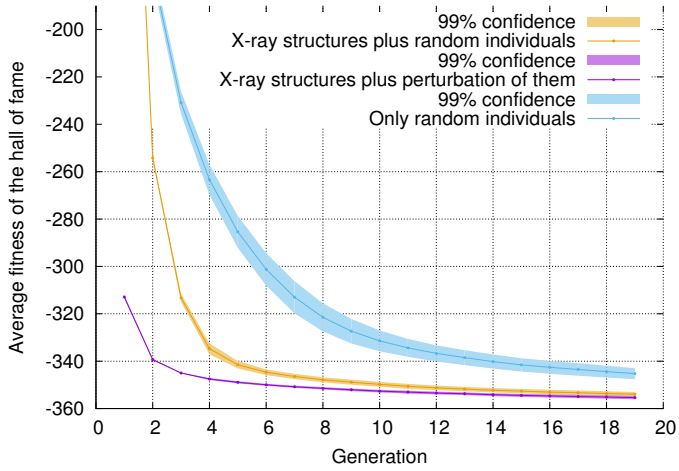

Figure 2: The average fitness of the hall of fame is tracked across generations, comparing the three different initialization mechanisms. The 99% confidence intervals over 3 independent runs are shown.

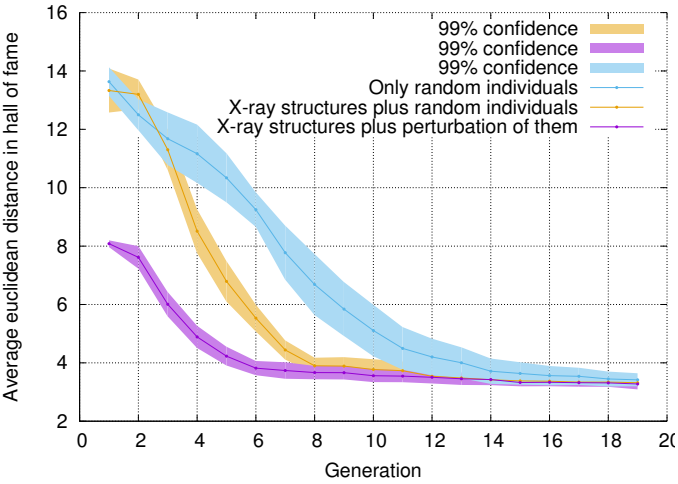

Figure 3: The average Euclidean distance among individuals in the hall of fame is tracked across generations, comparing the three different initialization mechanisms. The 99% confidence intervals over 3 independent runs are shown.

E. Projecting Halls of Fame to Visualize Energy Landscapes

Conformations in the hall of fame are projected onto the top two PCs for visualization and color-coded based on the Rosetta *score12* energy, effectively providing a low-dimensional map of the *score12* all-atom energy surface, the energy landscape. Fig. 4(a)-(b) shows this projected hall of fame for H-Ras. Fig. 4(a) is obtained from a run of the algorithm with a distance threshold of 2.10 (Manhattan distance) to determine inclusion in the hall of fame. The hall of fame in Fig. 4(b) is obtained with a smaller distance threshold of 0.31 but only calculated over the top two PCs as opposed to all m PCs used to draw conformations.

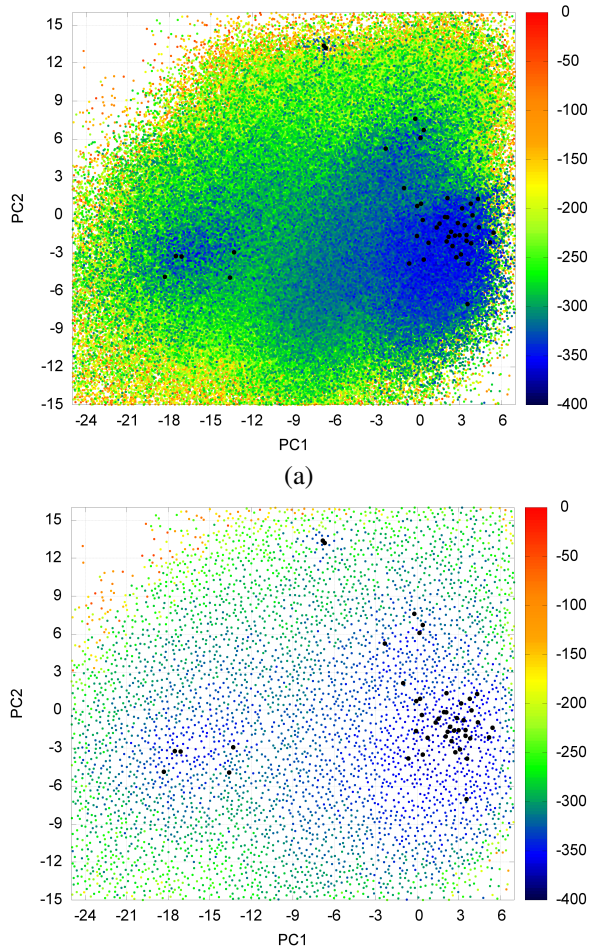

Figure 4: Projection of the H-Ras hall of fame onto the top two PCs, color-coded by *score12* energies (red-to-blue color scheme for high-to-low energies). Distance threshold for the hall of fame is 2.10 in (a) over the top ten PCs and 0.31 in (b) over the top two PCs. Projections of experimental structures are shown with black points.

Fig. 4(b) is a sparser representation of the H-Ras energy surface. This representation saves memory (from $634,691$ conformations when using a distance threshold of 2.10 over the top $m = 10$ PCs to $8,888$ conformations with a distance threshold of 0.31 over the top two PCs). The sparser representation is more amenable for analysis, including visualization of the basins. Such a visualization is provided in Fig. 5, which shows the energy landscape obtained for H-Ras. The visualization color-codes each grid cell by the average energy value over conformations that fall onto it and smooths the color of a grid cell with that of the neighbors. This technique is often employed to draw heat maps in visualization software and, in particular, to visualize protein energy landscapes.

Fig. 5 shows the energy landscape obtained for H-Ras and projections of the wet-laboratory structures. The wet-laboratory structures are organized in four subgroups, and three of these lie over the deepest (bluest) regions, which can be interpreted as three deep basins captured by the algorithm for H-Ras. The basin labeled *Off* corresponds to the known Off state of H-Ras (wet-laboratory structures that project to this basin are in the Off state). The basin labeled *On* corresponds to the known On state (wet-laboratory structures that project to this basin are in the On state). The separation of the energy landscape into these two main basins supports the known functional mechanism of H-Ras; this protein switches between the On and Off states to modulate its function in the cell. The energy landscape obtained by the algorithm additionally shows that the transition between these two states is likely to have a longer dwell time in the On state, given that the basin corresponding to the On state is deeper and broader than that corresponding to the Off state. The landscape also suggests that the transition overcomes an energy barrier, as conformations separating the On and Off basins have higher energies than those in the basins.

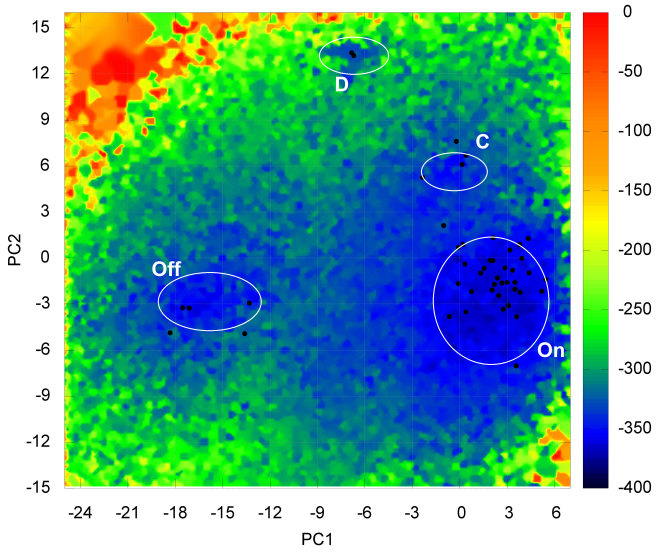

Figure 5: Energy landscape of H-Ras. Projections of experimental structures are shown with black points.

Two additional low-energy basins are visible in Fig. 5, labeled *C* and *D*. These two basins contain structures caught in the wet laboratory for variants of H-Ras. Comparing the location of basin *C* with the location of projections of wet-laboratory structures reveals that this basin maps the space populated by structures with PDB ids *2q21* and *1q21*. These structures are described are stable structures for an oncogenic variant of H-Ras in [28]. Basin *D* maps the space populated by wet-laboratory structures with PDB ids *1lf0* and *6q21(D)*. The structure with PDB id *1lf0* is the X-ray structure of yet another variant of H-Ras, which adopts a structure that is an intermediate between the On and Off states [29]. These two new low-energy basins capture

structural states that are possibly further stabilized in variants of H-Ras, affecting the actual transition path and rate of transition between the On and Off states and in turn the ability of H-Ras to regulate its functional mechanism. In light of the conformational selection mechanism, these new basins elucidated by the algorithm correspond to structural states populated with lower probabilities in the wildtype H-Ras but possibly higher ones in variants. The On-Off transition in H-Ras may spend significant time in any of these two new structural states in variants of H-Ras, giving a thermodynamic explanation for the misfunction caused by sequence mutations.

Fig. 6 shows the energy landscape obtained for SOD1 and projections of the wet-laboratory structures. The energy landscape exposes two well-separated, equally-deep basins, pointing towards the existence of two equally-stable structural states. The organization of wet-laboratory structures confirms this. It is interesting to note that the wet-laboratory structures include those of disease-involved variants of SOD1, and that these structures are in one of the two basins rather than high-energy regions. This points to the fact that sequence mutations in SOD1 possibly involve slight, in-basin structural changes, which reflects the current difficulty in the wet laboratory to understand how mutations in SOD1 are responsible for functional changes [30].

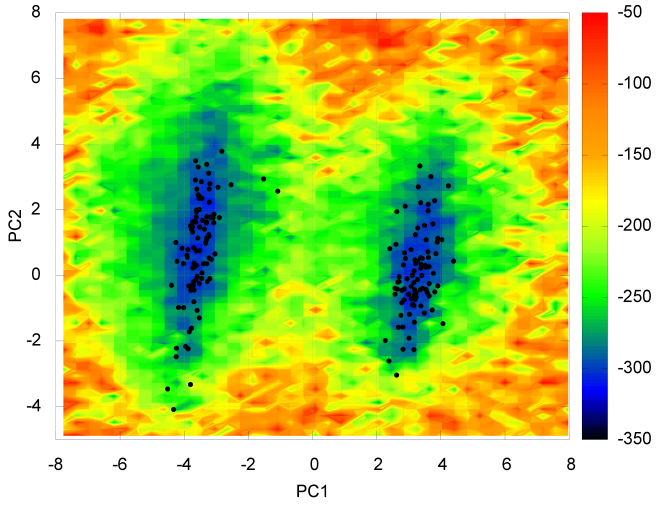

Figure 6: Energy landscape of SOD1. Projections of experimental structures are shown with the black points.

Fig. 7 shows the energy landscape obtained for HIV-I protease and projections of the wet-laboratory structures. The landscape contains a broad and shallow basin. In light of wet-laboratory knowledge that HIV-I protease has a fast mutation rate and yet forms stable monomers, the broad basin explains how HIV-I protease can undergo mutations and yet populate stable, functional structures. This is further affirmed by the fact that the majority of the wet-laboratory structures for HIV-I protease project on this broad basin, with very few off-basin structures. This result is in agreement with work in [26], which also shows a wide basin populated by many variants of HIV-I protease. The results obtained for HIV-I protease can be used to anticipate drug resistance; projecting new found structures of HIV-I protease on the obtained landscape may reveal similarity with already-characterized structures and provide information on stability, drug resistance, and possible, effective inhibitors.

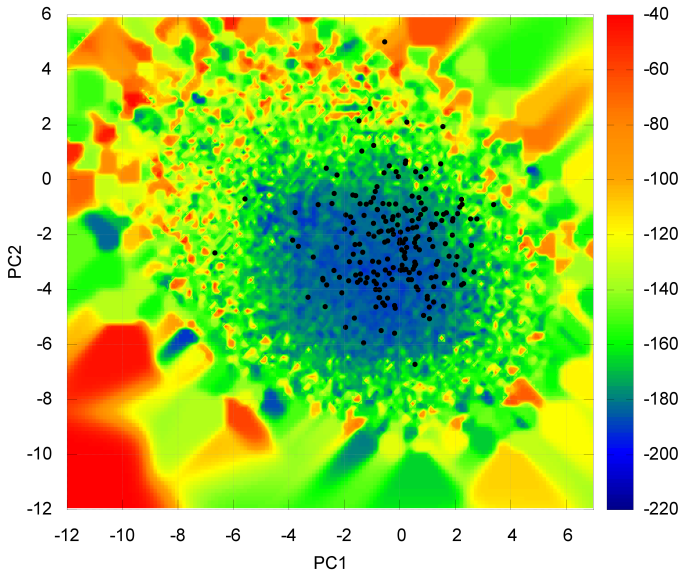

Figure 7: Energy landscape of HIV-I protease. Projections of experimental structures are shown with the black points.

IV. Conclusion

This paper introduces evolutionary search strategies for mapping complex, multiple-basin energy landscapes. While obtaining a detailed characterization of the protein energy surface remains a task of outstanding challenge in silico, the work here exploits the wealth of structure data and novel evolutionary search strategies to enhance exploration.

The exploitation of structure data is a powerful and timely mechanism to map the structure space of a protein. While there is merit in pursuing algorithmic treatments that operate in a *de novo* setting, the availability of semi-stable and stable structures for many proteins allows formulating algorithms that can map energy landscapes within a reasonable computational budget. The algorithm we have presented in this paper is a first step in this direction.

The work presented here paves the way for several future directions. One direction involves going beyond the linear dimensionality reduction exploited here to define a reduced variable space. Non-linear techniques may prove useful in this regard, but they need to allow direct sampling in the variable space. Also, applications of the proposed EA on longer protein chains will be considered. Another direction concerns the calibration of predictions on the locations

and depths of mapped basins by employing various energy functions. This direction aims to increase the reliability of *in-silico* predictions, as well as allow further applications of interest, where comparisons between landscapes mapped for different variants of a protein may provide the basis for understanding functional changes and disease.

ACKNOWLEDGMENT

This work is supported in part by NSF CCF No. 1421001, NSF IIS CAREER Award No. 1144106, and the Thomas F. and Kate Miller Jeffress Memorial Trust Award.

REFERENCES

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The energy landscapes and motion on proteins," *Science*, vol. 254, no. 5038, pp. 1598–1603, 1991.

[2] K. Jenzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, pp. 964–972, 2007.

[3] J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," *PLoS Comp Biol*, vol. 5, no. 8, p. e1000480, 2009.

[4] D. D. Boehr, R. Nussinov, and P. E. Wright, "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chem Biol*, vol. 5, no. 11, pp. 789–96, 2009.

[5] G. Zhu, Ed., *NMR of proteins and small biomolecules*, ser. Topics in Current Chemistry. Springer-Verlag, 2012, vol. 326.

[6] R. B. Fenwick, H. van den Bedem, J. S. Fraser, and P. E. Wright, "Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR," *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 4, pp. E445–E454, 2014.

[7] W. J. Greenleaf, M. T. Woodside, and S. M. Block, "High-resolution, single-molecule measurements of biomolecular motion," *Annu Rev Biophys Biomol Struct*, vol. 36, pp. 171–190, 2007.

[8] J. Hohlbein, T. D. Craggs, and T. Cordes, "Alternating-laser excitation: single-molecule FRET and beyond," *Chem Soc Rev*, vol. 43, pp. 1156–1171, 2014.

[9] D. Russel, K. Lasker, J. Phillips, D. Schneidman-Duhovny, J. A. Veláquez-Muriel, and A. Sali, "The structural dynamics of macromolecular processes," *Curr Opin Cell Biol*, vol. 21, no. 1, pp. 97–108, 2009.

[10] A. Shehu, "Probabilistic search and optimization for protein energy landscapes," in *Handbook of Computational Molecular Biology*, S. Aluru and A. Singh, Eds. Chapman & Hall/CRC Computer & Information Science Series, 2013.

[11] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations," *Proc. Natl. Acad. Sci. USA*, vol. 103, no. 32, pp. 11 844–11 849, 2006.

[12] W. L. Jorgensen and J. Tirado-Rives, "Monte carlo vs molecular dynamics for conformational sampling," *J Phys Chem*, vol. 100, no. 34, pp. 14 508–14 513, 1996.

[13] R. Unger, "The genetic algorithm approach to protein structure prediction," *Structure and Bonding*, vol. 110, pp. 153–175, 2004.

[14] A. Shehu, "A review of evolutionary algorithms for computing functional conformations of protein molecules," in *Computer-Aided Drug Discovery*, ser. Methods in Pharmacology and Toxicology, W. Zhang, Ed. Springer Verlag, 2015.

[15] K. A. De Jong, *Evolutionary Computation: A Unified Approach*. Cambridge, MA: MIT Press, 2006.

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

[17] Y. Zhang, "Progress and challenges in protein structure prediction," *Curr. Opinion Struct. Biol.*, vol. 18, no. 3, pp. 342–348, 2008.

[18] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[19] R. Clausen, E. Sapin, K. A. De Jong, and A. Shehu, "Mapping multiple minima in protein energy landscapes with evolutionary algorithms," in *Genet Evol Comput Conf (GECCO)*. New York, NY, USA: ACM, July 2015, pp. 923–927.

[20] B. Ma, S. Kumar, C. Tsai, and R. Nussinov, "Folding funnels and binding mechanisms," *Protein Eng*, vol. 12, no. 9, pp. 713–720, 1999.

[21] R. Clausen and A. Shehu, "A Data-Driven Evolutionary Algorithm for Mapping Multibasin Protein Energy Landscapes," *J Comput Biol*, vol. 22, no. 9, pp. 844–860, 2015.

[22] D. G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, 1984.

[23] A. E. Karnoub and R. A. Weinberg, "Ras oncogenes: split personalities," *Nat Rev Mol Cell Biol*, vol. 9, no. 7, pp. 517–531, 2008.

[24] A. Fernández-Medarde and E. Santos, "Ras in cancer and developmental diseases," *Genes Cancer*, vol. 2, no. 3, pp. 344–358, 2011.

[25] R. A. Conwit, "Preventing familial ALS: a clinical trial may be feasible but is an efficacy trial warranted?" *J Neurol Sci*, vol. 251, no. 1-2, pp. 1–2, 2006.

[26] M. W. Chang, "Computational structure-based methods to anticipate hiv drug resistance evolution and accelerate inhibitor discovery," Ph.D. dissertation, University of California, San Diego, 2008.

[27] M. Magrane and the UniProt consortium, "UniProt knowledgebase: a hub of integrated protein data," *Database*, vol. 2011, no. bar009, pp. 1–13, 2011.

[28] L. A. Tong, A. M. de Vos, M. V. Milburn, and S. H. Kim, "Crystal structures at 2.2 A resolution of the catalytic domains of normal ras protein and an oncogenic mutant complexed with GDP," *J Mol Biol*, vol. 217, no. 3, pp. 503–516, 1991.

[29] B. E. Hall, D. Bar-Sagi, and N. Nassar, "The structural basis for the transition from Ras-GTP to Ras-GDP," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 19, pp. 12 138–12 142, 2002.

[30] A. Nordlund and M. Oliveberg, "SOD1-associated ALS: a promising system for elucidating the origin of protein-misfolding disease," *HFSP J*, vol. 2, no. 6, pp. 354–364, 2008.