Mapping Multiple Minima in Protein Energy Landscapes with Evolutionary Algorithms

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

Many proteins involved in human proteinopathies exhibit complex energy landscapes with multiple thermodynamically-stable and semi-stable structural states. Landscape reconstruction is crucial to understanding functional modulations, but one is confronted with the multiple minima problem. While traditionally the objective for evolutionary algorithms (EAs) is to find the global minimum, here we present work on an EA that maps the various minima in a protein's energy landscape. Specifically, we investigate the role of initialization of the initial population in the rate of convergence and solution diversity. Results are presented on two key proteins, H-Ras and SOD1, related to human cancers and familial Amyotrophic lateral sclerosis (ALS).

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.6.3 [**Computing Methodologies**]: Simulation and Modeling—*Applications*

Keywords

evolution strategies; protein modeling; energy landscape.

1. INTRODUCTION

Traditionally, the focus of EAs in protein modeling has been the *de novo* structure prediction problem (PSP), where the goal is to discover the global minimum of the protein conformation space provided only information on its amino-acid sequence[4, 5, 6]. While this is a difficult problem, a more challenging setting concerns proteins with multiple minima in their energy landscapes. These minima correspond to stable or semi-stable structural states exploited for functional modulation. Many proteinopathies involve such proteins.

*Corresponding Author

GECCO '15, July 11 - 15, 2015, Madrid, Spain

 ISBN 978-1-4503-3472-3/15/07...$15.00
DOI: http://dx.doi.org/10.1145/2739482.2768439

Recently, we have proposed EAs to address such proteins and uncover multiple minima. However, to maintain feasibility, the *de novo* setting has been discarded. Instead, known experimental structures of wildtype and variant sequences of such proteins are employed to either define the dimensionality, shape, and bounds of the underlying variable space for a CMA-ES algorithm [2] or in addition to seed the initial population of a population-based memetic cellular EA [3, 1].

Here we explore in further detail the relationship between convergence rate and solution diversity conferred by the EA presented in [1]. Specifically, we investigate the role of the initial population, pursuing three settings that vary the amount of a priori information employed from known experimental structures. Testing is carried out on two proteins. Our analysis suggests an optimal setting to compromise between fast convergence but high solution diversity, motivating further investigation of other algorithmic components for mapping complex protein energy landscapes with EAs.

2. METHODS

The algorithm we investigate here is a population-based, memetic, cellular EA. The EA operates on a variable space extracted from Principal Component Analysis of existing X-ray structures of wildtype and variant sequences of a protein under investigation. Only principal components (PCs) that cumulatively contribute 90% of the total variance are retained as variables. This affords significant reduction in the dimensionality of the variable space over other cartesian-based or angle-based representations of protein chains.

At each generation, all parents are selected in turn to produce offspring. The coordinates of a selected parent are modified by a randomly-drawn vector in the variable space. The vector contains components for each of the variables, the underlying PCs, and is scaled to preserve the ratios of the bounds of each of the PCs relative to the first one. The goal is to perturb more along PCs that are responsible for more of the data variation. Each offspring undergoes a local improvement. The CA coordinates are first recovered, backbone atoms are reconstructed, and then side chain atoms are packed and minimized, so an all-atom conformation is obtained for each offspring, together with its all-atom Rosetta (score12) energy. Competition is limited between an offspring and only its structurally-similar parents to preserve diversity. Similarity is determined fast and coarsely, on a neighborhood-structure in a 2d embedding along the top two PCs. Further details can be found in [3, 1].

Here we investigate the effect of the initial population on the relationship between convergence and solution diversity. While the first is important for computational efficiency, the latter is central to discovering the multitude of potentially-relevant local minima in the landscape. In published work, the initial population is seeded with X-ray structures. To reach a desired population size, more individuals are generated by subjecting the X-ray structures to the perturbation operator described above. The generated offspring are subjected to the local improvement operator before being added to the initial population. One issue with such an initialization is that it is unlikely that individuals will be generated in regions of the variable space that are void and not populated by any X-ray structures. This may affect the performance of the algorithm. While it may confer faster convergence, potentially important minima may be missed.

For these reasons, we investigate here two additional initialization mechanisms. In one, we randomize the initial population by generating individuals at random in the variable space. The individuals are subjected to the local improvement operator before being added to the initial population. in the other initialization mechanism, we combine X-ray structures with individuals generated at random and improved to reach the desired population size. We report below our analysis on the convergence, solution diversity, and quality of the hall of fame for each of the initialization mechanisms. The size of the initial population is 250.

Each of the settings are run for 100 generations, and population size is set at 250. The analysis is conducted on two proteins, H-Ras and SOD1. The PCA on H-Ras reveals 10 PCs that contribute cumulatively 90% of the variance among the X-ray structures; thus the variable space for H-Ras is 10-dimensional. For SOD1, the variable space is 20-dimensional (these proteins are 165 and 90 amino acids long, respectively). We distribute the execution of the local improvement operator across 10 3.2GhZ HT Xeon CPUs with 9GB RAM, which allows a running time of $4-5$ hours for each of the proteins here. Each of the experiments is performed 5 times.

3. RESULTS

3.1 Convergence Rate Analysis

We track the average fitness of the hall of fame over generations. Results are shown in the top panel of Figure 1. Data are not drawn after generation 18, as convergence has clearly been reached by all 3 settings. The comparison shows that seeding the initial population with random individuals converges more slowly than the other settings. It takes longer for the algorithm to reconstruct fit individuals, as the probability that a randomly-packed chain will have low fitness is very low for proteins. The setting where X-ray structures are combined with individuals drawn at random reaches convergence faster, but is outdone by the setting where individual drawing is biased by the X-ray structures.

3.2 Solution Diversity Analysis

We track the structural diversity of the hall of fame across generations in terms of the mean Euclidean distance in the variable space. Results are shown in the middle panel of Figure 1. Solution diversity is measured as the mean Euclidean distance in the variable space. Again, data are not drawn after generation 18. The comparison shows that seeding the initial population with random individuals has higher diversity and preserves such diversity longer. The setting where the drawing of individuals is biased by the X-ray structures has the lowest diversity and loses it fast.

3.3 Quality of Hall of Fame Analysis

Figure 2 plots the hall of fame individuals of all generations, projecting them on the top two PCs (in terms of variance/eigenvalue), and color-coding them by their all-atom energy (Rosetta *score12*). The three population initialization settings are compared.

For H-Ras, the hall of fame in the setting where the initial population is seeded with X-ray structures and their offspring has settled over three major basins in the landscape. The major one to the right is the active structural state of H-Ras, and the one to the left is the inactive structural state. The other two clearly separated grouping of individuals are novel structural states deemed Conf1 and Conf2 in prior modeling work on H-Ras [1]. The other two settings have higher-energy individuals in the hall of fame across generation, as expected. The active structural state/basin is present in both, but the other three states are sparsely populated due to the presence of high-energy individuals.

For SOD1, the hall of fame in the setting where the initial population is seeded with X-ray structures and their offspring has settled over two major basins in the landscape, as observed in [1]. The setting where the initial population is seeded with X-ray structures and individuals drawn at random largely recovers the same basins, due to the fact that there are many more X-ray structures that can be used as opposed to H-Ras. The setting where individuals are drawn at random for the initial population has recovered the same basins, but has not had enough time to settle to the bottoms of these basins and discriminate against similarly-favorable structures that connect the basins.

These results point to the fact that while the average fitness results shown above suggest convergence by generation 18, longer generations and perhaps larger populations are needed in order for the algorithm to settle into all basins when employing less and less a priori information in its construction of the initial population . This is clearly a direction of future work.

4. CONCLUSION

The work presented here expands upon a novel direction of research on evolutionary algorithms designed for exploring protein energy landscapes and uncovering multiple minima. While it has been previously shown that a structure-driven EA can be designed to uncover diverse minima, here we have investigated the role of the initialization mechanism on convergence and solution diversity. Our analysis suggests an optimal setting to compromise between fast convergence but high solution diversity, motivating further work and investigation of other algorithmic components for mapping complex protein energy landscapes with EAs. However, detailed investigation of the known minima/basins recovered by the algorithm highlights further work is needed to determine appropriate population size and number of generations.

5. ACKNOWLEDGMENTS

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.

6. REFERENCES

[1] R. Clausen, K. A. De Jong, and Shehu. A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes. *J Comput Biol*, 2015. in press.

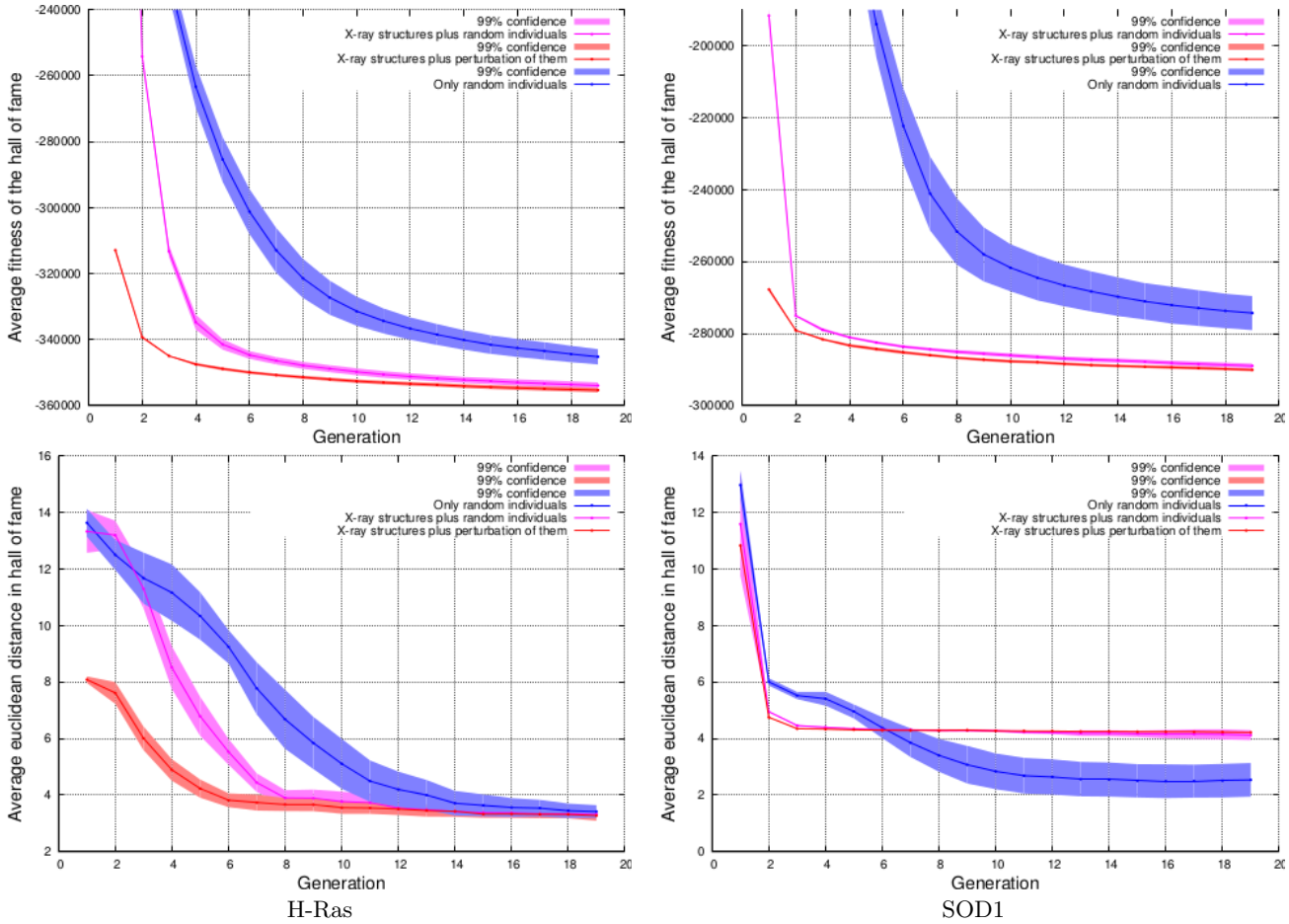
[2] R. Clausen, E. Sapin, K. A. De Jong, and A. Shehu. Evolution strategies for exploring protein energy landscapes. In *GECCO*. ACM, 2015.

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

[4] B. Olson, K. A. D. Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In *Conf on Genetic and Evolutionary Computation (GECCO)*, pages 287–294, New York, NY, 2013. ACM.

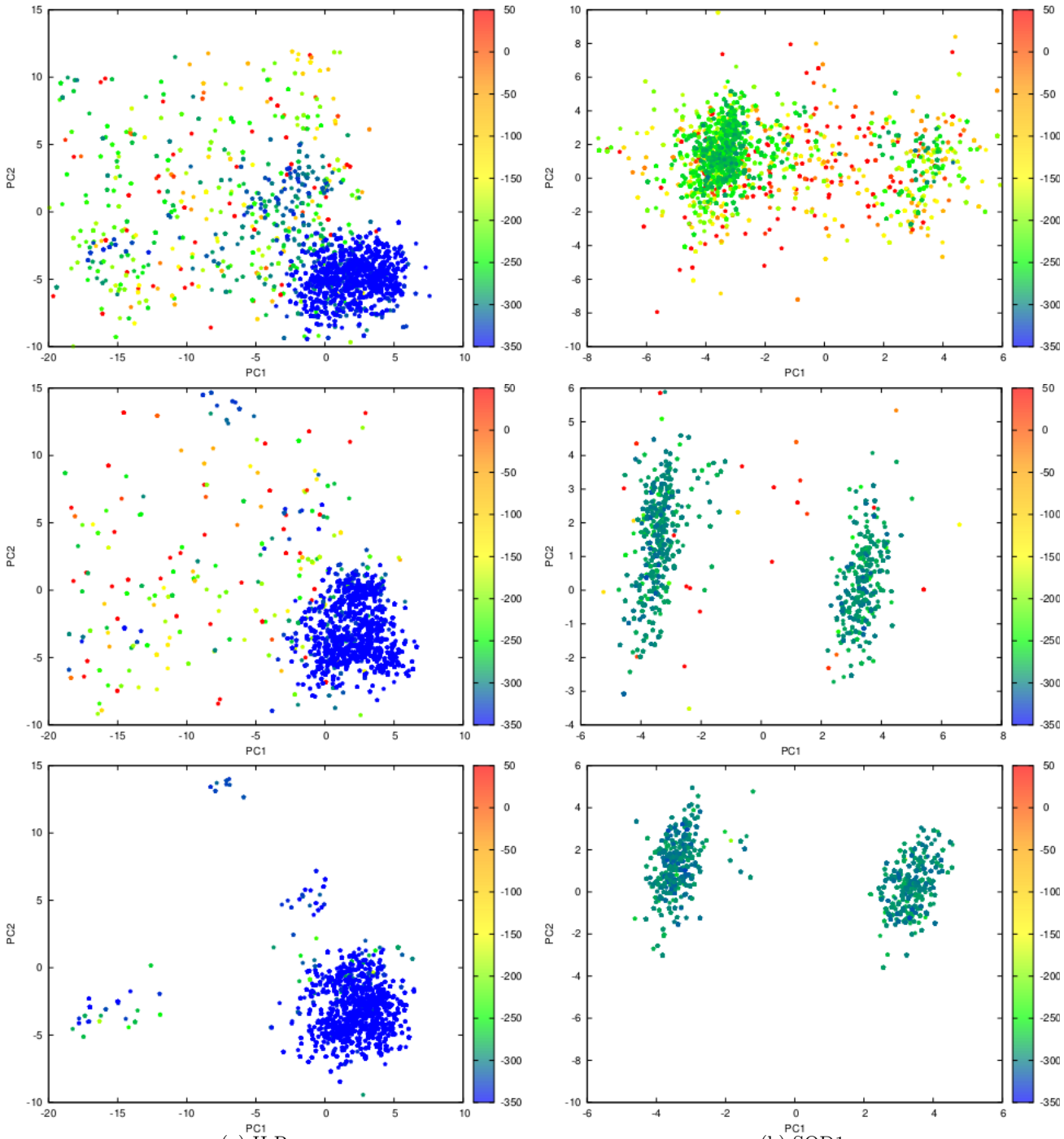
[5] B. Olson and A. Shehu. Multi-objective stochastic search for sampling local minima in the protein energy surface. In *ACM Conf on Bioinf and Comp Biol (BCB)*, pages 430–439, Washington, D. C., September 2013.

[6] J. Santos, P. Villot, and M. Dieguez. Emergent protein folding modeled with evolved neural cellular automata using the 3d hp model. *J of Comp Biol*, 21(11):823–845, 2014.



H-Ras SOD1

Figure 1: Results are shown in (a) for H-Ras and (b) for SOD1. The average fitness of the hall of fame is tracked across generations in the top panel. The average pairwise Euclidean distance in the hall of fame is tracked across generations in the middle panel.



(a) H-Ras (b) SOD1

Figure 2: Results are shown in (a) for H-Ras and (b) for SOD1. All individuals in the hall of fame at each generation are projected onto the top two PCs and color-coded by their all-atom Rosetta *score12* energies. The top panel shows the setting where the initial population is drawn at random; the middle panel shows the setting where X-ray structures are combined with individuals drawn at random; the bottom panel shows the setting where the perturbation operator is additionally used over the X-ray structures to initialize the population.