Evolutionary Search For Paths on Protein Energy Landscapes

Emmanuel Sapin
Dept of Computer Science
George Mason University
4400 University Dr.
Fairfax, Virginia 22030
esapin@gmu.edu

Kenneth De Jong
Dept of Computer Science
George Mason University
4400 University Dr.
Fairfax, Virginia 22030
kdejong@gmu.edu

Amarda Shehu*
Dept of Computer Science
George Mason University
4400 University Dr.
Fairfax, Virginia 22030
amarda@gmu.edu

ABSTRACT

Proteins are in perpetual motion, switching between structures to regulate interactions with molecular partners. These motions correspond to hops in an energy landscape that organizes the structures available to a protein by their potential energies. Here we introduce an evolutionary algorithm (EA) that computes structural excursions of a protein without the need to reconstruct its energy landscape a priori. The preliminary results are promising and suggest further directions of research.

CCS CONCEPTS

•**Applied computing** – **Molecular structural biology;**

KEYWORDS

evolutionary search, fitness landscape, structural excursions, protein modeling

ACM Reference format:
Emmanuel Sapin, Kenneth De Jong, and Amarda Shehu. 2017. Evolutionary Search For Paths on Protein Energy Landscapes. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3075599

1 INTRODUCTION

Protein modeling research aims to uncover the functionally-relevant structural excursions that a protein employs to tune its biological function. One direction of *in-silico* work involves first reconstructing energy landscapes (often with powerful memetic EAs [4, 5]) and then exploiting graph-based representations of such landscapes to answer path queries corresponding to structural excursions of interest. This direction has revealed key insights on many proteins [3] but has a large computational footprint due to the need to construct comprehensive and detailed representations of energy landscapes that are vast and high-dimensional [1, 2].

Here we explore a different direction. We propose an EA that computes paths without first reconstructing an energy landscape. The EA evolves a population of paths directly, exploits experimentally-known structures in its initialization, and uses novel selection and crossover operators. This path-evolving EA reproduces known structural excursions of a protein and its disease variants with a much smaller computational budget. The obtained paths are realistic and have fine granularity. The preliminary results suggest path-evolving EAs warrant further attention. Section 2 relates the salient ingredients of our EA. Section 3 showcases some preliminary results. The paper concludes in Section 4.

*Corresponding Author

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3075599

2 METHODS

Key building blocks in the path-evolving EA have been developed and analyzed in prior work [3–5]. They include exploiting known structures of a protein (of healthy and diseased sequence variants) to extract a lower-dimensional variable space for exploration. Unlike prior work, where an EA evolves individuals in this variable space, starting from a collection of individuals (points) corresponding to known structures, the new EA evolves paths utilizing only two given (experimentally-known) structures, which initialize the start and end points of paths. A path individual is represented as a (start-to-goal directed) list of points in the variable space. Initially, *n_points* are obtained by linear interpolation between the given start and end points. Each obtained point undergoes a transformation, which effectively converts it to an all-atom protein structure corresponding to a local minimum in the all-atom Rosetta energy landscape. The transformation utilizes stochastic optimization, so repeating it N times yields the initial population of N paths.

Once the initial population is defined, successive generations evolve as follows. First, new candidate path vertices are generated from the existing population of paths. For every two consecutive points in a path, a variation operator yields a new mid-point, which is then converted to a (local minimum) all-atom structure. All of these points are inserted into a nearest-neighbor graph (nngraph) which connects a point to others within a pre-specified radius r measured via the Euclidean distance in the variable space.

Dijkstra's algorithm is invoked on the nngraph to obtain the first lowest-cost path connecting the given pair of start and goal points.The algorithm is invoked N times in order to obtain the N lowest-cost paths with which to initialize the next generation. In order for such paths to be non-redundant, once a path $i \in [1, N]$ is identified, its internal points are removed from the nngraph, so that the next application of Dijkstra's algorithm to find the next lowest-cost path $i + 1$ operates on the induced subgraph.

Prior to each subsequent application, r is decreased from the value that resulted in path i and continues to decrease in a proportional regime until a candidate for path $i + 1$ cannot be found (the graph becomes disconnected). In that case, r is rolled back to the previous successful value and the resulting lowest-cost path is the one initializing individual $i + 1$ in the population. This selection

mechanism allows the algorithm to evolve both low-cost and high-resolution paths that better follow the actual energy landscape.

The EA operates under a fixed computational budget, tallying up the number of energy evaluations employed in the transformations from points to structures. The fitness of a path is its energetic cost, which sums up the energy increase between structures corresponding to consecutive points.

3 RESULTS

The performance of the path-evolving EA is showcased here on H-Ras, a protein central to cell growth and various human cancers. The EA is run to obtain paths connecting two known structures corresponding to two different functional states of H-Ras. The computational budget was fixed to 100, 000 fitness evaluations, 10 times less than that used in prior work that reconstructs energy landscapes with an EA and then uses graph-based representations to answer path queries [3]. The $N = 15$ paths obtained in the final generation with the path-evolving EA with this budget are rendered in Fig. 1.

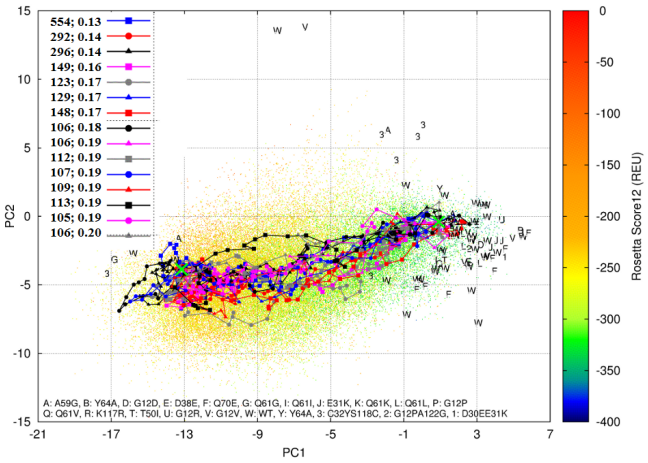

Figure 1: The 15 lowest-cost paths of the final generation obtained by the path-evolving EA are shown here (edges connect consecutive structures in a path). A structure is shown as a dot using as coordinates the values of the first two variables (prior work selects variables via principal component analysis). The other dots, which are color-coded, are those of structures generated during the execution of the algorithm. The blue-to-red color-coding scheme tracks low-to-high Rosetta all-atom energy values. Text annotations indicate experimentally-known structures, with WT referring to the healthy form and others pathogenic forms. The legend lists path costs and resolutions.

Table 1 juxtaposes the 10 lowest-cost paths obtained by the path-evolving EA with the 10 lowest-cost paths obtained by the EA in prior work [3]. The comparison is limited to 10 paths, as the resolution of the paths deteriorates afterwards. While all paths obtained by post-analysis of the reconstructed map in prior work have the same resolution, the ones obtained by the path-evolving EA have varying resolution. Table 1 orders the paths obtained by the path-evolving EA from high to low resolution (resolutions are rounded to at most two decimal places). The juxtaposition shows that the path-evolving EA is able to obtain very high-resolution (0.133Å, rounded to 0.13 in Table 1) paths with much less computational budget (and consequently fewer computed structures). Path costs at high resolutions typically increase due to the high ruggedness of protein energy landscapes. The best path found by the path-evolving EA has a cost of 292 Rosetta Energy Units (REUs) and a resolution of 0.143Å (rounded to 0.14 in Table 1). This is comparable to the best path produced by the EA in [3], which has a cost of 266 REUs and a resolution of 0.145Å (rounded to 0.15 in Table 1).

Table 1: Top ten paths obtained by each algorithm.

Path-evolving EA proposed here

Cost	554	292	296	149	123	129	148	106	106	112
Res	0.13	0.14	0.14	0.16	0.17	0.17	0.17	0.18	0.19	0.19

EA proposed in [3]

Cost	588	546	504	470	408	395	376	324	306	266
Res	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15

4 CONCLUSION

The initial evaluation of the path-evolving EA suggests that it represents an improvement over state-of-the-art methods for modeling protein structural excursions [3]. In addition, the proposed EA is able to further improve the quality of its paths when afforded more fitness evaluations (data not shown here). We intend to pursue the proposed path-evolving EA further as part of our goal of modeling protein structural dynamics with reasonable computational budgets. The emphasis on lower computational budgets is due to the foreseen applicability of this algorithm to obtain and then compare the structural dynamics of various forms of a protein. The latter would allow understanding the impact of mutation-altered dynamics on protein function. It is also worth noting that the techniques presented here are more general than the specific domain of protein modeling and thus potentially useful for a broad range of problems focused on landscape mapping and analysis.

5 ACKNOWLEDGMENTS

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

REFERENCES

[1] H. Frauenfelder, P. W. Fenimore, and R. D. Young. 2007. Protein Dynamics and Function: Insights from the Energy Landscape and Solvent Slaving. *IUBMB Life* 59, 8-9 (2007), 506–512.

[2] 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.

[3] 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.

[4] E. Sapin, K. A. De Jong, and A. Shehu. 2016. A Novel EA-based Memetic Approach for Efficiently Mapping Complex Fitness Landscapes. In *Conf on Genetic and Evolutionary Computation (GECCO)*. ACM, 85–92.

[5] E. Sapin, K. A. De Jong, and A. Shehu. 2017. From Optimization to Mapping: An Evolutionary Algorithm for Protein Energy Landscapes. *IEEE/ACM Trans Comput Biol & Bioinform* (2017). in press.