Memetic, Multi-Objective, Off-Lattice, and Multiscale Evolutionary Algorithms for De novo and Guided Protein Structure Modeling

Amarda Shehu[1,*] and Kenneth A. De Jong[1]
[1]Dept. of Computer Science, George Mason University, Fairfax VA 22003, USA
* Correspondence: amarda@gmu.edu

The goal of mapping out the biologically-active structural states of a protein is central to understanding the healthy and diseased cell, but it encompasses many challenging problems for an in-silico treatment. One of these is *de novo* protein structure prediction (PSP), where a single structure assumed to be representative of the active state (valid only in single-basin proteins) is sought for a given amino-acid sequence. Until recently, EAs for PSP were outperformed by Monte Carlo-based platforms, such as Rosetta and Quark.

Novel EAs for single-basin proteins: To address this issue, we have explored memetic EAs that employ fragment libraries, off-lattice backbone representations, and state-of-art energy functions [1]. We have investigated novel crossover operators [2] and multi-objective EAs [3, 4]. The resulting EAs have been demonstrated to be highly competitive with the current state-of-the-art in PSP. In particular, the multi-objective feature significantly enhances exploration and even improves model quality over Rosetta (see left panel of Fig. 1).

Novel EA for multi-basin proteins: While a *de novo* setting may be daunting for multi-basin proteins, known wet-lab structures can be used to define a lower-dimensional search space (via Principal Component Analysis) for an EA. Samples directly obtained in this space are subjected to an improvement operator that restores all-atom detail. A local selection operator over the inherent low-dimensional structurization delays convergence. The resulting EA (referred to as PCA-EA) maps the basins of wildtype and variants of the Ras oncogene (unpublished data summarized in right panel of Fig. 1). The EA shows that the variants lose access to various wildtype basins, providing a structural basis for their loss of function.

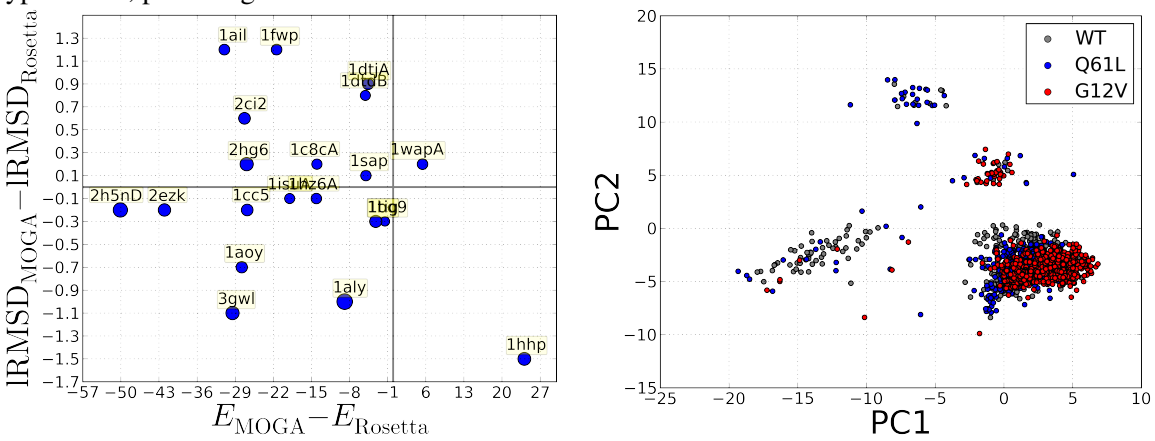

Figure 1: Left: Improvement in lowest sampled energy by MOGA [4] vs. Rosetta is plotted against improvement in lowest lRMSD to known structure (PDB ids shown, circle size indicates chain length). Right: Samples of final PCA-EA generation are projected on the top two principal components (PCs).

EAs are powerful algorithms for protein structure modeling when equipped with domain-specific insight. Our work in progress showcases their promise beyond the PSP setting.

References

[1] Saleh, S., Olson, B., Shehu, A.: A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. BMC Struct Biol **13** (2013) S4
[2] Olson, B., De Jong, K.A., Shehu, A.: Off-lattice protein structure prediction with homologous crossover. In: Conf on Genetic and Evolutionary Computation (GECCO), New York, NY, ACM (2013)
[3] Olson, B., Shehu, A.: Multi-objective stochastic search for sampling local minima in the protein energy surface. In: ACM Conf on Bioinf and Comp Biol (BCB), Washington, D. C. (2013) 430–439
[4] Olson, B., Shehu, A.: Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. In: Intl Conf on Bioinf and Comp Biol (BICoB), Las Vegas, NV (2014)