Multi-Objective Optimization Techniques for Conformational Sampling in Template-Free Protein Structure Prediction

Brian Olson¹,* Amarda Shehu²,³,⁴,‡
¹Lawrence Livermore National Lab, Livermore, CA
*Work was conducted when author was at George Mason University.
²Dept. of Computer Science, ³Dept. of Bioengineering, ⁴School of Systems Biology,
George Mason University, Fairfax, VA 22030
‡Correspondence: amarda@gmu.edu

Abstract

Template-free protein structure prediction continues to be a challenging problem in computational biology. State-of-the-art protocols are Monte Carlo-based and pay special emphasis on the set of moves and energy guidance. We report here on a complementary platform for decoy sampling that makes use of evolutionary search strategies. We propose that Evolutionary Algorithms (EAs) are effective platforms for the structure prediction problem as an optimization problem, outperforming the Monte Carlo-based sampling of protein conformational space. Moreover, these platforms allow casting the problem as a multi-objective optimization one to deal with the known imperfections in protein energy functions. We compare here different EAs to decoy sampling in the popular Rosetta protocol and show that multi-objective EAs have higher exploration capability and warrant further investigation.

1 Introduction

Characterizing protein structure is at the center of understanding the molecular basis of many proteinopathies [12]. Doing so in silico is ever more important due to the millions of sequences extracted from organismal genomes with no structural or functional characterization. Predicting the structure of a protein from its amino-acid sequence is known as template-free protein structure prediction (PSP) and is currently a challenging optimization problem [10].

PSP can be cast as an optimization problem, since the native structure of a protein (what is sought in PSP) is the thermodynamic equilibrium. A rich body of work has led to a Monte Carlo (MC)-based setting for this optimization problem, where the goal is to sample a large ensemble of low-energy (decoy) protein conformations at a coarse-grained representation (a conformation is an internal representation of protein structure, often in terms of only backbone dihedral angles). Structural clustering of this ensemble is used to identify the top-populated cluster as the one containing the native structure. Conformations in this cluster are subjected to intensive all-atom energetic refinements and then offered for the prediction.

How to best allocate computational resources to a stochastic optimization problem necessitates balancing global (exploration) and local (exploitation) search. Currently, in MC-based decoy sampling in top PSP protocols in CASP, global and local search are largely disjoint. Exploration is achieved in a multi-start setting, where essentially many intensive MC searches (that implement exploitation) are initiated from different conformations (possibly sampled at random). The balance between local and global search is limited to varying the temperature parameter in the Metropolis criterion in order to cross energy barriers for exploration. Great attention is paid to types of moves to improve exploitation.

Evolutionary Algorithms (EAs) allow addressing the exploration/exploitation balance in hard optimization problems. Exploration/global search in EAs is realized through a population of individuals (conformations here) that evolves over a number of generations towards fit individuals, as measured through some fitness function (an energy function in our setting). In each generation, the population is subjected to reproductive/variation operators, where a subset or all individuals are selected as parents to participate in asexual (through a mutation operator) or sexual (through crossover/recombination) reproduction to obtain offspring. Parents and offspring are evaluated according to the fitness function, and a subset is then selected through a selection mechanism to serve/survive as parents for the next generation. If a local search or improvement operator is additionally applied on an offspring to map it to a nearby minimum before evaluation, the resulting EA is a hybrid EA (HEA).

Work in [9] has shown that when employing the
molecular fragment replacement technique (popular and responsible for advances in accuracy and feasibility in decoy sampling [1]) in the mutation operator and the local search operator, the resulting HEA (using no recombination operator) compares favorable with the decoy sampling in Rosetta. Other work has investigated the role of the hybridization of global and local search in enhancing exploration capability [7,8,10]. Additional work in [6] has shown that EAs provide alternative venues to design and evaluate variation operators or moves on conformations. In particular, a genetic algorithm (GA) in [6] compares various recombination operators and proposes an effective one for highly-constrained systems, such as proteins.

The work in this manuscript makes the case that decoy sampling in PSP is suitable (and advantageous) to be cast as a multi-objective optimization problem, and that EAs provide a natural setting to do so and investigate various issues related to multi-objective optimization. In particular, in an EA-based setting, one can explore regroupings of the energy terms of an energy function into separate categories or objectives for optimization. Pursuing these directions is important to address inadequate exploration of the conformational space and known inaccuracies in protein energy functions, which are often cited as critical to why PSP remains challenging [5].

We propose multi-objective variants of hybrid and genetic EAs. Moreover, based on a proof-of-concept demonstration that Pareto-based measures such as Pareto rank are effective, resulting in what is referred to as MOEA in [9] (MOEA stands for HEA enhanced with multi-objective optimization, as measured through Pareto rank), we investigate various Pareto-based measures for how they allow balancing the exploration/exploitation tradeoff to further enhance exploration capability. Specifically, we investigate an additional measure, Pareto count, and refer to the three resulting multi-objective variants that incorporate Pareto rank and Pareto count (as detailed below) as MOEA-PC (HEA with Pareto rank and Pareto count), MOGA (GA with Pareto rank), MOGA-PC (GA with Pareto rank and Pareto count). We conduct a detailed comparative analysis of these EAs with one another and an in-house implementation of the decoy sampling algorithm in Rosetta.

2 Method

Briefly, we note that the EAs proposed and compared here make use of the Rosetta coarse-grained representation of a protein chain (where side chains are replaced with a CB atom and a centroid pseudo atom, and a conformation is represented as a vector of 3 angles $\phi, \psi, \omega$ – per amino acid) and the molecular fragment replacement technique. Fragment replacements of length 3, as in the Rosetta decoy sampling algorithm, are used to implement the mutation operator. In GA and MOGA, a novel recombination operator, homologous crossover, is additionally applied to combine two parents and obtain an offspring. The key idea of this operator is to carefully select recombination points based on structural comparison of parents (details are provided in [6]). The initial population is carefully constructed, using fragments of length 9 and 3 in order to obtain a good initialization (diverse but low-energy conformations) for the EAs. Further details can be found in [9].

Here we elaborate on Pareto measurements pursued in exploring alternative ways to inject multi-objective optimization in the proposed EAs. The Pareto rank of a conformation $C$ measures the number of conformations that dominate $C$ across all optimization objectives (in terms of strict $< \text{comparison}$). Pareto rank can be integrated in the evaluation mechanism as follows. Parent and offspring are ordered from low to high Pareto rank. Within this sorted order, those with the same Pareto rank are further ordered from low to high total (score4) energy. The lowest $p$ according to the sorted order are retained as parents for the next generation. What we refer to as MOGA here is a modification as described to obtain a multi-objective variant of the GA presented in [6].

It is not immediately clear how to group terms of a given energy function into optimization objectives or categories. The key rule is to keep the number of objectives small. Various studies, experimental and computational, on protein folding have shown that van der Waals interactions compete with hydrogen-bonding interactions. However, we further categorize hydrogen bonding interactions into two categories, short vs. long, based on analysis in [11] and our own experience that these two groups compete with one another. So, the Rosetta score4 function is organized in these three manageable objectives for optimization.

We explore an alternative multi-objective evaluation in this paper through the notion of Pareto count. The Pareto count of a conformation $C$ is the number of conformations that $C$ dominates (providing complementary information over Pareto rank). If a conformation has low Pareto rank and low Pareto count, this means that the conformation is probably an interesting local minimum, because it is close to being Pareto optimal but also resides in a region of the energy landscape not well explored (given its low Pareto count). Pareto count can be additionally added in the evalua-
tion mechanism by first sorting by Pareto rank (low to high), then by Pareto count (low to high), and then by total score (low to high). We investigate this modification in two proposed variants here, MOGA-PC (PC notes the addition of Pareto count in the evaluation) and MOEA-PC (MOEA was first proposed as a proof-of-concept that Pareto rank can be added to turn an HEA into a multi-objective variant in [9]).

Implementation Details We use an elitism rate of 100% (all parents compete with offspring for survival). Prior work in [9] has shown that, since parents can fall out of the Pareto front as offspring are added, MOEA does not suffer from the premature convergence observed in EAs guided by total energy with an elitism rate of 100%. The size of a population is \( p = 500 \) conformations. A fixed budget of 10,000,000 energy evaluations (including those in local search, which is a simple greedy search using fragments of length 3) is used. This results in 12–24 hours of CPU time on a 2.4Ghz Core i7 processor, depending on the number of amino acids in a protein sequence. Using such budget ensures fairness among chains of different lengths, since energy computations increases quadratically with chain length. The decoy ensemble reported for each algorithm here combines offspring obtained across all generations.

Experimental Setting We compare the performance of MOEA, MOEA-PC, MOGA, and MOGA-PC to an in-house implementation of the Rosetta decoy sampling in terms of three quantities: (i) lowest absolute energy reached, calculated through Rosetta score4, (ii) lowest lRMSD to the known native structure of each protein system considered here (lRMSD measures the weighted Euclidean distance between corresponding atoms – \( C_\alpha \) atoms here – after optimal superimposition that minimizes distances due to rigid-body motions [4]), and (iii) percentage of conformations with \( C_\alpha \) IRMSD to the known native structure below 5Å. These three quantities are calculated over an ensemble that combines 5 runs of each algorithm in order to take into account differences due to randomization. It is important to note that IRMSD is an imperfect structural dissimilarity measure and is not a metric. Its interpretation can be difficult beyond a certain value [3]. For instance, while an IRMSD of 1 vs. 5Å indicates closer proximity to the structure of reference, interpretation is difficult beyond 6Å. While typically an IRMSD > 5 – 6Å from the native structure is taken as indicative that the conformation is not similar to the native structure, exceptions exist, particularly on the borderline. Cases can be found where the dissimilarity is on specific regions, while the overall topology of the native structure has been captured.

We show such a case in the Results section.

Implementing Rosetta Decoy Sampling

In the results analyzed here, we refer by Rosetta to the classic implementation of the coarse-grained decoy sampling stage in the full Rosetta protocol. We have put together our own in-house version, which removes the variable temperature schedule and the combination of score2 and score5 in stage 3 of the decoy sampling in Rosetta. These modifications are made so that a fair comparison can be conducted between our EAs and the Rosetta decoy sampling and allow attributing improvements on performance measurements of interest to the novel features in the proposed EAs over the baseline MC-based search in Rosetta. One can always incorporate the variable temperature scheme, the combination of energy functions, and possibly other heuristics as they are introduced in future versions of Rosetta in the baseline algorithms proposed and analyzed here. We run our in-house version of Rosetta 1,5000 times to obtain a decoy ensemble of 1,5000 conformations. This uses 54,000,000 energy evaluations, which is in the same order as the total number of energy evaluations in the fixed budget used for our EAs (across all 5 runs).

3 Results

Target Protein Sequences Twenty protein sequences have been selected, with chain lengths varying from 53 to 146 amino acids and folds encompassing \( \alpha \), \( \beta \), and \( \alpha/\beta \) (no distinction is made between \( \alpha+\beta \) and \( \alpha/\beta \) folds). All these systems have known native structures in the PDB.

3.1 Comparative Analysis

Comparison of lowest energy over a decoy ensemble allows gaging the exploration capability of different algorithms. Fig. 1 shows the lowest energy sampled by EAs vs. Rosetta. In all systems, at least one of the EAs reaches lower values over Rosetta. The multi-objective optimization techniques that further incorporate Pareto count do not hamper the ability to reach low energies, doing better or comparably on at least 50% of the systems. MOEA-PC and MOGA-PC do worse than MOEA and MOGA, respectively, on 7 – 10 systems(with PDB ids 1sap, 1fwp, 1dtjA, 2ci2, 1tig, 3gw1, and 1aly for MOEA-PC vs. MOEA and PDB ids 1dtdB, 1hz6A, 1fwp, 1ail, 2ci2, 2ezk, 3gw1, 2hg6, 2h5nD, and 1aly for MOGA-PC vs. MOGA).

Since lowest energy may not capture the native basin, the second metric we employ is IRMSD. Fig. 2
shows the lowest lRMSD to the native structure obtained by each of the algorithms on each of the target protein sequences. Comparison of these values shows that at least one of the EAs does better than Rosetta on each system. There are no clear winners in terms of lowest lRMSD when comparing EAs with one another.

Fig. 3 shows the percentage of sampled conformations with lRMSD to the native structure less than 5Å as obtained by each of the algorithms on each of the protein sequences. These percentages are small for many systems (sampling is dominated by high-lRMSD conformations). In particular, while there are many systems where the EAs do better or comparably to Rosetta, there are three systems (with PDB ids 1c8cA, 1sap, aly) where Rosetta obtains more than twice the percentage of conformations with < 5Å to the native structure. This is a clear direction for further investigating how to balance ensemble diversification with improved sampling near the unknown native structure.

3.2 Analysis on Selected Systems

We provide some further details on selected systems. Fig. 4(a) shows the energy landscape sampled by MOEA vs. that sampled by MOEA-PC to illustrate that in many systems, MOEA-PC may not significantly improve the overall lowest lRMSD to the native structure but instead improves the funneling of the sampled landscape. Fig.4(b)-(c) highlights two systems on which there are dramatic improvements in terms of lowest lRMSD to the native structure achieved by our algorithms over Rosetta, from a 5.8 to 3.2Å for the system with native PDB id 3gwl to a 2.3 to 1.0Å for the system with native PDB id 1dtjA.

4 Conclusion

There are two conclusions that can be drawn from the presented results. First, EAs can improve the exploration capability over MC-based methods, when enough domain specific expertise accumulated over the years is properly incorporated. Second, there are benefits to decomposing an energy function and casting template-free protein structure prediction as a multi-objective optimization problem. The results presented in this paper suggest that a carefully-implemented EA framework and its multi-objective realizations can enhance exploration capability and warrant further investigation, not only for structure prediction, but in more general settings of structure modeling.

Acknowledgements

This work is supported NSF Award Nos. 1016995 and 1144106 (CAREER), and a Jeffress Trust Program in Interdisciplinary Research Award.

References

Figure 1: Comparison of lowest energies reached by EAs and Rosetta decoy sampling on all protein systems.

Figure 2: Comparison of lowest IRMSDs to the native structure of each system achieved by EAs and Rosetta decoy sampling.

Figure 3: Comparison of percentage of sampled conformations with IRMSDs less than 5 Å from known native structure on each system by EAs and Rosetta decoy sampling.
Figure 4: (a) Energies vs. lRMSDs from the native structure are shown on one selected system as obtained by MOEA and MOEA-PC. (b)-(c) Two systems are highlighted. The native structure is shown in gray. The lowest-lRMSD conformation obtained by Rosetta is shown in yellow. The lowest lRMSD obtained by MOGA-PC is shown in orange. Rendering is done with the VMD software [2].