Using Subpopulation EAs to Map Molecular Structure Landscapes

Ahmed Bin Zaman
Dept of Computer Science
George Mason University
azaman6@gmu.edu

Kenneth De Jong
Dept of Computer Science
George Mason University
kdejong@gmu.edu

Amarda Shehu*
Dept of Computer Science
George Mason University
amarda@gmu.edu

ABSTRACT

The emerging view in molecular biology is that molecules are intrinsically dynamic systems rearranging themselves into different structures to interact with molecules in the cell. Such rearrangements take place on energy landscapes that are vast and multimodal, with minima housing alternative structures. The multiplicity of biologically-active structures is prompting researchers to expand their treatment of classic computational biology problems, such as the template-free protein structure prediction problem (PSP), beyond the quest for the global optimum. In this paper, we revisit subpopulation-oriented EAs as vehicles to switch the objective from classic optimization to landscape mapping. Specifically, we present two EAs, one of which makes use of subpopulation competition to allocate more computational resources to fitter subpopulations, and another of which additionally utilizes a niche preservation technique to maintain stable and diverse subpopulations. Initial assessment on benchmark optimization problems confirms that stabler subpopulations are achieved by the niche-preserving EA. Evaluation on unknown energy landscapes in the context of PSP demonstrates superior mapping performance by both algorithms over a popular Monte Carlo-based method, with the niche-preserving EA achieving superior exploration of lower-energy regions. These results suggest that subpopulation EAs hold much promise for solving important mapping problems in computational structural biology.

CCS CONCEPTS

• **Computing methodologies** → **Bio-inspired approaches**; • **Applied computing** → **Molecular structural biology**; **Bioinformatics**.

KEYWORDS

mapping; multimodal landscape; subpopulation competition; niche preservation; molecular structure landscape; protein structure prediction; computational structural biology.

*Corresponding Author

GECCO '19, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6111-8/19/07…$15.00
https://doi.org/10.1145/3321707.3321777

ACM Reference Format:
Ahmed Bin Zaman, Kenneth De Jong, and Amarda Shehu. 2019. Using Subpopulation EAs to Map Molecular Structure Landscapes. In *Genetic and Evolutionary Computation Conference (GECCO '19), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3321707.3321777

1 INTRODUCTION

The driving thrust of research in computational structural biology is on understanding the structure-to-function relationship in biological molecules central to human biology and health. Due to their ubiquity and participation in virtually all processes in a cell's machinery, much of this research is focused on proteins. Decades of experimental, computational, and theoretical research have demonstrated that proteins are intrinsically dynamic molecules that reposition their covalently-linked atoms to assume different three-dimensional (tertiary) structures with which to bind molecular partners in the cell [17, 18, 21]. Structural rearrangements allow regulation of biological activity [2, 3] and correspond to hops on a vast and multimodal energy landscape that arise from physical interactions among constitutive atoms in different three-dimensional placements.

Protein energy landscapes are rich in minima that house tertiary structures. Deep and narrow minima constitute structural states with few but very low-energy structures, whereas shallow and broad minima constitute (more heterogeneous) states with many but not as low-energy structures [25]. When seeking to understand how structure(s) governs the biological activity/ies or function(s) of a protein (and, more broadly, an intrinsically-dynamic molecule), it is imperative to consider the diversity of structural states. To do so, the computational framework needs to shift from classic optimization to mapping of a multimodal landscape in order to identify the diverse minima that correspond to different biologically-active structures.

Retaining possibly numerous minima is also an important step toward advancing the treatment of a classic problem in molecular biology, the template-free protein structure prediction problem (PSP). In template-free PSP, one is provided only the identity and order of the constitutive building blocks, the amino acids, in a protein. Given the sequence of amino acids, the goal has traditionally been to discover *the* biologically-active structure. However, the now-recognized multiplicity of biologically-active structures is prompting researchers to expand their treatment of PSP. Moreover, inherent inaccuracies in energy/fitness functions that evaluate structures result in landscapes where the structure detected in the wet laboratory (which may be one of various biologically-active structures) may not reside in the deepest minimum. As such, classic optimization is indeed insufficient. Instead, the goal becomes

to retain diversity and then use complementary, domain-specific metrics to evaluate computed structures. The latter is treated as a separate problem, known as decoy selection, and is beyond the scope of this paper.

Here we revisit subpopulation-oriented Evolutionary Algorithms (EAs) as vehicles to switch the objective from classic optimization to mapping. The concept of a subpopulation is appealing, as it can be directly linked to a structural state. An EA that evolves and maintains multiple subpopulations at local minima while exploring new regions of the fitness landscape seems ideally suited for identifying multiple structural states. To do so, subpopulation EAs must maintain diversity, a recurring theme in Evolutionary Computation (EC) research. Many approaches have been pursued over the years, as summarized in Section 2. Diversity retainment strategies have also been motivated by problems in protein modeling, but none so far have considered subpopulation EAs.

In this paper, we develop a subpopulation EA by building on earlier work on effective representations of protein tertiary structure and representation-aware variation operators. We then further extend this baseline EA so as not only to allocate more computational resources to fitter subpopulations, but additionally maintain stable and diverse subpopulations via a niche preservation technique.

While our primary motivation is identifying the modality of unknown molecular structure landscapes, we first evaluate the two EAs on benchmark problems with fitness landscapes of known modalities. We then provide a comparative evaluation in the PSP setting and additionally compare the two EAs against a highly-popular method in the PSP community. Our results demonstrate that the subpopulation mechanism offers several advantages over the state-of-the-art, with the niche preservation technique yielding the best performance. These results motivate further investigation and development of subpopulation EAs to further advance treatment of protein structure modeling and, more broadly, molecular modeling problems in molecular biology.

The rest of this paper is organized as follows. After presenting a summary of related work in Section 2, the paper continues with a description of the two subpopulation EAs in Section 3. Section 4 then relates the evaluation of these EAs, and the paper concludes with a summary and discussion of further research in Section 5.

2 RELATED WORK

The need to map landscapes as key to understanding a wide range of molecular phenomena has long been recognized across computational physics, organic and inorganic chemistry, and biology [4, 5, 12–14, 21, 32, 34]. For instance, mapping the energy landscape of a cluster of 38 Lennard-Jones atomic particles reveals a double funnel that provides a microscopic basis for understanding how relaxation to the global minimum is diverted into a set of competing structures [34]. In [14], the mapped energy landscapes of small clusters of atoms are revealed to be highly heterogeneous and contain low-energy minima with large basins of attraction. In [12], the energy landscape is shown to facilitate the analysis and interpretation of supercooling and glass-formation phenomena. In [4, 21], various studies in computational chemistry, physics, and biology are summarized to propose and support the holistic view of the energy landscape as central to explaining the behavior of atomic clusters, glasses, and even proteins.

While great progress has been made in mapping energy landscapes of atomic clusters [32], glasses [5], and short peptides [13], mapping protein energy landscapes remains challenging due to the complexity of such landscapes. In glasses, atomic particles, and short peptides, the number of interacting atoms/particles does not exceed a few hundreds, and EAs that rely mainly on exploitation and limit exploration to naive strategies (e.g., random restart) can be useful. However, such approaches lose efficacy rapidly on landscapes of increasing modality, and sophisticated strategies are needed to balance between exploitation and exploration to avoid premature convergence.

Building on the pioneering efforts of Holland, De Jong, Goldberg, and Richardson [6, 10], various strategies have been proposed to address adequate exploration and diversity maintenance. Biased towards strategies that have been shown effective on computational structural biology problems, we highlight here three main techniques often used in combination: a hall of fame mechanism, multi-objective optimization, and hybridization. Work in [30] integrates a hall of fame mechanism in a hybridized/memetic EA to encode a detailed representation of the EA-explored landscape. Work in [27, 28] links the presence of multiple minima in protein energy landscapes to competing objectives in energy functions and demonstrates the utility of multi-objective optimization EAs. Work in [7–9] additionally debuts decentralized selection operators to retain diversity. Work in [26, 29] pursues various recombination strategies to promote generation of diverse candidates, hybridization for better exploitation, and non-local optimization operators to balance between exploration and exploitation.

While EC literature on subpopulation models is quite extensive, subpopulation EAs have not yet been considered for molecular modeling. Largely, existing research considers two scenarios, one where there is prior information on landscape modalities, and one where there is no such information. For the case of prior information, we highlight seminal work by Goldberg and Richardson [6], which assumes that the number of modalities/optima and their location are both known. This setting is not valid in molecular structure modeling, where the objective is to actually discover the diverse optima. An early survey by Spears [33] summarizes the use of restricted mating schemes to evolve subpopulations when no information about the optima is available. Work in [16] proposes a set of multi-population genetic algorithm (GA) operators for general landscape mapping. More recent work in [20] applies subpopulation EAs to the problem of feature selection but utilizes known information to organize the initial population into subpopulations (also referred to as tribes in [20]).

In this paper, we presume no *a priori* information regarding the number and/or location of optima, or the distinct characteristics that may allow organizing individuals in the initial population into distinct subpopulations. We note that in a discovery setting, the location of the competitive states would not be known in molecular modeling, though occasionally in computational physics or chemistry applications information would be available regarding the number of such states and attributes distinguishing them. In protein structure modeling, such information is not available, and one must proceed in more difficult blind settings.

3 METHODS

As described above, we first design a baseline subpopulation EA, to which we refer as SP-EA$^-$ from now on. In summary, SP-EA$^-$ organizes the initial population into subpopulations and then applies subpopulation competition to provide more resources to the fitter subpopulations during the evolutionary process. The second algorithm we propose to improve the stability of subpopulations via a niching technique is referred to as SP-EA$^+$ from now on. In summary, SP-EA$^+$ adds onto SP-EA$^-$ an additional mechanism to prevent genetic drift and maintain diversity of subpopulations.

3.1 A Baseline Subpopulation EA

SP-EA$^-$, shown in pseudocode in Algorithm 1, first initializes a running counter that keeps track of fitness function evaluations (line 1) so as to evaluate and compare the two different algorithms (SP-EA$^-$ and SP-EA$^+$) using the same user-defined budget FMAX of fitness evaluations.

Algo. 1 Baseline EA

Require: `FMAX` //total computational budget
`N` //population size
`CompFreq` //competition frequency
`ElitismRate` //elitism rate

1: $fcounter \leftarrow$ `FMAX` //counter of fitness evaluations
2: $i \leftarrow 0$ //generation counter
3: $\langle \mathcal{P}_i, budgetSpent \rangle \leftarrow$ InitOper(`N`) //generate //initial population
4: $fcounter \leftarrow fcounter - budgetSpent$
5: $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\} \leftarrow$ GenSubPops($\mathcal{P}_i$) //divide into //subpopulations
6: **while** $fcounter > 0$ **do**
7: **for** $S \in \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$ **do**
8: $C \leftarrow \emptyset$ //set of offspring
9: **for** $s \in S$ **do**
10: $c \leftarrow$ VarOper(s) //generate offspring
11: $\langle c', f', budgetSpent \rangle \leftarrow$ ImprovOper(c) //improve
12: $fcounter \leftarrow fcounter - budgetSpent$
13: $C \leftarrow C \cup \{c', f'\}$ //add improved offspring
14: $S' \leftarrow$ SelOper($S, C,$ `ElitismRate`) //select
15: $\mathcal{S} \leftarrow \mathcal{S}'$ //update subpopulation
16: **if** i mod `CompFreq` $= 0$ **then**
17: $\{\mathcal{S}_1, \ldots, \mathcal{S}_K\} \leftarrow$ SubPopCompete($\{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$)
18: $i \leftarrow i + 1$

The initial population is obtained via an initialization mechanism (line 3). For the benchmark problems studied here, coordinates for individuals are drawn uniformly at random from the given parameter ranges. On applications to proteins, different initialization mechanisms can be employed that take advantage of domain-specific knowledge. Work in [26] describes an effective initialization mechanism that makes use of the molecular fragment replacement technique that is popular in PSP.

3.1.1 Defining Subpopulations. Unlike a classic EA, where, once initialized, the population evolves over generations, the subpopulation EAs we present here first organize the initial population into subpopulations. Unlike other work, where information may be available on the attributes that can be leveraged for such organization, here we assume no *a priori* information. That is why line 5 in Algorithm 1 simply refers to a mechanism to generate subpopulations from the initial population. In this paper, we employ leader clustering, but other clustering algorithms can be utilized. The main idea behind leader clustering is that individuals are considered in order, and each individual either forms a new cluster (becoming its representative) or is assigned to the first cluster whose representative is within a distance threshold.

For the benchmark problems considered here, we utilize Euclidean distance to measure the distance between an individual yet to be assigned to a cluster and the representative individual of each cluster computed so far. In applications (and adaptation) of SP-EA$^-$ and SP-EA$^-$ to proteins, the distance function used is least root-mean-squared-deviation (lrmsd) [22]; lrmsd first removes differences between two three-dimensional structures due to rigid-body motions, and then averages the Euclidean distance over the number of atoms in each structure.

We note that the number of subpopulations in line 5 is not predetermined. Clustering algorithms that necessitate such determination can be used, but one of the reasons we prefer leader clustering is that the number of clusters follows based on the specified distance threshold. Once subpopulations are determined, they each undergo an evolutionary process. Lines 7-15 in Algorithm 1 evolve each subpopulation as follows. For each subpopulation, offspring are recorded in a set C that is initialized to the empty set (line 8). Each individual in the current subpopulation S under consideration (line 9) is selected to obtain an offspring c via a variation operator (line 10). The variation operator for the benchmark problems studied here is a Gaussian perturbation operator, which perturbs each coordinate of an individual by a value drawn from a zero-mean Gaussian distribution with a given variance. In applications on proteins, the variation operator is implemented as in [26] and is described in detail later.

3.1.2 Evolving Each Subpopulation. The obtained offspring is then subjected to a local search that seeks to improve the offspring (line 11). For the benchmark problems considered here, a naive local search chooses any of the coordinates of the offspring with equal probability and then applies a simple gradient descent on the chosen coordinate for a total of `budgetSpent` iterations/cycles. The local search utilized in the applications on proteins is implemented as in [26] and is detailed later. Note that all fitness evaluations that occur in the improvement operator are counted, and they are removed from the total budget (line 12).

Once the offspring of a subpopulation are generated and stored in C (line 13), they compete for survival with parents (line 14). An elitism-based truncation selection mechanism is employed for this purpose. Based on an user-defined elitism rate, a percentage of the fittest parents are selected to compete against the set of offspring. The selected parents and offspring are sorted by their fitness values, and the fittest $|S|$ individuals are selected for the next generation, where $|S|$ is the size of the current subpopulation.

3.1.3 Competition Among Subpopulations. Once this process completes for each subpopulation (line 7), subpopulations may now compete with one another. How frequently this occurs is determined via a user-defined competition frequency (line 16). In this paper, the competition takes place once every `CompFreq` generations. Algorithm 1 does not provide details of the competition process (line 17) which may update subpopulations. Different implementations of this process give rise to different variants of subpopulation EAs. Let us delay momentarily the implementations we consider here in the interest of first explaining how the competition among subpopulations takes places.

Provided a mechanism exists to associate a fitness with an entire subpopulation, a competition mechanism aims to accomplish the following. The fittest subpopulation is rewarded with more resources in hopes of affording better exploration of the landscape. This is operationalized by replicating the fittest individual in the fittest subpopulation; the size of the fittest subpopulation increases by 1. In addition, the worst (lowest fitness) subpopulation is penalized by discarding its worst (lowest fitness) individual; the size of the worst subpopulation decreases by 1. Note that it is possible under this mechanism for a subpopulation to gradually lose all its members, resulting in the elimination of a subpopulation.

As Algorithm 1 shows, the process of subpopulation evolution and subpopulation competition is repeated until the fitness evaluation budget is exhausted. At that point, the algorithm terminates. The competitive mechanism described above is greatly dependent on how the fitness of a subpopulation is defined. SP-EA$^-$ considers a straightforward definition of the fitness of a subpopulation as the average over the fitness values of individuals in the subpopulation:

$$F_S = \frac{\sum_{s \in S} f(s)}{|S|} \tag{1}$$

3.2 A Niche-Preserving Subpopulation EA

The population competition utilized in SP-EA$^-$ may result in a loss of population diversity in cases in which a subpopulation with the highest fit individuals may persist indefinitely, gradually acquiring more members, resulting in the loss of subpopulations containing less fit individuals. To provide some subpopulation stability, SP-EA$^+$ preserves niches in a population by redefining the fitness of a subpopulation to consider not only the fitness values of its members but also the size of the subpopulation. Specifically,

$$F_S = \frac{\sum_{s \in S} f(s)}{|S|} + T \cdot |S| \tag{2}$$

In Equation 2, the fitness of a subpopulation not only calculates the average over the fitness values of the members of the subpopulation, but also penalizes the subpopulation fitness by a factor (governed by the "temperature" parameter T) of the subpopulation size (number of members). Larger subpopulations have more penalty added to their score. This ensures that a large subpopulation can only win, if it really holds much fitter individuals than smaller subpopulations. Otherwise, smaller subpopulations get to increase their sizes. This way, a small subpopulation can also win if it holds good individuals, even if they are not the fittest. This helps preserve the niches, lets the algorithm map more of the subspaces in the search space, and gives the algorithm a better chance of finding diverse optima.

Note that the temperature parameter shifts the balance in the fitness of a subpopulation towards fitness or size. For instance, if $T = 0$, SP-EA$^+$ reverts to SP-EA$^-$ and does not consider the size of a population.

3.3 Adaptations for Application to Proteins

In the above exposition, we frequently refer to adaptations of various operators for application of SP-EA$^-$ and SP-EA$^+$ to proteins. Below, we briefly summarize each of these operators, as effective implementations for them have appeared in genetic algorithms, memetic EAs, and multi-objective memetic EAs in literature [26–28].

3.3.1 Initial Population. A structure (individual) of a protein is represented as a vector of (dihedral) angles. These angles basically determine the spatial arrangement of atoms that are covalently linked to form a chain that folds in different structures in three dimensions. We will refer to this vector representation of a structure as a conformation. Extended conformations can be computed trivially (due to characteristic values for the dihedral angles that "stretch" the chain of atoms in three dimensions). The initialization operator first generates N of such identical extended conformations, where `N` is the population size. Then, as described in [27], each extended conformation is subjected to a 2-stage Metropolis Monte Carlo (MMC) search. The first stage contains 200 moves and uses a temperature of 0 for the Metropolis criterion, which accepts a move if it decreases the energy of a conformation. This stage employs Rosetta's *score*0 [19] energy function to obtain conformations free of self collisions; we note that Rosetta is a popular PSP software package that contains representations, energy functions, and a protocol for generating low-energy conformations given an amino-acid sequence.

The second stage performs MMC search until l consecutive moves fail according to the Metropolis criterion, where l is the number of amino acids in the protein sequence. This stage uses a temperature of 2, which primarily accepts a move if it decreases the potential energy while also allowing small increases. This stage uses the *score*1 Rosetta energy function to reward secondary structure formation. Each move in this 2-stage search is a molecular fragment replacement of length 9.

3.3.2 Molecular Fragment Replacement. A fragment of length f is defined over amino acids at positions i through $i + f - 1$ in the chain. To perform molecular fragment replacement of length f on a conformation C, an uniformly random position i is sampled over the amino acid positions 1 to $l - f + 1$. Here, l is the number of amino acids in C. Then, a random matching fragment of length f is selected from the fragment configuration library and used to replace the $3f$ dihedral angles (ϕ, ψ, and ω per amino acid) of the previously selected fragment in C. As a result, a new conformation is achieved. We use the popular Rosetta fragment server [19] as the fragment configuration library.

3.3.3 Variation and Improvement Operators. The variation operator implements mutation by performing a molecular fragment replacement of length 3 on the parent. A generated offspring is

then improved by employing a local search. The goal is to map the offspring to a nearby local minimum in the energy landscape. The local search is greedy in nature and ensures that only the moves that lower energy are accepted. Each move is a molecular fragment replacement of length 3 and is evaluated with the Rosetta *score*3 energy function. This greedy search returns the lowest-energy conformation sampled when l consecutive moves fail to improve the conformation (l is the number of amino acids in the sequence).

4 RESULTS

In this section, we present a summary of our results. We first apply both algorithms on two generic landscapes with known global minima to analyze their performance in finding these minima as well as in the overall exploration of the subspaces. We also examine the stability of the subpopulations that they generate. Then, we execute both algorithms in the context of PSP on a benchmark dataset and compare them via different metrics against each other and against the popular Rosetta algorithm [19].

4.1 Analysis on Known Fitness Landscapes

We choose two benchmark problems to comparatively evaluate the behavior of SP-EA$^-$ and SP-EA$^+$:

- A sphere: $f(x) = \sqrt{\sum_{i=1}^{n} x_i^2}$.
- The product of two spheres: $f(x) = \sqrt{\sum_{i=1}^{n}(x_i - 200)^2} \times \sqrt{\sum_{i=1}^{n}(x_i + 200)^2}$.

where x is a D-dimensional vector. The landscape of the sphere function contains 1 global minimum, and the landscape of the product of two spheres contains 2 global minima. Each algorithm is run 1,000 times on each problem. On each run, we randomly pick the dimensionality D from $\{2, 5, 10, 20\}$. We set the temperature for SP-EA$^+$ to 6, 12, 25, and 50, respectively, for D = 2, 5, 10, and 20. We fix the range of values for each x_i to $[-500, 500]$. The population size is set to 200, elitism rate for selection to 25%, the frequency for subpopulation competition to 2, and the evaluation budget for each run to 10,000,000 fitness evaluations (this same budget is used in our evaluation on protein landscapes in the context of PSP).

We first evaluate the number of times each algorithm converges to the known global minima (or minimum). We consider an algorithm to have converged if for 1-sphere problem, the final population generated by the algorithm consists of a single subpopulation and that subpopulation contains the global minimum; and for the 2-sphere problem, the final population generated by the algorithm consists of only two subpopulations and each of the subpopulations contains one global minimum each. Table 1 shows the percentage of times the two EAs converge in 1000 runs on each problem. Both EAs converge in the 1-sphere problem to the only minimum in all the runs. On the 2-sphere problem, SP-EA$^-$ does not converge to both minima in the final subpopulations. In most cases, SP-EA$^-$ converges to a single subpopulation. This result indicates the genetic drift that occurs along the way, with the population losing diversity early. SP-EA$^+$ performs well and retains both minima the majority of the time, indicating that the niche-preserving technique is effective in preventing premature convergence.

We now provide a visual analysis of the stability of the subpopulations by examining the size of the subpopulations in the final

Table 1: Percentage of times (out of 1,000 runs) SP-EA$^-$ and SP-EA$^+$ converge to the 1 minimum and 2 minima in the known landscapes of the sphere problems considered here.

Algorithm	1-sphere	2-spheres
SP-EA$^-$	100%	0.13%
SP-EA$^+$	100%	71.2%

population of SP-EA$^+$. Fig. 1 shows the histogram of the smaller subpopulation sizes in the final populations for the 2-sphere problem in the cases where SP-EA$^+$ converges. In 75.9% cases, the smaller subpopulation has a size of 80 or more out of 200 individuals in the population. Only in 1.8% of the cases, the smaller subpopulation has a size of 20 or less. Considering the substantial budget, these results confirm that SP-EA$^+$ not only retains population diversity, but also produce stable subpopulations.

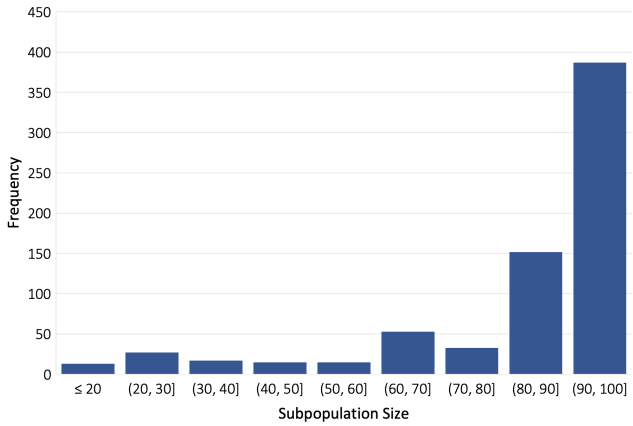

Figure 1: Histogram of smaller subpopulation sizes in the final population for the 2-sphere problem on the runs where SP-EA$^+$ produces 2 subpopulations that contain one minima each.

4.2 Analysis on Protein Structure Landscapes

We consider a benchmark dataset of 20 proteins of different folds (α, β, $\alpha + \beta$, and *coil*) and lengths (from 53 to 146 amino acids). Part of this dataset was originally introduced in [24] and then over time enriched with more protein targets [11, 27]. With regards to parameter values, differences from the above evaluation include the distance threshold, which is set to 5Å, and the temperature in SP-EA$^+$, which is set to 2, and the number of runs, which is set to 5 times on each protein sequence to account for stochasticity and we report the best performance over all 5 runs combined for each EA. We compare the two EAs to each other and the Rosetta algorithm. Note that the Rosetta algorithm implements simulated annealing and runs for a budget of 54,000,000 energy evaluations. Since the evaluation budget for each run of each of our EAs is fixed to 10,000,000 evaluations, this adds up to 50,000,000 over 5 runs.

As is practice in PSP evaluation [31], we measure performance in terms of the lowest reached energy and the lowest reached distance

(lrmsd) to a known biologically-active structure of a target under consideration. The former measures exploration capability. Since lower energies do not necessarily correlate with proximity to the active structure, it is important to also measure distance. It is worth noting that lrmsd is non-descriptive above 8Å and increases with sequence/chain length. An lrmsd within 5–6Å is considered to have captured the active structure under consideration.

Table 2 summarizes the performance of each of the three algorithms in terms of these two metrics; the lowest values on each target are marked in bold. The first column lists the test cases by identifying the Protein Data Bank Identifier (PDB ID) of the entry where an active structure known for each test case is deposited. We note that for many of these test cases, one active structure is reported in the wet laboratory though others may exist.

Table 2 shows that SP-EA$^+$ achieves the lowest energy in 12/20 targets, whereas SP-EA$^-$ and Rosetta do so on 2/20 and 6/20 targets, respectively. In a head-to-head comparison between SP-EA$^+$ and Rosetta, SP-EA$^+$ achieves lower energy in 14/20 targets over Rosetta. Between SP-EA$^+$ and SP-EA$^-$, the former wins in 17/20 cases. Finally, between SP-EA$^-$ and Rosetta, SP-EA$^-$ wins in 11/20 cases.

A similar comparison on lowest lrmsds reveals that SP-EA$^+$ achieves the lowest lRMSD in 12/20 targets, whereas SP-EA$^-$ and Rosetta do so on 2/20 and 10/20 targets, respectively. In a head-to-head comparison between SP-EA$^+$ and Rosetta, Rosetta achieves lower lRMSD in 8/20 targets than SP-EA$^+$. Between SP-EA$^+$ and SP-EA$^-$, the former wins in 15/20 cases. Between SP-EA$^-$ and Rosetta, Rosetta wins in 9/20 cases.

To give some insight into these low lrmsd values, Fig. 2 selects two proteins (with respective active structures under PDB IDs 1ail and 3gwl) and shows the lowest-lrmsd structure obtained by SP-EA$^+$ in each case. These structures (drawn in blue) are superimposed over the corresponding active structures (drawn in olive). The superimposition highlights the quality of the solutions obtained by SP-EA$^+$.

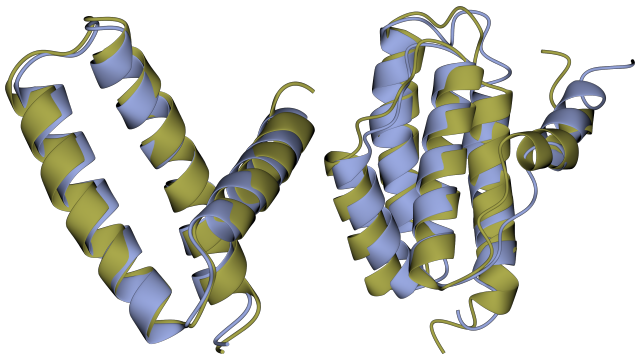

Figure 2: The lowest-lrmsd structure obtained by SP-EA$^+$ on each protein is drawn in blue, superimposed over the corresponding active structure (with PDB id and lrmsd shown), which is drawn in olive. Rendering is performed with the CCP4mg molecular graphics software [23].

The comparisons so far suggest that the subpopulation EAs outperform Rosetta on both metrics. We harden this result via statistical significance analysis tests. We use two statistical significance tests, Fisher's [15] and Barnard's [1] exact tests to determine if the results are statistically significant. We employ the tests over 2x2 contingency matrices generated from the results obtained using the comparison metrics. Fisher's test is conditional and Barnard's test is unconditional in nature. While Fisher's exact test is widely adopted for statistical significance, Barnard's test is generally considered more powerful than Fisher's test on 2x2 contingency matrices.

Table 3 shows the p-values for the 1-sided Fisher's and Barnard's tests for the lowest energy head-to-head comparison. All the values (< 0.05) that reject the null hypothesis with 95% confidence are marked in bold. Both null hypotheses (SP-EA$^+$ does *not* perform better than Rosetta and SP-EA$^+$ does *not* perform better than SP-EA$^-$) are rejected, confirming the superior performance of SP-EA$^+$. The null hypothesis that SP-EA$^-$ does *not* perform better than Rosetta is not rejected, indicating that the performance improvement of SP-EA$^-$ over Rosetta is not statistically significant with 95% confidence.

Similarly, Table 3 also shows the p-values for the 1-sided Fisher's and Barnard's tests for the lowest lrmsd head-to-head comparison. All the values (< 0.05) that reject the null hypothesis with 95% confidence are marked in bold. The null hypothesis that SP-EA$^+$ does *not* perform better than SP-EA$^-$ is rejected, confirming the superior performance of SP-EA$^+$ over SP-EA$^-$. The null hypotheses that SP-EA$^+$ does *not* perform better than Rosetta and that SP-EA$^-$ does not perform better than Rosetta are not rejected, indicating that the performance improvements of the two subpopulation EAs over Rosetta are not statistically significant with 95% confidence.

Taken altogether, the results presented above suggest a stronger exploration capability of the subpopulation EAs over Rosetta and a superiority of the niche-preservation technique in exploration. On the lrmsd-based comparison, none of the algorithms is a clear winner, but the subpopulation EAs perform comparably to Rosetta.

5 CONCLUSION

In this paper we revisit subpopulation-oriented EAs to switch the objective from classic optimization to mapping of fitness landscapes. The latter task is only relevant when the problem of interest is characterized by a multimodal landscape where the various modes contain information about the system being investigated. This is the case for most biological systems, and, in particular, protein molecules.

Since neither the number of subpopulations nor their distribution are known ahead of time for unknown molecular landscapes, we present here a baseline subpopulation EA that makes use of phenotypic clustering to define initial subpopulations and makes use of subpopulation competition to evolve subpopulations. We investigate two different strategies for such competition and show that taking into account not only the height/depth, but also the size/breadth of a local optimum allows better retaining diverse subpopulations that converge to the different modes of known landscapes. Evaluation on unknown landscapes in the context of

Table 2: Comparison of the lowest energy (in Rosetta Energy Units – REUs) obtained by each algorithm on each of the 20 test cases is shown in Columns 4, 5, and 6. Comparison of the lowest lrmsd (measured in Angstroms – Å) to a given active structure in each test case is shown in Columns 7, 8, and 9.

			Lowest Energy			Lowest lrmsd		
PDB ID	Length	Fold	Rosetta	SP-EA$^-$	SP-EA$^+$	Rosetta	SP-EA$^-$	SP-EA$^+$
1ail	73	α	−29.9	−74.1	**−81.3**	4.5	1.5	**1.2**
1aly	146	β	**−112.5**	−63.4	−74.6	12.4	11.1	**10.9**
1aoy	78	α	−73.3	−103.2	**−116.8**	4	**3.1**	**3.1**
1bq9	53	β	−46.9	−51.3	**−64.2**	**2.9**	4.3	4.7
1c8ca	64	β	**−101.4**	−69.5	−78.3	**2.2**	3.7	3.6
1cc5	83	α	**−82.5**	−67.6	−76.4	**3.7**	4.2	4.7
1dtdb	61	$\alpha + \beta$	−66.5	−57.5	**−69.6**	**4.2**	5.2	5
1dtja	76	$\alpha + \beta$	−72.5	**−85.5**	−72.6	**2.3**	**2.3**	2.5
1fwp	68	$\alpha + \beta$	−71.3	−66.9	**−72.1**	**2.8**	4	3.7
1hhp	99	β	**−106.3**	−87.8	−83.5	10.1	8.4	**8.2**
1hz6a	67	$\alpha + \beta$	−117.1	−122.5	**−122.8**	**1.9**	2.3	**1.9**
1isua	62	*coil*	−27	−38.8	**−41.8**	6.6	6.1	**5.8**
1sap	66	β	−107.8	−91.2	**−109.9**	**2.8**	4.4	4
1tig	88	$\alpha + \beta$	**−138.2**	−104.1	−112.2	**2.5**	3.5	3.7
1wapa	68	β	**−109**	−65.9	−71	6.5	5.9	**5.6**
2ci2	83	$\alpha + \beta$	−37.8	−72.7	**−82.7**	5.8	3.6	**3.5**
2ezk	93	α	−51.1	−126.4	**−135.2**	3.6	3	**2.9**
2h5nd	123	α	−82.5	−134.9	**−139.1**	**7.4**	7.8	**7.4**
2hg6	106	$\alpha + \beta$	−82.5	**−96.4**	−95.1	9.4	8.9	**8.7**
3gwl	106	β	−68.2	−112	**−117.8**	5.8	4.2	**2.9**

Table 3: p-values obtained by 1-sided Fisher's and Barnard's tests for head-to-head comparison of the algorithms on lowest energy (left) and lowest lrmsd (right). Top panel evaluates the null hypothesis that SP-EA$^+$ does *not* perform better than Rosetta. Middle panel evaluates the null hypothesis that SP-EA$^+$ does *not* perform better than SP-EA$^-$. Bottom panel evaluates the null hypothesis that SP-EA$^-$ does *not* perform better than Rosetta.

SP-EA$^+$ vs. Rosetta		
Test	Lowest energy	Lowest lrmsd
Fisher's	**0.01282**	0.3756
Barnard's	**0.008299**	0.3057

SP-EA$^+$ vs. SP-EA$^-$		
Test	Lowest energy	Lowest lrmsd
Fisher's	**0.000009693**	**0.0006159**
Barnard's	**0.000004182**	**0.0003401**

SP-EA$^-$ vs. Rosetta		
Test	Lowest energy	Lowest lrmsd
Fisher's	0.3762	0.5
Barnard's	0.3179	0.4373

protein structure prediction shows that niche preservation also confers higher exploration capability.

We believe this work is useful for a broad range of landscape mapping problems in various domains. Though our primary motivation in this paper is in the domain of protein modeling in computational structural biology, there are many problems in chemistry, physics, and network science that necessitate characterization of complex systems with diverse functional states. Prompted by the encouraging results presented here, we intend to further investigate subpopulation EAs and niche preservation techniques to advance research in molecular structure modeling. Further directions of work include investigating the impact of temperature in niche preservation, better regulating resources spent by larger subpopulations in the evolution process, as well as investigating additional variation, selection operators, and alternative mechanisms of subpopulation competition.

6 ACKNOWLEDGMENTS

This work is supported in part by NSF IIS Grant No. 1763233. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA (URL: http://orc.gmu.edu).

REFERENCES

[1] G. A. Barnard. 1945. A new test of 2x2 tables. *Nature* 156 (1945), 177.
[2] D. D. Boehr, R. Nussinov, and P. E. Wright. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol* 5, 11 (2009), 789–96.
[3] D. D. Boehr and P. E. Wright. 2008. How do proteins interact? *science* 320, 5882 (2008), 1429–1430.
[4] C. L. III Brooks, J. N. Onuchic, and D. J. Wales. 2001. Taking a Walk on a Landscape. *Science* 293, 5530 (2001), 612–613.
[5] P. Charbonneau, J. Kurchan, G. Parisi, P. Urbani, and F. Zamponi. 2013. Fractal free energy landscapes in structural glasses. *Nat Commun* 5, 4725 (2013), 3725.
[6] W. Chen and K. Y. Szeto. 1987. Complex energy landscape mapping by histogram assisted genetic algorithm. In *Intl Conf Genet Algorithms (ICGA)*. 44–49.
[7] R. Clausen, B. Ma, R. Nussinov, and A. Shehu. 2015. Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Comput Biol* 11, 9 (2015), e1004470.

[8] R. Clausen and A. Shehu. 2014. 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, 269–278.

[9] R. Clausen and A. Shehu. 2015. A Data-driven Evolutionary Algorithm for Mapping Multi-basin Protein Energy Landscapes. *J Comp Biol* 22, 9 (2015), 844–860.

[10] K. A. De Jong. 1975. *An analysis of the behavior of a class of genetic adaptive systems.* Master's thesis. University of Michigan.

[11] J. DeBartolo, G. Hocky, M. Wilde, J. Xu, K. F. Freed, and T. R. Sosnick. 2010. Protein structure prediction enhanced with evolutionary diversity: SPEED. *ProteinSci* 19, 3 (2010), 520–534.

[12] P. G. Debenedetti and F. H. Stillinger. 2001. Supercooled liquids and the glass transition. *Nature* 410, 6825 (2001), 259–267.

[13] D. Devaurs, K. Molloy, M. Vaisset, and A. Shehu. 2015. Characterizing Energy Landscapes of Peptides using a Combination of Stochastic Algorithms. *IEEE Trans. NanoBioSci.* 14, 5 (2015), 545–552.

[14] J. P. K. Doye. 2002. The network topology of a potential energy landscape: A static scale-free network. *Phys Rev Lett* 88, 23 (2002), 238701.

[15] R. A. Fisher. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J Roy Stat Soc* 85, 1 (1922), 87–94.

[16] Y. B. Guo and K. Y. Szeto. 2009. Landscape mapping by multi-population genetic algorithm. In *Nature Inspired Cooperative Strategies for Optimization*. Studies in Computational Intelligence, Vol. 236. Springer, Chapter 14, 165–176.

[17] J. S. Hub and B. L. de Groot. 2009. Detection of Functional Modes in Protein Dynamics. *PLoS Comp Biol* 5, 8 (2009), e1000480.

[18] K. Jenzler-Wildman and D. Kern. 2007. Dynamic personalities of proteins. *Nature* 450 (2007), 964–972.

[19] A. Leaver-Fay et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487 (2011), 545–574.

[20] B. Ma and Y. Xia. 2017. A tribe competition-based genetic algorithm for feature selection in pattern classification. *Applied Soft Computing* 58 (2017), 328–338.

[21] 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 Comp. Biol.* 12, 4 (2016), e1004619.

[22] A. D. McLachlan. 1972. A mathematical procedure for superimposing atomic coordinates of proteins. 26, 6 (1972), 656–657.

[23] S. McNicholas, E. Potterton, K. S. Wilson, and M. E. M. Noble. 2011. Presenting your structures: the CCP4mg molecular-graphics software. *Acta Cryst* D76 (2011), 386–394.

[24] Jens Meiler and David Baker. 2003. Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences of the United States of America* 100, 21 (2003), 12105–12110. https://doi.org/10.1073/pnas.1831973100

[25] R. Nussinov and P. G. Wolynes. 2014. A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys Chem Chem Phys* 16, 14 (2014), 6321–6322.

[26] B. Olson, K. A. De Jong, and A. Shehu. 2013. Off-Lattice Protein Structure Prediction with Homologous Crossover. In *Conf on Genetic and Evolutionary Computation (GECCO)*. ACM, New York, NY, 287–294.

[27] B. Olson and A. Shehu. 2013. 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., 430–439.

[28] B. Olson and A. Shehu. 2014. 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, 143–148.

[29] E. Sapin, K. A. De Jong, and A. Shehu. [n. d.]. From Optimization to Mapping: An Evolutionary Algorithm for Protein Energy Landscapes. *IEEE/ACM Trans Comput Biol and Bioinf* 15, 3), pages = 719-731, year = 2018, note = doi: 10.1109/TCBB.2016.2628745 ([n. d.]).

[30] E. Sapin, K. A. De Jong, and A. Shehu. 2015. Evolutionary Search Strategies for Efficient Sample-based Representations of Multiple-basin Protein Energy Landscapes. In *IEEE Intl Conf Bioinf and Biomed.* 13–20.

[31] A. Shehu. 2015. A Review of Evolutionary Algorithms for Computing Functional Conformations of Protein Molecules. In *Computer-Aided Drug Discovery*, W. Zhang (Ed.). Springer Verlag.

[32] L. C. Smeeton, J. D. Farrell, M. T. Oakley, D. J. Wales, and R. L. Johnston. 2015. Structures and Energy Landscapes of Hydrated Sulfate Clusters. *J Chem Theory Comput* 11, 5 (2015), 2377âĂŞ2384.

[33] W. M. Spears. [n. d.]. Simple Subpopulation Schemes. In *Evolutionary Programming Conf.* World Scientific, 1429–1430.

[34] D. J. Wales, M. A. Miller, and T. R. Walsh. 1998. Archetypal energy landscapes. *Nature* 394, 6695 (1998), 758–760.