EPiC Series in Computing

Volume 83, 2022, Pages 20–30

Proceedings of 14th International Conference
on Bioinformatics and Computational Biology

Guiding Protein Conformation Sampling with Conformation Space Maps

Ahmed Bin Zaman[1,5], Kenneth De Jong[1,5], and Amarda Shehu[1,2,3,4,5]*

[1] Dept. of Computer Science, George Mason University, Fairfax, VA 22030, USA
azaman6@gmu.edu, kdejong@gmu.edu, amarda@gmu.edu
[2] Center for Advancing Human-Machine Partnerships, George Mason University
[3] Dept. of Bioengineering, George Mason University
[4] School of Systems Biology, George Mason University
[5] Mailing Address: MS 4A5, 4400 University Dr., Fairfax, VA 22030, USA

Abstract

Deep learning research, from ResNet to AlphaFold2, convincingly shows that deep learning can predict the native conformation of a given protein sequence with high accuracy. Accounting for the plasticity of protein molecules remains challenging, and powerful algorithms are needed to sample the conformation space of a given amino-acid sequence. In the complex and high-dimensional energy surface that accompanies this space, it is critical to explore a broad range of areas. In this paper, we present a novel evolutionary algorithm that guides its optimization process with a memory of the explored conformation space, so that it can avoid searching already explored regions and search in the unexplored regions. The algorithm periodically consults an evolving map that stores already sampled non-redundant conformations to enhance exploration during selection. Evaluation on diverse datasets shows superior performance of the algorithm over the state-of-the-art algorithms.

1 Introduction

The recognition that the three-dimensional/tertiary structure of a protein molecule determines, to a large extent, its biological function and molecular mechanisms in the cell [3] has motivated the development of many computational approaches to protein structure modeling over the years [8]. A decade of work on deep neural networks shows that such networks can accurately predict contacts between the amino acids that are the building blocks of a protein molecule when its amino-acid sequence and databases of sequences and structures of known proteins. In particular, the powerful ResNet model [19] inspired and became a precursor to AlphaFold and AlphaFold2 [7], which recently demonstrated the ability to predict the native tertiary structure of a protein amino-acid sequence with extremely high accuracy.

Accounting for the structural plasticity of protein molecules remains challenging, despite increasing evidence that proteins harness their ability to assume different structures to regular

*Corresponding author

molecular interactions [2]. Obtaining a broader view of the structure space accessed by a protein that goes beyond one structure is important to understand molecular mechanisms and support the development of therapeutics. The literature on algorithms for exploring the protein structure space is rich in algorithms operate under the umbrella of optimization and enhance the sampling capability of Monte Carlo or Molecular Dynamics-based methods [8]. While these methods can provide great detail on specific systems, they are inferior in their sampling capability to evolutionary algorithms [13, 15].

In the complex and high-dimensional energy surface that accompanies the structure space, it is critical for sampling algorithms to explore a broad range of areas to increase the chances of sampling relevant structures. From now on we will utilize the concept of *a conformation*, which indicates a specific choice of representing a tertiary structure that facilitates certain operators in a sampling algorithm. In this paper, we present a novel evolutionary algorithm that guides its optimization process with a memory of the explored conformation space so that it can avoid searching already explored regions and search in the unexplored regions. The algorithm periodically consults an evolving map that stores already sampled non-redundant conformations to enhance exploration during selection. Evaluation on diverse datasets shows superior performance of the algorithm over the state-of-the-art algorithms. Before proceeding with a description and evaluation of the algorithm, we first provide an overview of related work.

2 Methods

We demonstrate the effects of guiding a conformation sampling algorithm with a memory of already explored spaces by building over a hybrid/memetic evolutionary algorithm (HEA), proposed in [14] and evaluated against Rosetta and others [14, 23, 24]. HEA contains the basic evolutionary ingredients and evolves a fixed-size population of individuals (conformations) for a number of generations. For the memory of the conformation space, we make use of the evolving map we developed previously [26, 25] which utilizes low-dimensional representations of protein conformations. The map represents the explored conformation space effectively by storing non-redundant diverse conformations and has considerably small storage requirement compared to a memory which stores all the conformations ever generated. We equip the HEA algorithm with the map and then change its selection operator which selects the individuals to construct the next generation. We propose a new selection mechanism that consults the map to select individuals in a way that allows sampling conformations from the unexplored parts of the conformation space.

2.1 Summary of HEA

In HEA, the initial population is obtained by applying an initial population operator. In each generation, individuals in the population are considered parents and offspring are produced from the parents via a variation operator. The offspring are then subjected to an improvement operator to improve their fitness. The improved offspring are then combined with the parents and a selection operator is utilized to select individuals for the next generation.

Initial Population Operator From a given amino-acid sequence, the initial population operator first creates p identical extended chains, where p is the size of the population, in Rosetta's centroid representation. For each amino-acid, the representation models only the heavy-backbone atoms and a pseudo-atom representing the centroid of the side chain atoms. To randomize each of these p extended chains, a two-stage MMC search is utilized. The goal for

the first stage is to randomize the extended chains while avoiding steric clashes (self collisions). To do so, it employs Rosetta *score*0 energy function. The second stage employs the *score*1 scoring function which encourages the formation of secondary structure elements. Each move in this MMC search is a molecular fragment replacement of length 9. We describe fragment replacement in the next section. The interested reader can find further details in Ref. [14].

Variation and Improvement Operators In HEA, each individual in the population is subjected to a variation operator to obtain an offspring. The variation operator applies a molecular fragment replacement of length 3 that introduces a small structural change over the parent. Molecular fragment replacement works as follows. A fragment of length m consists of m consecutive amino acids at positions i through $i+m-1$ in the chain. All fragment configurations ($3m$ backbone dihedral angles) of length m drawn out from known native conformations are stored in a fragment configuration library. To perform molecular fragment replacement of length m on a conformation, a uniformly random position i is sampled over the amino acid positions 1 to $l - m + 1$. Here, l is the number of amino acids in the conformation. Then, a random matching fragment of length m is extracted from the fragment configuration library and used to replace the $3m$ dihedral angles of the previously selected fragment in the chain. Any offspring generated by the variation operator is subjected to an improvement operator that employs local search to map the offspring to a nearby local minima in the energy landscape. The local search is greedy in nature and only the moves that lower Rosetta *score*3 energy are accepted. Each move in the local search is a molecular fragment replacement of length 3.

Selection Operator HEA uses an elitism-based truncation selection mechanism to select individuals for the next generation. All individuals (parents and improved offspring) are evaluated using Rosetta's full centroid scoring function *score*4 that considers short- and long-range hydrogen bonding in addition to the terms in *score*3. Top $n\%$ individuals from the parents are combined with the improved offspring to compete for survival; n is the elitism rate. The competing individuals are sorted in increasing order of their *score*4 and the top p individuals are selected to represent the population for the next generation (p is the size of the population).

2.2 Evolving Map of Explored Space

The map utilizes an energetic layer and a geometric layer to store generated conformations. The energetic layer is implemented as a 1-D grid defined over Rosetta *score*4 energy intervals in the range $[-200, 0]$. The choice of the bounds reflects the facts that positive energy conformations are of very low quality (conformation sampling algorithms start producing negative energy conformations very early) and our experiments reveal that the *score*4 energy of a generated conformation is comfortably over -200 Rosetta Energy Units (REUs). Each interval in the grid is set to a small value of 2 REUs.

For each such energy interval, a 3-D geometric grid of 3 shape-similarity features is defined. These features are the first momenta of atomic distance distributions from 3 reference points that summarize a conformation. The reference points are the molecular centroid (ctd), the farthest point from the centroid (fct), and the farthest point from fct (ffct). Each cube in the grid is represented by the integer levels of these first momenta.

For each conformation (d_i) generated by HEA, we consider including it in the map as follows. We first map it to a energy interval in the energetic layer based on its *score*4 energy. We then map d_i to a cube in the geometric layer residing in the energy interval based on its shape-similarity features. If d_i fall on an empty cube, it is stored in the cube. If there is already a

conformation (d_j) in the cube, we replace d_j with d_i in the cube if d_i has a lower energy than d_j; otherwise, d_i is excluded from the map.

2.3 Guiding with the Map

The selection operator in HEA is modified to allow consultation with the map to select individuals for the next generation. How often this consultation happens is governed by the consultation frequency f. In all the generations the map is not consulted, the selection operator works the same as the selection operator in HEA.

In any generation g, let the f generation earlier version of the ever-evolving map be denoted as MAP_{g-f}. The selection mechanism checks the MAP_{g-f} in every f generations during selection. The map consulted is always the f generation earlier version to provide the individuals in the MAP_{g-f} enough opportunities to reproduce and improve. This enables the algorithm to exploit the conformation space around these individuals. Starting with an empty selection pool, during each consultation, the parents and offspring that fall on empty cubes in the MAP_{g-f} are added to the selection pool. The parents and offspring that fall on already occupied cubes are excluded from the selection pool as the conformation space around these individuals have already been explored.

After all the individuals are checked, two scenarios can occur. First, the selection pool contains more individuals than the population size. In this case, we apply truncation selection to bring the selection pool down to the population size. Second, the selection pool contains less individuals than the population size. In this case, we randomly select the rest of the individuals from the map and apply molecular fragment replacement of length 9 on them once to have bigger structural change for exploration in the unknown parts of the landscape and get more diverse conformations.

When the above process is completed, the selection pool contains the same number of individuals as the population size. These individuals constitute the next generation.

2.4 Implementation Details

The population size p is set to 100 and the elitism rate n for elitism-based truncation selection is set to 25%, as in [14]. As is commonly done for evolutionary algorithms, the termination criterion is set to a total budget of fitness/energy evaluations. The algorithm is executed for a fixed budget of 10M energy evaluations. The consultation frequency f is set to 15. The algorithm is implemented in python and interfaces with the PyRosetta library. The algorithm runs for $2-5$ hours on one Intel Xeon E5-2670 CPU with 2.6GHz base processing speed and 100GB of RAM. The runtime differs mainly because of different lengths of the target proteins. The algorithm is run 5 times on each target to account for the variance due to stochasticity.

3 Results

3.1 Experimental Setup

We carry out our evaluation on two datasets. The first is a benchmark dataset, introduced in [11] and complemented with more targets [14, 4, 22]. The dataset contains 10 target proteins of varying lengths (varying from 53 to 123 amino acids) and folds (α, β, and $\alpha+\beta$). The second dataset consists of 10 hard, free-modeling targets from CASP12 and CASP13. This CASP dataset has also been used in recent work [24, 25, 21].

We refer to the algorithm described in Section 2 as HEA-Map. HEA-Map is compared to HEA for a baseline comparison. We also compare HEA-Map to two other state-of-the-art decoy generation algorithms. One is Rosetta's Simulated Annealing Metropolis Monte Carlo (SA-MMC) based decoy sampling algorithm. The other is a recent subpopulation EA, SP-EA$^+$ [20], that aims to prevent premature convergence and retain diversity during optimization by evolving and maintaining multiple subpopulations.

The HEA-Map, HEA, and SP-EA$^+$ algorithms are run 5 times on each target sequence, and what we report here is the best performance over 5 runs combined. Each run exhausts a fixed computational budget of $10M$ energy evaluations for a total of $50M$ energy evaluations for the 5 runs. Rosetta is run for $54M$ energy evaluations. As is practice in EAs for PSP evaluation [17], performance is measured by lowest reached energy and the lowest reached distance to the known native conformation of the target. The later is important because lower energies do not necessarily correlate with proximity to the native conformation. We use a popular proximity measure least root-mean-squared-deviation (lRMSD) [9]. After a decoy conformation and a given native conformation are optimally superimposed to remove differences due to rigid-body motions in 3D (rotations and translations), lRMSD measures the Euclidean distance averaged over the atoms under comparison; a lower score indicates a better proximity. In template-free PSP, the comparison typically focuses on the main carbon atoms or the CA atoms.

To present a principled evaluation, we further strengthen our comparison with statistical significance tests. We utilize Fisher's [6] and Barnard's [1] exact tests for this purpose. Although Fisher's conditional test is widely adopted for statistical significance, Barnard's unconditional exact test is generally considered more powerful than Fisher's test for 2x2 contingency matrices.

Finally, to provide a complete picture and measure how much better or worse performance is achieved on each target, we also employ performance profiles [5]. Performance profiles show the cumulative distribution functions for different performance ratios for a evaluation metric that reveal major performance characteristics. Let us briefly summarize the concept of performance profiles. Performance profiles provide us with a way of depicting how frequently a particular algorithm is within some distance of the best algorithm for a particular problem instance/target. So, for each problem instance, we first compute the best method, and then for every other method, we determine how far they are from optimal. We vary the *performance ratio* (pr) over a range for this analysis. Specifically, for a given pr, *measure reached* means that an algorithm comes within a factor of pr of the best measure over all algorithms on a given target. The number of targets where an algorithm does this is tallied up, and this becomes indicative of its performance, also referred to as number of problems *solved*, at a given performance ratio. In our case, problem instances are our targets in the dataset in consideration.

3.2 Evaluation on Benchmark Dataset

Table 1 shows the lowest *score*4 energy reached by each of the algorithms under comparison on the benchmark dataset. Table 1 shows that HEA-Map achieves lower energy than all other algorithms in 8/10 cases. In a head-to-head comparison, HEA-Map beats all other algorithms comfortably and achieves lower energy than Rosetta in 9/10 cases, than HEA in 8/10 cases, and than SP-EA$^+$ in 9/10 cases. Table 5(a) evaluates the 1-sided statistical significance tests of the performance of HEA-Map over the other algorithms. Table 5(a) shows that the performance improvements are statistically significant at 95% confidence level (p-values < 0.05) for both Fisher's and Barnard's tests.

Table 2 shows the lowest lRMSD to the native conformation reached by each of the algorithms under comparison on the benchmark dataset. Table 2 shows that HEA-Map achieves

Table 1: Comparison of the lowest energy obtained by each algorithm under comparison on each of the 10 benchmark targets is shown in Columns 4-7. The PDB ID of the known native, sequence length, and fold of each target are shown in Columns 1-3. The lowest energy value reached per target is marked in bold.

			Lowest Energy (REUs)			
PDB ID	Length	Fold	Rosetta	HEA	SP-EA$^+$	HEA-Map
1ail	73	α	−29.9	−56.1	−81.3	**−84.7**
1bq9	53	β	−46.9	−50.5	−64.2	**−71.1**
1c8ca	64	β	−101.4	−86.4	−78.3	**−105.7**
1cc5	83	α	−82.5	−68.6	−76.4	**−93.7**
1dtja	76	$\alpha+\beta$	−72.5	−82.2	−72.6	**−90.9**
1hhp	99	β	**−106.3**	−104.5	−83.5	−81.4
2ci2	83	$\alpha+\beta$	−37.8	**−109.8**	−82.7	−108.8
2ezk	93	α	−51.1	−100.7	−135.2	**−138**
2h5nd	123	α	−82.5	−129	−139.1	**−161.9**
3gwl	106	β	−68.2	−100	−117.8	**−133.7**

lower lRMSD than all other algorithms in 6/10 cases. In a head-to-head comparison, HEA-Map beats all other algorithms comfortably and achieves lower lRMSD than Rosetta in 7/10 cases, than HEA in 9/10 cases, and than SP-EA$^+$ in 7/10 cases. Table 5(b) evaluates the 1-sided statistical significance tests of the performance of HEA-Map over the other algorithms. Table 5(b) shows that the performance improvement over HEA is statistically significant at 95% confidence level (p-values < 0.05) for both Fisher's and Barnard's tests. Performance improvement over Rosetta and SP-EA$^+$ are not statistically significant at 95% confidence level but the p-values are close to 0.05.

Table 2: Comparison of the lowest lRMSD to the native conformation obtained by each algorithm under comparison on each of the 10 benchmark targets is shown in Columns 4-7. The PDB ID of the known native, sequence length, and fold of each target are shown in Columns 1-3. The lowest lRMSD value reached per target is marked in bold.

			Lowest lRMSD (Å)			
PDB ID	Length	Fold	Rosetta	HEA	SP-EA$^+$	HEA-Map
1ail	73	α	4.5	1.4	**1.2**	1.4
1bq9	53	β	2.9	3	4.7	**2.8**
1c8ca	64	β	**2.2**	4.8	3.6	3.7
1cc5	83	α	**3.7**	4.7	4.7	4.4
1dtja	76	$\alpha+\beta$	**2.3**	4.2	2.5	2.8
1hhp	99	β	10.1	8.8	8.2	**7.8**
2ci2	83	$\alpha+\beta$	5.8	3.7	3.5	**3.3**
2ezk	93	α	3.6	3.4	2.9	**2.7**
2h5nd	123	α	7.4	6.2	7.4	**5.3**
3gwl	106	β	5.8	5.4	2.9	**2.7**

Figure 1(a) shows the performance profiles of each algorithm over the benchmark dataset in terms of the lowest energy reached. Figure 1(a) shows that the probability of HEA-Map to be the optimal algorithm among these 4 algorithms is about 0.80, considerably more than any of the other algorithms. At $pr = 1.1$, HEA-Map succeeds for 90% targets. HEA-Map and SP-EA$^+$ reaches a success of 100% at a $pr = 1.4$, while HEA do so at $pr = 1.6$. Rosetta's performance profile rises very slowly and reaches 100% at $pr = 3.0$. Figure 1(b) relates a similar analysis

focusing on the lowest lRMSD to the native conformation and shows that the probability of HEA-Map to be the optimal algorithm among these 4 algorithms is about 0.60, considerably more than any of the other algorithms. At $pr = 1.2$ and $pr = 1.3$, HEA-Map succeeds for 80% and 90% targets respectively. HEA-Map and SP-EA$^+$ reaches a success of 100% at a $pr = 1.7$, while HEA do so at $pr = 2.2$. Rosetta saturates at $pr = 2.2$ with a success for 90% targets. These results clearly establish HEA-Map as the superior algorithm.

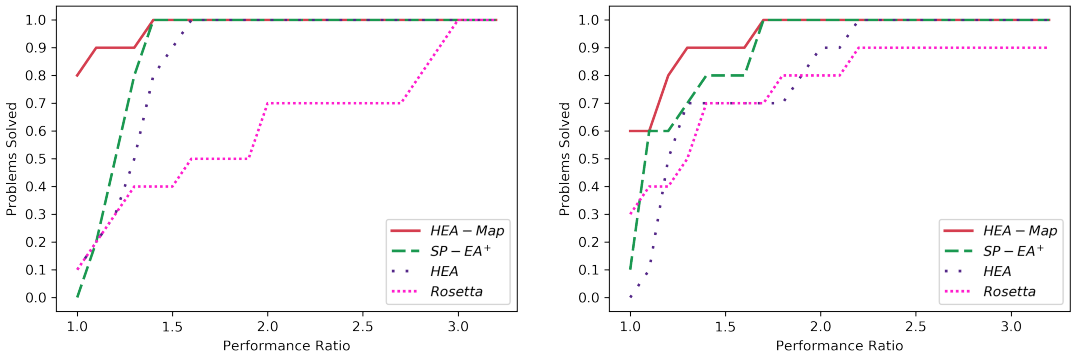


Figure 1: Performance profiles for the algorithms on (a) lowest energy and (b) lowest lRMSD metrics on the benchmark dataset.

These results show the utility of guidance by the map for conformation sampling. The superior performance of HEA-Map suggests the algorithm is able to sample from the parts of the conformation space missed by the algorithms that does not use the map to enhance exploration. The quality of the conformations obtained by HEA-Map is shown qualitatively in Fig. 3, which draws the lowest-lRMSD conformation obtained by HEA-Map (drawn in blue) in three selected targets, superimposing it over the known native (drawn in olive). Rendering is performed with the CCP4mg molecular graphics software [10].

3.3 Evaluation on CASP Dataset

Table 3 shows the lowest *score*4 energy reached by each of the algorithms under comparison on the CASP dataset. Table 3 shows that HEA-Map achieves lower energy than all other algorithms in 7/10 cases. In a head-to-head comparison, HEA-Map beats all other algorithms easily and achieves lower energy than Rosetta in 9/10 cases, than HEA in all cases, and than SP-EA$^+$ in 8/10 cases. Table 5(c) evaluates the 1-sided statistical significance tests of the performance of HEA-Map over the other algorithms. Table 5(c) shows that the performance improvements are statistically significant at 95% confidence level (p-values < 0.05) for both Fisher's and Barnard's tests.

Table 4 shows the lowest lRMSD to the native conformation reached by each of the algorithms under comparison on the benchmark dataset. Table 4 shows that HEA-Map achieves lowest lRMSD in 9/10 cases. In a head-to-head comparison, HEA-Map beats all other algorithms comfortably and achieves lower lRMSD than Rosetta in 9/10 cases, than HEA in all cases, and than SP-EA$^+$ in 8/10 cases. Table 5(d) evaluates the 1-sided statistical significance tests of the performance of HEA-Map over the other algorithms. Table 5(d) shows that the performance improvements are statistically significant at 95% confidence level (p-values < 0.05) for both Fisher's and Barnard's tests.

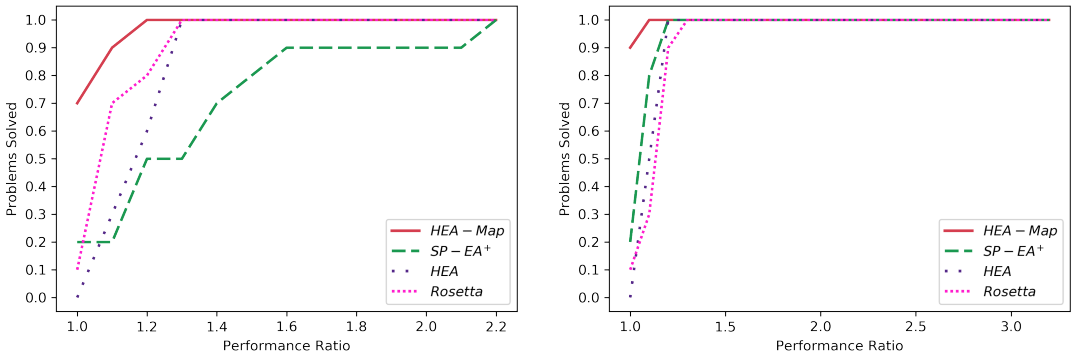
Table 3: Comparison of the lowest energy to the native conformation obtained by each algorithm under comparison on each of the 10 CASP targets is shown in Columns 3-6. The CASP ID of the native and the sequence length of each target are shown in Columns 1-2. The lowest energy value reached per target is marked in bold.

		Lowest Energy (REUs)			
Domain	Length	Rosetta	HEA	SP-EA$^+$	HEA-Map
T0859-D1	129	−99.5	−88	−92.4	**−103**
T0886-D1	69	**−89.2**	−69.9	−41.4	−83
T0892-D2	110	−101.8	−116.3	−76.7	**−120.8**
T0897-D1	138	−141.4	−135.2	−138.8	**−152.9**
T0898-D2	55	−65.5	−65.7	−51	**−70.1**
T0953s1-D1	67	−51.8	−55.8	**−67**	−60.7
T0953s2-D3	93	−53.1	−62.2	−44.5	**−66.3**
T0957s1-D1	108	−121.5	−102.6	−111.2	**−124.3**
T0960-D2	84	−79.7	−67.6	−63.2	**−87.5**
T1008-D1	77	−164.2	−148.4	**−170.9**	−167

Table 4: Comparison of the lowest lRMSD to the native conformation obtained by each algorithm under comparison on each of the 10 CASP targets is shown in Columns 3-6. The CASP ID of the native and the sequence length of each target are shown in Columns 1-2. The lowest lRMSD value reached per target is marked in bold.

		Lowest lRMSD (Å)			
Domain	Length	Rosetta	HEA	SP-EA$^+$	HEA-Map
T0859-D1	129	10.6	9.6	9.2	**9.1**
T0886-D1	69	6.3	6.4	6.2	**5.8**
T0892-D2	110	8	7.2	**6.7**	6.8
T0897-D1	138	9	9.3	8.4	**8.1**
T0898-D2	55	6.5	6.1	**5.8**	**5.8**
T0953s1-D1	67	7	6.2	5.7	**5.6**
T0953s2-D3	93	8.7	8	8	**7.6**
T0957s1-D1	108	6.9	7.4	7.2	**6.2**
T0960-D2	84	**7.2**	7.6	7.3	**7.2**
T1008-D1	77	3.2	3.6	3.6	**3**

Figure 2(a) shows the performance profiles of each algorithm over the CASP dataset in terms of the lowest energy reached. Figure 2(a) shows that the probability of HEA-Map to be the optimal algorithm among these 4 algorithms is about 0.70, considerably more than any of the other algorithms. At $pr = 1.1$, HEA-Map succeeds for 90% targets. HEA-Map and SP-EA$^+$ reaches a success of 100% at a $pr = 1.2$, while HEA and Rosetta do so at $pr = 1.3$. The performance profile of SP-EA$^+$ rises very slowly and reaches 100% at $pr = 2.2$. Figure 2(b) relates a similar analysis focusing on the lowest lRMSD to the native conformation and shows that the probability of HEA-Map to be the optimal algorithm among these 4 algorithms is about 0.90, considerably more than any of the other algorithms. HEA-Map reaches a success of 100% at a $pr = 1.1$, while SP-EA$^+$ and HEA do so at $pr = 1.2$. Rosetta reaches 100% at $pr = 1.3$. These results agree with the results in the benchmark dataset and emphasizes the effectiveness of guidance by the map to achieve more exploration of the energy landscape. The superior ability of HEA-Map to sample lower energy regions in the landscape also translates into better quality conformations closer to the native conformations.



(a) (b)

Figure 2: Performance profiles for the algorithms on (a) lowest energy and (b) lowest lRMSD metrics on the CASP dataset.

Table 5: Comparison of HEA-Map to other algorithms via 1-sided Fisher's and Barnard's tests. The tests evaluate the null hypothesis that HEA-Map does not achieve (a) lower lowest energy on benchmark dataset, (b) lower lowest lRMSD on benchmark dataset, (c) lower lowest energy on CASP dataset, (d) lower lowest lRMSD on CASP dataset, considering each of the other algorithms in turn. P-values less than 0.05 are marked in bold.

(a)

Test	Rosetta	HEA	SP-EA$^+$
Fisher's	**0.0005467**	**0.01151**	**0.0005467**
Barnard's	**0.0002012**	**0.005909**	**0.0002012**

(b)

Test	Rosetta	HEA	SP-EA$^+$
Fisher's	0.08945	**5.95e-05**	0.08945
Barnard's	0.05789	**2.00e-05**	0.05789

(c)

Test	Rosetta	HEA	SP-EA$^+$
Fisher's	**0.0005467**	**5.41e-06**	**0.01151**
Barnard's	**0.0002012**	**9.54e-07**	**0.005909**

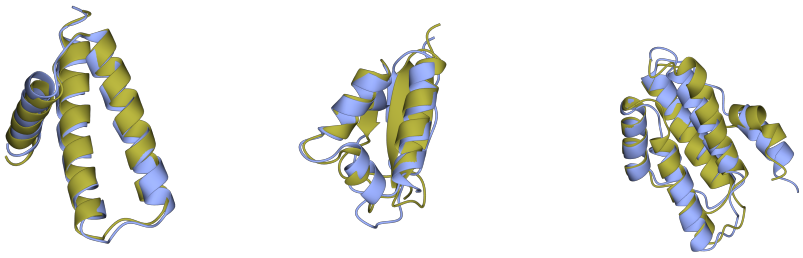
(d)

Test	Rosetta	HEA	SP-EA$^+$
Fisher's	**5.95e-05**	**5.41e-06**	**0.002739**
Barnard's	**2.00e-05**	**9.54e-07**	**0.001288**

4 Conclusion

In this paper, we present an EA that is guided by a map of the already explored parts of the conformation space. The EA is able to sample from unexplored regions of the conformation space through periodically excluding sampled individuals from selection and generating reasonably different new individuals. The results presented in the previous section demonstrates the effectiveness of the proposed EA for sampling better quality conformations and shows the potential of such mechanisms to enhance exploration. This work opens up a promising direction for further research. Future work will investigate the use of such maps to guide other conformation sampling algorithms in addition to the single-objective EA employed here and other ways to guide sampling with a memory of the conformation space.

It is worth noting that, recently, significant focus is placed on deep learning frameworks [16, 19], most prominently represented by AlphaFold2 [7], which leverage strong inductive bias to generate one high-quality conformation of a target protein sequence. The analysis in [18]



1ail (lRMSD = 1.4Å) 1dtja (lRMSD = 2.8Å) 3gwl (lRMSD = 2.7Å)

Figure 3: The conformation obtained by HEA-Map that is closest to the native conformation is shown for three selected cases, the protein with known native conformation under PDB ID 1ail (left), 1dtja (middle), and 3gwl (right). The HEA-Map conformation is in blue, and the known native conformation is in olive.

confirms that conformations generated for a variety of proteins are of high-quality but of varying confidence over regions of the protein sequence. Our own analysis (and that of others, data not shown) confirms that even though the AlphaFold2-advance implementation from ColabFold [12] allows generating a few dozen conformations for an input amino-acid sequence, the set is homogeneous and does not capture the possible diversity of the conformation space. We believe that conformation sampling algorithms and, in particular, evolutionary algorithms, which allow balancing between exploration and exploitation, are worth exploring and advancing to provide a broader view of the conformation space for a richer understanding of structure-based mechanisms and protein biological function.

Acknowledgments

This work is supported in part by NSF Grant No. 1900061. This material is additionally based on work by AS while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. 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. A new test of 2x2 tables. *Nature*, 156:177, 1945.

[2] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol*, 5(11):789–96, 2009.

[3] D. D. Boehr and P. E. Wright. How do proteins interact? *Science*, 320(5882):1429–1430, 2008.

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

[5] Elizabeth Dolan and Jorge Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91, 03 2001.

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

[7] J. Jumper, R. Evans, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021.

[8] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comp. Biol.*, 12(4):e1004619, 2016.

[9] A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst A*, 26(6):656–657, 1972.

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

[11] Jens Meiler and David Baker. Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences of the USA*, 100(21):12105–12110, 2003.

[12] Milot Mirdita, Sergey Ovchinnikov, and Martin Steinegger. Colabfold - making protein folding accessible to all. *bioRxiv*, 2021.

[13] B. Olson, K. A. De 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.

[14] 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.

[15] B. Olson and A. Shehu. Multi-objective optimization techniques for conformational sampling in template-free protein structure prediction. In *Intl Conf on Bioinf and Comp Biol (BICoB)*, pages 143–148, Las Vegas, NV, 2014.

[16] Andrew W. Senior, Richard Evans, John Jumper, et al. Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (casp13). *Proteins: Structure, Function, and Bioinformatics*, 87(12):1141–1148, 2019.

[17] A. Shehu. A review of evolutionary algorithms for computing functional conformations of protein molecules. In W. Zhang, editor, *Computer-Aided Drug Discovery*, Methods in Pharmacology and Toxicology. Springer Verlag, 2015.

[18] K. Tunyasuvunakool, J. Adler, Z. Wu, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596:590—596, 2021.

[19] J. Xu, M. McPartlon, and J. Lin. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Mach Intel*, 3:601–609, 2020.

[20] A. Zaman, K. A. De Jong, and A. Shehu. Using subpopulation eas to map molecular structure landscapes. In *Proceedings of The Genetic and Evolutionary Computation Conference (GECCO)*, pages 1–8, New York, NY, 2019. ACM.

[21] A. Zaman, P. Kamranfar, C. Domeniconi, and A. Shehu. Decoy ensemble reduction in template-free protein structure prediction. In *Proceedings of the 10th ACM Intl Conf on Bioinf and Comput Biol (BCB)*, pages 562–567, Niagara Falls, NY, 2019.

[22] A. Zaman, P. Kamranfar, C. Domeniconi, and A. Shehu. Reducing ensembles of protein tertiary structures generated de novo via clustering. *Molecules*, 25(9):2228, 2020.

[23] A. Zaman, P. Parthasarathy, and A. Shehu. Using sequence-predicted contacts to guide template-free protein structure prediction. BCB '19, page 154–160, New York, NY, USA, 2019. Association for Computing Machinery.

[24] A. Zaman and A. Shehu. Balancing multiple objectives in conformation sampling to control decoy diversity in template-free protein structure prediction. *BMC Bioinformatics*, 20(1):211, 2019.

[25] A. Zaman and A. Shehu. Building maps of protein structure spaces in template-free protein structure prediction. *Journal of Bioinformatics and Computational Biology*, 17(06):1940013, 2019.

[26] A. Zaman and A. Shehu. Equipping decoy generation algorithms for template-free protein structure prediction with maps of the protein conformation space. In *11th Intl Conf on Bioinf and Comput Biol (BICoB)*, volume 60 of *EPiC Series in Computing*, pages 161–169. EasyChair, 2019.