A Population-based Evolutionary Algorithm for Sampling Minima in the Protein Energy Surface

Sameh Saleh[1] and Brian Olson[1] and Amarda Shehu[1,2,3*]
[1]*Department of Computer Science*
[2]*Department of Bioinformatics and Computational Biology*
[3]*Department of Bioengineering*
George Mason University
Fairfax, VA, 22030, United States
ssaleh2@gmu.edu, bolson3@gmu.edu, amarda@gmu.edu
**Corresponding Author*

***Abstract*—Obtaining a structural characterization of the biologically active (native) state of a protein is a long standing problem in computational biology. The high dimensionality of the conformational space and ruggedness of the associated energy surface are key challenges to algorithms in search of an ensemble of low-energy decoy conformations relevant for the native state. As the native structure does not often correspond to the global minimum energy, diversity is key. We present a memetic evolutionary algorithm to sample a diverse ensemble of conformations that represent low-energy local minima in the protein energy surface. Conformations in the algorithm are members of an evolving population. The molecular fragment replacement technique is employed to obtain children from parent conformations. A greedy search maps a child conformation to its nearest local minimum. Resulting minima and parent conformations are merged and truncated back to the initial population size based on potential energies. Results show that the additional minimization is key to obtaining a diverse ensemble of decoys, circumvent premature convergence to suboptimal regions in the conformational space, and approach the native structure with lRMSDs comparable to state-of-the-art decoy sampling methods.**

***Keywords*-evolutionary computation; local minima; near-native conformations; greedy local search; molecular fragment replacement; protein native state.**

I. Introduction

Characterizing the biologically active or "native" three-dimensional structure of a protein from sequence alone remains a central challenge in computational structural biology [1]. The strong relationship between structure and biological function makes determination of protein structure critical to our understanding of biological processes. Experimental techniques for determining structure, such as X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy are time consuming and expensive and have failed to keep up with the exponential growth in protein sequence data. There is a growing need for computational methods to complement wet-lab efforts and improve the state of protein modeling across diverse applications [2]–[4].

A thermodynamic view of the protein conformational space posits that the native state is that of the lowest free energy and is in itself an ensemble of conformations [5]. A working assumption in the early days of ab-initio protein structure prediction was to regard the native state as a homogeneous ensemble of little conformational diversity; hence, entropy was disregarded, and instead the objective became locating the global minimum of the potential energy surface (the Global Minimum Energy Conformation - GMEC) with optimization techniques [6]. Computing the GMEC has proven to be NP-hard, since the size of the space of all possible conformations of a protein chain grows exponentially with chain length [7]. In addition, each conformation has a potential energy, yielding a funnel-like energy surface rich in local minima [8], [9]. To add to the complexity, some minima may be artifacts introduced by the potential energy functions available to us [10].

The high-dimensional protein conformational space and rugged potential energy surface require powerful optimization frameworks that often draw inspiration from diverse fields [11]. Powerful stochastic optimization strategies to better navigate the conformational space include Metropolis Monte Carlo (MMC) [12], Basing Hopping (BH) [13], and Evolutionary Strategies (ES) [14]. Some of these find their origins in the field of evolutionary computation. On the other hand, the most successful structure prediction protocols nowadays incorporate domain-specific knowledge originating in the computational biology community. Most salient developments in this community include the design of coarse-grained energy functions and protein chain representations and the molecular fragment replacement technique to simplify the conformational space and energy surface.

Evolutionary Algorithms (EA) are a class of powerful stochastic search and optimization algorithms developed in the evolutionary computation community to tackle an array of difficult search and optimization problems. An EA explores the protein conformational space by evolving a

978-1-4673-2747-3/12/$31.00 ©2012 IEEE

representative population of conformations towards a goal state, in this case the lowest potential energy. In a Memetic EA (MEA), the global EA search is combined with a short minimization phase. This approach allows explicitly probing local minima in a rugged energy surface by projecting each move at the global level to a nearby local minimum. Many studies have applied EAs to the problem of protein structure prediction [15]–[21]. Most fail to employ state-of-the-art domain-specific techniques, such as coarse graining and molecular fragment replacement, which have been proven effective in ab-initio structure prediction. In addition, current EA applications typically pursue optimization with the goal of finding the GMEC rather than examining how well the native structure is actually recovered. As a result, studies employing evolutionary strategies for protein structure prediction are restricted to very small molecules or toy models.

Advances in the computational structural biology community focus on simplifying and discretizing the search space. An emerging template among ab-initio structure prediction protocols combines a coarse-grained stochastic search with a second stage of fine-grained refinement [2], [22], [23], [23]–[27]. Stage one identifies a broad range of local minima in the energy surface. The hope is that if one of these is in the vicinity of the true native basin, heavy-duty refinement in the second stage will yield convergence to the native state.

In this two-stage protocol, the goal of stage one is no longer to realize the GMEC, as coarse graining often distorts the underlying energy surface [28]. Rather, the goal is capture a broad range of local minima. However, common approaches to stage one do not explicitly sample local minima. A typical approach is to run many independent MMC or molecular dynamics trajectories in parallel to sample a large number of low-energy "decoy" conformations. These decoys are then grouped by geometric similarity to reveal explored local minima in the energy surface.

In this paper we present an MEA algorithm to obtain an informative ensemble of low-energy decoys that represent low-energy minima in the protein energy surface. The contribution of this work is two-fold. First we combine evolutionary search strategies with a state-of-the-art coarse-grained chain representation and energy function and the molecular fragment replacement technique. The result is that a simple EA reaches its potential in generating useful decoys for the first stage of ab-initio structure prediction and is not limited to toy protein models. Its effectiveness is extended to small-to-medium length proteins, and its results are comparable to other conformational search methods proposed for stage one in literature.

We then show in realistic case studies and in great detail that the simple EA framework is highly exploitative. While it may by chance sample conformations near the native structure, its greedy nature makes it prone to early convergence to sub-optimal minima. While traditional optimization algorithms in the evolutionary computing community focus on finding the GMEC, this goal is not sufficient in the context of ab-initio structure prediction. Near-native conformations may often not be associated with the lowest-energy minima in the surface reconstructed by coarse-grained energy functions. The issue has been well-studied in the computational structural biology community and is the main reason why *ab-initio* structure prediction methods focus their sampling to obtain a broad view of the energy surface through diverse low-energy decoys. Here we show that an MEA framework is capable of offering such a view. Moreover, by sampling minima explicitly, the framework is more effective than methods that rely on clustering of sampled decoy conformations to detect minima. Our analysis shows that the MEA framework is less prone to convergence and obtains a broad view conformational space.

Section II describes the proposed EA and MEA frameworks. Results presented in section III show how the addition of a minimization step in MEA improves the ability to sample conformations near the native state. Results from the MEA are then compared compared to those obtained by MMC-based frameworks developed by us and others. Finally, section IV summarizes the findings in this work.

II. Methods

We first relate details on how a simple EA framework can be combined with domain-specific knowledge, such as representations, energy functions, and molecular fragment replacement. This combination, which is shown to extend applicability and effectiveness to small-to-medium size proteins is the first contribution of the work presented here.

The main steps of the basic EA are shown in Algorithm 1. In summary, the EA employs an evolutionary computing model to optimization. It essentially treats conformations as ever-improving solutions of a fitness function (in our case a state-of-the-art coarse-grained potential energy function described later in this section). Starting from conformations sampled at random for a given protein sequence α, the population P of conformations is evolved over a series of iterations or generations to obtain better, more fit, solutions to the fitness function (lines 2-6 in Algo 1).

The population has a constant capacity `PopSize` maintained through each generation. In each generation, a set of new children conformations C are sampled based on the current population (Algo 1, line 3). The children compete with the conformations in the current population P for survival (Algo 1, line 4). The surviving conformations that make up the evolved population are added to the running ensemble Ω_α of conformations that we analyze in section III.

Details now follow for each of the components of this basic EA framework and its modification to obtain an MEA framework that incorporates minimization.

Algo. 1 A canonical Evolutionary Algorithm (EA) is shown.

Input: Amino acid sequence, population size, and number of offspring as α, PopSize, and NumChild.
Output: the set of sampled of conformations, Ω_α

1: P $\leftarrow$ INITIALPOPULATION(PopSize)
2: **while** Eval$_{count}$ < Eval$_{max}$ **do**
3: $C \leftarrow$ SAMPLECONFORMATIONS(P,NumChild)
4: $P_{new} \leftarrow$SELECTPOPULATION(P,C)
5: $\Omega_\alpha \leftarrow \Omega_\alpha \cup P_{new}$
6: $P \leftarrow P_{new}$

A. A Basic EA Framework to Sample Low-energy Protein Conformations

In each generation, a set of `NumChild` children conformations are obtained by modifying parent conformations selected from the current population P. Parent conformations are selected from P using fitness-proportional selection, whereby conformations with lower energies have a higher probability of being selected as parents based on a linear weighting of energies. As a result, conformations with lower energies will be selected more often to spawn children (a parent can be selected more than once).

A new child conformation is sampled by modifying a parent through the molecular fragment replacement technique [29], [30]. A residue position i is sampled uniformly at random over the chain of the parent conformation. A fragment $[i, i+2]$ of 3 consecutive amino acids is then defined over the chain. The 6 ϕ, ψ backbone dihedral angles corresponding to this fragment in the parent conformation are then replaced with a configuration of 6 angles sampled uniformly at random over configurations recorded for that fragment in a fragment configuration library. The library stores fragment configurations extracted from known protein native structures. Therefore, the modifications are more likely to result in physically-realistic child conformations. More details on the construction of fragment configuration libraries can be found in [31].

The modification described above is over backbone dihedral angles, as this is the representation of a protein chain employed in this work. This is a well-studied representation in ab-initio structure prediction that allows reducing the dimensionality of the conformational space to $2n$ dimensions for a protein chain of n amino acids.

This backbone representation interfaces with a coarse-grained energy function known as the Associative Memory hamiltonian with Water (AMW). The AMW function has been well-studied by us and others [31]–[35]. AMW is a linear combination of five terms that include Lennard-Jones and hydrogen-bond interactions, and even allows formation of non-local contacts, a hydrophobic core, and water-mediated interactions. Further details can be found in [22], [36]. AMW operates over cartesian coordinates of atoms, which are obtained by accumulating and applying rotations formulated over backbone dihedral angles.

The EA framework summarized in Algo 1 maintains a fixed size population P from one generation to the next. To select a new population P_{new}, the current population is combined with the set of newly sampled children C. The resulting merged set is reduced back to $|P|$ conformations through truncation selection. This is a highly exploitative selection technique that picks the $|P|$ conformations with the lowest energy for the next population.

The molecular fragment replacement technique described above is used not only to obtain children conformations but also to provide the algorithm with a diverse set of low-energy conformations in its initial population. Starting from an extended conformation, $n-2$ fragment configuration replacements are conducted to obtain a random but realistic conformation. This process is repeated $|P|$ times to obtain $|P|$ conformations in the initial population P that seeds the basic EA framework.

B. A Memetic EA Framework to Sample Local Minima

The basic EA described above performs a highly exploitative search progressing towards lower energy conformations in pursuit of the GMEC. As discussed in section I, the goal in ab-initio structure prediction is to search for a diverse set of sufficiently low-energy decoy conformations rather than the absolute minimum due to energy calculation approximations and inaccuracies. Due to this exploitative nature, the population of decoys in the basic EA framework risks converging prematurely to narrow low-energy basins far from the native state. As a result, further exploration of the space and discovery of near-native decoy conformations can be limited in further generational propagation.

To enhance sampling of near-native decoy conformations, we investigate a Memetic EA (MEA) framework that allows explicitly sampling local minima in the protein energy surface. The MEA employs an additional minimization step to map a child conformation sampled by the basic EA framework to a nearby local minimum. In this work, the minimization is implemented as a greedy local search. The greedy search performs a series of fragment replacements only accepting modifications that decrease potential energy. The search continues until n consecutive fragment replacements have been rejected (recall that n is the length of the target protein), indicating that a local minimum has been reached. The corresponding conformation representing the minimum replaces the initial child conformation and is added instead to the set P_{new} of children conformations

It is important to note that children that survive the truncation selection and are therefore members of the evolved population represent local minima in the energy surface. The next iteration may select some of them to be parents. When that happens, a fragment configuration replacement applied to one of them (that is part of the process to generate

new children) is equivalent to a jump out of the current minimum represented by the parent. This resetting is crucial to obtain new nearby minima in the energy surface and reduce the exploitative nature of the basic EA framework. The resetting helps enhance conformational diversity in the MEA, as it essentially jumpstarts the framework with new starting points for minimization in the energy surface.

We note that the definition of a local minimum employed here is only an approximation of the true local minimum in the energy surface. However, previous work shows that, at a coarse-grained level of detail, this working definition is sufficient to sample low-energy conformations near the native state in the context of a basin hopping trajectory [35].

Both the basic and memetic evolutionary algorithms tested and analyzed in this paper are run for a fixed budget of energy evaluations. In the memetic version, each energy evaluation within the minimization contributes to the overall count of the fixed budget of energy evaluations.

III. Results

Both the basic EA and MEA are run on a testbed of 11 target protein systems with known native structures. The results focus on comparing the sampling ability of EA to MEA with respect to lowest energy and proximity to the native structure. Results obtained by MEA are compared to two other state-of-the-art MMC-based frameworks. The results show that the incorporation of domain-specific knowledge in EA make it competitive, but the addition of the minimization step significantly improves the ability of MEA to sample near-native conformations when compared to EA and other MMC-based frameworks.

A. Experimental Setup

Population size $|P|$ is set at 1000 conformations, which is large enough to maintain a structurally-diverse population, but small enough to ensure competition among members in a population. In MEA, $\mathrm{numChild}{=}250$ children are sampled at each generation. Setting $\mathrm{numChild} < |P|$ limits competition among children, thus increasing the percentage of children which make it to the next generation. In the simple EA, $\mathrm{numChild}{=}4000$ children. This is done for the following reason. Since the simple EA does not minimize each conformation, the majority of children will have high potential energies. A larger number of children in the basic EA thus increases the chance of obtaining some low-energy children, allowing for downward exploration of the energy surface in each generation.

Each experiment is repeated for 3 independent runs with a fixed budget of 10 million energy evaluations in each run. Taking up over 90% of CPU time, the calculation of potential energy for each conformation is the computationally intensive task that defines the runtime of the algorithm. Therefore, a fixed budget of energy evaluations maintains a fair comparison between each experiment across a range of target protein systems. In practice, each experiment runs for 30 to 50 hours of CPU time on a 2.4Ghz Core i7 processor, depending on the length of the target protein.

Results are compared by computing the least Root Mean Square Deviation (lRMSD) to the known native structure for each conformation in the output ensemble Ω_α (Ω_α is the union of the populations in each generation). The lRMSD aligns two conformations for comparison to remove differences due to rigid-body transformations and sums distances between corresponding atoms post-alignment.

B. Target Systems of Study

Each experiment is performed on a set of 11 diverse target protein systems, shown in Table I, ranging in size from 61 to 93 amino acids and including α, β, and α/β fold topologies. These targets all have experimentally-determined native structures deposited in the Protein Data Bank (PDB) [37] and so allow measuring the effectiveness of the proposed algorithms. The 11 proteins are selected to allow comparison to published results from other research groups [25].

Table I: Columns $2-4$ show the native PDB ID, number of amino acids and fold topology for each of the 11 target protein systems. Columns 5 and 6 break the fold topology down as the percentage of amino acids which are part of α-helices and β-sheets.

	Native PDB ID	Size	Fold Topology	% α	% β
1	1dtdB	61	α/β	15	46
2	1isuA	62	α/β	15	19
3	1c8cA	64	α/β	22	48
4	1sap	66	α/β	30	44
5	1hz6A	67	α/β	31	42
6	1wapA	68	β	0	62
7	1fwp	69	α/β	30	26
8	1ail	70	α	84	0
9	1aoy	78	α/β	41	10
10	1cc5	83	α	47	4
11	2ezk	93	α	68	0

C. Effectiveness of Minimization in MEA over EA

Table II shows the lowest lRMSD of conformations sampled by the basic EA and MEA. Comparison of columns 3 and 4 reveals that MEA is able to find a conformation at least 0.5Å closer to the native structure than the simple EA for all 11 proteins. This suggests that the addition of minimization enhances the sampling of near-native conformations.

Fig. 1 plots the energy of each conformation against its lRMSD to the native structure on four representative target protein systems. Fig. 1 shows that EA is able to reach much lower-energy levels than MEA. This is expected, because EA will tend to converge to a few basins in the energy surface and then continue to optimize them through a

Table II: The lowest lRMSD to the known native structure over conformations in the output ensemble Ω_α is reported for the 11 target protein systems. The average and minimum lowest lRMSD obtained over 3 independent runs is shown for the Evolutionary Algorithm (EA) in column 3 and the Memetic EA (MEA) in column 4. Column 5 shows values obtained by the MMC-based FeLTr decoy generation method [33], and column 6 shows lowest lRMSDs published the Sosnick lab for the ItFix method [25].

	Native	Avg(Min) Lowest lRMSD (Å)			
	PDB ID	EA	MEA	FeLTr	ItFix
1	1dtdB	8.1(7.3)	6.9(6.7)	7.7(7.6)	6.5
2	1isuA	7.5(7.1)	6.5(6.1)	6.8(6.7)	6.5
3	1c8cA	8.5(8.5)	7.2(6.8)	6.5(6.0)	3.7
4	1sap	8.0(7.2)	6.7(6.2)	7.1(6.5)	4.6
5	1hz6A	6.7(5.5)	6.2(6.0)	6.6(6.6)	3.8
6	1wapA	9.3(8.8)	7.5(6.8)	7.8(7.3)	8.0
7	1fwp	7.6(7.1)	6.8(6.6)	6.8(6.4)	8.1
8	1ail	4.8(4.0)	3.5(3.4)	4.7(4.5)	5.4
9	1aoy	7.0(6.2)	5.3(5.1)	5.1(4.6)	5.7
10	1cc5	7.1(6.4)	5.7(5.5)	6.4(6.2)	6.5
11	2ezk	6.1(5.0)	4.9(4.6)	6.4(6.0)	5.5

highly exploitative search. In contrast, MEA tends to capture a broader view of the energy surface due to its restart mechanism described in section II. This is advantageous, as many low-lRMSD conformations are associated with low energies but do not necessarily populate the lowest-energy regions in the energy surface.

Fig. 1 also shows that MEA yields more conformations with low lRMSDs to the native structure. Its conformational ensemble is more diverse than that of EA, and Fig. 1 indicates that it contains more near-native conformations. The point on diversity is important. In a complete ab-initio structure prediction protocol, computational constraints allow refinement of only a handful of decoy conformations in search of the true native state. While we do not conduct the second-stage refinement here, it is important to consider and compare not only the single lowest-lRMSD conformation from each method, but also the distribution of decoys that might be reasonably passed on to a second stage refinement.

Here we use a simple technique that selects only the 5% lowest-energy conformations from an output ensemble of decoys to create a reduced ensemble of refinement. Fig. 2 shows the distribution of lRMSDs to the native structure of this selected subset of decoys on four representative protein systems (with native PDB ids 1isuA, 1cc5, 1wapA, and 2ezk). The distribution obtained by MEA is superimposed in blue over that obtained by EA in red. On 9 of the 11 systems (Fig. 2 shows results in detail only for 4 selected systems), MEA not only finds the lowest-lRMSD conformation, but also samples significantly more low-lRMSD conformations than the simple EA. For 2 target proteins (1hz6A and 2ezk), the simple EA gets lucky in one of the three runs and converges to an energy basin containing low-lRMSD conformations to the native structure.

Fig. 2 shows that MEA (dashed blue line) not only finds the lowest-lRMSD conformation, but also samples significantly more low-lRMSD conformations than EA (solid red line). Fig. 2d illustrates the rare case when the simple EA samples more low-lRMSD conformations than MEA. These results confirm that, while the simple EA can occasionally get lucky, MEA is, on average, more effective at sampling conformations near the native state.

D. Comparison to state of the art

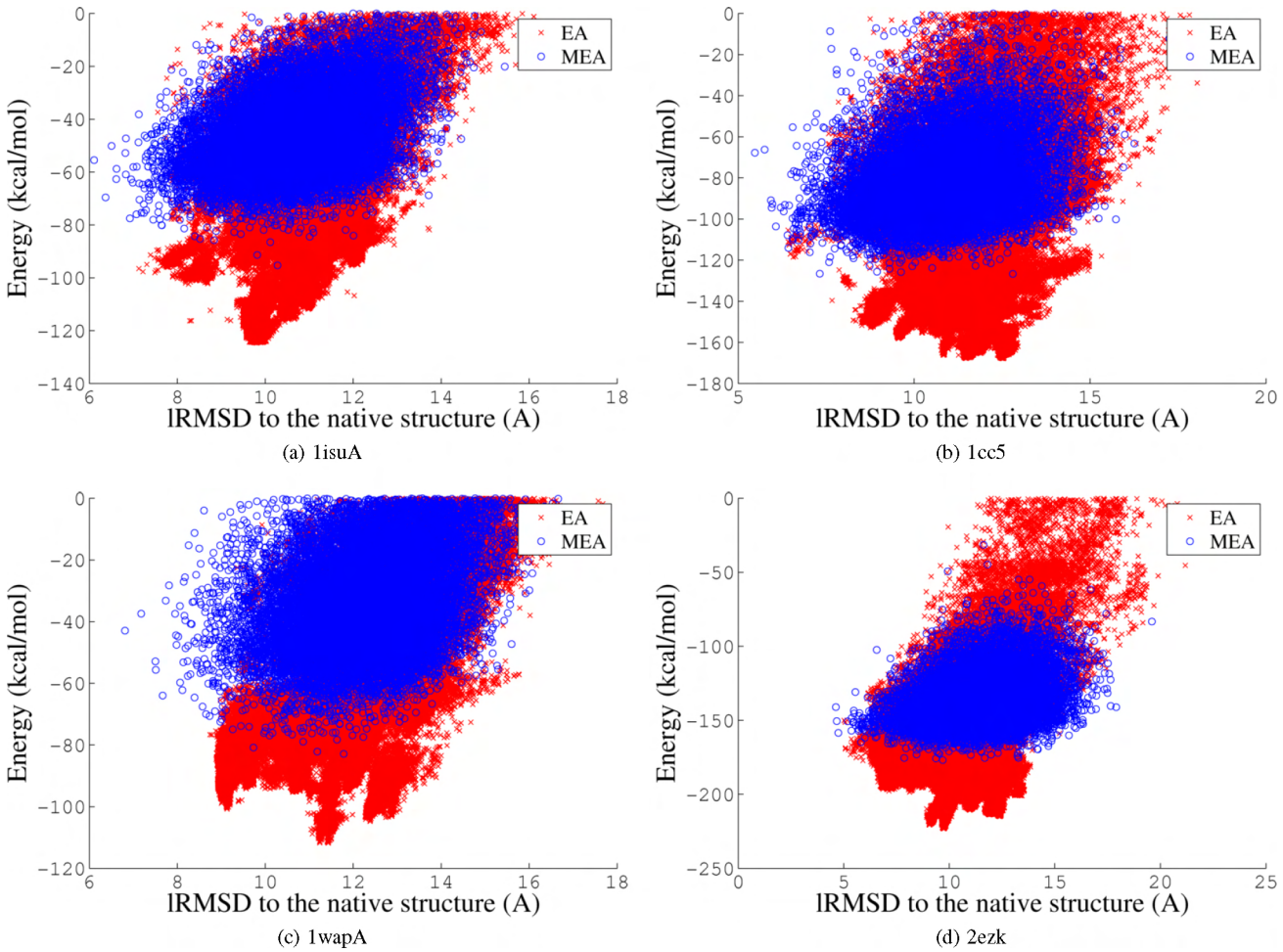
Table II compares results to those published by other decoy sampling methods in ab-initio structure prediction in terms of lowest lRMSD obtained to the native structure. The two methods chosen for comparison, FeLTr [33] and ItFix [25], are both MMC-based. In particular, FeLTr employs the same energy function and fragment replacement technique as the work presented here and thus allows a direct comparison between a state-of-the-art MMC-based algorithm and MEA. Comparison with ItFix, which uses the DOPE energy function [25], allows external validation of MEA as a viable decoy sampling framework for a structure prediction protocol. Results reported by EA, MEA, and FeLTr are averages and minima of lowest obtained lRMSDs over 3 independent runs in order to account for stochasticity.

Comparing columns 4 and 5 of Table II shows that MEA and FeLTr perform comparably in 6 of 11 protein systems. For 4 of the remaining target proteins (with native PDB ids 1dtdB, 1ail, 1cc5, and 2ezk), MEA reaches a significantly lower lRMSD to the native structure than FeLTr, with FeLTr only outperforming MEA in the single case of 1c8cA. This suggests that the minimization step in MEA provides a distinct advantage over MMC-based methods, particularly in the case of longer α-helix proteins.

Columns 4 and 6 in Table II show that MEA is also comparable to ItFix. MEA finds lower lRMSDs for 5 of the 11 target proteins, while ItFix finds lower lRMSDs for 3 proteins; the methods are comparable on the remaining 3 systems. These results are promising, as they show that employing a short greedy local search for minimization can make even a simple EA algorithm competitive with state-of-the-art structure prediction protocols.

IV. Conclusions

This work proposes a Memetic Evolutionary Algorithm, MEA, for sampling local minima in the protein energy surface. The results show that the addition of the minimization step allows the algorithm to more effectively sample near-native conformations than a canonical basic EA framework. The basic EA framework is nonetheless effective at optimizing the AMW energy function and reaches much lower-energy conformations than MEA. However, the simple



(a) 1isuA (b) 1cc5

(c) 1wapA (d) 2ezk

Figure 1: The Potential energy is plotted against lRMSD to the native structure for each conformation in the output ensemble Ω_α obtained during all 3 runs on 4 representative target proteins. The red "X"s are for the simple EA, and the blue circles for the MEA.

EA is highly exploitative, and it converges rapidly on a few particular basins in the energy surface. Once converged, the framework keeps drilling down. Conformations near the native structure are only achieved if the simple EA gets lucky in its convergence to an energy basin near the native state.

The exploitation in MEA, on the other hand, is limited by the greedy local search employed for minimization. MEA quickly reaches a low-energy floor, but then explores a breadth of conformations around that energy level, only gradually reducing the potential energy of its population. This occurs because the greedy local search does not probe too deeply before a fragment replacement allows it to escape a local minimum and jump to a conformation of higher energy. Since the effective moves in MEA are much larger (between local minima) and tend to result in similar energies, this also makes it much less likely for the entire population to converge as in the simple EA; hence, the population maintains diversity better in MEA.

Comparison of MEA to other conformational search methods shows that the addition of the minimization step along with domain-specific techniques from the computational structural biology community make evolutionary search strategies comparable to other state-of-the-art decoy sampling methods for ab-initio protein structure prediction. Comparison of MEA to EA shows that MEA improves diversity over the highly exploitative basic EA framework. Future work will investigate more advanced evolutionary search strategies that encourage greater diversity.

ACKNOWLEDGMENTS

This work is supported in part by NSF CCF No. 1016995 and NSF IIS CAREER Award No. 1144106.

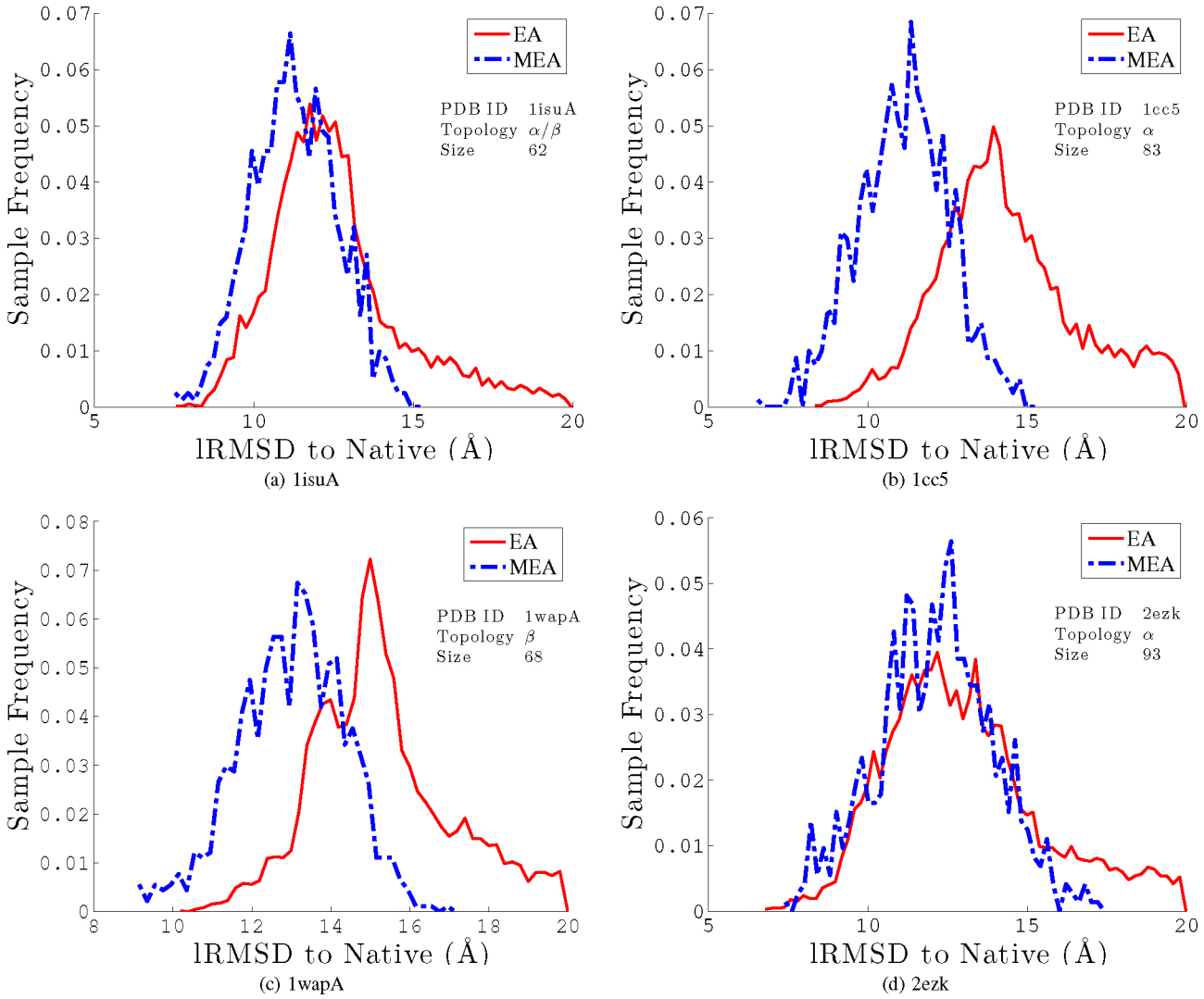


Figure 2: The distribution of lRMSDs to the native structure is shown for a selected ensemble containing the 5% lowest-energy conformations in the output ensemble Ω_α obtained during all 3 runs on 4 representative target proteins. The solid red line shows the distribution obtained by the simple EA, and the dashed blue line shows that obtained by the MEA.

REFERENCES

[1] K. A. Dill, B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev Biophys.*, vol. 37, pp. 289–316, 2008.

[2] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[3] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nat Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[4] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Curr. Opinion Struct. Biol.*, vol. 8, no. 1, pp. 91–97, 2004.

[5] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[6] Z. and S. H. A. Li, "Monte Carlo-minimization approach to the multiple-minima problem in protein folding," *Proc. Natl. Acad. Sci. USA*, vol. 84, no. 19, pp. 6611–6615, 1987.

[7] W. E. Hart and S. Istrail, "Robust proofs of np-hardness for protein folding: General lattices and energy potentials," *J. Comp. Biol.*, vol. 4, no. 1, pp. 1–22, 1997.

[8] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nat. Struct. Biol.*, vol. 4, no. 1, pp. 10–19, 1997.

[9] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Curr. Opinion Struct. Biol.*, vol. 14, pp. 70–75, 1997.

[10] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins: Struct. Funct. Bioinf.*, vol. 65, no. 3, pp. 712–725, 2006.

[11] A. Shehu, "Conformational search for the protein native state," in *Protein Structure Prediction: Method and Algorithms*, H. Rangwala and G. Karypis, Eds. Fairfax, VA: Wiley Book Series on Bioinformatics, 2010, ch. 21.

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[13] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," vol. 101, no. 28, pp. 5111–5116, 1997.

[14] K. A. D. Jong, *Evolutionary Computation : A Unified Approach*. Boston, MA: MIT Press, 2006.

[15] T. Dandekar and P. Argos, "Potential of genetic algorithms in protein folding and protein engineering simulations," *Protein Engineering*, vol. 5, no. 7, pp. 637–645, 1992.

[16] D. P. Djurdjevic and M. J. Biggs, "Ab initio protein fold prediction using evolutionary algorithms: Influence of design and control parameters on performance," *J. Comput. Chem.*, vol. 27, no. 11, pp. 1177–1195, 2006.

[17] A. Bazzoli and A. Tettamanzi, *A memetic algorithm for protein structure prediction in a 3D-lattice HP model*, ser. Lecture notes in computer science ; 3005. Berlin: Springer, 2004, pp. 1 – 10.

[18] M. Islam and M. Chetty, *Novel Memetic Algorithm for Protein Structure Prediction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5866, pp. 412–421.

[19] J. Keum, E.S., K. Kim, and E. Santos, "Local minima-based exploration for off-lattice protein folding," in *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, aug. 2003, pp. 615 – 616.

[20] N. Krasnogor, B. Blackburne, E. Burke, and J. Hirst, "Multimeme algorithms for protein structure prediction," in *Parallel Problem Solving from Nature PPSN VII*, ser. Lecture Notes in Computer Science, J. Guervs, P. Adamidis, H.-G. Beyer, H.-P. Schwefel, and J.-L. Fernndez-Villacaas, Eds. Springer Berlin / Heidelberg, 2002, vol. 2439, pp. 769–778.

[21] M. Mijajlovic and M. Biggs, "On potential energy models for ea-based ab initio protein structure prediction," *Evolutionary Computation*, Jan 2010.

[22] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Funct. Bioinf.*, vol. 76, no. 4, pp. 837–851, 2009.

[23] R. Bonneau and D. Baker, "De novo prediction of three-dimensional structures for major protein families," *J. Mol. Biol.*, vol. 322, no. 1, pp. 65–78, 2002.

[24] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins: Struct. Funct. Bioinf.*, vol. 73, no. 4, pp. 958–972, 2009.

[25] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 10, pp. 3734–3739, 2009.

[26] 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.*, vol. 19, no. 3, pp. 520–534, 2010.

[27] A. Shehu, L. E. Kavraki, and C. Clementi, "Unfolding the fold of cyclic cysteine-rich peptides," *Protein Sci.*, vol. 17, no. 3, pp. 482–493, 2008.

[28] A. Verma, A. Schug, K. H. Lee, and W. Wenzel, "Basin hopping simulations for all-atom protein folding," *J. Chem. Phys.*, vol. 124, no. 4, p. 044515, 2006.

[29] K. F. Han and D. Baker, "Global properties of the mapping between local amino acid sequence and local structure in proteins," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 12, pp. 5814–5818, 1996.

[30] N. Haspel, C. Tsai, H. Wolfson, and R. Nussinov, "Reducing the computational complexity of protein folding via fragment folding and assembly," *Protein Sci.*, vol. 12, no. 6, pp. 1177–1187, 2003.

[31] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *J. Bioinf. and Comp. Biol.*, vol. 9, no. 3, pp. 383–398, 2011.

[32] M. C. Prentiss, D. J. Wales, and P. G. Wolynes, "Protein structure prediction using basin-hopping." *The Journal of Chemical Physics*, vol. 128, no. 22, pp. 225 106–225 106, Jun. 2008.

[33] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–11 227, 2010.

[34] B. S. Olson, K. Molloy, S.-F. Hendi, and A. Shehu, "Guiding search in the protein conformational space with structural profiles," *J Bioinf and Comp Biol*, vol. 10, no. 3, p. 1242005, 2012.

[35] B. Olson and A. Shehu, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Sci*, vol. 10, no. Suppl 1, p. S5, 2012.

[36] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 10, pp. 3352–3357, 2004.

[37] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.