Off-Lattice Protein Structure Prediction with Homologous Crossover

Brian Olson
Dept of Computer Science
George Mason University
4400 University Drive
Fairfax, VA 22030
bolson3@gmu.edu

Kenneth De Jong
Dept of Computer Science
George Mason University
4400 University Drive
Fairfax, VA 22030
kdejong@gmu.edu

Amarda Shehu[*]
Dept of Computer Science
Dept of Bioinformatics
George Mason University
4400 University Drive
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

Ab-initio structure prediction refers to the problem of using only knowledge of the sequence of amino acids in a protein molecule to find spatial arrangements, or conformations, of the amino-acid chain capturing the protein in its biologically-active or native state. This problem is a central challenge in computational biology. It can be posed as an optimization problem, but current top ab-initio protocols employ Monte Carlo sampling rather than evolutionary algorithms (EAs) for conformational search. This paper presents a hybrid EA that incorporates successful strategies used in state-of-the-art ab-initio protocols. Comparison to a top Monte-Carlo-based sampling method shows that the domain=specific enhancements make the proposed hydrid EA competitive. A detailed analysis on the role of crossover operators and a novel implementation of homologous 1-point crossover shows that the use of crossover with mutation is more effective than mutation alone in navigating the protein energy surface.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.6.3 [**Computing Methodologies**]: Simulation and Modeling—*Applications*

General Terms

Algorithms

Keywords

evolutionary algorithms; conformational search; local minima; molecular fragment replacement; protein native state.

[*]Corresponding Author

GECCO'13, July 6-10, 2013, Amsterdam, The Netherlands.

1. INTRODUCTION

Knowledge of the three-dimensional biologically-active or *native structure* of a protein is central due to the promise that structure holds in revealing important information on protein function in the living and diseased cell. The native structure illustrated for a protein molecule in Fig. 1a corresponds to a particular spatial arrangement, or conformation, of the chain of amino-acids that make up a protein molecule. As Fig. 1b shows by zooming on three amino acids of this protein, amino acids are the building blocks of a protein molecule, and they link in a serial fashion to form the protein chain. Structure prediction refers to the problem of finding conformations of the native state. On many proteins, this state is unique, and the term native structure refers to an average over the native conformations.

When no sequence-homologous proteins of known native structures are available to construct a template structure for the protein sequence of interest, ab-initio protocols provide the only route to structure prediction [25]. These protocols use of an energy function for evaluating conformations and an algorithm capable of searching through conformations. The energy function biases the search towards low-energy conformations of the given sequence, based on the thermodynamic hypothesis that places native conformations at the global minimum of the protein energy surface [2, 32]. Yet, inaccuracies in energy functions, inadequate exploration of the conformational space, or a combination of the two are main reasons why ab-initio structure prediction remains challenging [25], particularly for protein chains more than 70 amino-acids long and/or certain native topologies [21, 33, 37].

One of the challenges is that the protein conformational space is vast and high-dimensional. All-atom representations of conformations offer greatest structural detail, but the ensuing conformational space would have $3N$ dimensions for a chain of N atoms and would be infeasible for search. Reduced representations improve feasibility but can be limiting. For instance, representations that limit atoms to a *lattice* cannot capture important structural detail and are not used by ab-initio protocols [6, 8, 17, 5, 36, 37, 27, 39]. Currently, lattice representations are only employed in studies assessing the use of Evolutionary Algorithms (EAs) for conformational search [11, 23, 12, 19]. Current ab-initio protocols employ reduced *off-lattice* representations also known as coarse-grained representations. These model mainly backbone atoms and a designated atom or pseudo-atom to track the location of each side chain. Spatial information for the modeled atoms is obtained through an application of forward kinematics on an internal angular representation that consists of the backbone ϕ, ψ, and ω dihedral angles for each amino acid [6] (see Fig. 1b for an illustration).

In this paper we present a hybrid EA that incorporates successful strategies used in state-of-the-art ab-initio protocols. The main premise for the work presented in this paper is that EAs that incorporate such ab-inito features hold great promise for their ability to improve the state-of-the-art in ab-initio structure prediction.

Our EA incorporates the coarse-grained representation used in the well-known Rosetta ab-initio package and the Rosetta energy function accompanying this representation [24]. This representation, consisting of only the ϕ, ψ, and ω angles, is adopted by many protocols routinely performing at the top of the CASP competition for ab-initio structure prediction [39, 24]. The Rosetta energy function is also considered among the most reliable coarse-grained functions for ab-initio structure prediction [36]. Our hybrid EA also uses the molecular fragment replacement technique which has been a major advance in ab-initio structure prediction [6]. The idea is to obtain a new conformation by replacing the configuration of an entire fragment of the chain in an old conformation with a configuration sampled from a pre-calculated fragment configuration library. To the best of our knowledge, no current EA makes use of coarse-grained representations and molecular fragment replacement.

We analyze in detail the role of reproductive operators, such as mutation and crossover in isolation and combination in sampling low-energy regions. The analysis, conducted over diverse proteins even longer than 70 amino acids, shows that the combination of crossover with mutation enhances sampling over mutation alone. An important final contribution of this paper is the proposal of a new implementation of homologous crossover. Our analysis reveals that homologous crossover allows our hybrid EA to more efficiently sample lower-energy regions than 1-point and 2-point crossover.

The rest of this paper is organized as follows. After a summary of related work in section 1.1, we proceed with a description of our hybrid EA in section 2. Analysis and results are presented in section 3, followed by a discussion in section 4.

1.1 Related Work

State-of-the-art ab-initio protocols split structure prediction in two stages. Typically, a search procedure that launches many independent Metropolis Monte Carlo (MMC) trajectories is used in the first stage is to obtain a broad view of low-energy regions [39, 6, 17]. A broad view is essential when employing a coarse-grained representation and energy function. In particular, the energy function may define an underlying energy surface that is rich in false local minima and may not place native conformations at the lowest-energy basin. These inaccuracies necessitate that the first stage not be as discriminating and retain more than just the lowest-energy conformations. The first stage terminates with end-points of MMC trajectories collected in an ensemble of conformations known as decoys. The second stage, known as selection, uses information on number of neighbors rather than energies through clustering to decide which subset of decoys are relevant for the native state. The operating assumption is that a coarse-grained energy function may preserve the breadth rather than the depth of the native basin. The subset of interest is then further optimized at all-atom structural detail with a more accurate and expensive all-atom energy function. Final energy rankings are used to make predictions on which resulting conformation is most likely to represent the sought native structure.

Great attention is paid to enhancing exploration capability in the first stage [8, 17, 5, 36, 35, 37]. While many search algorithms build on the basic procedure of launching many MMC trajectories by enhancing sampling of the conformational space through temperature schedules, conformation exchanges, deformations of the energy surface, multiscaling, and other techniques [8, 17, 5, 36, 37], other search algorithms explicitly enforce structural diversity among sampled conformations during exploration [35, 31].

While EAs should, in principle, provide high sampling capability, they are not adopted in the first stage in ab-initio protocols. Reasons include not making use of coarse-grained representations and molecular fragment replacement. To date, EAs struggle to report conformations closer than 6Å to the experimentally-determined native structure (see section 2 for details on the distance measure) on all but very short protein chains of up to 25 amino acids [38, 9, 10, 3, 20, 18].

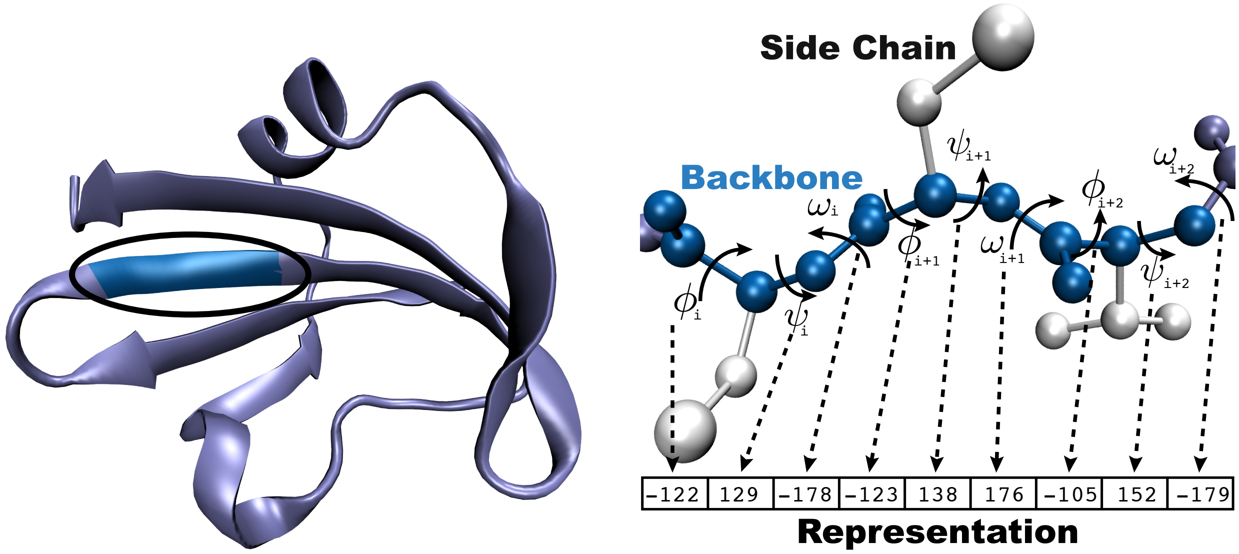
Other key limitations in current EAs are ineffective mutation and crossover operators. Even if one considers avoidance of self collisions as the only energetic constraint on a protein conformation, the likelihood that a conformation sampled at random will be invalid due to self collisions increases dramatically. Moreover, since low-energy conformations corresponding to the native state tend to be highly compact, even a small mutation to them can cause dramatic structural changes and energy increases. In this context, even mutations that essentially modify individual backbone angles can be highly ineffective. One of the key results in this paper is that the employment of molecular fragment replacement as a mutation operator is highly effective.

Similarly, when employing crossover on backbone dihedral angles from two compact conformations, the chance of collisions is even higher, especially around the crossover point(s). As a consequence, most current EAs for off-lattice ab-initio structure prediction do not employ crossover, leaving opened the question of how a powerful probabilistic sampling technique like crossover can best be leveraged for proteins [13, 14, 26]. The work we present in this paper provides a first step in addressing this question, namely, the design and evaluation of a homologous crossover operator that retains good structural segments of the two conformations being joined.

The EA presented in this paper is also a "memetic" EA in that it combines both global and local search to improve exploration capability. The employment of local search is inspired by related studies on EAs for conformational search that show that the added optimization of sampled conformations improves te quality of conformations and leads to lower-energy regions during search [1, 20, 38].

2. METHODS

The hybrid EA we describe in this paper to enhance sampling of low-energy regions in the protein conformational space for the purpose of ab-initio structure prediction makes use of the Rosetta coarse-grained representation and the Rosetta coarse-grained energy function summarized in sections 2.1 and 2.2. Details of the hybrid EA are provided in section 2.3. The local search that makes use of the molecular fragment replacement technique is described in section 2.5,



(a) Native Structure

(b) Amino Acids 11 to 13

Figure 1: (a) The native structure under id 1dtdB in the Protein Data Bank (PDB) [4] of deposited native structures is shown using the New Cartoon graphical representation (rendered by VMD [22]). (b) A short portion (amino acid positions 11−13) of this structure is shown in greater detail. The backbone atoms of these amino acids are drawn in dark blue, and the corresponding side-chain atoms are in light grey. The backbone dihedral angles are annotated for these three amino acids, and their values in the native structure are shown below in degrees. An angular coarse-grained representation would store only these angles for each conformation of the amino-acid chain. Forward kinematics would be used to obtain the cartesian coordinates.

whereas the different crossover operators we have evaluated in the context of our hybrid EA are detailed in section 2.6.

2.1 Coarse-grained Representation

A conformation is represented in this work as a vector of $3n$ angles for a chain of n amino acids. These are the ϕ, ψ, and ω dihedral angles (defined on the second of three consecutive bonds in the chain) illustrated in Fig. 1b. This representation idealizes protein geometry, assuming the main source of variation in native conformations is due to dihedral angles rather than bond lengths and angles (defined between two consecutive bonds). This assumption is valid, as statistical analysis over known native structures has shown bond lengths and bond angles adopt characteristic values depending only on the types of atoms involved. Given precomputed bond length and angles [7], dihedral angles can then yield cartesian coordinates for modeled atoms through application of forward kinematics [40].

As in the Rosetta ab-initio protocol, the atoms explicitly modeled are backbone atoms of an amino acid (shown in dark blue in Fig. 1b) and a centroid pseudo-atom for the amino-acid's side chain. Centroids are given steric-free initial locations in an extended conformation, and application of forward kinematics rigidly moves these centroids in cartesian space as the backbone deforms. It is useful to note that the employment of backbone dihedral angles for the internal representation is suitable for the ensuing molecular fragment replacement technique described in section 2.4.

2.2 Potential Energy Fitness Function

The fitness of a conformation is computed here through the Rosetta coarse-grained energy function estimating the conformation's potential energy (lower energy means higher fitness). The function operates over cartesian coordinates, which are calculated through forward kinematics over the internal representation of a conformation as described above. The Rosetta energy function we use here corresponds to the *score3* setting in the Rosetta ab-initio protocol, which is the full coarse-grained energy function.

The full Rosetta coarse-grained energy function is a linear combination of terms measuring repulsion, amino-acid propensities, residue environment, residue pair interactions, interactions between secondary structure elements, density, and compactness (see Ref. [34] for more details). It is worth pointing out that even a coarse-grained energy function is computationally expensive, as the steric repulsion term, for instance, measures distances of atom pairs. Moreover, even though the Rosetta energy function is considered state of the art, the energy surface it defines is full of local minima. Some potential artifacts are due to the estimation of weights scaling the contribution of each energy term during the design of this function. Structural changes to a conformation may lower the value of one term while increasing that of another. The result is a multimodal energy surface.

2.3 Our Hybrid EA

A schematic of the hybrid EA evaluated in this paper is shown in Fig. 2. An initial population, P_0, is constructed as p copies of an extended conformation subjected to n random mutations, where n is the length of the target protein system (the mutation employs molecular fragment replacement, detailed below). The population P_i in each subsequent generation i is obtained as follows. All conformations of the previous population P_{i-1} are first duplicated, then subjected to crossover, mutation, and finally projected to a nearby local minimum through a local search. Three different crossover implementations are considered: 1-point, 2-point, and homologous 1-point crossover. The result of this process is p child conformations that are added to population P_i. The $k\%$ conformations with highest fitness in P_{i-1} are also added to P_i in order to maintain low-energy conformations captured in previous generations. The resulting population is reduced down to the same constant size of p individuals through truncation selection. All experiments in this paper employ $p = 100$ and an elitism rate of $k = 25\%$ based on fitness. These values are as in related work on EAs.

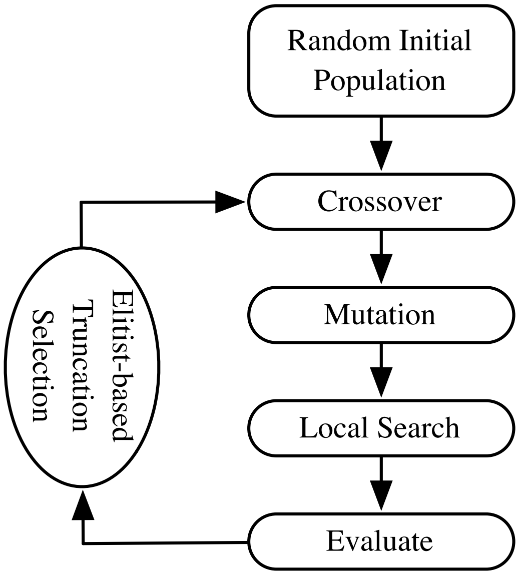


Figure 2: Flowchart of our hybrid EA. In each generation, the chosen crossover operator is followed by a mutation operator and a local search to optimize a conformation to a nearby local minimum. The new population of local minima competes with elite members of the parent population through truncation selection.

2.4 Implementing Mutation through Molecular Fragment Replacement

Our implementation of the mutation operator makes use of the molecular fragment replacement technique. The main motivation for using this technique is that the use of bond angles found in nature significantly improves sampling of near-native conformations over rotating angles by values sampled uniformly at random [6]. More importantly, the process of generating physically-realistic conformations is made much simpler when sampling values for a sequence of consecutive backbone dihedral angles simultaneously rather than sampling for one angle at a time. For these reasons, fragment replacement is now common practice in ab-initio structure prediction [39, 6, 8, 17].

Fragment length and construction of fragment configuration libraries is the subject of much ongoing research [21]. Here we use molecular fragments of length $l = 3$ and the latest fragment configuration libraries employed by the Rosetta ab-initio protocol [24]. A fragment configuration library for fragments of length l consists of $3l$ backbone dihedral angles extracted for each sequence of l consecutive amino acids found in proteins with known native structures. In the Rosetta libraries, clustering is used to store no more than 200 configurations for each possible fragment. Given a fragment configuration library, a mutation is performed on a conformation C as follows. First, an amino-acid position i is sampled uniformly at random over positions $[1, n - l + 1]$, where n is the total number of amino acids in the chain. The amino-acid sequence of the fragment from position i to $i + l - 1$ is used to find configurations available for that fragment in the fragment configuration library. A configuration is selected uniformly at random from those available. The selected configuration replaces the $3l$ backbone dihedral angles in the selected fragment in C, yielding a new conformation C_{new}.

2.5 Local Search Operator

The local search operator maps a conformation to a nearby local minimum in the protein energy surface. The local search operator accomplishes the goal of explicitly sampling local minima conformations while ensuring that these are in feasible regions of the search space. Our implementation is a greedy local search with molecular fragment replacement found to be effective in our previous studies on the basin hopping framework [28, 30, 29]. Essentially, the local search is implemented as a 1+1 EA with truncation selection using the same mutation operator described in section 2.4. Each local search continues until m consecutive fragment replacements fail to improve the fitness of a conformation. The local search encapsulates the precise definition of a local minimum with the value of m determining the depth at which each local minimum is probed. Analysis in previous work suggests setting m to the number of amino acids in the target protein sequence [29].

2.6 Crossover Operators

This work investigates the effectiveness of different crossover operators, comparing standard 1-point and 2-point crossover operators to a homologous 1-point crossover operator.

For each parent in the population, a second parent is selected uniformly at random for crossover, and each pair of parents produce a single offspring. In 1-point and 2-point crossover, the crossover points are selected uniformly at random over amino acid positions in the target protein sequence. In homologous 1-point crossover, the crossover point is selected uniformly at random over the set of amino acids for which both parents share the same ϕ and ψ dihedral bond angles. This effectively creates a bias towards selecting a crossover point based on the number of consecutive amino acids with matching ϕ and ψ angles. In the case that there are no matching pairs of ϕ and ψ angles, a standard 1-point crossover is performed instead.

To provide baseline comparisons, two additional experiments have been included: one with just the fragment mutation operator (i.e., no crossover), and a second experiment using only a random replacement operator that performs n fragment mutations (n is the number of amino acids), effectively restarting at a conformation randomly sampled from the fragment configuration library.

3. RESULTS

Systems of study and performance measurements: Performance is evaluated on 10 proteins with known native structures. These proteins range in length from 61 to 123 amino acids and represent a diverse set of α, β, and α/β native fold topologies, as listed in columns 2−4 in Table 1. On many of these systems, comparative results are available from other conformational search procedures consisting of multiple MMC trajectories.

Experimental setup: The parameters for each experiment are a particular crossover operator (1-point crossover, 2-point crossover, homologous 1-point crossover, random replacement, or mutation only) and a target protein system. Each experiment is an execution of our EA with the selected parameters for a fixed budget of $10,000,000$ energy function evaluations. Each run is repeated 30 times, with means and best runs reported. In practice, $10,000,000$ energy evaluations takes 7−24 hours of CPU time on a 2.4Ghz Core i7 processor, depending on protein length.

Because of the variable number of evaluations of the local search operator, the number of generations is not the same across experiments. Each experiment is run until the generation for which the number of energy evaluations exceeds $10,000,000$. As a result, the actual number of energy evaluations can also vary slightly between runs; however, this does not significantly affect results, since differences do not exceed 0.5% of total run time, and results are averaged over 30 independent runs.

Types of analyses and performance measurements: Section 3.1 compares reproductive operators on the lowest energy values reached in order to determine which operator results has higher sampling capability. Section 3.2 provides further details into how the fitness improvement of each operator relates to the sampling of low-energy conformations.

Finally, because one of the objectives of this work is to provide the first steps in making EAs competitive and thus useful in the context of ab-initio structure prediction, section 3.3 analyzes the lowest RMSD reached to the known native structure of a protein. RMSD stands for root-mean-squared-deviation and is a common measurement of dissimilarity between two conformations. RMSD scales Euclidean distance between corresponding atoms in two conformations by the number of atoms. Here we report RMSD after an optimal superposition is found of the two conformations under comparison that minimizes distances due to rigid-body motions. Additionally, in keeping with how results are reported by other search procedures, we calculate RMSD only over C_α atoms. This final analysis provides insight into the ability of different reproductive operators to approach the native structure and how that compares with established MMC-based search procedures for ab-initio structure prediction.

3.1 Navigating the Protein Energy Surface

Table 1 shows the lowest energy across all the conformations sampled during each experiment. The value shown is the mean over 30 runs, with the minimum of 30 shown in parentheses. Values shown for the random replacement operator in column 5 provide a baseline of comparison. Column 6 shows values reached by the mutation-only EA with no crossover. Columns 7−9 show values for the 1-point, 2-point, and homologous 1-point crossover, respectively (recall that mutation and local search follow after each crossover operator). Examination of these values shows that in nearly every instance, the addition of crossover results, on average, in lower energies. In only the single case of 2ezk does employment of mutation alone yield a slightly lower energy than 2-point crossover. These results make the case that use of a crossover operator allows our hybrid EA to more effectively navigate the protein energy surface and access lower-energy conformations.

A downside to employing crossover over mutation alone is that the larger moves in dihedral angle space tend to require a longer local search to return to a local minimum. As a result, our hybrid EA with mutation only is able to sample significantly more conformations for the same number of energy evaluations (data not shown here). While 1-point and 2-point crossover allow sampling on average 11% and 21% fewer conformations than mutation-only, respectively, homologous 1-point crossover averages only a 7% reduction and never more than 11% across all proteins. This suggests that homologous crossover is less "disruptive" to local structure than other crossover operators and thus increases sampling diversity without as much of an increase in the time required to map a conformation to a local minimum.

The effect of this "disruption" is particularly evident between the 1-point and 2-point crossover methods. As indicated in Table 1, improvements in energy over mutation alone are found to be statistically significant with 95% confidence by the Mann-Whitney U test for 6 of the proteins when employing 1-point and homologous 1-point crossover. However, statistical significance is only achieved for 4 proteins using 2-point crossover. This suggests that the increased disruption found in 2-point crossover makes it less suitable for use in protein structure prediction.

3.2 Fitness Improvement

Fig. 3 shows the mean fitness improvement (x-axis) versus the lowest energy (y-axis) reached for each experiment on all proteins. Fitness improvement is defined as the difference between parent energy and child energy (energy is averaged over the two parents in crossover). Fig. 3 shows a strong linear correlation, particularly at lower energies reached. This suggests that a more explorative reproductive operator, such as crossover, which is likely to have lower fitness improvement, will ultimately reach lower energy levels than a more exploitative operator, such as mutation.

3.3 Sampling Near-native Conformations

Section 3.3.1 evaluates the effectiveness of the crossover operators in terms of the lowest RMSD they reach to each of the known native structures. These RMSDs are then compared with published RMSDs by a study on a state-of-the-art MMC-based search procedure in section 3.3.2.

3.3.1 Effectiveness of Crossover

Table 2 shows the lowest C_α-RMSD to the known native structure over all conformations sampled during each experiment. The value given is the best result over all 30 runs. Examination of columns 6−9 in Table 2 shows no consistent improvement in C_α-RMSD between crossover methods and mutation alone; a difference of less than 0.5Å C_α-RMSD is not considered structurally significant. This suggests that while crossover is able to reach lower energy regions of the conformational space, inaccuracies in the Rosetta energy function prevent it from fully capitalizing on this success with respect to getting closer to the known native structure.

Column 11 in Table 2 quantifies energy function inaccuracies for each protein in terms of the Pearson correlation coefficient between the energy of each conformation and its

Table 1: Columns 2−4 show PDB Id of the native structure, number of amino acids, and native fold topology for each target protein, respectively. The remaining columns show the lowest energy sampled during each experiment, averaged over 30 runs. The minimum lowest energy over all runs is shown in parentheses. Results for which the mean difference between crossover and mutation only are statistically significant are given in bold (≥ 95% confidence according to the Mann-Whitney U test).

			Native	Lowest Energy Value Sampled				
			Fold	Random	Mutation	Crossover and Mutation		
	PDB Id	Size	Topology	Replacement	Only	1-point	2-point	Homologous
1	1dtdB	61	α/β	43.2(36.2)	-11.3(-38.8)	**-19.1**(-38.8)	-14.3(-37.7)	-12.7(-34.3)
2	1isuA	62	α/β	35.9(28.3)	1.1(-13.4)	-0.4(-13.4)	-1.5(-19.2)	**-2.9**(-15.3)
3	1c8cA	64	α/β	14.3(2.2)	-40.1(-67.0)	**-45.6**(-66.5)	**-45.6**(-64.2)	**-45.9**(-68.3)
4	1sap	66	α/β	5.6(-5.2)	-53.7(-76.5)	-58.6(-86.2)	-58.2(-83.2)	-57.8(-78.7)
5	1hz6A	67	α/β	21.3(-5.8)	-56.9(-81.3)	-63.0(-99.8)	-64.3(-84.8)	-64.7(-89.0)
6	1wapA	68	β	25.8(11.2)	-51.0(-93.3)	**-60.8**(-86.3)	-55.9(-74.0)	**-59.7**(-81.5)
7	1ail	70	α	30.7(26.0)	-6.2(-23.3)	**-12.6**(-25.8)	**-13.6**(-22.7)	**-13.3**(-27.3)
8	1aoy	78	α/β	19.3(5.8)	-28.6(-51.3)	**-37.4**(-52.5)	**-36.4**(-48.8)	**-41.7**(-56.9)
9	2ezk	93	α	22.8(16.9)	-18.7(-29.1)	-21.2(-33.8)	-18.4(-31.3)	-20.8(-31.6)
10	2h5nD	123	α	45.2(31.4)	-20.5(-40.1)	**-30.4**(-48.9)	-23.1(-39.7)	**-29.5**(-53.0)

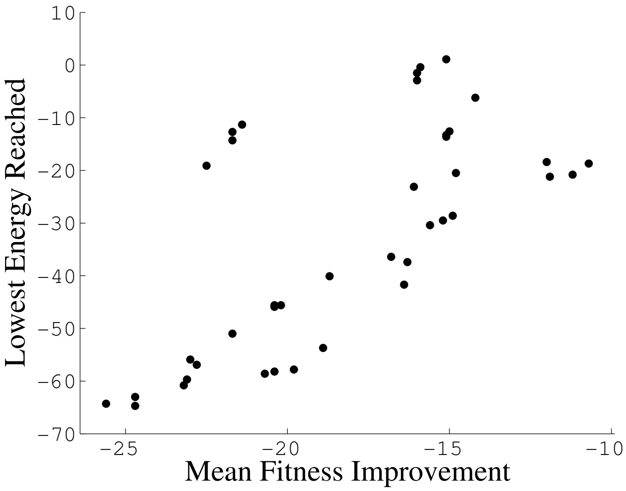


Figure 3: The mean fitness improvement between parents and children is given for each experiment versus the average lowest energy reached. A strong linear correlation is noted, suggesting a more explorative variation operator with lower mean fitness improvement will allow an EA to maintain breadth in search and access lower energies.

C_α-RMSD across all experiments for a given protein. In only a single protein (1hz6A) is this correlation above 52%, indicating the difficulty inherent in protein structure prediction as an optimization problem. On 1isuA, 1ail, and 2ezk where the correlation is bellow 40%, the random replacement method is actually competitive with crossover and mutation only. The fact that random replacement is able to find low RMSD structures for 1ail and 2ezk underscores the advantage of employing molecular fragment replacement.

3.3.2 *Comparison to Published Results*

Column 10 in Table 2 shows the lowest C_α-RMSD achieved by the ItFix algorithm as reported in [16]. ItFix is representative of search algorithms that consist of many independently-run MMC trajectories. Comparison of columns 9 and 10 shows that our hybrid EA with homologous 1-point crossover achieves a C_α-RMSD at least 0.5Å lower than ItFix for all 9 proteins where ItFix results are available. Since ItFix is not an EA and employs different algorithmic components, it is difficult to draw a further comparison to our EA. However, the results suggest that our hybrid EA is at least as effective as ItFix and can be an effective search procedure for ab-initio structure prediction protocols. It is also worth noting that homologous 1-point crossover is the only version of our hybrid EA which consistently finds C_α-RMSD values at least 0.5Å lower than the MMC-based method.

4. CONCLUSION

This work makes the case that EAs can offer effective and powerful search procedures for the coarse-grained stage in ab-initio structure prediction. In order to make EAs competitive with the MMC-based procedures typically employed in state-of-the-art ab-initio protocols, a coarse-grained representation (and the energy function that goes with it) and molecular fragment replacement are incorporated in the proposed EA. The employed representation and energy function are those used in the well-known Rosetta ab-initio protocol.

The proposed hybrid EA also employs local search to map each child conformation to a nearby local minimum. Different reproductive operators are evaluated according to various performance measurements. Analysis of sampled energies reveals that the use of crossover results in a higher exploration capability over mutation alone. Furthermore, results suggest that there is an advantage to 1-point crossover over 2-point crossover due to the more locally disruptive nature of 2-point crossover.

An interesting result in this paper is that, while crossover enhances sampling capability, accessing lower-energy regions does not necessarily translate to better proximity to the native structure. This result confirms other recent findings

Table 2: The lowest C_α-RMSDs to the known native structure over conformations sampled during each experiment are given. Columns 5−9 report results for experiments performed in this work, while Column 10 shows the lowest C_α-RMSD reported for each protein by the ItFix algorithm [16]. Column 11 shows the Pearson correlation coefficient between energies of sampled conformations and their C_α-RMSD across all experiments.

			Native	Lowest $C\alpha$-RMSD Conformation Found (Å)						Energy
			Fold	Random	Mutation	Crossover and Mutation			ItFix	to RMSD
	PDB Id	Size	Topology	Replacement	Only	1 point	2 point	homologous	[16]	Correlation
1	1dtdB	61	α/β	5.9	5.3	4.2	4.8	4.8	5.7	0.49
2	1isuA	62	α/β	6.1	5.9	6.3	5.7	5.9	6.5	0.31
3	1c8cA	64	α/β	5.9	2.5	3.5	4.5	3.2	3.7	0.52
4	1sap	66	α/β	5.5	3.5	2.4	2.7	3.6	4.6	0.48
5	1hz6A	67	α/β	4.0	1.9	1.8	2.3	1.9	3.8	0.68
6	1wapA	68	β	7.1	6.2	6.1	5.6	6.5	8.0	0.42
7	1ail	70	α	4.1	3.9	3.5	4.2	3.7	5.4	0.06
8	1aoy	78	α/β	5.4	3.6	4.0	3.7	3.2	5.7	0.50
9	2ezk	93	α	3.8	3.5	3.1	3.0	3.0	5.5	0.25
10	2h5nD	123	α	7.5	6.2	7.4	6.7	6.6	NA	0.48

that even enhanced sampling search procedures have to operate within the limitations of coarse-grained energy functions [30, 36, 15, 5, 35]. Our detailed analysis shows that even the Rosetta energy function can associate very low energies with non-native conformations, as confirmed by other studies [36, 15, 5].

Comparison with other MMC-based search procedures shows that the proposed hybrid EA is just as effective. We believe that further work, particularly on enhancing structural diversity, may open the way towards EAs becoming the search procedure of choice in ab-initio protocols. With their enhanced sampling capability, EAs can also aid researchers in improving energy functions on non-native low-energy conformations revealed by the EA.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] M. S. Abual-Rub, M. A. Al-Betar, R. Abdullah, and A. T. Khader. A hybrid harmony search algorithm for ab initio protein tertiary structure prediction. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, pages 1–17, 2012.

[2] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[3] D. Becerra, A. Sandoval, D. Restrepo-Montoya, and L. Nino. A parallel multi-objective ab initio approach for protein structure prediction. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 137–141, 2010.

[4] H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, 10(12):980–980, 2003.

[5] G. R. Bowman and V. S. Pande. Simulated tempering yields insight into the low-resolution rosetta scoring functions. *Proteins: Struct. Funct. Bioinf.*, 74(3):777–788, 2009.

[6] P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[7] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[8] T. J. Brunette and O. Brock. Guiding conformation space search with an all-atom energy potential. *Proteins: Struct. Funct. Bioinf.*, 73(4):958–972, 2009.

[9] J. Calvo and J. Ortega. Parallel protein structure prediction by multiobjective optimization. *Parallel, Distributed and Network-based Processing, 2009 17th Euromicro International Conference on*, pages 268–275, 2009.

[10] J. Calvo, J. Ortega, and M. Anguita. Comparison of parallel multi-objective approaches to protein structure prediction. In *Journal of Supercomputing*, pages 253–260. CITIC UGR Univ Granada, Dept Comp Architecture & Comp Technol, Granada, Spain, 2011.

[11] C. Chira, D. Horvath, and D. Dumitrescu. An Evolutionary Model Based on Hill-Climbing Search Operators for Protein Structure Prediction. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 38–49, 2010.

[12] Cutello, V, G. Morelli, G. Nicosia, M. Pavone, and G. Scollo. On discrete models and immunological algorithms for protein structure prediction. *Natural Computing*, 10(1):91–102, 2011.

[13] Cutello, V, G. Narzisi, and G. Nicosia. A class of pareto archived evolution strategy algorithms using immune inspired operators for ab-initio protein structure prediction. In *Applications of Evolutionary*

Computing, Proceedings, pages 54–63. Univ Catania, Dept Math & Comp Sci, I-95125 Catania, Italy, 2005.

[14] Cutello, V, G. Narzisi, and G. Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of The Royal Society Interface*, 3(6):139–151, 2006.

[15] R. Das. Four small puzzles that rosetta doesn't solve. *PLoS ONE*, 6(5):e20044, 2011.

[16] 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*, 106(10):3734–3739, 2009.

[17] 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.

[18] R. Faccioli, I. N. da Silva, L. O. Bortot, and A. Delbem. A mono-objective evolutionary algorithm for Protein Structure Prediction in structural and energetic contexts. pages 1–7, 2012.

[19] M. Garza-Fabre, G. Toscano-Pulido, and E. Rodriguez-Tello. Locality-based multiobjectivization for the HP model of protein structure prediction. In *GECCO '12: Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference.* ACM Request Permissions, July 2012.

[20] M. M. Goldstein, E. E. Fredj, and R. B. R. Gerber. A new hybrid algorithm for finding the lowest minima of potential surfaces: approach and application to peptides. *Journal of Computational Chemistry*, 32(9):1785–1800, July 2011.

[21] J. Handl, J. Knowles, R. Vernon, D. Baker, and S. C. Lovell. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. *Proteins: Struct. Funct. Bioinf.*, 80(2):490–504, 2011.

[22] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Mol. Graph. Model.*, 14(1):33–38, 1996. http://www.ks.uiuc.edu/Research/vmd/.

[23] M. K. Islam, M. Chetty, and M. Murshed. Novel Local Improvement Techniques in Clustered Memetic Algorithm for Protein Structure Prediction. pages 1–9, Apr. 2011.

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

[25] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP) round IX. *Proteins: Struct. Funct. Bioinf.*, Suppl(10):1–5, 2011.

[26] G. Narzisi, G. Nicosia, and G. Stracquadanio. Robust Bio-active Peptide Prediction Using Multi-objective Optimization. In *Biosciences (BIOSCIENCESWORLD), 2010 International Conference on*, pages 44–50, 2010.

[27] B. Olson, , and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, 2012. in press.

[28] B. Olson and A. Shehu. Populating local minima in the protein conformational space. In *IEEE Intl Conf on Bioinf and Biomed*, pages 114–117, Atlanta, GA, November 2011.

[29] B. Olson and A. Shehu. Efficient basin hopping in the protein energy surface. In J. Gao, W. Dubitzky, C. Wu, M. Liebman, R. Alhaij, L. Ungar, A. Christianson, and X. Hu, editors, *IEEE Intl Conf on Bioinf and Biomed (BIBM)*, pages 119–124, Philadelphia, PA, October 2012. IEEE.

[30] B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, 10(10):S5, 2012.

[31] 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*, 10(3):1242005, 2012.

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

[33] M. C. Prentiss, C. Hardin, M. P. Eastwood, C. Zong, and P. G. Wolynes. Protein structure prediction: The next generation. *J. Chem. Theory Comput.*, 2(3):705–716, 2006.

[34] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using rosetta. *Methods Enzymol.*, 383:66–93, 2004.

[35] A. Shehu and B. Olson. Guiding the search for native-like protein conformations with an ab-initio tree-based exploration. *Int. J. Robot. Res.*, 29(8):1106–11227, 2010.

[36] A. Shmygelska and M. Levitt. Generalized ensemble methods for de novo structure prediction. *Proc. Natl. Acad. Sci. USA*, 106(5):94305–95126, 2009.

[37] D. Simoncini, F. Berenger, R. Shrestha, and K. Y. J. Zhang. A probabilistic fragment-based protein structure prediction algorithm. *PLoS ONE*, 7(7):e38799, 2012.

[38] A.-A. Tantar, N. Melab, and E.-G. Talbi. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. *Soft Computing*, 12(12):1185–1198, 2008.

[39] D. Xu and Y. Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct. Funct. Bioinf.*, 80(7):1715–1735, 2012.

[40] M. Zhang and L. E. Kavraki. A new method for fast and accurate derivation of molecular conformations. *Chem. Inf. Comput. Sci.*, 42(1):64–70, 2002.