Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface

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

Amarda Shehu*
Department of Computer Science,
Department of Bioengineering
School of Systems Biology
George Mason University,
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

We present an evolutionary stochastic search algorithm to obtain a discrete representation of the protein energy surface in terms of an ensemble of conformations representing local energy minima. This objective is of primary importance in protein structure modeling, whether the goal is to obtain a broad view of potentially different structural states that are thermodynamically available to a protein system or to predict a single structure that is representative of a unique functional native state. In this paper, we focus on the latter setting and show how approaches from evolutionary computation for effective stochastic search and multi-objective analysis can be combined to result in protein conformational search algorithms with high exploration capability. From a broad computational perspective, the contributions of this paper are on how to balance global and local search of some high-dimensional search space and how to effectively guide the search in the presence of a noisy and inaccurate scoring/objective function. From an application point of view, the contributions are demonstrated in the domain of ab-initio protein structure prediction on the primary subtask of sampling diverse low-energy decoy conformations of a given amino-acid sequence. Comparison with the approach used for decoy sampling in the popular Rosetta protocol on 20 diverse protein sequences shows that the evolutionary search algorithm proposed in this paper is able to access lower-energy regions of the energy surface with similar or better proximity to the known native structure.

Keywords

stochastic search; hybrid evolutionary search algorithm; multi-objective guidance; protein structure modeling; ab-initio structure prediction; conformational space; decoy sampling

*Corresponding Author

ACM BCB 2012 Orlando, FL

1. INTRODUCTION

Understanding proteins is of primary importance in molecular biology. Many proteinopathies, such as as Alzheimer's, prion's, and Huntington's disease are protein conformational diseases occurring where protein loses its intended function due to its inability to assume an appropriate structure in the cell [31]. Protein structure is often regarded as a carrier of function, as proteins are known to adopt specific three-dimensional shapes to interact with other molecules. It is no surprise that the path to unraveling protein function goes through understanding and modeling protein structure [2].

Experimental techniques, such as X-ray crystallography typically capture a single representative/average structure of the protein native state associated with lowest free energy according to the thermodynamic hypothesis [1]. Nuclear Magnetic Resonance (NMR) extends the characterization to a few more structural models satisfying observed atomic distances. This incomplete dynamic characterization is additionally limited to small proteins (typically less than 100 amino acids). Experimental techniques capable of yielding structural models representing potentially different functional states of a protein include FRET and cryo-electron microscopy, but these techniques are either limited to very low-resolution models or a few structural states [15].

Currently, millions of protein sequences extracted from organismal genomes lack any structural or functional characterization. Modeling protein structure in silico is important to complement experimental techniques and additionally shed light on the structure-function relationship, help engineer novel proteins, predict protein stability, model molecular interactions, and design novel drug compounds [29].

The most powerful computational characterization of protein structure when assuming only knowledge of its amino-acid sequence is largely limited to one structure predicted to represent a unique protein functional state. When no homologs are available for the sequence under investigation, the problem is known as ab-initio structure prediction. Addressing it does not guarantee that a comprehensive structural characterization relevant for function will be obtained for the protein at hand. Many proteins are found to employ both small-scale and large-scale fluctuations to switch between different functional states, effectively modulating function [2]. However, the narrowed focus in ab-initio structure prediction is relevant for many proteins and has led to many advances in computational structural biology [28, 29].

While the ab-initio structure prediction problem remains

an outstanding challenge, much progress has been made [18]. The most successful framework so far employs a two-stage approach [25]. In stage one, a large ensemble of low-energy (decoy) conformations is obtained to provide a broad view of the protein energy surface relevant for function. The term conformation, though largely interchanged with structure, refers to an assignment of values to underlying parameters representing a spatial arrangement of the chain of amino acids. According to the Anfinsen experiments, the native state associated with protein function is also the thermodynamic equilibrium; that is, the state of lowest free energy [1]. However, due to the difficulty of measuring free energy in silico, the approach is often to ignore entropic considerations and only seek conformations of low potential energy. In the second stage, clustering is performed to group decoys in local minima, and representatives of top-populated clusters are followed up with further energetic refinement.

Two key developments have allowed the application of this two-stage framework for ab-initio structure prediction. First, the protein conformational space is simplified and reduced in dimensionality through a coarse-grained representation of the protein chain. Typically, side chains are ignored and the focus is primarily on modeling backbone atoms. It is in stage two that side-chain packing techniques [12] are used to add side chains prior to further energetic optimization to promising decoys. Second, the molecular fragment replacement technique is used to discretize the search space through bundling of backbone dihedral angles together. Values are assigned for a consecutive number of angles simultaneously according to structural pieces or fragment configurations pre-compiled over known native protein structures [13].

In this paper we focus on better understanding the first stage, as the quality of decoys determines whether the second stage will have any success. If no near-native (in close proximity to the native structure) conformations are obtained in stage one, then there is little chance that the local energetic refinement will lead to the native structure. In particular, we focus on the quality of the decoy ensemble obtained in ab-initio structure prediction. Since the purpose of obtaining this ensemble is to obtain a broad view of the energy surface relevant for function, our focus here allows using the proposed techniques and algorithms beyond ab-initio structure modeling to possible characterization of proteins with diverse functionally-relevant structural states.

We are interested in particular on algorithmic components for effective search of the protein conformational space with enhanced exploration capability. This is central to protein structure modeling, as most decoy sampling methods consist of two components: First, a component allows iterating between novel conformations. Given the high-dimensionality of the search space, this has to be stochastic/ sampling-based. A second component guides the search towards relevant conformations through a scoring or energy function that discriminates among conformations.

While the capability of current decoy sampling algorithms for ab-initio structure prediction is considered to be high, two issues are identified, representing an impasse in the computational structural biology community. How does one enhance sampling capability to reduce the likelihood that important regions are missed? How does one deal with the fact that all current scoring functions (which measure potential energy), particularly those that operate on coarse-grained representations, are known to be noisy and lead algorithms searching for the global minimum to non-native conformations? Inadequate exploration of the conformational space, and/or inaccuracies in energy functions are often cited as reasons why ab-initio structure prediction remains challenging, particularly on sequences more than 70 amino-acids long and/or certain native topologies [18].

In this paper, we present a stochastic search algorithm to address these two questions. First, rather than rely on multi-start Metropolis Monte Carlo (MMC) trajectories as in Rosetta [25] and many other decoy algorithms [4,8,30,32], we present a hybrid search that balances the exploration between global and local search. Second, rather than rely on noisy guidance by the energy function, we present a multi-objective approach in the context of this hybrid search that essentially optimizes multiple independent energetic criteria.

The novel algorithm we propose here is inspired by related work in evolutionary computation. Current Evolutionary search Algorithms (EAs) proposed for protein structure modeling use lattice-based representations or computationally-demanding all-atom representations [5–7,11,19]. Studies applying EAs for decoy sampling are thus limited to very small proteins or toy models; currently, EAs are not competitive against state-of-the-art methods in ab-initio structure prediction. In recent work we have provided a pathway for employing EAs in ab-initio structure prediction by incorporating coarse-grained representations and molecular fragment replacement, improving the exploration capability of single-trajectory or population-based EAs [21, 23, 24, 26] as well as tree-based search algorithms [22, 27]. In this paper we build upon this foundation and propose a more powerful novel algorithm.

We propose in this paper a hybrid multi-objective EA, referred to as MOEA. MOEA incorporates the latest coarse-grained chain representation, energy function, and fragment libraries used in Rosetta [13]. MOEA evolves a population of conformations, starting with a carefully-constructed initial population. Child conformations are obtained through asexual (mutation) reproduction. MOEA is guided by Pareto-based scoring metrics rather than total potential energy. Prior to adding a conformation to the population, the algorithm decomposes the energy of a conformation into various terms, the values of which are compared to those of other conformations maintained in an archive to decide whether the conformation should be added to the population.

MOEA explicitly samples local minima in the energy surface, as its procedure to generate child conformations combines global and local search. In this way, the ensemble of conformations obtained through MOEA is a discrete representation of the protein energy surface relevant for function in terms of local minima. Two different versions of the proposed MOEA (whether guiding by Pareto rank or count) are compared to each-other and to the baseline algorithm that is guided only by total energy rather than Pareto-based metrics. The comparison shows that Pareto-based metrics help explore lower-energy minima. Comparison with the decoy sampling algorithm used in the Rosetta protocol on 20 diverse protein sequences shows that MOEA is able to access lower-energy regions with similar or better proximity to the known native structure. We now proceed with details on the algorithm in section 2, followed by analysis of its performance in section 3, and conclusions and directions of future work in section 4.

2. METHOD

A fundamental challenge in stochastic search and optimization is how to balance limited computational resources between exploration of a space through global search with exploitation of local minima through local search. The common approach in decoy sampling methods is disjointed, achieving exploitation through intensive localized MMC search, while using multi-start or random-restart for global search. The balance between local and global search is primarily limited to tuning of the temperature parameter in the Metropolis criterion that determines whether the search will remain local or cross energy barriers for a more global view.

The MOEA algorithm we propose in this paper combines local and global search in a population-based evolutionary search framework. In MOEA, a fixed-size population of p of decoy conformations is "evolved" through a series of generations as guided by Pareto analysis (detailed below) rather than by the total energy of a conformation. We note that the primary source of "global search" in MOEA is its exploration of the conformational space with n simultaneous conformations that evolve over a series of generations.

The initial population P_0 in MOEA is carefully constructed to contain diverse randomized conformations with credible secondary structures. Details are provided in section 2.2. The population P_i in each subsequent generation i is then obtained as follows. All conformations of the previous population P_{i-1} are first duplicated, subjected to a single mutation, and then projected to a nearby local minimum through a local search. A conformation is mutated by making one fragment replacement move, using a fragment of length 3.

The result of this process is p child conformations, which are compared to an archive containing all child conformations ever computed by MOEA. The top l out of p children according to this comparison are then combined with the top $s\%$ of the p parents from population P_{i-1}. Out of these, only the top p conformations (according to same metrics used for comparison) comprise population P_i. The combination of parent and child conformations is known as elitism, and its purpose is to preserve good solutions captured in previous generations while preventing premature convergence of the entire population. Typically in EAs for decoy sampling, including our previous work [21, 23, 24, 26], determining which conformations in the combined set comprise P_i, known as population selection, uses truncation selection based on potential energies (top p from ascending order).

The primary contributions of this paper is that MOEA is not guided by total energy alone. An energy function is decomposed into groups of terms (the particular decomposition here is detailed in section 2.4). As section 2.4 describes, two Pareto metrics can be used, and so two algorithmic realizations of MOEA are pursued. MOEA uses only Pareto rank and total energy, whereas MOEA-PC uses Pareto rank, Pareto count, and total energy. Details are provided in section 2.4. MOEA and MOEA-PC are compared to the baseline algorithm that uses only total energy rather than Pareto-based metrics. We refer to this baseline algorithm as HEA for hybrid EA. We now provide further details, starting with the coarse-grained representation employed here to represent a protein chain, the molecular fragment replacement technique, and some details on the Rosetta energy function(s) employed in MOEA.

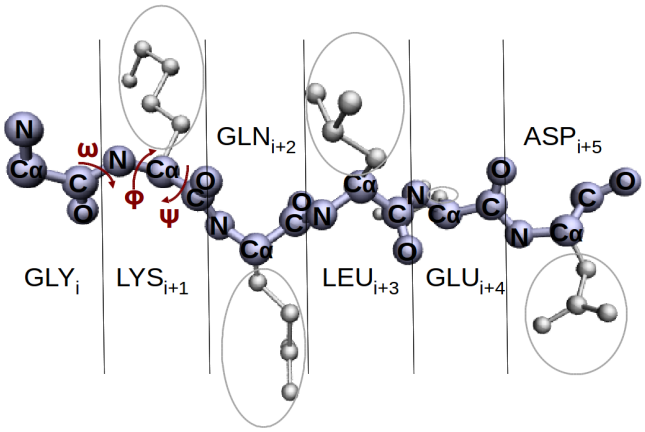

Figure 1: A fragment of a native structure is drawn to show backbone atoms (dark blue and annotated) and side-chain atoms (white and circled). Hydrogen atoms are not drawn for visibility. Boundaries between amino acids are drawn along the peptide bonds connecting them. Backbone dihedral angles ω, ϕ, ψ are annotated on a selected amino acid.

2.1 Preliminaries

Representation for Calculating Potential Energy.

Potential energy functions operate over interatomic distances. MOEA uses a coarse-grained one as in Rosetta and Quarkk. As in the Rosetta decoy sampling stage, the only atoms explicitly modeled in MOEA are backbone atoms of an amino acid and centroid pseudo-atom tracking the average location of an amino acid side-chain atoms (Fig. 1 illustrates backbone and side-chain atoms).

Kinematic Representation.

The internal (kinematic) representation employed in MOEA to speedily modify a conformation to obtain a new one models only backbone dihedral angles. This allows directly satisfying observed (equilibrium) constraints over bond lengths and valence angles on native structures and is the representation of choice in protocols for decoy sampling. A conformation of a protein sequence of n amino acids is represented as a vector of $3n$ angles (ϕ, ψ, and ω angles per amino acid, as dihedral illustrated in Fig. 1(b)). The representation is suitable for the molecular fragment replacement technique that directly modifies a selected group of backbone dihedral angles to obtain a new conformation. Dihedral angles can yield cartesian coordinates for modeled atoms through application of forward kinematics [33].

Molecular Fragment Replacement.

MOEA uses the molecular fragment replacement technique made popular by Rosetta [3]. A fragment is defined over amino acids at positions $[i, i+f-1]$ in the chain. A fragment configuration library stores all fragment configurations ($3f$ backbone dihedral angles) of fixed length f extracted from known protein native structures. MOEA uses libraries constructed as in Rosetta [13]. Given a fragment configuration library, a current conformation C is modified as follows to obtain a new one. First, an amino-acid position i is sampled uniformly at random over positions $[1, n-f+1]$, where n is the number of amino acids in the chain. The selected configuration replaces the $3f$ backbone dihedral angles in the

selected fragment in C, yielding a new conformation C_{new}. While fragment length is the subject of much research [9,10], here we use fragments of length 9 for the initial population and of length 3 for the mutation and local search.

Rosetta Energy Function(s).

MOEA employs a subset of the suite of energy functions used in the Rosetta decoy sampling stage. Rosetta uses 5 scoring functions, named `score0`, `score1`, and so on in its decoy sampling stage to gradually add more energetic constraints for the MMC trajectory initiated at an extended conformation. The functions are different weighted versions of the full Rosetta energy function. The full 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. [25] for more details). Even though the Rosetta energy function is considered state of the art, the energy surface it defines is full of non-native local minima [16,17,30]. Some potential artifacts are due to the estimation of weights that scale the contribution of each energy term during the design of this function. Additionally, structural changes to a conformation may lower the value of one term while increasing that of another, resulting in an energy surface rich in local minima. In MOEA, `score0` and `score1` are used to construct the initial population, while `score3` is used for the local search, mirroring the fact that in Rosetta, the majority of the search is done with `score3`. The comparison of conformations in MOEA, however, uses Pareto-based metrics and total energy value over `score4`, which is used in Rosetta not during search but to compare obtained decoys prior to further refinement in stage 2.

2.2 Initial Population in MOEA

The initial population P_0 is obtained by conducting p independent MMC trajectories from a fully extended conformation. The 200 first moves (fragment replacements) in a trajectory use `score0`. The rest of the moves use `score1` with a lower temperature close to room temperature that ensures a move is accepted primarily if it lowers potential energy, while allowing some small increases to be accepted per the Metropolis criterion. Moves are applied until n consecutive moves (n is number of amino acids in a given protein sequence) fail per the Metropolis criterion. Usage of `score0`, which consists of only a soft steric repulsion is to obtain a diverse population of conformations free of steric clashes. Usage of `score1` allows formation of secondary structure. This process, using a fragment length of 9 and employing different energy functions at different substages, mirrors that in Rosetta (though the length of the trajectories in the substages is much longer in Rosetta) to obtain credible initial conformations for the MOEA.

2.3 Local Search in MOEA

The local search operator maps a conformation to a nearby local minimum in the energy surface. The local search accomplishes the goal in MOEA of explicitly sampling local minima while ensuring sampled conformations are in feasible regions of the search space. We implement local search here as a greedy search where moves are fragment configuration replacements (per the molecular fragment replacement technique described above). This greedy search implementation has been found effective in our previous work on EAs [23]. The greedy search terminates when n consecutive replacements fail to lower total potential energy (n is number of amino acids). The energy function used for the local search is `score3`.

2.4 Multi-objective Guidance in MOEA

Energy Function Decomposition.

The Rosetta `score4` energy function used in conformation comparison in MOEA is the weighted sum of 13 distinct energy terms. Multi-objective analysis is most effective when employing only a few separate objectives. Therefore `score4` is decomposed into three separate terms measuring short range hydrogen bonding, long-range hydrogen bonding, and a third term summing the rest of the terms. Studies suggest that hydrogen bonding is especially important in identifying conformations of native topologies [30]; however, the additional energy terms are also important in guiding the search towards near-native regions. The decomposition employed in MOEA allows simultaneously optimizing conformations which are low in total energy and contain favorable local and non-local hydrogen bonding interactions.

Pareto-based Scoring Metrics.

The metrics employed in MOEA are based on the concept of Pareto dominance. A conformation C_i in the archive "dominates" another conformation C_j in the archive when each energy term of C_i (per some categorization as above) is lower than the corresponding energy term in C_j. If there is no conformation in the archive that dominates C_i, then C_i is "non-dominated" (its Pareto rank is 0, which is number of conformations that dominate it). C_i is said to belong to the Pareto front, which consists of all non-dominated conformations (with Pareto rank 0) in the archive. These are considered equivalent with respect to a multi-objective analysis. Fig. 2 illustrates the Pareto front for a simplified energy function containing only two terms. Membership in the Pareto front is a binary state, so additional metrics are necessary for a more granular ranking of conformations. In addition to Pareto rank, one can also measure the Pareto count of a conformation. The Pareto count of C_i measures the number of other conformations C_i dominates. Pareto rank and Pareto count are illustrated in Figure 2.

Pareto Archive in MOEA.

MOEA maintains an archive of every conformation resulting from local search in it in order to compute the Pareto-based metrics described above. The (Pareto) archive stores the *current* Pareto rank and count for each conformation, as well as the values of the energy terms per the decomposition described above. When a new child conformation is sampled, it is compared to every existing member of the archive to compute its Pareto rank and count and update these metrics for all existing members of the archive. This approach allows MOEA to have a more global view of the search space through the Pareto archive than the view provided by only the current members of the population. It should be noted that the Pareto rank (and Pareto count) of a conformation in the archive can change over time, so a conformation that starts in the Pareto front will likely fall out of the Pareto front over time. This is the reason for re-evaluating the entire archive after adding a child conformation to it.

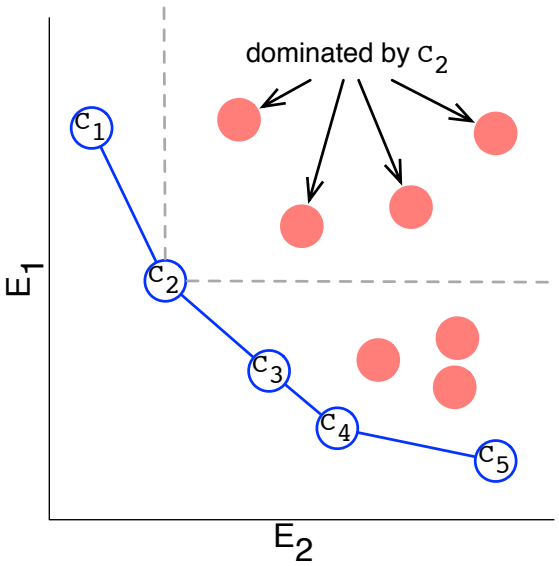

Figure 2: Conformations are plotted with respect to two energy terms, E_1 and E_2, to illustrate the Pareto-based metrics employed in MOEA. Conformations represented by empty blue circles are non-dominated and form the Pareto front (their Pareto rank is 0). The conformation C_2 dominates 4 other conformations, thus the Pareto count of C_2 is 4.

2.5 Population Selection in MOEA

MOEA employs truncation selection. Recall that in each generation, the parent conformations produce a same-size set of child conformations. A fraction (or all, based on the elitism rate employed) of the parent conformations are then combined with the child conformations, and the resulting set is sorted in ascending order. The sorting is conducted by Pareto rank, Pareto count, and total energy (conformations with the same Pareto rank are sorted by Pareto counts, and those with the same Pareto counts are sorted by total energy). We explore in this paper the importance of the Pareto count by comparing two versions of the algorithm, one that uses only Pareto front and total energy for the sorted order (which we refer to as MOEA), and another that uses Pareto front, Pareto count, and total energy (MOEA-PC).

The reason for additionally considering Pareto count is as follows. It is likely that, as the search progresses, the majority of combined parent and child conformations will be in the Pareto front. In this case, the Pareto count is used as a secondary ranking to distinguish between two conformations with the same Pareto rank. Pareto count is used to bias selection away from heavily-sampled regions of the Pareto front. The idea is that if a conformation in the Pareto front has a low Pareto count, then it is more likely that the search will be able to improve upon this conformation by finding a new conformation that dominates it. On the other hand, if the conformation has a high Pareto count, then this suggests than many attempts have already been made to improve upon this conformation.

Implementation Details.

In the HEA used for comparison, the elitism rate is set to 25%. MOEA (and MOEA-PC) uses an elitism rate of 100% ($s = p$) to ensure the greatest portion of the population will be taken from the Pareto front (it is unlikely that many of the child conformations will have a Pareto rank of 0). Since parents can fall out of the Pareto front as new children are added, MOEA does not suffer from premature convergence observed in EAs guided by total energy with an elitism rate of 100%. The size of a population is $p = 100$ conformations, and the number of generations is not fixed. Instead, MOEA (and MOEA-PC) is run for a fixed budget of $10,000,000$ energy evaluations (including those in local search). This results in $7-24$ hours of CPU time on a 2.4Ghz Core i7 processor, depending on protein length. Fixing the number of total energy evaluations ensures fairness among proteins of different lengths, as it is known that the time to evaluate potential energy increases quadratically with chain length.

MOEA uses fragments of length 9 for the initial population P_0 but switches to fragments of length 3 for higher resolution of the conformational space in subsequent generations. Additionally, while the initial population is obtained with a combination of `score0-score1` energy functions, as detailed above, local search uses `score3`, but `score4` is used for the Pareto metrics and potential energy in comparing a conformation to the archive and in population selection. Finally, Each realization (MOEA, MOEA-PC, and HEA) is run 5 times to remove differences in reported results due to stochasticity. The decoy ensemble reported for each combines child conformations generated across all generations.

3. RESULTS

Experimental Setting.

Here we analyze the performance of the proposed MOEA algorithm in the context of decoy sampling on a large and diverse list of target protein sequences. We measure the contribution of incorporating multi-objective optimization. In addition to comparing MOEA, MOEA-PC, and HEA to one another to determine which one has higher exploration capability and/or improves proximity to the known native structure, we compare all three to the MMC-based decoy sampling approach employed by Rosetta. Each algorithms is run 5 times, and the decoy ensemble reported and analyzed for each combines all 5 runs. A particular metric employed for comparison is minimum over all 5 runs (combined ensemble).

Target Protein Sequences.

We select 20 protein sequences, whose lengths vary from 53 to 146 amino acids, and folds encompass α, β, and α/β (we make no distinction between $\alpha+\beta$ and α/β folds). Details on these systems are listed in columns 1-3 in Table 1. The chosen protein sequences have known native structures, the Protein Data Bank (PDB), ids of which are also listed in Table 1 (column 1).

Comparison Metrics.

We use two metrics for comparison, lowest energy and lowest RMSD to the known native structure. Comparison of lowest potential energy over decoy ensemble allows comparing exploration capability. Since lowest potential energy does not necessarily capture the native basin, the second metric we employ is a popular dissimilarity metric, root-mean-squared-deviation (RMSD). RMSD sums deviations over atoms of corresponding amino acids in two conforma-

tions under comparison. We measure RMSD over C_α atoms as typical in ab-initio structure prediction. Prior to comparison, deviations due to rigid-body motions are removed. RMSD is non-descriptive above 8Å and increases with chain length. RMSD no higher than $5-6$Å is considered to have captured the native structure.

CRos: Implementation of Decoy Sampling in Rosetta.

What we refer to as CRos is the classic implementation of the coarse-grained decoy sampling stage in the full Rosetta ab-initio protocol. CRos is a long MMC search split into 4 substages, where the end conformation of a substage is the starting conformation for the following substage. Substage 1 uses `score0`, fragment length 9, and $2,000$ MMC moves to obtain a collision-free decoy conformation. Substage 2 switches to using `score1` to reward formation of secondary structures. Substage 3 uses `score2` and is longer, running for $20,000$ MMC moves to narrow in on an energy basin. Substage 4 switches to `score3`, fragment length 3, and uses $12,000$ MMC moves to optimize the decoy conformation though at coarse-grained level of detail.

Each substage employs a variable temperature schedule to allow probing into a minimum while not getting stuck; temperature is increased after a number of successive failed move attempts. As in a standard MMC search, each move is accepted with a probability given by the Metropolis Criterion, $p = exp(-\delta E/T)$, where δE is the difference in energy from proposed to current conformation, and T is a unitless measure of effective temperature, which serves to scale the change in energy. The value of T in CRos starts at 1 and increases to by 1 after 150 consecutive failed moves. Once a move is accepted, T is reset to 1.

Together, the substages result in a total of $36,000$ consecutive MMC moves, corresponding to the same number of energy evaluations in CRos. For the purpose of comparison, since Rosetta operates in a multi-start fashion, CRos is run for a total of $1,500$ times to obtain a decoy ensemble of $1,500$ conformations. This results in a total of $54,000,000$ energy evaluations which is slightly higher but in the same order as the total number of energy evaluations in the fixed budget in MOEA, MOEA-PC, and HEA (across all 5 runs).

3.1 Comparative Analysis of Lowest Energies

Table 1 shows the lowest energy sampled for the HEA, MOEA, MOEA-PC, and CRos on each of the target protein sequences in columns 4-7, respectively. For the purpose of this analysis, we consider two energy values within 2 kcal/mol of each-other as equivalent and use the following color scheme to facilitate comparison among the different algorithms. The lowest energy value in a row and other values in the row within 2kcal/mol of the lowest are drawn in red. The lowest energy value per row is highlighted in bold. The rows where CRos obtains lowest energy no higher than 2 kcal/mol of the lowest value in the row are highlighted in dark gray. The rows where either MOEA or MOEA-PC obtain lowest energy no higher than 2kcal/mol of the lowest value in the row are highlighted in lighter gray.

The color scheme used in Table 1 allows making a few observations. First, CRos achieves the lowest energy in only two systems and is within 2 kcal/mol of the lowest energy on only one other system (3 rows are colored in gray). On all the remaining 17 out of the 20 systems, the lowest energy is obtained by HEA, MOEA, or MOEA-PC. Either MOEA or MOEA-PC achieve the lowest energy or an energy no higher than 2 kcal/mol of the lowest value in a row on 14 systems (rows in light gray). Third, the lowest energy is obtained by MOEA-PC rather than MOEA on 9/14 systems (number of rows colored in light gray that have lowest value – in bold red – under MOEA-PC). Taken together, these results make the case that a hybrid evolutionary search algorithm led by Pareto-based metrics samples lower-energy regions than the coarse-grained decoy sampling in CRos (in fact, even the baseline HEA reaches lower lowest energies than Rosetta on the majority of systems). Moreover, the Pareto count is more effective than Pareto rank in this setting.

Table 1: Comparison of lowest energies.

PDB Id	n	Fold	lowest Rosetta Score4 Energy HEA	MOEA	MOEA-PC	CRos
1bq9	53	α/β	-50.5	-45.8	**-55.1**	-46.9
1dtdB	61	α/β	-55.0	-74.5	**-76.6**	-66.5
1isuA	62	α/β	-46.5	-48.4	**-76.7**	-27.0
1c8cA	64	α/β	-86.4	-98.4	**-101.5**	-101.4
1sap	66	α/β	**-121.4**	-120.1	-109.8	-107.8
1hz6A	67	α/β	-130.9	**-135.6**	-134.8	-117.1
1wapA	68	β	**-132.5**	-117.5	-121	-109.0
1fwp	69	α/β	-84.4	**-92.8**	-81.7	-71.3
1ail	70	α	-56.1	-67.1	**-71.1**	-29.9
1dtjA	76	α/β	-82.2	**-97.4**	-89.8	-72.5
1aoy	78	α/β	-98.1	-102	**-102.3**	-73.3
1cc5	83	α	-68.6	-67.8	-67.5	**-82.5**
2ci2	83	α/β	**-109.8**	-105.7	-102.4	-37.8
1tig	88	α/β	-128.0	**-151.7**	-136.1	-138.2
2ezk	93	α	-100.7	-93.4	**-101.1**	-51.1
1hhp	99	β	-104.5	-97.3	-96.0	**-106.3**
3gwl	106	α	**-100**	-95.3	-85.2	-68.2
2hg6	106	α/β	-102.6	-95.7	**-107.5**	-82.5
2h5nD	123	α	-129.0	-126.6	**-131.8**	-82.5
1aly	146	β	-81.1	**-117.1**	-103.6	-112.5

3.2 Comparative Analysis of Lowest RMSDs

Table 2 shows the lowest RMSD to native sampled for the HEA, MOEA, MOEA-PC, and CRos on each of the target protein sequences in columns 4-7, respectively. For the purpose of this analysis, we consider two RMSD values within 0.5Å of each-other as equivalent, and use the following color scheme to facilitate comparison among the different algorithms. The lowest RMSD value in a row and other values in the row within 0.5Å of the lowest are drawn in red. The lowest RMSD value per row is highlighted in bold. The rows where CRos obtains lowest RMSD no higher than 0.5Å of the lowest value in the row are highlighted in dark gray. The rows where either MOEA or MOEA-PC obtain lowest RMSD no higher than 0.5Å of the lowest value in the row are highlighted in lighter gray.

The color scheme used in Table 2 allows making a few observations. First, CRos achieves the lowest RMSD in 4/20 systems (rows where the other algorithms reach comparable RMSDs are highlighted in light gray). Lowest RMSDs reached are comparable on all other systems. Specifically, MOEA or MOEA-PC reach the lowest RMSD on 11 systems. However, as the coloring in Table 2 highlights, none of the algorithms show a consistent improvement in terms of the single lowest RMSD conformations sampled. Taken

Table 2: Comparison of lowest RMSDs

PDB Id	n	Fold	min $C\alpha$-RMSD (Å) HEA	MOEA	MOEA-PC	CRos
1bq9	53	α/β	3.0	3.4	3.3	**2.9**
1dtdB	61	α/β	4.4	5.3	5.0	**4.2**
1isuA	62	α/β	6.6	**6.4**	**6.4**	6.6
1c8cA	64	α/β	4.8	3.6	3.7	**2.2**
1sap	66	α/β	3.7	3.7	2.9	**2.8**
1hz6A	67	α/β	**1.9**	2.1	2.2	**1.9**
1wapA	68	β	**6.3**	6.4	6.4	6.5
1fwp	69	α/β	4.3	3.4	4.3	**2.8**
1ail	70	α	1.4	1.9	**1.1**	4.5
1dtjA	76	α/β	4.2	**2.3**	4.0	**2.3**
1aoy	78	α/β	3.9	**3.7**	3.9	4.0
1cc5	83	α	4.7	4.9	5.4	**3.7**
2ci2	83	α/β	**3.7**	3.9	**3.5**	5.8
1tig	88	α/β	3.2	**2.5**	3.9	**2.5**
2ezk	93	α	3.4	3.2	**2.9**	3.6
1hhp	99	β	8.8	**8.6**	8.9	10.1
3gwl	106	α	**5.4**	5.8	5.5	5.8
2hg6	106	α/β	9.3	9.6	**9.2**	9.4
2h5nD	123	α	6.2	7.5	**6.1**	7.4
1aly	146	β	**11.2**	11.4	**11.2**	12.4

together with the analysis on lowest energies, this shows that broader exploration may result in probing lower-energy regions but not necessarily lower-RMSD ones. This is due to the fact that the lowest-RMSD region is not necessarily the lowest energy one for a coarse-grained function. This has been elucidated for Rosetta in other studies focused on decoy sampling [17, 30].

3.3 Comparison of Conformational Ensembles

We now provide more detail on the distribution of RMSDs and energies. We first compare the distribution of RMSDs across HEA, MOEA, and MOEA-PC. The majority of the systems fall in one of two categories. In category one, illustrated on a representative protein in Fig. 3(a), both MOEA and MOEA-PC result in significantly more conformations with lower RMSDs to the native structure than HEA. In category two, illustrated on a representative protein in Fig. 4(a), HEA and MOEA are comparable but outperformed by MOEA-PC in Fig. 3(a). The third case, where the distributions are comparable among HEA, MOEA, and MOEA-PC, is only observed on protein with PDB id 1aoy (data not shown).

Energy vs. RMSD to native is now drawn for each conformation in the decoy ensemble for CRos, HEA, MOEA, and MOEA-PC on the two selected systems. These plots are shown in Fig. 3(c1, d1, e1) for protein with PDB id 1dtjA and in Fig. 4(c1, d1, e1), for protein with PDB id 1ail respectively. Fig. 3 shows that better funneling is obtained by MOEA and MOEA-PC (and better sampling of the native basin) over CRos and HEA. However, MOEA-PC seems to have explore in further detail a second non-native minimum (in agreement with summary results shown in Table 1. Fig. 4 shows that all algorithms, CRos included, reveal a rather flat basin for protein with PDB id 1ail (low correlation between low energies and low RMSDs) . Distinct funneling around the native structure is obtained by MOEA-PC, but non-native local minima are discovered, as well.

The three algorithms, HEA, MOEA, and MOEA-PC are further compared for the two selected systems in Fig. 3(c2, d2, e2) and in Fig. 4(c2, d2, e2). Energy vs. RMSD to native is plotted for each member of the population for a given generation. Conformations are colored based on how many generations they have remained in the population (results are shown only for the run achieving the lowest RMSD to the native over all 5 runs).

The investigation on the protein with PDB id 1dtjA in Fig. 3(c2, d2, e2) represents a common case, where there are many deep non-native local minima. Fig. 3(c2) shows that by using only total energy, HEA is unable to find the global minimum, while MOEA, shown in Fig. 3(d2) is able to follow multiple search paths and more effectively narrow in on the native basin. Fig. 3(e2) shows that the use of the Pareto count seems to hinder MOEA-PC, as the algorithm quickly narrows in on a low-RMSD basin but does not find the lower-energy and low-RMSD basin discovered by the MOEA.

The investigation on protein with PDB id 1ail in Fig. 4(c2, d2, e2) shows that, while the HEA samples lower-RMSD conformations early on, these conformations are soon discarded for lower-energy conformations (see Fig. 4(c2)). On the other hand, MOEA and MOEA-PC, shown in Fig. 4(d2), (e2), respectively, are able to effectively use the decomposed energy function to steer the search towards near-native conformations. In the case of MOEA-PC, the focus on sampling a more diverse set of conformations allows it to quickly identify the near-native portion of the search space.

4. CONCLUSION

The results in this paper show a distinct advantage to decomposing an energy function and employing multi-objective analysis to guide the exploration of the protein conformational space. The proposed MOEA is able not only to find lower-energy conformations but also sample significantly more near-native conformations than the baseline HEA. The use of the Pareto count to encourage further diversity in the search provides a dramatic improvement in sampling of near-native conformations in specific cases. Comparison with the decoy sampling stage in Rosetta suggests that incorporating latest representations and energy functions allows EAs to be competitive for decoy sampling. Moreover, multi-objective guidance enhances exploration capability and warrants further investigation. Work in computational structural biology, whether in the context of protein design or loop modeling, has started to investigate multi-objective optimization [14, 20]. We hope the work presented in this paper constitutes a first step into employment of EAs for decoy sampling and further investigation of these algorithms in more general settings of structure modeling for protein molecules.

Acknowledgment

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

5. REFERENCES

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

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

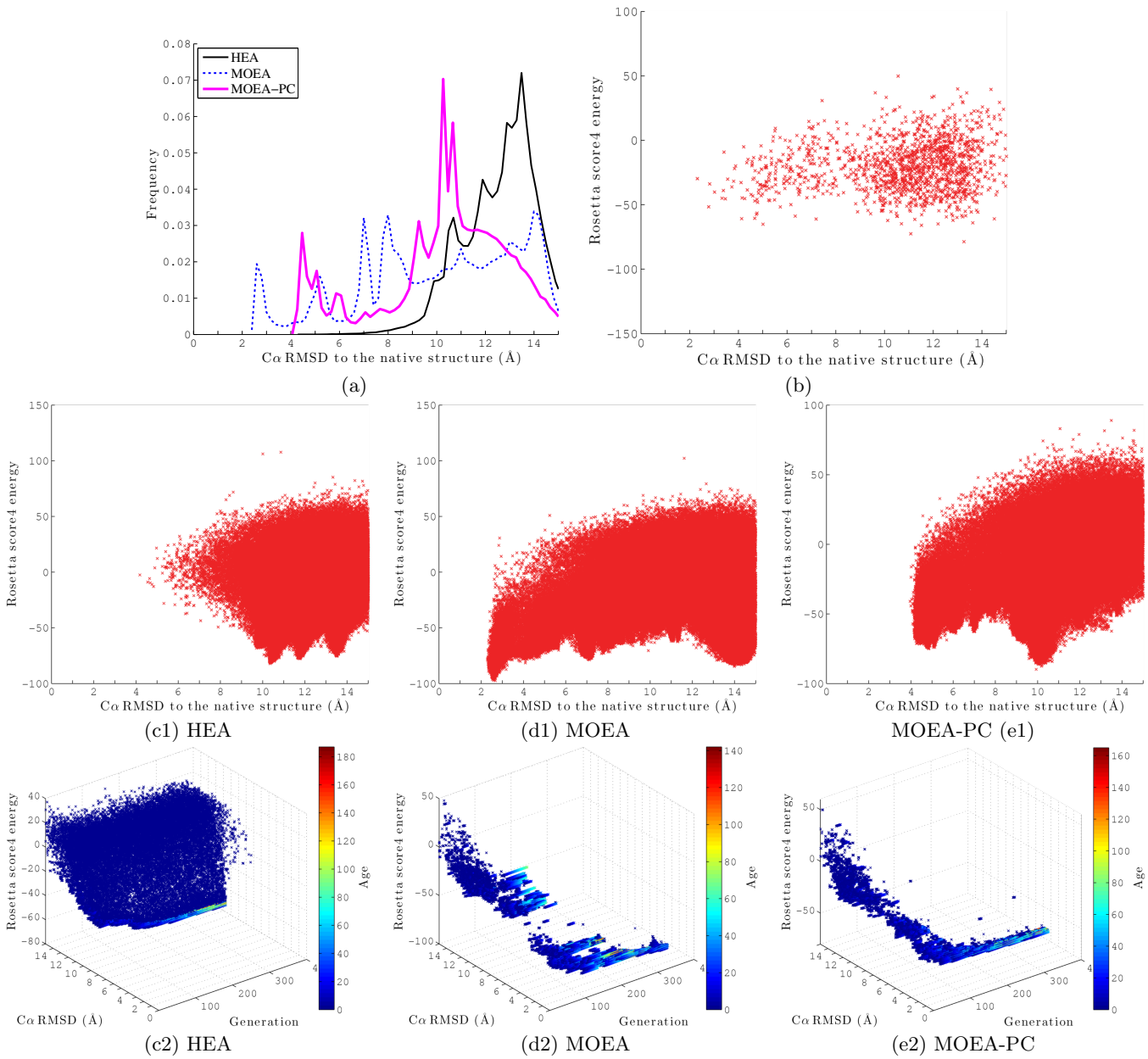

Figure 3: 1dtja, 76 aas, α/β: (a) Distribution of RMSDs of MOEA-obtained conformations from known native structure (dotted blue line) are superimposed over distribution obtained by MOEA-PC (purple line) and distribution obtained by HEA (black line). (b) Energy vs. RMSD plotted for CRos ensemble. (c-e)1) Energy vs. RMSD plotted for HEA, MOEA, and MOEA-PC, respectively. (c-e)2) Progress of Energy vs. RMSD over generations plotted, respectively.

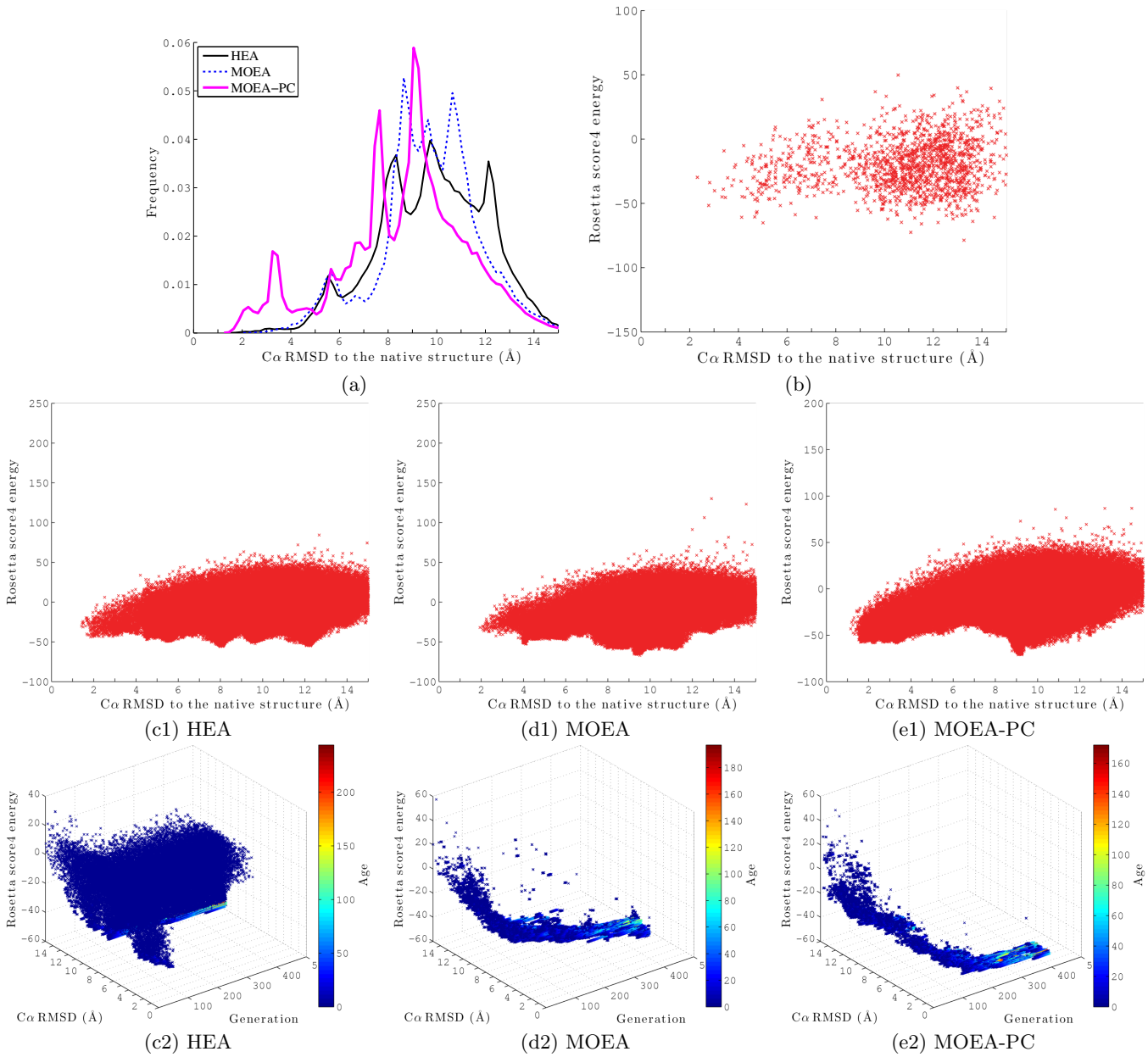

Figure 4: 1ail, 70 aas, α: (a) Distribution of RMSDs of MOEA-obtained conformations from known native structure (dotted blue line) are superimposed over distribution obtained by MOEA-PC (purple line) and distribution obtained by HEA (black line). (b) Energy vs. RMSD plotted for CRos ensemble. (c-e)1) Energy vs. RMSD plotted for HEA, MOEA, and MOEA-PC, respectively. (c-e)2) Progress of Energy vs. RMSD over generations plotted, respectively.

[4] 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.

[5] J. Calvo, J. Ortega, and M. Anguita. PITAGORAS-PSP: Including domain knowledge in a multi-objective approach for protein structure prediction. *Neurocomputing*, 74(16):2675–2682, 2011.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Heglera, J. Lätzera, A. Shehu, C. Clementi, and P. Wolynes. Restriction versus guidance in protein structure prediction. *PNAS*, 106(36):15302–15307, 2009.

[11] M. K. Islam, M. Chetty, and M. Murshed. Novel local improvement techniques in clustered memetic algorithm for protein structure prediction. In *IEEE Congress on Evol Comput (CEC)*, pages 1003–1011, 2011.

[12] G. G. Krivov, M. V. Shapovalov, and R. L. J. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *ProteinsSFB*, 77(4):778–795, 2009.

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

[14] Y. Li, I. Rata, and E. Jakobsson. Sampling multiple scoring functions can improve protein loop structure prediction accuracy. *J. Chem. Inf. Model.*, 51(7):1656–1666, 2011.

[15] S. J. Lutdke, D. H. Chen, J. L. Song, D. T. Chuang, and W. Chiu. Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. *Structure*, 12:1129–1136, 2004.

[16] K. Molloy and A. Shehu. Biased decoy sampling to identify near-native protein conformations. In A. Zhang, S. Ranka, T. Kahveci, M. Singh, and V. Honavar, editors, *ACM Bioinf. and Comp. Biol. (BCB)*, pages 131–138, Orlando, FL, October 2012.

[17] K. Molloy and A. Shehu. Probabilistic search and energy guidance for biased decoy sampling in ab-initio protein structure prediction. *IEEE Trans. Comput. Biol. Bioinform.*, 2013. in press.

[18] 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.

[19] G. Narzisi, G. Nicosia, and G. Stracquadanio. Robust Bio-active Peptide Prediction Using Multi-objective Optimization. In *Int Conf on Biosciences*, pages 44–50, 2010.

[20] L. G. Nivon, G. Moretti, and D. Baker. A pareto-optimal refinement method for protein design scaffolds. *PLoS One*, 18(4):e59004, 2013.

[21] B. Olson, K. A. De Jong, and A. Shehu. Off-lattice protein structure prediction with homologous crossover. In *Intl Conf on Genet. and Evol. Comput. Conf. (GECCO)*, Amsterdam, Netherlands, July 2013. in press.

[22] B. Olson, K. Molloy, and A. Shehu. In search of the protein native state with a probabilistic sampling approach. *J. Bioinf. and Comp. Biol.*, 9(3):383–398, 2011.

[23] B. Olson and A. Shehu. Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, 10(Suppl 1):S5, 2012.

[24] B. Olson and A. Shehu. Rapid sampling of local minima in protein energy surface and effective reduction through a multi-objective filter. *Proteome Sci*, 2013. in press.

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

[26] S. Saleh, B. Olson, and A. Shehu. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. *BMC Struct Biol.*, 2013. in press.

[27] A. Shehu. An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations. In *Robot: Sci. and Sys.*, pages 241–248, Seattle, WA, USA, 2009.

[28] A. Shehu. Conformational search for the protein native state. In H. Rangwala and G. Karypis, editors, *Protein Structure Prediction: Method and Algorithms*, chapter 21. Wiley Book Series on Bioinformatics, Fairfax, VA, 2010.

[29] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In A. Srinivas and M. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2 edition, 2013.

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

[31] V. N. Uversky. Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci*, 14:5188–5238, 2009.

[32] 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.

[33] 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.