Chapter []

A Review of Evolutionary Algorithms for Computing Functional Conformations of Protein Molecules

Amarda Shehu[†,‡,§]

[†]Dept. of Computer Science, George Mason University, Fairfax, VA, USA

[‡]Dept. of Bioengineering, George Mason University, Fairfax, VA, USA

[§]School of Systems Biology, George Mason University, Manassas, VA, USA

Abstract

The ubiquitous presence of proteins in chemical pathways in the cell and their key role in many human disorders motivates a growing body of protein modeling studies aimed at unraveling the relationship between protein structure and function. The foundation of such studies is the realization that knowledge of the structures a protein accesses under physiological conditions is key to a detailed understanding of its biological function and the design of therapeutic compounds for the purpose of altering misfunction in aberrant variants of a protein.

Dry laboratory investigations promise a holistic treatment of the relationship between protein sequence, structure, and function. Significant efforts are made in the dry laboratory to map protein conformation spaces and underlying energy landscapes of proteins. The majority of such efforts employ well-studied computational templates, such as Molecular Dynamics and Monte Carlo. The focus of this review is on a third emerging template, stochastic optimization under the umbrella of evolutionary computation. Algorithms based on such a template, also known as evolutionary algorithms, are showing promise in addressing fundamental computational challenges in protein structure modeling and are opening up new avenues in protein modeling research. This review summarizes evolutionary algorithms for novice readers, while highlighting recent developments that showcase current, state-of-the-art capabilities for experts.

Keywords: protein structure modeling, conformation space, energy landscape, conformational search, stochastic optimization, evolutionary computation, evolutionary algorithms

1 Introduction

Proteins are ubiquitous macromolecules in the cell as central components of cellular organization and function. Many diseases are due to misbehaving proteins, including critical human diseases, such as cancer, Amyotrophic lateral sclerosis, Alzheimer's, and other neurodegenerative disorders. The list of known gene mutations resulting in aberrant proteins misfunctioning in the cell is now growing (1,2). An important class of human diseases are due to proteins failing to adopt their native, biologically-active, three-dimensional (3d) structure with which they bind to small molecules or dock other macromolecules, giving rise to molecular interactions that make up all chemical reactions in the cell. While many such failures are deleterious, others lead to protein misfunction (3–9). Elucidating the long-lived structure or set of structures that an aberrant protein assumes and employs to interact with other molecules in the cell is key not only to a detailed understanding of how mutations impact function but also to pharmaceutical efforts to design effective compounds for blocking interactions of aberrant structures.

Investigations in the wet laboratory have elucidated by now over a hundred thousand active or functional structures of diverse proteins. As of May 2015, there are $108,957$ protein structures in the Protein Data Bank (PDB) (10). Increasingly, the focus is on rapidly resolving functional structures of possible protein constructs of decoded genomes and structures of proteins misfunctioning due to sequence mutations. Investigations in the dry laboratory are demonstrating their capability to complement wet-laboratory research and greatly enhance our understanding of the relationship between protein sequence, structure, and function. The role of dry-laboratory investigations is also expected to increase with the recently discontinued Protein Structure Initiative (11).

The foundation of dry-laboratory investigations is on the early experiments by Anfinsen, who demonstrated that a denatured protein spontaneously self-assembles into its native structure (12). The mechanistic treatment, which advocates that the native 3d structure of a protein determines its function, and that in turn the amino-acid sequence determines to a great extent the native 3d structure, is by now the basis of a growing body of protein research aiming to model structures and struc-

tural deformations relevant for biological function (13–15).

The most publicized work in protein modeling research is that on the protein structure prediction (PSP) problem, where the goal is to predict a structural representative of the active, functional state of a protein. The more challenging version of this problem, is known as *de novo* PSP (also referred to as *ab initio* PSP or *template-free* modeling, where *de novo* or *free* indicates the absence of a known structural template after which to model the unknown structure of the target protein sequence). While early investigations in silico pursued protein structure modeling with classic templates, such as Molecular Dynamics (MD), by now the most successful methods for *de novo* PSP build on the Metropolis Monte Carlo (MC) template (16). These methods aim to reveal the breadth of long-lived (and possibly functional) structures of a given protein sequence, as doing so constitutes a holistic treatment of the relationship between protein sequence, structure, and function. The holistic treatment also promises a detailed and comprehensive view of all possible long-lived structures a mutated protein sequence may assume to misbehave in the cell. However, at present, the computational demands of a holistic treatment are impractical for most existing methods (17).

The focus of this review is on an emerging group of methods known as evolutionary algorithms (EAs). EAs approach the problem of protein structure modeling under the umbrella of stochastic optimization and employ techniques from Evolutionary Computation (EC) to effectively find solutions of variable spaces with numerous and possibly inter-dependent variables. Though adaptations of EAs for the *de novo* PSP problem have been regularly pursued for decades, recent developments in computational structural biology regarding protein geometry and energetics have led to novel EAs with increased performance not only in terms of computational time but also in prediction quality. EAs are also gaining ground as effective tools to circumvent outstanding challenges in protein structure modeling and advance our ability to reveal detailed structure spaces of healthy versions, wildtype sequences of a protein, and unhealthy, aberrant variants.

This review appeals to both experts and novices. A short primer is provided first on protein geometry and energetics. Connections are then made between

protein structure modeling and stochastic optimization. A summary of EAs for *de novo* PSP is provided next, highlighting recent developments for experts. A growing group of EAs is also presented to showcase their ability to map structure spaces of healthy and aberrant versions of a protein. The review concludes with a summary of outstanding challenges and opportunities for the EC community in protein structure modeling.

2 Protein Geometry and Energetics

2.1 Protein Geometry

A protein molecule consists of one or more polypeptide chains. A polypeptide chain is comprised of many peptides, or reduced amino acids, bound in a serial fashion through covalent bonds. Many polypeptide chains can be held together through non-covalent interactions in quaternary structures of protein assemblies or complexes. In this review, we will focus on protein molecules comprised of only one chain. Single-chain proteins are predominantly the subject of PSP and structure modeling in general (structure modeling of protein complexes is known as protein-protein docking and is beyond the subject of this review).

Amino acids are the fundamental building blocks of proteins and come in twenty different naturally-occurring types. They are assigned 20 different names, which have abbreviated three-letter and one-letter codes. An amino acid consists of a central group of heavy atoms that is shared among all amino acids, a side-chain group of atoms that gives an amino acid its unique chemical properties (and its type), and hydrogens. The commonly-shared group of atoms consists of N, CA, C, and O. These are known as the backbone atoms. If one follows the thread of these atoms through the peptide bonds that links the terminal backbone C of one amino acid to the terminal backbone N atom of another, a skeleton or backbone can be traced that gives a protein chain its connectivity. The side chains dangle off the backbone and both guide and constrain its motions. Fig. 1(a) draws the native structure of a protein in 3d, highlighting the backbone. A short fragment of a few

amino acids is drawn in greater detail in Fig. 1(b).

2.1.1 Representation of a Protein Chain: Conformation Space

While the covalent bonds provide local rigidity, the protein chain (both backbone and side chains) is highly flexible. This intrinsic flexibility necessitates introducing the concept of a conformation, which refers to the spatial arrangement of the atoms in a protein chain.An all-atom conformation refers to the fact that detailed information is available at all times regarding all atoms. Not all atoms need to be explicitly modeled. Many algorithms for PSP primarily model the backbone, and once a set of (functional) conformations likely to comprise the native state of a protein are available, side-chain atoms are packed in optimal configurations with side-chain packing techniques. When not all the atoms of an amino acid are modeled explicitly by neglecting some atoms or grouping atoms together in pseudo-atoms, the conformation is said to employ a reduced, or coarse-grained representation of a protein chain. Research into effective reduced representations for protein conformations is active (18).

A conformation does not need to be represented internally in a computer code as a list of atomic coordinates. Generally, a conformation is a list of instantiated variables or parameters. These variables can be discrete, such as positions on a lattice or bits encoding positions or angles, or continuous, such as cartesian coordinates or angles. The former are known as discrete representations, whereas the latter are known as continuous representations. There are advantages and disadvantages with either, as summarized briefly below.

Discrete Representations: Atoms on a Lattice The earliest representations of protein chains were on a 2d or 3d lattice. Typically, only the CA atom of each amino acid is explicitly modeled and restricted to lie on the lattice (19). The lattice restricts bonded atoms to neighboring cells and allows both fast integer-math evaluations of conformational energies, as well as enumeration of all self-avoiding walks on the lattice (20–22). On-lattice deterministic search algorithms were useful to elucidate various organizing principles of amino acids during fold-

ing; the number of amino acid types was often restricted to 2, hydrophobic (H) vs. hydrophilic/polar (P), resulting in the popular HP model, to allow calculations on chains of more than a dozen amino acids. On-lattice representations allowed also obtaining a theoretical understanding of the PSP problem and facilitated various complexity results (23–25). While the majority of conformation sampling algorithms nowadays has moved away from on-lattice models, significant research on EAs for PSP still employs them. Several types of lattice models are pushed forward in the EC community, such as triangular, cubic, and face-centered-cubic. In general, the top-performing algorithms for PSP employ off-lattice representations. The reason is that on-lattice representations sacrifice a lot of structural detail and have been shown to reproduce the backbone of a known native structure with accuracy no greater than half the lattice spacing (26); some lattice representations have also been shown to bias towards specific secondary structures (27). It is worth noting that on-lattice representations are nowadays the only computationally-reasonable representations for very large proteins of hundreds of amino acids (28).

Continuous, Off-lattice Representations The majority of MC-based conformation sampling algorithms for PSP employ continuous representations, where atoms are not forced to occupy a limited number of positions on a 2d or 3d lattice. Two popular representations are the cartesian-based and the angular-based ones. There are advantages and disadvantages with either.

Cartesian-based representations: Cartesian-based representations are straightforward for conducting energetic calculations, as summarized below, because most protein energy functions contain terms that are distance-based. Cartesian-based representations are also typically computationally demanding. In naive cartesian-based representations, the number of variables (cartesian coordinates) for a protein chain of N atoms is $3N$. A small protein chain contains hundreds of atoms. Cartesian-based representations do not allow trivial satisfying explicit constraints on locations of atoms, such as those imposed by covalent bonds. While

covalently-bound atoms do oscillate, large oscillations carry heavy energetic penalties. So, it is often computationally desirable to preserve the lengths of covalent bonds in computed conformations. However, a computer algorithm that modifies cartesian-based variables to compute new conformations will move atoms independently of one another and break bonds. Less naive cartesian-based representations that preserve local constraints exist, and they often build on statistical analysis techniques to define variables that encode collective motions of atoms.

Angular-based representations: In addition to covalent bonds, there are other local constraints in native protein structures whose violation results in high energetic penalties. Optimal valence angles (between two consecutive covalent bonds) observed to remain constant and depend only on the types of atoms involved in protein functional conformations would also be changed if one were to employ cartesian-based representations. Instead, angular-based representations that model only dihedral angles (see Fig. 1(b) for an illustration of such angles) save on both the number of variables (on average, there are $3N/7$ dihedral angles in a protein chain of N atoms (29)) and the number of constraints that are violated. The only constraints that angular-based representations cannot readily satisfy are long-range ones resulting from energetically-constrained motions of non-bonded atoms. Typically, the violation of such constraints is evaluated through energy functions. Increasingly, while cartesian-based representations are employed by MD-based methods for simulating the dynamics of a protein molecule, angular-based representations are the preferred ones in MC-based methods and those particularly designed for *de novo* PSP. Angular-based representations necessitate that a transformation occur from angles to cartesian coordinates (30) in order to evaluate a computed conformation with specified energy function.

Distance Functions for Conformations

Measuring the distance between two conformations is key to the ability to report the performance of a PSP algorithm to reproduce a known native structure. Depending on the representation employed, several distance functions can be useful.

For instance, Hamming and Manhattan distance can be useful for discrete representations, such as on-lattice ones. Continuous representations allow the employment of Euclidean-based distance functions. The majority of PSP algorithms that employ off-lattice representations make use of the popular RMSD function to measure the distance between two conformations.

Root-Mean-Squared Deviation (RMSD) RMSD is a Euclidean-based dissimilarity metric to measure the distance between two conformations. Given two conformations C and $C^{'}$ of N atoms, where $p_i(C)$ indicates the position of atom i in conformation C, $\text{RMSD}(C, C^{'}) = \sqrt{\sum_{i=1}^{N} |p_i(C) - p_i(C')|^2}/N$. Prior to measuring the RMSD between a computed conformation and the native structure, an algebraic procedure is carried out that determines the optimal superimposition removing differences due to rigid-body motions in 3d (translations and rotations of the whole conformation) (31). The term "least" is sometimes explicitly added, as in least RMSD (lRMSD), to indicate that the conformations have undergone rigid-body motions so as to report their lowest RMSD from the known native structure. It is generally understood that even when RMSD is reported, it is lRMSD. It is also worth noting that even conformation sampling algorithms that use continuous, angular-based representations make use of RMSD to report results. The homogeneous transformations encoded by the angles in a conformation represented as a list of angles can be accumulated to obtain the cartesian coordinates of the atoms over which angles are defined. While angular-based distance functions are available, they are not widely adopted in structure modeling literature.

2.2 Protein Energetics

Current protein energy functions are based on molecular mechanics, summing over favorable and unfavorable atomic interactions (and possibly with surrounding solvent) to associate a potential energy value with a conformation. Interactions between atoms are classified as bound (local, due to covalent bonds) or non-bound (non-local due to non-covalent interactions). Local interactions concern bonds, bond angles, and the periodicity of dihedral angles. Non-local interactions are di-

vided based on their physical nature, electrostatic (measured through the Coulomb function) or van der waals (measured through the Lennard-Jones – LJ– function). The latter interactions are estimated via distance-based power terms responsible for the computational cost and nonlinearity of protein energy functions. Equation 1 shows the functional form of the CHARMM potential energy function (32).

$$
\begin{aligned}
E_{\text{CHARMM}} = & \sum_{\text{bonds}} k_b \cdot (b - b_0)^2 & + \\
& \sum_{\text{UB}} k_{UB} \cdot (S - S_0)^2 & + \\
& \sum_{\text{angles}} k_\theta \cdot (\theta - \theta_0)^2 & + \\
& \sum_{\text{dihedrals}} k_\chi \cdot 1 + cos(n\chi - \delta) & + \\
& \sum_{\text{impropers}} k_{\text{imp}} \cdot (\phi - \phi_0)^2 & + \\
& \sum_{\text{non-bond}} \epsilon_{ij}[(\frac{Rmin_{ij}}{r_{ij}})^{12} - (\frac{Rmin_{ij}}{r_{ij}})^6] & + \\
& \sum_{\text{non-bond}} \frac{q_i q_j}{\epsilon r_{ij}}
\end{aligned}
\tag{1}
$$

In Equation 1, the k_* terms are constants, and the $*_0$ terms are ideal values of variables. The Urey Bradley (UB) term is calculated over pairs of atoms separated by two covalent bonds (known as the 1,3 interaction), and S is the distance between the atoms. The n and δ variables in the fourth term are the multiplicity and the phase angle, respectively. In CHARMM, improper dihedral angles are penalized according to the formula shown in the fifth term. The sixth term measures the LJ interactions: r_{ij} measures the Euclidean distance between two non-bonded atoms (that are not covered by the UB term), ϵ_{ij} is the LJ well depth, and $Rmin_{ij} = (Rmin_i + Rmin_j)/2$ is the minimum interaction radius between the atoms, measured as half the sum of the known van der waals radii. The LJ term in CHARMM has a $12 - 6$ functional form, whereas other physics-based functions

may have a $12-10$ functional form. The final term measures electrostatic interactions via the Coulomb functional form: q_i measures the known partial charge of atom i and ϵ is the dielectric constant encoding the type of environment (vacuum, solvent).

Different energy functions may have different functional forms and even employ a different list of terms; for instance, some treat hydrogen interactions differently. This is particularly the case for knowledge-based functions, which may also contain additional terms based on statistically-observed interactions calculated over databases of protein structures. Whether physics-based, knowledge-based, or hybrid functions that combine both physics-based and knowledge-based terms, current protein energy functions are semi-empirical. In addition to decisions on how many terms and what the terms should capture, important decisions are made to weight the contribution of each term so as to reproduce experimental measurements on specific subsets of protein structures. Moreover, most energy functions limit interactions to pairwise ones. Energy functions that calculate multi-body interactions often outperform pairwise-based functions in reproducing experimental kinetic data, but their computational cost remains high to be practical for most protein structure modeling algorithms (18).

2.2.1 Protein Energy Surfaces

If one were to organize all conformations of a protein chain on horizontal axes and the potential energy corresponding to each conformation in a vertical axis, the view that would emerge would be a funnel-like (multidimensional) energy surface (33, 34). If one were to find few collective (also referred to as reaction) coordinates that discriminate among the important structural states, projecting the surface onto these coordinates would give rise to the energy landscape, where thermodynamically-available states would be easily discerned as basins (35). A classic landscape is shown in Fig. 1(c).

Horizontal cuts would reveal energetic states (and thus conformations of comparable potential energies). The width of these cuts goes down as energy decreases in a true protein energy surface. This width is captured in the concept of entropy,

which measures the degree to which a protein chain can assume different conformations while maintaining the same potential energy (within a dE). An entropy value can be associated with each energetic state; thermodynamically-stable states are those with low free energy F, measured as F = ⟨E⟩ - TS, where ⟨E⟩ is average potential energy over conformational ensemble corresponding to the state, T is temperature, and S is entropy. Evolutionary bias has been found to be the reason for why the native state in naturally-occurring proteins has lowest free energy (12).

Long-lived states in proteins correspond to deep and wide basins. The exact details of the contribution of potential energy versus entropy are what determine whether a basin corresponds to the most thermodynamically-stable (longest-lived) and thus the native state of a protein or a semi-stable state. In many proteins, complex energy surfaces are emerging, where more than one structural state is employed in conformational switch mechanisms that modulate function and gives rise to rich functionality. In many aberrant versions of a protein, energy barriers between stable and semi-stable states drastically change and modify the underlying detailed structural mechanism regulating function, resulting in misfunction.

In view of complex protein energy surfaces, conformational search algorithms need to elucidate not just one putative global minimum but map the breadth of low-energy conformations corresponding to local minima in the underlying energy surface. Visualization of this surface via a low-dimensional energy landscape may reveal a multitude of basins that are worth considering, whether the goal is to select from them the one corresponding to the predicted native state or understand the mechanisms through which a protein and its aberrant versions may make use of more than one basin for function modulation and misfunction.

3 PSP as an Optimization Problem

Whether cartesian-based or angular-based representations are employed, it is generally expected that the representation of a protein chain of hundreds of amino acids will result in hundreds of variables. Thus, the conformation space is expected to be high-dimensional. The space can be discretized, but the number of

variables makes enumeration impractical as a way of computing conformations of a protein chain. It is worth noting that in the early days, when short polypeptide chains were being investigated to extract physical principles of protein folding, bonded atoms were forced to occupy neighboring cells in a 2d or 3d lattice (19). Combinatorial techniques could be employed to enumerate conformations of short protein chains (20–22), but finding the lowest-energy conformation on a lattice has been proved to be NP-hard (23–25). While lattice representations allow tackling very large proteins of several hundred amino acids, PSP methods designed for small-to-medium size proteins not exceeding 200 amino acids can afford to produce higher-quality conformations with more accurate backbones by employing off-lattice representations.

When off-lattice representations are considered, the conformation space is expected to be vast, high-dimensional, and continuous. As described later, discretization can still be employed (as in the popular molecular fragment replacement technique), but the exponential explosion in the number of resulting conformations makes enumeration impractical. As a result, only stochastic or probabilistic algorithms can be employed to essentially sample the conformation space one conformation at a time. Such algorithms essentially build sample-based representations of conformation spaces. Since the goal is to map low-energy regions of the underlying energy surface where deep and wide basins may be found that correspond to the native, longest-lived structural state, such algorithms implement stochastic optimization of complex, nonlinear and multimodal energy functions.

Stochastic optimization algorithms forsake completeness, as no guarantees can be made even on their ability to find, for instance, the global minimum energy conformation or GMEC (a term coined by Scheraga and colleagues (36)). It is worth noting that the GMEC may not correspond to the native structure, after-all. One reason is due to inaccuracies in energy functions. The global minimum of even state-of-the-art all-atom energy functions can be more than 4Å off the true native structure (37). Another reason is due to the thermodynamic argument that the native state is not necessarily the one with the lowest energy but the one that compromises between potential energy and entropy. A deep but narrow basin may

not only be an artifact in an energy function but also possibly a kinetic trap. The native basin may be deep enough and wide enough to prevent fast escapes and allow structural flexibility at equilibrium.

There are currently three groups of stochastic optimization methods. The first group builds on the MD template and essentially follows the negative gradient of a given energy function to find local minima. The second group builds on the MC template and makes use of repeated moves or perturbations to hop between conformations while generally lowering potential energy. The second group of methods and the subject of this review, EAs, builds on the EC template and shares many characteristics in its core functional units with MC-based methods. Indeed, as our exposition shows, MC-based methods can be classified as specific EAs. We summarize the popular MD and MC templates before proceeding with EAs in the next section in order to better appreciate the algorithmic differences among these three groups of methods.

MD-based methods simulate the dynamics of a physics-based system. An MD trajectory is initiated from a protein conformation and systematically searches the conformation space one conformation at a time by numerically solving Newton's equations of motion. These are integrated to obtain a conformation $C_{t+\delta t}$ at time $t + \delta t$ from a current conformation C_t at time t in the trajectory. All atoms in C_t are modified in the direction of the calculated forces, allowing the MD trajectory to essentially follow the slope of the potential energy function. Newton's equations of motions are used to update the position and velocity of each atom in time. While velocity is initialized at some random value, accelerations are computed from the (negative) gradient of the potential energy function. Repeated application of the equations of motions dictates that a small timestep δt in the order of 1-2 femtoseconds be used so that the calculated gradient at each conformation in the trajectory closely follow the curvature of the potential energy function. The small timestep limits the breadth of conformation space that an MD trajectory can explore. Significant advances in dedicated, specialized hardware for MD simulations, parallelizations, and other enhanced sampling techniques have pushed the capability of MDs and their ability to capture molecular processes in the order of

micro-to-milli seconds (38–40).

It remains hard for MD trajectories to reach the length and time scales needed to follow transitions between unfolded and folded states, or vice-versa, or between other stable states. Moreover, gradient calculations are more easily conducted in cartesian space, which results in a vast search space. Modifications to conduct MD search over the space of dihedral angles have been proposed (41–43). In essence, an MD trajectory realizes local search, and it is bound to converge to a local minimum in the energy surface. For these reasons, most methods with high sampling capability employ a thermodynamic rather than a kinetic treatment in the interest of computational efficiency. Monte Carlo (MC)-based methods fall in this category.

In contrast to MD, conformations in an MC trajectory are not obtained by following the slope of the energy function but are the direct result of designed moves. The moves change values of the underlying variables and do not have to be physically-realistic as long as they are coupled with the Metropolis criterion $e^{-\Delta E/T}$. ΔE is difference in energy between the conformation resulting from the move, and T (effective temperature) is a scaling parameter that determines whether an energetic increase can be accepted or not. The result is a series of conformations that still converges to a local minimum but has the ability to cross over energetic barriers by controlling T. The MC template has higher sampling capability than the MD one, as moves can be designed to allow larger jumps in conformation space. However, because designed moves do not have to encode realistic (physics-based) motions, any information on whether and when the protein chain can diffuse from a computed conformation to another (thus, actual timescale information) is lost.

4 EAs for PSP and Mapping Energy Landscapes

We first summarize the unifying template of EAs and even show how MC can be regarded as a specialized EA. We then provide a summary of EAs for PSP, organizing it around principal algorithmic components. Recent EAs with high

performance on PSP are highlighted in greater detail. The section concludes with an exposition of a recent group of EAs that go beyond the narrow focus of single-structure prediction in PSP and instead mapping the breadth of functionally-relevant structures and corresponding basins in the energy landscape.

4.1 Unifying Template of EAs for PSP

The realization that protein energy functions are nonlinear and multimodal, and that PSP can be cast as a global optimization problem has motivated many researchers in the EC community to approach PSP with specialized EAs. One of the first works demonstrating the promise of EAs for PSP proposes a genetic algorithm (GA) (44). Several lattice-based and off-lattice EAs have been proposed since then. Before summarizing the developments in a little over a decade, the unifying template that EAs follow is summarized first.

The basic EA template mimics the process of evolution and natural selection to find local minima of a complex objective/fitness function. The template evolves a population of conformations (generally referred to as individuals) over a number of generations. A mechanism needs to be specified to generate the initial population, which can consist of conformations sampled at random over the employed variables (especially in the context of PSP, where only the amino-acid sequence is provided) or conformations provided by domain experts, such as wet-laboratory investigators, corresponding to known structures (in other applications that go beyond PSP and aim to map energy landscapes).

The population is allowed to evolve over generations such that individual (conformations in EAs for protein structure modeling) with low potential energy values (high fitness) are repeatedly selected and improved upon. In each generation, a selection mechanism is specified to select parent conformations for producing new conformations (offspring). The mechanism can be based on energies or other measures, incorporating various heuristics on what is more likely to lead to low-energy (high fitness) and possibly (structurally) diverse offspring. Popular selection mechanisms are truncation-based, fitness-proportional, tournament-based, and others (45). The injection of structural diversity in the selection mechanism is

particularly important to diversity a population and often credited with avoiding premature convergence to select local minima.

Once parents are selected, asexual perturbation or reproductive operators that modify/mutate one parent at a time or sexual operators that combine parents through cross-over are invoked to compute new individuals, offspring. A survival mechanism determines which individuals survive to the next generation. In non-overlapping or generational survival mechanisms, the offspring replace the parents. In overlapping ones, a subset of individuals from the combined parent and offspring pool are selected. Survival mechanisms may be based on fitness or consider other criteria (such as for instance, structural similarity of conformations) in order to steer the algorithm over the generations to optima of the fitness landscape.

EAs that employ crossover in addition to the mutation operator are often referred to as genetic algorithms, or GAs. EAs that additionally incorporate a meme, which is a local search/improvement operator to optimize a child and effectively map it to a nearby local minimum, are referred to as hybrid or memetic EAs (MAs). The employment of multiple objective functions as opposed to a single fitness function results in multi-objective EAs (MO-MAs). Specific variants that build over GA are respectively referred to as MGAs and MO-GAs.

Customized EAs for PSP contain many other evolutionary strategies and metaheuristics, such as the employment of a hall of fame to preserve "good" solutions, tabu search to improve the performance of a meme, co-evolving memes, niching, crowding, twin removal for population diversification, structuring of the solution space to facilitate distributed implementations capable of exploiting parallel computing architectures, and more. The combination of all these strategies and more (Ref. (45) provides a comprehensive review on EAs for stochastic optimization) give rise to different, powerful EAs.

EAs have been demonstrated effective for sampling low-energy protein conformations. Though MC-based methods for PSP are generally more accepted and popular in computational structural biology, EAs have less of a chance of getting stuck at local minima compared to MC search (44). Recent adaptations of EAs

that employ state-of-the-art domain-specific knowledge on proteins, such as off-lattice, coarse-grained, angular representations of protein chains, state-of-the-art protein energy functions, and the popular molecular fragment replacement technique in perturbation and improvement operators, have been demonstrated to be competitive with state-of-the-art MC-based methods (37, 46–49).

4.2 Performance Measurements of EAs for PSP

There are typically two measurements used to assess the performance of an EA. When the goal is to compare the addition of novel algorithmic components and heuristics in a customized EA for PSP against a baseline EA, performance is assessed based on the lowest energy reached by each algorithm over the course of the execution. The termination criterion is set in terms of number of generations or total energy evaluations allowed. The latter allows for fair comparison of EAs with MAs. The second performance metric assesses the ability of the algorithm to compute the known global optimum, that is, the known native structure of a protein. The metric of choice is the least RMSD metric summarized in section 2. The lowest RMSD from the native structure over conformations obtained by a conformation sampling algorithm is recorded and reported as the closest that the algorithm comes to the known native structure. In EAs, this calculation is often carried out over all conformations ever computed (over all generations) as opposed to only those in the final population or those in the hall of fame (if a hall of fame is employed). The reason for doing so is that it is not uncommon for a good solution to be lost in later generations.

4.3 MC as 1+1 EA: Basin Hopping as a Specialized MA

We now provide further understanding on why EAs are promising avenues to approach PSP by first demonstrating that they encapsulate MC-based methods.

MC can be cast as a 1+1 EA. Since an MC trajectory is a series of conformations, where C_{i+1} is the result of applying a move on C_i in the trajectory, an MC trajectory of n conformations can be viewed as an EA of n generations. In each

generation i, the only individuals C_i is subjected to a perturbation operator that employs a designed move, and the result of that operator is conformation C_{i+1}. A non-overlapping, generational model replaces C_i with C_{i+1}; that is, C_{i+1} is the only conformation retained in the population of the next generation. This is a standard MC algorithm. In a specialized version, known as the Metropolis MC, a probabilistic criterion is employed to determine whether C_{i+1} is retained in the trajectory or another attempt/move needs to be made on C_i. Even Metropolis MC can be viewed as a 1+1 EA. Instead of the generational model, the parent and offspring are combined, and a probabilistic criterion is employed to determine which one survives in the next generation.

An interesting adaptation of Metropolis MC has been proposed to address PSP in the computational structural biology community. The adaptation concerns additionally subjecting each generated conformation C_{i+1} to an energetic minimization technique that maps C_{i+1} to a local minimum in the energy surface. The conformation representing the local minimum, $C_{i+1}*$ is the one considered for addition into the trajectory through the Metropolis criterion. Essentially, the MC trajectory is a series of local minima, or basins, and the algorithm has also been referred to as basin hopping (BH). BH is a specialized EA. The energetic minimization can be considered a local improvement iterator, thus making BH a 1+1 MA.

The history of BH in computational structural biology is rich and can be traced to work by Wales and Doye on obtaining the LJ minima of small atomic clusters (50). When considering that BH is an MC with minimization, its roots go even deeper to the "MC with minimization" methods proposed by Scheraga and colleagues (36,51). Simultaneous work on BH for addressing challenging real-life problems has appeared in the AI community. In particular, in the EC community, BH is also known as Iterated Local Search and is popular for solving discrete optimization problems (52).

Recently, BH algorithms has seen a comeback in computational structural biology. BH algorithms essentially differ in how they implement the perturbation and improvement operators. For instance, the perturbation predominantly modifies atomic coordinates, and minimization is either a gradient descent or a

Metropolis MC at low temperature (37,53–55). Application for PSP in (49) shows that cartesian-based representations are the culprit of decreased efficiency and efficacy on capturing the native structure on protein sequences longer than 75 amino acids. Adaptations of BH algorithms to employ angular-based representations and the fragment replacement technique in the perturbation operator have resulted in competitive performance with top MC-based conformational sampling algorithms in PSP. We highlight one such algorithm below. The reader is referred to Ref. (56) for a review on BH algorithms for protein structure modeling.

4.3.1 Highlight: Basin Hopping for PSP

Work in (46) extends the applicability of BH for PSP in proteins more than 120 amino acids long. This is mainly a result of employing the molecular fragment replacement technique in the perturbation operator. The technique is popular with the top conformational sampling algorithms for PSP and other structure-related problems and central to their performance (46, 47, 57–66). Its popularity is due to the fact that it allows rapidly computing conformations with credible secondary structures. Below we briefly summarize it.

Molecular Fragment Replacement The basic idea is to bundle together consecutive dihedral angles of k amino acids (k is typically 3 and 9). The segment $[i, i + k - 1]$ in the protein chain is referred to as a fragment, and the dihedral angles corresponding to the fragment are referred to as the fragment configuration. A move consists of replacing values of all these angles simultaneously with values obtained from a pre-compiled library (often referred to as a library of fragment configurations). The library is pre-compiled from known, non-redundant protein structures. The chain of each structure is excised in overlapping fragments, and configurations are stored organized by their amino-acid sequence. Making a move on a conformation C to generate a new conformation consists of the following three steps: a position i in the protein chain/sequence is sampled at random. A fragment $[i, i + k - 1]$ is then defined. The library is queried with the amino-acid sequence of the fragment. Configurations of fragments with the identical (or sim-

ilar) sequence are then collected, and one is selected at random to replace that of the fragment in conformation C. Details on constructing fragment configuration libraries are presented in (67). A representative PSP package that employs MC-based conformational sampling algorithms with the molecular fragment replacement technique is the Rosetta package (68).

PSP-BH with Molecular Fragment Replacement The BH-based algorithm in (46) samples local minima within $5 - 6$Å at most of the native structure on diverse proteins and is thus competitive with MC-based state-of-the-art sampling algorithms for PSP. Work in (46) shows a strong correlation between RMSD to the native structure and RMSDs between consecutive local minima. Based on this finding, later work in (47) introduces techniques to control the distance between consecutive local minima and thus further improve proximity to the native structure. Work in (47) also shows that simple greedy search in the local improvement operator is just as effective but more efficient that MC-based improvement.

4.4 Population-based Off-lattice and On-lattice EAs for PSP

We now summarize state-of-the-art population-based EAs and algorithmic components responsible for recent advances.

The popularity of lattice-based representations in the early 1980s in protein structure modeling motivated development and adaptations of EAs for a simplified instantiation of the PSP problem. Significant work in the EC community on PSP still employs a lattice-based HP model of a protein chain, where an amino acid is modeled as a bead in 2d or 3d, two types of beads are considered (hydrophobic/H versus hydrophilic/P), and amino acid beads are positioned on a 2d or 3d lattice. In essence, the PSP problem is simplified, and the objective becomes finding an on-lattice self-avoiding walk that minimizes the interaction energy among amino-acid beads. Lattice-based representations facilitate the design of simple perturbation operators and are amenable to simplistic energy functions for summing up interactions and scoring conformations. The employment of lattice-based representations reduces the typical computational demands of PSP and allows fo-

cusing on algorithmic design and analysis, particularly regarding an optimal the interplay of exploration versus exploitation. A comprehensive review of on-lattice EAs can be found in Ref. (69). In the following, we review salient search strategies demonstrated effective on on-lattice and off-lattice EAs and then highlight recent developments that position EAs as competitors with top MC-based PSP conformation sampling algorithms.

4.4.1 Hybridization to Balance Global and Local Search

In the EC community it is well understood that, for complex optimization problems, simple EAs are not sufficient for achieving the necessary balance between exploration and exploitation. As a consequence, there continues to be an interest in developing more complex EAs capable of achieving this balance on complex fitness landscapes rich in local minima. One direction concerns the design and implementation of hybrid EAs that combine population-based global search techniques with domain-specific local search methods.

There are a variety of ways in which local search methods have been embedded in EAs. MA is the most well-known hybridization approach, based on the idea that a top-level EA manages a population of local searches (memes) with the goal of maintaining a diverse set of memes (exploration) while exploiting efficient local search methods with memes. Other less well-known approaches include Baldwinian EAs, Lamarkian EAs, cultural algorithms, and genetic local search (70–73). MAs have been first adapted for conformation sampling in PSP for toy or short peptides, employing the lattice-based HP model (74–79).

Work on on-lattice EAs has has demonstrated that the addition of local searches or memes is particularly effective when crossover is employed to combine features from multiple previously-sampled conformations (80). In a rugged landscape, offspring obtained through crossover are highly likely to have low fitness. This is particularly the case for protein conformations, where variable coupling makes it difficult to obtain offspring that readily satisfy implicit energetic constraints imposed by long-range interactions. Studies show that the use of short memes improves the ability of an MA to sample low-energy conformations (81). This in turn allows

reaching significantly lower-energy conformations in a shorter amount of time and has been demonstrated both in on-lattice and off-lattice MAs (74–79, 82, 83).

In (80), where the meme is a hill-climbing local search, the MA is shown both more efficient and more effective at finding near-native conformations over a standard EA. A recent study in (84), which extends the EA-based Harmony algorithm to use a hybrid local search shows similar improvements over the original algorithm. However, improvements are often reported in terms of energetic values reached, with lower values taken as indication of higher exploration capability (85, 86). While such a metric is important for comparing novel algorithmic ingredients to baseline EAs, in itself it is not indicative of the utility of sampled conformations for PSP. In the computational structural biology community, the focus is often on the ability of a conformation sampling algorithm to reach the known native structure; that is, the metric underpinning performance is RMSD or other distance-based metrics between sampled conformations and the known native structure. When judged by this metric, many on-lattice EAs fall short when compared to the state-of-the-art MC-based conformation sampling algorithms that have moved beyond lattice models. In addition, due to the popularity of a legacy benchmark dataset, the majority of on-lattice MAs are tested on proteins no longer than 61 amino acids. On longer chains, prediction quality suffers; for instance work in (84) reports the inability to find conformations below 6Å RMSD to known native structures on chains longer than 60 amino acids (84).

Highlight: Population-based, Off-lattice MAs for PSP MAs capable of reaching similar or better prediction quality than state-of-the-art MC-based conformation sampling algorithms incorporate key domain-specific insight on proteins. These include off-lattice, backbone-based angular representations, state-of-the-art energy functions, such as the suite of Rosetta energy functions, and the popular molecular fragment replacement technique in perturbation operators and memes (46, 48, 87, 88). Work in (82) introduces a fixed-size MA that makes use of such domain-specific insights. The greedy local search that constitutes the meme makes use of the fragment replacement technique; conformations are evaluated with to

the Rosetta score3 function, and elitism is employed to pitch the top offspring against the top parents. The survival mechanism is truncation-based. A representative result of the performance of this MA is provided in Fig. 2, which shows that this MA beats the (multistart) MC-based conformation sampling algorithm employed in the popular Rosetta structure prediction protocol in terms of exploration capability while achieving similar or lower RMSDs to the known native structure. Work in (82) additionally considers injecting crossover into this MA and studies the interplay between various crossover operators and the local search. A novel crossover operator is proposed that preserves local structural features and results in offspring with fewer constraint violations.

Highlight: Memes and Move Sets in EAs for PSP With the realization that the local search/improvement operator is key to obtaining optimal conformations, significant efforts are spent in designing customized operators for both on-lattice and off-lattice EAs for PSP. Due to the demonstrated superiority of the molecular fragment replacement technique in MC-based conformation sampling algorithms, related efforts in off-lattice EAs have pursued memes that are hill climbers, MC local search, Metropolis MC local search over fragment replacement moves (46–48, 82, 87–89).

Work in on-lattice EAs has also revealed a variety of effective memes for EAs on 2d square and triangular, and 3d cubic, triangular, and face-centered-cubic (FCC) lattices. The majority of memes for lattices employ the concept of move sets, such as diagonal moves and tilt moves (90), moves that preserve local, secondary structures (91), pull moves (92, 93), end moves, corner moves, three-bead and end flip for single-point moves and crankshaft for double point moves (94), rotation moves (23). In particular, recent work in (95) investigates in detail the geometric properties of the 3D FCC lattice and proposes several local search operators that build on lattice rotation and generalized move sets to achieve optimal conformations much faster than baseline EAs.

Highlight: Co-evolving Memes in GAs for PSP Another interesting direction in MAs is the co-evolving of memes. Early work in MAs for PSP pursued dynamically modifying memes. The idea is that a single static local search may not be effective for all protein sizes and topologies or stages of an EA. Work in (96) proposes an on-lattice GA with a Metropolis MC-based local search (an HP lattice model is employed). Temperature in the Metropolis criterion is varied in a reactive scheme so that the method balances between exploration and exploitation. When the population of conformations is deemed diverse, the temperature is lowered to focus on exploitation of local minima. As the population converges, the temperature is increased to shift the focus on exploration. The method is reported to find high-fitness conformations faster than a baseline EA. Extensions in (97) co-evolve memes alongside conformations (variables such as length of the local search and type of mutation are added to the variables employed to represent a conformation) (97). Co-evolving memes is shown to improve both time performance and fitness of computed conformations over baseline EAs (97–99).

4.4.2 Evolutionary Strategies to Avoid Premature Convergence

The issue of premature convergence or stalling due to loss of population diversity, long known in the EC community to accompany GAs, has also been observed in adaptations for PSP. The GA crossover and mutation operators can become ineffective over time, leading to growing similarity among individuals in a population (100, 101). Stalling is an expected phenomenon, as earlier generations are expected to cover a broader search space while later generations are expected to converge to specific regions in the fitness landscape. With growing similarity in a population, the crossover operator becomes implicitly controlled and fails to produce offspring that are significantly different from their parents. In effect, the crossover operator produces more twins.

Stalling is responsible for GAs getting frequently trapped in local minima (102). This is exacerbated for longer protein sequences and is credited as one of the major reasons why GAs, though effective, are not competitive with state-of-the-art MC-based conformation sampling algorithms for PSP (103).

Some of the earliest work addresses this phenonemon by using genotypic diversity for selection and replacement of individuals in a population (104). The original on-lattice GA proposed in (96) is extended in (104) so that parents selected for crossover have maximal genotypic difference, measured via the Manhattan distance metric. Experiments show that significantly lower energy values are obtained over the original GA (104).

In other studies, a twin removal approach is employed instead. Twins are energetically- and structurally-similar conformations and they are determined based on phenotypic distance measured via Hamming distance or distance over contact maps (105). There are several twin removal strategies. One strategy is to periodically remove and replace twins with new conformations sampled at random (77, 105, 106). Other strategies relax the definition of twins to include not only identical but also highly-correlated individuals (107). Work in (107) shows that such an approach significantly improves performance of a number of on-lattice, GA-based methods. Crowding, a strategy originally introduced in (108), can also be seen as a specific implementation of twin removal, though restricted to an offspring replacing an individual most similar to itself (109). Another strategy known as niched-penalty (110) does not explicitly remove similar individuals but imposes a penalty to discourage their participation as parents for producing offspring for the next generation. Though promising, the strategy has yet to be evaluated in EAs for PSP.

Another popular approach for increasing diversity in EAs for PSP is to avoid redundant conformations all together. This approach, known as tabu search, keeps track of conformations recently visited by the local search to avoid their revisitation by other local searches (111). Work in (112) employs a subset of already-sampled conformations to avoid revisitation at the global level in an MA. Comparison of these two distinct employments of tabu search to avoid revisitation shows that tabu search at the global level is more effective than at the local level on HP-lattice models (112). Customizations of tabu search for on-lattice EAs are regularly pursued (113,114). Integrations of tabu search in off-lattice MAs are also beginning to be pursued, though currently limited to hydrophobic-hydrophilic toy

models (115, 116).

4.4.3 Multi-Objective Optimization in EAs for PSP

Casting PSP as a multi-objective rather than a single-objective PSP problem is proving powerful and effective at avoiding premature convergence and overall improving the performance of EAs regarding their ability to reproduce known native structures.

Multi-objective optimization (MOO) lies in the ability to decouple and group terms in an energy function in a few categories considered as separate objectives. MOO originates from the Pareto analysis in economic theory on simultaneous optimization of conflicting objectives. Casting PSP as an MOO problem is particularly suitable, because terms in protein energy functions compete with one another. For instance, slight fluctuations in the backbone of a protein conformation may simultaneously lower the value of the energy terms measuring local interactions but increase the value of the terms measuring non-local interactions.

A simple way to cast PSP as an MOO is to do so in the survival mechanism through the concept of dominance. Suppose that the terms in a protein energy function are grouped into two categories, NB (non-bonded) and B (bonded). A conformation C_i is said to dominate C_j when the value of every category in C_i is strictly less than the value of the corresponding category in C_j. This is known as strong dominance (weak dominance allows equivalent values). If strong dominance is employed, the number of conformations that dominate a conformation C_i is known as the Pareto count of C_i. Non-dominated conformations (those with Pareto count 0) constitute the Pareto front of a set of conformations. A Pareto rank can also be associated with a conformation C_i by counting the number of conformations that C_i dominates (conformations in the Pareto front have Pareto rank 0 by definition).

Before highlighting several EAs that treat PSP as an MOO problem, it is worth noting that the concept of Pareto dominance has been shown useful also for *decoy selection* techniques that select a subset of computed conformations for further refinement and then decide among those which ones represent the unknown native

state in true blind prediction setting. The majority of current decoy selection techniques in PSP make use of RMSD-based clustering of conformations; typically, the cluster with the largest number of members is predicted as the native state.

MOO analysis provides an alternative route. Recent work has shown that the Pareto front or various thresholds of Pareto counts are effective at reducing the ensemble of sampled conformations while retaining near-native ones (47). The selection of the Pareto knees is also shown effective (117, 118). In (118), a knee-based selection technique retains conformations within 0.3Å of the actual best conformation in the ensemble (best in terms of RMSD to the native structure). Work in (119) provides contradictory results and shows that including knees makes little difference. Testing is conducted on short peptides up to 20 amino acids long, which probably do not benefit from MOO analysis.

Decoupling energy terms into separate objectives and employing MOO reduces the complexity of the energy landscape in short polypeptides by reducing the number of local minima (120). MOO has already been integrated in on-lattice EAs for conformation sampling (121, 122). MOO in off-lattice EAs decomposes terms of all-atom energy functions, such as CHARMM and AMBER (123, 124). Typically, terms of the energy function are grouped in two categories, with one category consisting of the Lennard-Jones term that measures non-bonded interactions, and the other category summing up all other terms. Doing so in the fast messy Genetic Algorithm (fmGA), which represents dihedral angles as 10-bit strings, is shown to result in lower-energy conformations over the baseline fmGA for short protein chains of 5-14 amino acids (123, 124). Other studies employ three rather tan two categories and show that doing so results in more conformations closer to the native structure. Testing is limited, however, to a short peptide and a medium-length protein (125–127).

Highlight: State-of-the-art 1+1 MO-MA for PSP Work in (128) integrates MOO in a 1+1 EA, using the same CHARMM bonded vs. non-bonded categories as work in (123, 124). Conformations in the Pareto front are recorded in an archive or hall of fame, and secondary structure prediction from a given amino-

acid sequence and side-chain rotamer libraries are used to bias sampling toward physically-relevant conformations. At each generation, the offspring and parent compete for survival. If neither dominates the other, the one which dominates more of the archive survives. Later work extends the I-PAES and shows it effective on several longer protein chains up to 70 amino acids (118). Conformations below 5Å lRMSD to the known native structure are found for sequences up to 70 amino acids in length, and results are shown to outperform several other MOO EAs and standard EAs on shorter chains, as well. Some representative results are shown in Fig. 3.

Highlight: State-of-the-art population-based MO-MAs and MO-MGAs for PSP Work in (129) integrates MOO in the population-based MAs and MGAs proposed and shown competitive in (82) to the Rosetta MC-based conformation sampling algorithm. Three categories are defined that group together terms of the Rosetta score4 function; the first category measures short range hydrogen bonding,; the second measures long-range hydrogen bonding; and the third term summing the rest of the terms. It is worth noting that the employment of backbone dihedral angles as variables preserves bond lengths and valence angles; thus, energetic differences between conformations are due to non-bonded interactions. A novel truncation selection mechanism is employed, which sorts parent and offspring at the end of each generation first by their Pareto rank and then by total energy (for conformations with the same Pareto rank). The injection of this MOO technique is compared against the baseline MA and MGA algorithms originally introduced in (82), resulting in the MO-MA and MO-MGA algorithms presented in (129). The addition of Pareto count in the truncation-based mechanism is also tested (conformations are first sorted by Pareto rank, then Pareto count, then total energy). The resulting extensions are referred to as MO-MA-PC and MO-MGA-PC (130). Fig. 4 shows that these four algorithms outperform the multistart MC-based conformation sampling algorithm in Rosetta and represent the state of the art in off-lattice EAs for PSP. Fig. 5 showcases the capability of these algorithms by rendering the lowest-RMSD (best) conformations obtained by these algorithms

on a variety of proteins and comparing these to the best conformations found by multistart MC-based conformation sampling algorithm in Rosetta.

4.5 EAs for Mapping Protein Energy Landscapes

EAs obtain a discrete representation of the potential energy surface of a protein chain. It is thus easy to see the promise of EAs for more than PSP. A central challenge in molecular biology is to understand functional changes upon single-point mutations in proteins. EAs hold significant promise for providing a detailed characterization of structure spaces and underlying energy landscapes, which currently challenge methods based on MC and those based on Molecular Dynamics. However, a truly *de novo* setting currently proves too challenging, given that the objective is to retain diversity and map multiple basins of a protein's energy landscape. In recent work, EA-based methods are proposed to map multi-basin energy landscapes of complex proteins over 100 amino acids long. These methods make use of known, experimentally-available long-lived structures of healthy (wildtype) and aberrant versions of a protein. These structures are leveraged to transform a discrete problem into a continuous one, subjecting them to Principal Component Analysis to reveal a few collective variables constituting the search space for the EA. In (109, 131), mutation operators are defined over the variables and a family-based crowding mechanism is used to retain diverse conformations longer. The evaluation operator lifts individuals from the reduced representation to an all-atom representation prior to subjecting them to a meme for improvement; the latter uses a simulated annealing MC local search currently implemented as the *relax* protocol in Rosetta.

An implementation of the above method is available in `http://www.cs.gmu.edu/~ashehu/?q=OurTools`. Applications on several proteins up to 165 amino acids long show that this MA is able to reveal multiple basins of proteins known switch between different structural states for function. In particular, work in (132) builds on this MA and introduces a method that is capable of mapping energy landscapes of wildtype and oncogenic sequences of the H-Ras catalytic domain and explaining via comparison of the landscapes the reasons for func-

tional changes. While these methods explicitly seed the initial population with experimentally-available structures, an adaptation of the popular CMA-ES technique is introduced in (133) employs these structures only to extract the reduced search space via PCA and initialize the multivariable distribution.

5 Future Prospects

As stochastic optimization now represents the only computationally-feasible approach to PSP, work on improving the capability of EAs for PSP is expected to continue. Work on on-lattice EAs is expected to advance PSP for very large protein chains of several hundred amino acids. On protein chains of up to 200 amino acids, the goal is to increase prediction accuracy, and in this domain, more returns are expected from off-lattice EAs that make use of state-of-the-art protein energy functions.

While this review highlighted several evolutionary techniques adapted from the EC community to address the exploration vs. exploitation issue in the multimodal protein energy landscapes, there are several opportunities to design more complex EAs. As the review has highlighted, there are several known mechanisms for population diversification that have yet to be adapted and tested for PSP. There are several opportunities to further investigate dynamic, co-evolving memes, particularly for more complex local searches. There is a growing interest in the EC community to dynamically make decisions on allocation of computational resources to computation-heavy memes (134). Other evolutionary strategies for structurization of EAs also present interesting new avenues to enhance exploration capability. Interestingly, structured EAs have been debuted in macromolecular modeling but have been limited to sequence-function prediction problems (135). Further investigation of multi-objective optimization and Pareto-based measures is expected to improve accuracy, particularly in the context of inherently approximate protein energy functions. Finally, given the importance of injecting domain-specific insight in EAs for PSP, efforts on designing novel, representation-specific perturbation operators are expected to improve performance.

As this review has highlighted, the potential of EAs beyond PSP for the more general and challenging problem of mapping protein energy landscapes is only beginning to be realized (109, 131–133, 136). Evolutionary strategies that hold off premature convergence are key to the ability of EAs to reproduce a multitude of possible basins in a complex landscape.

EC researchers tempted by the richness and complexity of scientific questions posed by protein structure modeling now have strong foundations to venture into this domain. Work in protein structure modeling is challenging, as it often requires researchers to attain working knowledge in a new discipline. For those willing to do so, however, the payoff is significant. It is worth considering that, while at the moment EAs are not the top methods for PSP and modeling of single protein chains, there is one domain where they have dominated and replaced MC-based algorithms. In protein-ligand binding and protein-protein docking, the top algorithms are complex EAs. The hope is that in a few years, one will be able to say the same for for PSP and protein structure modeling in general.

Acknowledgement

Funding for this work is provided in part by the National Science Foundation (Grant No. 1421001 and CAREER Award No. 1144106) and the Thomas F. and Kate Miller Jeffress Memorial Trust Award.

References

1. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;1(33):D514–D517.

2. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014;133(1):1–9.

3. Ratovitski T, Corson LB, Strain J, Wong P, Cleveland DW, Culotta VC,

et al. Variation in the biochemical/biophysical properties of mutant superoxide dismutase 1 enzymes and the rate of disease progression in familial amyotrophic lateral sclerosis kindreds. Human Molecular Genetics. 1999;8(8):1451–1460.

4. DiDonato M, Craig L, Huff ME, Thayer MM, Cardoso RM, Kassmann CJ, et al. ALS mutants of human superoxide dismutase form fibrous aggregates via framework destabilization. J Mol Biol. 2003;332(1):601–615.

5. Soto C. Unfolding the role of protein misfolding in neurodegenerative diseases. Nat Rev Neurosci. 2003;4(1):49–60.

6. Soto C. Protein misfolding and neurodegeneration. JAMA Neurology. 2008;65(2):184–189.

7. Uversky VN. Intrinsic disorder in proteins associated with neurodegenerative diseases. Front Biosci. 2009;14:5188–5238.

8. Neudecker P, Robustelli P, Cavalli A, Walsh P, Lundstrm P, Zarrine-Afsar A, et al. Structure of an intermediate state in protein folding and aggregation. Science. 2012;336(6079):362–366.

9. Fetics SK, Guterres H, Kearney BM, Buhrman G, Ma B, Nussinov R, et al. Allosteric Effects of the Oncogenic RasQ61L Mutant on Raf-RBD. Structure. 2015;23(3):505–516.

10. Berman HM, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol. 2003;10(12):980–980.

11. Reardon S. Large NIH projects cut. Nature. 2013;503(7475):173–174.

12. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(4096):223–230.

13. Boehr DD, Wright PE. How do proteins interact? Science. 2008;320(5882):1429–1430.

14. Dill KA, Ozkan B, Shell MS, Weikl TR. The Protein Folding Problem. Annu Rev Biophys. 2008;37:289–316.

15. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nature Chem Biol. 2009;5(11):789–96.

16. Zhang Y. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins. 2014;82(Suppl 2):175–187.

17. Amaro RE, Bansai M. Editorial overview: Theory and simulation: Tools for solving the insolvable. Curr Opinion Struct Biol. 2014;25:4–5.

18. Clementi C. Coarse-grained models of protein folding: Toy-models or predictive tools? Curr Opinion Struct Biol. 2008;18:10–15.

19. Taketomi H, Ueda Y, , Go N. Studies on protein folding, unfolding and

fluctuations by computer simulation: The effect of specific amino acid sequence represented by specific interunit interactions. Int J Peptide Prot Res. 1975;7(6):445–459.

20. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol. 1994;243(4):668–682.

21. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins: Struct Funct Genet. 1994;18(4):338–352.

22. Ishikawa K, Yue K, Dill KA. Predicting the structures of 18 peptides using Geocore. Protein Sci. 1999;8(4):716–721.

23. Unger R, Moult J. Finding lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. Bull Math Biol. 1993;55(6):1183–1198.

24. Hart WE, Istrail S. Robust Proofs of NP-Hardness for Protein Folding: General Lattices and Energy Potentials. J Comp Biol. 1997;4(1):1–22.

25. Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M. On the complexity of protein folding. J Comput Biol. 1998;5(3):423–465.

26. Reva BA, Finkelstein AV, Sanner MF, Olson AJ. Adjusting potential energy functions for lattice models of chain molecules. Proteins: Struct Funct Genet. 1996;25(3):379–388.

27. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol. 1995;249(2):493–507.

28. Dotu I, Cebrian M, Van Hentenryck P, Clote P. On lattice protein structure prediction revisited. IEEE Trans Comput Biol Bioinform. 2011;8(6):1620–1632.

29. Abayagan R, Totrov M, Kuznetsov D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem. 1994;15(5):488–506.

30. Zhang M, Kavraki LE. A New Method for Fast and Accurate Derivation of Molecular Conformations. Chem Inf Comput Sci. 2002;42(1):64–70.

31. McLachlan AD. A mathematical procedure for superimposing atomic coordinates of proteins. Acta Crystallogr A. 1972;26(6):656–657.

32. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem. 1983;4(2):187–217.

33. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annual Review of Physical Chemistry. 1997;48:545–600.

34. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nat Struct Biol. 1997;4(1):10–19.

35. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opinion Struct Biol. 1997;14:70–75.

36. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci USA. 1987;84(19):6611–6615.

37. Verma A, Schug A, Lee KH, Wenzel W. Basin hopping simulations for all-atom protein folding. J Chem Phys. 2006;124(4):044515.

38. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How fast-folding proteins fold. Science. 2011;334(6055):517–520.

39. Vendruscolo M, Dobson CM. Protein dynamics: Moore's law in molecular biology. Curr Biol. 2011;21(2):R68–R70.

40. Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. Proceedings of the National Academy of Sciences. 2013;110(15):5915–5920.

41. Stein EG, Rice LM, Bruenger AT. Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. J Magn Reson. 1997;124(1):154–164.

42. Rice LM, Bruenger AT. 277-290. Proteins: Struct Funct Bioinf. 2004;19(4):277–290.

43. Chen J, Im W, Brooks C. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. J Comput Chem. 2005;26(15):1565–1578.

44. Unger R. The Genetic Algorithm Approach to Protein Structure Prediction. Structure and Bonding. 2004;110:153–175.

45. De Jong KA. Evolutionary Computation : A Unified Approach. Boston, MA: MIT Press; 2006.

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

47. Olson B, Shehu A. Rapid Sampling of Local Minima in Protein Energy Surface and Effective Reduction through a Multi-objective Filter. Proteome Sci. 2013;11(Suppl1):S12.

48. Saleh S, Olson B, Shehu A. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. BMC Struct Biol. 2013;13(Suppl1):S4.

49. Prentiss MC, Wales DJ, Wolynes PG. Protein structure prediction using basin-hopping. J Chem Phys. 2008;128(22):225106–225106.

50. Wales DJ, Doye JPK. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. J Phys Chem A. 1997;101(28):5111–5116.

51. Nayeem A, Vila J, Scheraga HA. A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. J Comput Chem. 1991;12(5):594–605.

52. Lourenco HR, Martin OC, Stutzle T. Glover F, Kochenberger G, editors. Iterated Local Search. Kluwer Academic Publishers; 2002.

53. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol. 1994;235(3):983–1002.

54. Mortenson PN, Evans DA, Wales DJ. Energy landscapes of model polyalanines. J Chem Phys. 2002;117(3):1363–1376.

55. Iwamatsu M, Okabe Y. Basin hopping with occasional jumping. Chem Phys Lett. 2004;399:396–400.

56. Olson B, Hashmi I, Molloy K, Shehu A. Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules. Advances in AI J. 2012;2012(674832).

57. Bradley P, Misura KMS, Baker D. Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science. 2005;309(5742):1868–1871.

58. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004;383:66–93.

59. Brunette TJ, Brock O. Guiding conformation space search with an all-atom energy potential. Proteins: Struct Funct Bioinf. 2009;73(4):958–972.

60. DeBartolo J, Colubri A, Jha AK, Fitzgerald JE, Freed KF, Sosnick TR. Mimicking the folding pathway to improve homology-free protein structure prediction. Proc Natl Acad Sci USA. 2009;106(10):3734–3739.

61. Shehu A, Olson B. Guiding the Search for Native-like Protein Conformations with an Ab-initio Tree-based Exploration. Int J Robot Res. 2010;29(8):1106–1127.

62. Simoncini D, Berenger F, Shrestha R, Zhang KYJ. A Probabilistic Fragment-Based Protein Structure Prediction Algorithm. PLoS ONE. 2012;7(7):e38799.

63. Handl J, Knowles J, Vernon R, Baker D, Lovell SC. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. Proteins: Struct Funct Bioinf. 2011;80(2):490–504.

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

65. Shehu A, Kavraki LE, Clementi C. Multiscale Characterization of Protein Conformational Ensembles. Proteins:

Struct Funct Bioinf. 2009;76(4):837–851.

66. Molloy K, Shehu A. Elucidating the Ensemble of Functionally-relevant Transitions in Protein Systems with a Robotics-inspired Method. BMC Struct Biol. 2013;13(Suppl 1):S8.

67. Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci USA. 1996;93(12):5814–5818.

68. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–574.

69. Hoque M, Chetty M, Sattar A. Genetic Algorithm in Ab Initio Protein Structure Prediction Using Low Resolution Model: A Review. Biomedical Data and Applications. 2009;p. 317–342.

70. Hart WE, Krasnogor N, Smith JE, editors. Recent advances in memetic algorithms. Vol 166 of Studies in Fuzziness and Soft Computing. Heidelberg, Germany: Springer; 2004.

71. Ong YS, Keane AJ. Meta-Lamarckian learning in memetic algorithms. IEEE Trans on Evol Comp. 2004;8(2):99–110.

72. Ong YS, Krasnogor N, Ishibuchi H. Special issue on memetic algorithms. IEEE Trans on Systems, Man, and Cybernetics, Part B. 2004;37(1):2–5.

73. Ong Y, Lim M, Neri F, Ishibuchi H. Special issue on emerging trends in a soft computing: Memetic Algorithms. Soft Computing - A Fusion of Foundations, Methodologies, and Applications. 2004;p. 739–740.

74. Lopes HS, Scapin MP. An enhanced genetic algorithm for protein structure prediction using the 2d hydrophobic-polar model. In: Intl Conf on Artificial Evolution. Springer-Verlag; 2005. p. 238–246.

75. Berenboym I, Avigal M. Genetic Algorithms with Local Search Optimization for Protein Structure Prediction Problem. In: Intl Conf Genet Evol Comput (GECCO). ACM; 2008. p. 1097–1098.

76. Islam M. Novel Memetic Algorithm for Protein Structure Prediction. AI 2009: Advances in Artificial Intelligence. 2009;.

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

78. Tsay J, Su S. Ab initio protein structure prediction based on memetic algorithm and 3D FCC lattice model. In: Intl Conf. on Bioinformatics and Biomedicine (BIBM). IEEE; 2011. p. 315–318.

79. Su S, Lin C, Ting C. An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction. Proteome Sci. 2011;9(Suppl 1):S19–S19.

80. Cooper L, Corne D, Crabbe M. Use of a novel Hill-climbing genetic algorithm in protein folding simulations. Computational Biology and Chemistry. 2003;27(6):575–580.

81. Cotta C. Protein structure prediction using evolutionary algorithms hybridized with backtracking. Artificial Neural Nets Problem Solving Methods. 2003;p. 1044–1044.

82. Olson B, Jong KAD, Shehu A. Off-Lattice Protein Structure Prediction with Homologous Crossover. In: Intl Conf Genet Evol Comput (GECCO). New York, NY: ACM; 2013. p. 287–294.

83. Olson B. Evolving Local Minima in the Protein Energy Surface. George Mason University; 2013.

84. Abual-Rub MS, Al-Betar MA, Abdullah R, Khader AT. A hybrid harmony search algorithm for ab initio protein tertiary structure prediction. Network Modeling and Analysis in Health Informatics and Bioinformatics. 2012;p. 1–17.

85. Tantar AA, Melab N, Talbi E. A grid-based genetic algorithm combined with an adaptive simulated annealing for protein structure prediction. Soft Computing. 2008;12(12):1185–1198.

86. Goldstein M, Fredj E, Gerber R, Benny RB. A new hybrid algorithm for finding the lowest minima of potential surfaces: approach and application to peptides. J Comput Chem. 2011;32(9):1785–1800.

87. Olson B, Shehu A. Populating Local Minima in the Protein Conformational Space. In: IEEE Intl Conf on Bioinf and Biomed. Atlanta, GA; 2011. p. 114–117.

88. Saleh S, Olson B, Shehu A. A Population-based Evolutionary Algorithm for Sampling Minima in the Protein Energy Surface. In: IEEE Intl Conf on Bioinf and Biomed Workshops (BIBMW). Philadelphia, PA; 2012. p. 64–71.

89. Olson B, Shehu A. Efficient Basin Hopping in the Protein Energy Surface. In: IEEE Intl Conf on Bioinf and Biomed. Philadelphia, PA; 2012. p. 119–124.

90. Hoque T, Chetty M, Dooley LS. A Guided Genetic Algorithm for Protein Folding Prediction Using 3D Hydrophobic-Hydrophilic Model. Evolutionary Computation, 2006 CEC 2006 IEEE Congress on. 2006;p. 2339–2346.

91. Huang C, Yang X, He Z. Protein folding simulations of 2D HP model by the genetic algorithm based on optimal secondary structures. Comput Biol Chem. 2010;34(3):137–142.

92. Böckenhauer HJ, Dayem UA, Kapsokalivas L, Steinhöfel K. A local move set for protein folding in triangular lattice models. LNCS: Algorithms in Bioinformatics. 2008;11:369–381.

93. Lesh N, Mitzenmacher M, Whitesides S. A complete and effective move set for simplified protein folding. In: Seventh Annual Intl Conf on Res in Comp Mol Biol (RECOMB). ACM; 2003. p. 188–195.

94. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, et al. Principles of protein folding–a perspective from simple exact models. Protein Sci. 1995;4(4):561–602.

95. Tsay J, Su S. An effective evolutionary algorithm for protein folding on 3D FCC HP model by lattice rotation and generalized move sets. Proteome Sci. 2013;11(Suppl 1):S19.

96. Krasnogor N, Smith J. A memetic algorithm with self-adaptive local search: TSP as a case study. In: Intl Conf Genet Evol Comput (GECCO); 2000. p. 987–994.

97. Krasnogor N, Blackburne B, Burke E, Hirst J. Multimeme algorithms for protein structure prediction. In: Parallel Problem Solving from Nature (PPSN) VII. Lecture Notes in Computer Science. Springer-Verlag; 2002. p. 769–778.

98. Smith JE. Protein structure prediction with co-evolving memetic algorithms. In: Congress on Evolutionary Computation (CEC). vol. 4. IEEE; 2003. p. 2346–2353.

99. Smith JE. The co-evolution of memetic algorithms for protein structure prediction. Recent Advances in Memetic Algorithms. 2005;p. 105–128.

100. Fogel DB. Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. 3rd ed. Wiley-IEEE Press; 2005.

101. Deb K, Goldberg DE. An Investigation of Niche and Species Formation in Genetic Function Optimization. In: Intl Conf Genet Algorithms. ACM; 1989. p. 42–50.

102. Deb K, Goldberg DE. Simple Subpopulation Schemes. In: Evol Prog Conf. ACM; 1994. p. 296–397.

103. Corne DW, Fogel GB. An Introduction to Bioinformatics for Computer Scientists. In: Fogel GB, Corne DW, editors. Evolutionary Computation in Bioinformatics. Elsevier, India; 2004. p. 3–18.

104. Bazzoli A, Tettamanzi A. A Memetic algorithm for protein structure prediction in a 3D-lattice HP model. Applications of Evolutionary Computing. 2004;3005:1–10.

105. Chira C. A hybrid evolutionary approach to protein structure prediction with lattice models. In: Congress on Evolutionary Computation. IEEE; 2011. p. 2300–2306.

106. Chira C, Horvath D, Dumitrescu D. Hill-Climbing search and diversification within an evolutionary approach to protein structure prediction. BioData Mining. 2011;4(1):23.

107. Hoque MT, Chetty M, Lewis A, Sattar A. Twin removal in genetic algorithms for protein structure prediction using low-resolution model. IEEE/ACM Trans Comput Biol Bioinf. 2011;8(1):234–245.

108. De Jong KA. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. University of Michigan, Ann Arbor, MI; 1975.

109. Clausen R, Shehu A. A Multiscale Hybrid Evolutionary Algorithm to Obtain Sample-based Representations of Multi-basin Protein Energy Landscapes. In: ACM Conf on Bioinf and Comp Biol (BCB). Newport Beach, CA; 2014. p. 269–278.

110. Deb K, Agrawal S. Niched-Penalty Approach for Constraint Handling in Genetic Algorithms. In: Artificial Neural Nets and Genetic Algorithms. Springer-Verlag; 1999. p. 235–243.

111. Dotu II, Cebrián MM, Van Hentenryck PP, Clote PP. On lattice protein structure prediction revisited. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2011 Nov;8(6):1620–1632.

112. Swakkhar S, A HM, Pham DN, Sattar A. Memory-Based Local Search for Simplified Protein Structure Prediction. In: ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB); 2012. p. 1–8.

113. Liu J, Sun Y, Li G, Song B, Huang W. Heuristic-based tabu search algorithm for folding two-dimensional AB off-lattice model proteins. Comput Biol Chem. 2013;47:142–148.

114. Zhou C, Hou C, Zhang Q, Wei X. Enhanced hybrid search algorithm for protein structure prediction using the 3D-HP lattice model. J Mol Model. 2013;19(9):3883–3891.

115. Zhang X, Wang T, Luo H, Yang JY, Deng Y, Tang J, et al. 3D Protein structure prediction with genetic tabu search algorithm. BMC Systems Biol. 2010;4(Suppl 1):S6.

116. Zhou C, Hou C, Wei X, Zhang Q. Improved hybrid optimization algorithm for 3D protein structure prediction. J Mol Model. 2014;20(7):2289–2300.

117. Becerra D, Sandoval A, Restrepo-Montoya D, Nino LF. A parallel multiobjective ab initio approach for protein structure prediction. In: Intl Conf on Bioinformatics and Biomedicine (BIBM). IEEE; 2010. p. 137–141.

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

119. Narzisi G, Nicosia G, Stracquadanio G. Robust Bio-active Peptide Prediction Using Multi-objective Optimization. In: Biosciences (BIOSCIENCESWORLD), 2010 International Conference on; 2010. p. 44–50.

120. Handl J, Lovell S, Knowles J. Investigations into the effect of multiobjectivization in protein structure prediction. Parallel Problem Solving from Nature–PPSN X. 2008;p. 702–711.

121. Garza-Fabre M, Rodriguez-Tello E, Toscano-Pulido G. Multiobjectivizing the HP model for protein structure prediction. Evolutionary Computation in Combinatorial Optimization. 2012;p. 182–193.

122. Garza-Fabre M, Toscano-Pulido G, Rodriguez-Tello E. Locality-based multiobjectivization for the HP model of protein structure prediction. In: Intl Conf Genet Evol Comput (GECCO). ACM; 2012. p. 473–480.

123. Day RO, Zydallis JB, Lamont GB, Pachter R. Solving the protein structure prediction problem through a multiobjective genetic algorithm. Nanotechnology. 2002;2:32–35.

124. Day RO. A Multiobjective Approach Applied to the Protein Structure Prediction Problem. Storming Media; 2002.

125. Calvo JC, Ortega J. Parallel protein structure prediction by multiobjective optimization. Euromicro Intl Conf on Parallel, Distributed and Network-based Processing. 2009;p. 268–275.

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

127. Calvo JC, Ortega J, Anguita M. Comparison of parallel multi-objective approaches to protein structure prediction. Supercomputing. 2011;p. 253–260.

128. Cutello V, Narzisi G, Nicosia G. A class of pareto archived evolution strategy algorithms using immune inspired operators for ab-initio protein structure prediction. In: Applications of Evolutionary Computing; 2005. p. 54–63.

129. Olson B, Shehu A. Multi-Objective Stochastic Search for Sampling Local Minima in the Protein Energy Surface. In: ACM Conf on Bioinf and Comp Biol (BCB). Washington, D. C.; 2013. p. 430–439.

130. Olson B, Shehu A. Multi-Objective Optimization Techniques for Conformational Sampling in Template-Free Protein Structure Prediction. In: Intl Conf on Bioinf and Comp Biol (BICoB). Las Vegas, NV; 2014. .

131. Clausen R, Shehu A. A Data-driven Evolutionary Algorithm for Mapping Multi-basin Protein Energy Landscapes. J Comput Biol. 2015;In press.

132. Clausen R, Ma B, Nussinov R, Shehu A. Mapping the Conformation Space

of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. PLoS Comput Biol. 2015;In press.

133. Clausen R, Sapin E, A DK, Shehu A. Evolution Strategies for Exploring Protein Energy Landscapes. In: Intl Conf Genet Evol Comput (GECCO). ACM; 2015. .

134. Ong YS, Lim M, Wong K. Classification of adaptive memetic algorithms: a comparative study. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics. 2006;36(1):2–5.

135. Kamath U, Kaers J, , Shehu A, De Jong KA. A Spatial EA Framework for Parallelizing Machine Learning Methods. In: Coello C, Cutello V, Deb K, Forrest S, Nicosia G, Pavone M, editors. Parallel Problem Solving from Nature - PPSN XII. vol. 7491 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg; 2012. p. 206–215.

136. Sapin E, Clausen R, A DK, Shehu A. Mapping Multiple Minima in Protein Energy Landscapes with Evolutionary Algorithms. In: Intl Conf Genet Evol Comput (GECCO) Workshop. ACM; 2015. .

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

Fig. 1

Illustration of Protein Geometry and Energetics

Top panel: The 3d structure shown on the left is a wet-lab snapshot of the biologically-active state of the ubiquitin protein. The Visual Molecular Dynamics (VMD) (137) is used for rendering. The backbone is drawn in opaque, with the local secondary structures drawn in different colors and side chains in transparent to easily visualize the backbone. A small fragment from amino acid at position 47 to position 52 in the 76-aa chain of ubiquitin is highlighted in greater detail on the right. The chain is drawn in the ball-stick representation with VMD. Backbone atoms in each amino acid are annotated. The side chain atoms are drawn in silver. The ϕ, ψ dihedral angles are shown, as well. Bottom panel: A model energy surface with a single deepest basin is illustrated here, adapted from (34). The surface is nonlinear and multimodal. The deepest basin is populated by conformations of the biologically-active state, illustrated here by superimposing over one another conformations of the ubiquitin NMR ensemble deposited in the PDB

under id 1d3z.

Fig. 2

Illustration of Performance of State-of-the-art MAs for PSP

Top panel: Given the same computational budget, the lowest energy value (measured with the score4 energy function in Rosetta) reached by the on-lattice MA in (129) and the multistart MC-based conformation sampling algorithm in Rosetta are measured. The y axis shows the difference. Bars below the 0 line indicate where MA reaches lower energy regions in the variable space. MA does so for 75% of the 20 different protein sequences used for the comparison. The PDB ids of the native structures of these sequences are shown on the x axis. Results combine many independent runs of each algorithm under comparison. Bottom panel: The y axis shows the difference in the lowest RMSD reached by many runs of each algorithm to the known native structure. The difference shows that MA is competitive with the conformation sampling algorithm in Rosetta even on the unforgiving RMSD metric.

Fig. 3

Illustration of Best Models Obtained by a State-of-the-art MO-MA

Four proteins of varying sizes from 34 to 70 amino acids are chosen to illustrate the high quality of the lowest-RMSD conformations obtained by the MO-GA algorithm presented in (118). The left panel shows the native structure and its PDB ID, whereas the right panel shows the computed lowest-RMSD conformation for each protein and its CA RMSD (calculated over CA atoms) from the native structure. Rendering is done with Pymol, showing secondary structures of the backbone and drawing side chains with thin lines. Figures are kindly provided by Giuseppe Nicosia and Giuseppe Narzisi.

Fig. 4

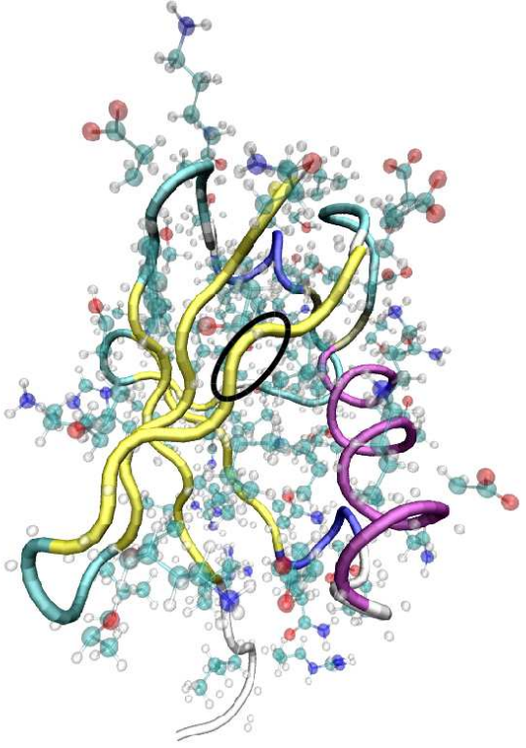
Illustration of Performance of State-of-the-art MO-M(G)As for PSP

The performance of MO-MA and MO-MGA presented in (129) and MO-MA-PC and MO-MGA-PC presented in (130) is shown here, compared to the MC-based conformation sampling in Rosetta, on 20 proteins. The PDB ids of the native structures of these sequences are shown on the x axis. The top panel shows the lowest energy reached by each algorithm. The bottom panel shows the lowest RMSD to the native structure reached by each algorithm. Results combine many independent runs of each algorithm under comparison.

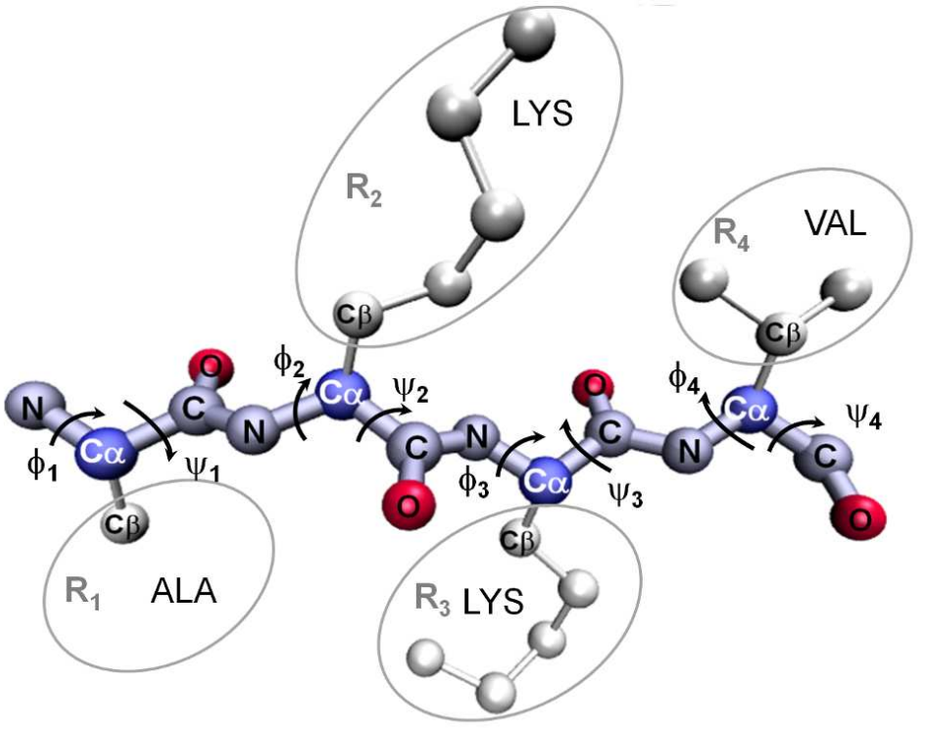
Fig. 5

Illustration of Best Models Obtained by a State-of-the-art MO-GA

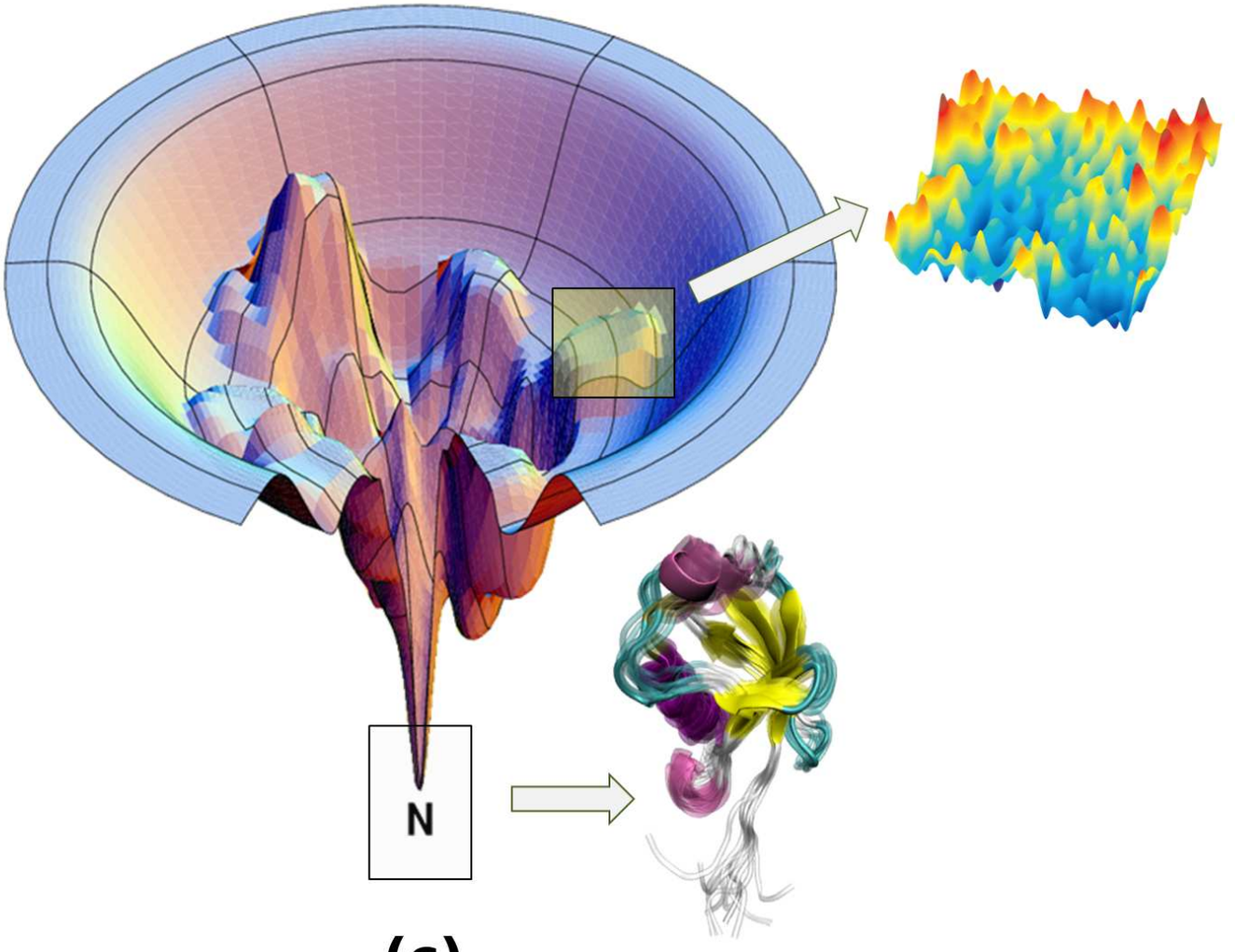
Three proteins of varying sizes from 70 to 106 amino acids are chosen to illustrate the high quality of the lowest-RMSD conformations obtained by the MO-GA algorithm presented in (129). The left panel superimposes the best conformation produced by the MC-based conformation sampling algorithm in Rosetta (drawn in lemon green) over the known native structure (drawn in gray). The right panel superimposes the best conformation produced by MO-GA (drawn in orange) over the native structure. The PDB IDs of each native structure are shown. The RMSD of each lowest-RMSD conformation to the known native structure is shown for each algorithm on each of the three selected proteins. The reported RMSD is computed over backbone atoms. Rendering is performed with VMD (137), using the NewCartoon graphical representation that shows the local secondary structures.



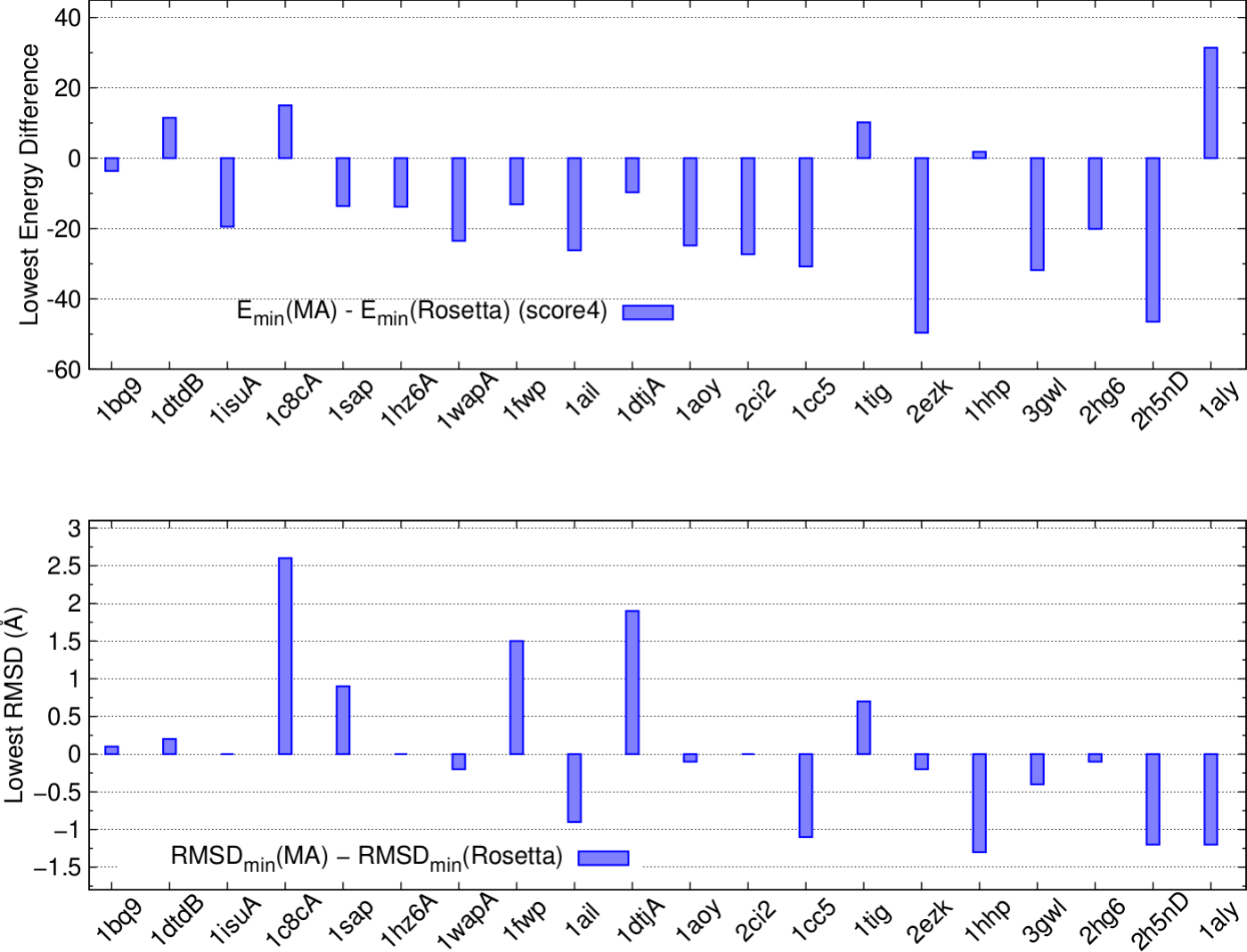
(a)

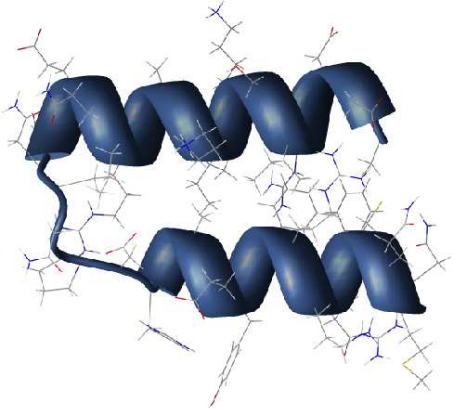

LYS
R2
VAL
R4
Cβ
Cβ
O
φ2
Cα
ψ2
N
C
N
C
O
φ4
Cα
ψ4
N
C
O
φ1
Cα
ψ1
φ3
Cα
ψ3
Cβ
Cβ
R1 ALA
R3 LYS

(b)

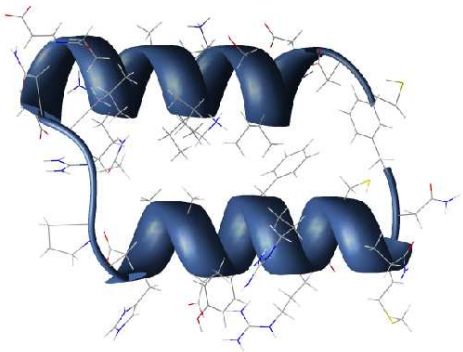

N

(c)

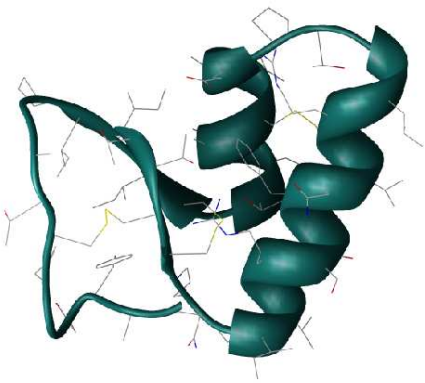

Lowest Energy Difference
40
20
0
-20
-40
-60
E_min(MA) - E_min(Rosetta) (score4)
1bq9
1dtdB
1isuA
1c8cA
1sap
1hz6A
1wapA
1fwp
1ail
1dtjA
1aoy
2ci2
1cc5
1tig
2ezk
1hhp
3gwl
2hg6
2h5nD
1aly
Lowest RMSD (Å)
3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
RMSD_min(MA) – RMSD_min(Rosetta)
1bq9
1dtdB
1isuA
1c8cA
1sap
1hz6A
1wapA
1fwp
1ail
1dtjA
1aoy
2ci2
1cc5
1tig
2ezk
1hhp
3gwl
2hg6
2h5nD
1aly



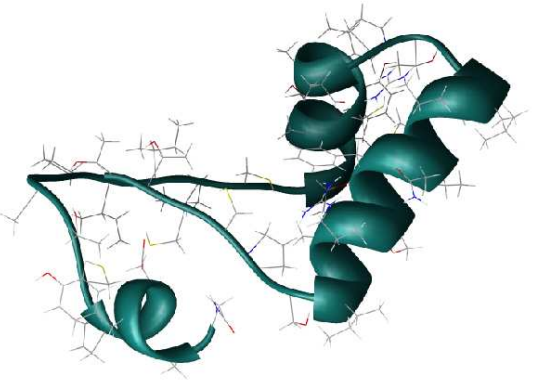
PDB ID 1zdd



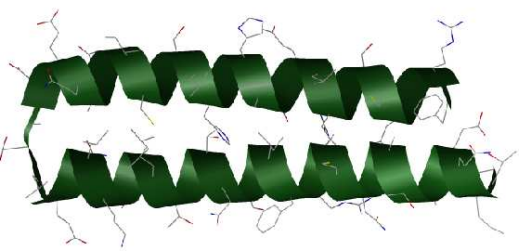
RMSD: 2.27



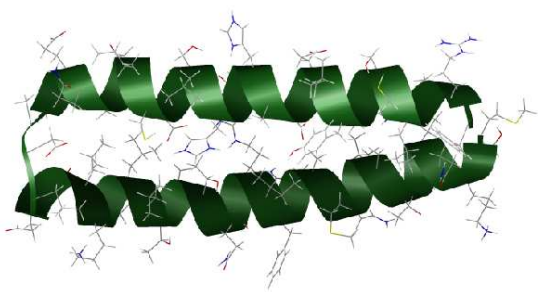
PDB ID 1crn



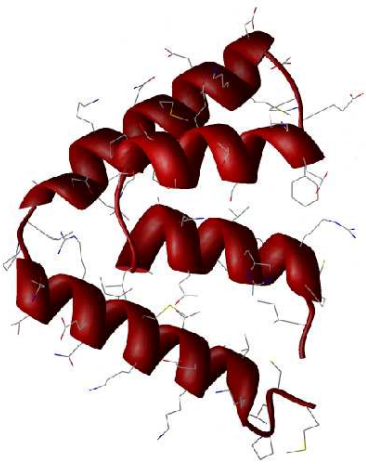
RMSD: 4.43



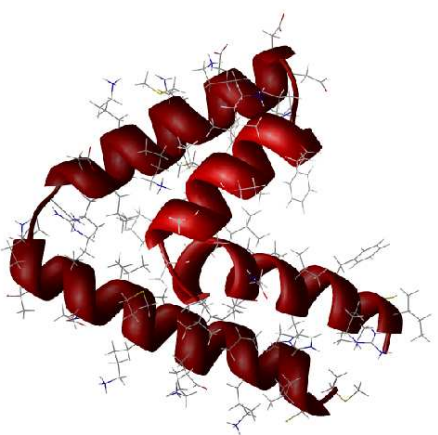
PDB ID 1rop



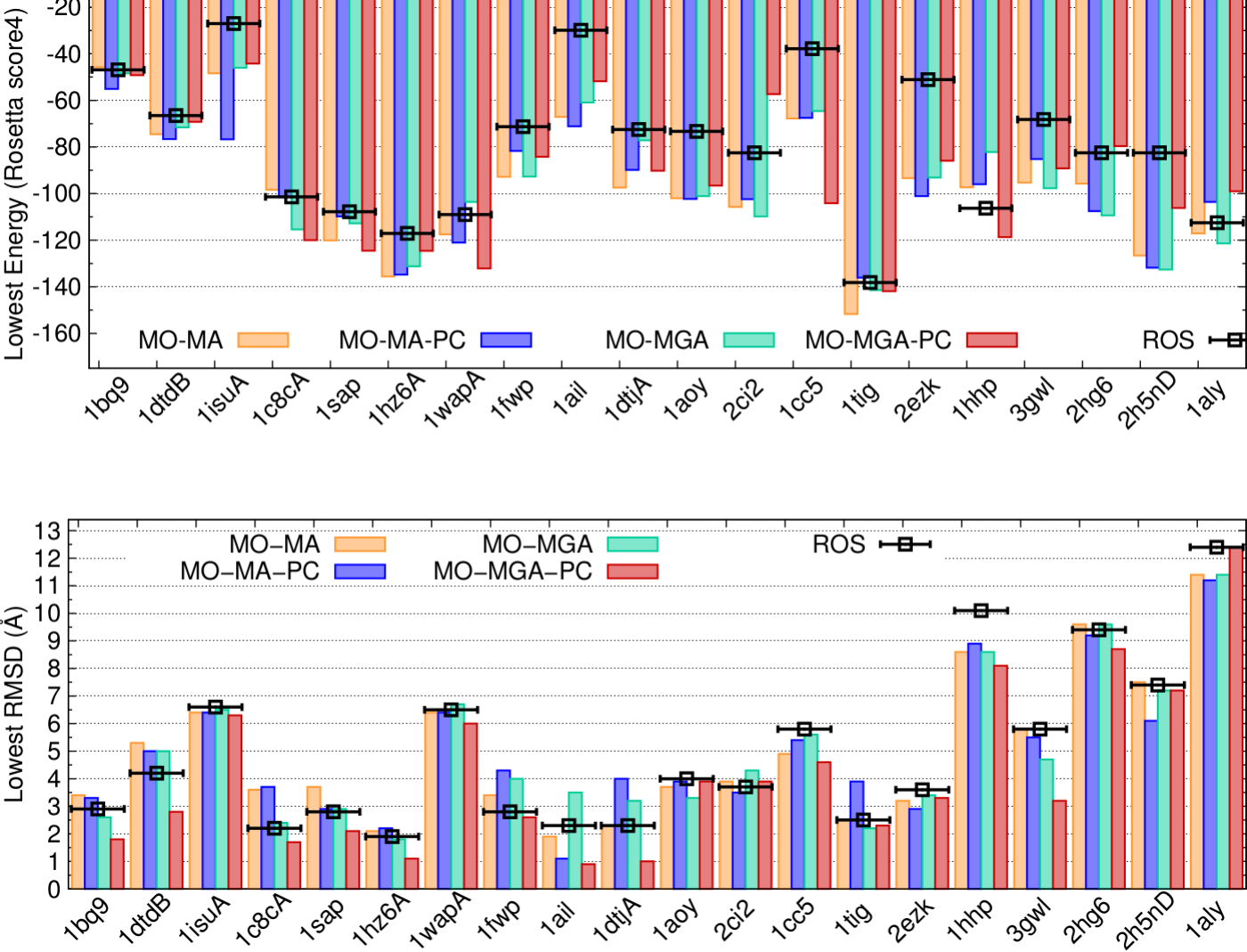
RMSD: 3.70



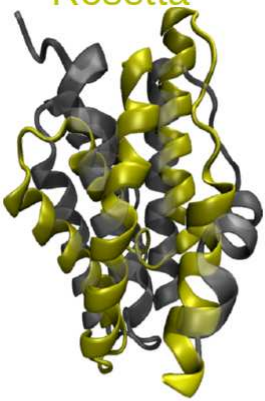
PDB ID 1utg



RMSD: 4.60


Lowest Energy (Rosetta score4)
-20
-40
-60
-80
-100
-120
-140
-160
MO-MA
MO-MA-PC
MO-MGA
MO-MGA-PC
ROS
1bq9
1dtdB
1isuA
1c8cA
1sap
1hz6A
1wapA
1fwp
1ail
1dtjA
1aoy
2ci2
1cc5
1tig
2ezk
1hhp
3gwl
2hg6
2h5nD
1aly
Lowest RMSD (Å)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
MO−MA
MO−MA−PC
MO−MGA
MO−MGA−PC
ROS
1bq9
1dtdB
1isuA
1c8cA
1sap
1hz6A
1wapA
1fwp
1ail
1dtjA
1aoy
2ci2
1cc5
1tig
2ezk
1hhp
3gwl
2hg6
2h5nD
1aly

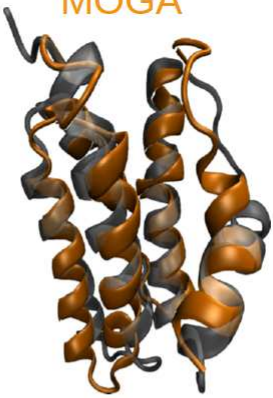
Rosetta



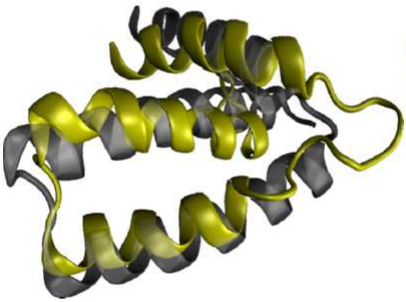
5.8 Å RMSD to native

PDB ID
3gwl

MOGA

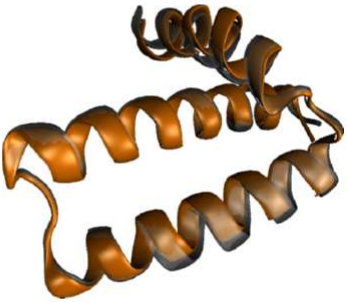


3.2 Å RMSD to native

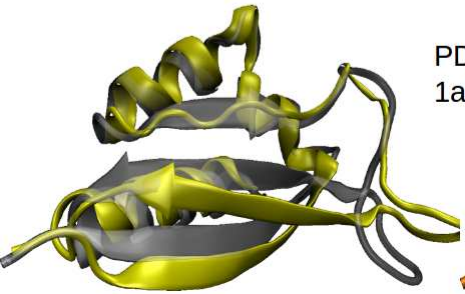


4.5 Å RMSD to native

PDB ID
1dtjA

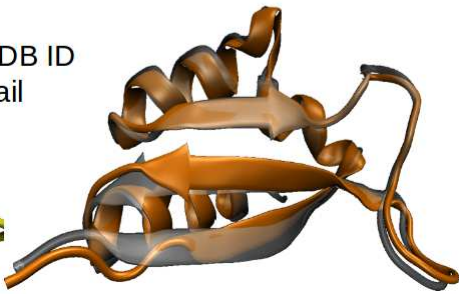


0.9 Å RMSD to native



2.3 Å RMSD to native

PDB ID
1ail



1.0 Å RMSD to native