Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab Initio Protein Structure Prediction

Kevin Molloy, Sameh Saleh, and Amarda Shehu

Abstract—Adequate sampling of the conformational space is a central challenge in ab initio protein structure prediction. In the absence of a template structure, a conformational search procedure guided by an energy function explores the conformational space, gathering an ensemble of low-energy decoy conformations. If the sampling is inadequate, the native structure may be missed altogether. Even if reproduced, a subsequent stage that selects a subset of decoys for further structural detail and energetic refinement may discard near-native decoys if they are high energy or insufficiently represented in the ensemble. Sampling should produce a decoy ensemble that facilitates the subsequent selection of near-native decoys. In this paper, we investigate a robotics-inspired framework that allows directly measuring the role of energy in guiding sampling. Testing demonstrates that a soft energy bias steers sampling toward a diverse decoy ensemble less prone to exploiting energetic artifacts and thus more likely to facilitate retainment of near-native conformations by selection techniques. We employ two different energy functions, the associative memory Hamiltonian with water and Rosetta. Results show that enhanced sampling provides a rigorous testing of energy functions and exposes different deficiencies in them, thus promising to guide development of more accurate representations and energy functions.

Index Terms—Protein structure prediction, probabilistic conformational search, near-native conformations, energy bias

1 Introduction

Ab initio protein structure prediction or template-free modeling is by now recognized as one of the most difficult problems in computational structural biology [23], [49]. Ab initio protocols are in principle more broadly applicable for protein structure prediction than homology-based methods that rely on the availability of homologs of known structure to reconstruct a template structure. In the absence of a template structure, ab initio protocols rely on a conformational search procedure guided by some energy function to obtain an ensemble of low-energy (decoy) conformations. In a thermodynamics treatment [1], [27], the sought native structure of the target sequence is expected to be present among the lowest energy decoys. There are many reasons why this treatment often fails to lead to the native structure [23].

Despite the challenges, over the last years considerable advancements have been made, most notably recently by the well-known Quark [46] and Rosetta protocols [20]. Key to these advancements has been the employment of the molecular fragment replacement technique [5], [7], [11]. The technique allows obtaining realistic decoy conformations by essentially assembling them with short structural blocks known as fragment configurations. These are extracted from known native structures of nonredundant sequences and are stored as ϕ, ψ, and ω backbone torsion angles in a library of fragment configurations [14]. The assembly is implemented in the context of a Metropolis Monte Carlo (MMC) trajectory, where an MC move replaces the configuration of a fragment selected at random over the currently assembled decoy conformation with a configuration selected for that fragment from the library. The replacement is accepted if it satisfies the metropolis criterion, resulting over time in low-energy decoys.

The predominant ab initio protocol consists of generation of a large number of decoy conformations at low resolution through the molecular fragment replacement technique followed by energetic refinement of selected decoys at higher resolution [23]. Typically, decoys are end points of MMC trajectories initiated from extended or random conformations. Many global and local search strategies are proposed to enhance the sampling capability of the core MMC exploration in the low-resolution stage. Typical protocols employ simulated annealing or replica exchange to enhance sampling [7], [11], [4], [38], [35], [39], [24]. Work in [36] employs a different strategy that directly enforces conformational diversity through discretization layers of the search space.

The molecular fragment replacement technique is directly related to the ability of a conformational search procedure to enhance sampling. Employment of fragment configurations reduces the size and complexity of the search space. Fragment configurations also capture local propensities of amino acid segments directly, allowing an energy function to focus on scoring instead longer range interactions. While many studies focus on exploring the relationship between fragment length, size of ensuing conformational space, and accuracy [18], [15], fragment

- *The authors are with the Department of Computer Science, George Mason University, 4400 University Dr., Fairfax, VA 22030.*
  E-mail: amardagmu.edu.

Manuscript received 21 Jan. 2013; accepted 19 Mar. 2013; published online 27 Mar. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2013-01-0023.
Digital Object Identifier no. 10.1109/TCBB.2013.29.

lengths typically employed by ab initio protocols are in the 3-20 range [5], [46].

In addition to the molecular fragment replacement technique, these protocols employ low-resolution representational detail for sampled decoy conformations. The low resolution reduces the dimensionality of the search space thus controlling the computational demands of sampling many conformations. However, the energy functions capable of scoring low-resolution conformations typically contain many artifacts that result in associating low energies with conformations sometimes 4-8 Å away from the known native structure of a protein sequence [44], [4], [10], [38].

Despite the employment of fragment configurations and low resolution, the efficiency and accuracy of ab initio protocols decrease with target size [17]. Many studies show that current protocols do not reliably scale to longer protein chains of more than 70 amino acids and topologies with many long-range contacts (often present in all β or α/β proteins) [15], [30], [31], [39]. It is currently unknown whether the thermodynamics treatment in ab initio structure prediction protocols fails due to the quality of the fragment libraries, the sampling capability of the conformational search procedure, or a combination of the three [23], [15], [4], [10], [38].

In this paper, we do not focus on investigating the quality of fragment libraries but employ instead the most recent ones available in the Rosetta package [20]. Other studies have shown that while the current quality of the protein data bank (PDB) [2] allows putting together high-quality fragment libraries, different fragment lengths can be more effective for different protein topologies [15]. In this paper, we employ fragments of length 9 and 3, constructed through the utilities available in the Rosetta package [20] as described in [13].

In this paper, we investigate the interplay between low-resolution sampling and low-resolution energy functions. We do so in the context of a robotics-inspired framework with high sampling capability that allows directly investigating the role that energy should play in guiding sampling. Our investigation of the interplay between sampling and energy builds on several observations by other studies on the importance of enhanced sampling versus the accuracy of the energy function.

The generation of a decoy ensemble that contains near-native conformations, albeit at low resolution, is of primary importance in a blind prediction setting. If the sampling is inadequate, the region containing the native structure may be missed altogether. However, even if the native structure is reproduced by some decoy conformation(s), the energy function may not rank near-native conformations as having lowest energies. Indeed, this is common with many low-resolution energy functions and has been demonstrated by many studies [36], [44], [4], [10], [38]. For instance, work in [4] demonstrates that the Rosetta force field not only are weakly funneled (that is, many geometrically dissimilar conformations can have comparable energies at low resolution), but they can also lead optimization initiated at the native structure away from it [4]. Other studies show that significant deviations of as much as 4 Å can exist between the global minimum of even an all-atom energy function and the experimentally available native structure [44].

Inaccuracies in a low-resolution energy function can lead a highly effective search or optimization technique to exploit artifacts of the energy function and so populate the decoy ensemble with very low-energy conformations of nonnative topologies. In fact, the hallmark of an effective search procedure is often the presence of nonnative conformations with much lower energies than the known native structure [10], [25], [24]. Even if the native structure is captured by some conformation(s) in the ensemble, the ensuing selection stage that picks a subset of decoys for further refinement may discard near-native conformations if they are high energy and/or insufficiently represented in the ensemble.

Effective selection of decoys for high-resolution refinement is a significant area of research [50]. The selection can be improved by building better low-resolution energy functions to score decoys. Since it is challenging to define such functions and find an energy threshold below which all the best decoys lie, the predominant strategy in selection techniques is to ignore energy altogether. Decoys are clustered by some measure of geometric similarity (predominantly, least root mean squared deviation—lRMSD), and the most populous or lowest energy cluster of decoys is selected for further refinement [33], [3]. Other approaches employ distance matrices [12] or structural profiles extracted from contact matrices [45], take into account correlations between decoys [42], or filter decoys through NMR data [9], [37].

The appropriateness and success of the selection technique are closely tied to the conformational search procedure employed for the generation of decoys in the first place. For instance, density-based clustering relies on the assumption that the sampled conformations are redundant; that is, the search procedure has sampled many geometrically similar decoys. This is certainly the case when numerous independent MMC trajectories are launched to obtain decoys. Moreover, the extraction of the most populous rather than the lowest energy cluster for refinement is based on the working assumption that the coarse-grained energy function may not preserve the depth of the native basin (the true global minimum) but may preserve its width [33], [5].

The predominant approach in many ab initio protocols is to essentially rely on numerous and long MMC trajectories to simultaneously obtain a broad view of the energy surface and converge to a region near the native structure. Achieving both comes with great computational cost. Massively parallel architectures are sometimes employed to manage the cost [32]. Improving efficiency suggests sacrificing redundancy, which, if implemented improperly, may affect both the broad view of the energy surface and convergence to the native structure. In turn, it may render selection techniques that essentially exploit redundancy less effective. Sacrificing redundancy, however, brings the focus back onto the exploration method and its ability to enhance the sampling of near-native conformations.

In this paper, we propose to separate the objective of the low-resolution exploration into two subgoals. We first propose to obtain a broad, nonredundant view of the energy surface. We do so through a robotics-inspired exploration framework that is efficient and allows, moreover, investigating the role of energy bias in the sampling of nonredundant decoys beyond the metropolis criterion. Unlike the predominant template, where numerous independent MMC trajectories implement energy bias locally through the metropolis criterion, our framework incorporates energy bias at a global level. Sampling is centralized into a tree search structure, whose branches are short MMC trajectories that employ molecular fragment replacement. Previous work by us on a particular realization of this framework has shown that biasing the growth of the tree allows effectively biasing sampling [36], [25]. Here, we show that the framework allows obtaining a broad view of the energy surface in terms of diverse low-energy decoy conformations.

Different techniques are investigated to tune the strength of the energy bias in the framework and steer sampling toward a particular distribution of decoys. A soft energy bias lowering the average energy of a growing ensemble of decoys is shown most effective in obtaining a distribution that facilitates the subsequent selection of good-quality decoys. This setting guards against the framework exploring deep energy minima representative of nonnative topologies that are artifacts of a given energy function. Energetic analysis of the decoy ensemble shows that soft rather than strong energy bias during the exploration allows retaining many near-native conformations even if a nonparametric energetic criterion is used for selection. Clustering of retained conformations shows that the ensemble obtained through a soft energy bias is structurally more diverse than when employing a strong energy bias.

We conduct our investigation in the context of two different well-known energy functions, the associative memory Hamiltonian with water (AMW) and the Rosetta energy function. We have used AMW in our previous work to guide different search procedures [35], [36], [24]. The Rosetta energy function is available to us from the Rosetta package [20]. While the ab initio protocol in Rosetta employs a suite of energy functions in a hierarchical scheme, all these functions are scaled versions of the full Rosetta energy function of 10 terms. We employ the full function for the purpose of our analysis. A couple of observations can be drawn from analysis of our results. First, no energy function is the clear winner according to various metrics on a list of proteins of different lengths and topologies. However, while the AMW energy surface seems easier to explore and saturate, the Rosetta energy function seems more complex and benefits more from enhanced sampling. Higher quality decoys closer to the known native structure can be obtained with the Rosetta energy function on proteins with β sheets, whereas AMW seems capable of providing similar quality on α proteins.

Finally, we investigate possible convergence to a region near the native structure in the low-resolution stage. Since MMC trajectories are well suited for convergence, we use them to optimize the nonredundant ensemble of low-energy decoys obtained with the robotics-inspired exploration framework. While the framework employs fragments of length 9 to efficiently obtain a broad view of a simplified energy surface, the MMC trajectories employ fragments of length 3 to further populate a more complex surface. Analysis shows that while convergence is easily reached for some proteins, the energy function can steer the search away to other low-energy regions in some cases. In all proteins, the top 10 populous clusters identified through density-based clustering contain near-native conformations that can be reliably used in a blind prediction setting for ab initio structure prediction. However, many nonnative topologies can be found populated sufficiently well among the top clusters. In a comparison of AMW to the Rosetta energy function, conformations closer to the native structure can be found among the top 10 clusters when using the Rosetta energy function.

The results presented in this paper make a first step into exploring the relationship between sampling and energy guidance in conformational search for ab initio structure prediction. Taken together, the results suggest that search frameworks with enhanced sampling capability are important to better understand current deficiencies in ab initio modeling to develop more accurate representations and energy functions. Our results confirm that the robotics-inspired search framework is promising in this direction and deserves further study.

2 Methods

We first describe the main ingredients of the robotics-inspired framework employed to obtain a broad view of the energy surface. We then relate details on the representation employed and the energy functions investigated in this paper. The energy biasing techniques investigated to control the distribution of decoys during exploration are described next. Finally, we describe how MMC trajectories are employed to study convergence.

2.1 Obtaining a Broad View of the Energy Surface with a Robotics-Inspired Sampling-Based Framework

The robotics-inspired framework we investigate here for its ability to provide a broad view of the energy surface has been proposed by our group before [36]. Instead of launching independent long MMC trajectories, the framework integrates many short MMC trajectories into a tree search structure. The tree maintains the growing ensemble of decoys and so provides a discrete representation of the sampled conformational space. The short MMC trajectories employ molecular fragment replacement to efficiently obtain protein-like conformations. The tree search structure allows the framework to make decisions on the fly about which trajectories should be extended. This is an important feature, as it allows the framework to adapt its exploration and bias it away from regions of the conformational space and energy surface that are already well represented in the tree.

To bias its exploration, the framework employs two discretization layers that facilitate analysis of the explored conformational space and energy surface. The employment of discretization layers is inspired by sampling-based motion-planning work in robotics [41], [29], [47], [19], [43]. The first discretization is over energies of sampled conformations, and the second is over their geometries. A 1D grid is associated with energies of conformations in the tree. The issue of finding coordinates to efficiently group

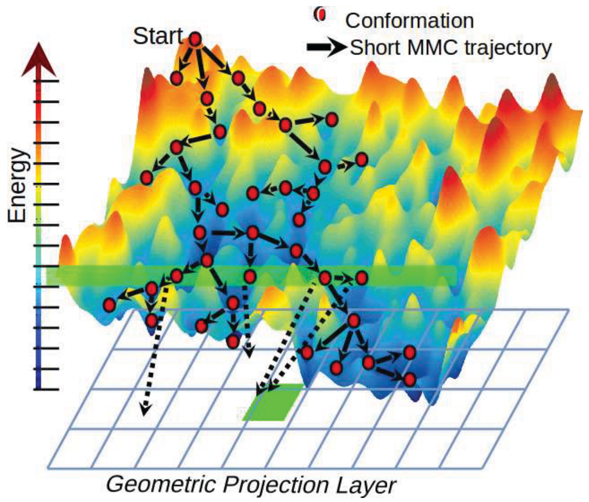


Fig. 1. An energy level is first selected as described. A second probability distribution function over all cells mapping to the selected energy level is then used to select a cell. The tree expands from a conformation selected uniformly at random over conformations mapping to the selected cell. Two coordinates instead of the three employed by the framework for the geometric projection layer are shown here for ease of visualization.

together structurally similar conformations is resolved by employing coarse geometric features about a conformation (average distance from the centroid, average distance from the point farther from the centroid, and so on). These features allow associating a 3D grid with conformations in the tree. Probability distribution functions can be defined over the discretization layers to bias the growth of the tree.

Application of the framework in previous work has focused on expediting the process of biasing the search toward lowest energy conformations and investigating different projection coordinates [36], [25], [26]. The 1D energy grid has been used to bias the selection toward conformations in lower energy levels (2 kcal/mol wide each) through the quadratic weight function $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell) + \epsilon$, where ϵ is a small value that ensures high-energy conformations have a nonzero probability of selection. A level ℓ is selected with probability $w(\ell)/\sum_{\ell' \in \text{Layer}_E} w(\ell')$. In this paper, we will refer to this probability distribution as the QUAD distribution. Once an energy level is selected, a cell belonging to it in the 3D geometric projection grid can be selected according to another probability distribution. A second weight function $1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$, where nsel records how often a cell is selected, and nconfs is the number of conformations projected to the cell. This function avoids cells that have been selected for expansion many times before and are already populated by many conformations. Once a cell is selected, any conformation in it can be selected at random for expansion; a short MMC trajectory from that conformation constitutes a new branch of the tree. This process is illustrated in Fig. 1.

The probability distributions over the discretization of the energy surface and over the discretization of the conformational space are shown to help the framework quickly populate low-energy regions [36]. The framework has been shown to have higher sampling capability than a long MMC trajectory, and the combination of both discretization layers are shown to improve sampling over using one of them in isolation or none at all (when both layers are turned off, the tree degenerates to an MMC trajectory) [36]. Fragments of length 3 and the AMW energy function have been employed in previous work. On many proteins, the exploration has been found to approach the native structure within 5 Å [36], [25], [26].

The objective in previous work has been to demonstrate that the framework improves coverage of the conformational space over independently running MMC trajectories. While QUAD biases the tree toward lower energies, employment of QUAD for the purpose of decoy generation risks exploiting minima that are artifacts of the energy function. However, the employment of probability distribution functions to ultimately control the distribution of sampled conformations make the framework particularly versatile for the purpose of decoy sampling and the study of deficiencies in ab initio modeling. Here, we show the first steps in this direction. We propose different probability distribution functions to implement the energy bias and show that one of them, corresponding to a soft energy bias, is better suited to obtain a broad nonredundant view of the energy surface through low-energy distinct decoys. We do so on two different state-of-the-art low-resolution energy functions and show that, while both allow capturing near-native conformations in the decoy ensemble, both are capable of associating very low scores with nonnative decoys. We now detail the implementation of the energy bias.

2.1.1 Implementing Energy Bias

The QUAD probability distribution function defined over weights $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell) + \epsilon$ described above essentially implements a strong energy bias that controls the growth of the tree through the expansion of lowest energy decoys to obtain even lower energy decoys. Note that the geometric projection grid is employed as above in conjunction with the energy bias. This setting can be very greedy and lead the framework, despite the bias away from oversampled cells in the conformational space, to deep energy minima that are artifacts of a given energy function. In contrast, one can ignore energy bias altogether. Essentially, all conformations can be treated as energetically equivalent and projected to the same energy level. Only the geometric projection grid and the probability distribution function defined on it (defined above over weights $1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$) can be employed. Let us refer to this probability distribution function as COV, as it essentially allows ignoring the energy surface and only steers the search to coverage of unsampled regions of the conformational space.

A new probability distribution function can be defined to implement a soft energy bias instead. As the tree and its conformational ensemble Ω grow, the mean (μ_Ω) and standard deviation (σ_Ω) can be updated over the energies of decoys. The mean tends to go lower over time, as the MMC trajectories that constitute the tree branches guide the tree toward lower energies through the metropolis criterion. The energy level whose average energy is closest to a sample drawn from the Gaussian distribution ($\mu_\Omega, \sigma_\Omega$) can be selected for expansions. The geometric projection grid is employed as above. We refer to this third realization of the framework as NORM. Unlike QUAD, NORM does not greedily bias the search tree toward the lowest energy decoys. Instead, the tree slowly grows toward low-energy decoys and associates low probabilities of selection to energy levels on either tail of the energy distribution.

2.2 Employed Representation and Energy Functions

2.2.1 Representational Detail

As we focus only on the low-resolution stage in ab initio modeling, the representation of a conformation sacrifices some structural detail. When employing the AMW energy function, the representation reduces side chains to only the C_β atom (with exception of glycine). When employing Rosetta energy function, the C_β atom is swapped for a centroid per side chain. Internally, two representations are maintained, one angular and another consisting of Cartesian coordinates. The angular representation maintains only three backbone dihedral angles $(\phi, \psi, omega)$ per amino acid, as sampled from the fragment configuration library. This representation is essentially the idealized geometry model, which fixes bond lengths and angles to idealized (native) values (taken from CHARMM22 [6]). The conversion from backbone dihedral angles to atomic coordinates, necessary for calculation of an energy score, employs forward kinematics [48]. Cartesian coordinates are only calculated for the backbone N, C, C_α, O atoms and either the C_β atom for amino acids with side-chain heavy atoms when using AMW or the side-chain centroid pseudo-atom when using the Rosetta energy function.

2.2.2 AMW Energy Function

This function, a modification of the low-resolution potential originally proposed in [28], has been used previously by us and others in the context of ab initio structure prediction [34], [36], [25], [26], [35], [16]. AMW sums nonlocal terms (local interactions are kept at ideal values in the idealized geometry model): $E_{\text{AMW}} = E_{\text{Lennard-Jones}} + E_{\text{H-Bond}} + E_{\text{contact}} + E_{\text{burial}} + E_{\text{water}} + E_{\text{Rg}}$. The $E_{\text{Lennard-Jones}}$ term is implemented after the 12-6 Lennard-Jones potential in AMBER9 [8] allowing a soft penetration of van der Waals spheres. The $E_{\text{H-Bond}}$ term allows modeling hydrogen bonds and is implemented as in [12]. The other terms, E_{contact}, E_{burial}, and E_{water}, allow formation of nonlocal contacts, a hydrophobic core, and water-mediated interactions and are implemented as in [30]. The E_{Rg} favors collapse by penalizing conformations with radius of gyration significantly different from theoretically calculated values [35].

2.2.3 Rosetta Energy Function

The Rosetta ab initio protocol uses a suite of different scoring functions in a hierarchical scheme. A total of six different scoring functions are used in the low-resolution stage in Rosetta. These correspond to different assignments to the weights that measure the contribution of different local and nonlocal energy terms. What we refer to as the Rosetta energy function is a linear combination of all possible 10 energy terms, which measure repulsion, amino acid propensities, residue environment, residue pair interactions, three terms measuring interactions between secondary structure elements, and three other terms measuring density and compactness of structure (see [33] for more details).

The low-resolution stage in the Rosetta protocol consists of four different substages, each with different scoring functions. The first substage conducts one to two cycles of 2,000 MMC moves each starting with an extended chain and using the score0 assignment. The only energy term modeled is a soft steric repulsion, and its purpose is to yield a random starting conformation. The second substage of 2,000 MMC moves uses score1 to accumulate secondary structure. The third substage uses five cycles of 2,000 MMC moves each with score2 followed by a cycle of 2,000 MMC moves with score5; score2 includes terms to favor hydrophobic collapse and beta strand pairings, whereas score5 lacks these two terms to allow relaxation. The fourth and final substage consists of three cycles of 4,000 MMC moves each and uses score3, which has all the possible energy terms except for hydrogen bonding. The ensuing selection analysis in preparation for side-chain packing and energetic refinement uses score4 to rank low-resolution conformations; score4 does not have any compaction or beta-strand pairing terms.

In light of this intricate protocol of different scoring functions, what we refer to as the Rosetta energy function in the comparison analysis in this paper corresponds to score3, as this is the one that has the highest number of Rosetta energy terms in the low-resolution stage, and all other scoring function in the low-resolution stage can be viewed as a scaled variant of score3.

2.3 Ensemble Analysis

We now describe techniques to compare the three different realizations of the exploration framework implementing the different energy biases.

2.3.1 Energetic Reduction

Reducing the ensemble Ω produced by the tree through an energetic criterion allows removing high-energy decoys added to the tree during the exploration. We employ a nonparametric threshold that discards any sampled conformation with energy higher than the mean. This threshold is not protein dependent and reduces the size of the ensemble by about 50 percent. While discarding about half the ensemble may sacrifice a few decoys with low lRMSDs to the native structure, the majority of low-lRMSD decoys are generally maintained in the reduced ensemble Ω_E. The results in Section 3 show that more low-lRMSD conformations are maintained when reducing the ensemble produced through QUAD and NORM. This is expected, as these two probability distribution functions implement an energy bias, and near-native conformations, while not among the lowest energy decoys, are associated with low energies. The results in Section 3 also show that more near-native conformations are retained when reducing the ensemble produced through NORM than QUAD, and this is particularly pronounced when using the AMW versus the Rosetta energy function.

2.3.2 Geometric Reduction

The framework employs coarse projection coordinates to efficiently group together similar conformations and bias the search on the fly away from oversampled regions. Employing lRMSD-based (see Section 3 for a description of lRMSD) comparisons and clustering would provide more detail and accuracy, but it would not be efficient. However, lRMSD-based clustering can be performed on the energetically reduced ensemble Ω_E both to analyze and compare the diversity of decoys across the three realizations of the framework and to further reduce the ensemble to a subset of distinct regions from which exploration can resume at greater detail.

We utilize an adaption of the bisecting K-means algorithm [40] on the Ω_E ensemble. Medioids instead of centroids are chosen to represent clusters so as to avoid irregular local structures resulting from angle averaging [51]. Initially, a conformation is selected at random to serve as the representative of the first cluster that encompasses all conformations in the ensemble. The essential process in bisecting K-means clustering is that a cluster is broken into two new ones if the minimum lRMSD from their cluster representative is above an ϵ threshold. Two random conformations are selected to serve as the representatives of the two new clusters. When conformations are reassigned, the representatives selected at random are replaced with the cluster medioids. The proximity of the conformations in each cluster is reevaluated. If the minimum lRMSD is above ϵ, the process begins anew (hence, bisecting). In the end, the medioids of the clusters are essentially a reduced representation of the Ω_E ensemble and constitute the $\Omega_{E,C}$ ensemble.

The bisecting K-means algorithm is less susceptible to initialization issues and does not require a priori determining the number of clusters. It requires, however, setting the maximum intracluster distance ϵ. In this work, we analyze the effect of two different values, 3 and 5 Å on the diversity of the resulting $\Omega_{E,C}$ ensemble.

2.4 Exploration Convergence

The reduced ensemble $\Omega_{E,C}$ can now be used to drive the exploration toward possible convergence on a more complex search space. A long MMC trajectory is launched from each conformation in $\Omega_{E,C}$. The trajectory length is a compromise between reaching convergence and controlling the overall computational cost. The fragment length employed here is 3 (9 is used by the framework above to obtain Ω). The shorter fragment length increases the complexity of the conformational space but also allows adding more detail to the energy surface.

The end points of the trajectories are analyzed through density-based clustering analysis [51]. An end point is assigned the number of neighbors that are within an lRMSD threshold of it (we use the same ϵ threshold above). The end point with the largest number of neighbors is considered to be the representative of the most populous cluster. This point and its neighbors are removed, and the process continues until all conformations have been exhausted. An exploration that started with obtaining a broad view of the energy surfaces terminates with revealing decoys in regions of the conformational space where many MMC trajectories converge. The results in Section 3 show that near-native conformations are retained among the top populous clusters; that is, the corresponding decoys are near native and as such are good candidates for high-resolution refinement.

TABLE 1
The PDB ID, nr. of Amino Acids, and Known Native Topology Are Shown for the 10 Proteins Studied Here

ID	1gb1	1sap	1wapa	1fwp	1ail	1aoy	1cc5	2ezk	3gwl	2h5nD
N	56	66	68	69	70	78	83	93	106	123
Fold	α/β	α/β	β	α/β	α	α/β	α	α	α	α

3 Materials

Systems of study. Ten protein systems are employed here, listed in Table 1. The systems range from 61-123 amino acids in length, cover α, β, and α/β folds, and include CASP targets. The list includes sequences longer than 70 amino acids and α/β native topologies known to be challenging for ab initio modeling.

Measurements. The main measurement used in the analysis below is lRMSD, which is the weighted euclidean distance between corresponding atoms after optimal superposition of two conformations under comparison. The optimal superposition refers to the rigid-body motion or transformation in SE(3) minimizing the weighted euclidean distance [22]. lRMSD captures structural dissimilarity but is not a euclidean metric, as it does not obey the triangle inequality. Low values indicate high similarity, and high values indicate high dissimilarity, but interpretation of intermediate values is difficult and the subject of many studies [21]. For instance, lRMSD has been found to depend on system size. A 5 Å lRMSD between a computed conformation and the native structure of a short protein chain of no more than 30 amino acids is considered a large deviation, but the same dissimilarity is less significant for a protein of 70 amino acids or more. In general, if the lowest lRMSD obtained over computed conformations to the known native structure is more than 6 Å, the native structure is not considered to have been captured.

High values of lRMSD do not necessarily indicate significant structural dissimilarity. Since lRMSD weighs each atom equally, it overly penalizes cases where differences are localized to a specific region, say a loop in different orientations in the two conformations under comparison. In such cases, other measurements, such as GDT_TS (global distance test total score), can be more appropriate. GDT_TS essentially locates a maximum subset of atoms between two conformations under comparison that are close in space after optimal superposition and minimizes an overall lRMSD-based error. GDT_TS is reported in percent and captures similarity, so higher values are better. As employed in CASP, $\mathrm{GDT\_TS} = (\mathrm{GDT\_P}_1 + \mathrm{GDT\_P}_2 + \mathrm{GDT\_P}_4 + \mathrm{GDT\_P}_8)/4$, where $\mathrm{GDT\_P}_d$ is the fraction of maximum amino acid subsets in a conformation superimposing on the reference (native, in our comparisons) structure with an lRMSD $\leq d$ Å. Some of our detailed analysis below employs GDT_TS scores in addition to lRMSD.

Implementation details. Each biasing scheme using each of the two energy functions is applied on each protein for 24 CPU hours on a 2.66-GHz Opteron processor with 8 GB of memory. This is repeated three times to obtain three ensembles per setting. Results and further analysis are presented on the ensemble that yields the median value in terms of lowest lRMSD from the native structure (lRMSD is

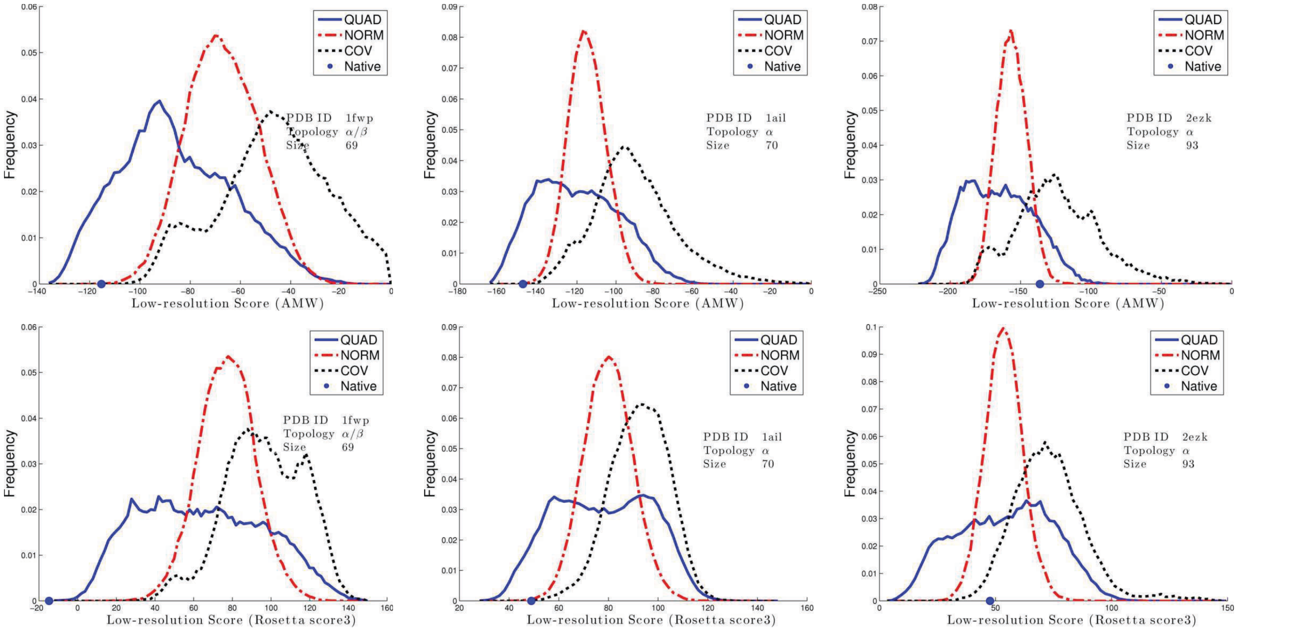


Fig. 2. Distributions of energies of Ω resulting from QUAD, COV, and NORM are superimposed over one another. The energy of the native structure is marked by a blue circle on the *x*-axis. While the top row shows results obtained with AMW, the bottom row shows results obtained with the Rosetta score3 function.

calculated over heavy backbone atoms). Clustering is conducted on a 2.4 Intel Xeon E5620 processor with 24 GB of memory. The MMC trajectories that optimize each decoy in the resulting ensemble $\Omega_{E,C}$ are limited to 20,000 steps and are run on a 2.66-GHz Opteron processor with 8 GB of memory. This second stage lends itself to embarrassing parallelization and takes 12-36 hours on 80 CPU cores depending on the size of $\Omega_{E,C}$ and protein length.

3.1 Analysis of Decoy Ensembles Obtained with Different Biasing Schemes

The distribution of conformational energies in Ω is shown for QUAD, COV, and NORM in Fig. 2 on three selected proteins. Superimposition of the distributions shows that, as expected, QUAD results in lower energies (distribution is shifted to the left), whereas COV results in higher energies. The distribution obtained with NORM is expectedly Gaussian, and its mean energy is between the means of QUAD and COV. Each of the three distributions can contain lower energies than the native structure, whose energy is shown for reference.

Fig. 2 shows these results when either AMW or Rosetta score3 are employed. Due to detailed fine tuning in calculations of the Rosetta energy functions, the setting with Rosetta score3 runs six to seven times faster than when employing our in-house version of AMW. To conduct a fair comparison, the size of the conformational ensemble obtained when using Rosetta score3 is limited to the size obtained in 24 hrs with AMW on a particular protein and biasing scheme. For instance, if within 24 CPU hours, the ensemble obtained with AMW on the system with PDB ID 1fwp1 is 51K when using QUAD and 95K when using NORM, the ensemble sampled when using Rosetta score3 and QUAD is then limited to 51K conformations, and the ensemble sampled when using Rosetta score3 and NORM is limited to 98K conformations.

It is worth noting that one cannot directly compare values between the AMW and Rosetta energy functions. However, the location of the known native structure shows that both energy functions can associate low or high energies with a native structure. For instance, on the protein systems with PDB IDs 1fwp and 1ail, the native structure has lower energy than the mean of the energy distribution obtained under NORM whether AMW or Rosetta score3 are employed. On the system with PDB ID 2ezk, the native structure has higher energy than the mean under AMW but not Rosetta score3. On all three systems, lower energies than that of the native structure can be obtained under QUAD under each energy function due to the strong energy bias in QUAD driving the exploration toward deep nonnative minima.

Table 2 shows the lowest lRMSD obtained under each biasing scheme when using AMW or Rosetta. As in Fig. 2, the data are presented on the median ensemble (over three runs for each biasing scheme). Lowest lRMSDs under 6 Å are obtained by all three biasing schemes on most protein

TABLE 2
The Lowest lRMSD from the Native Structure Is Shown for Each of the Three Biasing Schemes

	lowest lRMSD(Å) over Ω					
	AMW			Rosetta score3		
ID	COV	QUAD	NORM	COV	QUAD	NORM
1gb1	4.7	5.0	4.6	4.4	**3.8**	4.1
1sap	6.8	6.5	5.2	5.9	5.9	**4.5**
1wapa	7.6	7.4	6.9	**6.4**	6.8	6.6
1fwp	6.6	6.9	6.1	5.8	5.1	**4.6**
1ail	3.5	2.5	**1.9**	4.7	4.7	4.6
1aoy	5.5	5.6	5.8	**5.0**	5.2	5.4
1cc5	5.9	**5.7**	5.8	6.5	5.9	5.8
2ezk	4.5	3.7	4.1	3.2	**3.1**	3.5
3gwl	6.1	5.5	6.0	**4.6**	6.0	6.5
2h5nD	9.0	**6.9**	9.0	8.9	9.9	11.1

Results are shown for both AMW and Rosetta score3.

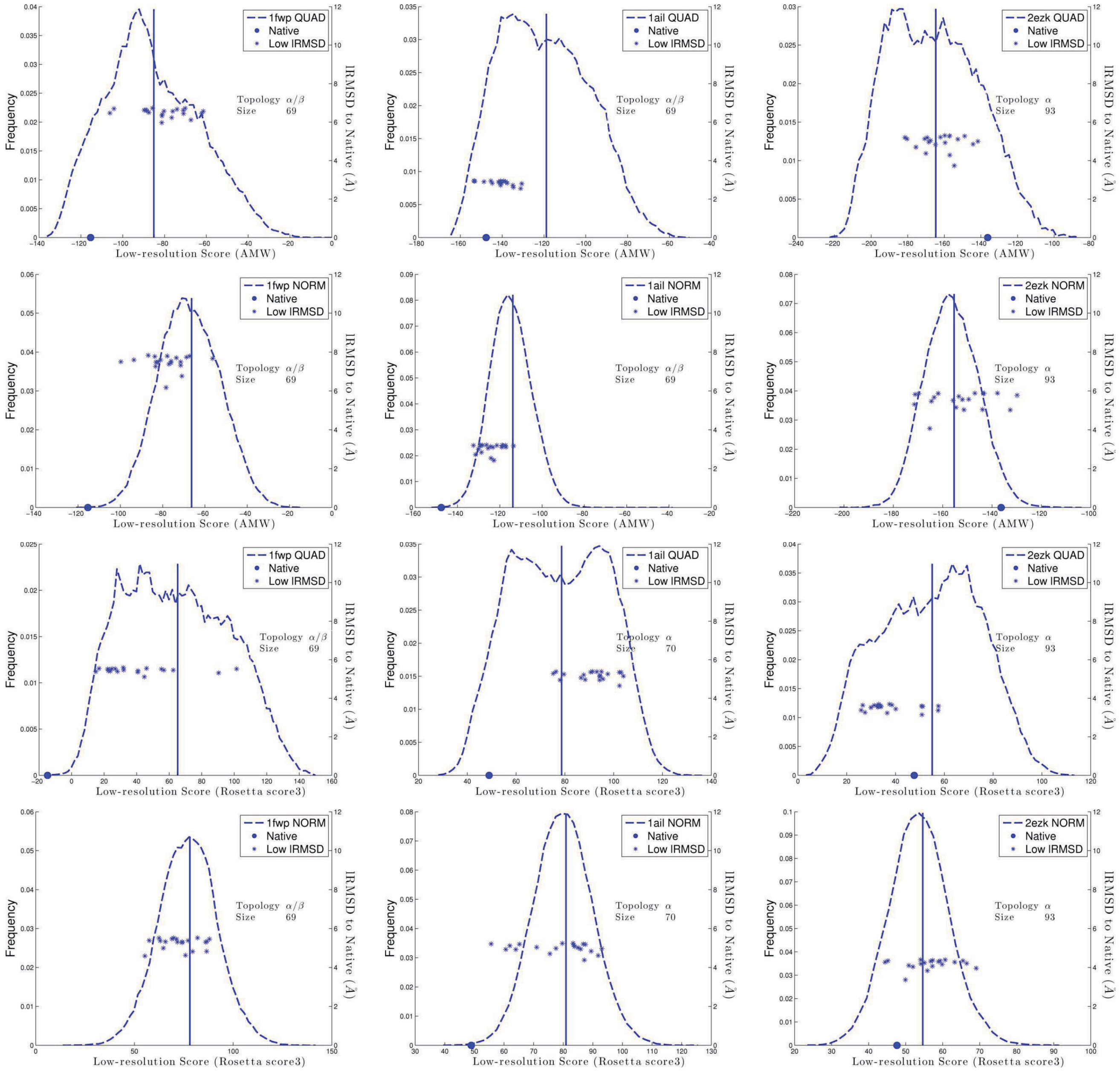


Fig. 3. The 20 lowest-lRMSD conformations are shown as blue circles over the distribution of energies in Ω for three selected protein systems. Their lRMSDs from the native structure are shown on the right hand axis. Results are shown for both AMW and Rosetta score3.

systems, whether AMW or Rosetta score3 are used. The global energy bias present in QUAD and NORM but not in COV improves proximity to the native structure (lower lowest lRMSDs are obtained overall). Moreover, when using AMW, lower lowest lRMSDs are obtained on 50 percent of the systems with NORM than QUAD, comparable lowest lRMSDs within 0.2 Å are obtained on 20 percent of the systems, and increases are observed on the rest. When using the Rosetta energy function, differences in lowest lRMSDs between NORM and QUAD are less pronounced, suggesting than the Rosetta energy surface is more complex than AMW and can benefit from further sampling. A comparison between AMW and Rosetta score3 reveals that the lowest lRMSD is obtained by Rosetta score3 (in bold), whether COV, NORM, or QUAD are used. AMW seems to have a significant advantage on 1ail and obtains comparable results on 1cc5, both all-α proteins. Results are uniformly poor on 2h5nD, suggesting this longest protein may benefit from further sampling.

Focusing on the lowest lRMSD may be misleading, as the conformation realizing it may not be sufficiently represented in the decoy ensemble or may be missed altogether by a selection technique. Fig. 3 analyzes Ω in some more detail for the three selected protein systems. The 20 decoys with the lowest lRMSDs from the native structure are marked in the distribution of conformational energies obtained with each biasing scheme.

Fig. 2 shows that many of the 20 lowest-lRMSD conformations can be lost if the selection criterion discards those with energies above the mean in the ensembles obtained with AMW and QUAD. Many of these conformations would be retained if using NORM. Differences between

TABLE 3
AMW and Rosetta Energy Functions Are Compared over Entire Ω Ensemble Obtained with NORM

	lRMSD$_{\min}$(Å)		GDT_TS$_{\max}$(%)		lRMSD$_{\mu,p90}$(Å)		GDT_TS$_{\mu,p90}$(%)	
ID	AMW	Rosetta (score3)	AMW	Rosetta (score3)	AMW	Rosetta (score3)	AMW	Rosetta (score3)
1gb1 (α/β)	4.6	**4.1**	0.63	**0.69**	11.4	**9.3**	0.39	**0.49**
1sap (α/β)	5.2	**4.5**	**0.52**	**0.52**	**10.6**	11.9	**0.34**	**0.32**
1wapa (β)	6.9	**6.6**	0.39	**0.43**	**13.0**	13.7	0.23	**0.28**
1fwp (α/β)	6.1	**4.6**	0.48	**0.53**	12.4	**11.1**	0.30	**0.35**
1ail (α)	**1.9**	4.6	**0.84**	0.65	**9.8**	11.0	**0.43**	0.37
1aoy (α/β)	5.8	**5.4**	0.57	**0.62**	**9.9**	12.2	**0.40**	0.36
1cc5 (α)	**5.8**	**5.8**	**0.45**	**0.46**	**12.3**	13.2	0.28	**0.30**
2ezk (α)	4.1	**3.5**	0.56	**0.70**	11.7	**8.4**	0.34	**0.50**
3gwl (α)	**6.0**	6.5	**0.44**	**0.46**	**13.4**	15.6	**0.30**	**0.31**
2h5nD (α)	**9.0**	11.1	**0.33**	0.24	**15.5**	16.7	**0.23**	0.19

In addition to the lowest lRMSD and maximum GDT_TS to the known native structure, the comparison includes mean lRMSD and mean GDT_TS over the 90th percentile (p90) of low-energy conformations in Ω.

QUAD and NORM are less pronounced when using Rosetta, suggesting again that the Rosetta energy surface is more complex. We point out that the system with PDB ID 1ail, an all α protein, seems to be an easier case for AMW than Rosetta. Whether using QUAD or NORM with AMW, the 20 lowest-lRMSD conformations have energies not only below the mean but also close to that of the native structure. On the other hand, the system with PDB ID 2ezk seems to be more challenging for AMW than Rosetta. When using AMW, the 20 lowest lRMSD conformations have energies that place them above the mean whether using QUAD or NORM. In contrast, when using Rosetta score3, many of these conformations are close in energy to the native structure, which also falls below the mean both under NORM and QUAD. We note that this system is a longer α protein of 93 amino acids.

A further comparison between AMW and the Rosetta energy function can be conducted by comparing not only the lowest lRMSDs or the highest GDT_TS scores to the known native structure obtained on each system but also the mean lRMSD and the mean GDT_TS score on the 90 percent percentile of low-energy conformations. The results shown in Table 3 fix the biasing scheme to NORM and limit the source of variation to the energy function employed. Values in bold indicate either lower or comparable lRMSDs between AMW and Rosetta or higher or comparable GDT_TS scores between AMW and Rosetta. If focusing on lowest lRMSDs, Rosetta provides scores that are lower or comparable than those obtained with AMW on $7/10$ of the systems. Looking at GDT_TS scores brings the number of systems with higher or comparable GDT_TS scores in Rosetta to $8/10$. Interestingly, the majority of the improvements are on proteins with β or α/β folds. On the majority of the all-α proteins, AMW provides better or similar results.

Comparing mean lRMSDs and mean GDT_TS scores over the 90th percentile of low-energy conformations reveals that differences between Rosetta and AMW in terms of representation of near-native conformations are less stark. Rosetta has lower or comparable mean lRMSDs or higher or comparable mean GDT_TS scores on this subensemble of conformations on 30 and 70 percent of the systems, respectively. Taken together, these results provide a detailed insight into AMW and Rosetta. While Rosetta seems capable of better recognition of conformations in close proximity to the native structure, neither energy function has a distinct advantage for the purpose of a selection technique driven by an energy cutoff.

3.2 Ensemble Reduction and Analysis

Since our goal for the robotics-inspired exploration is to obtain a broad nonredundant view of the energy surface, QUAD and NORM are further investigated in terms of the geometric diversity of the Ω_E ensembles they yield (discarding any conformation with energy above the mean). Since the bisecting K-means clustering employed for this purpose makes use of an $N \times N$ matrix to store pairwise lRMSDs between the N decoys in Ω_E, the size of Ω_E can pose computational and memory issues. We impose a limit of 40K conformations. When the limit is exceeded, uniform sampling over Ω_E is used to obtain 40K conformations. Table 4 shows $|\Omega|$ and $|\Omega_E|$ for each protein in columns 2-3 for QUAD and 6-7 for NORM. Larger Ω ensembles are obtained on all proteins with NORM, confirming that it becomes increasingly harder to satisfy the metropolis criterion (and so expand selected conformations) from the lowest energy levels selected by QUAD. The difference in $|\Omega|$ between QUAD and NORM becomes less pronounced on the longer proteins, where energy evaluations become the bottleneck.

The reduction in size of $\Omega_{E,C}$ resulting from the clustering of Ω_E is shown in columns 4-5 and 8-9 of Table 4 for QUAD

TABLE 4
$|\Omega|$ and $|\Omega_E|$ Obtained When Using AMW Are Shown in Units of 10^3

ID	AMW							
	QUAD				NORM			
	$\lvert\Omega\rvert$	$\lvert\Omega_E\rvert$	ΔC_3	ΔC_5	$\lvert\Omega\rvert$	$\lvert\Omega_E\rvert$	ΔC_3	ΔC_5
1gb1	101	40	57%	83%	168	40	28%	65%
1sap	70	40	76%	90%	105	40	35%	51%
1wapa	45	26	78%	86%	84	42	37%	52%
1fwp	51	33	73%	88%	95	40	31%	51%
1ail	73	38	76%	90%	94	40	58%	80%
1aoy	57	31	73%	90%	71	35	47%	72%
1cc5	37	33	71%	83%	55	28	32%	43%
2ezk	38	20	63%	87%	42	21	43%	85%
3gwl	23	12	70%	85%	28	14	47%	75%
2h5nd	15	8	61%	76%	18	9	55%	69%

Δ_C shows $|\Omega_E| - |\Omega_{E,C}|$ as a percent of Ω_E. Subscripts 3 and 5 refer to ϵ values 3 and 5 Å.

TABLE 5
$|\Omega|$ and $|\Omega_E|$ Obtained When Using Rosetta score3 Are Shown in Units of 10^3

ID	Rosetta score3							
	QUAD				NORM			
	$\|\Omega\|$	$\|\Omega_E\|$	ΔC_3	ΔC_5	$\|\Omega\|$	$\|\Omega_E\|$	ΔC_3	ΔC_5
1gb1	101	50	70%	92%	168	40	65%	90%
1sap	70	33	69%	84%	105	52	54%	76%
1wapa	45	25	85%	95%	84	41	38%	65%
1fwp	51	26	70%	84%	95	47	50%	75%
1ail	73	36	71%	85%	94	47	57%	78%
1aoy	57	29	70%	89%	71	36	49%	83%
1cc5	37	18	80%	87%	55	28	64%	77%
2ezk	38	18	70%	93%	42	21	64%	94%
3gwl	23	12	72%	91%	28	14	48%	67%
2h5nd	15	8	76%	88%	18	9	52%	78%

Δ_C shows $|\Omega_E| - |\Omega_{E,C}|$ as a percent of Ω_E. Subscripts 3 and 5 refer to ϵ values 3 and 5 Å.

TABLE 6
The Lowest lRMSD from the Native Structure over Conformations in Top i Clusters ($i \in 1, 5, 10$) Are Shown in Columns 2-4, Respectively

ID	lRMSD to Native (Å)						
	T_1	T_5	T_{10}	B_{10}	B_1	G_{f9}	G_{f3}
1gb1	11.2	11.2	10.7	6.6	6.1	3.7	9.0
1sap	6.4	6.4	6.4	6.8	5.7	8.4	6.4
1wapa	10.4	10.4	9.0	7.5	6.1	17.8	6.3
1fwp	11.9	9.5	9.5	6.7	5.9	11.0	17.0
1ail	7.2	4.1	4.1	3.9	3.4	2.1	1.5
1aoy	7.1	7.1	6.9	6.0	5.0	12.9	11.5
1cc5	8.9	8.9	8.2	6.3	5.6	6.0	5.6
2ezk	7.9	7.4	7.4	5.9	4.8	10.4	9.8
3gwl	9.1	6.8	6.5	6.3	5.5	16.2	10.7
2h5nd	12.0	11.4	11.4	9.4	8.4	7.8	8.0

The 10th lowest and the lowest lRMSD over the entire $\Omega_{E,C}$ are shown for reference in columns 5-6, respectively. The lRMSD of the conformation resulting from global fit with fragment lengths of 9 and 3 is shown in columns 7-8, respectively.

and NORM. Results are shown for ϵ values of 3 and 5 Å (a higher value would degenerate the quality of the clusters). As expected, a higher ϵ value results in a more significant reduction over Ω_E. Moreover, comparison between QUAD and NORM for a given ϵ shows that clustering is able to achieve a more substantial reduction on the Ω_E ensemble resulting from QUAD. This suggests that NORM results in a more diverse set of low-energy decoys, and so it is better suited to be employed for the purpose of obtaining a broad view of the energy surface. The improved diversity of low-energy decoys implies increased coverage of the conformational space, which is a critical component, especially if it is to be followed by further more detailed exploration or studies focusing on improvements of energy functions on a diverse set of decoys. The results shown in Table 4 are overall reproduced when using Rosetta score3, shown in Table 5. A more substantial reduction is obtained on the ensemble obtained with QUAD using Rosetta score3, as well, further suggesting that the soft energy bias in NORM is more appropriate at yielding a diverse nonredundant decoy ensemble not exploiting artifacts of an energy function.

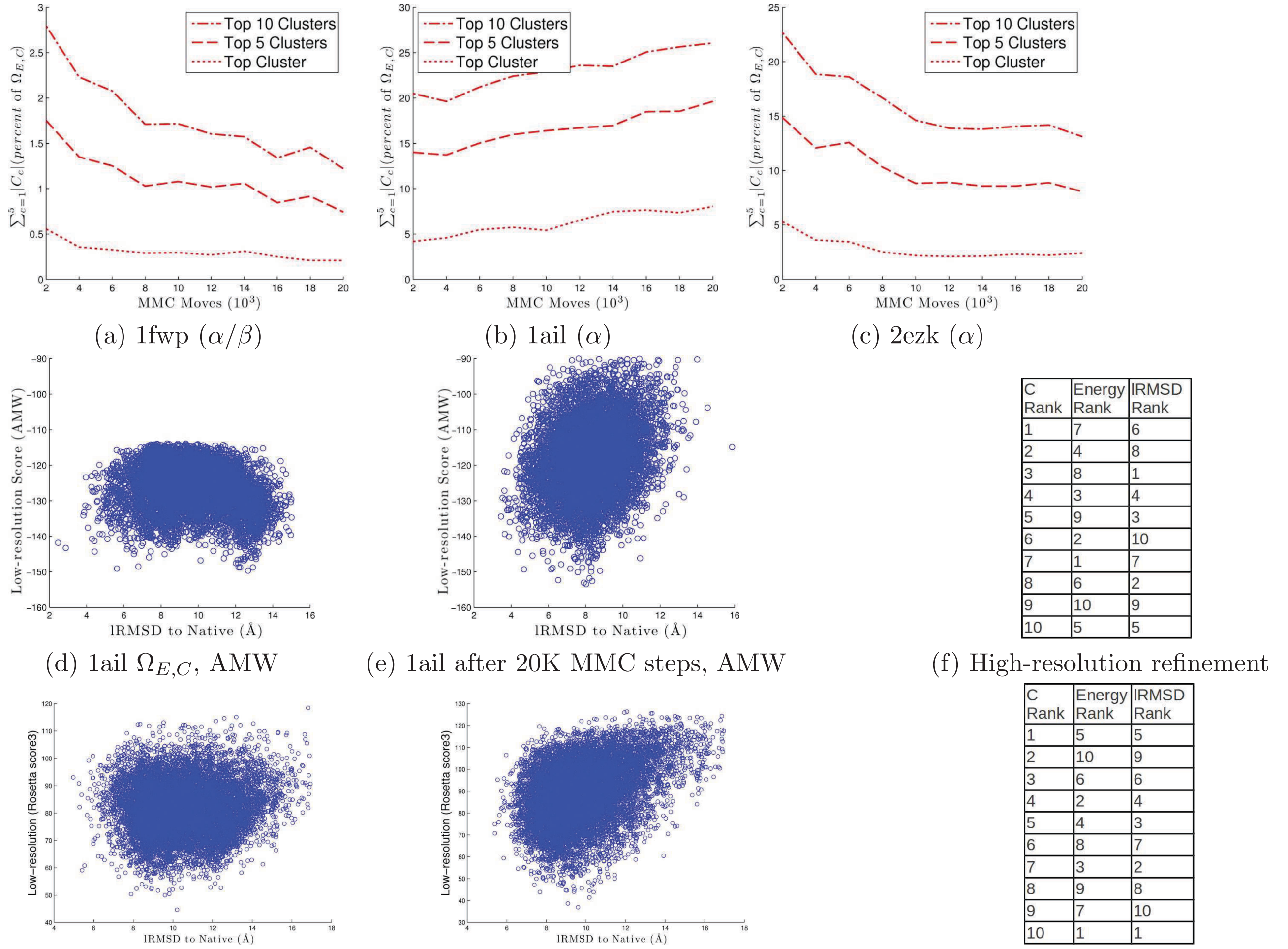
3.3 Convergence Analysis

Here, we conduct further analysis and optimization of obtained decoys. The conformations in Ω_{E_C} (medioids of clusters) resulting from NORM now serve as starting points for MMC trajectories (20,000 steps long). Unlike the previous stage, which uses fragments of length 9, the MMC trajectories use fragments of length 3. The end points of the trajectories constitute the final set of conformations subjected to density-based analysis to detect possible regions of convergence.

The quality of the top 10 clusters resulting from the density-based analysis with $\epsilon = 5$ is shown for each of the protein systems in Table 6a. The results shown in Table 6a are obtained with AMW. Columns 2-4 show the lowest lRMSD from the native structure over the representatives of the top i populous clusters, where i varies from 10, 5, down to 1, respectively. For reference, columns 5-7 show the lowest lRMSD and the 10th lowest lRMSD over the entire $\Omega_{E,C}$ ensemble. Additionally, columns 8-9 show the lRMSD of the conformation that can be assembled if the fragment configuration selected from the library for each fragment is the one that is closest to the actual fragment configuration in the native structure (a process known as global fit [36]).

Comparison of these columns allows drawing a few conclusions. If either the top five or top 10 populous clusters are employed for further refinement, near-native decoys (in terms of low lRMSDs) are preserved after the selection, promising recovery of the native structure in great detail and accuracy. Comparison of columns 4 and 5 shows that at most the selection loses ≈4 Å in terms of proximity to the native structure and on average loses 1.5 Å. In general, there is good correlation between cases when low lRMSDs are maintained by the selection and low lRMSDs obtained by global fit. Lower lRMSDs obtained over global fit suggest that sometimes suboptimal fragment configurations are needed locally to obtain a better global conformation. Similar observations can be drawn from the density analysis over ensembles obtained with Rosetta score3. The Rosetta score3 improves the quality of the lowest lRMSD among the top 10 clusters on some systems but it offers no distinct advantage overall (data not shown).

Further detailed analysis is showcased on three representative systems. The density-based analysis is repeated on the set of conformations resulting after every 2,000 MMC steps (AMW is used), and the aggregate size of the top i populous clusters $i \in \{1, 5, 10\}$ is shown in Figs. 4(a), 4(b), and 4(c) for each system. The results in Figs. 4(a), 4(b), and 4(c) showcase that this aggregate size can decrease, settle, or grow. A decrease is the result of MMC trajectories diverging in the energy surface. In Fig. 4(b), which shows results for the system with PDB ID 1ail, the most populated clusters grow in size, signaling convergence of many MMC trajectories to nearby regions for this system; the clusters contain a large percentage of the decoys when $\epsilon = 5$ Å. Repeating the analysis with $\epsilon = 3$ Å shows that 3 Å is too small to measure convergence (data not shown). Convergence on the system with PDB ID 1ail suggests that the widest low-energy basins captured by the robotics-inspired framework with AWM and NORM are also deep enough for the ensuing MMC runs to remain trapped. This result provides further insight into why it is that the low-resolution exploration of the AMW energy surface for this system can capture decoys within 2 Å of the



(a) 1fwp (α/β) (b) 1ail (α) (c) 2ezk (α)

(d) 1ail $\Omega_{E,C}$, AMW (e) 1ail after 20K MMC steps, AMW

C Rank	Energy Rank	lRMSD Rank
1	7	6
2	4	8
3	8	1
4	3	4
5	9	3
6	2	10
7	1	7
8	6	2
9	10	9
10	5	5

(f) High-resolution refinement

C Rank	Energy Rank	lRMSD Rank
1	5	5
2	10	9
3	6	6
4	2	4
5	4	3
6	8	7
7	3	2
8	9	8
9	7	10
10	1	1

(g) 1ail $\Omega_{E,C}$, Rosetta (score3) (h) 1ail after 20K MMC steps, Rosetta (score 3) (i) High-resolution refinement

Fig. 4. (a)-(c) The aggregate size of the top i clusters $i \in \{1, 5, 10\}$ resulting from density-based analysis with $\epsilon = 5$ Å is shown every 2K MMC steps (red lines). (d)-(f) Energy versus lRMSD from the native structure is plotted for system with PDB ID 1ail for conformations in $\Omega_{E,C}$ in (d) and for the end points of the MMC trajectories in (e). These results are obtained with AMW and NORM. (f) also shows the energetic and lRMSD ranking of the top 10 populous cluster representatives after a short high-resolution refinement. (g)-(i) Energy versus lRMSD from the native structure is plotted for system with PDB ID 1ail for conformations in $\Omega_{E,C}$ in (g) and for the end points of the MMC trajectories in (h). These results are obtained with the Rosetta score3 energy function and NORM. (i) also shows the energetic and lRMSD ranking of the top 10 populous cluster representatives after a short high-resolution refinement.

native structure. In contrast, the other two systems have shallower basins in the AMW energy surface.

Figs. 4(d), 4(e), and 4(f) provide some more detail on the system with PDB ID 1ail. The distribution of energies versus lRMSDs from the native structure of the conformations (medioids) in $\Omega_{E,C}$ in Fig. 4(d) shows that AMW is weakly funneled over the 9-mer space. Fig. 4(e) shows that the correlation between low energies and low lRMSDs improves after the MMC trajectories populate the 3-mer space. Moreover, a proof-of-concept analysis takes the top 10 clusters resulting from the density based for this system and subjects them to short high-resolution refinement through the Rosetta relaxation protocol. The resulting energetic and lRMSD ranks shown in Fig. 4(f) make the case that the top 10 clusters are good-quality candidates for further refinement. The same analysis is repeated over ensembles obtained with Rosetta score3 on this system, shown in Figs. 4(g), 4(h), and 4(i). In contrast to AMW, Rosetta yields stronger funneling on the 3-mer space despite the lowest lRMSD to the native structure being higher than what is obtained with AMW. Results showing ranks after high-resolution refinements of the top 10 clusters in Fig. 4(i) are similar to those obtained with AMW.

4 Discussion

We propose a new approach to obtain promising decoys for ab initio protein structure prediction protocols. Instead of launching numerous long MMC trajectories to obtain both a broad view of the energy surface and convergence to regions that are promising for further refinement, we propose to separate this objective into two subgoals. A broad nonredundant view of the energy surface is first obtained through a robotics-inspired exploration framework. The framework employs discretization layers over the explored energy surface and conformational space to bias its exploration.

Our analysis of different probability distribution functions over the discretization layers shows that a Gaussian distribution is more suitable for a diverse ensemble of low-energy decoys. This distribution effectively implements a soft energy bias that guards the framework from converging too fast to deep energy minima. While additionally enforcing structural diversity through the geometric projection layer, the combination of a soft energy bias and coverage result in a diverse ensemble of low-energy decoys. A nonparametric energetic reduction and a K-means bisecting clustering algorithm allow further reducing the ensemble and show that near-native conformations are more likely to be retained when using the soft energy bias rather than more greedy schemes.

Comparison of ensembles obtained with AMW versus Rosetta allow drawing a few observations. First, Rosetta seems to allow improvements in terms of closer proximity to the known native structure by as much as 1.5 Å over AMW. This is more pronounced for proteins with all β or α/β folds. AMW seems better suited for all α proteins. This observation confirms recent analyses of versions of AMW in [30], [16], [31] that the function seems well equipped to capture the basin of all α fold proteins. In line with other studies of the Rosetta energy function [4], [10], [38], our analysis shows, however, that like AMW, Rosetta is capable of ranking lower in energy decoys with significantly nonnative topologies. A comparison of the different energy biasing schemes when using the Rosetta energy function indicates that the function results in a probably more complex surface than AMW. While the AMW surface is saturated more speedily by the framework, the Rosetta energy surface may benefit from further sampling.

The convergence analysis is conducted by applying long MMC trajectories to the reduced ensemble. Shorter fragment lengths of 3 instead of 9 are used to access a more detailed energy surface and further populate the regions indicated as promising by the above exploration. Switching from longer to shorter fragments during exploration is employed by other methods for structure prediction [5]. These methods perform this switch in the context of very long independent MMC trajectories. In this framework, longer fragments are used to gain a broader view of conformational space. Once the areas of interest are identified via energetic reduction and geometric clustering, shorter fragments are employed to optimize the energy function on the remaining ensemble. Density-based clustering over the end points of the trajectories shows that the top populous clusters retain near-native conformations that can be used for further refinement in a blind prediction setting for ab initio structure prediction.

Taken together, results presented in this paper suggest that the proposed framework for decoy sampling is versatile and allows exploring current open issues and deficiencies in ab initio modeling. Higher accuracy in energy functions is one direction to pursue. The density clustering analysis showcases that the enhanced sampling by the robotics-inspired framework results in many regions, including nonnative topologies, being sufficiently populated to be reported among the top 10 populated clusters. This result effectively indicates that the framework leads to a diverse set of highly populated energy basins of conformations. These basins can be used for further development of scoring functions to improve recognition of nonnative topologies.

Other directions that merit further investigation concern the balancing of different energetic objectives during exploration or the employment of different-length fragments to suit these objectives. Currently, in the Rosetta ab initio protocol, different versions of the Rosetta energy function are used to scale interactions during the progression of an MMC trajectory. It is interesting to pursue this direction further in a more unified way. Moreover, while the effective temperature employed in this work for the metropolis criterion is a medium-range temperature, incorporating a simulated annealing or an adaptive temperature schedule in the exploration will be considered. The adaptive schedule can allow the framework to expand to difficult regions and so further enhance sampling.

Acknowledgments

This work was supported in part by the US National Science Foundation (NSF) CCF No. 1016995 and NSF IIS CAREER Award No. 1144106.

References

[1] C.B. Anfinsen, "Principles That Govern the Folding of Protein Chains," *Science,* vol. 181, no. 4096, pp. 223-230, 1973.

[2] H.M. Berman, K. Henrick, and H. Nakamura, "Announcing the Worldwide Protein Data Bank," *Nature Structural Biology,* vol. 10, no. 12, pp. 980-980, 2003.

[3] M.R. Betancourt and J. Skolnick, "Finding the Needle in a Haystack: Educating Native Folds from Ambiguous Ab Initio Protein Structure Predictions," *J. Computational Chemistry,* vol. 22, no. 3, pp. 339-353, 2001.

[4] G.R. Bowman and V.S. Pande, "Simulated Tempering Yields Insight into the Low-Resolution Rosetta Scoring Functions," *Proteins: Structure Function Bioinformatics,* vol. 74, no. 3, pp. 777-788, 2009.

[5] P. Bradley, K.M.S. Misura, and D. Baker, "Toward High-Resolution De Novo Structure Prediction for Small Proteins," *Science,* vol. 309, no. 5742, pp. 1868-1871, 2005.

[6] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations," *J. Computational Chemistry,* vol. 4, no. 2, pp. 187-217, 1983.

[7] T.J. Brunette and O. Brock, "Guiding Conformation Space Search with an All-Atom Energy Potential," *Proteins: Structure Function Bioinformatics,* vol. 73, no. 4, pp. 958-972, 2009.

[8] D.A. Case, T.A. Darden, T.E.I. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, R.C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K.F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C. Schafmeister, W.S. Ross, and P.A. Kollman, *Amber 9.* Univ. of California, San Francisco, 2006.

[9] A. Cavalli, X. Salvatella, C.M. Dobson, and M. Vendruscolo, "Protein Structure Determination from NMR Chemical Shifts," *Proc. Nat'l Academy of Sciences USA,* vol. 104, no. 23, pp. 9615-9620, 2007.

[10] R. Das, "Four Small Puzzles That Rosetta Doesn't Solve," *PLoS One,* vol. 6, no. 5, article e20044, 2011.

[11] J. DeBartolo, G. Hocky, M. Wilde, J. Xu, K.F. Freed, and T.R. Sosnick, "Protein Structure Prediction Enhanced with Evolutionary Diversity: SPEED," *Protein Science,* vol. 19, no. 3, pp. 520-534, 2010.

[12] H. Gong, P.J. Fleming, and G.D. Rose, "Building Native Protein Conformations from Highly Approximate Backbone Torsion Angles," *Proc. Nat'l Academy of Sciences USA,* vol. 102, no. 45, pp. 16227-16232, 2005.

[13] D. Gront, D. Kulp, R. Vernon, C. Strauss, and Baker, "Generalized Fragment Picking in Rosetta: Design, Protocols and Applications," *PLoS One,* vol. 6, no. 8, article e23294, 2011.

[14] K.F. Han and D. Baker, "Global Properties of the Mapping between Local Amino Acid Sequence and Local Structure in Proteins," *Proc. Nat'l Academy of Sciences USA,* vol. 93, no. 12, pp. 5814-5818, 1996.

[15] J. Handl, J. Knowles, R. Vernon, D. Baker, and S.C. Lovell, "The Dual Role of Fragments in Fragment-Assembly Methods for De Novo Protein Structure Prediction," *Proteins: Structure Function Bioinformatics,* vol. 80, no. 2, pp. 490-504, 2011.

[16] J.A. Hegler, J. Laetzer, A. Shehu, C. Clementi, and P.G. Wolynes, "Restriction vs. Guidance: Fragment Assembly and Associative Memory Hamiltonians for Protein Structure Prediction," *Proc. Nat'l Academy of Sciences USA,* vol. 106, no. 36, pp. 15302-15307, 2009.

[17] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, and N.V. Grishin, "CASP9 Assessment of Free Modeling Target Predictions," *Proteins: Structure Function Bioinformatics,* vol. 79, no. 10, pp. 59-73, 2011.

[18] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small Libraries of Protein Fragments Model Native Protein Structures Accurately," *J. Molecular Biology,* vol. 323, no. 2, pp. 297-307, 2002.

[19] H. Kurniawati and D. Hsu, "Workspace-Based Connectivity Oracle: An Adaptive Sampling Strategy for PRM Planning," *Proc. Int'l Workshop Algorithmic Foundations of Robotics (WAFR '06),* pp. 35-51, 2006.

[20] A. Leaver-Fay, M. Tyka, S.M. Lewis, O.F. Lange, J. Thompson, R. Jacak, K. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, I.W. Davis, S. Cooper, A. Treuille, D.J. Mandell, F. Richter, Y.E. Ban, S.J. Fleishman, J.E. Corn, D.E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovi, J.J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J. Gray, B. Kuhlman, D. Baker, and P. Bradley, "ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules," *Methods Enzymology,* vol. 487, pp. 545-574, 2011.

[21] V.N. Maiorov and G.M. Crippen, "Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins," *J. Molecular Biology,* vol. 235, no. 2, pp. 625-634, 1994.

[22] A.D. McLachlan, "A Mathematical Procedure for Superimposing Atomic Coordinates of Proteins," *Acta Crystallographica A,* vol. 26, no. 6, pp. 656-657, 1972.

[23] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical Assessment of Methods of Protein Structure Prediction (CASP) Round IX," *Proteins: Structure Function Bioinformatics,* vol. 79, no. 10, pp. 1-5, 2011.

[24] B. Olson and A. Shehu, "Evolutionary-Inspired Probabilistic Search for Enhancing Sampling of Local Minima in the Protein Energy Surface," *Proteome Science,* vol. 10, no. Suppl 1, article S5, 2012.

[25] B. Olson, K. Molloy, and A. Shehu, "In Search of the Protein Native State with a Probabilistic Sampling Approach," *J. Bioinformatics and Computational Biology,* vol. 9, no. 3, pp. 383-398, 2011.

[26] B.S. Olson, K. Molloy, S.-F. Hendi, and A. Shehu, "Guiding Search in the Protein Conformational Space with Structural Profiles," *J. Bioinformatics and Computational Biology,* vol. 10, no. 3, article 1242005, 2012.

[27] J.N. Onuchic and P.G. Wolynes, "Theory of Protein Folding," *Current Opinion Structural Biology,* vol. 14, pp. 70-75, 1997.

[28] G.A. Papoian, J. Ulander, M.P. Eastwood, Z. Luthey-Schulten, and P.G. Wolynes, "Water in Protein Structure Prediction," *Proc. Nat'l Academy of Sciences USA,* vol. 101, no. 10, pp. 3352-3357, 2004.

[29] E. Plaku, L. Kavraki, and M. Vardi, "Discrete Search Leading Continuous Exploration for Kinodynamic Motion Planning," *Robotics: Science and System,* MIT Press, 2007.

[30] M.C. Prentiss, C. Hardin, M.P. Eastwood, C. Zong, and P.G. Wolynes, "Protein Structure Prediction: The Next Generation," *J. Chemical Theory Computation,* vol. 2, no. 3, pp. 705-716, 2006.

[31] M.C. Prentiss, D.J. Wales, and P.G. Wolynes, "Protein Structure Prediction Using Basin-Hopping," *J. Chemical Physics,* vol. 128, no. 22, pp. 225106-225106, June 2008.

[32] S. Raman, D. Baker, B. Qian, and R.C. Walker, "Advances in Rosetta Protein Structure Prediction on Massively Parallel Systems," *IBM J. Research and Development,* vol. 52, nos. 1/2 pp. 7-17, 2008.

[33] C.A. Rohl, C.E. Strauss, K.M. Misura, and D. Baker, "Protein Structure Prediction Using Rosetta," *Methods Enzymology,* vol. 383, pp. 66-93, 2004.

[34] A. Shehu, "An Ab-Initio Tree-Based Exploration to Enhance Sampling of Low-Energy Protein Conformations," *Robotics: Science and Systems,* pp. 241-248, MIT Press, 2009.

[35] A. Shehu, L.E. Kavraki, and C. Clementi, "Multiscale Characterization of Protein Conformational Ensembles," *Proteins: Structure Function Bioinformatics,* vol. 76, no. 4, pp. 837-851, 2009.

[36] A. Shehu and B. Olson, "Guiding the Search for Native-Like Protein Conformations with an Ab-Initio Tree-Based Exploration," *Int'l J. Robotic Research,* vol. 29, no. 8, pp. 1106-11227, 2010.

[37] Y. Shen, O. Lange, F. Delaglio, P. Rossi, J.M. Aramini, G. Liu, A. Eletsky, Y. Wu, K.K. Singarapu, A. Lemak, A. Ignatchenko, C.H. Arrowsmith, T. Szyperski, G.T. Montelione, D. Baker, and A. Bax, "Consistent Blind Protein Structure Generation from NMR Chemical Shift Data," *Proc. Nat'l Academy of Sciences USA,* vol. 105, no. 12, pp. 4685-4690, 2008.

[38] A. Shmygelska and M. Levitt, "Generalized Ensemble Methods for De Novo Structure Prediction," *Proc. Nat'l Academy of Sciences USA,* vol. 106, no. 5, pp. 94305-95126, 2009.

[39] D. Simoncini, F. Berenger, R. Shrestha, and K.Y.J. Zhang, "A Probabilistic Fragment-Based Protein Structure Prediction Algorithm," *PLoS One,* vol. 7, no. 7, article e38799, 2012.

[40] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," *Proc. KDD Workshop Text Mining,* 2000.

[41] M. Stilman and J.J. Kuffner, "Planning among Movable Obstacles with Artificial Constraints," *Int'l J. Robotics Research,* vol. 12, no. 12, pp. 1295-1307, 2008.

[42] A.W. Stumpff-Kane and M. Feig, "A Correlation-Based Method for the Enhancement of Scoring Functions on Funnel-Shaped Energy Landscapes," *Proteins: Structure Function Bioinformatics,* vol. 63, no. 1, pp. 155-164, 2006.

[43] J.P. van den Berg and M.H. Overmars, "Using Workspace Information as a Guide to Non-Uniform Sampling in Probabilistic Roadmap Planners," *Int'l J. Robotics Research,* vol. 24, no. 12, pp. 1055-1071, 2005.

[44] A. Verma, A. Schug, K.H. Lee, and W. Wenzel, "Basin Hopping Simulations for All-Atom Protein Folding," *J. Chemical Physics,* vol. 124, no. 4, article 044515, 2006.

[45] K. Wolff, M. Vendruscolo, and M. Porto, "Efficient Identification of Near-Native Conformations in Ab Initio Protein Structure Prediction Using Structural Profiles," *Proteins: Structure,* vol. 78, pp. 249-258, Jan. 2010.

[46] D. Xu and Y. Zhang, "Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-Based Force Field," *Proteins: Structure Function Bioinformatics,* vol. 80, no. 7, pp. 1715-1735, 2012.

[47] Y. Yang and O. Brock, "Efficient Motion Planning Based on Disassembly," *Robotics: Science and Systems,* pp. 97-104, MIT Press, 2005.

[48] M. Zhang and L.E. Kavraki, "A New Method for Fast and Accurate Derivation of Molecular Conformations," *J. Chemical Information Computer Sciences,* vol. 42, no. 1, pp. 64-70, 2002.

[49] Y. Zhang, "Progress and Challenges in Protein Structure Prediction," *Current Opinion Structural Biology,* vol. 18, no. 3, pp. 342-348, 2008.

[50] Y. Zhang and J. Skolnick, "Scoring Function for Automated Assessment of Protein Structure Template Quality," *Proteins: Structure, Function, and Bioinformatics,* vol. 57, no. 4, pp. 702-710, 2004.

[51] Y. Zhang and J. Skolnick, "Spicker: A Clustering Approach to Identify Near-Native Protein Folds," *J. Computational Chemistry,* vol. 25, no. 6, pp. 865-871, 2004.

Kevin Molloy received the BS and MS degrees in computer science from George Mason University in 1998 and 2011, respectively. He is currently working toward the PhD degree in computer science. His research interests include computational biology, analytical performance modeling, and parallel computation. He is a member of the ACM.

Sameh Saleh is currently a senior undergraduate student in the Department of Computer Science, George Mason University. His research interests include ab initio structure prediction, evolutionary computation, and conformational search. He is a member of the ACM.

Amarda Shehu received the BS degree in computer science and mathematics from Clarkson University, Potsdam, New York, and the PhD degree in computer science from Rice University, Houston, Texas, where she was a National Institutes of Health fellow of the Nanobiology Training Program of the Gulf Coast Consortia. She is an assistant professor in the Department of Computer Science, George Mason University. She holds affiliated appointments at the School of Systems Biology and the Department of Bioengineering, George Mason University. Her research contributions are in computational structural biology, biophysics, and bioinformatics with a focus on issues concerning the relationship between sequence, structure, dynamics, and function in biological molecules. Her research on probabilistic search algorithms for protein conformational spaces is supported by the US National Science Foundation (NSF). She received the NSF CAREER Award in 2012 for her research on a unifying framework for protein modeling. She is a member of the IEEE.