Biased Decoy Sampling to Aid the Selection of Near-native Protein Conformations

Kevin Molloy
Department of Computer Science,
George Mason University,
Fairfax, VA 22030
kmolloy1@gmu.edu

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

ABSTRACT

A central challenge in ab-initio protein structure prediction is the selection of low-resolution decoy conformations whose subsequent refinement leads to high-resolution near-native conformations. Successful selection strategies are tightly coupled with the exploration method employed to obtain decoys. Density-based clustering is often used to identify regions of the energy surface that are highly sampled by exploration trajectories. The trajectories are often numerous and long, because the goal is to obtain both a broad view of the energy surface and to converge to regions that are promising for further refinement. In this paper we separate this into two subgoals. We first investigate a robotics-inspired exploration framework and demonstrate its ability to steer sampling towards diverse decoy conformations. Once a broad view of the energy surface is obtained, Metropolis Monte Carlo trajectories continue the exploration from selected decoys. Density-based clustering then identifies regions where trajectories converge. The two exploration stages both employ molecular fragment replacement but gradually add more detail through different fragment lengths. Results on a diverse list of proteins show that highly-sampled regions contain near-native conformations that are worthy of further refinement for use in a blind prediction setting.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms

Algorithms

Keywords

protein structure prediction, probabilistic conformational search, near-native conformations, energy bias

*Corresponding Author

ACM BCB'12 October 7-10, 2012, Orlando, FL, USA

1. INTRODUCTION

The predominant protocol in ab-initio protein structure prediction is the generation of a large number of decoy conformations at low resolution followed by refinement of selected decoys at higher resolution [14]. The generation of a relevant decoy set that contains near-native conformations, albeit at low resolution, is a primary challenge. Another challenge has to do with identifying and selecting decoys whose subsequent refinement will lead to high-resolution near-native conformations. Ab-initio protocols are in principle more broadly applicable for structure prediction than homology-based methods that rely on the availability of homologs of known structure, but the efficiency and accuracy of ab-initio protocols decrease with target size [12]. Over the last years, however, considerable advancements have been made through the molecular fragment replacement technique [4–6, 8, 9].

The fragment replacement technique allows obtaining realistic decoy conformations by essentially assembling them with short structural pieces. The pieces, or fragment configurations, are extracted from known native structures of proteins and compiled into a non-redundant library [11]. 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 success of the subsequent refinement of selected decoys in recovering the native structure relies on two criteria. First, the ensemble of generated decoys needs to contain near-native conformations; the extent of the proximity of the decoys to the target's native structure often determines the final prediction accuracy. Second, the selection technique that analyzes the ensemble needs to recognize near-native conformations, so that their high-resolution refinement can successfully recover the native structure.

The typical strategy employed to enhance the sampling of near-native conformations is to generate a very large number of low-energy decoys. Efficiency concerns restrict methods to usage of coarse-grained energy functions. Since near-native conformations are associated with the lowest energies in the energy surface [17], most methods employ such functions to bias their sampling towards low-energy decoys. Predominantly, the conformations are end points of Molecular Dynamics (MD) or MC trajectories that optimize a given coarse-grained energy function. The MC-based methods

that essentially discretize the conformational space through the fragment replacement technique exhibit enhanced sampling over MD-based methods [21] and are currently the most successful in ab-initio structure prediction [5, 6, 9, 22]. Their success is tightly related to obtaining a broad view of the energy surface through numerous MC trajectories.

The techniques that subsequently analyze the large ensemble of decoy conformations produced by the exploration method to select the best decoys for further high-resolution refinement are either driven by energy or ignore it altogether. While coarse-grained energy functions employed during exploration allow efficiently obtaining low-energy conformations, they are often not useful in identifying the best decoys in the ensemble. Most functions are weakly-funneled; that is, many geometrically-dissimilar conformations can have comparable energies at low resolution. Moreover, due to inherent approximations in these functions, significant deviations of up to 4Å can exist between the global minimum and the experimentally-available native structure [29].

Effective selection of the best decoys is a significant area of research. The selection can be improved by building better low-resolution energy functions to score decoys [3, 32]. Since it is challenging to define such functions and find an energy threshold below which all the best decoys lie, the predominant strategy 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 [2, 20]. Other approaches employ distance matrices [10] or structural profiles extracted from contact matrices [30], take into account correlations between decoys [27], or filter decoys through NMR data [7,24].

The appropriateness and success of the selection technique are closely tied to the exploration method 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 exploration method has sampled many geometrically-similar decoys. This is certainly the case when numerous independent MC 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. Empirical evidence suggests this is often the case [5, 20].

The predominant approach in many exploration methods is to essentially rely on numerous and long MC 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 [19]. 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 exploration into two subgoals. We first propose to obtain a broad, non-redundant view of the energy surface. We do so through a robotics-inspired exploration framework that is efficient and allows further investigating the role of energy bias in the sampling of non-redundant decoys beyond the Metropolis criterion. Unlike the predominant template, where numerous independent MC trajectories implement energy bias locally through the Metropolis criterion, our framework incorporates energy bias at a global level. The sampling of decoys is centralized into a tree search structure, whose branches are short MC 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 [15, 23]. 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 towards 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. We show that the majority of near-native conformations are retained even if clustering is used.

Once a broad view of the energy surface is obtained, the second subgoal of the exploration, convergence in regions that are promising for high-resolution refinement, is then addressed. Since MC trajectories are well-suited for convergence, they are employed to optimize the non-redundant ensemble of low-energy decoys obtained with the robotics-inspired exploration framework. This process is conducted at a finer level of detail. While the robotics-inspired exploration framework employs fragments of length 9 to efficiently obtain a broad view of a simplified energy surface, the MC trajectories employ fragments of length 3. Results on a diverse list of proteins show that the top populous clusters identified through density-based clustering contain near-native conformations which can be reliably used in a blind prediction setting for ab-initio structure prediction.

2. METHODS

We first describe the main ingredients of the framework employed to obtain a broad view of the energy surface. We then relate details on the biasing techniques investigated to control the distribution of decoys. Finally, we describe how MMC trajectories are employed to achieve convergence.

2.1 Obtaining a Broad View of the Protein Energy Surface

The robotics-inspired framework we investigate here for its ability to provide a broad view of the energy surface has been proposed before by our group [23]. Instead of launching independent, long MMC trajectories, the framework essentially 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 [13, 18, 26, 28, 31]. 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 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 towards lowest-energy conformations and investigating different projection coordinates [15, 16, 23]. The 1d energy grid has been used to bias the selection towards 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 constitues a new branch of the tree.

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 [23]. Fragments of length 3 and the coarse-grained Associative Memory Hamiltonian with Water (AMW) energy function have been employed in previous work. On many proteins, the exploration approaches the native structure [15, 16, 23].

As is, the framework is not directly useful for decoy generation. The quick convergence to lowest-energy regions through QUAD risks exploring 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 ab-initio structure prediction protocols. Here we show the first steps in this direction. We propose two different probability distribution functions, compare them to QUAD, and show that one of them is better suited to obtain a broad non-redundant view of the energy surface through low-energy distinct decoys.

2.1.1 Implementing Energy Bias

The `QUAD` probability distribution function 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. One can ignore any 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) can be employed. Let us refer to this probability distribution function as `COV`, as it essentially allows ignoring the energy surface and only biases 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. 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 towards 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 towards the lowest-energy decoys. Instead, the tree slowly grows towards low-energy decoys.

2.2 Ensemble Analysis

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

2.2.1 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 non-parameteric 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%. 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 are associated with low energies.

2.2.2 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 comparisons and clustering would provide more detail and accuracy, but it would not be efficient. However, lRMSD-based clustering can be performed on 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 [25] 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 [33]. 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 bisect-

ing 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 mediods. 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 does require, however, setting the maximum intra-cluster distance ϵ. In this work, we analyze the effect of two different values, 3 and 5Å on the diversity of the resulting $\Omega_{E,C}$ ensemble.

2.3 Exploration Convergence

The reduced ensemble $\Omega_{E,C}$ can now be used to drive the exploration towards convergence. 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 [33]. 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 the top populous clusters are also characterized by low-lRMSDs; that is, the corresponding decoys are near-native and as such are good candidates for high-resolution refinement.

3. RESULTS

Results are shown for the 10 protein systems in Table 1, which range from 61-123 amino acids in length, cover α, β, and α/β folds, and include CASP targets. Each of the three realizations of the framework 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. The Ω ensemble employed for the subsequent analysis and exploration is the one that yields the median value in terms of lowest lRMSD from the native structure (lRMSD is calculated over heavy backbone atoms). Clustering over this ensemble 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 embarrasing parallelization and takes 12-36 hours on 80 CPU cores depending on the size of $\Omega_{E,C}$ and protein length.

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

Table 1: The Protein Data Bank [1] (PDB) ID, nr. of amino acids, and fold are shown for the 10 proteins studied here.

3.1 Ensemble Reduction and Analysis

The distribution of conformational energies in Ω is shown for each the three realizations `QUAD`, `COV`, and `NORM` in Figure 1 on three selected proteins. Superimposition of the distributions shows that `QUAD` results in lower energies (the distribution is shifted to the left), whereas `COV` results in higher energies. The distribution of energies obtained with `NORM` is expectedly Gaussian, and its mean energy is between the means of `QUAD` and `COV`. Each of the three distributions contain lower energies than the native structure, whose energy is shown for reference.

Table 2 summarizes the distribution of lRMSDs from the native structure by showing the lowest lRMSD obtained by each realization of the framework. As in Fig. 1, the data are presented on the median ensemble (over three runs for each realization). Comparison of the lowest lRMSDs suggests that low-lRMSD conformations are obtained by all three realizations on all protein systems (with the exception of the longest system with PDB ID 2h5nd, where a longer exploration may be needed). In addition, the global energy bias, present in `QUAD` and `NORM` but not in `COV`, improves proximity to the native structure (lower lowest lRMSDs are obtained). Comparison of `QUAD` to `NORM` highlights that the soft energy bias in `NORM` allows obtaining overall lower lowest lRMSDs.

ID	lowest lRMSD (Å) over Ω		
	`COV`	`QUAD`	`NORM`
1gb1	4.7	5.0	4.6
1sap	6.8	6.5	5.2
1wapa	7.6	7.4	6.9
1fwp	6.6	6.9	6.1
1ail	3.5	2.5	1.9
1aoy	5.5	5.6	5.8
1cc5	5.9	5.7	5.8
2ezk	4.5	3.7	4.1
3gwl	6.1	5.5	6.0
2h5nD	9.0	6.9	9.0

Table 2: The lowest lRMSD from the native structure is shown for each of the three realizations of the framework.

Fig. 2 analyzes Ω in some more detail for a selected protein system. In addition to the native structure, the ten decoys with the lowest lRMSDs from the native structure are marked in the distribution of conformational energies obtained with each realization of the framework. The majority of the ten lowest-lRMSD conformations have energies below the mean only in `NORM`. This suggests that `NORM` would be more effective and allow maintaining near-native conformations if an energy criterion is employed to reduce Ω. The mean can be employed as a non-parametric threshold to reduce Ω by about half in size. The remaining decoys in the resulting Ω_E ensemble can then be utilized for geometric analysis to seed the next stage of the exploration.

Since our goal for the robotics-inspired exploration is to obtain a broad non-redundant view of the energy surface,

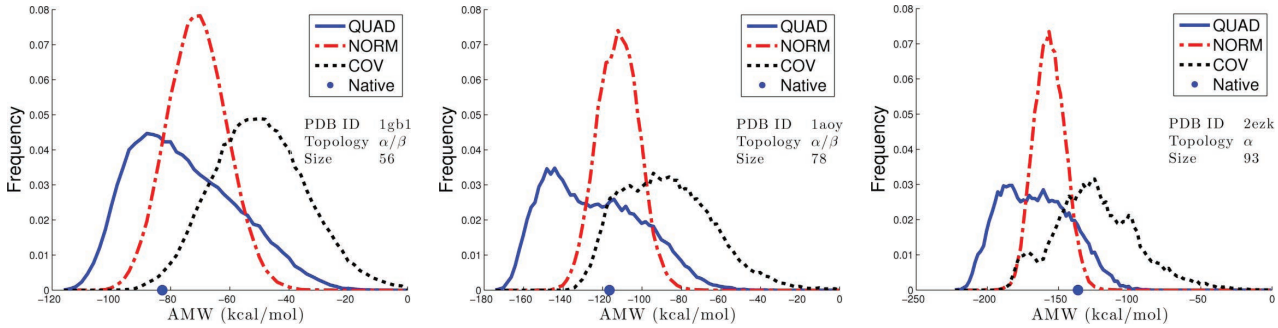


Figure 1: Distributions of energies of Ω resulting from QUAD, COV, and NORM are superimposed over one another. The energy of the native structure is marked by a blue circle on the x-axis.

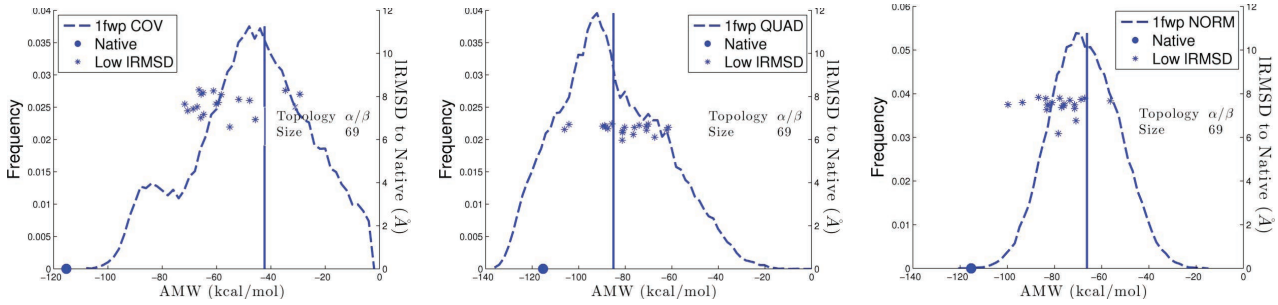


Figure 2: The ten lowest lRMSD conformations are shown over the distribution of energies in Ω for a selected protein system. The conformations are marked as blue circles. Their lRMSDs from the native structure are shown on the right hand axis.

the two realizations, QUAD and NORM are now investigated in terms of the geometric diversity of the ensembles they generate. Since the bisecting K-Means clustering employed for this purpose makes use of an $N \times N$ matrix to store the 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 40,000 conformations. When the limit is exceeded, uniform sampling over Ω_E is used to obtain 40,000 conformations. This happens on the shorter proteins where 24 CPU hours yields large Ω ensembles. Table 3 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, which confirms that it becomes increasingly harder to satisfy the Metropolis criterion (and so expand selected conformations to grow the tree) from the lowest-energy levels selected by QUAD. The difference in $|\Omega|$ between QUAD and NORM becomes less pronounced on the longer protein chains, 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 3 for QUAD and NORM. Results are shown for ϵ values of 3 and 5Å(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.

ID	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%

Table 3: $|\Omega|$ and $|\Omega_E|$ are shown in units of 10^3. Δ_C shows $|\Omega_{E,C}|$ as a percentage of Ω_E. Subscripts 3 and 5 refer to ϵ values 3 and 5Å employed during clustering.

3.2 Convergence Analysis

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 decoys subjected to density-based analysis to detect possible regions of convergence. Conformations are collected and analyzed every 2,000 MMC steps to obtain a dynamic picture of whether trajectories converge to certain regions.

Our detailed analysis is showcased on four representative protein systems selected from our test set. The density-based analysis is repeated on the set of conformations resulting after every 2,000 MMC steps and the aggregate size

of the top i populous clusters $i \in \{1, 5, 10\}$ is shown in Figure 3(a)-(d) for each system. The results in (a)-(d) showcase that this aggregate size can decrease, settle, or grow. A decrease is the result of the MMC trajectories converging to different regions of the energy surface. In (d), the most populated clusters grow in size, signaling convergence of many MMC trajectories to nearby regions for this system; the clusters also contain a large percentage of the decoys when ϵ=5 Å. On this system, results shown when the analysis is repeated with $\epsilon =$ 3Å make the case that 3Å is too small for the purpose of capturing convergence.

Figure 3(e)-(f) provide some more detail on this last system. The distribution of energies vs. lRMSDs from the native structure of the conformations (medioids) in $\Omega_{E,C}$ in (e) shows that the employed energy function is weakly funneled. Figure 3(f) shows that the correlation between low energies and low lRMSDs improves after the MMC trajectories. Moreover, a proof of concept analysis takes the top ten clusters resulting from the density-based analysis over the end points of the trajectories for this system and subjects them to short high-resolution refinement through the Rosetta relaxation protocol [5]. The resulting energetic and lRMSD ranks shown in Figure 3(g) make the case that the top 10 clusters are good-quality candidates for further refinement.

The quality of the top 10 clusters resulting from the density-based analysis is shown for each of the protein systems in Table 4. 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 tenth 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 [23]).

Comparison of these columns allows drawing a few conclusions. If either the top 5 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 $\approx$ 4Å 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 in order to obtain a better global conformation.

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

Table 4: 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. The tenth 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 are shown in columns 7-8, respectively.

4. CONCLUSION

We propose a new exploration method to obtain promising decoy conformations in the context of ab-initio protein structure prediction protocols. Unlike most protocols, which employ numerous and long MMC trajectories to obtain both a broad view of the energy surface and convergence to regions that are promising for further refinement, the method separates this objective into two subgoals.

A broad non-redundant 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 to obtaining an ensemble low-energy decoys that is broad and non-redundant. A non-parametric energetic reduction and a K-means bisecting clustering algorithm allow further reducing the ensemble.

The second goal of convergence is reached by applying long MMC trajectories to the reduced conformational ensemble. Shorter fragment lengths are used at this stage to access a more detailed energy surface. The technique of switching from longer to shorter length fragments during the 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 which can be used for further refinement in a blind prediction setting for ab-initio structure prediction.

Future work will consider several directions. An interesting direction is the comparison of different coarse-grained energy functions in the context of a common protocol as the one presented in this paper. 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.

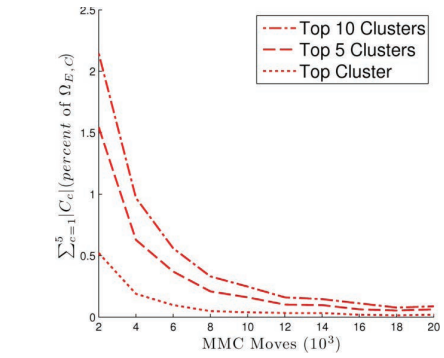
Acknowledgment

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

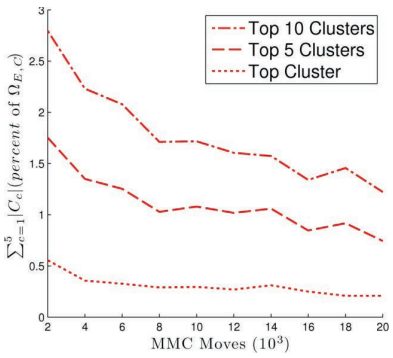
5. REFERENCES

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

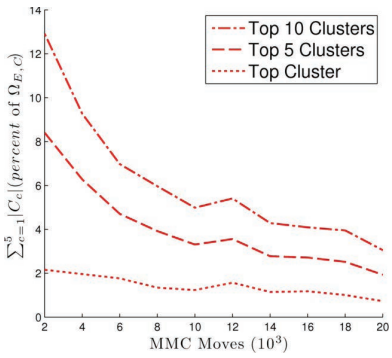
[2] M. R. Betancourt and J. Skolnick. Finding the needle in a haystack: educating native folds from ambiguous



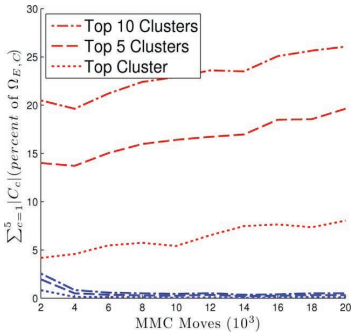
(a) 1wapa (β)



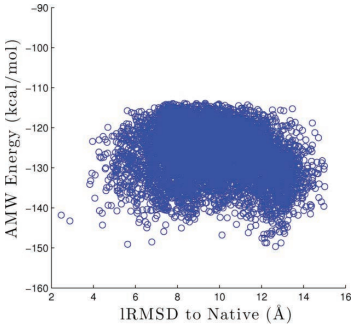
(b) 1fwp (α/β)



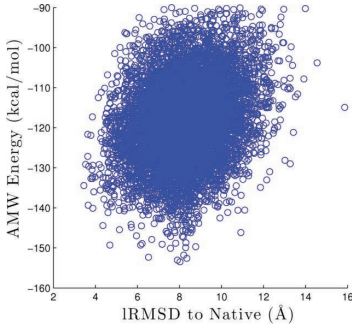
(c) 3gwl (α)



(d) 1ail (α)



(e) 1ail $\Omega_{E,C}$



(f) 1ail after 20K MMC steps

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

Figure 3: (a)-(d)The aggregate size of the top i clusters $i \in \{1, 5, 10\}$ resulting from density-based analysis with $\epsilon = 5$Åis shown every 2K MMC steps (red lines). (d) The aggregate size is shown for $\epsilon = 3$Å, as well (blue lines). (e)-(f) Energy vs. lRMSD from the native structure are plotted for system with PDB id 1ail for conformations in $\Omega_{E,C}$ in (e) and for the end points of the MMC trajectories in (f). (f) also shows the energetic and lRMSD ranking of the top 10 populous cluster representatives after a short high-resolution refinement.

ab initio protein structure predictions. *J. Comput. Chem.*, 22(3):339–353, 2001.

[3] R. Bonneau and D. Baker. Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. and Biomolec. Struct.*, 30(1):173–189, 2001.

[4] R. Bonneau and D. Baker. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.*, 322(1):65–78, 2002.

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

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

[7] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo. Protein structure determination from NMR chemical shifts. *Proc. Natl. Acad. Sci. USA*, 104(23):9615–9620, 2007.

[8] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc. Natl. Acad. Sci. USA*, 106(10):3734–3739, 2009.

[9] J. DeBartolo, G. Hocky, M. Wilde, J. Xu, K. F. Freed, and T. R. Sosnick. Protein structure prediction enhanced with evolutionary diversity: SPEED. *Protein Sci.*, 19(3):520–534, 2010.

[10] H. Gong, P. J. Fleming, and G. D. Rose. Building native protein conformations from highly approximate backbone torsion angles. *Proc. Natl. Acad. Sci. USA*, 102(45):16227–16232, 2005.

[11] K. F. Han and D. Baker. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl. Acad. Sci. USA*, 93(12):5814–5818, 1996.

[12] L. Kinch, Y. S., Q. Cong, H. Cheng, Y. Liao, and N. V. Grishin. CASP9 assessment of free modeling target predictions. *Proteins: Struct. Funct. Bioinf.*, Suppl(10):59–73, 2011.

[13] H. Kurniawati and D. Hsu. Workspace-based connectivity oracle: An adaptive sampling strategy for PRM planning. In *WAFR*, volume 47 of *Springer Tracts in Advanced Robotics*, pages 35–51. New York, NY, 2006.

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

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

[16] B. S. Olson, K. Molloy, S.-F. Hendi, and A. Shehu. Guiding search in the protein conformational space with structural profiles. *J. Bioinf. and Comp. Biol.*, 10(3):1242005, 2012.

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

[18] E. Plaku, L. Kavraki, and M. Vardi. Discrete search leading continuous exploration for kinodynamic motion planning. In *Robotics: Sci. and Syst.*, Atlanta, GA, USA, 2007.

[19] S. Raman, D. Baker, B. Qian, and R. C. Walker. Advances in rosetta protein structure prediction on massively parallel systems. *IMB J Res & Dev*, 52(1-2):7–17, 2008.

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

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

[22] A. Shehu, L. E. Kavraki, and C. Clementi. Unfolding the fold of cyclic cysteine-rich peptides. *Protein Sci.*, 17(3):482–493, 2008.

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

[24] 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. Natl. Acad. Sci. USA*, 105(12):4685–4690, 2008.

[25] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.

[26] M. Stilman and J. J. Kuffner. Planning among movable obstacles with artificial constraints. *Int. J. Robot. Res.*, 12(12):1295–1307, 2008.

[27] A. W. Stumpff-Kane and M. Feig. A correlation-based method for the enhancement of scoring functions on funnel-shaped energy landscapes. *Proteins: Struct. Funct. Bioinf.*, 63(1):155–164, 2006.

[28] J. P. van den Berg and M. H. Overmars. Using workspace information as a guide to non-uniform sampling in probabilistic roadmap planners. *Int. J. Robot. Res.*, 24(12):1055–1071, 2005.

[29] A. Verma, A. Schug, K. H. Lee, and W. Wenzel. Basin hopping simulations for all-atom protein folding. *J. Chem. Phys.*, 124(4):044515, 2006.

[30] K. Wolff, M. Vendruscolo, and M. Porto. Efficient identification of near-native conformations in ab initio protein structure prediction using structural profiles. *Proteins: Structure*, Jan 2010.

[31] Y. Yang and O. Brock. Efficient motion planning based on disassembly. In *Robotics: Sci. and Syst.*, pages 97–104, Cambridge, MA, 2005.

[32] Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

[33] Y. Zhang and J. Skolnick. Spicker: A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry*, 25(6):865–871, 2004.