Decoy Database Improvement for Protein Folding[†]

(*submitted to the Computational Structural Bioinformatics Workshop (CSBW) special issue)*

Hsin-Yi (Cindy) Yeh*
Parasol Lab
Dept. of Comp. Sci. & Engr.
Texas A&M University
College Station, TX, USA
hyeh@cse.tamu.edu

Aaron Lindsey
Parasol Lab
Dept. of Comp. Sci. & Engr.
Texas A&M University
College Station, TX, USA
alindsey@cse.tamu.edu

Chih-Peng Wu
Parasol Lab
Dept. of Comp. Sci. & Engr.
Texas A&M University
College Station, TX, USA
chinuy@cse.tamu.edu

Shawna Thomas
Parasol Lab
Dept. of Comp. Sci. & Engr.
Texas A&M University
College Station, TX, USA
sthomas@cse.tamu.edu

Nancy M. Amato
Parasol Lab
Dept. of Comp. Sci. & Engr.
Texas A&M University
College Station, TX, USA
amato@cse.tamu.edu

Keywords: decoy databases, protein folding, sampling methods

Abstract

Predicting protein structures and simulating protein folding are two of the most important problems in computational biology today. Simulation methods rely on a scoring function to distinguish the native structure (the most energetically stable) from non-native structures. Decoy databases are collections of non-native structures used to test and verify these functions.

[*] Corresponding author.

We present a method to evaluate and improve the quality of decoy databases by adding novel structures and removing redundant structures. We test our approach on 20 different decoy databases of varying size and type and show significant improvement across a variety of metrics. We also test our improved databases on two popular modern scoring functions and show that for most cases they contain a greater or equal number of native-like structures than the original databases, thereby producing a more rigorous database for testing scoring functions.

1 Introduction

Two important problems in computational biology are predicting protein structures and simulating protein folding motions. The protein's most energetically stable structure, the native structure, determines its function and how it interacts with other proteins. Because a protein's structure and function are so intimately related, predicting a protein's structure is of paramount importance. In addition, errors in the protein folding process (i.e., folding from an unstructured chain of amino acids to the native structure) cause a protein to fold incorrectly thereby altering its functional ability and possibly lead to many devastating diseases. Thus, the folding process itself remains an important area of study.

Many computational tools have been developed to study these problems because they are either too difficult or too expensive to tackle experimentally. Protein structure prediction [Sternberg, 1996] is a widely studied area. One notable method is Rosetta [Rohl et al., 2004] which uses a simplified model to predict the low-resolution protein structure. In response to increased research in protein structure prediction, the CASP [Moult et al., 1995] competition emerged as a platform to test structure prediction methods. Molecular dynamics [Levitt, 1983] and Monte Carlo simulations [Covell, 1992] have been widely used to simulate protein motion. All of these methods rely on a scoring function, typically an energy function. A scoring function attempts to distinguish between native and non-native structures, ranking them in terms of their energetic feasibility. Thus, the accuracy of these methods largely depends on the accuracy of the scoring function used.

Decoy databases have been developed to test and verify these scoring functions [Samudrala and Levitt, 2008]. A decoy is a computer-generated protein structure that is similar to the native structure. Decoys test the ability of a scoring function to identify the protein's native structure from a set of incorrect protein structures. If the scoring function can correctly identify the native structure, the function is said to be correct. Such tests where decoys attempt to "fool" the scoring function are commonly used to test protein folding algorithms. Thus, if we can create higher quality decoy databases, we can improve protein folding algorithms by improving the scoring functions they rely on.

Many large decoy databases for specific proteins have already been compiled for the purpose of testing and improving scoring functions [Samudrala and Levitt, 2008, Moult et al., 1995, Tsai et al., 2003]. However, there is not currently a good way to take these existing databases and improve them so that they are more effective at testing modern scoring functions. Here, we strive to generate higher quality decoy sets in order to more rigorously test these functions.

Contribution. This work presents a method for evaluating the quality of decoy databases and improving them by adding novel structures and removing redundant structures. Our specific contributions are as follows:

- We test on 20 different decoy databases and show that we are able to generate higher quality decoy databases across a variety of metrics.

- We find that most improvement stems from the addition of structures by our sampling methods.

- We test our improved databases on two popular modern scoring functions, QMEAN [Benkert et al., 2008] and RWplus potential [Zhang and Zhang, 2010], and show that for most cases they contain a greater or equal number of structures ranked more native-like than the actual native state than the original databases.

2 Related Work

In this section we discuss related work in the areas of protein models, scoring functions, and existing decoy databases. We also discuss existing methods for sampling conformations as we will use these to add conformations to existing decoy sets.

2.1 Protein Models

A protein is composed of a chain of amino acids that determine its function. Amino acids are distinguished by their side chain. When hydrogen bonds form between atoms on the protein backbone, secondary structures can develop. α helices and β sheets make up the majority of secondary structures.

The most accurate protein model is the all-atoms model. However, in many cases the all-atoms model is too computationally expensive, particularly for larger proteins. Therefore, some coarse-grained models [Rader and Bahar, 2004, Amato and Song, 2002, Go, 1983] have been developed to ease the computational complexity by ignoring some detail information. For example, the Gaussian Network Model (GNM) [Rader and Bahar, 2004] models amino acids as beads connected by elastic strings. Lattice models [Go, 1983] constrain the protein as a rigid lattice and each amino acid is represented as a bead on the lattice.

Another coarser-grained approach models a protein as a series of ϕ and ψ torsional angles. All other bond angles and all bond lengths remain fixed. This is a common modeling assumption as bond lengths and angles typically only undergo small fluctuations [Sternberg, 1996]. In this $\phi - \psi$ model, a protein conformation with n amino acids has $2n$ degrees of freedom. Side chains are modeled as spheres with zero degrees of freedom located at the C_β position. This model has been successfully used to simulate the correct order of large folding events for several small proteins [Amato and Song, 2002].

Regardless of the model, a protein's potential energy landscape is defined as the set of all conformations and their associated potential energies. Figure 1 shows a theoretical energy landscape where all possible conformations are represented by the $x - y$ plane. The native structure of a protein is thought to be located at the lowest point on a funnel-shaped landscape. The landscape may contain several local minima which are distinct from the native energy basin.

2.2 Energy Functions

The protein's atoms interact with each other and with the surrounding solvent through bonds and non-bond interactions such as electrostatic interaction and van der Waals forces. A potential energy function determines conformation validity by taking into account these different atom interactions.

Generally, potential functions that compute all pairwise interactions are called all-atoms functions, e.g., CHARMM [Brooks et al., 1983] and AMBER [Weiner and Kollman, 1981]. These are the most accurate since they consider all possible interactions. However, they are computationally expensive and infeasible for many large proteins.

Instead of modeling all possible interactions, coarse-grained functions only consider side chain contributions to approximate the potential energy. In the $\phi - \psi$ model, each side chain is modeled as a sphere located at the C_β atom. If two side chains are too close (i.e., less than 2.4 Å, the conventional van der Waals contact distance), the conformation is rejected [Levitt, 1983]. Otherwise, the energy may be calculated as:

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + E_{hp}, \tag{1}$$

where K_d is 100 kJ/mol, d_i is the distance of a hydrogen or disulphide bond in native structure, and $d_0 = d_c = 2$Å [Levitt, 1983]. E_{hp} represents the hydrophobic effect. This energy function has been shown to produce high-level folding simulation results similar to an all-atoms function in a fraction of the time [Song et al., 2003]. It is the energy function used here in these results.

2.3 Decoy Databases

Decoys are computer-generated protein structures. Decoy databases have been used to improve the accuracy of scoring functions [Handl et al., 2009, Subramani et al., 2009]. A scoring function is the component of a protein folding algorithm that distinguishes between native and non-native structures. Thus, the performance of the algorithm is dependent on the accuracy of the scoring function. Decoy databases attempt to "fool" a scoring function into choosing a non-native structure as the native. Some existing decoy databases include (i) the Decoys 'R' Us set [Samudrala and Levitt, 2008], (ii) the Rosetta set [Tsai et al., 2003], and (iii) the Critical Assessment of Protein Structure Prediction (casp) set [Moult et al., 1995].

The Decoys 'R' Us set contains three subsets: the single decoy set, the multiple decoy set, and the loop decoy set. The single decoy set only contains the native structure and one decoy structure. The purpose of this set is to test whether a scoring function can distinguish between these two structures. The multiple decoy set and the loop decoy set each contain many decoy structures, and they are both used to verify that a scoring function can select a conformation with low RMSD to the native structure.

The Rosetta set is generated by the Rosetta protein structure prediction method developed in the Baker Laboratory. It can generate low-resolution structures by adding side chains and making structure adjustments [Bonneau et al., 2001].

casp is a protein structure prediction competition held every other year. Competition submissions are collected as a decoy database. Participants use their own approaches to predict the three-dimensional structure of the given amino acid sequence. In order to evaluate the results, the distances between the C_α positions in the predicted model and the target structure are calculated and a score is assigned showing how similar the prediction is compared to the target [Zemla, 2003].

Some work has been proposed to improve protein decoy sets. For example, the Rosetta set has been improved by adding back the side chains and running the structures through an energy minimizer [Tsai et al., 2003]. Other work uses a library of short fragments to generate protein decoys by assembling them together given the protein's geometric constraints [Kolodny and Levitt, 2003]. Most assembled proteins are 6Å from the native structure. Fragments of varying lengths are used in [Molloy and Shehu, 2012] to refine near-native protein decoy structures. While this multi-level approach produces decoy structures closer to the native structure, this method is dependent on the quality of the input fragments.

2.4 Sampling Conformations

Algorithms in the field of motion planning and robotics use sampling-based methods to generate valid robot configurations. Some examples include the Probabilistic Roadmap (PRM) method [Kavraki et al., 1996] and Rapidly-Exploring Random Trees (RRT) [LaValle and Kuffner, 1999]. Both of these strategies rely on a sampling method to find valid configurations for a robot in its environment.

In the context of protein folding, sampling methods generate protein conformations by setting a value for each degree of freedom in the protein model. Thus, for the $\phi - \psi$ protein model used in this work, a conformation q is generated by assigning a value to each ϕ and ψ angle. The conformation q is accepted based on its potential energy $E(q)$ with the following probability:

$$P(\text{accept } q) = \begin{cases} 1 & \text{if } E(q) < E_{min} \\ \frac{E_{max} - E(q)}{E_{max} - E_{min}} & \text{if } E_{min} \leq E(q) \leq E_{max} \\ 0 & \text{if } E(q) > E_{max} \end{cases}. \tag{2}$$

Values for each degree of freedom may be generated in a variety of ways. For example, they can be obtained from molecular dynamics or Monte Carlo simulations, global optimization techniques such as basin-hopping [Wales and Doye, 1997], or robotics-inspired search methods [Amato et al., 2003, Shehu et al., 2009, Cortés and Al-Bluwi, 2012].

3 Methods

In this section we describe our approach to evaluating and improving the quality of decoy databases. We first discuss how to evaluate a decoy set using various metrics. We then present two types of improvement operations: adding novel structures to the set and removing redundant structures from the set.

3.1 Decoy Set Evaluation

Because our methods improve existing decoy sets, we first develop strategies for analyzing the quality of decoy sets. These are used later to show what advantages the improved set provides over the original. We present several quantitative metrics to compare decoy sets and describe how their values are calculated in the experiments.

Z-Score. The z-score (or standard score) indicates the number of standard deviations between the native structure energy and the average energy of a decoy set [Tsai et al., 2003]. Researchers frequently use the z-score to determine the likelihood that a scoring function would pick the native structure from the other structures in the set. The z-score of a decoy set D is:

$$\textsc{ZScore}(D) = \frac{\mathrm{E}(D.\mathrm{native}) - \mathrm{E}_{avg}(D)}{\mathrm{E}_{std}(D)} \tag{3}$$

where $\mathrm{E}(d)$ is the energy of a structure d, $\mathrm{E}_{avg}(D)$ is the average energy of D, and $\mathrm{E}_{std}(D)$ is the standard deviation of the energies in D.

A positive z-score means the native structure has a higher energy than the average energy of the set, and a negative z-score means the native structure energy is lower than the average energy. A z-score of zero indicates the native structure energy and the average decoy energy are exactly the same. A desirable decoy set has structures with low energies close to the native structure. Thus, we would like to see the z-score approach zero after improvement which indicates that it contains structures with similar energies to the native.

Improvement Score. Given an original decoy set D and an improved decoy set D', the improvement score returns the change in z-score per sample between the two sets. The improvement score between D and D' is:

$$\textsc{Improvement}(D, D') = \frac{\textsc{ZScore}(D')}{|D'|} - \frac{\textsc{ZScore}(D)}{|D|} \tag{4}$$

Higher values indicate greater changes in z-score. Note that this score is most meaningful when the two sets are roughly equal in size which is the case in the results presented here, see Section 4.

Minimum Distance. The minimum distance metric measures the average minimum distance from each decoy structure to any other decoy structure in the set. In other words, it is the average distance of each structure to its closest neighbor measured by some distance metric δ.

This metric measures the diversity of structures in the set. As the minimum distance increases, the diversity of structures included in the set also increases. Possible distance functions include Euclidean distance in $\phi - \psi$-space and CαRMSD. In this work, we use Euclidean distance.

3.2 Decoy Set Improvement

There are two main phases in the improvement of decoy sets. First, samples are generated on the protein's energy landscape. This set may be generated in a variety of ways and is discussed in further detail below in Section 3.2.1. In the decoy selection phase, some structures are chosen from the original set D to be removed and some are chosen from the sample set S to be added. Decoy selection is discussed below in Section 3.2.2.

3.2.1 Sample Set Generation

To improve decoy sets by adding structures, we must first generate a set of samples from which to select. We can use one or more of the methods discussed below to generate structures and retain the energetically feasible ones as given by Equation 2.

Sampling Methods. Many different sampling methods exist. In this paper, we combine the following methods in a hybrid sampler:

- *Uniform Sampling.* This is the simplest strategy. It returns a structure at a random point on the energy landscape by simply selecting values for each ϕ and ψ angle uniformly at random. This will generate many unwanted high-energy structures but provides good coverage of the landscape. However, it does not provide dense coverage around more interesting regions, e.g., the native energy basin or local minima. Thus, it should be used in conjunction with other sampling methods.

- *Sampling with Native Bias.* Iterative Gaussian sampling [Amato et al., 2003] biases samples by iteratively perturbing the native state. It samples torsional angles from a set of normal distributions centered around previously sampled structures, starting with the native state. This approach has been successfully applied to simulate the folding process on larger proteins. It generates many low energy samples, but they are usually confined to the native energy basin. Figure 2 illustrates this process.

- *Biased Sampling from Low-Energy Decoys.* Instead of starting from the native structure as iterative Gaussian sampling [Amato et al., 2003], this approach begins the iteration from the decoy structures with the lowest energy. To our knowledge, this is a novel approach to generating low-energy structures. As with native bias sampling, perturbations are selected from a set of normal distributions. Here, generated structures have low energies and are not confined to the native energy basin. However, it typically produces samples near the energy basins of selected decoys.

Any sampling method can be incorporated in the hybrid sampler. In our experiment, we combine the above three methods to form a hybrid sampler that exploits the strengths of each. Such a sampler first adds the native structure to the set of seeds as in iterative Gaussian sampling [Amato et al., 2003]. For the remaining seeds in the set, it selects half from the lowest energy decoy structures and half from uniform sampling. This ensures that there are plenty of low-energy structures in the final set that are located throughout the energy landscape in many different local minima. Such structures are important to include because they are most likely to confuse a scoring function.

Calculating Sample Set Size. For each sample set, we must specify n, the number of sample structures to generate. We would like to have an adequate sample set size which can efficiently provide high quality decoy candidates. After some preliminary experiments monitoring how n affects the z-score rate of change, we found that doubling the original set size provides informative structures efficiently.

3.2.2 Decoy Selection

Given an existing decoy set D and a set of sample structures S, we would like to add viable structures from S to D and remove redundant structures from D. To select such structures, we apply a filter to each one, keeping those structures that pass the filter. In this work, we investigate the following filters:

- *Energy Filter.* This filter chooses all structures whose energy is less than some threshold. For the results in this work, we use the energy function in [Song et al., 2003].
- *Minimum Distance Filter.* This filter selects structures whose distance to their closest neighbor as determined by some distance metric δ is greater than some threshold. Here we use Euclidean distance.

For each filter, we must specify a threshold or cutoff value that decides which samples to keep and which to discard. We set the filter's threshold to be one standard deviation above (for the energy filter) or below (for the minimum distance filter) the average value of the set.

4 Results and Discussion

We apply our methods to existing decoy sets and show that they are able to generate sets with lower energies and more diverse structures that are more likely to "fool" protein folding scoring functions. All decoy sets were obtained from the existing Decoy 'R' Us databases [Samudrala and Levitt, 2008], CASP8 [Moult et al., 2009], CASP9 [Moult et al., 2011], and CASP10 [Moult et al., 2014] and are listed in Table 1. We study various types of proteins (e.g., α, β, and α/β mixed) including larger proteins (e.g., 4fle with 192 residues) and larger decoy sets (e.g., 1eh2 with 2413 conformations). All results are averaged over 10 runs.

4.1 Decoy Selection

The original decoy set D and the sample set S can be broken down into four subsets:

- redundant decoy structures D_D from D,
- viable decoy structures D_V from D,
- redundant sampled structures S_D from S, and
- viable sampled structures S_V from S.

Figure 3 shows the relationship of these four subsets. The final improved decoy set is $D_V \cup S_V$ and is a combination of two operations: removing redundant structures (yielding D_V) and adding new structures (yielding S_V).

Table 1 provides the resulting set sizes after improvement. To generate S, we combine uniform sampling, sampling with native bias, and decoy-biased sampling into a hybrid sampler containing all three methods. We then apply both the energy filter (using the coarse-grained energy function from [Song et al., 2003]) and the minimum distance filter (using Euclidean distance in $\phi - \psi$-space) to S and to the original set D to create the set of viable structures $D_V \cup S_V$. For all proteins, the resulting size is comparable to the original.

Figure 4 summarizes the resulting z-score, improvement score, and minimum distance value for each protein. For each metric, we show the contribution from each operation (removing redundant decoys (D_V) and adding new samples ($D \cup S_V$)) and from their combination ($D_V \cup S_V$).

When the z-score approaches zero, the native structure energy is harder to distinguish among the other structures in the set. For every protein in Figure 4(a), the z-scores of D and D_V are very similar. Hence, simply removing structures does not greatly impact z-score. Once we add new structures from our sampling approach ($D \cup S_V$), the z-score drops drastically with scores comparable to the final set ($D_V \cup S_V$). Thus, the main contributors to z-score improvement are the structures generated by our sampling approach.

Recall that the improvement score shows the change in z-score per sample between two sets. A higher value indicates that the change (either structure addition, removal, or both) has a greater impact on the z-score. Figure 4(b) displays the improvement scores across all tested proteins. We again see that adding structures provides a decoy set with better quality than simply removing redundant ones. Proteins 1ash and 1gdm with the smallest original sets show the largest improvement scores. Since the original set sizes are small, removing structures causes significant decrease in the improvement scores.

The last metric we examine is the minimum distance between neighboring structures which indicates set diversity. A larger distance signifies greater structural diversity and implies a greater ability to “fool” different scoring functions. Figure 4(c) shows how this metric changes for each operation. As expected, when decoys are removed (D_V), the minimum distance increases, and when adding decoys ($D \cup S_V$), the minimum distance decreases. For all proteins studied, the minimum distance is not affected significantly by adding decoys ($D \cup S_V$) implying that they are informative structures.

Table 1 also shows the contributions from each type of sample (e.g., from the original decoy set, from uniform sampling, etc.) to the final improved decoy set. In the final improved decoy set, most samples come from the original set (D_V) and the remaining largely come from native bias sampling. Uniform sampling and decoy bias sampling play a very small role because they typically fail to pass the filters employed in the decoy selection phase, either because their energies are too high or they are too similar to structures already in the final set. Although uniform sampling doesn't contribute in the improved decoy set, it generates samples efficiently and covers the energy landscape evenly. We believe uniform sampling is the baseline method to be included in the hybrid sampler. Decoy bias sampling is considered as a strong approach to cover the local minimum regions in the energy landscape. It provides some good candidates to improve the decoy set quality for 1sn3, 4fr9, 1r69, and 2k4x. Decoy bias sampling performs pretty well particularly in the 2k4x case which contributes 10.67% of samples in the final improved set. This is something interesting that we would like to study the sample locations in the energy landscape further in the future.

Figures 5 and 6 show the potential vs. CαRMSD from the native structure distribution for each protein. In each plot, three subsets are displayed: D_V, D_D and S_V. We see that in general, the added structures have lower energies and cover a wider range of CαRMSD than the original structures. While most added structures are located near the native structure, there are also some structures with high CαRMSD. These structures are especially valuable because they could be located in a local minima of the energy landscape and may be more likely to "fool" scoring functions. Decoy bias sampling was able to generate samples that passed the filters for 1sn3, 4fr9, 1r69, and 2k4x; these are highlighted with brown boxes.

4.2 Improved Decoy Sets in Practice

Here we assess the ability of our improved decoy sets to "fool" modern scoring functions. We study two different scoring functions: Qualitative Model Energy ANalysis (QMEAN) and

RWplus potential. QMEAN [Benkert et al., 2008] is a composite scoring function incorporating several different structural descriptors including local geometry features for discriminating native-like torsional angles from others, secondary structure features for long-range interactions, burial status, and solvent accessibility. RWplus potential [Zhang and Zhang, 2010] is a distance-dependent atomic statistical potential which uses an ideal random walk chain as a reference state and incorporates orientation-dependent energy arising from side chains. RWplus potential favors capturing short-range interactions and has been shown to accurately identify native-like structural features. QMEAN and RWplus potential were selected for study because they showed statistically significant improvement over other well-established scoring functions and are both publicly available.

Table 2 compares the number of structures QMEAN and RWplus potential ranked higher than the native state between the original decoy dataset and our improved decoy dataset, respectively. We used the QMEAN webserver [Benkert et al., 2009] and the RWplus potential executable from the Zhang lab website [Zhang,] to generate rankings.

In 8 out of 20 proteins studied, our improved decoys sets were able to produce more structures that "fooled" at least one of the scoring functions than the original set, sometimes finding a large number of new structures as for 1eh2. RWplus potential was less likely to be "fooled" than QMEAN as a more decoy sets contained a greater number of structures ranked higher than the native in the improved set over the original set for QMEAN. This is a bit different from what was found in [Zhang and Zhang, 2010] where QMEAN and RWplus potential had comparative performance. QMEAN is a composite potential combining various resources, and RWplus potential claims to be strong at recognizing strong signals at short-range interactions. We suspect that local secondary structure changes in our sampling set may be one factor causing RWplus potential to outperform QMEAN here.

It is interesting that both scoring functions misrank a large number of structures for 1eh2. We note that decoy set improvement performance is less for the three β proteins studied than for the other proteins in the set. This may be attributed to the Euclidean distance filter not being sensitive enough to all-β structures. Consider that two β structures (or near β structures) may be close in $\phi - \psi$ space but have large RMSD as they do not contain the tight contact pattern typical of α helices to constrain them. Thus, the Euclidean distance filter may mistakenly discard a structure with large RMSD but small Euclidean distance. Instead, a distance filter based on RMSD may be more appropriate for β proteins. (This is a subject of future investigation.) However, even on sophisticated, modern scoring functions, our improved decoy sets are able to indicate areas of weakness for several of the proteins studied.

To understand why the improved decoy set produces more structures with higher rankings than the native structure, we look at the structural differences between the original and improved decoy sets. As an example, Figure 7 shows the superimposition between the native structure (displayed in green) and the highest QMEAN ranked decoy (displayed in blue) for 1sn3. 1sn3 is one example where QMEAN is “fooled” but RWplus potential is not. Even though the decoy misses a β sheet (left, circled in red) and has a shorter β sheet (middle), QMEAN scores it higher and would incorrectly select this decoy as the native structure. Since the local secondary structures have been changed, RWplus potential is able to capture the difference and recognize it as a non-native structure.

5 Conclusion

We describe a new method for evaluating and improving the quality of decoy databases. Our method removes redundant structures and generates new low energy structures in varied locations on the energy landscape resulting in higher quality decoy sets that are more likely to "fool" the scoring functions of modern protein folding algorithms. We tested our approach on 20 different decoy databases of varying size and type and showed significant improvement over the original set. Interestingly, most of the improvement came from adding structures not originally covered by the set indicating a capacity to "fool" more scoring functions. We also show that our improved databases produced a greater number of structures ranked more native-like by two popular modern scoring functions than the original databases for many of the proteins studied. In the future, we plan to implement a web service to improve user-submitted decoy databases. Our hope is that others can use these improved databases to develop better protein folding algorithms and more accurate folding simulations.

6 Acknowledgments

This work is supported in part by NSF awards CRI-0551685, CCF-0833199, CCF-0830753, IIS-096053, IIS-0917266 by THECB NHARP award 000512-0097-2009, by Chevron, IBM, Intel, Oracle/Sun and by Award KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). A preliminary version of this work appeared in [Lindsey et al., 2014].

7 Author Disclosure Statement

No competing financial interests exist.

References

[Amato et al., 2003] Amato, N. M., Dill, K. A., and Song, G. (2003). Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, 10(3-4):239–255. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.

[Amato and Song, 2002] Amato, N. M. and Song, G. (2002). Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2):149–168. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.

[Benkert et al., 2009] Benkert, P., Künzli, M., and Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic Acids Res.*, 37:W510–W514.

[Benkert et al., 2008] Benkert, P., Tosatto, S. C. E., and Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, 71:261–277.

[Bonneau et al., 2001] Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C. E. M., and Baker, D. (2001). Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins Struct. Funct. Bioinf.*, Suppl 5:119–126.

[Brooks et al., 1983] Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S., and Karplus, M. (1983). Charmm: a program for macromolecular energy, minimization and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217. http://yuri.harvard.edu/.

[Chan and Dill, 1998] Chan, H. S. and Dill, K. A. (1998). Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins: Structure, Function, and Bioinformatics*, 30(1):2–33.

[Cortés and Al-Bluwi, 2012] Cortés, J. and Al-Bluwi, I. (2012). A robotics approach to enhance conformational sampling of proteins. In *Proc. ASME Mech. and Rob. Conf.*

[Covell, 1992] Covell, D. G. (1992). Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Bioinf.*, 14(3):409–420.

[DeLano, 2002] DeLano, W. (2002). The pymol molecular graphics system (2002). *DeLano Scientific, Palo Alto, CA, USA.*

[Go, 1983] Go, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.*, 12:183–210.

[Handl et al., 2009] Handl, J., Knowles, J., and Lovell, S. C. (2009). Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics*, 25(10):1271–1279.

[Kavraki et al., 1996] Kavraki, L. E., Švestka, P., Latombe, J. C., and Overmars, M. H. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580.

[Kolodny and Levitt, 2003] Kolodny, R. and Levitt, M. (2003). Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*, 68(3):278–285.

[LaValle and Kuffner, 1999] LaValle, S. M. and Kuffner, J. J. (1999). Randomized kinodynamic planning. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 473–479.

[Levitt, 1983] Levitt, M. (1983). Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, 170:723–764.

[Lindsey et al., 2014] Lindsey, A., Yeh, H.-Y. C., Wu, C.-P., Thomas, S., and Amato, N. M. (2014). Improving decoy databases for protein folding algorithms. In *Proc. ACM Conf. Bioinform., Comp. Bio., and Health Inform.: Comp. Struct. Bioinform. Wkshp. (CSBW)*, pages 717–724. ACM.

[Molloy and Shehu, 2012] Molloy, K. and Shehu, A. (2012). Biased decoy sampling to aid the selection of near-native protein conformations. In *BCB*, pages 131–138. ACM.

[Moult et al., 2009] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., and Tramontano, A. (2009). Critical assessment of methods of protein structure predictionround viii. *Proteins Struct. Funct. Bioinf.*, 77(S9):1–4.

[Moult et al., 2014] Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure predictionround x. *Proteins Struct. Funct. Bioinf.*, 82(S2):1–6.

[Moult et al., 2011] Moult, J., Fidelis, K., Kryshtafovych, A., and Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (casp)round ix. *Proteins Struct. Funct. Bioinf.*, 79(S10):1–5.

[Moult et al., 1995] Moult, J., Pedersen, J. T., Judson, R., and Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinf.*, 23(3):ii–iv.

[Rader and Bahar, 2004] Rader, A. and Bahar, I. (2004). Folding core predictions from network models of proteins. *Polymer*, 45:659–668.

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

[Samudrala and Levitt, 2008] Samudrala, R. and Levitt, M. (2008). Decoys 'R' Us: A database of incorrect conformations to improve protein structure prediction. *Protein Sci.*, 9(7):1399–1401.

[Shehu et al., 2009] Shehu, A., Kavraki, L. E., and Clementi, C. (2009). Multiscale characterization of protein conformational ensembles. *Proteins*, 76(4):837–851.

[Song et al., 2003] Song, G., Thomas, S., Dill, K., Scholtz, J., and Amato, N. (2003). A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. In *Proc. Pacific Symposium of Biocomputing (PSB)*, pages 240–251.

[Sternberg, 1996] Sternberg, M. J. (1996). *Protein Structure Prediction.* OIRL Press at Oxford University Press.

[Subramani et al., 2009] Subramani, A., DiMaggio, P. A., and Floudas, C. A. (2009). Selecting high quality protein structures from diverse conformational ensembles. *Biophys. J.*, 97(6):1728–1736.

[Tsai et al., 2003] Tsai, J., Bonneau, R., Morozov, A. V., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins Struct. Funct. Bioinf.*, 53(1):76–87.

[Wales and Doye, 1997] Wales, D. J. and Doye, J. P. K. (1997). Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A*, 101(28):5111–5116.

[Weiner and Kollman, 1981] Weiner, P. and Kollman, P. (1981). Amber: Assisted model building with energy renement, a general program for modeling molecules and their interactions. *J. Comp. Chem.*, 2:287–303.

[Zemla, 2003] Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, 31(13):3370–3374.

[Zhang and Zhang, 2010] Zhang, J. and Zhang, Y. (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One*, 5(10):1–13.

[Zhang,] Zhang, Y. RW potential. `http://zhanglab.ccmb.med.umich.edu/RW/`. Accessed: 2014-08-13.

Table 1: Decoy sets studied and final improved set sizes with distribution breakdown from Decoys 'R' Us [Samudrala and Levitt, 2008], CASP8 [Moult et al., 2009], CASP9 [Moult et al., 2011], and CASP10 [Moult et al., 2014]. D is the original decoy set, D_D is the set of deleted structures from D, and D_V is the set of viable (retained) structures from D. S is the set of sampled structures and S_V is the set of viable (retained) structures from S.

					Improved			% Samples in $D_V \cup S_V$			
				Original	Size		D_D/D	Uniform	Native	Decoy	D_V
Type	Protein	Residue	Source	Size	Avg.	Std.	(%)		Bias	Bias	D_V
α/β	1fca	55	Decoys 'R' Us [Samudrala and Levitt, 2008]	2001	2024.90	21.19	12.26	0.00	13.30	0.00	86.70
	4pti	58	Decoys 'R' Us [Samudrala and Levitt, 2008]	334	361.80	23.12	19.28	0.00	25.48	0.00	74.52
	1igd	61	Decoys 'R' Us [Samudrala and Levitt, 2008]	501	512.30	9.53	9.52	0.00	11.52	0.00	88.48
	1sn3	65	Decoys 'R' Us [Samudrala and Levitt, 2008]	660	630.50	4.03	11.53	0.00	7.38	0.02	92.61
	1ctf	68	Decoys 'R' Us [Samudrala and Levitt, 2008]	630	604.50	6.25	10.67	0.00	6.90	0.00	93.10
	4icb	76	Decoys 'R' Us [Samudrala and Levitt, 2008]	500	579.70	8.01	5.50	0.00	18.49	0.00	81.51
	1eh2	79	Decoys 'R' Us [Samudrala and Levitt, 2008]	2413	2546.40	13.88	8.65	0.00	13.43	0.00	86.57
	4fr9	141	CASP10 [Moult et al., 2014]	406	496.90	4.30	6.16	0.00	23.28	0.04	76.68
	4gb5	148	CASP10 [Moult et al., 2014]	217	228.90	4.89	0.00	0.00	5.20	0.00	94.80
	4f54	184	CASP10 [Moult et al., 2014]	322	310.90	3.67	11.49	0.00	8.33	0.00	91.67
	4fle	192	CASP10 [Moult et al., 2014]	182	183.40	0.66	0.33	0.00	1.09	0.00	98.91
α	1r69	63	Decoys 'R' Us [Samudrala and Levitt, 2008]	676	744.70	9.59	11.64	0.00	18.61	1.18	80.21
	2cro	65	Decoys 'R' Us [Samudrala and Levitt, 2008]	501	619.20	9.11	7.29	0.00	24.98	0.00	75.02
	1nkl	78	Decoys 'R' Us [Samudrala and Levitt, 2008]	1995	2293.80	21.41	10.63	0.00	22.27	0.00	77.73
	1jwe	114	Decoys 'R' Us [Samudrala and Levitt, 2008]	1407	1452.40	29.27	12.14	0.00	14.89	0.00	85.1
	1ash	147	Decoys 'R' Us [Samudrala and Levitt, 2008]	30	36.00	1.41	3.33	0.00	19.44	0.00	80.56
	lgdm	153	Decoys 'R' Us [Samudrala and Levitt, 2008]	30	33.20	1.72	6.67	0.00	15.66	0.00	84.34
β	2k4x	55	CASP8 [Moult et al., 2009]	461	744.10	16.34	12.04	0.00	34.90	10.67	54.50
	2kyw	79	CASP9 [Moult et al., 2011]	458	447.10	9.19	11.33	0.00	9.17	0.00	90.83
	3mx7	90	CASP9 [Moult et al., 2011]	499	494.80	4.26	10.36	0.00	9.60	0.00	90.40

Table 2: Comparison of the number of structures ranked higher than the native state by QMEAN [Benkert et al., 2008] and RWplus potential [Zhang and Zhang, 2010].

Type	Protein	# Structures Ranked Higher than Native by QMEAN		# Structures Ranked Higher than Native by RWplus potential	
		Original	Improved	Original	Improved
α/β	1fca	0	8	0	0
	4pti	0	0	0	0
	1igd	0	0	0	0
	1sn3	0	10	0	0
	1ctf	0	2	0	0
	4icb	0	0	0	0
	1eh2	0	45	1	43
	4fr9	0	0	26	26
	4gb5	0	0	0	0
	4f54	0	1	0	0
	4fle	0	0	1	0
α	1r69	0	0	0	0
	2cro	0	0	0	0
	1nkl	0	3	0	0
	1jwe	7	13	0	0
	1ash	0	0	0	0
	1gdm	0	0	0	0
β	2k4x	93	77	1	0
	2kyw	7	0	9	0
	3mx7	2	3	9	6

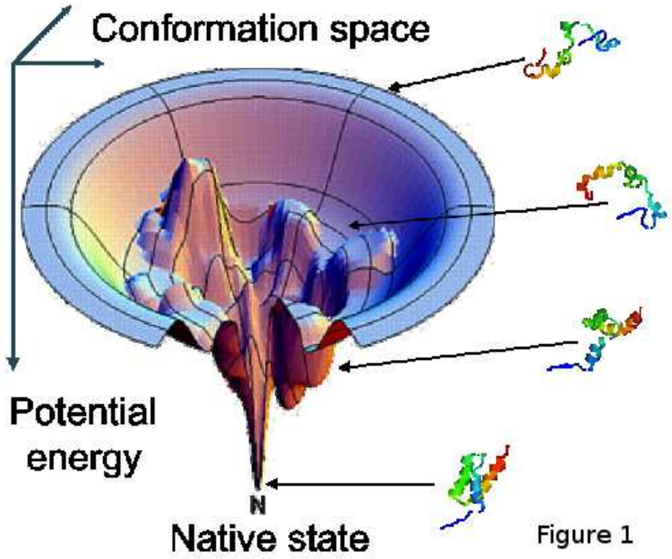
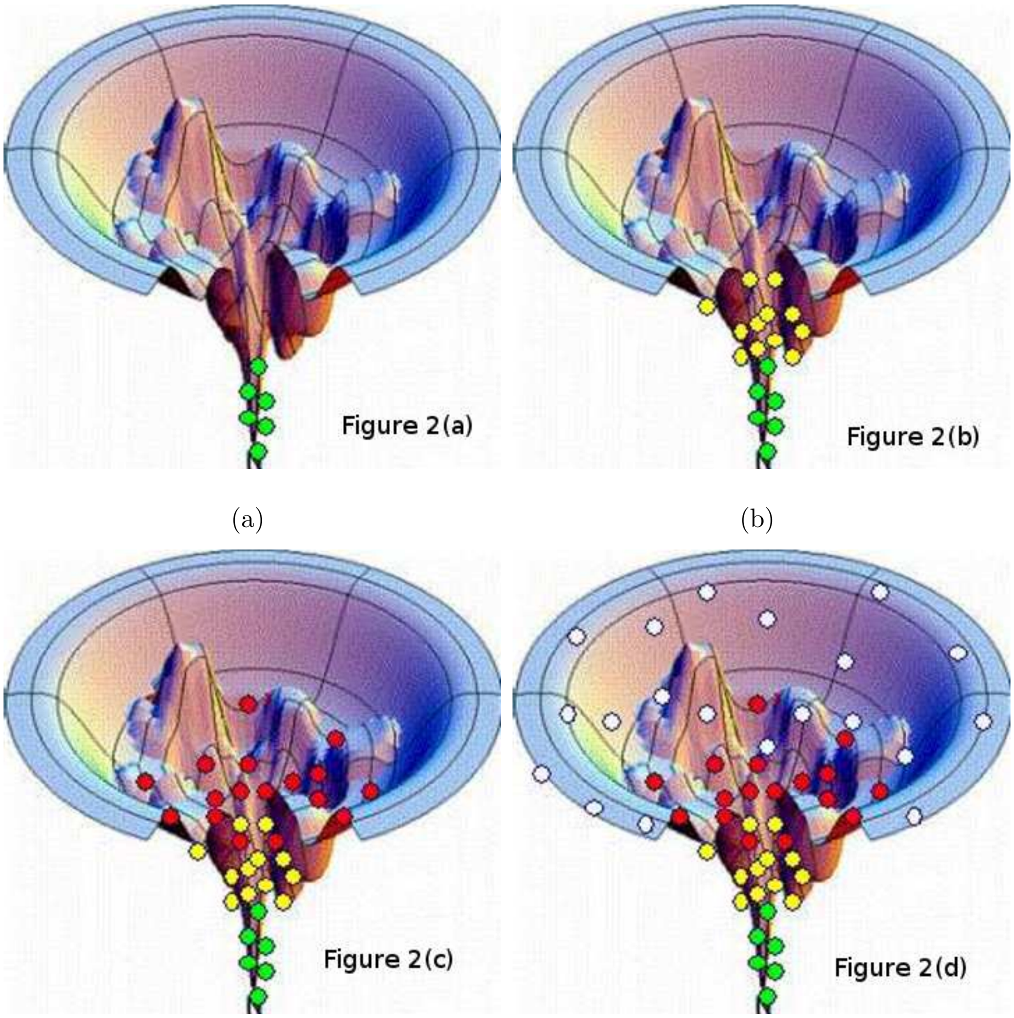


Figure 1: The potential energy landscape of a protein is the set of all conformations and their associated potential energies [Chan and Dill, 1998]. The conformation space of a protein can span hundreds of degrees of freedom; it is not limited to only the $x - y$ plane.



(a)

(b)

(c)

(d)

Figure 2: Iterative Gaussian sampling gives a denser distribution near the native structure at the bottom of the energy landscape funnel and a sparser distribution as they grow outward.

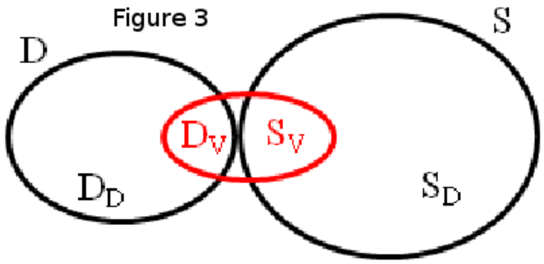
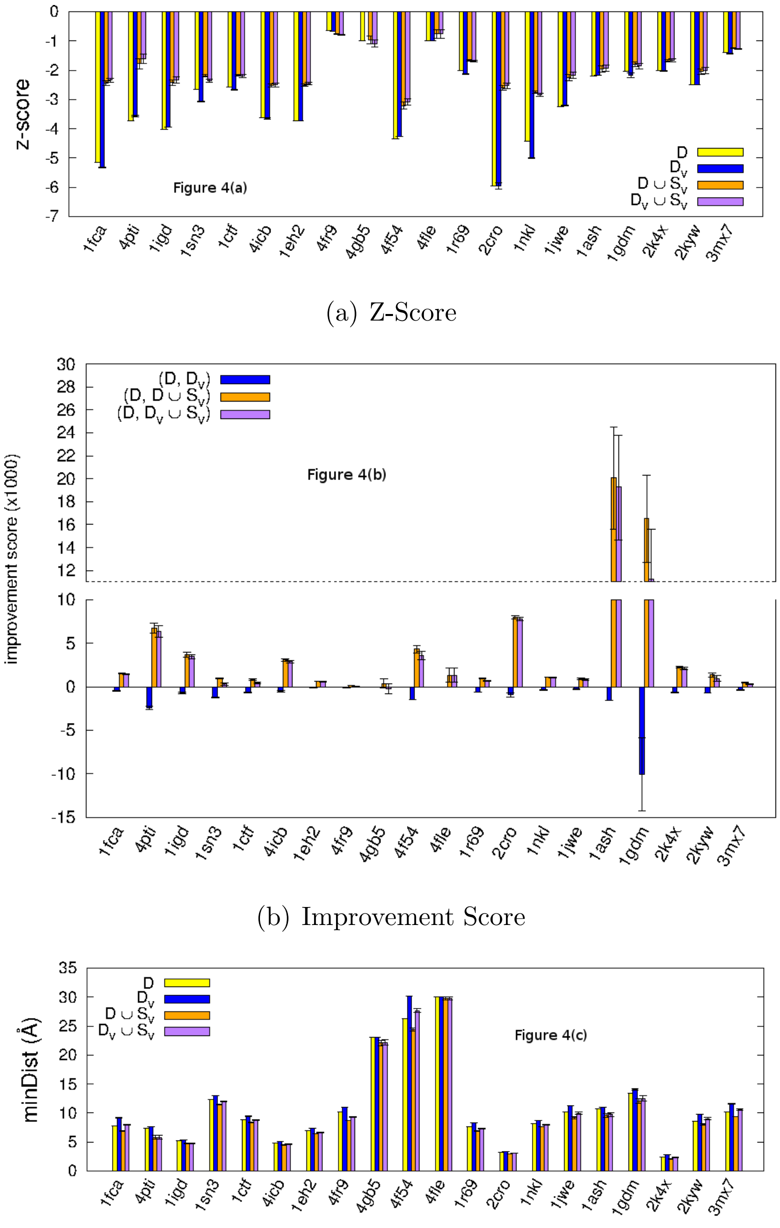


Figure 3: The relationship between each of the subsets D_D, D_V, S_D and S_V from the original decoy set D and the sample set S. The final set, in red, is $D_V \cup S_V$.

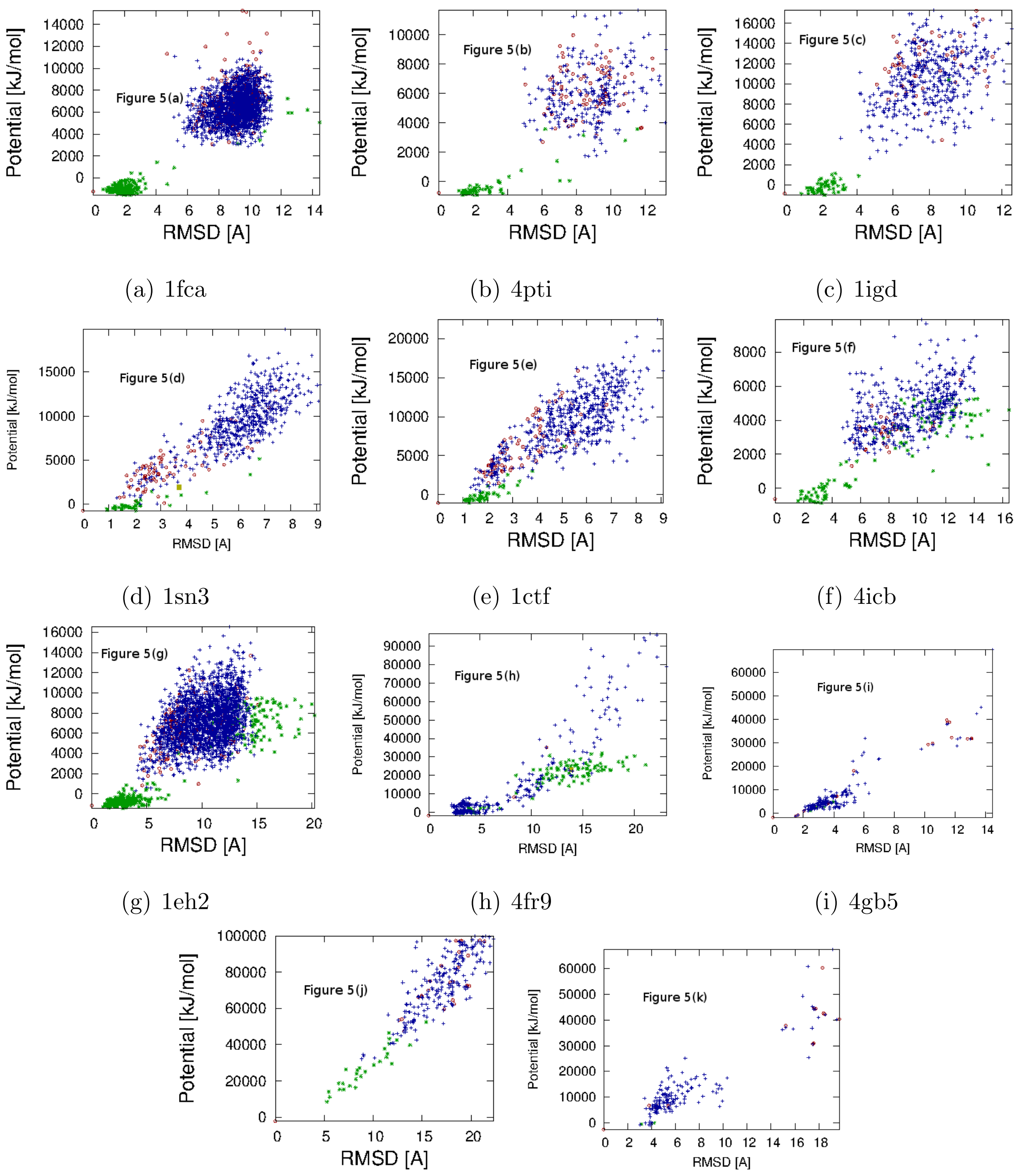


(a) Z-Score

(b) Improvement Score

(c) Minimum Distance

Figure 4: Resulting metrics of improved decoy sets and their subsets, where D is the original set, D_V is after redundant structures are removed, and S_V is the set of sampled structures to be added.



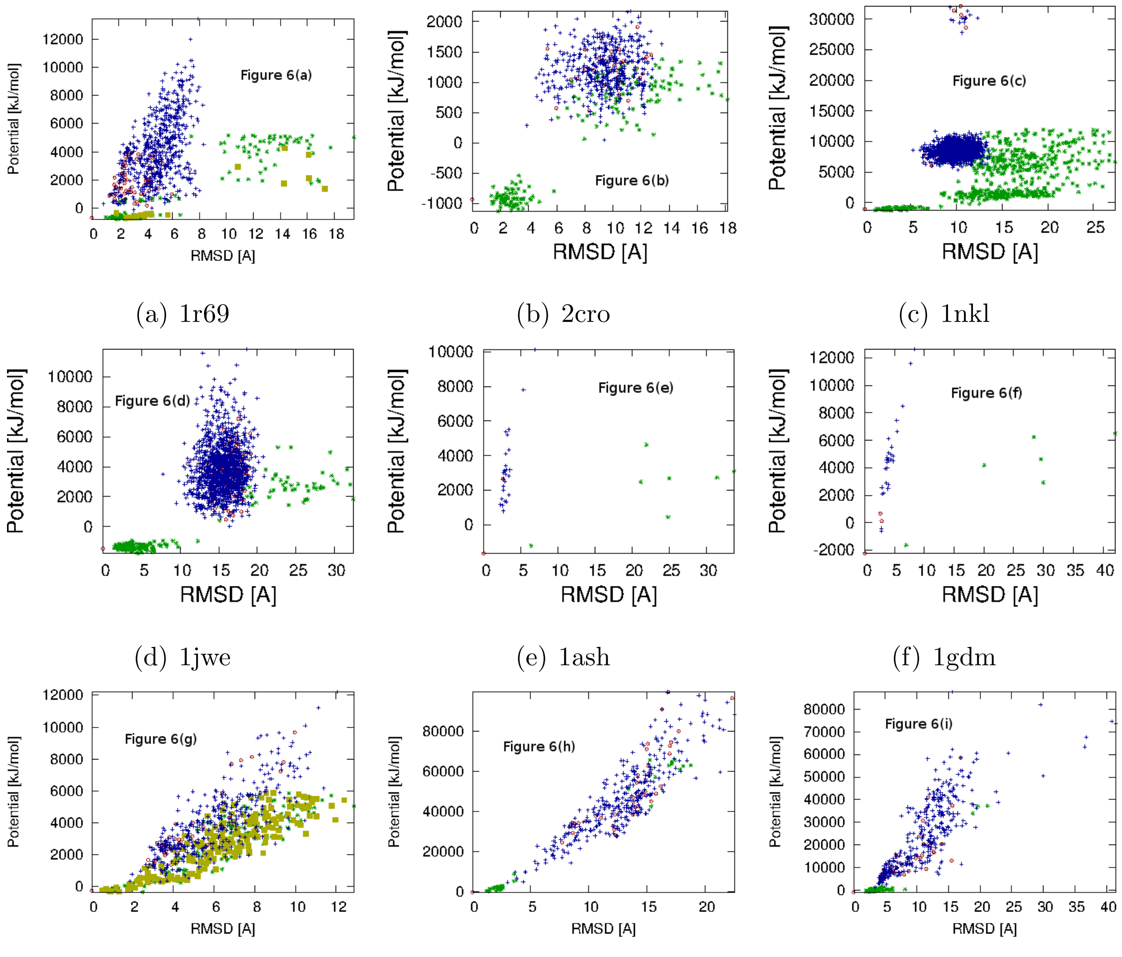
(a) 1fca (b) 4pti (c) 1igd

(d) 1sn3 (e) 1ctf (f) 4icb

(g) 1eh2 (h) 4fr9 (i) 4gb5

(j) 4f54 (k) 4fle

Figure 5: Potential vs. CαRMSD for D_D (red circles), D_V (blue '+'s), and S_V (green '*'s) for α/β mixed proteins. Decoy bias sampling was able to generate samples that passed the filters for 1sn3 and 4fr9; these are highlighed with brown boxes.



(a) 1r69

(b) 2cro

(c) 1nkl

(d) 1jwe

(e) 1ash

(f) 1gdm

(g) 2k4x

(h) 2kyw

(i) 3mx7

Figure 6: Potential vs. CαRMSD for D_D (red circles), D_V (blue '+'s), and S_V (green '*'s) for α and β proteins. Decoy bias sampling was able to generate samples that passed the filters for 1r69 and 2k4x; these are highlighted with brown boxes.

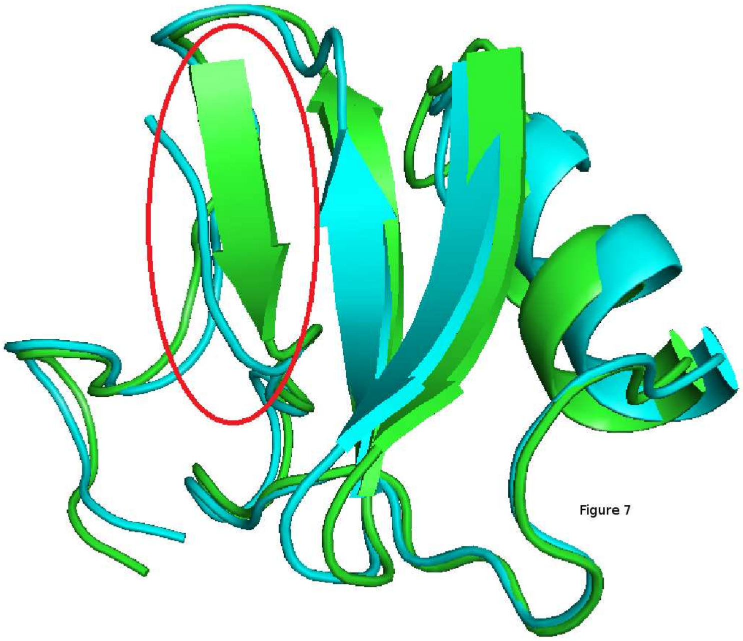


Figure 7: Superimposition of the native structure (shown in green) and the decoy with the highest QMEAN score (shown in blue) for 1sn3 by PyMOL [DeLano, 2002]. The decoy misses a piece of secondary structure, circled in red.

Response Document

1 Summary Changes

We thank the reviewers for providing meaningful and insightful review that helped in making the paper better. Following is the response to each reviewer's comments.

2 Reviewer 1

1. Are these (three measurements) standard in decoy improvement articles?
   Response: We have three evaluation metrics in the paper: z-score, improvement score, and minimum distance. Z-score is some standard metric people always used and is developed in [2]. Improvement score and minimum distance are the metrics we developed in this work which is also one of our contribution.

2. While low energy structures are generated, they mostly come from adjustments to the native structure, whereas the sampling scheme seems not able to identify lower-energy structures from adjustments of existing decoys.
   Response: Decoy bias sampling uses the low-energy decoy structure as the seed to generate the samples which is able to cover some local minimum region in the energy landscape. And it does contribute some samples in the final improved set for few proteins, e.g., 1sn3, 4fr9, 1r69, and 2k4x.

3. It is not clear why the authors choose their own sampling scheme as opposed to other perhaps established methods that maybe can generate lower-energy structures. What is the reason for doing so?
   Response: Our decoy bias sampling method is inspired by the native bias sampling method which has been shown that is able to cover the energy landscape well and also can correctly simulate the protein folding pathways. Instead of using the native structure as the seed, the decoy bias sampling uses some low-energy decoy structure as the seed and tries to perturb the torsional angles and generate new samples from it.

4. The energy in reference 27 seems old.
   Response: This energy function is shown to be able to correctly simulate the protein folding pathway in [1] and it is fairly simple to calculate.

3 Reviewer 2

1. How did the authors evaluate the performance of the minimum distance metric?
   Response: The minimum distance metric calculates the average distance of each structure to its closest neighbor measured by some distance metric. We use Euclidean distance in our experiment.

2. How is Euclidean distance defined in this case?
   Response: The Euclidean distance is calculated in the space of torsional angles between protein backbone.

3. The sentence in page 14 starting with "RWplus potential was less likely..." is not clear.
   Response: We find that QMEAN picks the wrong native structure more times than RWplus potential. Therefore, RWplus potential is less likely to be "fooled" than QMEAN in our experiment.

4. According to Table 1, the uniform sampling did not contribute anything to the decoy set and the decoy bias sampling only slightly improved the set in two out of the 20 cases. Please justify the use of these methods if they hardly contribute.
   Response: The reason we include uniform sampling in our hybrid sampling approach is that uniform sampling is served as the baseline method. Although uniform sampling can not provide good decoy candidates, it still can generate samples very efficiently. Decoy bias sampling uses the low-energy decoy structure as the seed to generate the samples which is able to cover some local minimum region in the energy landscape. And it does contribute some samples in the final improved set for few proteins, e.g., 1sn3, 4fr9, 1r69, and 2k4x.

5. What does MinDist stand for in Figure 4? Is that the Euclidean distance described above?
   Response: MinDist stands for the minimum distance which measures the average distance of each structure to its closest neighbor measured by some distance metric. In this work, we use Euclidean distance.

6. What is the justification for using a very simplified representation for the protein representation?
   Response: It has been shown that the coarse-grained model we use in this work is able to correctly simulate the protein folding process although it only considers the ϕ and ψ angles of the protein backbone. A coarse-grained protein model basically helps us to simplify the problem and further improve the computational time while maintaining the accuracy.

4 Reviewer 3

1. Wouldn't it be better if the z-score is as negative as possible?
   Response: The z-score indicates the number of standard deviations between the native structure energy and the average energy of a decoy set. As our goal here is to "fool" the scoring function, the ideal decoy set should have its average energy almost the same as the native structure. The smaller the standard deviations between the native structure energy and the average energy of a decoy set, the higher possibility that a scoring function can be "fooled". Therefore, we would like to see the z-score approach zero after improvement.

2. Why is the improvement score normalized by the set of the decoy sets?
   Response: Normalization helps us to understand the change in z-score per sample between two sets. As the two sets are roughly equal sized, this score is meaningful.

3. The authors write that they use Euclidean distance over angles. I don't understand why the author wouldn't account for the wraparound at 2*π.
   Response: The Euclidean distance is computed in the space of the torsional angles between protein backbone. The angle is actually modeled in the range from 0 to 2π.

4. Please include the number of runs used and perhaps also the standard deviations of the values reported.
   Response: The results reported in the paper are averaged over 10 runs which is stated in the first paragraph in the Results and Discussion section.

References

[1] G. Song, S. L. Thomas, K. A. Dill, J. M. Scholtz, and N. M. Amato. A Path Planning-based Study of Protein Folding With a Case Study of Hairpin Formation in Protein G and L. *Proc. Pac. Symp. of Biocomputing (PSB)*, 240–251, 2003

[2] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins Struct. Funct. Bioinf.*, 53(1):76–87, 2003.