Using Sequence-Predicted Contacts to Guide Template-free Protein Structure Prediction

Ahmed Bin Zaman
Dept of Computer Science
George Mason University
azaman6@gmu.edu

Prasanna Venkatesh Parthasarathy
Dept of Computer Science
George Mason University
ppartha@gmu.edu

Amarda Shehu*
Dept of Computer Science
George Mason University
amarda@gmu.edu

ABSTRACT

The primary challenge in template-free protein structure prediction is controlling the quality of computed tertiary structures, also known as decoys. While research on how to do so is highly active, the main rule of thumb is to generate as many decoys as can be afforded. This rule acknowledges that more decoys increase the likelihood that some will reside near the sought biologically-active/native structure. Generating large numbers of decoys imposes time and space costs. These costs percolate down to decoy selection algorithms that need to select from the generated decoys a few sufficiently near-native. In this paper, we evaluate the hypothesis that the generated decoy ensemble can be significantly reduced without sacrificing decoy quality. Evaluation on diverse proteins shows that drastic reductions can be achieved in the number of preserved decoys while retaining the quality of generated decoys via clustering. The presented results suggest that decoy ensemble reduction promises to aid protein structure prediction.

CCS CONCEPTS

• **Applied computing → Molecular structural biology**; **Bioinformatics**;

KEYWORDS

template-free protein structure prediction; decoy generation; decoy ensemble reduction; clustering algorithms.

ACM Reference Format:
Ahmed Bin Zaman, Prasanna Venkatesh Parthasarathy, and Amarda Shehu. 2019. Using Sequence-Predicted Contacts to Guide Template-free Protein Structure Prediction. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19), September 7–10, 2019, Niagara Falls, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3307339.3342175

*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6666-3/19/09…$15.00
https://doi.org/10.1145/3307339.3342175

1 INTRODUCTION

While it is now well-recognized that the three-dimensional/tertiary structure of a protein is key to determining its array of activities in the cell [5], protein structure determination (PSP) poses many challenges [13]. Increasingly faster and cheaper high-throughput gene sequencing technologies have yielded millions of protein-encoding gene sequences that are now stored in genomic databases [4]. In contrast, as of June 2019, the number of known native structures determined in wet laboratories and deposited in the Protein Data Bank (PDB) [3] is 152, 500. This discrepancy continues to motivate computational research in PSP. Increasingly, the focus is on template-free PSP, where target protein sequences with no known structures do not have sufficiently-similar protein sequences with known structures that could otherwise serve as templates [11].

Template-free PSP is carried out in two stages. In the first stage, the focus is on decoy generation; the tertiary structures are referred to as decoys to highlight the fact that it is unclear which ones are sufficiently close to the sought native structure. While decoy generation algorithms effectively address an optimization problem where they seek tertiary structures that minimize the interaction energy among the atoms of a given target protein, one cannot infer that a lower-energy decoy is more similar to the sought native structure; the energy functions designed in computational laboratories are inherently inaccurate [1, 6, 15]. The focus of the second stage, known as decoy selection, is to tease out the decoys that are near-native among the many generated in the first stage.

Good advances have been made in decoy generation [10, 16–18, 23–26]. These advances are documented in the Critical Assessment of protein Structure Prediction (CASP), which is a biennial community experiment/competition that assesses progress in PSP in several categories, including the template-free/free modeling category [9]. In most state-of-the-art decoy generation algorithms, the amino-acid units that comprise a protein are represented at a reduced level of detail. The representation of choice models explicitly only the backbone atoms and either a centroid pseudo-atom or the beta carbon for the side-chain atoms of each amino acid. The actual variables are typically dihedral angles on bonds connecting the explicitly-modeled atoms. The energy functions that evaluate decoys operate over Cartesian coordinates of modeled atoms due to their dependence on (pairwise) interatomic distances. The reader is referred to the review [11] for more details.

Despite clever choices in variable selection, the search space explored by a decoy generation algorithm is vast and high-dimensional; e.g., on a protein not exceeding 100 amino acids, modeling three backbone dihedral angles per amino acid results in around 300 variables; the search space has 300 dimensions. In addition, the

energy function is noisy. These challenges make it exceptionally difficult to control the quality of generated decoys, and algorithmic research on how to achieve this under the umbrella of stochastic optimization is highly active [21].

Users of popular platforms, such as Rosetta and Quark, are advised to generate as many decoys as can be afforded. More decoys mean higher likelihood that some will reside near the sought native structure. This recommendation is impractical. While generating decoys used to be significantly more expensive than analyzing them, now this relationship is less imbalanced. Great progress in software and hardware has made it less costly to generate decoys. Algorithms operating under the umbrella of evolutionary computation can generate hundreds of thousands of decoys [17, 24–26]. Decoy selection algorithms tasked with analyzing decoys now may have to additionally deal with a data size issue.

In addition, the end of the decoy generation stage adds back the side-chain atoms on each decoy and carries out local improvements on the resulting all-atom decoys prior to handing them off to the decoy selection stage. Adding atomistic detail is computationally expensive, as the energy function employed has to handle a large number of atoms per decoy (that includes all side-chain atoms and all hydrogen atoms per amino acid). This is exacerbated when the recommendation is to collect large numbers of decoys.

The focus of this paper is on the interaction between the first and second stage and asks a fundamental question: Can the ensemble of generated decoys be reduced, thus lowering the computational burden on refinement and selection, all the while without sacrificing decoy quality? This paper addresses this question via clustering. To the best of our knowledge, there is little work on reducing decoy ensembles. A common approach is to discard higher-energy decoys, which requires setting a threshold. The pitfall is that similar structures can have different energies. The threshold would also have to be determined on a per-target basis. Some related attempts employ Principal Component Analysis to reduce the dimensionality of the structure space [19, 20]. Employing dimensionality reduction techniques may be useful for visualization of decoy ensembles, but it is not exactly clear how they would apply to ensemble reduction.

Here we propose a clustering-based approach. Our focus is not on comparing clustering algorithms but on evaluating representative ones for their utility in reducing decoy ensembles. A rigorous analysis is conducted to determine optimal settings for employed clustering algorithms. Evaluations on protein datasets that include CASP targets show drastic reductions in ensemble size while retaining decoy quality. The presented results suggest that research on decoy ensemble reduction is a promising direction to aid template-free PSP and can generally be useful in reducing molecular structure data. The rest of this paper is organized as follows. We describe the proposed methodology in Section 2. Evaluation is presented in Section 3. The paper concludes in Section 4.

2 METHODS

From now on, we will employ the terms for generated ensemble Ω_{gen} and reduced ensemble Ω_{red}. The generated ensemble Ω_{gen} consists of decoys generated from a decoy generation algorithm. The reduced ensemble Ω_{red} is the objective of the methodology we propose in this paper; Ω_{red} retains only a fraction of the decoys in the original Ω_{gen}. To efficiently produce a reduced-size decoy ensemble Ω_{red} that retains the quality of the original, full-size decoy ensemble Ω_{gen}, the proposed methodology leverages fast shape similarity of the decoys. It consists of three stages: (i) A *featurizer* extracts decoy features that summarize the shape of a decoy. (ii) These features are utilized by a clustering algorithm to group generated decoys based on their shape similarity. (iii) Finally, a *selector* populates the reduced ensemble Ω_{red} with selected decoys from each cluster/group identified over Ω_{gen}.

2.1 Decoy Ensemble Generation

While the focus of this work is not on decoy generation algorithms, it is worth summarizing the algorithm employed to generate the ensemble Ω_{gen} over which we evaluate our objective of reducing it while retaining quality. We make use of the hybrid evolutionary algorithm (HEA) that has been previously published [17] and evaluated against Rosetta and other algorithms [16–18]. HEA carries out a biased exploration of the space of backbone dihedral angles, using the same representation as in the Rosetta decoy generation algorithm. As an evolutionary algorithm, HEA evolves a population of decoys towards lower-energy regions of the Rosetta energy surface (evaluated with a Rosetta coarse-grained energy function) while delaying premature convergence via evolutionary search strategies. The interested reader is referred to Ref. [17] for more details on the HEA. We note that any decoy generation algorithm can be used to generate a decoy ensemble for our purposes. Specifically, we employ HEA to generate hundreds of thousands of decoys for a target protein (given its amino-acid sequence).

2.2 Featurizer: Extraction of Shape-based Features of Generated Decoys

We leverage the Ultrafast Shape Recognition (USR) metrics that were originally introduced in [2] to summarize three-dimensional molecular shapes. The purpose of the USR metrics was to speed up searches for similar structures in molecular structure databases. We employ such metrics as features by which to encode a generated decoy. USR metrics are momenta of distance distributions (of atoms) from four chosen reference points in a tertiary structure. These reference points are the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom from ctd (fct), and the farthest atom from fct (ftf). For each reference point, the distance of each atom is computed, and the resulting distribution is summarized with three momenta, the mean, the variance, and the skewness. In this way, each decoy in Ω_{gen} is summarized by 12 features.

We note that the motivation for encoding each decoy via features is two-fold. First, reducing the number of coordinates needed to represent each decoy reduces the computational time needed by any algorithm expected to process generated decoys in some fashion. Second, it is well-known that clustering algorithms, which we utilize in the second stage of our approach, are less effective in high-dimensional spaces [8, 12, 22].

2.3 Clustering: Grouping Generated Decoys by Shape Similarity

With each decoy in Ω_{gen} represented by 12 features as described above, clustering algorithms are now applied to group the decoys

into clusters. We choose two popular clustering algorithms, hierarchical and k-means. The choice reflects the fact that these algorithms can handle large data.

Our application of k-means is as follows. We consider two hyperparameters, the decoys that can serve as cluster centroids and the number of clusters k. For a given value of k, we initially select k decoys uniformly at random over Ω_{gen} to serve as the centroids. For this particular grouping C of the dataset into k clusters, one can now measure the loss function via the within-cluster scatter: $L(C) = \frac{1}{2}\sum_{l=1}^{k}\sum_{i\in C_l}\sum_{j\in C_l, j\neq i} D(x_i, x_j)$, where $D(x_i, x_j)$ measures the Euclidean distance between two points/decoys $x_i \neq x_j$ in the same cluster C_l, where $l \in \{1, \ldots, k\}$. One can vary the decoys serving as cluster centroids and among the different options choose the ones that result in the smallest loss. For a given k, we do so 10 times, each time selecting decoys at random to serve as centroids, and retaining the assignment resulting in smallest loss.

The optimal number of clusters, k, is determined via the popular knee-finding approach. Specifically, for a given k, after the centroids of clusters are determined as above, the squared distance of each decoy in a cluster from the centroid of the cluster can be recorded, and the sum of these squared distances can be obtained over the clusters k. This sum of squared distances, also often referred to as the sum of squared errors (SSE), can be plotted for different values of k, as shown in Fig. 1 on a particular target protein. The knee (also referred to as elbow) in the curve indicates the optimal number of clusters. We want a small SSE. Naturally, the SSE approaches 0 as one increases k; it is exactly 0 when $k = |\Omega_{\text{gen}}|$. The goal is to choose a small value of k that results in a low SSE. The knee or the elbow in the curve that tracks SSE as a function of k corresponds to the region where increasing k starts yielding diminishing returns.

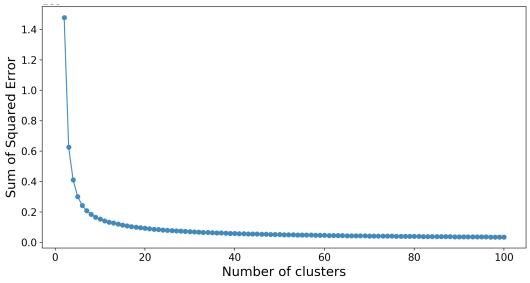

Figure 1: The SSE is plotted as a function of the number of clusters identified via k-means on decoys generated via HEA on a target protein. This target is part of our evaluation dataset that we relate in Section 3. Specifically, it is the target protein with known native structure in the PDB entry with identifier (id) 1ail.

Hierarchical clustering does not require *a priori* specifying the number of clusters. We note that hierarchical clustering refers to a family of clustering algorithms that build nested clusters by merging or splitting them successively. We make use of the merging/agglomerative approach, where every decoy starts as its own cluster, and clusters are successively merged together until the root is reached; the root is the unique cluster that contains all the decoys. There are different linkage criteria depending on the metric used for the merge strategy. We use single linkage, which sets the distance between two clusters as the distance between the two closest points across the clusters.

The hierarchy of clusters is represented as a tree/dendrogram. "Cutting" at different places in the tree corresponds to selecting a particular partition of the dataset into clusters. We utilize a cached implementation, so that cutting at different places in the hierarchical tree does not require recomputation of the clusters. In order to determine where to cut the tree, we employ the Davies-Bouldin (DB) index [7], which is a popular clustering validation technique in the absence of ground truth labels that is computed on features inherent to the dataset. A lower DB index relates to better separation between the clusters. Specifically, the DB index is defined as the average similarity between each cluster and its most similar one; it evaluates intra-cluster similarity and inter-cluster differences to provide a non-negative score. In our application of hierarchical agglomerative clustering with single linkage, the DB index is evaluated at every height of the tree, and the height that results in the smallest DB is the one that is selected as the optimal partition (and optimal corresponding number of clusters).

2.4 Selector: Populating the Reduced Ensemble

After the application of a clustering algorithm over Ω_{gen}, the decoys in it are grouped/partitioned into clusters. The *selector* selects a subset of decoys from each cluster to populate the reduced ensemble Ω_{red}. Specifically, the selector makes use of both the identified clusters and the Rosetta *score*4 energy function. This function evaluates not only Lennard-Jones interactions, but also short-range and long-range hydrogen bonding in a decoy. The decoys in a cluster are organized into levels/bins. The decoys in a bin are those with score4 energies that are identical up to two digits after the decimal sign. One decoy is selected at random from each bin and placed in the reduced ensemble Ω_{red}. This process is repeated for each identified cluster. This approach indirectly biases the contribution to the reduced ensemble by cluster size. Larger clusters have more decoys in them, which results in more energy levels for those clusters, and so more decoys selected from larger clusters for placement in the reduced ensemble. As a result, this approach indirectly biases towards diversity, as well, though we note that the width of a bin/level can be modified/tuned to control the size of the reduced ensemble.

3 RESULTS

Evaluation is carried out on two datasets. The first is a benchmark dataset of 10 proteins of varying lengths and folds that are used widely for evaluation [17, 25–27]. The second contains 10 hard, free-modeling target domains from CASP12 and CASP13. On each target, HEA is run 5 times to account for its stochasticity. Decoys generated over these runs are collected, resulting in a Ω_{gen} ensemble of 250,000 decoys for each target. As described in Section 2, Ω_{gen} is subjected to k-means or hierarchical clustering to obtain a reduced ensemble Ω_{red}. The number of clusters obtained from each algorithm with the SSE- or DB-guided process detailed in Section 2 varies per target. Fig. 2(a) shows the distribution of the number of clusters determined by the DB index-based approach for the hierarchical clustering algorithm over all target proteins (in the combined benchmark and CASP dataset). Fig. 2(b) shows

the distribution of the number of clusters determined by the SSE-guided approach for k-means over all target proteins. Visualizing these distributions reveals that on the majority of the targets, the number of clusters is between $20-35$, indicating the presence of structure in the decoy dataset (that is, a large number of decoys are similar) that is leveraged in this paper to reduce the generated decoy ensemble while retaining quality.

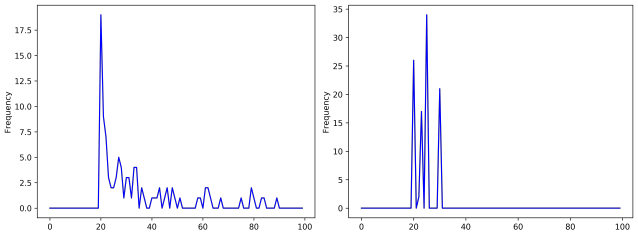

Figure 2: Distribution of the number of clusters identified via (a) hierarchical and (b) k-means clustering over target proteins in the benchmark and CASP datasets.

The Ω_{gen} and Ω_{red} ensembles are compared in terms of size and quality. For the latter, we focus on the proximity of decoys to the known native structure of a target protein. Specifically, we make use of the popular least root-mean-squared-deviation (lRMSD) to measure the dissimilarity of a decoy to a known native structure [14]. lRMSD reports the average over the Euclidean distances of corresponding atoms in two given structures once differences due to rigid-body motions (whole-body translation and rotation in three dimensions) are removed. Our comparison considers the main carbon (C_α) atom of each amino acid.

To compare quality, the minimum and standard deviation of the lRMSDs of each decoy from the native structure in each ensemble is reported. To provide a baseline, the reduced ensemble identified via k-means and hierarchical clustering is compared to a reduced ensemble identified via truncation selection. Given a target size M, the M lowest-energy decoys in Ω_{gen} are selected in the truncation-based approach to populate the reduced ensemble. The target size M is the maximum over the reduced ensembles obtained via hierarchical or k-means clustering. As the results presented below will make clear, k-means yields larger reduced ensembles, so all truncation-based reductions end up matching in size the reduced ensemble obtained via k-means. Finally, a subset of the targets is selected for visualization of the original and reduced ensembles. The visualization plots a decoy in each ensemble via two coordinates, its lRMSD from the known native structures and its Rosetta score4 energy.

3.1 Reduction versus Quality

Fig. 3 compares Ω_{gen} and Ω_{red} in terms of size for the benchmark and CASP datasets. Specifically, Fig. 3 plots the reduction $(1-\frac{|\Omega_{\text{red}}|}{|\Omega_{\text{gen}}|})\cdot 100\%$ obtained by k-means and hierarchical clustering. The reductions obtained by hierarchical clustering are drastic, over 77% on all targets, and over 80% on 9/10 of the targets. Those obtained by k-means range from 54% to 71%. Fig. 3 shows similar results for the CASP dataset, with reductions of 59% and higher obtained via k-means and higher reductions of 80% and higher obtained via hierarchical clustering.

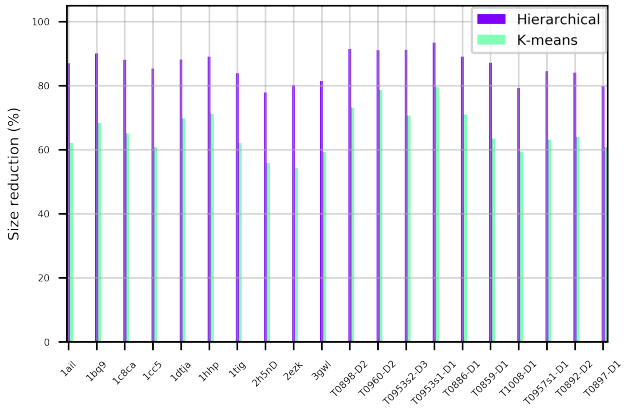

Figure 3: The reduction $(1-\frac{|\Omega_{\text{red}}|}{|\Omega_{\text{gen}}|})\cdot 100\%$ obtained by k-means and hierarchical clustering is shown for each protein in the benchmark and CASP datasets.

Fig. 4 compares Ω_{red} to Ω_{gen} in terms of the minimum (top panel) and standard deviation (bottom panel) of lRSMDs of decoys in each ensemble to the known native structure on each target (of the benchmark and CASP datasets). Specifically, Fig. 4 plots the difference of the minimum or standard deviation in Ω_{red} over the corresponding quantity in Ω_{gen}, comparing hierarchical clustering, k-means, and truncation selection.

Fig. 4 reveals that truncation selection achieves the worst performance; differences in minimum lRMSD range from 0.73Å to 5.12Å. That is, the best decoy retained via truncation selection can be 5.12Å further from the native structure than the best decoy in the original ensemble. It is clear that quality cannot be maintained via truncation selection. In contrast, the differences in minimum lRMSD obtained in the case of k-means are 0Å, and those obtained by hierarchical clustering range from 0Å to 0.29Å. The slight increase in the case of hierarchical clustering is not surprising, given that hierarchical clustering yields more drastic reductions in size over k-means, as related above.

Fig. 4 shows that differences on lRMSD standard deviation in the case of k-means range from 0.02Å to 0.26Å, and those obtained in the case of hierarchical clustering range from 0Å to 0.36Å, with less than 0.1Å on 5/10 targets. Fig. 4 allows making similar observations on the CASP dataset, with quality lost in the reduced ensemble obtained via truncation selection, quality preserved on both the reduced ensembles obtained via the clustering algorithms, with the best results obtained via k-means. On standard deviation, hierarchical and k-means clustering perform comparably.

To provide greater detail, the actual distribution of decoy lRMSDs from the native structure is shown for the Ω_{gen} ensemble of generated decoys, as well as the reduced ensembles Ω_{red} obtained via k-means and hierarchical clustering. Fig. 5 does so for one target protein (with native structure under PDB id 1ail) and shows that the reduced ensembles obtained by each clustering algorithm contain decoys with similar relative frequencies of lRMSD as Ω_{gen}. Altogether, these results allow concluding that, both clustering

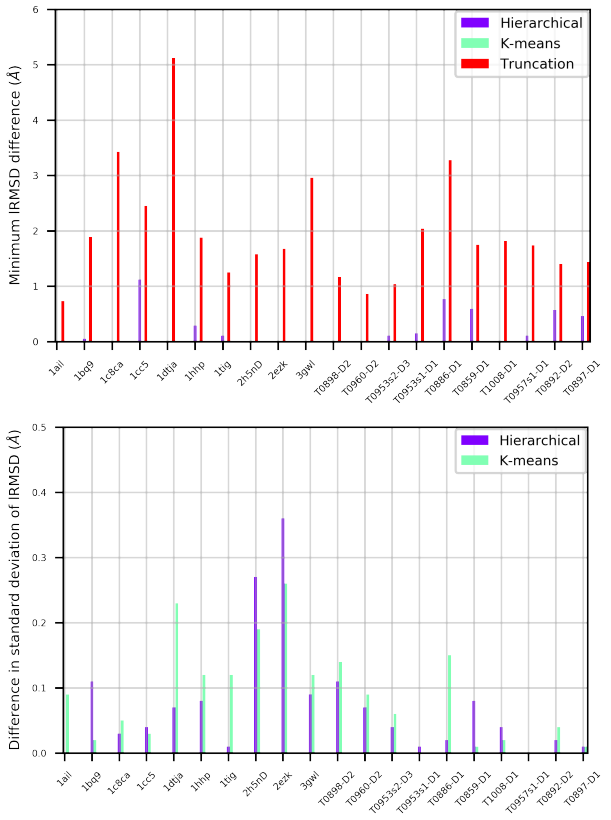

Figure 4: Comparison of the minimum and standard deviation of the distribution of lRMSDs (to the known native structure) of decoys in the Ω_{gen} and Ω_{red} ensembles of each target in the benchmark and CASP datasets. Comparison of minimum lRMSDs includes the ensemble reduced via truncation selection. Differences between the minimum and standard deviation obtained over Ω_{red} from those obtained over Ω_{gen} are related.

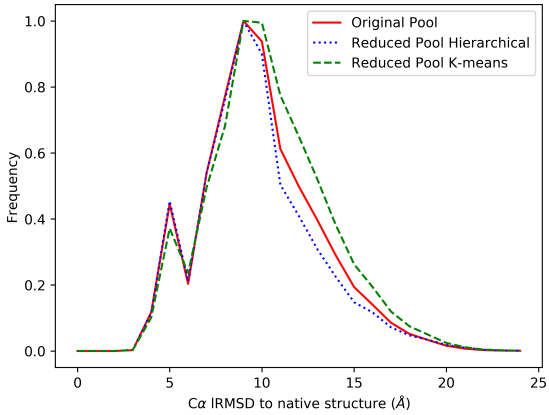

Figure 5: The distribution of decoy lRMSDs from the native structure is shown for the Ω_{gen} ensemble (in red) and the reduced ensembles Ω_{red} obtained via k-means (green) and hierarchical clustering (in blue). Results are shown for the target protein with native structure under PDB id 1ail.

algorithms allow obtaining drastic reductions in the decoy ensemble size while preserving quality, with further reductions in size provided by hierarchical clustering.

3.2 Ensemble Visualization

Fig. 6 (top panel) visualizes the Ω_{gen} and Ω_{red} ensembles for a selected target in the benchmark dataset. Decoys in Ω_{gen} are shown in red, and those in Ω_{red} are superimposed in blue. Fig. 6 (bottom panel) provides a similar visual comparison for a selected target in the CASP dataset. Fig. 6 shows that the reduced ensemble Ω_{red} retains decoys from all the regions in the structure space probed by the original ensemble Ω_{gen}. In particular, k-means does so better than hierarchical clustering (practically all red dots are occluded by the superimposition), which is not surprising, as k-means yields larger reduced ensembles than hierarchical clustering.

4 CONCLUSION

The results presented here suggest that it is possible to significantly reduce the number of generated decoys retained for further analysis without sacrificing quality. A clustering-based approach is shown effective at doing so. Hierarchical clustering is shown more effective at reducing ensemble size. While the presented methodology and evaluation serves as a proof-of-concept, the work presented here opens up many venues of further research. It is naturally appealing to integrate the proposed approach in decoy selection methods that operate under the umbrella of machine learning. In addition, while one can investigate the utility and effectiveness of different features via which to represent decoys, advances in subspace clustering can be leveraged to address the high-dimensionality of molecular structure spaces.

5 ACKNOWLEDGMENTS

This work is supported in part by NSF Grant No. 1763233 and a Jeffress Memorial Trust Award. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University.

REFERENCES

[1] N. Akhter, W. Qiao, and A. Shehu. 2018. An Energy Landscape Treatment of Decoy Selection in Template-free Protein Structure Prediction. *Computation* 6, 2 (2018), 39.
[2] P. J. Ballester and W. G. Richards. 2007. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J Comput Chem* 28, 10 (2007), 1711–1723.
[3] H. M. Berman, K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. 10, 12 (2003), 980–980.
[4] C. E. Blaby-Haas and V. de Crécy-Lagard. 2013. Mining high-throughput experimental data to link gene and function. *Trends Biotechnol* 29, 4 (2013), 174–182.
[5] D. D. Boehr and P. E. Wright. 2008. How do proteins interact? *Science* 320, 5882 (2008), 1429–1430.
[6] R. Das. 2011. Four small puzzles that Rosetta doesn't solve. *PLoS ONE* 6, 5 (2011), e20044.
[7] D. L. Davies and D. W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (1979), 224–227.
[8] C. Domeniconi, D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, and D. Papadopoulos. 2007. Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery* 14, 1 (2007), 63–97.
[9] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, T. Schwede, and A. Tramontano. 2017. Assessment of model accuracy estimations in CASP12. *Proteins: Struct, Funct, Bioinf* 86, Suppl 1 (2017), 345–360.

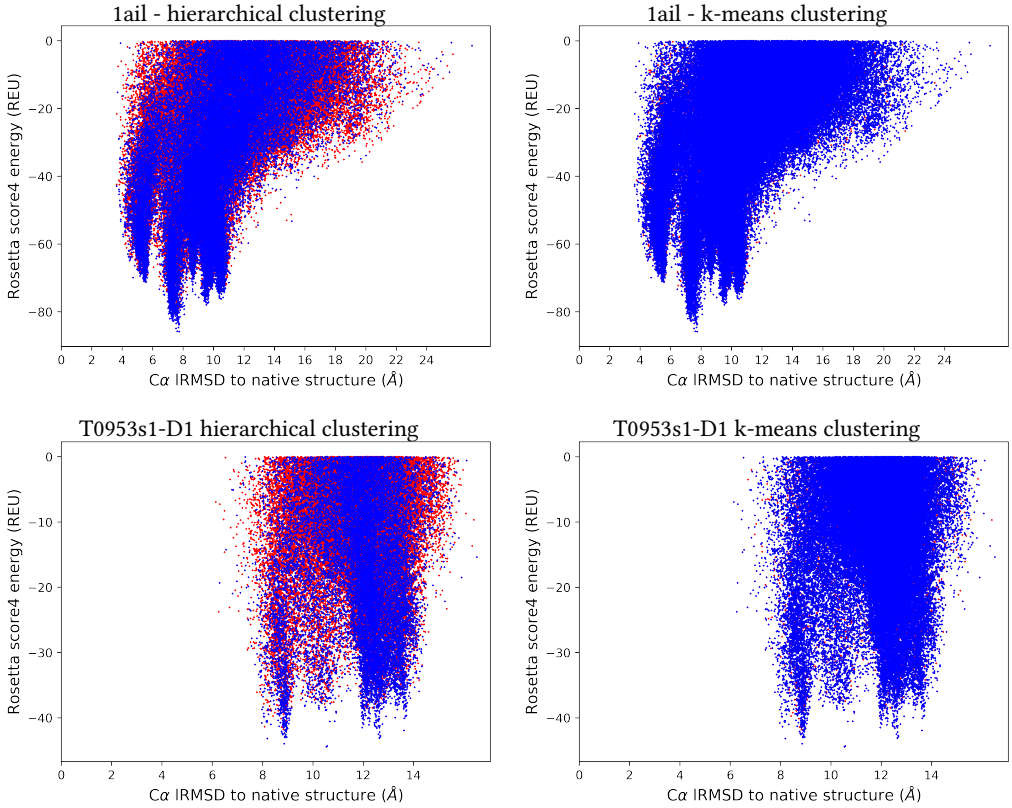

Figure 6: Decoys in the Ω_{gen} ensemble are plotted in red in terms of their lRMSD (Å) from the native structure (x-axis) versus their Rosetta score4 energy function (y-axis) measured in Rosetta Energy Units (REUs). Decoys in the Ω_{red} ensemble obtained via hierarchical and k-means clustering are superimposed in blue. Targets are selected from the benchmark and CASP datasets.

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

[11] J. Lee, P. Freddolino, and Y. Zhang. 2017. Ab initio protein structure prediction. In *From Protein Structure to Function with Bioinformatics* (2 ed.), D. J. Rigden (Ed.). Springer London, Chapter 1, 3–35.

[12] P. Mani, M. Vazquez, J. R. Metcalf-Burton, C. Domeniconi, H. Fairbanks, G. Bal, E. Beer, and S. Tari. 2019. The Hubness Phenomenon in High-Dimensional Spaces. In *Research in Data Sciences*, Gasparovic E. and Domeniconi C. (Eds.). Association for Women in Mathematics, Vol. 17. Springer, Cham, Switzerland, 15–45.

[13] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comp. Biol.* 12, 4 (2016), e1004619.

[14] A. D. McLachlan. 1972. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst A* 26, 6 (1972), 656–657.

[15] K. Molloy, S. Saleh, and A. Shehu. 2013. Probabilistic Search and Energy Guidance for Biased Decoy Sampling in Ab-initio Protein Structure Prediction. *IEEE/ACM Trans Comput Biol and Bioinf* 10, 5 (2013), 1162–1175.

[16] B. Olson, K. A. De Jong, and A. Shehu. 2013. Off-Lattice Protein Structure Prediction with Homologous Crossover. In *Conf on Genetic and Evolutionary Computation (GECCO)*. ACM, New York, NY, 287–294.

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

[18] B. Olson and A. Shehu. 2014. Multi-Objective Optimization Techniques for Conformational Sampling in Template-Free Protein Structure Prediction. In *Intl Conf on Bioinf and Comp Biol (BICoB)*. Las Vegas, NV, 143–148.

[19] W. Qiao, N. Akhter, X. Fang, T. Maximova, E. Plaku, and A. Shehu. 2018. From Mutations to Mechanisms and Dysfunction via Computation and Mining of Protein Energy Landscapes. *BMC Genomics* 19, Suppl 7 (2018), 671.

[20] W. Qiao, T. Maximova, X. Fang, E. Plaku, and A. Shehu. 2017. Reconstructing and Mining Protein Energy Landscape to Understand Disease. IEEE, Kansas City, MO.

[21] A. Shehu. 2015. A Review of Evolutionary Algorithms for Computing Functional Conformations of Protein Molecules. In *Computer-Aided Drug Discovery*, W. Zhang (Ed.). Springer Verlag.

[22] M. Steinbach, L. Ertöz, and V. Kumar. 2004. The Challenges of Clustering High Dimensional Data. In *New Directions in Statistics Physics*, L. T. Wille (Ed.). Springer, Berlin, Heidelberg, 273–309.

[23] D. Xu and Y. Zhang. 2012. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct, Funct, Bioinf* 80, 7 (2012), 1715–1735.

[24] A. Zaman, K. A. De Jong, and A. Shehu. 2019. Using Subpopulation EAs to Map Molecular Structure Landscapes. In *Conf on Genetic and Evolutionary Computation (GECCO)*. ACM, New York, NY, 1–8.

[25] A. Zaman and A. Shehu. 2019. Balancing multiple objectives in conformation sampling to control decoy diversity in template-free protein structure prediction. *BMC Bioinformatics* 20, 1 (2019), 211. DOI: https://doi.org/10.1186/s12859-019-2794-5

[26] G. Zhang, L. Ma, X. Wang, and X. Zhou. 2018. Secondary Structure and Contact Guided Differential Evolution for Protein Structure Prediction. *IEEE/ACM Trans Comput Biol and Bioinf* (2018). DOI: https://doi.org/10.1109/TCBB.2018.2873691 preprint.

[27] G. J. Zhang, G. Zhou, X, X. F. Yu, H. Hao, and L. Yu. 2017. Enhancing protein conformational space sampling using distance profile-guided differential evolution. *IEEE/ACM Trans Comput Biol and Bioinf* 14, 6 (2017), 1288–1301.