Assembly of Low-Energy Protein Conformations with Heterogeneous Fragments

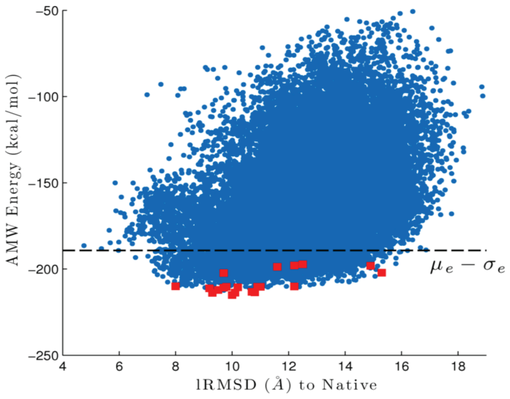
Kevin Molloy[1], Amarda Shehu[1,2]
[1]Department of Computer Science [2]Department of Bioinformatics and Computational Biology
George Mason University, Fairfax VA, 22030
http://www.cs.gmu.edu/~ashehu
amarda@gmu.edu

Abstract

The most successful ab-initio protein structure prediction methods employ fragment-based assembly. This technique is usually implemented in the context of a Metropolis Monte Carlo template. Fragment length is an important consideration, as it controls the complexity of the discretization of the search space and the associated energy surface. This work measures the impact of employing multiple, heterogeneous fragment lengths in the context of a multi-stage probabilistic search framework.

Clustering

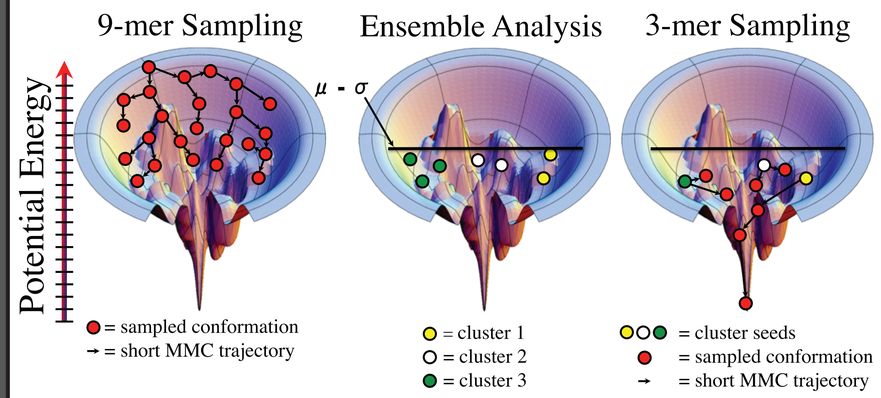
In our multi-stage exploration, an ensemble analysis is performed to select promising conformations from the first stage. Four clustering techniques were evaluated: leader clustering, radius of gyration (Rg) clustering, K-Means clustering, and USR-Leader clustering[3,6]. All methods exhibited similar performance with Rg clustering offering the best weighted performance. The ensemble of conformations generated from the first stage for the *3gwl* system is shown below. This figure highlights the low correlation between low lRMSD structures and a low coarsed-grain energy evaluation. The conformations selected with the Rg clustering technique are highlighted with red squares.



Methods

Our lab's probabilistic structure prediction framework is used to perform a multi-stage search for the protein native state[1,3]. This framework couples the use of fragment based assembly using heterogeneous lengths with a powerful probabilistic search algorithm. The search algorithm provides a broad view of the conformational space by biasing the search towards under explored regions.

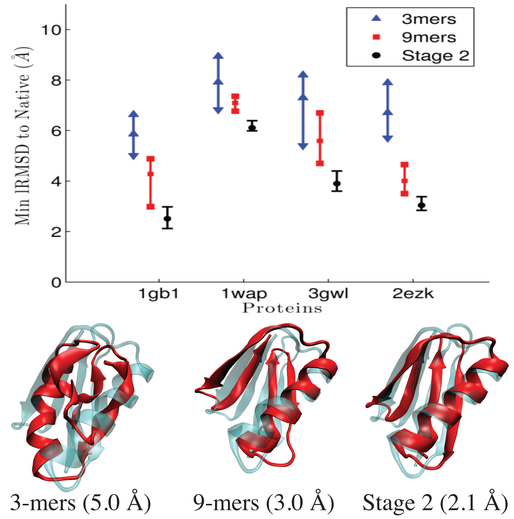
Fragment libraries of length 9 and 3 are constructed using methods that exploit local feature characteristics (secondary structure and sequence similarity)[1,2,4]. As shown below, the first stage grows a search tree in conformational space using 9-mers. To focus on low-energy structures, the ensemble from the first stage is reduced via an energetic cutoff. The remaining conformations are analyzed via clustering algorithms to identify *k* conformations which seed the second stage of the search performed with 3-mers.



Results

Four protein systems are selected to test our new method with the results shown in the table to the right. These results show that compared to our prior work (3-mer column), using 9-mers improves our results. Our two-stage method performed similarly to a single stage 3-mer search when seeds were selected using our radius of gyration clustering method.

		minimum lRMSD(Å)				
PDB ID	Length Fold	3-mer	9-mer	2-Stage Cluster	2-Stage Optimal	Rosetta
1gb1	56 (α/β)	5.0	3.0	6.2	2.5	2.7
1wap	68 (β)	6.8	6.8	7.5	6.1	6.8
2ezk	93 (α)	5.7	3.5	5.2	2.8	3.2
3gwl	106(α)	5.4	4.7	5.3	3.6	9.1



To fully evaluate the strength of this approach, the second stage is seeded with the optimal structure obtained from the first stage of the search. This experiment was performed 10 times with the minimum lRMSD structure recorded for each exploration. The figure to the left shows the results, comparing our prior work (3-mers) to this two-stage approach. The error bars show the variance between executions and the symbol on each bar shows the mean lRMSD value. This approach illustrates improved lRMSD to native values for all proteins tested. The progress of each stage can be seen visually in the structures to the left for the *1gb1* system (red = predicted structure, transparent blue = native structure)[5].

Conclusions

We present methods that employ the use of heterogeneous fragment lengths within a powerful probabilistic search framework. Employing longer fragment lengths allow the exploration to quickly obtain conformations of quality that exceed our prior work. Refining these structures further using shorter fragments improves the quality of the predicted structure when good seed conformations are selected from the first stage. Selecting good seed conformations is a challenging problem. The techniques proposed in this work do not adequately capture the lowest-lRMSD conformations from the initial stage. This work shows that if a method for selecting better seeds can be identified, the two-stage exploration approach with different fragment lengths guides the exploration to elucidate conformations of higher quality compared to a single stage approach.

Acknowledgements

Grant No. 1016995

References

1 B. Olson, K. Molloy, and A. Shehu. In Search of the Protein Native State with a Probablistic Sampling Approach. *J Bioinf and Comp Biol* 2011.

2 K. Molloy. "Variable-Length Fragment Assembly within a Probablistic Protein Structure Prediction Framework", Fairfax, VA. 2011.

3 A. Shehu and B. Olson, Guiding the search for native-like protein conformations with an ab-inito tree-based exploration. *Int J. Robot Res* 29(8):1106-11 227, 2010.

4 P. Bradley, K.M.S. Misura, and D. Baker. "Towards High-Resolution de novo structure prediction for small proteins", *Science*, 309(5742), 1868-1871, 2005.

5 W. Humphrey, A. Dalke, and K. Schulten, "VMD - Visual Molecular Dynamics", *J. Mol Graph Model*, vol. 14, no. 1, pp. 33-38, 1996.

6 P. J. Ballester and G. Richards. Ultrafast shape recognition to search compound databases for similar molecular shapes. J. Comput. Chem., 28(10):1711–1723, 2007.