Protein-protein Docking Using Information from Native Interaction Interfaces

Irina Hashmi
Department of Computer Science
4400 University Drive
Fairfax, VA 22030
ihashmi@gmu.edu

Amarda Shehu*
Department of Computer Science
Department of Bioengineering
School of Systems Biology
4400 University Drive
Fairfax, VA 22030
amarda@gmu.edu

ABSTRACT

We present a probabilistic search algorithm for rigid-body protein-protein docking. The algorithm is a realization of the basin hopping framework for sampling low-energy local minima of a given energy function. To save computational resources, the algorithm employs a machine learning model to score bound configurations prior to subjecting promising configurations to a local optimization with a sophisticated force field. The machine learning model is a decision tree trained on 138 known native dimeric interactions to learn features that constitute a *true* interaction interface. The FoldX force field is employed only on dimeric configurations sampled by the algorithm that are determined by the decision tree model to contain true interaction interfaces. The preliminary results are promising and motivate us for further investigation of such an informatics-driven approach to protein-protein docking.

Categories and Subject Descriptors

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

General Terms

Algorithms

Keywords

Protein-protein rigid docking; machine learning; native interaction; basin hopping

1. BACKGROUND

Proteins take specific three dimensional shapes that they use to bind with other molecules and so perform specific cellular tasks. Modeling these bound complexes is key to characterizing supramolecular assemblies and so understanding the molecular basis of biological function. Template-free structural characterization of protein assemblies entails searching a high-dimensional configuration space. Simplifying the problem of protein-protein docking to its rigid-body dimeric version brings the dimensionality of the search space down to the 6 parameters needed to encode spatial arrangements of the moving unit around the reference unit.

Though significant computational efforts are devoted to pairwise rigid-body protein-protein docking, the problem remains challenging. Primarily, the difficulty lies with either search algorithms of limited exploration capability or with the accuracy of the criterion used to guide these algorithms to the true native assembly, or a combination of both. Lately, a lot of work has resulted in probabilistic search algorithms with high exploration capability [13, 14, 7, 8]. However, guidance of these algorithms by an energy function presents a problem, as all current energy functions, even physics-based ones, contain errors and distort the true underlying energy surface. To address this issue, a complementary direction of research focuses on learning aspects of native interaction interface and encoding them either explicitly in the search process itself [5, 6, 7] or implicitly in a pseudo-energy function [10, 1, 2].

In this preliminary investigation we present a hybrid approach that employs a probabilistic search algorithm of high exploration capability but guides the algorithm in a computationally efficient manner towards native interaction interfaces. Rather than employ a costly energy function, the algorithm is guided in two steps, first ranking configurations with a machine learning model, then refining promising configurations with a sophisticated force-field.

2. METHODS

The search algorithm here builds upon the basin hopping framework in our previous work [7]. The framework is an iterative applications of structural perturbation followed by energetic minimization. Both the perturbation and energetic minimization operate over the space of rigid-body transformations. In previous work, these transformations are limited to matching geometrically-complementary and evolutionary-conserved regions of molecular surfaces [7]. Local optimization seeks the minimum of a simple interaction energy function [7]. In this work, we remove the evolutionary conservation consideration, and encapsulate it instead in a more general machine learning model. The perturbation

*corresponding author

BCB '13, September 22 - 25, 2013, Washington, DC, USA

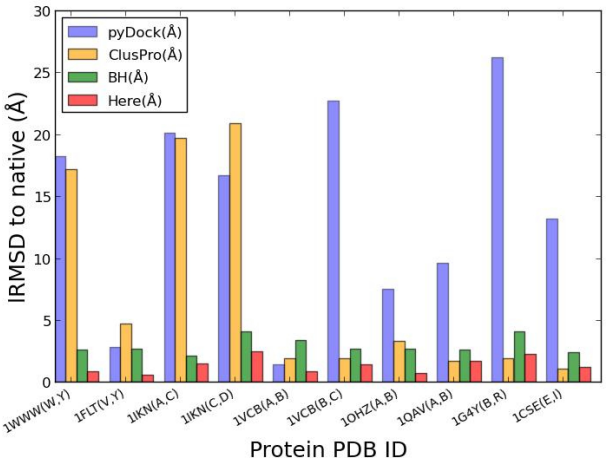

Figure 1: Comparison of approach proposed here to other methods in terms of lowest RMSD to known native dimeric structure.

modifies a transformation to obtain a new one. The local optimization proceeds in two steps, first expediently ranking a sampled interface with a learned model and then subjecting the corresponding dimeric configuration, if determined promising by the model, to refinement with the FoldX energy function. The model is learned a priori on known native interfaces, and comparison of various classifiers leads us to use a decision tree (DT) model that also facilitates designing a score for ranking configurations. Further details are available in [9].

3. RESULTS

A testing dataset of 11 protein dimers is selected here to compare the performance of the proposed approach with other known methods for docking. Our approach is run to obtain $10,000$ configurations that are representative of local minima in the FoldX energy surface. Fig. 1 shows a representative result that compares methods in terms of lowest RMSD to the known native structure. The methods chosen for comparison includes our previous basin hopping algorithm in [7], pyDock [3], and ClusPro [4]. Results show that the proposed informatics-guided approach is capable of obtaining near-native configurations within an average range of < 2Åto native configurations. Results are not only comparable, but better than other methods on 70% of the systems.

4. CONCLUSION

The results prompt us to further investigate the combination of a probabilistic search with a learned model. It is worth noting that the use of such a model postpones the usage of computationally-demanding energy functions to the critical part of guiding to the vicinity of the native structure. Ongoing work is considering expanding the scope of our investigation to a larger testing dataset and considering more powerful search frameworks, building on related work in our lab on robotics-inspired and evolutionary search algorithms for free-template protein structure prediction [11, 12].

5. ACKNOWLEDGMENTS

This work is supported in part by NSF IIS CAREER Award No. 1144106. We thank members of the Shehu lab for their valuable comments on this work and Dr. H. Rangwala for his feedback on a class project pursuing a preliminary investigation of classification models for interfaces.

6. REFERENCES

[1] B. Akbal-Delibas, I. Hashmi, A. Shehu, and N. Haspel. Refinement of docked protein complex structures using evolutionary traces. In *IEEE Intl Conf on Biomed and Bioinf Workshops*, pages 400–404, November 2011.

[2] B. Akbal-Delibas, I. Hashmi, A. Shehu, and N. Haspel. An evolutionary conservation based method for refining and reranking protein complex structures. *J of Bioinf and Comp Biol*, 10(3):1242008, 2012.

[3] T. M. Cheng, T. L. Blundell, and J. Fernandez-Recio. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins*, 68(2):503–515, 2007.

[4] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho. ClusPro: a fully automated algorithm for protein-protein docking. *Nucl. Acids Res.*, 32(S1), 2004.

[5] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Protein docking with information on evolutionary conserved interfaces. In *IEEE Intl Conf on Biomed and Bioinf Workshops*, pages 358–365, November 2011.

[6] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Guiding protein docking with geometric and evolutionary information. *J Bioinf and Comp Biol*, 10(3):1242002, 2012.

[7] I. Hashmi and A. Shehu. A basin hopping algorithm for protein-protein docking. In *IEEE Intl Conf on Bioinf and Biomed*, pages 466–469, October 2012.

[8] I. Hashmi and A. Shehu. Hopdock: A probabilistic search algorithm for decoy sampling in protein-protein docking. *Proteome Sci*, 2013. in press.

[9] I. Hashmi and A. Shehu. Informatics-driven protein-protein docking. In *ACM Cong on Bioinf and Comp Biol Workshops*, pages 1–8, September 2013.

[10] E. Kanamori, Y. Murakami, Y. Tsuchiya, D. Standley, H. Nakamura, and K. Kinoshita. Docking of protein molecular surfaces with evolutionary trace analysis. *proteinssfb*, 69:832–838, 2007.

[11] B. Olson, K. Molloy, and A. Shehu. Enhancing sampling of the conformational space near the protein native state. In *BIONETICS: Intl. Conf. on Bio-inspired Models of Network, Information, and Computing Systems*, Boston, MA, December 2010.

[12] S. Saleh, B. Olson, and A. Shehu. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction. *BMC Struct Biol*, 2013. in press.

[13] 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.

[14] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and A. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2013.