Protein Docking with Information on Evolutionary Conserved Interfaces

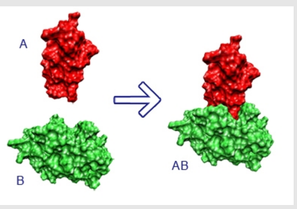
Irina Hashmi∗, Bahar Akbal-Delibas†, Nurit Haspel†, and Amarda Shehu∗
∗ Department of Computer Science, George Mason University, Fairfax, VA 22030
† Department of Computer Science, University of Massachusetts at Boston, Boston, MA, 02125

ABSTRACT

Structural modeling of molecular assemblies lies at the heart of understanding molecular interactions and biological function. We present a method for docking protein molecules and elucidating native-like structures of protein dimers. Our method is based on geometric hashing to ensure the feasibility of searching the combined conformational space of dimeric structures. The search space is narrowed by focusing the sought rigid-body transformations around surface areas with evolutionary-conserved amino-acids. Recent analysis of protein assemblies reveals that many functional interfaces are significantly conserved throughout evolution.

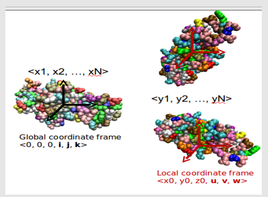
We test our method on a broad list of sixteen diverse protein dimers and compare the structures found to have lowest lRMSD to the known native dimeric structures to those reported by other groups. Our results show that focusing the search around evolutionary-conserved interfaces results in lower lRMSDs.

PROTEIN DOCKING



Docking refers to a computational method that tries to predict the binding orientation between two biomolecules that is more likely to be present in nature.

COMPUTATIONAL CHALLENGES



High dimensional search space:
N X M + 6

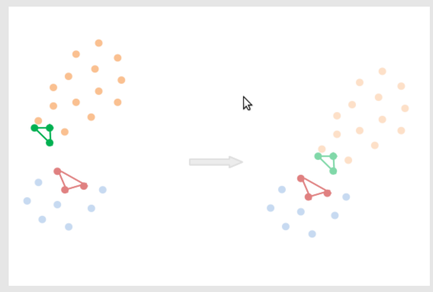
N, M = number of parameters to represent the unbound protein structures
6 = three translations followed by three rotations

METHODS

Connolly Surface A1 → Critical Points A2 → Active Triangles B → Rigid Body Transformation C

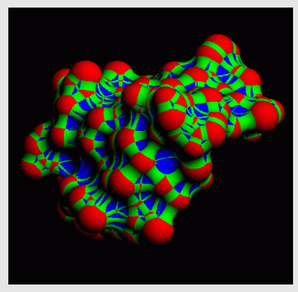
Regions Relevant for Matching

The JET method [1], which relies on multiple sequence analysis, is employed to identify conserved amino acids. The JET score calculated for each amino acid can range from 0.0 (least conserved) to 1.0 (most conserved). Amino acids with at least 50% surface accessibility and JET score above a predefined threshold are deemed 'active' and assumed to participate in the interaction interface. The rest of the amino acids are treated as passive.

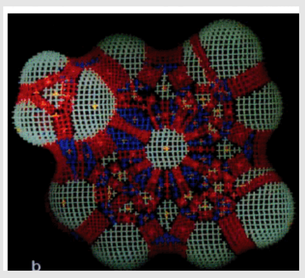


From Active Triangles to Rigid-body Transformation

A. Molecular Surface Representation



Convex
Saddle
Concave



Cap/ Pit: Critical point of convex/concave surface
Belt: Critical point of saddle surface

A1. Connolly Surface[2]:

- Dense Representation
- Solvent accessible surface area.
- Each point is represented by 3D coordinate, normal mode, type of surface.
- Type ranges from convex, to saddle, to concave.

A2. Critical Points[3]:

- Sparse Representation.
- Projection of the center of gravity of a Connolly face.
- Types are 'caps', 'pits', or 'belts' to correspond to convex, concave, or saddle faces.
- Takes the conservation score of its closest amino acid on the molecular surface.

C. Rigid Body Transformation

- Active triangles are considered as reference frames for each transformation.
- For each unique active triangle selected from a monomer A, a matching active triangle is selected from the second one B.
- The features for matching are only geometric at this point, as in [4].
- The two corresponding frames define a transformation.
- The transformation aligns the frames by superimposing their origins and rotating B on A.

B. Critical Points to Active Triangles

Input: Critical points of a subunit
Output: Active Triangles – Triangle with at lease one critical point

1. while total number of critical points:
2. randomly pick one critical point p_1 with conservation score > 0.5
3. perform a query on p_1 around a threshold radius (2-5 Å)
4. pick p_2 and p_3 randomly from this query result list where p2!=p3

Uniqueness of the Triangles:
- A lexicographic ordering of a triangle's vertices is employed to ensure that no two triangles share the first vertex in the ordering.
- No two triangles share their center of mass.

RESULTS

We selected 16 different diverse set of dimers with known native structures as our systems of study. Results obtained after the experiments make the case that on all selected systems of study the method is able to reproduce the native structure with an accuracy of less than 5Å within a feasible amount of time.

1. This table shows two set of results: The number of active triangles goes down as the conservation threshold increases. The lowest lRMSD generally goes down as the conservation threshold goes up.

PDB ID	Threshold	Nr. Triangles	lRMSD (Å)
1FLT (V,Y)	0.25	2417	2.06
	0.50	2338	1.12
	0.75	2080	1.03
1WWW (W, Y)	0.25	2900	2.29
	0.50	2911	2.24
	0.75	2854	2.60
1VCB (A,C)	0.25	2002	3.03
	0.50	1964	4.44
	0.75	1954	2.84

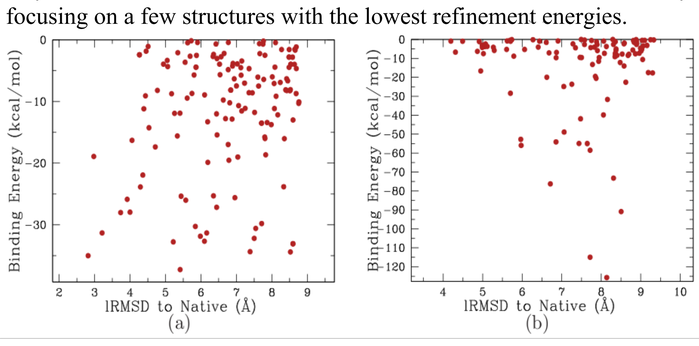
2. Lowest lRMSDs by our method are compared to BUDDA[5] and Kanamori [6]. Size refers to number of atoms. On about 12 out of the 16 dimers, the method presented here achieves similar low lRMSDs to other two methods. On the remaining systems, our method outperforms these other two.

PDB ID (Chains)	Size	BUDDA (Å)	Kanamori (Å)	Our Algorithm (Å)
1C1Y (A, B)	1376, 658	1.2	NA	1.29
1G4U (R, S)	1398, 2790	1.03	NA	2.20
1DS6 (A, B)	1413, 1426	1.18	NA	1.87
1TX4 (A, B)	1579, 1378	1.37	NA	2.42
1WWW (W, Y)	862, 782	11.4	NA	2.24
1FLT (V, Y)	770, 758	1.55	NA	1.12
1IKN (A, C)	2262, 916	1.19	NA	2.04
1IKN (A, D)	2262, 1589	1.43	NA	2.04
1IKN (C, D)	916, 1589	2.05	NA	2.01
1VCB (A, B)	755, 692	0.75	NA	2.06
1VCB (B, C)	692, 1154	13.1	NA	4.62
1VCB (A, C)	755, 1154	1.73	NA	1.27
1OHZ (A, B)	1027, 416	1.77	0.66	1.70
1T6G (A, C)	2628, 1394	1.64	3.8	2.55
1ZHI (A, B)	1597, 1036	25.3	3.4	1.75
2HQS (A, C)	3127, 856	29.1	2.55	2.19

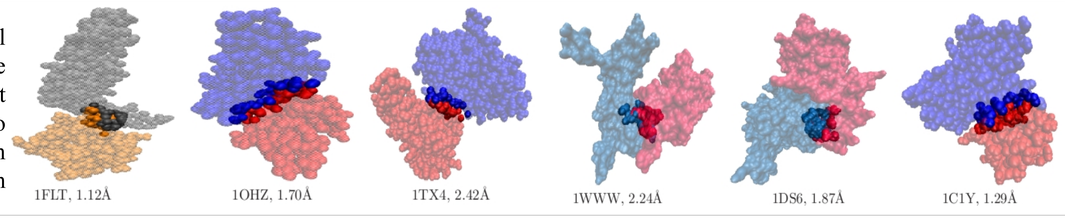
3. This results shows the effect of the number of active triangles on time and accuracy of the method. In the first setting all unique triangles have been selected and the second settings chooses one third of the number of active triangles. This construction of active triangles achieves similarly low lRMSDs while improving the feasibility of the method.

PDB ID (Chains)	Nr. triangles ratio	Run Time ratio	lRMSD Diff. (Å)
1C1Y (A, B)	1 : 2.80, 1 : 2.86	1 : 6.57	1.02
1G4U (R, S)	1 : 2.89, 1 : 2.88	1 : 6.44	−0.72
1DS6 (A, B)	1 : 2.86, 1 : 2.88	1 : 7.16	1.42
1TX4 (A, B)	1 : 2.88, 1 : 2.91	1 : 6.44	0.30
1WWW (W, Y)	1 : 2.94, 1 : 2.94	1 : 9.57	−0.05
1FLT (V, Y)	1 : 3.23, 1 : 2.56	1 : 10.5	1.69
1IKN (A, C)	1 : 2.90, 1 : 2.89	1 : 7.93	−0.89
1IKN (A, D)	1 : 2.90, 1 : 2.90	1 : 11.9	1.01
1IKN (C, D)	1 : 2.89, 1 : 2.90	1 : 7.26	0.99
1T6G (A, C)	1 : 2.90, 1 : 2.86	1 : 6.82	−0.27

4. The figure plots for two system PDB ID (a) 1FLT and (b) 1WWW energy values obtained from Firedock[7] for the refined structures versus lRMSDs from the native structure. Only negative energy values are shown. The results show that many of the lowest-energy structures are also low in lRMSD from to the native structure. This suggests that short refinements may allow the method to detect low-lRMSD structures automatically by focusing on a few structures with the lowest refinement energies.



5. Lowest-lRMSD structures and the actual lRMSD achieved are shown for some of the dimers. Chains are drawn in different colors in transparent. Conserved amino acids in contact with one another are drawn in opaque. Structures are drawn with VMD[8].



CONCLUSIONS

Our approach is a promising first step towards efficiently computing physically relevant structures. Our ongoing work focuses on the following:

1. Incorporate scoring, clustering and ranking.
2. Combining the method with more detailed refinement procedure.
3. Investigate the approach in the context of a Monte Carlo based exploration.
4. Extension of the method for arbitrary number of molecular assemblies.

REFERENCES

1. *S. Engelen, A. T. Ladislas, S. Sacquin-More, R. Lavery, and A. Carbone, "A large-scale method to predict protein interfaces based on sequence sampling," PLoS Comp Bio, vol. 5, no. 1, p. e1000267, 2009.*
2. *M. L. Connolly, "Analytical molecular surface calculation," J. Appl. Cryst., vol. 16, no. 5, pp. 548–558, 1983.*
3. *R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov, "Examination of shape complementarity in docking of unbound proteins," Proteins: Struct. Funct. Bioinf., vol. 36, no. 3, pp. 307–317, 1999.*
4. *H. L. Wolfson and I. Rigoutsos, "Geometric hashing: an overview," IEEE Comp Sci and Engineering, vol. 4, no. 4, pp. 10–21, 1997.*
5. *V. Polak, "Polak v 2003 Budda: backbone unbound docking application master's thesis school of computer science, tel-aviv university," Master's thesis, Computer Science, Tel-Aviv University, Tel-Aviv,Israel, 2003.*
6. *E. Kanamori, Y. Murakami, Y. Tsuchiya, D. Standley, H. Nakamura, and K. Kinoshita, "Docking of protein molecular surfaces with evolutionary trace analysis," Proteins: Struct. Funct. Bioinf., vol. 69, pp. 832–838, 2007.*
7. *N. Andrusier, R. Nussinov, and H. J. Wolfson, "Firedock: fast interaction refinement in molecular docking," Proteins: Struct. Funct. Bioinf., vol. 69, no. 1, pp. 139–159, 2007.*
8. W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," J. Mol. Graph. Model., vol. 14, no. 1, pp. 33–38, 1996, http://www.ks.uiuc.edu/Research/vmd/.

ACKNOWLEDGEMENTS

The authors would like to thank members of the Computational Biology group Department of Computer Science, GMU for valuable comments on this work.
Shehu Lab: http://www.cs.gmu.edu/~ashehu/
Email: amarda@gmu.edu