A Basin Hopping Algorithm for Protein-protein Docking

Irina Hashmi[1] and Amarda Shehu[1,2,3*]
[1]*Department of Computer Science*
[2]*Department of Bioinformatics and Computational Biology*
[3]*Department of Bioengineering*
George Mason University
Fairfax, VA, 22030, United States
ihashmi@gmu.edu, amarda@gmu.edu
∗Corresponding Author

***Abstract*—We present a novel probabilistic search algorithm to efficiently search the structure space of protein dimers. The algorithm is based on the basin hopping framework that repeatedly follows up structural perturbation with energy minimization to obtain a coarse-grained view of the dimeric energy surface in terms of its local minima. A Metropolis criterion biases the search towards lower-energy minima over time. Extensive analysis highlights efficient and effective implementations for the perturbation and minimization components. Testing on a broad list of dimers shows the algorithm recovers the native dimeric configuration with great accuracy and produces many minima near the native configuration. The algorithm can be employed to efficiently produce relevant decoys that can be further refined at greater detail to predict the native configuration.**

***Keywords*-protein docking; basin hopping; geometric hashing; rigid-body transformations; evolutionary conservation**

I. Introduction

Protein-protein interactions are central to cellular organization, signal transduction and protein degradation [1]. In a mechanistic view of macromolecules, protein assemblies assume specific structures to achieve their biological function. Structure determination of these assemblies, also known as protein-protein docking, is a central problem to understanding molecular interactions and obtaining a molecular basis for biological function in the living and diseased cell [2].

Computational approaches promise to complement wet-laboratory efforts. Several protein-protein docking methods are now available, such as Haddock [3], Zdock [4], ClusPro [5], PatchDock and SymmDock [6], Combdock [7], [8] ,SKE-DOCK [9], and others [10]. The community-wide CAPRI (Critical Assessment of PRedicted Interactions) experiment shows that, while accuracy is improving, no single method is sufficient to successfully predict native assemblies in every test case [11]. About 30-58% of the correct interface is predicted in any given target [12] due to the high dimensionality of the parameter space.

Many methods approach pairwise docking as an optimization problem on a continuous parameter space and seek regions in the energy surface that correspond to low-energy near-native docked configurations [3], [13]. Other methods, such as SKE-DOCK [9], PatchDOCK [6], and Combdock [7], [8], take a geometry- rather than an energy-driven approach; the parameter space is discretized through rigid-body transformations that dock two monomers by aligning geometrically-complementary patches on their surfaces. The number of transformations can be reduced by narrowing the focus to evolutionary-conserved patches [14], [15].

Here we propose a novel probabilistic search algorithm that combines a geometry-driven approach with the basin hopping (BH) [16] framework to efficiently generate low-energy decoys for docking. The BH framework is employed to explicitly sample local minima of an energy function. The algorithm essentially consists of repeated applications of a structural perturbation followed by an energy minimization. A Metropolis criterion biases the resulting trajectory of local minima towards minima that reside in the lower-energy regions of the energy surface. Different implementations for perturbation and minimization components as well as various effective temperatures for the Metropolis criterion are analyzed. Testing on a broad list of dimers shows that the algorithm recovers the near-native configuration with great accuracy and produces a large number of minima near the native configuration. This suggests the approach can be employed to efficiently produce near-native decoys. Coupled with ranking and further refinement [3], [17], this work promises to advance protein-protein docking.

II. Methods

A. Overview of the Main Ingredients in the BH algorithm

The BH algorithm we propose here obtains a trajectory of n dimeric configurations $C_1, \ldots, C_n$ that correspond to minima of a chosen energy function. The algorithm hops between two consecutive minima C_i and C_{i+1} through an intermediate $C_{\text{perturb,i}}$ conformation. The perturbation component modifies the current minimum C_i to escape it

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

through the resulting conformation $C_{\text{perturb,i}}$. The minimization component follows a series of modifications, starting with $C_{\text{perturb,i}}$, to reach a new conformation C_{i+1} that represents the nearest minimum to $C_{\text{perturb,i}}$. C_{i+1} is added to the trajectory according to the Metropolis criterion based on the energetic difference between C_i and C_{i+1} and the effective temperature employed. Different implementations are analyzed for the perturbation and minimization. Straightforward ones build over the rigid-body transformations that align geometrically-complementary and evolutionary-conserved patches on molecular surfaces. We briefly define these patches before relating further details.

B. From Molecular Surfaces to Rigid-body Transformations

The Connolly molecular surface [18] can be simplified by representing only key points known as critical points [19]. In previous work, we only consider critical points near evolutionary-conserved amino acids to define "active" triangles. Aligning geometrically-complementary triangles on the molecular surfaces of the monomers involved results in a rigid-body transformation that docks the monomers. Further details can be found in our previous work [14], [15].

C. Structural Perturbation

The perturbation in our BH algorithm modifies minimum C_i by seeking a new rigid-body transformation. Research on the BH framework in other domains [20], [21] has shown that perturbation needs to preserve the good features of C_i in $C_{\text{perturb,i}}$. One way to do so is by limiting the neighborhood over which active triangles are sought for the new transformation. We do so as follows. Let the active triangles that define the transformation resulting in C_i be T_A, T_B (for monomers A and B). A new active triangle $T_A^{'}$ is sampled uniformly at random in a d-neighborhood of T_A over the surface of A. (d is distance between the center of mass of T_A and $T_A^{'}$). A new active triangle $T_B^{'}$ is obtained similarly near T_B. The process is repeated until $T_A^{'}$ and $T_B^{'}$ are geometrically-complementary, and a new rigid-body transformation can be defined to obtain $C_{\text{perturb,i}}$.

Studies show that BH applications are successful when the magnitude of the perturbation, measured through some distance function over C_i and $C_{\text{perturb,i}}$, is neither small nor large [21], [22]. In section III we show the role of d on the magnitude of the perturbation and the ability of our algorithm to sample near-native minima. We also show that controlling d yields better results than minimization with random restarts (where d is essentially infinite).

D. Energy Minimization

The minimization modifies $C_{\text{perturb,i}}$ to obtain a new nearby minimum C_{i+1}. Its goal is to correct the structural features that the perturbation changed from C_i in $C_{\text{perturb,i}}$ to obtain a new set of good structural features that correspond to another minimum C_{i+1}. The minimization here consists of at most m structural modifications until k consecutive modifications fail to lower energy.

Let us represent a transformation as $\langle t, u, \theta \rangle$, where t is the translation component, and $\langle u, \theta \rangle$ are to the orientation component in an axis-angle representation (implemented through quaternions here). In each modification, a new translation is sampled in a δ_t neighborhood of t, and a new rotation is obtained by sampling a new axis $u^{'}$ rotating around the axis u by a sampled angle value δ_ϕ; a new angle is obtained by sampling in a δ_θ neighborhood around θ.

The minimization employs a simple energy function $E = E_{VdW} + E_{electrostatic} + E_{hydrogen-bonding}$. The first two terms, capturing van der Waals and electrostatic interactions, are implemented as in the CHARMM22 force field [23], and the 12-10 hydrogen-bonding term is as in [24].

III. Results

Testing is carried out on a 2.66GHz Opteron processor with 8GB of memory. 17 dimers with known native structures are selected for testing, because they vary in size, represent diverse functional classes, and have been tested by other methods (some are also CAPRI targets). While we have experimented with sampling $5,000$ to $20,000$ dimeric configurations, results do not change after $10,000$. Results below are on $10,000$ configurations.

We present four sets of results. The effect of the temperature in the Metropolis criterion is measured first over three representative systems. The same systems are selected to determine an effective neighborhood distance d for the perturbation. Different implementations of the minimization are tested on the same systems and analyzed next. Finally, the concluded effective temperature and perturbation and minimization implementations are employed on the rest of the dimers in our final set of results.

A. Choice of Effective Temperature

The effective temperature T_e in the Metropolis criterion affects the probability $e^{-\delta E T_e}$ with which an energy increase is accepted. We have tested various temperatures, but the results in Table I show the effect on sampling of two main temperatures on three representative systems (T_0 accepts a 2 kcal/mol increase with probability 0.39, whereas T_1 does so with probability 0.16). Column 1 shows PDB ids of native structures. Column 3 shows lowest lRMSD (least Root-Mean-Squared-Deviation) from the native structure over sampled configurations. Column 4 shows lowest sampled energy. Column 5 shows the percentage of configurations with lRMSD $<$ 5Å from the native structure. On two systems, T_0 allows getting closer to the native structure. T_1 expectedly achieves lower energies, but not necessarily more near-native configurations. Column 3 indicates T_0 is more effective, because it results in more or at least the same number of near-native ($<$ 5Å) configurations. The rest of the experiments employ T_0 in the Metropolis criterion.

Table I: Effective Temperature on Representative Systems.

PDB ID	T_e	lRMSD (Å)	Energy (kcal/mol)	$< 5\mathring{A}$ (%)
1FLT	T_0	1.69	-1.73	0.21
	T_1	2.10	-0.67	0.09
1WWW	T_0	2.98	-0.88	0.14
	T_1	2.16	-21.39	0.12
1C1Y	T_0	1.05	-0.83	0.78
	T_1	1.42	-10.50	0.80

Table II: Effect of Perturbation d on Representative Systems.

PDB ID	d (Å)	l_{m} (Å)	l<5Å(%)	l<10Å(%)	lRMSD (Å)	<5Å(%)
1FLT	$d=\infty$	16.73	0.29	3.82	3.37	0.11
	$d=7$	14.76	2.09	16.57	2.48	0.12
	$d=5$	13.48	4.73	18.20	1.69	0.23
1WWW	$d=\infty$	17.83	0.27	2.86	2.33	0.12
	$d=7$	14.62	2.19	13.32	3.63	0.08
	$d=5$	14.22	4.86	14.91	2.89	0.08
1C1Y	$d=\infty$	14.31	0.51	9.87	2.17	0.70
	$d=7$	11.23	4.86	30.02	2.09	0.71
	$d=5$	11.06	6.49	31.00	1.05	0.86

B. Controlling Perturbation Distance

We now compare 3 perturbation implementations. The first is random restart, where $d=\infty$. The second and third control d to 7 and 5Å, respectively. Table II compares the following statistics. We define l to be the lRMSD between two consecutive perturbed conformations, $C_{\mathrm{perturb,i}}$ and $C_{\mathrm{perturb,i+1}}$; l measures perturbation magnitude. Column 3 shows the median of l over perturbed conformations obtained during our algorithm. Columns 4, 5 show the percentage of l in $0-5$Å and $5-10$Å, respectively. Columns 6, 7 show resulting lowest lRMSD to the native structure and percentage of minima with lRMSD < 5Å to the native structure. The results in Table II show that controlling d allows reducing the median perturbation distance l. The number of consecutive perturbed conformations within < 5Å and < 10Å lRMSD of each-other also increases with lower d. Also, the number of minima with low lRMSDs to the native structure is not impacted negatively by lowering d. The rest of our experiments employ $d=5$Å in the perturbation.

C. Implementations for the Minimization Component

Here we compare different implementations for the minimization component. In all implementations, m=100 and k=20, but we vary the translation distance t in $\{1.5, 2.0, 2.5\}$Å. Table III compares the following statistics. We define i to be the lRMSD between $C_{\mathrm{perturb,i}}$ and C_{i+1}. Column 3 shows the median of i over conformations minima obtained by the BH algorithm. Columns 4, 5 show the percentage of i in $0-5$Å and $5-10$Å, respectively. Columns 6, 7 show resulting lowest lRMSD to the native structure and the percentage of minima with lRMSD < 5Å to the native structure. The results in Table III show that lower values of t overall keep consecutive perturbed and minima conformations close. Comparing number of minima close to the native structure suggests a translation threshold t=1.5Å can be employed on the rest of the protein systems.

Table III: Effect of Minimization on Representative Systems.

PDB ID	t (Å)	i_{m} (Å)	i<5Å(%)	i<10Å(%)	lRMSD (Å)	<5Å(%)
1FLT	1.5	10.14	19.30	29.69	2.70	0.07
	2.0	9.81	20.56	30.66	1.90	0.15
	2.5	9.98	20.26	29.82	1.69	0.23
1WWW	1.5	9.74	21.12	30.30	2.63	0.21
	2.0	9.68	21.50	30.55	3.88	0.12
	2.5	9.50	21.85	31.43	2.98	0.08
1C1Y	1.5	9.04	24.41	31.94	1.79	0.92
	2.0	9.40	21.99	31.78	2.14	0.51
	2.5	9.56	21.62	31.71	1.05	0.86

Table IV: Final Results of the BH algorithm.

PDB ID (Chains)	Size	Others(Å)	Here (Å)	$<$ tRMSD (%)	$< 5\mathring{A}$ (%)
1C1Y (A,B)	1376, 658	1.2-1.3	1.8	1.22	0.92
1DS6 (A,B)	1413, 1426	1.2-1.9	3.4	3.25	0.10
1TX4 (A,B)	1579, 1378	1.4-2.4	2.4	26.6	0.17
1WWW (W,Y)	862, 782	2.3-11.4	2.6	2.30	0.21
1FLT (V,Y)	770, 758	1.1-1.6	2.7	22.67	0.07
1IKN (A,C)	2262, 916	1.2-2.1	2.1	3.62	0.18
1IKN (C,D)	916, 1589	2.0-2.1	4.1	0.98	0.02
1VCB (A,B)	755, 692	0.8-2.1	3.4	3.41	0.11
1VCB (B,C)	692, 1154	1.3-13.1	2.7	13.82	0.16
1OHZ (A,B)	1027, 416	1.7-1.8	2.7	0.06	0.76
1T6G (A,C)	2628, 1394	1.7-2.6	3.6	15.85	0.06
1ZHI (A,B)	1597, 1036	1.8-25.3	4.6	5.28	0.01
2HQS (A,C)	3127, 856	2.2-29.1	2.6	7.46	0.40
1QAV (A,B)	663, 840	1.1-1.5	2.6	5.01	0.07
1G4Y (B,R)	682, 1156	0.8-2.3	4.1	37.94	0.04
1CSE (E,I)	1920, 522	0.7-1.5	2.4	0.86	0.91
1G4U (R,S)	1398, 2790	1.0-2.2	3.19	47.84	0.02

D. Application of the BH Algorithm on Diverse Systems

Table IV shows the lowest lRMSD to the native structure obtained by our BH algorithm on all 17 dimers in column 4. Lowest lRMSDs reported by other methods [8], [14] are shown in column 3. Columns 1-2 show PDB ids of native structure and dimer size. Let us order configurations by energy and define a tolerance lRMSD (tlRMSD) as 2Å higher than the lowest lRMSD achieved on a system by our algorithm. Column 5 shows the number of lowest-energy configurations that would have to be selected (as a percentage of the total size of the generated ensemble of local minima) in order to have that tlRMSD achieved by some configuration in the selected subset. Column 6 removes energy considerations and shows the percentage of minima with lRMSD less than 5Å to the native structure.

Table IV suggests that the BH algorithm achieves the low lRMSDs to the native structure on each system. The algorithm produces other configurations within 5Å of the native structure. Low lRMSDs are found among the lowest-energy configurations. These configurations, if selected in the course of a multi-stage docking protocol, will allow obtaining the native structure in great detail.

IV. Conclusion

We have presented a novel basin hopping algorithm to efficiently generate low-energy decoys for pairwise docking. The algorithm conducts its search over rigid-body transformations that align geometrically-complementary and evolutionary-conserved patches of molecules surfaces. The analysis here highlights efficient and effective implementations for the perturbation and minimization. In future work, the minimization can conduct its search for the nearest local minimum in a more detailed, possibly continuous, parameter space. Another interesting direction involves the employment of coarse-grained energy functions to efficiently project a perturbed dimeric configuration to its nearest local minimum. The presented results suggest that the proposed algorithm can be employed as the first stage in blind docking to reveal informative decoys that can then be clustered, ranked, and fed to further refinement protocols.

Acknowledgment

The authors would like to thank members of the Computational Biology group for valuable comments on this work.

References

[1] D. S. Goodsell and A. J. Olson, "Structural symmetry and protein function," *Annu. Rev. Biophys. and Biomolec. Struct.*, vol. 29, pp. 105–153, 2000.

[2] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking," *Curr. Opinion Struct. Biol.*, vol. 19, pp. 164–170, 2009.

[3] C. Dominguez, R. Boelens, and A. M. Bonvin, "Haddock: A protein-protein docking approach based on biochemical orbiophysical information," *J. Am. Chem. Soc.*, vol. 125, pp. 1731–1737, 2003.

[4] R. Chen, L. Li, and Z. Weng, "ZDock: an initial-stage protein-docking algorithm," *Proteins: Struct. Funct. Bioinf.*, vol. 52, no. 1, pp. 80–87, 2003.

[5] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: a fully automated algorithm for protein-protein docking," *Nucl. Acids Res.*, vol. 32, no. S1, 2004.

[6] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking," *Nucl. Acids Res.*, vol. 33, no. S2, pp. W363–W367, 2005.

[7] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson, "Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies," *J. Phys. Biol.*, vol. 2, pp. S156–S165, 2005.

[8] ——, "Prediction of multimolecular assemblies by multiple docking," *J. Mol. Biol.*, vol. 349, no. 2, pp. 435–447, 2005.

[9] G. Terashi, M. Takeda-Shitaka, K. Kanou, M. Iwadate, D. Takaya, and H. Umeyama, "The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation," *Proteins: Struct. Funct. Bioinf.*, vol. 69, no. 4, pp. 866–887, 2007.

[10] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "Geometry based flexible and symmetric protein docking," *Proteins: Struct. Funct. Bioinf.*, vol. 60, no. 2, pp. 224–231, 2005.

[11] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, "Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go," *British J Pharmacology*, vol. 153, no. S1, pp. S7–S27, 2009.

[12] M. F. Lensink and S. J. Wodak, "Blind predictions of protein interfaces by docking calculations in CAPRI," *Proteins: Struct. Funct. Bioinf.*, vol. 78, no. 15, pp. 3085–3095, 2010.

[13] S. Lyskov and J. J. Gray, "The RosettaDock server for local protein-protein docking," *Nucl. Acids Res.*, vol. 36, no. S2, pp. W233–W238, 2008.

[14] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Protein docking with information on evolutionary conserved interfaces," in *IEEE BIBM Comp. Struct. Biol. Workshop*, 2011, pp. 358–365.

[15] ——, "Guiding protein docking with geometric and evolutionary information," *J. Bioinf. and Comput. Biol.*, vol. 10, no. 3, p. 1242008, 2012.

[16] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *J. Phys. Chem. A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[17] B. Akbal-Delibas, I. Hashmi, A. Shehu, and N. Haspel, "Refinement of docked protein complex structures using evolutionary traces," in *Comput Struct Biol Workshop*, 2011, pp. 400–404, accepted.

[18] M. L. Connolly, "Analytical molecular surface calculation," *J. Appl. Cryst.*, vol. 16, no. 5, pp. 548–558, 1983.

[19] R. Norel, D. Petrey, H. J. Wolfson, and R. Nussinov, "Examination of shape complementarity in docking of unbound proteins," *Proteins*, vol. 36, no. 3, pp. 307–317, 1999.

[20] B. Olson and A. Shehu, "Populating local minima in the protein conformational space," in *IEEE Intl Conf on Bioinf and Biomed (BIBM)*, 2011, pp. 114–117.

[21] ——, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Sci*, 2012, in press.

[22] H. R. Lourenço, O. C. Martin, and T. Stützle, "Iterated local search," vol. 57, no. 513, pp. 321–353, 2002.

[23] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, 1983.

[24] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein–protein complexes," *Proc. Natl Acad of Sci USA*, vol. 99, no. 22, pp. 14 116–14 121, 2002.