Journal of Bioinformatics and Computational Biology
Vol. 10, No. 3 (2012) 1242008 (16 pages)

DOI: 10.1142/S0219720012420085

Imperial College Press
www.icpress.co.uk

GUIDING PROTEIN DOCKING WITH GEOMETRIC AND EVOLUTIONARY INFORMATION

IRINA HASHMI*, BAHAR AKBAL-DELIBAS†, NURIT HASPEL†
and AMARDA SHEHU*,‡,§,¶

*Department of Computer Science
George Mason University
Fairfax, VA, 22030, USA

†Department of Computer Science
University of Massachusetts at Boston
Boston, MA, 02125, USA

‡Department of Bioinformatics and Computational Biology
George Mason University
Fairfax, VA, 22030, USA

§Department of Bioengineering
George Mason University
Fairfax, VA, 22030, USA
¶ashehu@gmu.edu

Received 12 February 2012
Revised 2 April 2012
Accepted 8 April 2012
Published 21 May 2012

Structural modeling of molecular assemblies promises to improve our understanding of molecular interactions and biological function. Even when focusing on modeling structures of protein dimers from knowledge of monomeric native structure, docking two rigid structures onto one another entails exploring a large configurational space. This paper presents a novel approach for docking protein molecules and elucidating native-like configurations of protein dimers. The approach makes use of geometric hashing to focus the docking of monomeric units on geometrically complementary regions through rigid-body transformations. This geometry-based approach improves the feasibility of searching the combined configurational space. The search space is narrowed even further by focusing the sought rigid-body transformations around molecular surface regions composed of amino acids with high evolutionary conservation. This condition is based on recent findings, where analysis of protein assemblies reveals that many functional interfaces are significantly conserved throughout evolution. Different search procedures are employed in this work to search the resulting narrowed configurational space. A proof-of-concept energy-guided probabilistic search procedure is also presented. Results are shown on a broad list of 18 protein dimers and additionally compared with data reported by other labs. Our analysis shows that focusing the search around evolutionary-conserved interfaces results in lower lRMSDs.

Keywords: Protein docking; geometric hashing; evolutionary conservation.

¶Corresponding author.

1. Introduction

Molecules come together in molecular assemblies in order to achieve their biological function in the living cell. Modeling structural aspects of these assemblies is important to improve our understanding of molecular interactions and our ability to target molecules with drug compounds. Due to the ubiquity and central cellular role of protein molecules, significant computational efforts go toward predicting structures of protein-based assemblies, a problem known as protein docking.[1–5]

Protein docking is a challenging problem.[6–8] Even when docking only two protein molecules onto each other, the process involves searching for low-energy dimeric structures in a space of $N \times M + 6$ dimensions; N and M are the number of parameters employed to represent the unbound protein structures, and 6 is the number of translation and rotation parameters that correspond to the different placements of one monomer onto another. The large number of parameters results in a high-dimensional search space. As a result, most methods focus on rigid-body docking, where the monomeric structures are considered rigid. In this way, the focus is on finding the placements that result in low-energy dimeric structures. From now on, we will refer to these placements as rigid-body transformations, and to the dimeric structures that result after applying such a transformation as configurations.

The approach proposed here searches the space of rigid-body transformations. The approach is guided by geometry, as it only considers transformations that match geometrically complementary regions on the involved molecular surfaces. This focusing allows narrowing the configuration space that one would have to explore in search of dimeric configurations that reproduce the native structure. Based on findings that evolutionary-conserved regions are good predictors of functional interfaces,[9] the proposed approach further limits matching regions of interest to those that are evolutionary conserved. This greatly reduces the number of transformations attempted to obtain dimeric configurations.

The proposed approach searches for bound configurations that match geometrically complementary surface regions, essentially matching concave with convex regions. Geometric features are hashed in order to expedite the search for complementary regions, a process known as geometric hashing.[10,11] The employment of geometric hashing in this paper is due to its demonstrated success in allowing docking methods to feasibly compute configurations that through further energetic refinement reproduce known native structural assemblies.[12]

Instead of geometry, other docking methods are guided by energy and do not explicitly conduct their search for dimeric configurations over the space of rigid-body transformations. Classic search frameworks, such as Monte Carlo or Molecular Dynamics, or other energy minimization protocols are conducted to find minima of a designed energy function.[13] With a realistic energy function, these minima correspond to native-like structural assemblies.[6,8,14] Designing accurate energy functions to capture molecular interactions is a challenging research area.[15]

Several methods as web servers or/and software are now available for protein docking, such as Zdock,[16] Haddock,[17] ClusPro,[18] PatchDock and SymmDock,[19] Combdock,[20,21] FiberDock,[22] RosettaDock,[23] and others.[24] Summaries of CAPRI (Critical Assessment of PRedicted Interactions) results show that, while the accuracy of docking methods is improving, no single method is currently sufficient to successfully predict native-like assemblies in every test case.[6] Even top methods predict only 30%−58% of the correct interface in any given target.[25]

Detecting the correct interaction interface[26] is a fundamental challenge in protein docking. Some studies show that this interface exhibits a higher degree of evolutionary conservation than other regions on the molecular surface.[9,27] However, conserved residues may form a small part of interaction interfaces for various reasons.[26] Taken together, these findings suggest that the ranking of amino acids by the evolutionary conservation is a reasonable approach to locate the interaction interface, even if partially. Two representative methods are currently available for ranking amino acids by evolutionary conservation, the original evolutionary trace (ET) method,[28] and the joint evolutionary trace method (JET) based on ET.[9]

The extent to which interaction interfaces contain evolutionary-conserved amino acids is being employed as a scoring function to rank computed bound configurations.[29] Other methods have started to incorporate knowledge of the location of conserved residues to guide the search for bound configurations. For instance, the energy function employed for minimization can include terms that reward matching of surface regions with high conservation.[30]

In contrast to work in Ref. 30 that proposes an energy-based approach, this paper presents a geometry-based one. The JET method is used to rank amino acids of the protein monomers involved in the assembly by their degree of evolutionary conservation. This information is then employed to filter geometrically complementary surface regions on the monomers. Matching geometrically complementary and evolutionary-conserved regions results in rigid-body transformations that bring one monomer onto the other. Applying the transformation produces a bound dimeric configuration. Details on this approach and the different search procedures employed to select transformations are related in Sec. 2.

The proposed approach is applied to a list of 18 dimeric systems with known native structures. Extensive analysis is conducted in Sec. 3 in order to evaluate the dimeric configurations obtained for each system. Obtained results are compared to data reported by other labs in terms of lRMSD values of computed configurations to the known native dimeric structure (lRMSD refers to least root mean squared deviation and is a measure of the average distance between corresponding atoms in aligned configurations). Comparisons are made with a geometry-based method that does not employ evolutionary conservation[29] and the energy-based method in Ref. 30. On a majority of the systems, lower lRMSDs are reported by our approach. Our results further demonstrate that evolutionary conservation is a good predictor of functional interfaces. A detailed analysis is presented in Sec. 3.

2. Methods

2.1. *Definition of regions relevant for matching*

The JET method, which relies on multiple sequence analysis, is employed to identify conserved amino acids.[9] The JET score calculated for each amino acid can range from 0.0 (least conserved), to 1.0 (most conserved). We employ the iterative version, iJET, which repeats the analysis 50 times to obtain an average score for each amino acid. Amino acids determined to be on the surface (detailed later) that also have a 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." The active/passive designation is inspired by the work in Ref. 17.

Different thresholds of conservation scores are considered. The lower the threshold, the larger the surface area for docking and the higher the number of rigid-body transformations considered. The higher the threshold, the smaller the surface area and the more targeted the docking process. Our experiments (shown in Sec. 3) suggest that thresholds of 0.25−0.75 do not affect the accuracy of the method in reproducing the native assembly.

2.2. *Molecular surface representation and critical points*

Two representations are employed for the molecular surface. The Connolly surface is first employed to represent the solvent accessible surface area, calculated through the Molecular Surface Package (MSP).[32] This representation is dense. For each surface point, the Connolly representation maintains the 3D coordinate, the normal mode, and a numerical value to indicate the type of the surface. Types range from convex, saddle, to concave.

A sparse representation that simplifies the Connolly surface is calculated as in Ref. 11. This representation consists of a series of critical points. A critical point is defined as the projection of the center of gravity of a Connolly face on the molecular surface. Critical points are nicknamed *caps*, *pits*, or *belts* to correspond to convex, concave, or saddle faces. The collection of critical points cover key locations on the molecular surface to represent the shape of a molecule. In our approach, a critical point inherits the conservation score of its closest amino acid on the molecular surface.

2.3. *From critical points to active triangles*

Critical points are employed to define active triangles. A critical point p_1 with conservation score above a predefined threshold (we employ 0.5 in the implementation presented here) is selected first. Two more critical points, p_2 and p_3 (not necessarily conserved) are then selected from the molecular surface. Their selection satisfies both angle and distance constraints. The angle constraints ensure that the points are not collinear. Points p_2 and p_3 are also selected to lie no closer than 2 Å and no further than 5 Å from p_1. The minimum distance of 2 Å ensures that two points

are not on the same van der Waals (vdw) sphere of an amino acid. The maximum distance of 5 Å ensures that a triangle does not cover a lot of the molecular surface. The employed angle and distance parameter values are as in Ref. 12.

Our approach employs unique active triangles to limit the number of attempted transformations. A lexicographic ordering is first applied over the triangle's vertices. Given that triangles capture a small surface area, two triangles that share the first vertex in the lexicographic ordering essentially represent the same region in the molecular surface. Therefore, no two triangles are allowed to share their first vertex in the lexicographic ordering. Additionally, triangles are hashed by their center of mass. This reduces the number of unique active triangles even further. Given n critical points, ensuring satisfaction of the distance, angle, and the two uniqueness constraints described here results in fewer than n active triangles.

2.4. *From active amino acids to rigid-body transformations*

The calculation of a rigid-body transformation requires defining a local coordinate frame for each monomer. Active triangles are employed for this purpose. First, one of the monomers, let us refer to it as A, is arbitrarily selected as the "base" monomer. Let the other "moving monomer" be B. For each unique active triangle selected from A, a matching active triangle is selected from B. The features considered for matching are only geometric at this point, as in Ref. 10. Other physico-chemical features can be incorporated in later implementations. The two triangles selected for matching define two local coordinate frames. The rigid-body transformation aligns the frames by superimposing their origins and rotating B on A.

2.5. *Searching configurational space*

The main results shown in Sec. 3 are obtained with a simple exhaustive search procedure that essentially matches unique active triangles. Two settings are employed, as described in Sec. 3, in order to determine the effect of the number of triangles on the ability of this approach to obtain dimeric configurations with low lRMSDs to the known native structure. The obtained results show that each setting allows reproducing the known native structure within a few angstroms.

2.5.1. *Energy-guided probabilistic search*

When the number of unique active triangles is small, an enumeration-based procedure like the one described above, which essentially iterates over all the possible or a carefully selected subset of transformations, is feasible. However, as the size of the molecular assemblies grows and potentially the number of monomeric units as well (a direction we will explore in future work), enumeration-based search is infeasible. Probabilistic search procedures are needed instead. Here we present our first steps toward such a procedure that guides its search through energy.

A simple energy function is employed that models only Lennard-Jones (LJ) interactions between the involved monomers. The functional formula is that of the LJ potential in CHARMM22.[33] Only interactions among backbone atoms are modeled, since a refinement procedure like the one in Ref. 34 can then place side chains in favorable configurations. The LJ potential is employed to evaluate the feasibility of a generated configuration at a coarse-grained level of detail.

The search procedure generates configurations through a sampling-based process. It essentially samples the configurational space one rigid-body transformation at a time. An active critical point is selected uniformly at random from the molecular surface of one of the monomers, and an active triangle is then constructed as described earlier. A second active critical point is then sampled uniformly at random from the molecular surface of the second monomer and another active triangle is constructed. This process may be repeated until a geometrically complementary active triangle is obtained from the molecular surface of the second monomer. Matching of the sampled triangles results in a rigid-body transformation as detailed above, and the resulting transformation is applied to obtain a random dimeric configuration.

The LJ energy is employed to guide this probabilistic process similar to how the Metropolis criterion guides a Monte Carlo trajectory toward low-energy configurations.[35] It is important to point out that, while the current configuration in a Monte Carlo trajectory is the result of a perturbation of the previously obtained configuration, the configurations obtained by our search procedure are not guaranteed to reside in nearby regions on the configurational space. They are not consecutive points in a trajectory. In the classic employment of the Metropolis criterion, a perturbation is proposed to obtain a new configuration. The perturbation is evaluated according to the ΔE energetic difference after the perturbation. The perturbation is accepted (and the resulting configuration is added to the Monte Carlo trajectory) with probability $e^{-\Delta E/(K_b T)}$. The $K_b T$ term is a temperature scaling factor, where T is the effective chosen temperature of the simulation. The effective temperature determines the extent to which high-energy perturbations are allowed in the trajectory. Lower temperatures drive the search greedily toward low-energy configurations, but they may cause the trajectory to get stuck in a local minimum.

The success of a Metropolis Monte Carlo search in sampling low-energy configurations depends to a great extent on the fact that consecutive configurations in the trajectory are also near in configurational space. That is, a low-energy configuration will probably result in another low-energy configuration after a perturbation. In the probabilistic search procedure employed in this work, the rigid-body transformation may match different regions over the molecular surfaces. Hence, there is no dependence between consecutive conformations sampled by the search procedure. However, the Metropolis criterion is still useful, and it is employed here in order to bias the search toward conformations within an energetic range of the current energy. Rather than determine an arbitrary energetic cutoff for what are considered low-energy configurations (a value that depends on the specific system

under investigation), the Metropolis criterion allows the search procedure a natural way to bias toward low-energy configurations.

The selection of the effective temperature is important because it determines the extent to which the search will allow energetic increases. The results shown in Sec. 3 showcase two different temperatures (from high temperatures down to room temperature), selected from a proportional cooling schedule we have employed in previous work for Simulated Annealing search.[34] The selected temperatures showcase that medium-range effective temperatures allow balancing the search toward low-energy but diverse configurations. Other modifications that can be made to this proof-of-concept search procedure for a more effective exploration of the configurational space are discussed in Sec. 4.

3. Results

Our experiments are carried out on a 2.66 GHz Opteron processor with 8GB of memory. We select 18 different dimers with known native structures as our systems of study. These systems are selected because they cover different functional classes and have been investigated by other computational groups, as well. Results obtained after the experiments summarized below make the case that on all selected systems of study the approach presented in this paper is able to reproduce the native structure within a few angstroms in a feasible amount of time.

We present five sets of results. First, we present detailed results on two systems selected from our set of 18 to showcase the effect of different conservation thresholds on the number of active triangles and proximity of computed configurations from the known native structures. These results justify our employment of a 0.5 conservation threshold on the rest of the dimers. The second set of results shows the lowest lRMSD obtained for each system from the corresponding native structure. These results are compared to those published by two other labs. The lowest lRMSD configurations obtained by the method are also shown on select systems. In the third set of results, nine dimers are selected to showcase the effect of different schemes for construction of active triangles on the number of triangles and computational time. These results make the case that the method is able to achieve low lRMSD values while improving computational cost. In the fourth set of results, energetic refinement is carried out on two selected systems to analyze resulting energy values and rank on a subset of configurations selected from those computed by the exhaustive procedure. Finally, configurations obtained with our energy-guided probabilistic search procedure are presented and analyzed in the fifth set of results.

3.1. *Detailed analysis of representative systems*

Two signaling proteins are selected to measure the effect of the conservation threshold on the lowest lRMSDs obtained. The first system consists of the vascular endothelial growth factor and FMS-like tyrosine kinase-1. The native structure has

Table 1. Effect of varying the conservation threshold.

PDB ID	Threshold	No. of 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

PDB ID 1FLT. The second system, the Nerve Growth Factor/TRKA receptor, is responsible for the development and maintenance of the sympathetic and sensory nervous systems. Its native structure can be found under PDB entry 1WWW.

The following experiment is performed on each system. Three conservation thresholds, 0.25, 0.5, and 0.75 are employed to define three sets of active triangles. The exhaustive procedure is employed to obtain dimeric configurations with each threshold, essentially resulting in three sets of configurations. The lowest lRMSD is recorded for each setting, and these are reported for each system in Table 1. The first column shows the PDB ID of each dimer and their chains in brackets. The second column shows the varying conservation threshold. The third column shows the number of active triangles defined on the reference chain under each conservation threshold. The last column shows the lowest lRMSDs obtained in each setting.

Table 1 shows that the number of active triangles goes down as the conservation threshold increases, as expected, with no significant changes to the lowest lRMSD. The detailed distribution of lRMSDs from the native structure for configurations computed with a threshold of 0.5 is shown for each of the two selected systems in Figs. 1(a1)−1(b1). The distributions show that the method produces many configurations with less than 5 Å of the native structure. Given these results, the threshold of 0.5 is employed to obtain configurations for all other systems in the rest of the results.

3.2. *Comparison of proximity to known native structures with results obtained by other methods*

In this set of results, configurations computed with the exhaustive search procedure for each of the 18 dimers are compared to the respective native structures in terms of lRMSD. lRMSDs are calculated over C_α atoms. The lowest lRMSD configuration is shown for selected systems in Fig. 4. The amino-acids with conservation scores above 0.5 that are part of the contact interface are also shown. Two amino acids are considered in contact if their Euclidean distance is no larger than 4.0 Å. Figure 4 illustrates that the lowest lRMSD structures are those where the contact interface overlaps well with the predicted interaction interface.

The lowest lRMSD value obtained on each system is compared with those reported by two other methods, BUDDA,[31] and the method in Ref. 30. BUDDA is a geometric method that relies on geometric hashing, whereas the method in Ref. 30 is

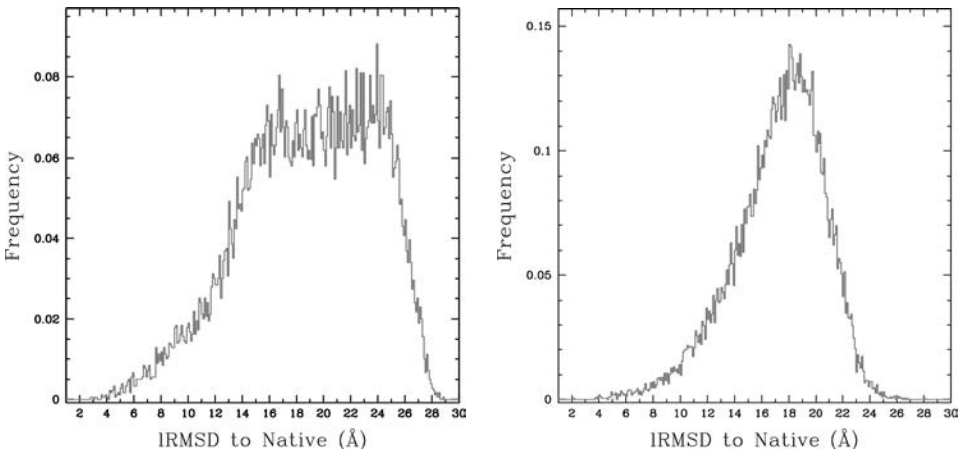


Fig. 1. (a1) and (b1) Histograms (for system with PDB ID 1FLT left and 1WWW right) show distributions of lRMSDs from the native structures for dimeric configurations computed with our exhaustive protocol and a conservation threshold of 0.5.

an energy-based method that incorporates ET conservation scores in the energy function. The comparison is shown in Table 2.

When differences in lowest lRMSDs achieved by different docking methods are within 2 Å these results are considered equivalent, since energetic refinement can potentially reduce these differences. On about 13 of the 18 dimers, the approach presented here achieves similar low lRMSDs to BUDDA and the method in Ref. 30. On the remaining systems, the approach outperforms these other two methods. This result is encouraging, as it shows that better or equivalent results can be obtained. While the energy-based method in Ref. 30 relies on long energy minimizations, the BUDDA method in Ref. 31 exhaustively matches all geometrically complementary triangles without making use of putative interaction interfaces. The next set of results shows how the number of triangles employed for matching affects computation time.

The results in Table 3 measure the effect of the number of active triangles on both time and accuracy. Nine systems are selected for this purpose. Two different settings are employed. In the first setting, all unique active triangles from the base monomer are employed to define transformations, resulting in the data shown in Table 2. In the second setting, the number of unique active triangles is reduced by roughly one-third by essentially ensuring that no critical point is used more than once in the construction of active triangles over the molecular surface.

Table 3 shows the savings in the number of triangles and computation time over the setting in Table 2. Column 2 in Table 3 shows for each of the monomers the ratio of the number of triangles in this setting over the number in Table 2. Column 3 shows the ratio of the time requirements here over the time requirements in Table 2. Finally, column 4 shows the difference in lRMSD of this setting from that shown in

Table 2. Lowest lRMSDs by our method, reported in column 5, are compared to those published by Polak *et al.* and Kanamori *et al.* reported in columns 3, and 4, respectively. Size in column 2 refers to the number of atoms in each chain.

PDB ID (Chains)	Size	Ref. 31 (Å)	Ref. 30 (Å)	Here (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 (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	**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**
1QAV (A,B)	663, 840	1.44	N/A	1.04
1G4Y (B,R)	682, 1156	0.83	N/A	2.31
1BDJ (A, B)	979, 919	15.4	N/A	**2.65**
1CSE (E,I)	1920, 522	0.73	N/A	1.49

Table 3. Effect of number of active triangles on time and lowest lRMSDs.

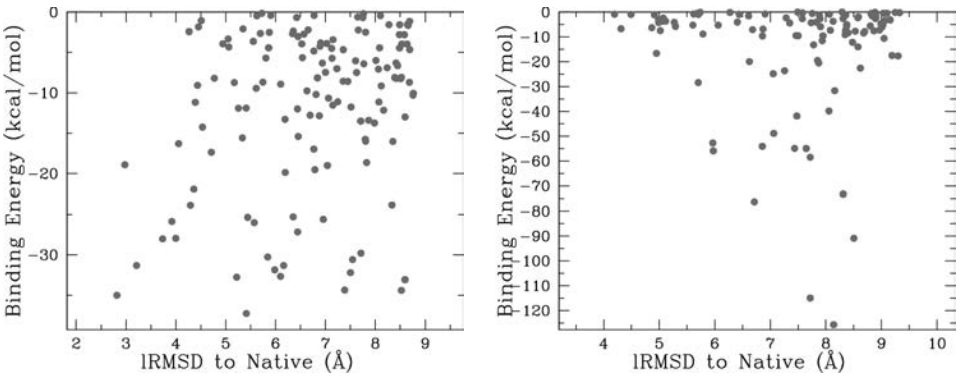
PDB ID (Chains)	No. of triangles ratio	Time ratio	lRMSD Diff. (Å)
1C1Y (A, B)	1:2.80, 1:2.86	1:6.57	1.02
1G4U (R, S)	1:2.89, 1:2.88	1:6.44	−0.72
1DS6 (A, B)	1:2.86, 1:2.88	1:7.16	1.42
1TX4 (A, B)	1:2.88, 1:2.91	1:6.44	0.30
1WWW (W, Y)	1:2.94, 1:2.94	1:9.57	−0.05
1FLT (V, Y)	1:3.23, 1:2.56	1:10.5	1.69
1IKN (A, C)	1:2.90, 1:2.89	1:7.93	−0.89
1IKN (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

Table 2. The results in Table 3 make the case that this construction of active triangles achieves similar low lRMSDs while improving feasibility.

3.3. *Energetic refinement of select configurations*

Computed configurations for two representative systems with PDB IDs 1FLT and 1WWW are now selected for refinement. Out of all dimeric configurations computed for each system with our exhaustive procedure, the 500 configurations with lowest lRMSDs from the known respective native structures are refined with Firedock.[13] Firedock is chosen due to its fast interaction refinement protocol, as a first step toward detailed refinement of computed structures.

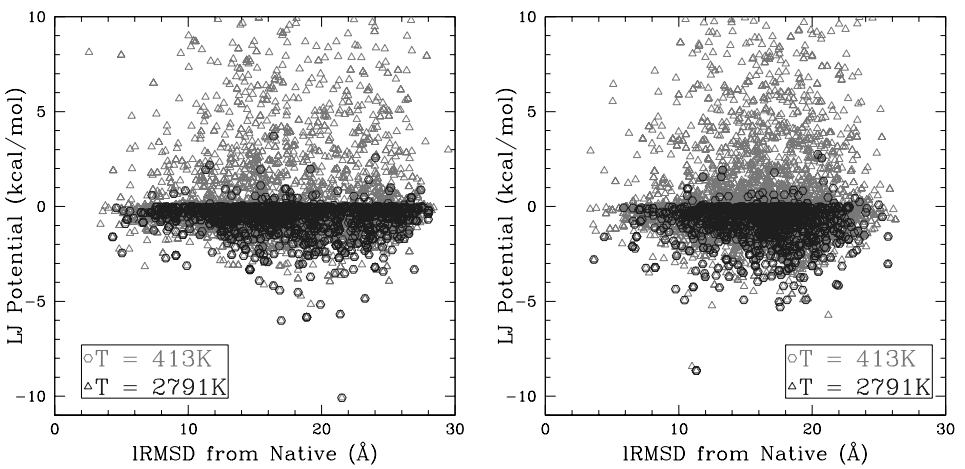
Figures 2(a)−2(b) plots for each system the Firedock-reported binding energy values for the refined configurations versus the lRMSDs of these configurations from



(a) (b)

Fig. 2. Firedock energies of refined structures are plotted against lRMSDs from the native structure. Results for 1FLT are shown in (a) and 1WWW in (b).

the known native structure. Only negative energy values are shown. The results in Fig. 2 show that many of the lowest-energy structures are also low in lRMSD from the corresponding native structure. This suggests that short refinements may allow detecting low-lRMSD configurations by selecting a few configurations with lowest refinement energies for prediction. These in turn may be refined in further detail in order to recover the native structure among the lowest-energy ones.



(a) (b)

Fig. 3. The LJ potential energies of conformations sampled by the energy-guided search are plotted against lRMSD values from the known native structure. Two different configurational ensembles are shown, corresponding to two different values of temperatures shown in the legend. The results obtained with the different temperatures are superimposed over one another. Results in (a) are for the system with PDB ID 1FLT, whereas those in (b) are for the system with PDB ID 1WWW.

3.4. *Analysis of configurations obtained with energy-guided probabilistic search*

We present here results obtained by our proof-of-concept probabilistic search procedure that employs the Metropolis criterion to bias its exploration toward low-energy configurations. Results are presented for the two selected systems with PDB

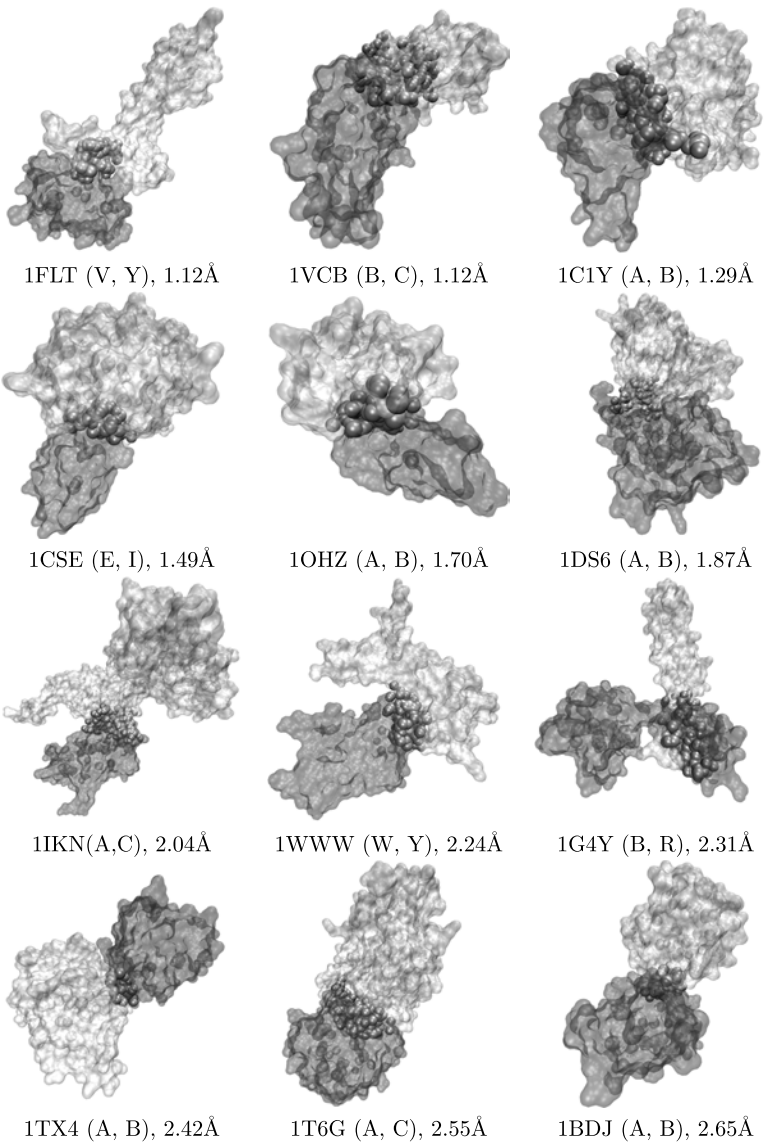


Fig. 4. Lowest-lRMSD structures and the actual lRMSD achieved are shown for 12 selected systems. Chains are drawn in different shades of gray in transparent. Conserved amino acids in contact with one another are drawn in opaque.

IDs 1FLT and 1WWW in terms of LJ energies vs. lRMSDs from the known native structure for sampled configurations. The results are shown in Fig. 3.

Figure 3 superimposes the analysis of two different configurational ensembles obtained with the effective temperatures shown in Fig. 3. The superimposition illustrates that high temperatures result in a broader ensemble of configurations in terms of energy values. However, lowering the temperature allows focusing the search to low-energy configurations. These results suggest that medium-range temperatures may provide a good compromise and focus the search toward low-energy configurations while allowing to obtain low lRMSDs to the native structure. It is worth noting that the size of the configurational ensemble that can be obtained at higher temperatures is larger than that obtained at lower temperatures, as it becomes harder to satisfy the Metropolis criterion at lower temperatures. While the number of attempts is kept fixed at around 100,000, the actual number of configurations accepted depends on the Metropolis criterion.

At this degree of resolution, where the only term modeled in the binding energy is the LJ potential, no significant correlation is expected between low energies and low lRMSD values. However, even at this resolution, the energetic bias in the search allows avoiding configurations with unfavorable interactions. As our discussion in Sec. 4 points out, more sophisticated coarse-grained energy functions that incorporate additional interactions will be considered in future work.

4. Discussion

We have presented a geometry- and evolutionary-guided approach to protein docking that focuses the search for bound configurations through rigid-body transformations that match surface regions deemed to be both geometrically complementary and evolutionary conserved. Our results show that this focusing narrows the configurational search space and allows obtaining low-lRMSD configurations for many protein systems. The different search procedures employed here illustrate both the relative ease at obtaining low-lRMSD configurations with little computational cost and the promise of the approach in tackling larger systems with probabilistic search.

Like Zdock, the proposed approach can be employed as the first stage in docking software. The obtained configurations can be clustered and ranked in a second stage. Low-scoring configurations of selected clusters can then be further refined in a third stage. The refinement can be carried out with protocols like the one used in Ref. 13 or with more detailed and powerful protocols like the one in Ref. 34.

Our ongoing work focuses in combining the approach presented here with the refinement procedure presented in Ref. 34. In future work we will consider different established procedures for predicting interaction interfaces. We will additionally investigate more powerful probabilistic search procedures and their applications beyond dimers to multimeric assemblies.

Acknowledgments

We thank members of the Haspel and Shehu labs for useful comments on this work.

References

1. Lensink MF, Mendez R, Wodak SJ, Docking and scoring protein complexes: CAPRI 3rd edition, *Proteins: Struct Funct Bioinf* **69**(4):704−718, 2007.
2. Gray JJ, High-resolution protein-protein docking, *Curr Opinion Struct Biol* **16**(2):183−193, 2006.
3. Mendez R, Leplae R, Lensink MF, Wodak SJ, Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures, *Proteins: Struct Funct Bioinf* **60**(2), 2005.
4. Vajda S, Kozakov D, Convergence and combination of methods in protein-protein docking, *Curr Opinion Struct Biol* **19**:164−170, 2009.
5. Smith GR, Sternberng MJE, Prediction of protein-protein interactions by docking methods, *Curr Opinion Struct Biol* **12**(1):28−35, 2002.
6. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR, Towards the development of universal, fast and highly accurate docking/scoring methods: A long way to go, *British J Pharmacol* **153**(S1):S7−S27, 2009.
7. Halperin I, Ma B, Wolfson H, Nussinov R, Principles of docking: An overview of search algorithms and a guide to scoring functions, *Proteins* **47**(4):409−443, 2002.
8. Camacho CJ, Gatchell DW, Kimura SR, Vajda S, Scoring docked conformations generated by rigid-body protein-protein docking, *Proteins* **40**(1):525−537, 2000.
9. Engelen S, Ladislas AT, Sacquin-More S, Lavery R, Carbone A, A large-scale method to predict protein interfaces based on sequence sampling, *PLoS Comp Bio* **5**(1):e1000267, 2009.
10. Wolfson HL, Rigoutsos I, Geometric hashing: An overview, *IEEE Comp Sci Engineering* **4**(4):10−21, 1997.
11. Norel R, Lin SL, Wolfson HJ, Nussinov R, Examination of shape complementarity in docking of unbound proteins, *Proteins* **36**(3):307−317, 1999.
12. Fischer D, Lin SL, Wolfson HL, Nussinov R, A geometry-based suite of molecular docking processes, *J Mol Biol* **248**(2):459−477, 2005.
13. Andrusier N, Nussinov R, Wolfson HJ, Firedock: Fast interaction refinement in molecular docking, *Proteins: Struct Funct Bioinf* **69**(1):139−159, 2007.
14. OTóole N, Vakser IA, Large-scale characteristics of the energy landscape in protein-protein interactions, *Proteins: Struct Funct Bioinf* **71**(1):128−132, 2007.
15. Murphy J, Gatchell DW, Prasad JC, Vajda S, Combination of scoring functions improves discrimination in protein-protein docking, *Proteins* **53**(4):840−854, 2003.
16. Chen R, Li L, Weng Z, ZDock: An initial-stage protein-docking algorithm, *Proteins: Struct Funct Bioinf* **52**(1):80−87, 2003.
17. Dominguez C, Boelens R, Bonvin AMJJ, Haddock: A protein-protein docking approach based on biochemical orbiophysical information, *J Am Chem Soc* **125**:1731−1737, 2003.
18. Comeau SR, Gatchell DW, Vajda S, Camacho CJ, ClusPro: A fully automated algorithm for protein-protein docking, *Nucl Acids Res* **32**(S1):W96−9, 2004.
19. D. Duhovny-Schneidman, Inbar Y, Nussinov R, Wolfson HJ, PatchDock and SymmDock: Servers for rigid and symmetric docking, *Nucl Acids Res* **33**(S2):W363−W367, 2005.
20. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ, Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies, *J Phys Biol* **2**: S156−S165, 2005.

21. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ, Prediction of multimolecular assemblies by multiple docking, *J Mol Biol* **349**(2):435−447, 2005.
22. Mashiach E, Nussinov R, Wolfson HJ, Fiberdock: Flexible induced-fit backbone refinement in molecular docking, *Proteins: Struct Funct Bioinf* **78**(6):1503−1519, 2010.
23. Lyskov S, Gray JJ, The RosettaDock server for local protein-protein docking, *Nucl Acids Res* **36**(S2):W233−W238, 2008.
24. Schneidman-Duchovny D, Inbar Y, Nussinov R, Wolfson HJ, Geometry based flexible and symmetric protein docking, *Proteins: Struct Funct Bioinf* **60**(2):224−231, 2005.
25. Lensink MF, Wodak SJ, Blind predictions of protein interfaces by docking calculations in CAPRI, *Proteins: Struct Funct Bioinf* **78**(15):3085−3095, 2010.
26. Fernandez-Recio J, Prediction of protein binding sites and hot spots, *Wiley Interdisciplinary Reviews: Comput Mol Chem* **1**(5):680−698, 2011.
27. Kim WK, Henschel A, Winter C, Schroeder M, The many faces of protein-protein interactions: A compendium of interface geometry, *PLoS Comp Bio* **2**:e124, 2006.
28. Lichtarge O, Bourne HR, Cohen FE, An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol* **257**(2):342−58, 1996.
29. Tress M, de Juan D, Graña O, Gómez MJ, Gómez-Puertas P, González JM, López G, Valencia A, Scoring docking models with evolutionary information, *Proteins: Struct Funct Bioinf* **60**(2):275−280 2005.
30. Kanamori E, Murakami Y, Tsuchiya Y, Standley DM, Nakamura H, Kinoshita K, Docking of protein molecular surfaces with evolutionary trace analysis, *Proteinssfb* **69**:832−838, 2007.
31. Polak V, Budda: backbone unbound docking application, Master's thesis, Tel-Aviv University, Tel-Aviv, Israel, 2003.
32. Connolly ML, Analytical molecular surface calculation, *J Appl Cryst* **16**(5):548−558, 1983.
33. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M, CHARMM: a program for macromolecular energy, minimization, and dynamics calculations, *J Comput Chem* **4**(2):187−217, 1983.
34. Akbal-Delibas B, Hashmi I, Shehu A, Haspel N, Refinement of docked protein complex structures using evolutionary traces, *Comput Struct Biol Workshop (CSBW)*, pp. 400−404, 2011.
35. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E, Equation of state calculations by fast computing machines, *J Chem Phys* **21**(6):1087−1092, 1953.
36. Shehu A, Kavraki LE, Clementi C, Multiscale characterization of protein conformational ensembles, *Proteins: Struct Funct Bioinf* **76**(4):837−851, 2009.

Irina Hashmi is pursuing her Ph.D. in Computer Science at George Mason University. She received her B.S. in 2007 and M.S. in 2009 in Computer Science and Engineering from University of Dhaka, Bangladesh. Her research interests are in computational structural biology, evolutionary computation, and reversible computing. Her current work focuses on high-dimensional search space problems related to protein docking.

Bahar Akbal-Delibas is pursuing her Ph.D. in Computer Science at University of Massachusetts Boston. She received her M.S. in Computer Science from University of Massachusetts Boston and B.S. in Computer Engineering from Fatih University in Turkey. Her research interests are in better understanding protein−protein interactions through modeling of conformational changes in protein chains and protein-based assemblies.

Nurit Haspel is an Assistant Professor in the Department of Computer Science at the University of Massachusetts Boston. She received her B.Sc. in Chemistry and Computer Science, her M.Sc in Human Genetics, and Ph.D in Computer Science from Tel Aviv University in Israel. She was trained as a postdoctoral research associate at Rice University in Houston, TX. Haspel's research contributions are in structural bioinformatics with emphasis on studying the structure, function, and dynamics of proteins and protein complexes, as well as the design of novel nano-structures and the study of molecular self-assembly. Her research is funded in part by the National Science Foundation.

Amarda Shehu is an Assistant Professor in the Department of Computer Science at George Mason University. She holds affiliated appointments in the Department of Bioinformatics and Computational Biology and the Department of Bioengineering at George Mason University. She received her B.S. in Computer Science and Mathematics from Clarkson University in Potsdam, NY, and her Ph.D. in Computer Science from Rice University in Houston, TX, where she was an NIH fellow of the Nanobiology Training Program of the Gulf Coast Consortia. Shehu's research contributions are in computational structural biology, biophysics, and bioinformatics, with a focus on issues concerning the relationship between sequence, structure, dynamics, and function in biological molecules. Shehu is a recent recipient of an NSF CAREER award for her research on probabilistic search algorithms for protein conformational spaces.