Chapter 14

DOCK/PIERR: Web Server for Structure Prediction of Protein–Protein Complexes

Shruthi Viswanath, D.V.S. Ravikant, and Ron Elber

Abstract

In protein docking we aim to find the structure of the complex formed when two proteins interact. Protein–protein interactions are crucial for cell function. Here we discuss the usage of DOCK/PIERR. In DOCK/PIERR, a uniformly discrete sampling of orientations of one protein with respect to the other, are scored, followed by clustering, refinement, and reranking of structures. The novelty of this method lies in the scoring functions used. These are obtained by examining hundreds of millions of correctly and incorrectly docked structures, using an algorithm based on mathematical programming, with provable convergence properties.

Key words Protein–protein docking, FFT-based docking, Knowledge-based potential, Atomic potential, Residue potential, Scoring function, Mathematical programming, Refinement and reranking

1 Introduction

The DOCK/PIERR protein docking server predicts the quaternary structure of the complex formed by two proteins, given their individual tertiary (3D) structures. The structures of the complexes can be useful in obtaining molecular details of protein function and biochemical pathways. Examples are interactions between an enzyme and its inhibitor or between an antibody and antigen. Further, given structural details of the interface between proteins, experiments can be designed to alter the strength and specificity of binding by introducing mutations at the interface. Finally, complexes can also aid in structure-based drug design, where designed small molecules can inhibit the interaction between two proteins by preferentially binding to one partner and thus affecting the pathways involving them [1, 2].

Protein–protein docking algorithms in general work in two stages. In the first stage, various possible conformations of the complex are examined and scored, treating the proteins as rigid bodies. The most frequently used methods for the search stage are

Daisuke Kihara (ed.), *Protein Structure Prediction*, Methods in Molecular Biology, vol. 1137,
DOI 10.1007/978-1-4939-0366-5_14,

Fast Fourier Transforms [3–5], which enables fast exhaustive sampling of the search space, Monte-Carlo [6, 7] and Geometric Hashing [8]. In the second stage of refinement and reranking, some limited flexibility in the models is introduced through techniques like energy minimization [5, 9] and Monte-Carlo [7, 10], and structures are reranked with fine-grained scoring functions.

In DOCK/PIERR, the conformational space of complexes is sampled exhaustively using Fast Fourier Transforms, and the encountered structures are scored using a residue scoring function. This is followed by side-chain rearrangement of the proteins at the docking site and a short energy minimization. The structures are then rescored using a combination of residue and atomic scores. The novelty of this algorithm and its accuracy lies in the scoring functions used. These scoring functions are parameterized using mathematical programming [11] and provably optimal structural SVM algorithms [12]. Hundreds of millions of models encountered from docking hundreds of complexes are used in the learning, and the models include both correctly and incorrectly docked structures. Constraints that stipulate that the energy of a misdocked structure should be higher than the energy of a correctly docked structure are derived from these models. The set of constraints derived from all the models in the learning set is solved through methods like linear programming or structural SVMs, to produce the parameters of the scoring function. The docking algorithm has been tested on docking benchmark datasets and is found to perform comparable to the state-of-art docking algorithms [13], ranking fourth in the server category in the CAPRI assessment of 2013 [14].

2 Materials

2.1 Input

The server takes as input the PDB structures of the two proteins to predict the structure of the complex. *See* **Note 1** on details of how to prepare the PDB structures.

2.2 Program Description

One of the proteins (called receptor) is kept fixed. All possible rigid rotations and translations of the second protein (called ligand) with respect to the receptor are explored using Fast Fourier Transforms. Each conformation is scored using a linear combination of an interface residue-contact based scoring function, PIE, and a van der Waals-like term for shape complementarity. The top scoring models are then clustered and filtered for interface clashes. The top 1,000 models are then refined. The refinement involves side-chain remodeling of the interface residues using rotamers (SCWRL4 [15]) and a short rigid energy minimization in vacuum with the OPLS force field using the molecular dynamics package MOIL [16]. The last procedure removes bad contacts and makes the structures more chemically reasonable. The refined structures

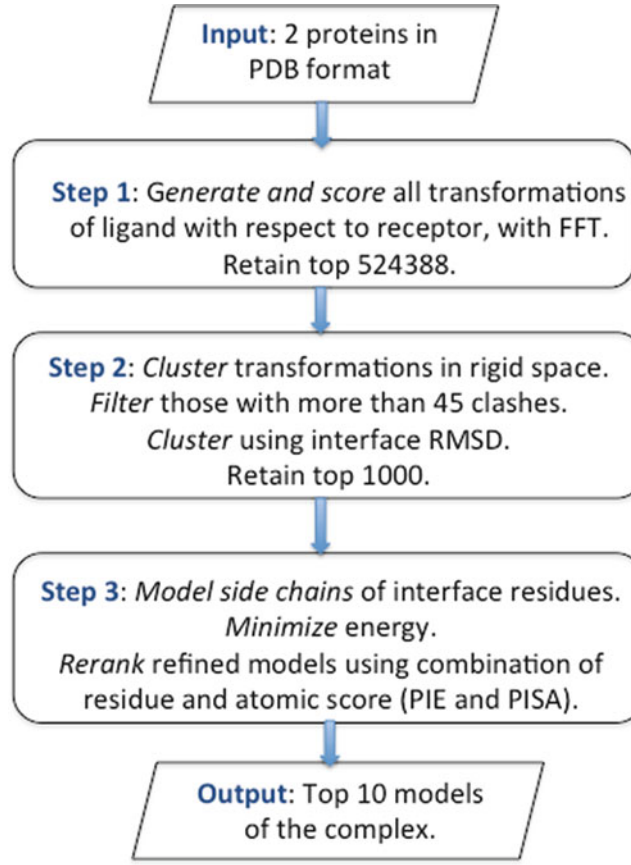

Fig. 1 Flowchart representing steps taken for docking two proteins using DOCK/PIERR

are then reranked using a combination of the residue potential, PIE, and an atomic potential, PISA, that was trained on refined models. It is to be noted that the adjustments during refinement are very small and typically of the order of ~0.1 A. Nevertheless they remove bad contacts and hence significantly improve the rescoring. The ten best ranked models of the complex are then made available as server output to the user. On tests on standard benchmarks and independent test sets, the algorithm as described above, obtains a near-native structure in the top ten models about 40–60 % of the time, and is comparable in accuracy to other leading docking algorithms. Figure 1 explains in detail the steps taken by the server to dock two proteins. For further details regarding the algorithm the reader is referred to [12, 13, 17].

2.3 Server Availability

The server is available at http://clsb.ices.utexas.edu/web/dock.html. It is implemented using HTML frontend and a PHP backend. The PHP script sends a mail to the server, which launches the

docking jobs on 16 cores (Intel Xeon X5460, 3.16 GHz) of a Linux cluster at the University of Texas at Austin. While the entire docking package is in C++, the server also uses external programs such as SCWRL4 and MOIL.

2.4 Scoring Function Downloads

A user, who wishes to rank a set of structures obtained from a single server run or multiple-related runs, can also download and use our scoring functions, PIE (residue-based) and PISA (atomic). The source code and Linux executables for these are provided at http://clsb.ices.utexas.edu/web/dock_details.html. Scoring a model of a complex simply requires the structure of the complex in PDB format and the receptor and ligand chain names.

3 Methods

1. The server requires as input the structures of the two proteins in PDB format. The PDB files can be simply uploaded and submitted. *See* **Note 1** for potential sources of error in the input. Also *see* **Note 2** for cases where the user has only the sequence and not structure for an input protein.
2. For computational efficiency, the larger of the two proteins should be uploaded in the receptor field and the smaller one in the ligand field.
3. After submitting, the user gets a confirmation email with the job number. This job number denotes the submission ID and is referenced in the output email.
4. Jobs generally take about 4–5 h to complete. They may take more time if the proteins are large, i.e., longer than 400 residues, or if the server is experiencing high traffic.
5. Once the job is completed, a zipped file containing the ten best scoring docked conformations in PDB format is emailed to the user. The name of the zipped file corresponds to the submission ID or job number that the user was provided with, during submission. The chain names in the output PDB are alphabetically ordered, starting from the receptor chains.
6. Visualization of the models of the complexes can be performed with any structure visualization software like PyMol [18].
7. The accuracy of the docking method is between 40 % and 60 % currently, i.e., a near-native structure, a structure within 4 Å interface RMSD to the native, is in the top 10 docked structures about 40–60 % of the time. Cases where this docking method can be inaccurate are when the actual complex has a small number of contacts. Since (on the average) more contacts mean lower energy in our model, complexes with a small number of contacts are missed.

4 Case Studies

An early version of the docking software has been used previously in a biological study to suggest oligomeric conformations of a four-domain orange-fluorescent protein (Ember) [19]. Below we describe a case study of docking using DOCK/PIERR.

Unbound docking of Textilinin-1, a serine protease inhibitor with bovine trypsin. [*PDB 3D65*]

Here we dock bovine trypsin with the serine protease inhibitor, Textilinin-1, derived from the Australian Common Brown snake. This complex has been experimentally determined (PDB 3D65) [20]. Trypsin is an enzyme found in the pancreas and involved in proteolysis and digestion, while the protease inhibitor binds to trypsin to down-regulate its enzymatic activity.

To dock trypsin with its inhibitor, we perform unbound docking. That is, we model the tertiary structure of one or both of the constituent proteins using their homolog structures as templates. We then perform docking on the homology-modeled proteins. The trypsin molecule is chain E of the complex 3D65 and 223 residues long. The inhibitor molecule is chain I of 3D65 and 57 residues long. We use the structure of trypsin as in the bound form for docking, i.e., chain E of 3D65. To model the inhibitor, we perform a search for homologs using PSI-BLAST [21], searching the PDB database for structures homologous to chain I of PDB 3D65. We find that the chain I of PDB entry 3BTM is a good match, with *E*-value of 9×10^{-13} and sequence identity of 44.8 %. We next obtain pairwise alignments between the sequences of 3D65, chain I and 3BTM, chain I. A pairwise alignment can be obtained using dynamic programming, and is implemented in alignment servers such as the EMBOSS server (http://www.ebi.ac.uk/Tools/psa/). We then use the program Modeller [22, 23] to produce the structure of the inhibitor from the template structure of 3BTM, and the pairwise sequence alignment between 3BTM chain I and 3D65 chain I. We use the new PDB file obtained from Modeller for docking. Note that Modeller produces PDB files with no chain names by default, and hence it is recommended to add chain names to the PDB files before submitting files to the docking server.

We then submit the PDB files for the trypsin in the receptor field and the newly obtained inhibitor structure in the ligand field of the DOCK/PIERR server submission form. Upon completing the docking, we obtain the top ten models of the complex. Figure 2 shows the input proteins we docked, and Fig. 3 is a superposition of one of the top ten models obtained from the DOCK/PIERR server with the actual complex, 3D65. The model has an interface RMSD of 3.63 Å to 3D65.

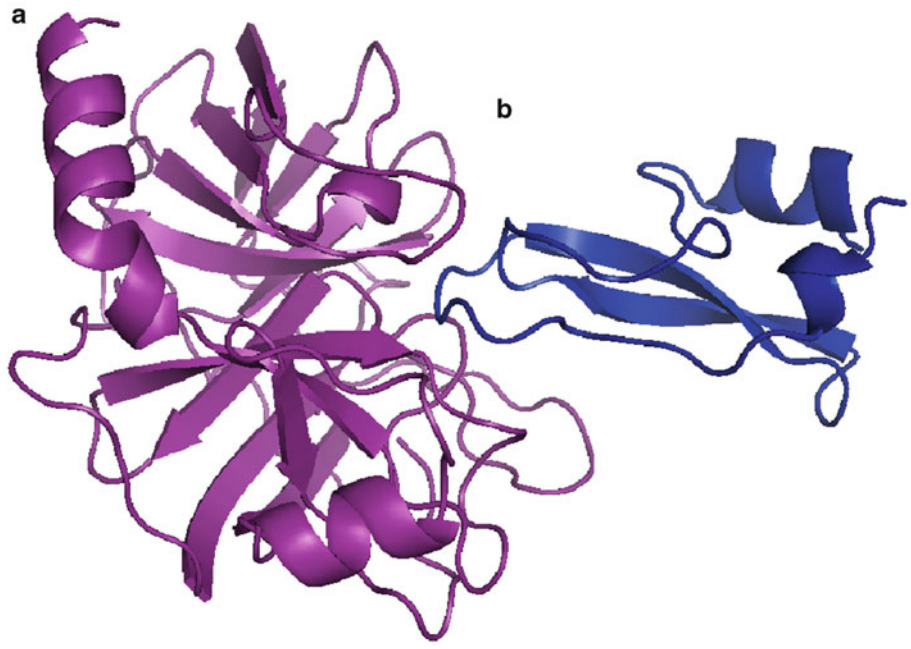

Fig. 2 (**a**) Chain E (bovine trypsin) and (**b**) Chain I (Textilinin-1, serine protease inhibitor) of complex 3D65 to be docked. These structures are inputs to the DOCK/PIERR server

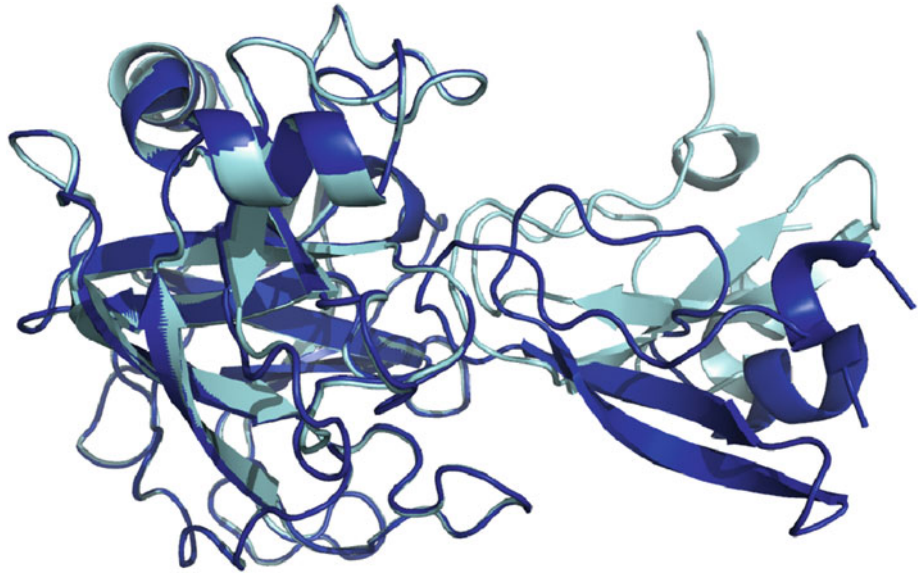

Fig. 3 One of the top ten models produced by DOCK/PIERR server (*cyan*) superposed with the native structure of the 3D65 trypsin–inhibitor complex (*blue*). The model has an interface RMSD of 3.63 Å to 3D65. The difference in tertiary structures between the native PDB and the model for the inhibitor is due to unbound docking (Color figure online)

5 Notes

1. The most common problems with the input PDB files that cause server failures are as follows:
   (a) Missing atoms in the PDB files. The missing atoms may be side-chain atoms or main chain atoms. For missing side-chain atoms, it is recommended to use the program SCWRL [15] or a similar program for side-chain placement. For missing main chain atoms, DOCK/PIERR is able to dock the proteins but the structures may not be refined, since the molecular dynamics program used in the refinement stage needs the coordinates of all the atoms. Failure to refine the models might result in less than optimal docking results.
   (b) Nonstandard atom names. These might be ignored in the initial docking stage and the structures may not be refined, as our molecular dynamics program is not capable of dealing with nonstandard atoms. These too might lead to suboptimal docking results if left unchanged.
   (c) Nonstandard residue names. Sometimes, some residues have nonstandard amino acid names. In many of these cases, the residue is chemically modified and the name is adjusted. For example the residue HIS is named differently as HSD, HSE, HSP depending on the protonation state. In such a case, the user is advised to rename such residues to their standard label.
   (d) Negative residue numbering. Some structures use negative residue numbers, for example when a tail is added to the native N-terminal. This causes problems during the refinement stage and the user is advised to index all residues with positive numbers.
   (e) Missing chain names for either protein, or identical chain names for both proteins. These can cause problems in the initial stages of docking. Also if the receptor and/or ligand have multiple chains, care must be taken to make all chain names between the receptor and ligand nonidentical. For example, if the receptor has chains A, B and the ligand has chain A, it is recommended to rename the ligand chain to C.
   (f) If a PDB file containing multiple NMR models is submitted, only the first model is considered for docking.
   (g) Some atoms in the PDB have multiple locations specified, using the alternate location field in the PDB. The docking program ignores the alternate locations. It also ignores HETATM records.

(h) Both the receptor and ligand molecules need to be proteins. Other molecules like DNA/RNA, or small molecule compounds are not supported as our scoring functions are tailored for protein interactions.

(i) If the user deals with very flexible peptides, it is recommended to dock one flexible conformation of the peptide at a time. This is because our algorithm performs rigid docking. At present it does not combine docking and internal motions.

2. It is most straightforward if one has the PDB structures of the two proteins to be docked. But if one just has the sequence for one (or both) input proteins, then structural model(s) need to be built from sequence. Examples of servers that produce models from the sequence include LOOPP [24, 25] (http://clsb.ices.utexas.edu/loopp/web/) and i-TASSER [26] (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). If one has already a template structure on which to model the sequence, homology modeling packages such as Modeller [22, 23] (http://salilab.org/modeller) can be used to predict the structure. Note that using different templates or different modeling methods for structure prediction can affect docking results.

Acknowledgements

The authors acknowledge funding from NIH grant GM59796 and Welch grant F-1783.

References

1. Gray JJ (2006) High-resolution protein-protein docking. Curr Opin Struct Biol 16(2):183–193. doi:10.1016/J.Sbi.2006.03.003
2. Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure. Q Rev Biophys 41(2):133–180. doi:10.1017/S0033583508004708
3. Chen R, Li L, Weng ZP (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins Struct Funct Genetics 52(1):80–87. doi:10.1002/Prot.10389
4. Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004) ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res 32:W96–W99. doi:10.1093/Nar/Gkh354
5. Tovchigrechko A, Vakser IA (2005) Development and testing of an automated approach to protein docking. Proteins 60(2):296–301. doi:10.1002/Prot.20573
6. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331(1):281–299. doi:10.1016/S0022-2836(03)00670-3
7. Wang C, Bradley P, Baker D (2007) Protein-protein docking with backbone flexibility. J Mol Biol 373(2):503–519. doi:10.1016/J.Jmb.2007.07.050
8. Duhovny D, Nussinov R, Wolfson HJ (2002) Efficient unbound docking of rigid molecules. Lect Notes Comput Sci 2452:185–200
9. Li L, Chen R, Weng ZP (2003) RDOCK: refinement of rigid-body protein docking predictions. Proteins Struct Funct Genetics 53(3):693–707. doi:10.1002/Prot.10460

10. Wang C, Schueler-Furman O, Baker D (2005) Improved side-chain modeling for protein-protein docking. Protein Sci 14(5): 1328–1339. doi:10.1110/Ps.041222905
11. Wagner M, Meller J, Elber R (2004) Large-scale linear programming techniques for the design of protein folding potentials. Math Program 101(2):301–318. doi:10.1007/S10107-004-0526-7
12. Ravikant DVS, Elber R (2011) Energy design for protein-protein interactions. J Chem Phys 135(6):065102. doi:10.1063/1.3615722
13. Viswanath S, Ravikant DVS, Elber R (2013) Improving ranking of models for protein complexes with side chain modeling and atomic potentials. Proteins 81(4):592–606. doi:10.1002/prot.24214
14. Lensink M, Wodak SJ (2013) Docking, Scoring and Affinity Prediction in CAPRI. 81(12):2082–2095. doi: 10.1002/prot.24428
15. Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77(4):778–795. doi:10.1002/Prot.22488
16. Elber R, Roitberg A, Simmerling C, Goldstein R, Li HY, Verkhivker G, Keasar C, Zhang J, Ulitsky A (1995) Moil—a program for simulations of macromolecules. Comput Phys Commun 91(1–3):159–189
17. Ravikant DVS, Elber R (2010) PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. Proteins 78(2):400–419. doi:10.1002/Prot.22550
18. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.
19. Hunt ME, Modi CK, Aglyamova GV, Ravikant DVS, Meyer E, Matz MV (2012) Multi-domain GFP-like proteins from two species of marine hydrozoans. Photochem Photobiol Sci 11(4):637–644. doi:10.1039/c1pp05238a
20. Millers E-KI, Lavin MF, de Jersey J, Masci PP, Guddat LW. Crystal structure of textilinin-1, a Kunitz-type serine protease inhibitor from the Australian Common Brown snake venom, in complex with trypsin. *RCSB PDB entry 3D65*, http://www.rcsb.org/rxplore/explore.do?structureId=3d65. Accessed 5 June 2013
21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
22. Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
23. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31(13):3375–3380. doi:10.1093/Nar/Gkg543
24. Vallat BK, Pillardy J, Majek P, Meller J, Blom T, Cao B, Elber R (2009) Building and assessing atomic models of proteins from structural templates: learning and benchmarks. Proteins 76(4):930–945. doi:10.1002/prot.22401
25. Vallat BK, Pillardy J, Elber R (2008) A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins. Proteins 72(3):910–928. doi:10.1002/prot.21976
26. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4): 725–738. doi:10.1038/nprot.2010.5