BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 22 2008, pages 2634–2635
doi:10.1093/bioinformatics/btn497

Structural bioinformatics

DOCKGROUND protein–protein docking decoy set

Shiyong Liu[1], Ying Gao[1] and Ilya A. Vakser[1,2,*]

[1]Center for Bioinformatics and [2]Department of Molecular Biosciences, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, USA

Received on April 9, 2008; revised on September 12, 2008; accepted on September 16, 2008

Advance Access publication September 23, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Summary: A protein–protein docking decoy set is built for the DOCKGROUND unbound benchmark set. The GRAMM-X docking scan was used to generate 100 non-native and at least one near-native match per complex for 61 complexes. The set is a publicly available resource for the development of scoring functions and knowledge-based potentials for protein docking methodologies.

Availability: The decoys are freely available for download at http://dockground.bioinformatics.ku.edu/UNBOUND/decoy/decoy.php

Contact: vakser@ku.edu

1 INTRODUCTION

Computational techniques for structural modeling of protein–protein interactions are rapidly developing, both in terms of methodology and computing power (Gray, 2006; Vajda and Camacho, 2004). An important activity in the field of protein–protein docking is the community-wide Critical Assessment of Predicted Interactions (CAPRI; http://capri.ebi.ac.uk; Wodak, 2007), which allows comparison of different computational methods on a set of prediction targets.

A number of databases of protein–protein complexes have been compiled and used to investigate physicochemical and structural preferences at protein–protein interfaces (Davis and Sali, 2005; Douguet *et al*., 2006; Gao *et al*., 2007; Keskin *et al*., 2004; Kundrotas and Alexov, 2007; Lu *et al*., 2003). It is essential for the protein–protein databases to be comprehensive, automatically updated and fully querying, like the ones in the DOCKGROUND project (Douguet *et al*., 2006; Gao *et al*., 2007).

Benchmark sets of complexes with both bound and unbound structures have been developed for validation of docking approaches (Gao *et al*., 2007; Mintseris *et al*., 2005). The sets contain ∼100 crystallographically determined pairs of proteins. An important part in developing intermolecular potentials and scoring functions is decoy sets of structures (false positive matches). Reliable docking procedures have to distinguish between decoys and correct matches. Development of protein–protein docking decoys started in our lab in 1998. The number of decoys was further expanded by Sternberg and co-workers, and then by Baker, Gray and co-workers (RosettaDock, http://depts.washington.edu/bakerpg), Weng and co-workers (ZDOCK, http://zlab.bu.edu) and others. Currently available decoy sets typically are ranked by scoring functions that involve force field terms, statistical potentials, etc. The ZDOCK set contains tens of thousands of matches per complex, which complicates testing and optimization of computationally expensive scoring functions. The RosettaDock set consists of minimized structures with replaced side chains, targeted for high-resolution (post-refinement) scoring, which may be inappropriate for low-resolution scoring of post-scan/pre-refinement complexes with structural clashes and gaps. Some complexes in the above sets do not contain near-native matches. The decoy set presented in this article, built within the DOCKGROUND project (http://dockground.bioinformatics.ku.edu), involves post-scan matches based on shape complementarity alone and contains 100 decoys per complex plus near-native matches for each complex. Thus, it is an unbiased set that it is optimally suited for testing and optimization of the post-scan scoring functions.

2 METHODS

The docking was performed by our GRAMM-X FFT docking procedure (Tovchigrechko and Vakser, 2005). The procedure performs exhaustive sampling of the translation/rotation space with the soft Lennard–Jones potential, based on our GRAMM algorithm, which has been extensively published and validated over the years (Katchalski-Katzir *et al*., 1992; Vakser, 1995, 1997; Vakser *et al*., 1999). The scan stage grid translation step was 1.5 Å and rotation step 6°.

DOCKGROUND project is an expanding resource for the development of docking techniques and studies of protein interfaces (http://dockground.bioinformatics.ku.edu; Douguet *et al*., 2006; Gao *et al*., 2007). The docking decoys were built for the unbound docking benchmark set Version 2, which contains structures with crystallographically determined bound (co-crystallized) and unbound (crystallized separately) forms. The set was built based on the following selection criteria: sequence identity between bound and unbound structures >97%, sequence identity between complexes <30%, deleted homomultimers (sequence identity between chains <70%) and deleted crystal packing complexes and structures in wrong format. The total number of complexes in the set was 99.

GRAMM-X scan was applied to the set to build docking decoys. The following characteristics from the CAPRI evaluation protocol were computed for 500 000 matches per complex: RMSD of the backbone atoms of the ligand (the smaller the component of the complex; the receptor being the larger one), RMSD of the backbone atoms of the interface residues, the number of native residue–residue contacts in the predicted complex divided by the number of contacts in the native complex and the number of non-native residue–residue contacts in the predicted complex divided by the total number of contacts in the complex. Matches with ligand RMSD < 5.0 Å were defined as the near-native ones. The set contains 100 lowest energy non-native structures and at least one near-native structure per complex. The total number of complexes in the decoy set is 61 and includes only complexes where at least one near-native match was found.

[*]To whom correspondence should be addressed.

Table 1. Average statistics on protein–protein docking decoys

Classification	Ligand RMSD[a]	Receptor RMSD[b]	Near-native RMSD[c]	Hits[d]	Number of complexes
enzyme/inhibitor	1.69	1.49	2.77	9.1	21
antibody/antigen	1.04	0.92	3.37	7.4	5
others	1.46	1.87	3.23	7.8	35

[a] Unbound/bound ligand C^{α} RMSD (Å).
[b] Unbound/bound receptor C^{α} RMSD (Å).
[c] Ligand backbone RMSD (Å) in the closest to the native structure near-native match.
[d] Number of near-native matches per complex.

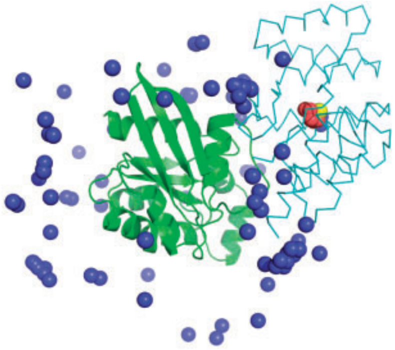

Fig. 1. Example of docking decoys. Matches represented by the ligand's center of mass are shown for 1e96 enzyme-inhibitor complex. the receptor (in green) and the ligand (in cyan) are shown in co-crystallized configuration. The native match is in yellow (not part of the decoy set), 10 near-native matches are in red and 100 non-native matches are in blue.

3 RESULTS

The RMSD between bound and unbound structure reflects the degree of conformational change upon the complex formation. Table 1 shows the average statistics for the three groups of complexes. The average RMSDs between bound and unbound structure are rather small. This corresponds to the earlier estimates indicating that the majority of protein complexes have small backbone conformational change between bound and unbound forms (Gao *et al.*, 2007).

GRAMM-X was unable to detect near-native matches in complexes with large conformational changes (primarily due to the domain shifts). Thus such complexes are not present in the decoy set.

The native structures, as opposed to the near-native ones, were deliberately excluded from the set because they are never achievable in practical docking and thus would be an unrealistic reference point for the development of docking methodologies. An example of docking decoys for a particular complex is shown in Figure 1. Application of popular scoring functions ZRANK (http://zdock.bu.edu/software.php) and DFIRE (http://sparks.informatics.iupui.edu) placed the near-native structure in top 10 matches in 40–50% of complexes.

4 CONCLUSION

A protein–protein docking decoy set is built for the DOCKGROUND unbound benchmark set. The GRAMM-X docking scan was used to generate 100 non-native and at least one near-native match per complex for 61 complexes. The set is a publicly available resource for the development of scoring functions and knowledge-based potentials for protein docking methodologies.

ACKNOWLEDGEMENTS

The authors wish to thank Andrey Tovchigrechko for assistance with GRAMM-X docking.

Funding: National Institutes of Health (grant R01 GM074255).

Conflict of Interest: none declared.

REFERENCES

Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.

Douguet,D. *et al.* (2006) DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, **22**, 2612–2618.

Gao,Y. *et al.* (2007) DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. *Proteins*, **69**, 845–851.

Gray,J.J. (2006) High-resolution protein–protein docking, *Curr. Opin. Struct. Biol.*, **16**, 183–193.

Katchalski-Katzir,E. *et al.* (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl Acad. Sci. USA*, **89**, 2195–2199.

Keskin,O. *et al.* (2004) A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. *Protein Sci.*, **13**, 1043–1055.

Kundrotas,P.J. and Alexov,E. (2007) PROTCOM: searchable database of protein complexes enhanced with domain–domain structures. *Nucleic Acids Res.*, **35**, D575–D579.

Lu,H. *et al.* (2003) Development of unified statistical potentials describing protein-protein interactions. *Biophys. J.*, **84**, 1895–1901.

Mintseris,J. *et al.* (2005) Protein-protein docking benchmark 2.0: an update. *Proteins*, **60**, 214–216.

Tovchigrechko,A. and Vakser,I.A. (2005) Development and testing of an automated approach to protein docking. *Proteins*, **60**, 296–301.

Vajda,S. and Camacho,C.J. (2004) Protein–protein docking: is the glass half-full or half-empty?. *Trends Biotechnol.*, **22**, 110–116.

Vakser,I.A. (1995) Protein docking for low-resolution structures. *Protein Eng.*, **8**, 371–377.

Vakser,I.A. (1997) Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. *Proteins* (Suppl. 1), 226–230.

Vakser,I.A. *et al.* (1999) A systematic study of low-resolution recognition in protein-protein complexes. *Proc. Natl Acad. Sci. USA*, **96**, 8477–8482.

Wodak,S.J. (2007) From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent. *Proteins*, **69**, 697–698.