An Automatic Procedure for the Search and Identification of New Unbound Docking Examples

Oliver Martin, Philipp Heuser, Frank Steinacker and Dietmar Schomburg

CUBIC (Cologne University BioInformatics Center),
Institute of Biochemistry, University of Cologne

Abstract

We developed an automatic procedure to detect unbound-unbound testcases for protein-protein docking. From a set of ~600 protein complexes of known structure, 37 test cases were derived. This dataset is available at:
http://hnb-cologne.uni-koeln.de:8080/uuppd/ .

Introduction

The term protein-protein docking refers to the computational prediction of the natural conformation of a protein complex starting from individual substructures of the complexes components. The most challenging field of protein-protein docking are the so called unbound docking cases, in which individually crystallised structures with high similarity to the subunits of a complex of known structure are subjected to the docking procedure. One of the major problems in the field of unbound protein-protein docking is the low number of unbound docking cases that are presently known. The largest available collection of test cases presently contains 31 entries for unbound docking[1]. Most of the publications concerning docking are therefore only tested on low data fundamentals. Since protein structure databases like the PDB[2] are constantly growing, it is our aim to develop and apply an automatic procedure for the search and identification of new qualified unbound docking examples. The collected unbound docking examples are accessible to the scientific community via a web interface.

Methods

For a *new* unbound docking test case, the following criteria must apply:

a) It should be non redundant to any other known unbound case as far as the interface region is concerned.

b) The quality of the structure should be as high as possible, respectively the resolution with which it has been determined as low as possible.

c) In the individually crystallised structures, the region that refers to the interface in the corresponding bound structure has to be solvent accessible (only applies to proteins with multiple chains).

In order to find new unbound docking examples, i.e. individually crystallised homologues to the sub-structures of protein complexes of know structure, appropriate seeds for this search had to be selected. A set of 431 protein complexes of known structure, derived from the COMBASE[3] was used as well as other complex data taken from the literature. For each complex used as input, a *five step procedure* as described in Figure 1 is applied.

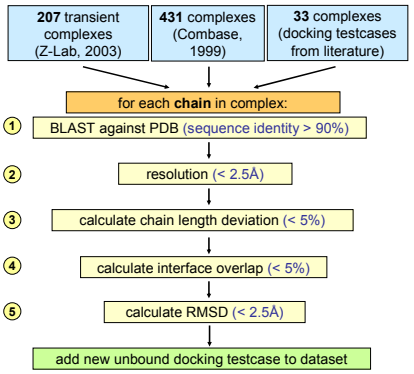

Figure 1: Schematic illustration of the method

While the input or seed data is shown in light blue boxes with the number of actual input complexes given in bold digits, the five individual steps of the procedure are depicted in beige boxes, each stating the actual calculation step with the individual cutoff criterion written in blue letters.

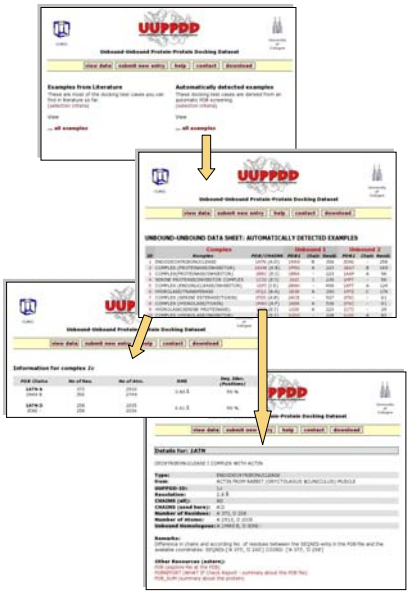

Figure 2: Illustration of the web interface to the UUPPDD, the unbound unbound protein-protein docking dataset

The dataset contains one table each for a collection of unbound test cases extracted from literature as well as the automatically detected examples. Each table is non redundant in itself and furthermore provides detailed information about the test case itself as well as the involved protein chains.

1 Starting with a protein chain of a complex a sequence alignment is performed against all chains in a non-redundant sequence database derived from the PDB. All protein chains with a sequence identity above a cut off criterion are then retrieved from the respective structures.

2 During step two of the procedure, all those chains whose structures have been determined with a resolution above 2.5 Å are omitted.

3 In the third step the chain lengths of the remaining structures are compared to those of the seed. Chains with a percentage deviation in length of more than a cut off value are neglected.

4 In case that there are multiple chains available in the candidate, the interface overlap is calculated in a fourth step, i.e. the percentage value of interface atoms by which the candidate differs from the seed. In order to achieve this, the interface atoms have to be determined. Again a cut off criterion is applied to reduce the number of candidates.

5 Finally a structural alignment between the remaining candidates and the seed is calculated in order to remove those candidates with an RMSD value larger than a selected cut off criterion.

The remaining candidates for possible unbound docking test cases are then ranked according to their values for sequence identity, RMSD and resolution (in the given order). The first rank becomes the representative for the new unbound docking case. The following external programs were used: BLAST[4] for sequence alignment, CE[5] for structure alignment and RMSD calculation as well as NACCESS[6] for the calculation of accessible surface areas which were used to determine the interface regions[7].

Results

The results are available in full detail to the scientific community via a web interface: http://hnb-cologne.uni-koeln.de:8080/uuppdd/ (see Figure 2). As input we used 431 non redundant protein complexes of known structure as derived from the COMBASE, 207 transient protein complexes as described in [8], as well as a collection of complexes taken from known unbound docking examples extracted from various sources. For these examples we found 37 unbound docking test cases. Therefore we used the following parameters: Minimum sequence identity 90%, maximum chain length deviation 5%, maximum interface overlap 5% and maximum RMSD 2.5 Å.

References

[1] R. Chen, J. Mintseris, J. Janin and Z. Weng, A protein-protein docking benchmark. *Proteins*, 52(1):88-91, 2003

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, *Nucleic Acids Research*, 28 pp. 235-242, 2000

[3] I. Vakser and A. Sali, http://salilab.org/sub-pages/combase.html.

[4] S.F.Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* 215:403-410, 1990

[5] I.N. Shindyalov and P.E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* 11(9) 739-747, 1998

[6] S.J. Hubbard and J.M. Thornton, 'NACCESS' Computer Program, Department of Biochemistry and Molecular Biology, University College London, 1993

[7] P. Chakrabarti and J. Janin, Dissecting protein-protein recognition sites. *Proteins*, 47(3):334-43, 2002

[8] J.Ministeris and Z. Weng, Atomic Contact Vectors in Protein-Protein Recognition. *Proteins* (In Press)

Universität zu Köln, Zülpicher Strasse 47, 50674 Köln (Germany),
E-Mail: [Oliver.Martin|Philipp.Heuser|D.Schomburg]@uni-koeln.de

SPONSORED BY THE
Federal Ministry of Education and Research