

# An automatic procedure for the search and identification of new unbound docking examples

Frank Steinacker, Oliver Martin , Philipp Heuser and Dietmar Schomburg

CUBIC (Cologne University BioInformatics Centre), Universität zu Köln, Zùlpicher Strasse 47,  
50674 KÖLN - Germany  
[Oliver.Martin|Philipp.Heuser|D.Schomburg]@uni-koeln.de

**Keywords:** Protein structure, Docking, Interface, Benchmark

## 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 crystallized 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 will be 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 in terms of identity of sequence and structure
- 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 crystallized 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 crystallized homologues to the substructures of protein complexes of known structure, an appropriate seed for this search had to be selected. A set of 950 protein chains, originating from 431 protein complexes of known structure, was derived from the COMBASE[3].

For each of the 431 complexes in the dataset used as input, a five step procedure as described below is applied. 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 in a first step. All protein chains with a sequence identity above a cut off criterion are then retrieved from the respective structures. During step two of the procedure, all those chains whose structures have been determined with a resolution above 2.5 Å are omitted. In the third step the chain lengths of the remaining structures are compared to the seeds'. Chains with a percentage deviation in length of more than a cut off value are omitted. 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. Finally a structural alignment between the remaining candidates and the seed is calculated in order to remove those candidates with an RMSD value above this step's 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

Since the search for new unbound docking test cases following the procedure described above is currently still running, we are not able to present the full amount of our results yet (June 2003). First tests are quite encouraging though and we expect to find new unbound docking examples for about 15-20% of the input data.

The results will be available to the scientific community via a web interface linked at the groups homepage: <http://www.uni-koeln.de/math-nat-fak/biochemie/ds/>

For 378 non redundant binary protein complexes of known structure used as input we found 68 new unbound docking test cases including 43 homodimers. Homodimers are not likely to represent the future application of protein-protein docking but are nevertheless valuable for the testing and evaluation of docking algorithms. These results have been produced while scanning only for binary complexes consisting of no more than two chains. 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. We expect to obtain further unbound docking test cases for the currently running scans for complexes with multiple chains.

## 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, The Protein Data Bank. *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