CAN WE USE HOMOLOGY MODELS IN PROTEIN-PROTEIN DOCKING?

Christoph Schneider, christoph@accelrys.com
Accelrys, Inselkammerstr.1, 82008 Unterhaching, Germany

Introduction

Structures of macromolecular complexes are necessary for a mechanistic description and understanding of biochemical and cellular processes. The formation of a protein-protein complex normally has a functional consequence (e.g. signal transduction), but may also be responsible for the development of pathological processes (e.g. Alzheimer's and prion disease). Genome-wide proteomic studies [1] provide a growing list of putative protein-protein interactions and demonstrate that most, if not all, proteins have interacting partners in the cell. Protein-protein complexes comprise only a few percent of structures in the Protein Data Bank (PDB) [2], since databases contain more sequence than structural information. Thus, it is important to develop computational docking methods that, starting from the structures of component proteins, can determine the structure of their complexes with an accuracy close to that provided by X-ray crystallography.
The impact of protein flexibility in protein-protein docking can in principle be approached by different methods. Ehrlich's approach [3] is a combination of rigid body and torsion angle dynamics. Zacharias [4] has used a reduced protein model to account for the side-chain flexibility. The torsion angle dynamics or the reduced model is a consequence of the fact that a docking combination with an all-atom flexibility of both partners is, due to computational limitations, currently not feasible.
Here, we go one step back and start from the sequence to evaluate the applicability of homology models in protein-protein docking. So the main focus of this poster is to find out to which extend we can use homology models in protein-protein docking.

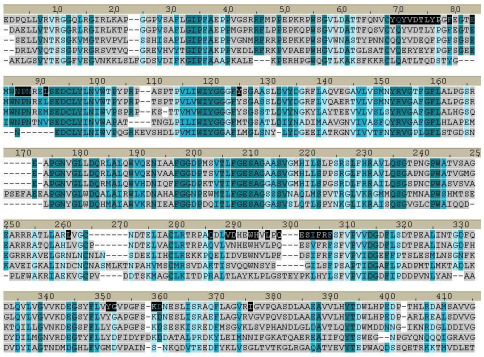

Figure 1 Multiple Sequence Alignment for the homology model building. Residues with black background are part of the receptor interface from 1MAH.

Methodology

Therefore, we need a protein-protein complex where the bigger protein, called the receptor, has some structural and sequence homologs that can be used as templates in the homology model building process. We have chosen an acetylcholinesterase (AChE) complexed with fasciculin2 (Fas)(pdb code 1MAH [5]) as a model system, although the resolution is not that good with 3.5Å. The interface accessible surface area (ASA) is approx. 1030 Å^2, which is 5% of the receptors and 27% of the ligand surface respectively. The number of residues that are in contact are 35 on the receptor side and 21 residues on the ligand side. Furthermore, it consists of 12 hydrogen bonds, no salt bridges, the percentage of polar atoms at the interface is 42%, for the non-polar atoms, it is 58%.

a) Homology Modeling

We have used the sequence from 1MAH to build 4 different models with the following templates (**Table 1**).

Table 1 Sequence Identity/similarity and overall RMSd of the models compared to 1MAH in Å. RMSd values are based on CA atoms.

pdb code	Seq Id	Seq Sim	RMSD [Å]
1b41	87.8	95.3	0.93
1fss	59.3	78.0	1.16
1dx4	39.0	60.7	2.21
2bce	32.8	49.6	4.36

The AChE from 1MAH has 543 amino acids (AA), Fas has 61 AA. We have used the Discovery Studio 1.1 [6] suite with Align123, which is a ClustalW hybrid, for the sequence alignment and Modeler 7v06 [7] for the homology model building. The different homology models were then used as receptor models in the protein-protein docking process. Fas has not been edited or modified. It is important to check the templates and homology structures carefully, because of the sensibility of the protein-protein docking process towards missing side chains, gaps or even missing atoms.

b) Protein-Protein docking

We have used a 2 stage procedure, called ZDOCKpro1.0 [8]. The ZDOCKpro package is based on the ZDOCK and RDOCK programs developed at Boston University. ZDOCK, is a rigid body search with an initial stage algorithm [9], followed by RDOCK, an interface refinement minimization algorithm [10].
In the initial stage, the protein receptor and protein ligand are treated as rigid bodies and all six rotational and translational degrees of freedom are fully explored with scoring functions that are tolerant to conformational changes, also known as soft docking functions. No contact information was used, which can filter hits or block residues during the search. An angular step of 6 deg was used, which results in 54000 poses. ZDOCK uses a shape complementarity method for scoring, called Pairwise Shape Complementarity (PSC) [11]. For the rotational search, evenly distributed Euler angles are used [12].
In the refinement stage RDOCK, the 2000 best poses of near native structures obtained in the initial stage were refined and re-ranked using a more detailed energy function that takes into account conformational changes as well as a solvation term [10].

Results

a) Homology modeling

The homology models were build according to the sequence alignment in **Figure 1**. The templates 1B41, 1FSS and 1DX4 resulted in good models while 2BCE was not a good model, because it resulted in too big structural differences in the binding interface (black residues in **Figure 1**), as can be already seen from the sequence alignment around amino acids 80-90 and 290-305. The 2BCE template therefore produced a model that had a loop pointing directly into the binding site and couldn't therefore give any good results in the docking part. Therefore we will leave out a closer look for the 2BCE model. **Table 2** shows the overall RMSd of the model C alpha atoms compared to the reference structure 1MAH. As we will see, the most important thing for a successful docking is the correct interface. Thats why the difficulties begin, when the template residues in the binding interface differ from the original, because then, Modeller has to build side chain rotamers that might differ significantly from the original crystal structure, which has already an impact on the first stage in the docking process.

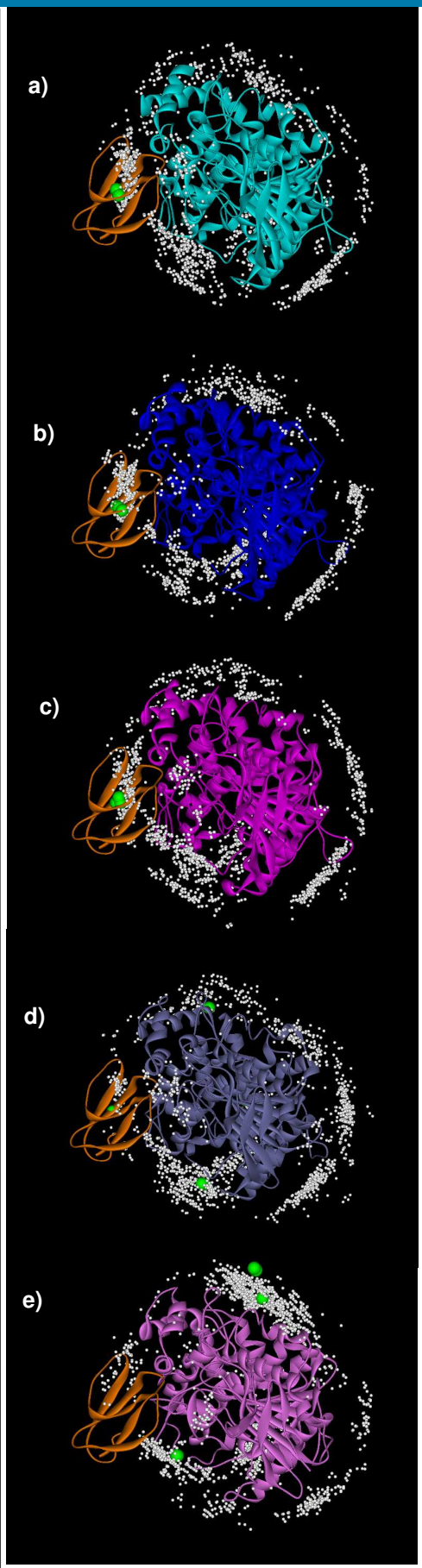

Figure 2 a) Results of the run against the original complex receptor 1MAH as a blind test. White balls represent the center of mass of the 2000 poses, while the 5 best ranked poses are shown as green CPK. In orange is Fas in ribbon. b) 1B41, c) 1FSS, d) 1DX4 and e) 2BCE

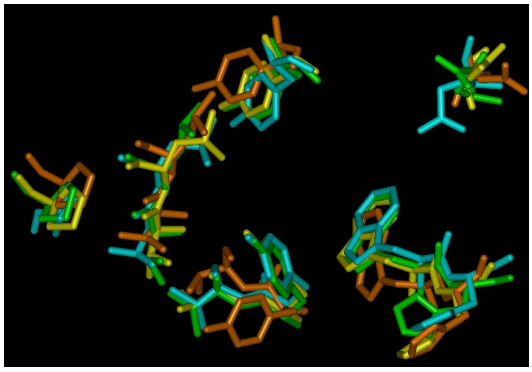

Figure 3 Superimposed are important residues in the binding site with more than 5% ASA. 1MAH in green, 1B41 in yello, 1FSS in light blue and 1DX4 in orange.

Table 3 RMSd differences of 9 binding site residues that show more than 5 % interface accessible surface area per residue. These are Y72, V73, T75, L76, P78, W286, H287, E292 and Y341 from 1MAH.

1MAH-1B41	3.13
1MAH-1FSS	3.16
1MAH-1DX4	3.34
1MAH-2BCE	6.50
1B41-1FSS	1.42
1B41-1DX4	1.65
1B41-2BCE	6.35
1FSS-1DX4	2.04
1FSS-2BCE	6.30
1DX4-2BCE	6.48

b) Protein-protein docking

Protein-protein docking algorithms generate numerous possible complex structures with only a few of them resembling the native structure. The major challenge is choosing the near-native structures from the generated set.
For the purpose of visualizing the docking results, we have reduced all ligands to a center of mass and show the distribution of the 2000 retained poses from ZDOCK, because the orientation and position of the center of mass do not differ significantly between ZDOCK and RDOCK. Although it might be possible that a near-native pose has another orientation, it is quite unlikely. Thats why the best 5 poses taken from the RDOCK output and shown as green CPK in the figures are near native, when they are centered at the Fas (see **Figure 3**).

This is true for the models of 1B41 and 1FSS (**Figure 3b and 3c**) . All 5 top ranked poses for these two models are near-native, which is already an excellent result. 1MAH has been calculated as a blind probe (**Figure 3a**) to show that we can also get good results for the original complex.
It is also obvious from Figure 3, that the interface residues and the top 5 poses from 1DX4 (**Figure 3d**) differ already to a higher degree from 1MAH, than those from 1B41 and 1FSS, but we still can get a very good hit in the top 5 ranked poses. For the 2BCE model (**Figure 3e**) no near-native pose could be retained at all. This is due to the fact that the binding interface looks very different from 1MAH. Rajamani et al. [13] have already shown that often a single residue on the ligand side, called the anchor residue, binds into a structurally constrained groove of the receptor. Therefore even a rotameric state of such an anchor residue that differs from the original, can have a big influence on the docking results.

Table 4 Results from the re-ranking of the RDOCK run. The pose numbers come from the ZDOCK output, which is a ranking according to the PSC only [11]. Energies are in kcal/mol.

	RANK	POSE #	ENERGY		RANK	POSE #	ENERGY
1MAH	1	40	-39.81	1FSS	1	44	-51.52
1MAH	2	17	-39.26	1FSS	2	809	-49.74
1MAH	3	94	-39.02	1FSS	3	5	-46.57
1MAH	4	413	-38.72	1FSS	4	763	-46.06
1MAH	5	60	-38.60	1FSS	5	288	-46.01
1B41	1	2	-42.33	1DX4	1	813	-29.34
1B41	2	1	-41.75	1DX4	2	1066	-27.17
1B41	3	3	-41.24	1DX4	3	768	-26.29
1B41	4	74	-39.71	1DX4	4	465	-25.87
1B41	5	129	-39.36	1DX4	5	1956	-25.76

Summary

Since we know that the density of conserved residue positions is higher at the interface regions of interacting protein surfaces, except for antibody-antigen complexes, it is of high value for the complete docking process, if such evolutionary information can be used to filter the hits.
The sequence identity can be low as long as we can be sure that the binding site is well conserved, especially in structure. Therefore it is of great advantage to collect all possible information that might give hints to the area/residues of interaction.
We also go in line with the results from Ehrlich et al. [3], who have stated that even small (2Å) local loop deformations on the backbone level can affect the contact formation upon docking. So using homology models in protein-protein docking is feasible, when the binding site is very similar to the original structure. The results underline the fact that protein-protein docking of unbound complexes is still a challenging task.

References

[1] Lander et al., *Nature* **409**, 860 (2001).
[2] Berman et al., *Nucleic Acids Research* **28**, 235 (2000).
[3] Ehrlich et al., *Proteins* **58**, 126 (2005).
[4] Zacharias, M. *Protein Sci.* **12**, 1271 (2003).
[5] Bourne et al. *Cell* **83**, 503 (1995).
[6] Discovery Studio 1.1, Accelrys Inc., San Diego, CA, USA. http://www.accelrys.com/dstudio
[7] Sali et al. *Proteins* **23**, 318 (1995).
[8] ZDOCKpro1.0, Accelrys Inc., San Diego, CA, USA. http://www.accelrys.com/zdockpro
[9] Chen et al. *Proteins* **52**, 80 (2003).
[10] Li et al. *Proteins* **53**, 693 (2003).
[11] Chen, R., and Weng, Z. *Proteins* **51**, 397 (2003).
[12] Chen, R., and Weng, Z. *Proteins* **47,** 281 (2002).
[13] Rajamani et al. *Proc Natl Acad Sci U S A.* **101**, 11287 (2004).