pyDock performance in 5th CAPRI edition: from docking and scoring to binding affinity predictions and other challenges

Chiara Pallara, Brian Jiménez-García, Miguel Romero, and Juan Fernandez-Recio
Joint BSC-IRB Research Programme in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain
chiara.pallara@bsc.es

Abstract- *Proteins form the executive machinery underlying all the biological processes that occur within and between cells, from DNA replication to protein degradation. Although genome-scale technologies enable to clarify their large, intricate and highly dynamics networks, they fail to elucidate the detailed molecular mechanism that underlies the protein association process. Therefore, one of the most challenging objectives in biological research is to functionally characterize protein interactions by solving 3D complex structures.*

This is, however, not a trivial task as confirmed by the large gap that exist between the number of complexes identified by large-scale proteomics efforts and those for which high-resolution 3D experimental structures are available. For these reasons, computational docking methods, aimed to predict the binding mode of two proteins starting from the coordinates of the individual subunits, are bound to become a complementary approach to solve the structural interactome.

Given its importance, the field of protein docking has experienced an explosion in recent years partially propelled by CAPRI (http://www.ebi.ac.uk/msd-srv/capri/). CAPRI (Critical Assessment of PRedicted Interaction) is a community-wide blind experiment aimed at objectively assessing the performance of computational methods for modeling protein interactions by inviting developers to test their algorithms on the same target system and quantitatively evaluating the results.

In order to test pyDock,[1] a docking scoring algorithm developed in our group, the PID (Protein Interaction and Docking) group of the BSC Life Science Department, we have participated in all the 15 targets (T46 to T58) of the 5th CAPRI edition (2010-2012). Our automated protocol confirmed to be highly successful to provide correct models in easy-to-medium difficulty protein-protein docking cases placing among the Top5 ranked groups out of more than 60 participants.

Key words: Complex structure, CAPRI, protein-protein docking, pyDock, protein interactions.

I. Introduction

One of the major challenges in structural biology is to provide structural data for all complexes formed between proteins and other macromolecules. Current structural coverage of protein-protein interactions (i.e. available experimental structures plus potential models based on homologous complex structures) is below 4% of the estimated number of possible complexes formed between human proteins.[2] The pace of experimental determination of complex structures is still behind the determination of individual protein structures. In addition, many of these interactions will never be determined by x-ray crystallography because of their transient nature. For these reasons, computational docking methods aim to become a complementary approach to solve the structural interactome. The field of protein docking has experienced an explosion in recent years, partially propelled by the CAPRI (http://www.ebi.ac.uk/msd-srv/capri/). CAPRI (Critical Assessment of PRedicted Interaction) is a community-wide blind experiment aimed at objectively assessing the performance of computational methods for modeling protein interactions by inviting developers to test their algorithms on the same target system and quantitatively evaluating the results. This involves sampling putative association modes and modeling their atomic structure (the docking problem), and identifying those likely to be stable out of a very large pool of decoys (the scoring problem). Models submitted by participants are finally evaluated in comparison with experimental coordinates made available by their authors to the CAPRI assessors according to some criteria as described in Figure 1 of Lensink et al. Proteins 2007 69:704.

In order to test pyDock,[1] a docking scoring algorithm developed in our group, we have participated in all the 15 targets of the 5th CAPRI edition (2010-2012). In addition to the standard prediction of protein-protein targets, this edition has entered into related areas including binding affinity predictions and free energy changes upon mutation, as well as prediction of sugar binding and interface water molecules. Our overall experience has been highly rewarding and we describe here the details of our participation and the key factors of our success.

II. Materials and Methods

A. *Generation of rigid-body docking poses for the predictors experiment*

In all targets, we used FTDock[3] and ZDOCK 2.1[4] to generate 10,000 and 2,000 rigid-body docking poses, respectively. For the final four targets of this edition (T53, T54, T57 and T58) we generated an additional pool of flexible docking poses using SwarmDock. [5]

B. *Scoring of rigid-body docking poses for both the predictors and the scorers experiment*

We scored the docking models generated by the above described methods with our pyDock protocol, based on energy terms previously optimized for rigid-body docking. The binding energy is basically composed of ASA-based desolvation, Coulombic electrostatics and van der Waals energy (with a weighting factor of 0.1 to reduce the noise of the scoring function). Cofactors, water molecules and solvent ions were not considered for scoring.

C. *Removal of redundant docking poses*

After scoring, we eliminated redundant predictions in order to increase the variability of the predictions and maximize the success chances by using a simple clustering algorithm with a distance cutoff of 4.0 Å, as previously described. [6]

D. *Minimization of final models*

The final ten selected docking poses were minimized in order to improve the quality of the docking models and reduce the number of interatomic clashes. In the majority of the targets we used TINKER. In targets T53 and T54 we used

CHARMM[7] while in target T58 we used AMBER10 with AMBER parm99 forcefield. [8]

E. Modeling of subunits with no available structure

For several targets, the structures of the subunits were not available and needed to be modeled. We used Modeller 9v6[9] with default parameters based on the template/s suggested by the organizers or on other homologue proteins found by BLAST search (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The final selected model was that with the lowest DOPE score.

III. RESULTS AND DISCUSSION

In this CAPRI edition we submitted predictions for all the proposed targets. Our results for the standard protein-protein docking assessment are summarized in Table I and Fig.1.

TABLE I

Target	Type[a]	Predictors		
		Submission rank[b]	Quality[c]	Successful Groups[d]
T46	HH	-	-	2 (40)
T47	HU	1	***	25 (29)
T48	UU	3	*	14 (32)
T49	UU	4	*	14 (33)
T50	UH	1	**	18 (40)
T51	DHD	-	-	3 (46)
T53	UH	3[f]	**	20 (42)
T54	UH	-	-	4 (41)
T58	UU	5	**	11 (23)

Table I. Results of pyDock protocol for all protein-protein targets in predictors.
a B: bound; U: unbound; H: homology-based model.
b Rank of the best model within our submission to CAPRI.
c Quality of our best model according to CAPRI criteria: acceptable (*), medium (**), and high (***)
d Number of successful groups for each target; in brackets, total number of participants.
e Model rank 1 had acceptable accuracy (*).

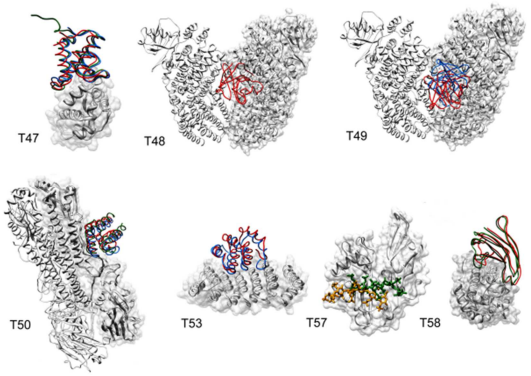

Fig. 1. Representation of our best models for targets T47, T48, T49, T50, T53, T57 and T58. For each target, receptors are superimposed and shown in white. Ligand in our best model as predictors is shown in red, and as scorers in blue. For comparison, the structure of the experimental complex (if available) is represented in green.

For the generation of docking poses, the better grid resolution used for FTDock and the use of flexible SwarmDock for the last targets were key for the success. In selected targets (T47, T48 and T58), distance restraints were used, but in most cases this did not make a difference. In the target T58, SAXS data was used for complementary scoring with pyDockSAXS,[10] which slightly improved the scoring. We obtained consistently good models for all non-difficult cases, although they were far from being trivial, since their subunits were unbound or needed to be modeled based on homology templates. In all cases but one our successful models were ranked within our first five submitted solutions, being ranked 1st in several cases.

IV. CONCLUSIONS

In this CAPRI edition we learned that our automated protocol is useful to provide correct models in easy-to-medium difficulty protein-protein docking cases, but we need further methodological development for difficult cases, especially when subunits need to be modeled based on homologues with low sequence identity. Our overall experience has been highly rewarding, pyDock docking scheme confirmed its high performance in protein complexes prediction placing among the Top5 ranked groups out of more than 60 participants (Table II).

TABLE II

Rank	Group	Summary: #Targets / *** + ** + *
1	Bonvin	9 / 1 *** + 3 ** + 5 *
2	Bates	8 / 2 ** + 6 *
3	Vakser	7 / 1 *** + 6 *
4	Vajda	6 / 2 *** + 3 ** + 1 *
5	**Fernandez-Recio**	**6 / 1 *** + 3 ** + 2 ***
5	Shen	6 / 1 *** + 3 ** + 2 *
7	Zou	6 / 1 *** + 2 ** + 3 *
8	Zacharias	6 / 1 *** + 5 *
9	ClusPro	6 / 4 ** + 2 *
10	Eisenstein	5 / 1 *** + 2 ** + 2 *

Table II. Overall pyDock performance among the Top10 ranked groups. Predictions are classified as acceptable (*), medium (**), and high (***).

REFERENCES

1. Cheng, T. M.-K., Blundell, T. L., & Fernandez-Recio, J. (2007). *Proteins* 68, 503-15.
2. Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., et al. (2009). *Nat Methods* 6, 83-90.
3. Gabb, H. A., Jackson, R. M., & Sternberg, M. J. (1997). *J Mol Biol* 272, 106-20.
4. Chen, R., Li, L., & Weng, Z. (2003). *Proteins* 52, 80-7.
5. Moal, I. H. & Bates, P. A. (2010). *Int J Mol Sci* 11, 3623-48.
6. Pons, C., Solernou, A., Perez-Cano, L., Grosdidier, S., & Fernandez-Recio, J. (2010). *Proteins* 78, 3182-8.
7. Brooks, B. R., Brooks, 3rd, C. L., Mackerell, Jr, A. D., Nilsson, L., et al (2009). *J Comput Chem* 30, 1545-614.
8. Case, D. A., Cheatham, 3rd, T. E., Darden, T., Gohlke, H., Luo, R., Merz, Jr, K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). *J Comput Chem* 26, 1668-88.
9. Sali, A. & Blundell, T. L. (1993). *J Mol Biol* 234, 779-815.
10. Pons, C., D'Abramo, M., Svergun, D. I., Orozco, M., Bernadó, P., & Fernández-Recio, J. (2010). *J Mol Biol* 403, 217-30.