pyDock scoring for the new modeling challenges in docking: protein-peptide, homo-multimers and domain-domain interactions

Chiara Pallara,[1#] Brian Jiménez-García,[1#] Miguel Romero,[1] Iain H. Moal,[1,2] and Juan Fernández-Recio[1*]

[1] Joint BSC-IRB Research Programme in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain

[2] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

# Equal contribution

* Correspondence to: Juan Fernández-Recio, Barcelona Supercomputing Center (BSC), Jordi Girona 29, 08034 Barcelona, Spain. E-mail: juanf@bsc.es

Short title: New modeling challenges in docking

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

ABSTRACT

The 6th CAPRI edition included new modelling challenges, such as the prediction of protein-peptide complexes, and the modelling of homo-oligomers and domain-domain interactions as part of the first joint CASP-CAPRI experiment. Other non-standard targets included the prediction of interfacial water positions and the modelling of the interactions between proteins and nucleic acids. We have participated in all proposed targets of this CAPRI edition both as predictors and as scorers, with new protocols to efficiently use our docking and scoring scheme pyDock in a large variety of scenarios. In addition, we have participated for the first time in the server section, with our recently developed webserver, pyDockWeb. Excluding the CASP-CAPRI cases, we submitted acceptable models (or better) for 7 out of the 18 evaluated targets as predictors, 4 out of the 11 targets as scorers, and 6 out of the 18 targets as servers. The overall success rates were below those in past CAPRI editions. This shows the challenging nature of this last edition, with many difficult targets for which no participant submitted a single acceptable model. Interestingly, we submitted acceptable models for 83% of the evaluated protein-peptide targets. As for the 25 cases of the CASP-CAPRI experiment, in which we used a larger variety of modelling techniques (template-based, symmetry restraints, literature information, etc.), we submitted acceptable models for 56% of the targets. In summary, this CAPRI edition showed that pyDock scheme can be efficiently adapted to the increasing variety of problems that the protein interactions field is currently facing.

INTRODUCTION

The detailed energetics and structural knowledge of all biomolecular interactions occurring in living organisms would contribute to our understanding of biological processes and pathological conditions at the molecular level and give raise to uncountable applications in biomedicine and biotechnological fields. Unfortunately, current experimental knowledge of complex structures covers only a tiny fraction of the total estimated number of possible complexes.[1-2] In this context, computational docking can help to complement experimental efforts in the quest to solve the structural interactome. The number of docking algorithms that have been developed and made available to the scientific community has been continuously growing, especially during the last decade. The CAPRI international experiment, from its very beginning, has been an excellent catalyzer for the field of protein docking. In the most recent editions, the experiment has been extended to an increasing variety of challenges related to the structural modeling of protein interactions. This has continued in the sixth CAPRI edition, which has consisted in the modeling of protein-protein complexes of special difficulty, protein-peptide and protein-nucleosome interactions, as well as interfacial water predictions. In addition, a series of targets from the first joint CASP-CAPRI experiment included the modeling of homo-oligomers and domain-domain interactions. We have participated in all targets of this CAPRI edition, and present here the detailed description of our modeling efforts and the new protocols we have developed to adapt our approaches to the new challenges that the field is facing.

MATERIALS AND METHODS

Generation of rigid-body docking poses for the predictors experiment

In all targets (except for T100-101), we used FTDock[3] (with electrostatics and 0.7 Å grid resolution) and ZDOCK 2.1[4] to generate 10,000 and 2,000 rigid-body docking poses, respectively, in the same conditions as previously described.[5] For six targets of this edition (T59, T96-97, T103-105) we

generated an additional pool of flexible docking poses using SwarmDock. For these runs, the standard protocol was employed,[6-8] with the DFIRE score used as the objective function,[9] but without the final rescoring phase. Target T95 included DNA, so we used an *ad-hoc* protocol based on SwarmDock (see more details in Results section). For target T106, we used a work-in-progress version of our new docking protocol LightDock (unpublished) to generate over 3,000 additional docking poses. This new protocol included explicit backbone flexibility by using Anisotropic Network Model (ANM)[10] during the sampling process, and made use of DFIRE[9] scoring function. Cofactors, water molecules and solvent ions were not included in our docking calculations. T100 and T101 models were built assuming that the chromatin remodeler component INO80G had the same orientation as the proteasome regulatory subunit Rpn13 in the T99 and T98 submitted models, respectively.

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 default pyDock protocol,[11] 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). Electrostatics and van der Waals were limited to -1.0/+1.0 and 1.0 kcal/mol for each inter-atomic energy value, respectively, in order to avoid excessive penalization from possible clashes derived of the rigid-body approach. For some of the targets we found experimental information on possible interface residues, which were included in the final scoring as distance restraints with pyDockRST[12] (T60-64, T98-99 and T103), or used as a final distance-based filtering step (T104-105). The same protocol used in predictors was applied in the scorers experiment to score all the docking models that were proposed, except for target T59, where a final RMSD-based filtering step was applied only as scorers. Cofactors, water molecules and solvent ions were not considered for scoring. After scoring, we eliminated redundant predictions by using a BSAS algorithm[13] with a distance cutoff of 4.0 Å, as previously described.[14] The final ten

selected docking poses were minimized by using AMBER12[15] with AMBER parm99 force field[16] in order to improve the quality of the docking models and reduce the number of interatomic clashes, as previously described.[17] The minimization protocol consisted in a 500-cycle steepest descent minimization with harmonic restraints applied at a force constant of 25 kcal/(mol·Å2) to all the backbone atoms in order to optimize the side-chains, followed by another 500-cycle conjugate gradient minimization without restraints.

Modeling of subunits with no available structure

For several targets, the structures of the subunits were not available and needed to be modeled. In most of the targets, we used Modeller 9v6 with default parameters[18] based on the template/s suggested by the organizers or on other homologue proteins found by BLAST[19] search. The final selected model was that with the lowest DOPE score.[20] HHpred server[21] was used to model the artificial alpha-repeat eGFP A in T96, as well as the missing carboxy-terminal peptide (313-329 residues) of the Ubiquitin carboxyl-terminal hydrolase L5 (UCH-L5) in T98-99, which was previously reported to be involved in the binding to the proteasome regulatory subunit Rpn13.[22-23] MUSTER on-line server[24] was used to model the UBE2Z protein in target T103.

Servers experiment

For the servers experiment, we participated in 13 of the evaluated targets (T59-67, T96-97, T103-105, T107) with our pyDockWeb server (https://life.bsc.es/servlet/pydock).[25] The generation of docking poses and further scoring were done in a fully automatic manner by FTDock and pyDock, as previously described. In cases for which additional experimental data were available, we added distance restraints with the pyDockRST module included in the server. Finally, the best-scored server predictions were clustered and minimized according to our default protocol before submission to CAPRI.

Modelling of protein-peptide complexes

For the prediction of the complexes between importin-α and nuclear signaling peptides (T60-64), we applied two different strategies. On the one side, the initial peptide structures were modelled by 500-cycle minimization with GB model using AMBER12 package[15] and AMBER parm99 force field,[16] followed by 20-ns unrestrained molecular dynamics (MD), from which 5 representative snapshots were selected. Then, these peptide structures were used for docking simulations with our standard protocol for predictors, after which the results of the independent docking runs were merged, scored by pyDock and clustered (see above). On the other side, we applied a template-based approach. We first superimposed 27 peptide-bound importin-α structures, and identified the residue correspondence in the peptides at both binding sites (Figure 1A). We threaded the target sequences through the peptide sequence and identified alignments which gave good agreements with the residue propensities in the homologues. We then used this as a basis for template modeling. For each target/alignment pair, peptide fragments were joined together with averaging of the atomic coordinates of overlapping fragments, keeping the side-chain conformations where possible. The amalgamated partial models were then superimposed into position in the binding sites of all 27 homologues, missing side-chains were rebuild with SCWRL4,[26] and the structures minimized with CHARMM.[27] The large set of models was then scored with pyDock (see above). For each target, we submitted the five best models generated by each of these two strategies. The server submissions were automatically built by the pyDockWeb docking server, using as input the conformations of the peptides generated by homologous templates (PDB 3UL1 and 3UKZ), followed by side-chain rebuilding with SCWRL in the context of the PDB 1EJL complex, and a subsequent 500-cycle minimization with GB model with Amber using AMBER12 package[15] and AMBER parm99 force field.[16]

For the prediction of the rest of protein-peptide complexes (T65-67), we applied an *ad-hoc* template-based homology protocol, For T65 and T66, we rigidly docked by FTDock and then scored the DIPF binding motif of the SBB peptide, which is structurally conserved in other SSB interactions

(3C94, 3Q8D, 3SXU, 3UF7 and 3UFM). For our 10 top hits, we then built the missing WMDFDD fragment by iteratively building towards the N-terminus by sampling putative conformations from neighbor-dependent φ and ψ distributions[28] and a backbone-dependent rotamer library,[29] selecting configurations using DFIRE.[9] A similar protocol was undertaken for T67, after docking of the PSY domain of Commissureless (2EZ5), and building both the N- and C-terminal flanking residue outwards from this motif. We scored the final models with pyDock. For the server submissions with pyDockWeb, we used FTDock with the peptide models obtained by template-based homology modeling (for T65-66: 8 templates with the following PDB code and chain ID: 3C94_B, 3C94_C, 3Q8D_E, 3Q8D_F, 3UF7_B, 3UF7_C, 3SXU_C, and 3UFM_B; for T67: 2 templates with the following PDB code and chain ID: 2KQ0_B, 2KPZ_B) or 5 representative models from 100 ns MD trajectories. The results from all docking runs were merged and we submitted the top 10 models as scored by pyDock.

Prediction of interfacial water positions

The prediction of interfacial water positions was performed by using the protocol previously reported by Zacharias group,[30] as follows.

Step 1: Each complex structure (previously minimized as described above) was first solvated in explicit TIP3P water,[31] and Na+ and Cl- counterions were added to the solvent bulk to neutralize the system by using the leap module of the AMBER12 package and the parm03 force field. Then, the solvated complex underwent a short energy minimization of 2,000 steps of conjugated gradient method.

Step 2: A 100-ps MD simulation was run after raising the temperature to 300K and applying positional restraints on all the solute heavy atoms to their location in the predicted docked complex with a force constant of 25 kcal/(mol·Å^2). The long range electrostatic interactions were computed by the Particle Mesh Ewald (PME) method[32] with an integration time step of 2 fs.

Step 3: Only the water molecules located at the interface between the two proteins were selected from the structure obtained after the MD simulation. Interface water molecules were defined as those within 4 Å atomic distance from any of the docking partners of the protein-protein complex. Then, the complex plus the interface waters were further minimized by 2,000 additional steps of conjugated gradient minimization.

Step 4: Finally, only the water molecules involved in a polar contact to at least one protein atom were selected for the final prediction of interface water location.

RESULTS AND DISCUSSION

In this CAPRI edition we submitted predictions for all the proposed targets. Our results are summarized in Table I. Below are the details of our submissions for the most relevant targets.

Successful predictions

Targets T60-64 (unbound / peptide models)

This set of targets consisted in the interaction between mouse importin-α and five different nuclear signaling peptides: Gu-α (T60), a28 (T61), a58 (T62), b6 (T63), or b141 (T64). These five complexes were evaluated as three targets: i) the major binding site; ii) the minor binding site; and iii) the 6 central peptide residues in the minor binding site. Coordinates of importin-α were taken from PDB 1EJL, bound to a large T antigen 7-residue peptide. The structures of the five target peptides were not available and needed to be modelled by using a dual strategy, based on *ab initio* molecular dynamics or on homologous templates (see Methods). For each submission, five of the protein-peptide models were built by template-based homology modeling (see Methods), and the other five, by docking, using as input structures the models generated by molecular dynamics. Complex structures for all the complexes are now available with the following PDB codes: 3ZIN[33] (T60), 3ZIO[33] (T61), 3ZIP[33] (T62), 3ZIQ[33] (T63), and 3ZIR[33] (T64) respectively. Overall, we obtained excellent results, with medium or better quality models for the major binding site and the 6 central peptide residues in the minor binding site of the five protein-peptide complexes as predictors, and acceptable or better quality models for these two binding sites as servers (Figure 1B). More in detail, as predictors we submitted 1 high- and 4 medium-accuracy models for the major binding site in the 5 peptide complexes; 1 acceptable model for the complete minor binding site of one of the peptide complexes (T63); and 5 medium models for the 6-residue minor binding site of the 5 peptide complexes. As servers, we submitted 1 acceptable model for the major binding site of one of the peptide complexes (T63); and 1 medium and 2 acceptable models for the 6-residue minor binding site of 3 of the peptide complexes (T61, T62, T63). Interestingly, all of

the correct protein-peptide models submitted as predictors were directly built based on homologous templates, which shows that the use of unrestrained molecular dynamics to build the conformations of the peptides did not yield suitable input structures for docking. Remarkably, the correct models submitted as servers were automatically built by the docking server, using minimized template-based peptide models as input for the docking.

[INSERT TABLE I HERE]

[INSERT FIGURE 1 HERE]

Targets T65-66 (unbound / peptide model)

These two targets consisted in the Ct peptide of ssDNA binding protein (SBB-Ct) in complex with RNaseH (T65) or DNA helicase (T66). The structure of RNaseH was available as unbound (PDB 2RN2), while the coordinates of the DNA helicase were provided as unbound by the CAPRI organizers (now available as PDB 4NL4). The structure of the peptide was not available. As predictors, we applied an *ad-hoc* template-based homology modeling procedure (see Methods). As servers, we modelled the structure of the peptide following a dual strategy, template-based and molecular dynamics sampling (see Methods), and then the peptide models were automatically used as input for the pyDockWeb server. For the T66 complex (complex structure is now available as PDB 4NL8),[34] we submitted one acceptable model as predictors. As servers, we submitted two acceptable models, which were generated using as input one peptide structure built by template-based modeling and another by MD sampling (Figure 1C).

However, for the T65 complex (complex structure is now available as PDB 4Z0U)[35] we were not able to submit any correct model, either as predictors or as scorers. This case was highly difficult for the majority of participants, since there was only one successful group out of more than 40 participants. Indeed, RNaseH binding to Ct peptide involved a large conformational change: RNaseH

interface atoms, defined as those within 5 Å distance from Ct peptide in the complex structure, showed 4.2 Å RMSD between the unbound and bound structures.

Target 67 (unbound / peptide model)

Target T67 consisted in the interaction between Nedd4 WW3 domain and the PPxY motif of ARRDC3. The unbound structure for the protein was provided, and the peptide structure needed to be modeled. As predictors, we applied an *ad-hoc* template-based homology modeling procedure (see Methods). As servers, we modeled the structure of the peptide following a dual strategy, template-based and molecular dynamics sampling (see Methods), and then the peptide models were automatically used as input for the pyDockWeb server. After evaluation by the organizers (complex structure is now available as PDB 4N7H),[36] we found we submitted acceptable models both as predictors and as servers (Figure 1C). Interestingly, the three successful models submitted as servers were generated using as input template-based modeled peptides.

Targets T96-97 (model / model)

Targets T96-97 consisted in the interaction between eGFP and the artificial α-repeat eGFP-binder A (T96) or C (T97). The coordinates of eGFP were available as an unbound structure (PDB 1JBZ), but given the high number of missing residues in the crystal, we decided to model it based on a FRET-optimized cerulean fluorescent protein (PDB 4EN1, 92% sequence identity). On the other side, the structures of eGFP-binder A and C were not available and were modeled based on a homologous template (PDB 3LTJ) with 82% and 74% sequence identity, respectively. The complex structures for these targets are now available with PDB codes 4XL5[37] (T96) and 4XVP[37] (T97). We submitted acceptable models for T97 as servers and as scorers (Figure 1C), while we failed to submit any correct model for T96 (our submitted model #2 as servers was almost acceptable, with ligand RMSD 8.2 Å and interface RMSD 3.2 Å, but unfortunately it was classified as incorrect due to the fraction of native

contacts 0.091). The main reason for the different performance of these two targets could be related to the larger deviation of the modeled α-repeat eGFP-binder A protein with respect to the bound structure in T96 as compared to that of the α-repeat eGFP-binder C protein in T97 (interface RMSD 5 Å and 2 Å, respectively). This larger deviation in T96 subunit could be due either to modeling issues or to conformational rearrangement upon binding. From *a posteriori* analysis of our initial sets of decoys, we found that in T97 there were many more acceptable solutions than in T96, both as predictors and as servers, which suggests that the large deviation in T96 subunit had some kind of effect in sampling. In the case of scorers, we also obtained better results for T97. Interestingly, in the initial set of scorers provided by the organizers there were only two acceptable poses in T96, as compared to 18 acceptable poses in T97, which again points to the existence of global sampling difficulties in T96. In general, our performance in these two targets was consistent with the results of the rest of the CAPRI participants, which showed that T96 was a more difficult target than T97.

Target T103 (model / model)

Target T103 consisted in the Ube2Z protein in complex with Fat10. The structures of the Ube2Z and Fat10 proteins were modeled based on homologous templates (PDB 3CEG and 3U30, with 43% and 32% sequence identity, respectively). We submitted acceptable models only as scorers (Figure 1C). As predictors and servers, we submitted models that were only slightly worse than those as scorers, but they were not classified as acceptable.

Targets T104-105 (model / model)

These targets were trivial to model, since a homologous structure was available for the complex (actually a previous CAPRI target T47), so the real challenge was to predict the interfacial water positions. Target T104 consisted in the interaction between pyoAP41 and ImAP41 proteins. As none of these structures were available, they had to be modeled based on homologous templates, colicin E9

(48% sequence identity) and Im9 immunity protein (46% sequence identity), respectively. Both template structures were extracted from the PDB 1BXI. Target T105 consisted in the interaction between pyoS2 and ImS2 proteins, whose structures were modeled based on colicin E2 DNase (52% sequence identity) and Im2 immunity protein (59% sequence identity), as found in PDB 3U43 chains B and A, respectively. Given the existence of the above-mentioned homologous complex structures, the binding mode for both the targets would be easy to determine by template-based docking. However, we performed the template-free docking calculations to assess the automatic docking protocol. We applied distance restraints after pyDock protocol by selecting those docking poses in which two key contacting residues, pyoAP41 Y81 and ImAP41 F59 (equivalent to colicin E9 F86 and Im9 Y54), or pyoS2 Y85 and ImS2 Y55 (equivalent to colicin E2 F86 and Im2 Y54), were within an arbitrary distance of 6 Å (same distance as that used by default in pyDockRST module).[12] After evaluation by the organizers (complex structures for the targets are now available with PDB codes 4UHP[38] and 4QKO,[38] respectively), we found we had submitted acceptable (or better quality) predictions for complex structure and water positions in the two targets, both as predictors and as scorers, but only for T105 as servers. There was a clear correlation between the quality of our predictions for the complex structure and that of the interfacial water positions. In the case of T104, the best model was incorrect (12% native contacts, 14.9 Å ligand RMSD, and 7.7 Å interface RMSD) mostly because we used automatic docking, which included only a minimal information on the homologous complex structure as distance restraints as above explained, and this model was insufficient to correctly predict interfacial water positions. This is consistent with the previous finding that high- to medium-quality protein complex structural models are required for successful interface water predictions.[30]

Unsuccessful cases

In most of the protein-protein cases (excluding the CASP-CAPRI experiment) we were not able to submit any correct model, either as predictors or as scorers. In general, these cases seemed to be highly

difficult for the majority of participants, and actually, for many of them there were very few (if any) successful groups (Table I).

T59 (unbound / model)

Target T59 consisted in the interaction between the LSm domain of Edc3 protein and the ribosomal protein Rps28b. The NMR structure of Edc3 was available both as unbound (PDB 4A53) and in complex with a short helical leucine-rich motif (HLM) from Dcp2m mRNA Decapping Complex (PDB 4A54), i.e., SxxLLxLL, involving S258, L260, L261, L263, L264, which is expectedly responsible for binding to the LSm domain of Edc3 protein. The structure of the Rps28b was not available and had to be modeled. We note that the CAPRI organized suggested the NMR structure PDB 1NE3 as a template (SI 53%), but we decided to use a closer template, the cryo-EM structure PDB 3IZB (superseded by 4V6I) (SI 85%). Both have significant structural differences (global RMSD > 4.6 Å), so we initially speculated that our choice of template was the reason for our wrong predictions, but this did not seem the case after all (see below). Interestingly we found a motif on Rps28b similar to the short helical leucine-rich motif (HLM) of Dcp2m mRNA Decapping Complex (i.e., ILxLL, I54, L55, L57, L58). We applied different protocols as servers, predictors and scorers. As servers, we used only one structure of the LSm domain of Edc3 protein (namely the first NMR model in 4A53) and one modeled structure of the ribosomal protein Rps28b. As predictors, we merged the docking poses generated starting from several NMR models of Edc3, in which the C-term half of the protein was fully disordered, and one modeled structure of Rps28b. As scorers, we combined our standard energy-based protocol with the selection of those docking poses in which the ligand leucine-rich motif orientation in the binding site was similar to that of the short helical leucine-rich motif (HLM) of Dcp2m mRNA Decapping Complex bound to Edc3 (PDB 4A54) (i.e., RMSD calculated on all heavy atoms within 5 Å distance from the corresponding atoms located on the HLMs motif of Dcp2m mRNA Decapping Complex). Unfortunately, none of these different protocols managed to generate any correct model within the top

10 ranked docking solutions. However, we submitted at least one almost acceptable solution as predictors (ranked 10 model, with 13% native contacts, 15.6 Å ligand RMSD, and 6.3 Å interface RMSD with respect to the complex crystal structure), as servers (ranked 9 model, with 13% native contacts, 13.9 Å ligand RMSD, and 5.5 Å interface RMSD with respect to the complex crystal structure), and as scorers (ranked 10 model, with 5.1% native contacts, 15.1 Å ligand RMSD, and 7.5 Å interface RMSD). From the *a posteriori* analysis of the scorer results, we found that in several of our selected poses (ranked 2, 5, 8) the Rps28b protein was probably modelled upon 1NE3 (RMSD < 1 Å from 1NE3), and yet these poses were also incorrect. Therefore, the choice of template perhaps was not as critical as we initially speculated, and there must be additional reasons for our incorrect predictions, perhaps related to scoring issues, to incorrect inclusion of the expected binding motif in Rps28b, or to the fact that we included the disordered C-term half of Edc3 in docking.

T95 (unbound / unbound)

Target T95 consisted in the interaction between PRC1 ubiquitylation module and the nucleosome core particle, whose coordinates were available as unbound structures (PDB 3RPG and 3LZ0, respectively). Once the complex structure was released (PDB 4R8P),[39] we found that the molecules did not show large conformational changes upon binding (RMSD of less than 1 Å to the bound conformation calculated on all the carbon-alpha atoms and roughly 1.5 Å on all the DNA atoms with respect to the complex structure). Nevertheless, this was a challenging case in which only three participants submitted acceptable models. Most likely, the presence of DNA made docking and scoring extremely difficult. For this target we tried an *ad-hoc* procedure, using SwarmDock with the standard DFIRE potential,[9] in combination with the DDNA3[40] scoring function for the DNA interactions. The fact that the near-native solution was not sampled within the top-scoring poses could have been caused by scoring problems, given that DFIRE and DDNA3 were not optimally balanced due to time constraints. However, we cannot disregard that using the DDNA3 potential in a non-optimized manner might have

also had a negative effect on sampling, which in our implementation is strongly guided by the scoring function.

T98-101 (unbound / model)

Targets T98-101 consisted in the interaction between the Ubiquitin Carboxyl-terminal hydrolase L5 or L5Ub (with ubiquitin covalently bound) proteins and either RPN13 activator or Ino80G inhibitor. The structures for these complexes were later released with the following PDB codes: 4UEM,[41] 4UEL,[41] 4UF6,[41] and 4UF5,[41] respectively. For none of these targets we were able to submit any acceptable model, either as predictors or as scorers. Indeed, these cases were highly difficult for all participants, since there was not found a single acceptable model among all the participants. The main challenges in these cases were the large conformational changes of both the interacting proteins upon binding, the inaccuracy in the modelling of Ubiquitin carboxyl-terminal hydrolase L5 interface, as well as the small interface area between the docking partners. Moreover, for Targets T100-101 we assumed the same binding orientation as T98-99, which turned out to be wrong.

T107 (unbound / unbound).

Target T107 consisted in the interaction between the hemopexin binding protein and hemopexin. For this target, we were not able to submit any acceptable model, either as predictors, servers or scorers (the structure of the complex is now available as PDB 4RT6).[42] Indeed, this case was highly challenging for all participants, since there was not a single acceptable model among all participants. The main reason for the difficulty of this target lies on the large conformational changes of the hemopexin binding protein upon hemopexin binding, especially involving a large loop (residues 707-730) located within the complex interface (unbound-to-bound Cα-RMSD 16.2 Å). Another potential reason for the target difficulty could be the large size of the hemopexin binding protein, composed of

around 800 amino acid residues, for which our methodology cannot provide sufficient sampling, as it was found during the last CAPRI edition.[17]

Prediction of protein-water interactions

In targets T104-105, participants were asked to predict the location of the water molecules within the complex structure. In a similar past CAPRI target T47, we used DOWSER *ab initio* optimization procedure.[43] Although this choice was reasonably successful, *a posteriori* analysis showed that the most successful approaches were based on the combination of molecular mechanics force fields with some conformational sampling step and a final energy minimization.[30] Thus, for Targets T104 and T105 we decided to use a protocol based on that previously described by Zacharias et al (see Methods). For Target T104, as predictors we submitted 5 fair models (+), and as scorers 7 good (++) models and 2 fair (+) ones. For Target T105, we submitted 2 fair models (+) as predictors, and 8 fair (+) ones as scorers. This target was similar to past target T47, and although the protocol used here is supposed to be more robust than the one we used that past target, the prediction success did not improve. This suggests that the most important determinant for the prediction of interfacial water positions is the accuracy in the prediction of the complex structure to a greater extent than the protocol used.

CASP-CAPRI experiment

CAPRI round 30, the first joint CASP-CAPRI experiment, consisted in 25 targets of homo- and hetero-oligomers from the CASP11 2014 Round (targets T68-94, excluding T76 and T86, which were cancelled). We submitted at least one acceptable model in 11 out of the 12 easy homo-dimer targets, either as predictors or as scorers. In addition, as scorers we successfully predicted two out of the six difficult homo-dimer targets, and one out of the two hetero-complex targets. On the contrary, we did not submit any successful model for any of the five tetrameric targets, where the inaccuracy of the homology-built subunit models and the smaller pair-wise interfaces severely limited the ability to

derive the correct assembly mode. Globally, pyDock predictions were placed among the top 10 ranked groups out of about 25 predictors, and among the top 5 ranked groups out of about 12 scorers participating in this experiment. More details on our protocols and results have been already described in a recent publication.[44]

CONCLUSIONS

We have continued our participation in CAPRI, submitting models for all targets as predictors and scorers, and for most of the targets as servers. Our scoring scheme pyDock has been used to rank models generated by different approaches (FTDock, ZDOCK, SwarmDock). The protein-protein targets in this 6th CAPRI edition showed to be highly challenging, since for most of them only a few (if any) participants submitted correct models. In most of the failed cases, the main problem was the large deviation between the structure of the subunit/s used in docking and the bound state. This could be due to conformational rearrangement of the unbound state upon binding (T107), but also to the added uncertainty of modeling a subunit when the unbound structure was not available (T96, T98-101, T103). Other difficulties were the fact that half of a protein was completely disordered in T59, and the presence of DNA in T95. On the other side, we had quite successful predictions for the protein-peptide targets, using a combination of docking and template-based modeling. This shows that, when using a correct conformation for the peptide, especially if based on a homologous template, the pyDock scoring function is very efficient for the identification of near-native binding modes in protein-peptide interactions, which opens new research and methodology development possibilities for our software. Although evaluated in a separate way, we should mention here our successful results for the first joint CASP-CAPRI experiment. Overall, our results in CAPRI confirm the applicability of the pyDock approach, in combination with other state-of-the-art tools, to an increasing variety of targets, including protein-peptide complexes, homo-oligomers and domain-domain interactions. On the other side, the CAPRI experiment shows also the limitations of current docking approaches in difficult cases with

large conformational movements upon binding, interacting subunits without available structure, or multi-molecular complexes including DNA molecules.

ACKNOWLEDGEMENTS

This work has been funded by grant number BIO2013-48213-R from the Spanish Ministry of Economy and Competitiveness. B.J.-G. was supported by FPI fellowship, from the Spanish Ministry of Economy and Competitiveness. I.H.M. was supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement PIEF-GA-2012-327899.

REFERENCES

1. Venkatesan K, Rual JF, Vazquez A, Stelzl U, Lemmens I, Hirozane-Kishikawa T, Hao T, Zenkner M, Xin X, Goh KI, Yildirim MA, Simonis N, Heinzmann K, Gebreab F, Sahalie JM, Cevik S, Simon C, de Smet AS, Dann E, Smolyar A, Vinayagam A, Yu H, Szeto D, Borick H, Dricot A, Klitgord N, Murray RR, Lin C, Lalowski M, Timm J, Rau K, Boone C, Braun P, Cusick ME, Roth FP, Hill DE, Tavernier J, Wanker EE, Barabasi AL, Vidal M. An empirical framework for binary interactome mapping. Nat Methods 2009;6:83-90.
2. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, Wiuf C. Estimating the size of the human interactome. Proc Natl Acad Sci U S A 2008;105:6959-64.
3. Gabb HA, Jackson RM, Sternberg MJ. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106-20.
4. Chen R, Weng Z. A novel shape complementarity scoring function for protein-protein docking. Proteins 2003;51:397-408.
5. Grosdidier S, Pons C, Solernou A, Fernandez-Recio J. Prediction and scoring of docking poses with pyDock. Proteins 2007;69:852-8.
6. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 2010;11:3623-48.
7. Torchala M, Moal IH, Chaleil RA, Fernandez-Recio J, Bates PA. SwarmDock: a server for flexible protein-protein docking. Bioinformatics 2013;29:807-9.
8. Li X, Moal IH, Bates PA. Detection and refinement of encounter complexes for protein-protein docking: taking account of macromolecular crowding. Proteins 2010;78:3189-96.
9. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 2004;56:93-101.
10. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins 2000;40:512-24.

11. Cheng TM, Blundell TL, Fernandez-Recio J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 2007;68:503-15.

12. Chelliah V, Blundell TL, Fernandez-Recio J. Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. J Mol Biol 2006;357:1669-82.

13. Theodoridis SK, Konstantinos. Pattern Recognition. London (England): Academic Press; 1999. 984 p.

14. Pons C, Solernou A, Perez-Cano L, Grosdidier S, Fernandez-Recio J. Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain-domain assembly and protein-RNA binding. Proteins 2010;78:3182-8.

15. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr., Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668-88.

16. Cheatham TE, 3rd, Cieplak P, Kollman PA. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. J Biomol Struct Dyn 1999;16:845-62.

17. Pallara C, Jimenez-Garcia B, Perez-Cano L, Romero-Durana M, Solernou A, Grosdidier S, Pons C, Moal IH, Fernandez-Recio J. Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges. Proteins 2013;81:2192-200.

18. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779-815.

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403-10.

20. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507-24.

21. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and

structure prediction. Nucleic Acids Res 2005;33:W244-8.

22. Chen X, Lee BH, Finley D, Walters KJ. Structure of proteasome ubiquitin receptor hRpn13 and its activation by the scaffolding protein hRpn2. Mol Cell 2010;38:404-15.

23. Hamazaki J, Iemura S, Natsume T, Yashiroda H, Tanaka K, Murata S. A novel proteasome interacting protein recruits the deubiquitinating enzyme UCH37 to 26S proteasomes. EMBO J 2006;25:4524-36.

24. Wu S, Zhang Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72:547-56.

25. Jimenez-Garcia B, Pons C, Fernandez-Recio J. pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics 2013;29:1698-9.

26. Krivov GG, Shapovalov MV, Dunbrack RL, Jr. Improved prediction of protein side-chain conformations with SCWRL4. Proteins 2009;77:778-95.

27. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem 2009;30:1545-614.

28. Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL, Jr. Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. PLoS Comput Biol 2010;6:e1000763.

29. Shapovalov MV, Dunbrack RL, Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 2011;19:844-58.

30. Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, Karaca E, Schmitz C, van Dijk M, Bonvin AM, Eisenstein M, Jimenez-Garcia B, Grosdidier S, Solernou A, Perez-Cano L, Pallara C, Fernandez-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G,

Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ. Blind prediction of interfacial water positions in CAPRI. Proteins 2014;82:620-32.

31. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 1983;79:926-935.

32. Darden T, York D, Pedersen L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. The Journal of Chemical Physics 1993;98:10089-10092.

33. Chang CW, Counago RM, Williams SJ, Boden M, Kobe B. Distinctive conformation of minor site-specific nuclear localization signals bound to importin-alpha. Traffic 2013;14:1144-54.

34. Bhattacharyya B, George NP, Thurmes TM, Zhou R, Jani N, Wessel SR, Sandler SJ, Ha T, Keck JL. Structural mechanisms of PriA-mediated DNA replication restart. Proc Natl Acad Sci U S A 2014;111:1373-8.

35. Petzold C, Marceau AH, Miller KH, Marqusee S, Keck JL. Interaction with Single-stranded DNA-binding Protein Stimulates Escherichia coli Ribonuclease HI Enzymatic Activity. J Biol Chem 2015;290:14626-36.

36. Qi S, O'Hayre M, Gutkind JS, Hurley JH. Structural and biochemical basis for ubiquitin ligase recruitment by arrestin-related domain-containing protein-3 (ARRDC3). J Biol Chem 2014;289:4743-52.

37. Chevrel A, Urvoas A, de la Sierra-Gallay IL, Aumont-Nicaise M, Moutel S, Desmadril M, Perez F, Gautreau A, van Tilbeurgh H, Minard P, Valerio-Lepiniec M. Specific GFP-binding artificial proteins (alphaRep): a new tool for in vitro to live cell applications. Biosci Rep 2015;35.

38. Joshi A, Grinter R, Josts I, Chen S, Wojdyla JA, Lowe ED, Kaminska R, Sharp C, McCaughey L, Roszak AW, Cogdell RJ, Byron O, Walker D, Kleanthous C. Structures of the Ultra-High-Affinity Protein-Protein Complexes of Pyocins S2 and AP41 and Their Cognate Immunity Proteins from Pseudomonas aeruginosa. J Mol Biol 2015;427:2852-66.

39. McGinty RK, Henrici RC, Tan S. Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome. Nature 2014;514:591-6.

40. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein and protein-DNA complexes. J Med Chem 2005; 48:2325-35.

41. Sahtoe DD, van Dijk WJ, El Oualid F, Ekkebus R, Ovaa H, Sixma TK. Mechanism of UCH-L5 activation and inhibition by DEUBAD domains in RPN13 and INO80G. Mol Cell 2015;57:887-900.

42. Zambolin S, Clantin B, Chami M, Hoos S, Haouz A, Villeret V, Delepelaire P. Structural basis for haem piracy from host haemopexin by Haemophilus influenzae. Nat Commun 2016;7:11590.

43. Zhang L, Hermans J. Hydrophilicity of cavities in proteins. Proteins 1996;24:433-8.

44. Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Rie Lee G, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RA, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JP, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond AS, Visscher K, Kastritis PL, Bonvin AM, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jimenez-Garcia B, Moal IH, Fernandez-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Marze N, Kuroda D,

Roy Burman SS, Gray JJ, Chermak E, Cavallo L, Oliva R, Tovchigrechko A, Wodak SJ. Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins 2016; 84 Suppl 1:323-48.

FIGURE LEGENDS

Figure 1. Successful models submitted to CAPRI. (A) The modeling scheme for T60-64. We first aligned and characterized the binding site bound to other peptides in the PDB, before identifying putative alignments with the targets. We then modeled the peptides by stitching together peptide fragments, building side-chains in the context of the bound PDB structures, minimizing, scoring and filtering. (B) Predicted poses and pose qualities for targets T60-64. (C) Representation of our best models for targets T66, T67, T97, T103, T104 and T105. For each target, receptors are superimposed and shown in white. Ligand in our best model as predictors is shown in red, as servers in yellow, and as scorers in blue. For comparison, the structure of the experimental complex (if available) is represented in green.

Table I

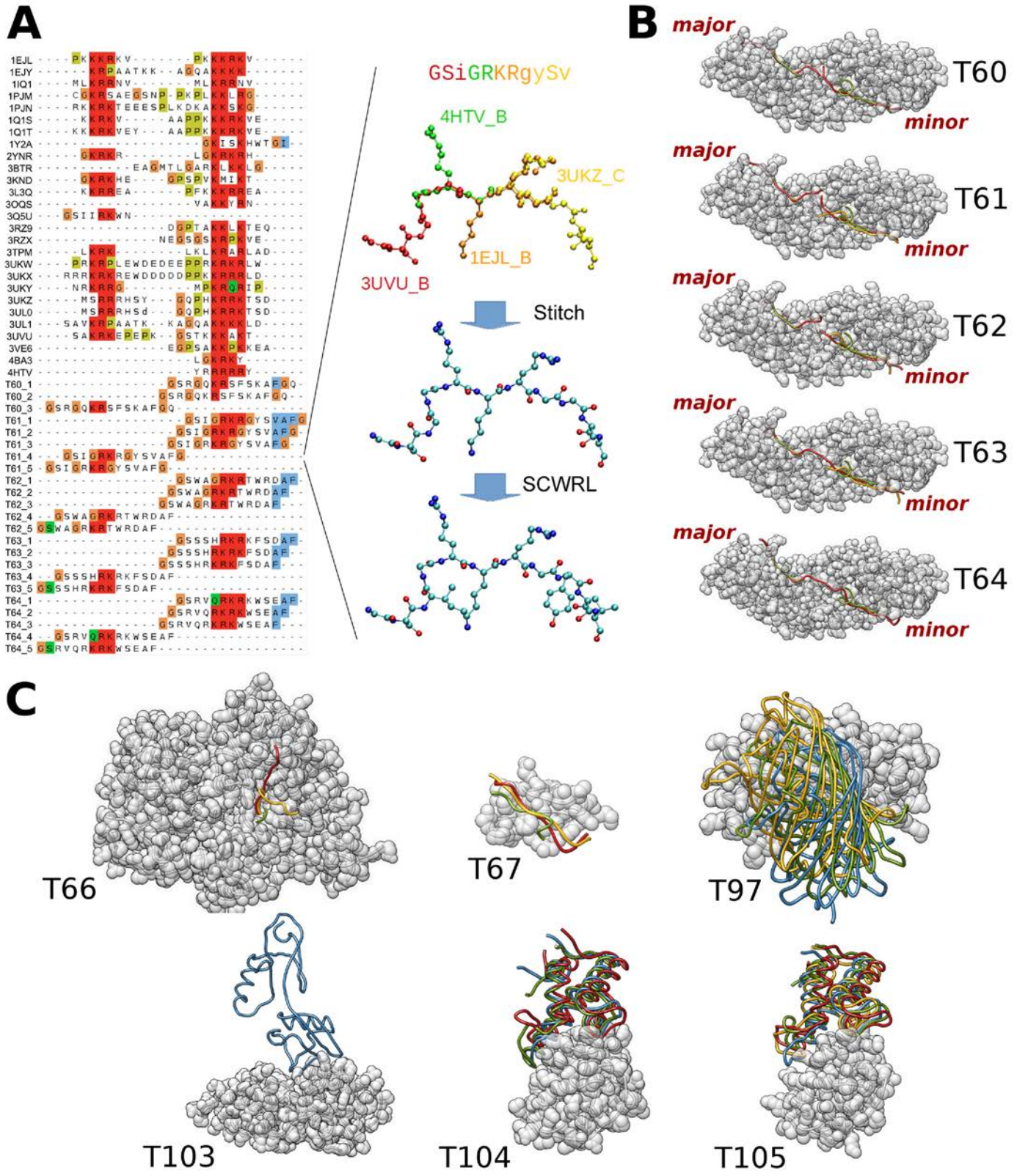
Summary of the evaluation results of our participation in the 6th CAPRI edition.

#Target[a]	Type[b]	Predictors[c]	Servers[c]	Scorers[c]
T59	Prot/Prot (U/H)	0	0	0
T60-64 (Major binding site)	Prot/Pep (U/H)	**M02** (T60)** **M03** (T61)**, M04* (T61) **M03*** (T62)**, M04* (T62) **M03** (T63)**, M04* (T63) **M03** (T64)**, M04* (T64)	**M08* (T63)**	*N/A*
T60-64 (Minor binding site)	Prot/Pep (U/H)	**M01* (T63)**	0	*N/A*
T60-64 (Minor binding site, 6-residue)	Prot/Pep (U/H)	**M01** (T60)** **M01** (T61)** **M01** (T62)** **M01** (T63)** **M01** (T64)**	**M03* (T61)** **M02** (T62)**, M08** (T62), **M06* (T62)**, M10* (T62) **M01* (T63)**, M07* (T63)	*N/A*
T65	Prot/Pep (U/H)	0	0	*N/A*
T66	Prot/Pep (U/H)	**M01* (EF)**	**M04* (EF)**, M08* (EF)	*N/A*
T67	Prot/Pep (U/H)	**M10* (all)** M01-M10* (PPxY)	**M06* (PPxY)**, M07* (PPxY), M09* (PPxY)	*N/A*
T95	Prot-DNA/Prot (U/U)	0	X	*N/A*
T96	Prot/Prot (H/H)	0	0	0
T97	Prot/Prot (H/H)	0	**M10***	M05*, M09*, M10*
T98	Prot/Prot (U/U)	0	X	0
T99	Prot/Prot (U/U)	0	X	0
T100	Prot/Prot (U/H)	0	X	0
T101	Prot/Prot (U/H)	0	X	0
T103	Prot/Prot (H/H)	0	0	**M03* (Ct)**, M05* (Ct)
104	Prot/Prot (H/H)	M03*, **M06****, M07*, M10**	0	**M01**-02*****, M03**, M04***, M05**, M06***, M07-08**, M09***, M10**
105	Prot/Prot (H/H)	**M02****, M10**	M02*, **M06****	**M01****, M02*, M03-05**, M07-10**
107	Prot/Prot (U/U)	0	0	0

[a] Underscored: target of special difficulty, with only 3 or fewer groups that submitted correct models.

[b] B: bound; U: unbound; H: homology-based model.

[c] Correct models submitted to CAPRI by our group. Each model is numbered according to its rank within our submission. The quality of each model is indicated, following CAPRI criteria: * acceptable; ** medium quality; *** high quality. In bold, our best-quality model for each target. "0": no correct model submitted. "X": no submissions. "N/A": experiment not available (i.e. target was not proposed for the scorer experiment).


A
GSiGRKRgySv
4HTV_B
3UKZ_C
1EJL_B
3UVU_B
Stitch
SCWRL
B
major
minor
T60
T61
T62
T63
T64
C
T66
T67
T97
T103
T104
T105