Data-Driven Docking: HADDOCK's Adventures in CAPRI

A. D. J. van Dijk,[1] S. J. de Vries,[1] C. Dominguez,[1†] H. Chen,[2] H.-X. Zhou,[3] and A. M. J. J. Bonvin[1*]
[1]*Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands*
[2]*Department of Physics, Drexel University, Philadelphia, Pennsylvania*
[3]*Department of Physics and Institute of Molecular Biophysics and School of Computational Science, Florida State University, Tallahassee, Florida*

ABSTRACT **We have shown previously that given high-resolution structures of the unbound molecules, structure determination of protein complexes is possible by including biochemical and/or biophysical data as highly ambiguous distance restraints in a docking approach. We applied this method, implemented in the HADDOCK (High Ambiguity Driven DOCKing) package (Dominguez et al., J Am Chem Soc 2003;125:1731–1737), to the targets in the fourth and fifth rounds of CAPRI. Here we describe our results and analyze them in detail. Special attention is given to the role of flexibility in our docking method and the way in which this improves the docking results. We describe extensions to our approach that were developed as a direct result of our participation in CAPRI. In addition to experimental information, we also included interface residue predictions from PPISP (Protein–Protein Interaction Site Predictor; Zhou and Shan, Proteins 2001;44:336–343), a neural network method. Using HADDOCK we were able to generate acceptable structures for 6 of the 8 targets, and to submit at least 1 acceptable structure for 5 of them. Of these 5 submissions, 3 were of medium quality (Targets 10, 11, and 15) and 2 of high quality (Targets 13 and 14). In all cases, predictions were obtained containing at least 40% of the correct epitope at the interface for both ligand and receptor simultaneously. Proteins 2005;60:232–238.** © 2005 Wiley-Liss, Inc.

Key words: flexible docking; HADDOCK; biomolecular complexes; interface prediction

INTRODUCTION

Biochemical and biophysical experiments such as mutagenesis, NMR, and mass spectrometry are widely used to gain insight into biomolecular interactions. The information generated in this way can in principle be used to model the structure of the corresponding complex when conventional NMR and crystallographic approaches fail. Taking the step from experimental data to modeling is, however, not common practice. This can be done using docking approaches that model the structure of a complex based on the structure of the constituents. Although clear progress has been achieved in the field of "ab-initio docking," as illustrated by the previous rounds of CAPRI,[1] most current approaches have difficulties in generating consistently reliable predictions. However, as highlighted in a recent review,[2] in many cases of biological interest, some kind of experimental information is available that can be used to filter docking solutions or even to drive the docking. We developed for this purpose the data-driven docking method HADDOCK that can incorporate any kind of information about interface residues.[3] Using HADDOCK, we participated in Rounds 4 and 5 of CAPRI. Our method and its performance within CAPRI are presented here.

MATERIALS AND METHODS

HADDOCK Data

Experimental and/or prediction information is incorporated in HADDOCK by defining active residues (which, based on the data, are supposed to be part of the interface) and passive residues (surface neighbors of active residues). The docking is driven by ambiguous interaction restraints (AIRs) defined between any atom of the active residues and all atoms of all active and passive residues on the partner protein.[3] In cases where the data were very fuzzy or judged unreliable, we randomly removed 25% of the data for each docking trial. In the absence of experimental data, we experimented with the use of all accessible residues as active and/or passive residues in the combinations: A-active–B-passive, A-passive–B-active, and both A- and B-active.

HADDOCK Flexible Docking Protocol

Flexibility is introduced at several levels in the algorithm:

1. By docking from ensembles of structures and taking all possible pairwise combinations.
2. By introduction of flexibility in the side chain at the interface.

Grant sponsor: National Institutes of Health; Grant number: GM58187 (to H.-X. Zhou). Grant sponsor: Netherlands Organization for Scientific Research (NWO) "Jonge Chemici" grant (to A. M. J. J. Bonvin).

†Current address: Institute for Molecular Biology and Biophysics, ETH Zurich, 8093 Zurich, Switzerland.

*Correspondence to: A. M. J. J. Bonvin, Department of NMR Spectroscopy, Bijvoet Center for Biomolecular Research, Utrecht University, 3584, Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@chem.uu.nl

Received 13 January 2005; Accepted 4 February 2005

DOI: 10.1002/prot.20563

3. By allowing both side-chain and backbone flexibility in the final refinement stage.

This is then followed by a final refinement in explicit solvent.

In the rigid-body docking stage we typically save to disk 1000 to 2000 solutions. The best 200 are then subjected to a refinement step consisting of 3 consecutive simulated annealings, first treating the molecules as rigid bodies and then with introduction of flexibility (steps 1 and 3 above). For details, see Dominguez et al.[3]

For each target we typically performed a number of docking runs with various definitions of the AIRs. The initial scoring was based on the sum of intermolecular van der Waals, electrostatic, and AIRs energies. The nonbonded energies are calculated with a 8.5 Å cutoff using the OPLS parameters.[4] For CAPRI Round 5, we set epsilon to 10 during the vacuum part of the docking (rigid-body docking and semiflexible refinement). The electrostatic energy contribution was scaled down by a factor 0.1 in the final scoring after water refinement. The scoring was performed on a cluster basis considering only the best 5 to 10 structures of each cluster to remove cluster size effects. The clustering is based on pairwise root-mean-square deviations (RMSDs). The lowest energy structure of the lowest energy cluster is considered the highest ranking solution. Due to the inclusion of experimental information, only a limited number of solutions is typically obtained (<25). For CAPRI, the selection from the clusters from different docking runs was performed manually based on energy considerations and visual inspection.

Interface Predictions

In PPISP, sequence profiles produced by PSI-BLAST and solvent accessibility of spatially neighboring surface residues calculated by Dictionary of Protein Secondary Structures[5] (DSSP) were used as input to a neural network.[6] The network was trained on interface residues collected from the protein–protein complexes in the Protein Data Bank (PDB). A residue is considered as a surface residue if at least 10% of its surface area is solvent accessible. A surface residue is considered as an interface residue if at least 1 of its heavy atoms is within 5 Å of a heavy atom of the partner protein. Since the original publication of PPISP, extensive improvements have been made (Chen and Zhou, to appear in Proteins). The training set now consists of 1156 protein chains with less than 30% identity. These chains each have at least 20 interface residues. The training set contains a total of 225,238 surface residues, of which 52,624 (or 23%) are interface residues. Test on a different set of 100 nonhomologous protein chains shows that the overall prediction accuracy is ~80%. One problem was that interface residues were overpredicted for some proteins but underpredicted for others. To tackle this problem, we developed a consensus method that combines predictions from a series of neural networks with different levels of accuracy and coverage of native interface residues (F_{IR}). Usually a neural network with higher F_{IR} predicts more interface residues but with lower accuracy, and vice versa. Each predicted interface residue was ranked by consensus score (the number of neural networks predicting it as interface) and the top-ranked ones were mapped onto the protein surface. The residues clustered together were collected. This process was stopped early if there were enough interface residues collected (to prevent overprediction), or extended to less confidently predicted residues if there were only a few predictions made (to avoid underprediction).

RESULTS AND DISCUSSION

HADDOCK Data

Information to drive the docking process was derived from literature searches, in combination with interface residue predictions by PPISP[6] (see below). This information is highly ambiguous in the sense that it gives information about the interface but not about contacts made across it. In order to use such information to drive the docking, we distinguish between "active" and "passive" residues: Active residues correspond to solvent accessible, experimentally identified or predicted interface residues, while passive residues correspond to their surface neighbors. These are used to define AIRs between each active residue of one chain and all active and passive residues of the other chain (for further discussion, see Dominguez et al.[3]).

Performance of the PPISP Interface Residues Predictions

PPISP is a neural network based method that is trained on known structures of protein complexes. The input of the neural network consists of sequence profiles and solvent accessibility of spatially neighboring residues. Sequence profiles, obtained from multiple sequence alignment by PSI-BLAST, capture characteristics of interface residues such as conservation and hydrophobicity. The PPISP interface predictions for CAPRI Targets 4 and 5 are summarized in Table I. For a total of 11 proteins in the targets (no prediction was necessary for the 2 antibodies), 4 [tick-borne encephalitis virus (TBEV) monomer of T10, cohesin of T11/T12, myosin phosphatase–targeting subunit (MYPT1) of T14, and colicin D of T15] had higher than 50% prediction accuracy, and 2 others (dockerin of T11/T12 and the xylanase of T18) had very good values for F_{IR} (1 and 0.75, respectively) but moderate accuracy. Interface predictions were poor for the 2 antigens (in T13 and T19) and 2 large proteins [protein phosphatase-1 (PP1) of T14 and *Triticum aestivum* xylanase inhibitor (TAXI) of T18]. These predictions were used to defined AIRs (see above) to drive the docking.

HADDOCK Results for CAPRI Targets 4 and 5

For 5 of the 8 targets (including the canceled Target 15), we had at least 1 acceptable solution in our submissions; of these, 3 were of medium quality and 2 of high quality. The corresponding models superimposed on their respective targets are shown in Figure 1 (the still unpublished Target 13 is shown as a cartoon). The ranking we report for the various targets is based on the fraction of correct native contacts. In all 5 cases, we ended within the top 10 of all

TABLE I. Results of Interface Residue Prediction on CAPRI Targets

Target	Chain	Seq. length	#Native int. res.	PPISP predictions			HADDOCK best model	
				#Pred.	F_{IR}[a]	Accuracy[b]	F_{IR}[a]	Accuracy[b]
10	TBE monomer	381	95	19	0.11	0.53	0.44	0.90
11/12	Cohesin	140	25	13	0.48	0.92	0.89	0.91
	Dockerin	56	13	32	1	0.41	0.74	0.67
13	Antigen	245	25	5	0	0.0	0.88	0.79
14	PP1	309	49	10	0	0.0	0.81	0.85
	MYPT1	291	60	36	0.32	0.53	0.67	0.85
15	Colicin D	107	22	20	0.59	0.65	0.85	0.85
	Immunity protein	87	19	14	0.11	0.14	0.85	0.79
18	TAXI	370	23	25	0.04	0.04	0.42	0.38
	Aspergillus niger xylanase	182	24	51	0.75	0.35	0.62	0.71
19	Ovine prion	102	21	15	0.14	0.20	0.44	0.42

[a]F_{IR}, fraction of correctly predicted (by PPISP) or observed (HADDOCK best) interface residues among all native interface residues. Native interface residues were defined as those forming native contacts (<5 Å).
[b]Accuracy, fraction of correct interface residues among all predicted (by PPISP) or observed (HADDOCK best model) interface residues.

submissions. In 3 cases, our best prediction was correctly ranked number 1 within our 10 submissions. In the following, we discuss our results for the individual targets.

Target 10

For Target 10,[7] we modified the docking protocol in order to be able to deal with 3 molecules and impose C3 symmetry. This was done by using symmetry restraints as introduced for NMR structure calculation of symmetrical dimers.[11] To impose C3 symmetry, we defined triplets of distance pairs (AB/BC, BC/AC, and CA/AB), requiring that the distances must be equal within a pair. We used in addition noncrystallographic symmetry restraints in the Crystallography & NMR System[12] (CNS), which ensure that the molecules are similar without defining any symmetry operation between them.

The AIRs used to drive the docking were derived from various sources[13–16] and included epitope mapping, mutagenesis, and sequence conservation data, in combination with interface predictions.

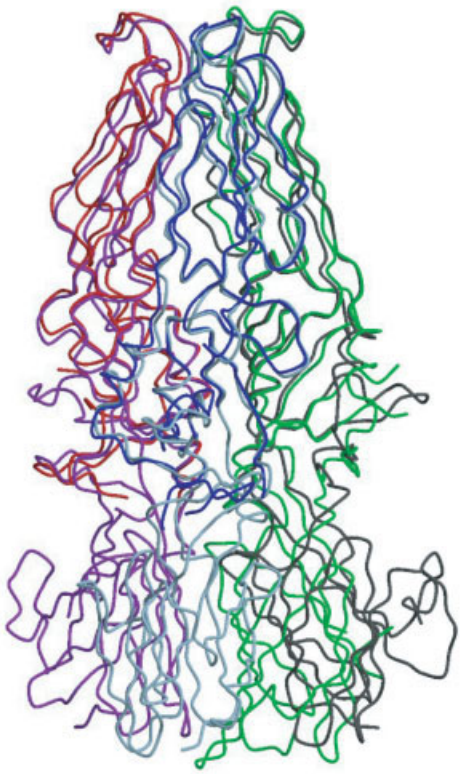
The results of the docking could be partitioned into 2 classes: one with rather "flat" triangular arrangements of the 3 monomers, and the other consisting of "spikes." We based our selection on electron microscopy data[17] that suggested a triangular flat shape of the timer. Accordingly, we mainly selected such structures and only put one "spike" conformation in our submission, which we ranked tenth. The latter turned out to be a very close prediction, with a ligand RMSD (l-RMSD) of 2.9 Å, an interface RMSD (i-RMSD) of 1.9 Å and a fraction of native contacts of 0.3. It is worth noting that this "spike" solution had the best intermolecular energy but that misjudgment of literature data led us to rank it last.

Target 11/12

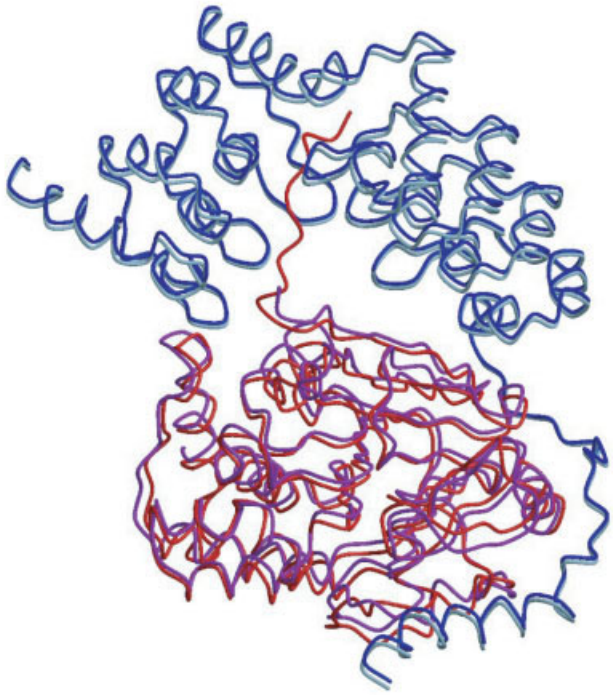
For Target 11,[8] a homology model of dockerin was built using SWISSMODEL[18] and WHATIF.[19] The model was subjected to a 200 ps molecular dynamics (MD) simulation in explicit solvent using Gromacs3.1.4.[20] The starting structure plus 10 structures taken every 20 ps were selected as input for the docking. The backbone RMSD of our starting homology model from the bound form was 4.3 Å, whereas it ranged from 4.1 Å to 4.8 Å for the structures taken from the MD trajectory. These values however decrease to 2.3 Å for the model and 2.2–2.9 Å for the MD structures if only the 2 Ser/Thr pairs that are known to be important for the binding are considered. This means that some of the MD structures had moved closer to the bound form than the homology model itself. The unbound form of cohesin was also subjected to 200 ps MD in explicit solvent. Again, 11 structures were selected for the docking. These were also used for Target 12.

To drive the docking, we used previously published mutagenesis data,[21–24] in particular, 2 amino acid pairs Ser11/Thr12 and Ser45/Thr46, and other predicted interface residues. As discussed in the article describing the experimental crystal structure of the complex,[8] the dockerin sequence contains a tandem repeat, with residues 1–23 showing high homology to residues 35–57. These 2 stretches of sequences also adopt very similar 3-dimensional (3D) structures (main-chain atom RMSD of 0.36 Å). Moreover, dockerin contains near perfect internal 2-fold symmetry, such that residues 1–22 overlay onto residues 35–56, and vice versa (see Fig. 5 in Carvahlo et al.[8]). This symmetry is also reflected in the mutagenesis data that we used for docking. As stated,[8] it would seem likely that both symmetry-related halves of dockerin could interact with cohesin in almost identical manners. The accuracy of our docking results, especially for Target 12, were mainly hampered by the fact that both sites were included in our restraints, while only the second one is actually involved in binding in the crystal structure. Since both sites were defined, we also obtained several docking solutions corresponding to a 180° rotated binding mode.

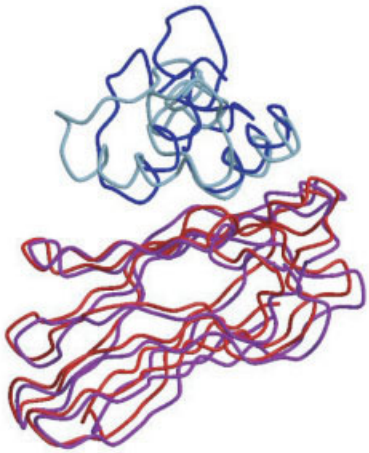
For Target 11, we nevertheless had 1 medium and 3 acceptable solutions in our submission (with the best one having an l-RMSD of 6.0 Å, an i-RMSD of 2.0 Å, and a fraction of native contacts (F_{nat}) of 0.4, ranking at the third position overall). For Target 12, no acceptable solutions were submitted, although acceptable ones within 2.3 Å



T10: TBE virus envelope protein (**, #10)
l-RMSD 2.9 Å / i-RMSD 1.9 Å / F_{nat} 0.3

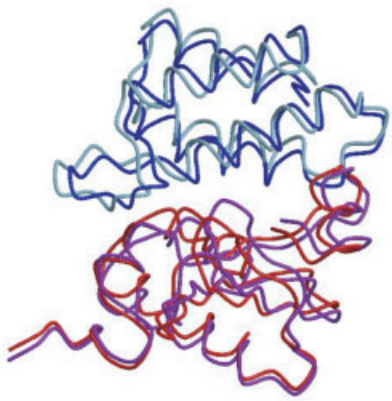


T14: PP1-MYPT1 (***, #1)
l-RMSD 2.3 Å / i-RMSD 0.96 Å / F_{nat} 0.6



T11: cohesin - dockerin (**, #10)
l-RMSD 6.0 Å / i-RMSD 2.0 Å / F_{nat} 0.4

T13: (***, #1)
l-RMSD 3.8 Å
i-RMSD 0.3 Å
F_{nat} 0.8



T15: colicin D - immunity protein (**, #1)
l-RMSD 5.4 Å / i-RMSD 1.8 Å / F_{nat} 0.5

Fig. 1. Best HADDOCK models and corresponding experimental structures. Closest models overlaid with corresponding crystal structures for Target 10 (PDB code: 1urz[7]), Target 11 (PDB code: 1ohz[8]), Target 14 (PDB code: 1s70[9]), and Target 15 (PDB code: 1v74[10]). Color coding: red and blue for crystal structures, purple and light-blue for model; for Target 10 in addition: green for crystal structure, black for model. Note that Target 10 is a trimer; for clarity, for 2 of its (C3-symmetric) chains, we show only part of the chain. Segments 148–159 and 204–209 are missing in the crystal structure. The still unpublished Target 13 is shown as a cartoon. The quality of our best model (stars) and its rank within our 10 submissions is indicated between brackets. The CAPRI quality criteria are: "high" (***): $F_{nat} > 0.5$, l-RMSD or i-RMSD < 1.0 Å and "medium" (**): $F_{nat} > 0.3$, l-RMSD < 5.0 Å or i-RMSD < 2.0 Å, where F_{nat} is the fraction of native contacts, l-RMSD is the ligand backbone RMSD from the target, after superimposition on the receptor, and i-RMSD the interface residues backbone RMSD from the target.

interface RMSD were generated. However, if the symmetry-related binding mode were considered, all our submissions for Targets 11 and 12 would be within 10 Å l-RMSD.

Target 13

For Target 13, the bound form of the antibody and the unbound form of the antigen were provided. The latter was subjected to a 200 ps MD in explicit solvent, and 11 structures were selected as input for the docking as described for Targets 11/12. In this case, the unbound form was 0.5 Å away from the bound form, and it drifted away to 0.9 Å during the MD simulation.

For the antibody, we defined the complementarity determining region (CDR) as active residues. For the antigen, several epitopes were identified from literature data.[25–27] This information was introduced differently in various runs: either using all epitopes simultaneously or only a subset of them to define AIRs. We also performed a docking run in which all accessible residues of the antigen were defined as passive.

The clusters from all runs were pooled and ranked mainly based on favorable values of van der Waals energy and on packing arguments after visual inspection. Our number 1 submission was of high-quality (l-RMSD = 3.8 Å; i-RMSD = 0.3 Å; F_{nat} = 0.8).

Target 14

The docking was performed from the bound form of MYPT1, and a homology model of PP1 created using SWISSMODEL[18] and WHATIF.[19] For this target,[9] excellent information was available from literature,[28–30] including a crystal structure of the RRVSFA peptide bound to PP1. In addition, interface predictions from PPISP involving the N-terminal helix of MYPT1 were used. Because of the rather closed, "embracing" conformation of MYPT1, which might prevent a proper rigid-body docking, we turned off the intermolecular interactions of PP1 with the C-terminal residues of MYPT1 starting from residue 52 and scaled down the intermolecular interactions by a factor 0.01. Interaction with the C-terminal residues was reintroduced in the second flexible refinement stage of the docking. As the quality of the data in this case was really high, it was not surprising that 1 of our predictions was of high quality (l-RMSD = 2.3 Å; i-RMSD = 0.96 Å; F_{nat} = 0.6), with 2 additional acceptable ones out of our 10 submitted structures.

Target 15

Target 15[10] was a bound docking case, but with shaved surface side-chains. The latter were generated with CNS,[12] and the structures were submitted to a short refinement in explicit solvent,[31] with position restraints on the backbone. For each protein, 10 structures were generated and used as starting point for the docking. Only few data were found in the literature.[32] We used as always predicted interface residues but also experimented (as for Target 13, but now for both proteins), taking all accessible residues as passive and/or active, in the combinations active–passive, passive–active, and active–active. The results of these 3 docking runs were analyzed in terms of the frequency that a given residue contacts the partner protein. The top 10% most frequently found residues at the interface were then selected for a new docking run.

Our number 1 submission, resulting from a run with the latter restraints, was a close hit (l-RMSD = 5.4 Å; i-RMSD = 1.8 Å; F_{nat} = 0.5) corresponding to a medium-quality prediction. Out of the 3 runs with a total of 600 structures from which we selected our submissions, only 5 structures had l-RMSD below 10 Å (resulting from the runs with all accessible residues on the immunity protein as passive). Our selection criterion (based on E_{vdw} + 0.1 E_{elec}) thus correctly led to the best conformation in this case.

Target 18

For Target 18,[33] the predicted interface residues were quite good for the xylanase, with 75% of the epitope correctly predicted (see F_{IR}, fraction of correctly predicted native interface residues, in Table I). In addition, some mutagenesis information was available.[34] However, for TAXI, the interface prediction was poor (F_{IR} = 0.04), and we did not find any experimental data for defining the interface. Although we tried the same kind of active/passive protocols as for Target 15, in this case, our docking runs did not generate any acceptable solution. We did, however, obtain solutions with more than 42% of the native epitope of each partner simultaneously at the interface, but these corresponded to rotated solutions.

Target 19

For Target 19,[35] the ovine prion was modeled as described above from the NMR entry 1DWY.[36] We found various epitopes in the literature,[37,38] which we used again in various combinations of active and passive residues, together with the CDR residues of the antibody. A docking run was also performed using all accessible residues of the antigen. However, no acceptable solution was obtained. As for Target 18, our submissions, however, contained rotated solutions containing as much as 44% and 79% of the correct epitopes at the interface for the antibody and the antigen, respectively.

The Effect of Flexibility on Our Docking Results

In our docking protocol, flexibility is introduced stepwise, first for interface side-chains and then for both backbone and side-chains at the interface (see Materials and Methods section). To analyze in detail if it is worth paying the additional computational price, we compared various quality parameters for structures after rigid-body docking and after final refinement in explicit solvent (Table II). Although there is some variability between targets, it is clear that the number of good solutions as monitored by l-RMSDs increases when comparing the rigid-body docking results with the results after flexibility has been introduced. The quality of the prediction also improves significantly, as can clearly be seen from the fraction of native contacts. Finally, flexibility also considerably improved the ranking of structures.

TABLE II. Impact of Flexibility on Docking Results

Target	10	11	12	13	14	15
Number of solutions within the given 1-RMSD range: rigid body/refined[a]						
0–5 Å	9/8	36/63	33/36	26/26	11/12	2/3
Average fraction of native contacts[a]						
Rigid body: 0–5 Å	0.14 ± 0.05	0.27 ± 0.11	0.11 ± 0.03	0.57 ± 0.07	0.54 ± 0.15	0.45 ± 0.07
Refined: 0–5 Å	0.18 ± 0.09	0.36 ± 0.11	0.12 ± 0.03	0.74 ± 0.07	0.50 ± 0.10	0.35 ± 0.13
Ranking of best submitted structure: rigid body/refined[b]						
	6/3	181/51	—	36/1	23/2	8/2

[a]For Target 11, the RMSD range is 5–7.5 Å; for Targets 12 and 15, it is 5–10 Å.
[b]Ranking of best submitted structure according to our scoring scheme. Note that lower ranking indicates better result.

In addition to the explicit inclusion of flexibility in the refinement stage, we also implicitly included it in the rigid-body docking stage by starting from ensembles of structures obtained from short MD simulations in explicit solvent. Depending on the quality of the starting model, the RMSD to the bound form increased or decreased during the MD. Still, we often observed that some conformations preferentially lead to better docking solutions. This effect must originate from better side-chain conformations, since large backbone conformational changes cannot be expected within such short simulation times (200 ps). For example, for Target 11, our best model originated from MD structures taken at 40 ps for cohesin (backbone RMSD to bound form was 0.5 Å) and 200 ps for dockerin (backbone RMSD to bound form was 4.6 Å). For Target 13, good solutions were obtained mainly from the unbound form of the antigen (backbone RMSD to bound form was 0.5 Å) and from a MD snapshot taken at 140 ps (backbone RMSD to bound form was 0.75 Å).

What Did We Learn From CAPRI?

CAPRI was a very good stimulus to develop new tools, as well as a possibility to validate new features in an unbiased way. As explained above, the fact that we did often not have very reliable experimental data (or no data at all) inspired us to experiment with all accessible residues as active–active, active–passive, or passive–active. For Targets 13 and 15, this was successful, while for Targets 18 and 19 this did not generate any acceptable solution. Our solutions for the latter 2 cover, however, the correct epitopes but correspond to rotated solutions. Although there are several possible reasons for this difference, it might be related to the fact that Targets 18 and 19 consisted of bigger proteins, which means that the configuration space that needs to be covered is larger; it is therefore possible that our sampling was insufficient. In cases where data were fuzzy or scarce, we used all accessible residues and also experimented with the random removal of a fraction (typically 25%) of our data for each solution that was generated. In this way, the fuzziness of the interface definition is increased, and wrong or inconsistent data need not have a disastrous influence. We started using this feature after the results for Round 4 were known, when we realized the effect that the wrong definition of active residues for dockerin had on our docking of Target 12. Note that random removal of as much as 50% of the restraints leads to high-quality predictions for Target 12 (data not shown). Considering the interface predictions, the blind test results on the CAPRI targets were consistent with results on other test sets (Chen and Zhou, to appear in Proteins). In general, predictions for small proteins are much better than for large proteins. Complexes of small proteins on which our training set is based are indeed well represented in the PDB. In terms of classes of protein complexes, the PPISP method generally makes good prediction for enzyme–inhibitor interfaces but is not suited for antigen–antibody interactions. The former complexes have presumably evolved over time to optimize the interface. In contrast, antigen–antibody interactions are not subjected to evolutionary optimization. A strength of PPISP is that it is relatively insensitive to conformational changes accompanying complex formation. Results on CAPRI targets also showed that PPISP performed equally well with homology models (dockerin in Target 11).

We also analyzed in detail the influence of flexibility on our docking results. As described here and in related studies,[39,40] flexibility is a very important factor for docking. Indeed, we found that in general the docking results improve after the flexible refinement stages. The largest improvement is found in the fraction of native contacts, while the effect of flexibility on the RMSDs from the target is the most pronounced when the unbound form is further away from the bound form. Finally, we also observed that flexibility improves the ranking of correct solutions. Taken all together, this indicates that the inclusion of flexibility in docking is clearly beneficial, even if it increases the required computational time.

Our participation to CAPRI revealed both the strengths and weaknesses of our data-driven docking approach HADDOCK. Good results can be expected when the data are of high quality. Using very fuzzy or ambiguous data, HADDOCK sometimes still generated acceptable results (as for Targets 13 and 15), but in other cases failed to do so (Targets 18 and 19). The scoring of flexible docking solutions, which have a significantly larger number of degrees of freedom, remains a difficult process that can clearly be improved in the future.

REFERENCES

1. Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJE, Vajda S, Vasker I, Wodak SJ. CAPRI: A Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.

2. van Dijk ADJ, Boelens R, Bonvin AMJJ. Data-driven docking for the study of biomolecular complexes. FEBS J 2005;272:293–312.
3. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A protein–protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125:1731–1737.
4. Jorgensen WL, Tirado-rives J. The OPLS Potential functions for proteins: energy minimizations for crystals of cyclin peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
5. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
6. Zhou HX, Shan YB. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
7. Bressanelli S, Stiasny K, Allison SL, Stura EA, Duquerroy S, Lescar J, Heinz FX, Rey FA. Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. EMBO J 2004;23:728–738.
8. Carvalho AL, Dias FMV, Prates JAM, Nagy T, Gilbert HJ, Davies GJ, Ferreira LMA, Romao MJ, Fontes CMGA. Cellulosome assembly revealed by the crystal structure of the cohesin–dockerin complex. Proc Natl Acad Sci USA 2003;100:13809–13814.
9. Terrak M, Kerff F, Langsetmo K, Tao T, Dominguez R. Structural basis of protein phosphatase 1 regulation. Nature 2004;429:780–784.
10. Graille M, Mora L, Buckingham RH, van Tilbeurgh H, de Zamaroczy M. Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. EMBO J 2004;23:1474–1482.
11. Nilges M. A calculation strategy for the structure determination of symmetrical dimers by H-1-NMR. Proteins 1993;17:297–309.
12. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR System: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54:905–921.
13. Stiasny K, Allison SL, MarchlerBauer A, Kunz C, Heinz FX. Structural requirements for low-pH-induced rearrangements in the envelope glycoprotein of tick-borne encephalitis virus. J Virol 1996;70:8142–8147.
14. Allison SL, Schalich J, Stiasny K, Mandl CW, Heinz FX. Mutational evidence for an internal fusion peptide in flavivirus envelope protein E. J Virol 2001;75:4268–4275.
15. Stiasny K, Allison SL, Schalich J, Heinz FX. Membrane interactions of the tick-borne encephalitis virus fusion protein E at low pH. J Virol 2002;76:3784–3790.
16. Allison SL, Stiasny K, Stadler K, Mandl CW, Heinz FX. Mapping of functional elements in the stem-anchor region of tick-borne encephalitis virus envelope protein E. J Virol 1999;73:5605–5612.
17. Ferlenghi I, Clarke M, Ruttan T, Allison SL, Schalich J, Heinz FX, Harrison SC, Rey FA, Fuller SD. Molecular organization of a recombinant subviral particle from tick-borne encephalitis. Mol Cell 2001;7:593–602.
18. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–3385.
19. Rodriguez R, Chinea G, Lopez N, Pons T, Vriend G. Homology modeling, model and software evaluation: three related resources. Bioinformatics 1998;14:523–528.
20. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 2001;7:306–317.
21. Miras I, Schaeffer F, Beguin P, Alzari PM. Mapping by site-directed mutagenesis of the region responsible for cohesin–dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. Biochemistry 2002;41:2115–2119.
22. Schaeffer F, Matuschek M, Guglielmi G, Miras I, Alzari PM, Beguin P. Duplicated dockerin subdomains of *Clostridium thermocellum* endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity. Biochemistry 2002;41:2106–2114.
23. Mechaly A, Fierobe HP, Belaich A, Belaich JP, Lamed R, Shoham Y, Bayer EA. Cohesin–dockerin interaction in cellulosome assembly—a single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. J Biol Chem 2001;276:9883–9888.
24. Mechaly A, Yaron S, Lamed R, Fierobe HP, Belaich A, Belaich JP, Shoham Y, Bayer EA. Cohesin–dockerin recognition in cellulosome assembly: experiment versus hypothesis. Proteins 2000;39:170–177.
25. Godard I, Estaquier J, Zenner L, Bossus M, Auriault C, Darcy F, Grasmasse H, Capron A. Antigenicity and immunogenicity of P-30-derived peptides in experimental-models of toxoplasmosis. Mol Immunol 1994;31:1353–1363.
26. Velge-Roussel F, Dimier-Poisson I, Buzoni-Gatel D, Bout D. Anti-SAG1 peptide antibodies inhibit the penetration of *Toxoplasma gondii* tachyzoites into enterocyte cell lines. Parasitology 2001;123:225–233.
27. Velge-Roussel F, Chardes T, Mevelec P, Brillard M, Hoebeke J, Bout D. Epitopic Analysis of the *Toxoplasma gondii* major surface-antigen SAG1. Mol Biochem Parasit 1994;66:31–38.
28. Egloff MP, Johnson DF, Moorhead G, Cohen PTW, Cohen P, Barford D. Structural basis for the recognition of regulatory subunits by the catalytic subunit of protein phosphatase 1. EMBO J 1997;16:1876–1887.
29. Toth A, Kiss E, Herberg FW, Gergely P, Hartshorne DJ, Erdodi F. Study of the subunit interactions in myosin phosphatase by surface plasmon resonance. Eur J Biochem 2000;267:1687–1697.
30. Hirano K, Phan BC, Hartshorne DJ. Interactions of the subunits of smooth muscle myosin phosphatase. J Biol Chem 1997;272:3683–3688.
31. Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ, Nilges M. Refinement of protein structures in explicit solvent. Proteins 2003;50:496–506.
32. de Zamaroczy M, Mora L, Lecuyer A, Geli V, Buckingham RH. Cleavage of colicin D is necessary for cell killing and requires the inner membrane peptidase LepB. Mol Cell 2001;8:159–168.
33. Sansen S, De Ranter CJ, Gebruers K, Brijs K, Courtin CM, Delcour JA, Rabijns A. Structural basis for inhibition of *Aspergillus niger* xylanase by *Triticum aestivum* xylanase inhibitor-I. J Biol Chem 2004;279:36022–36028.
34. Gebruers K, Brijs K, Courtin CM, Fierens K, Goesaert H, Rabijns A, Raedschelders G, Robben J, Sansen S, Sorensen JF, Van Campenhout S, Delcour JA. Properties of TAXI-type endoxylanase inhibitors. Biochim Biophys Acta 2004;1696:213–221.
35. Eghiaian F, Grosclaude J, Lesceu S, Debey P, Doublet B, Treguer E, Rezaei H, Knossow M. Insight into the PrPC→PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. Proc Natl Acad Sci USA 2004;101:10254–10259.
36. Garcia FL, Zahn R, Riek R, Wuthrich K. NMR structure of the bovine prion protein. Proc Natl Acad Sci USA 2000;97:8334–8339.
37. Peretz D, Williamson RA, Matsunaga Y, Serban H, Pinilla C, Bastidas RB, Rozenshteyn R, James TL, Houghten RA, Cohen FE, Prusiner SB, Burton DR. A conformational transition at the N terminus of the prion protein features in formation of the scrapie isoform. J Mol Biol 1997;273:614–622.
38. Korth C, Stierli B, Streit P, Moser M, Schaller O, Fischer R, SchulzSchaeffer W, Kretzschmar H, Raeber A, Braun U, Ehrensperger F, Hornemann S, Glockshuber R, Riek R, Billeter M, Wüthrich K, Oesch B. Prion (PrPSc)-specific epitope defined by a monoclonal antibody. Nature 1997;390:74–77.
39. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. Proc Natl Acad Sci USA 2004;101:11287–11292.
40. Ehrlich LP, Nilges M, Wade RC. The impact of protein flexibility on protein–protein docking. Proteins 2005;58:126–133.