WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

PREDICTION REPORT

Docking and scoring protein complexes: CAPRI 3rd Edition

Marc F. Lensink,[1] Raúl Méndez,[2,3] and Shoshana J. Wodak[4,5]*

[1] Centre de Biologie Structurale et Bioinformatique, CP 263, BC6, Université Libre de Bruxelles, Blvd du Triomphe, 1050 Bruxelles, Belgium

[2] Service de Conformation de Macromolécules Biologiques, et Bioinformatique, CP 263, BC6, Université Libre de Bruxelles, Blvd du Triomphe, 1050 Bruxelles, Belgium

[3] Grup Biomatemàtic de Recerca, Institut de Neurociències, Unitat de Bioestadística, Facultat de Medicina, Universitat Autonòma de Barcelona, 08193 Bellaterra, Spain

[4] Structural Biology Program, Hospital for Sick Children, 555 University Av. Toronto, Ontario, Canada M5G 1X8

[5] Department of Biochemistry, University of Toronto, Toronto Ontario, Canada

ABSTRACT

The performance of methods for predicting protein–protein interactions at the atomic scale is assessed by evaluating blind predictions performed during 2005–2007 as part of Rounds 6–12 of the community-wide experiment on Critical Assessment of PRedicted Interactions (CAPRI). These Rounds also included a new scoring experiment, where a larger set of models contributed by the predictors was made available to groups developing scoring functions. These groups scored the uploaded set and submitted their own best models for assessment. The structures of nine protein complexes including one homodimer were used as targets. These targets represent biologically relevant interactions involved in gene expression, signal transduction, RNA, or protein processing and membrane maintenance. For all the targets except one, predictions started from the experimentally determined structures of the free (unbound) components or from models derived by homology, making it mandatory for docking methods to model the conformational changes that often accompany association. In total, 63 groups and eight automatic servers, a substantial increase from previous years, submitted docking predictions, of which 1994 were evaluated here. Fifteen groups submitted 305 models for five targets in the scoring experiment. Assessment of the predictions reveals that 31 different groups produced models of acceptable and medium accuracy-but only one high accuracy submission-for all the targets, except the homodimer. In the latter, none of the docking procedures reproduced the large conformational adjustment required for correct assembly, underscoring yet again that handling protein flexibility remains a major challenge. In the scoring experiment, a large fraction of the groups attained the set goal of singling out the correct association modes from incorrect solutions in the limited ensembles of contributed models. But in general they seemed unable to identify the best models, indicating that current scoring methods are probably not sensitive enough. With the increased focus on protein assemblies, in particular by structural genomics efforts, the growing community of CAPRI predictors is engaged more actively than ever in the development of better scoring functions and means of modeling conformational flexibility, which hold promise for much progress in the future.

Proteins 2007; 69:704–718.

Key words: *protein–protein interactions; docking; CAPRI; predictions; force fields; conformational changes; structural genomics.*

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/
The authors state no conflict of interest.
Grant sponsor: the Actions de Recherché Concertées; Grant number: convention 02/07-291; Grant sponsor: the GeneFun project; Grant number: EU contract 503567; Grant sponsors: the Canada Institute of Health Research; the Ontario Research Fund; Department of Biochemistry, Hospital for Sick Children, Structural Genomics Consortium; Connaught foundation.
*Correspondence to: Shoshana J. Wodak, Centre for Computational Biology, Room 1300, 180 Dundas St. W., Toronto, Ontario, Canada M5G 1X8.
E-mail: shoshana@sickkids.ca
Received 11 July 2007; Revised 31 July 2007; Accepted 2 August 2007
Published online 5 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21804

INTRODUCTION

Significant progress has been achieved towards the goal of deriving comprehensive catalogues of protein complexes formed in model organisms, as witnessed by several recent genome scale analyses in yeast.[1,2] It has also been shown recently, that structural information on interacting protein pairs is very helpful in deriving meaningful interpretations from genome-wide protein interaction networks.[3] Unfortunately, however, protein assemblies are still poorly represented in the protein data bank (PDB).[4,5] Structural genomics programs have so far mainly focused on large-scale structure determination of individual gene products, but not on larger assemblies.[6,7] Although new large-scale efforts to determine the structures of multi-component complexes are underway (e.g., SPINE2: http://www.spine2.eu), they require combining crystallography, NMR, and cryo electron microscopy and remain low throughput. Hence, while the repertoire of individual protein structures is increasingly well covered, the catalogue of three-dimensional (3D) protein assemblies will remain incomplete in the near future.

Computational procedures capable of reliably generating structural models of multi-protein assemblies starting from the atomic coordinates of the individual components, the so-called "docking" methods, should therefore play an important role in helping to bridge the gap. But as with all predictive approaches, objective tests are needed to monitor their performance. The Critical Assessment of PRedicted Interactions (CAPRI), a community wide experiment modeled after CASP[8] was set up in 2001, with precisely this goal. In CAPRI, individual groups that develop docking procedures, predict the 3D structure of a protein complex from the known structures of the components. The predicted structure is subsequently assessed by comparing it to the experimental structure—the target—determined most commonly by X-ray diffraction, which is deposited with CAPRI prior to publication. The predictions are thus made blindly, without any knowledge of the correct answer, and the evaluation is carried out by an independent team from whom the identity of the predictors is concealed.

In the 6 years of CAPRI's existence, 12 prediction Rounds were completed with a total of 28 targets. The results of Rounds 1–5 were presented at 10 evaluation meetings held in 2002 and 2004 and described in the literature.[9–11]

Here we describe the prediction results for Rounds 6–12, which were completed between January 2005 and March 2007. These results were presented at the 3rd CAPRI evaluation meeting held in Toronto earlier this year.[12] These latest Rounds had a total of nine targets representing X-ray structures of eight protein–protein complexes, two of which were homodimers. In stark contrast to previous CAPRI Rounds these nine targets included none of the classical enzyme-inhibitor and antigen-antibody complexes, but represented new systems involved in a variety of cellular processes. Round 8 (2 targets) had to be canceled because images of the experimental structures were available on the Internet, unbeknownst by their authors. Predictor groups were allowed to submit ten models for each target, and their submissions were assessed by comparison to the X-ray structure. Rounds 9–12 also included a new scoring experiment. In this experiment, predictors were invited to upload larger sets of 100 predicted models for each target. Those were shuffled to conceal their origins and made available to groups developing functions for scoring predicted association modes. These "scorer" groups reranked the uploaded set using their preferred scoring function and submitted their own 10-models from that set for assessment in the usual way.

In the docking experiment a total of 1994 models for seven targets were evaluated. These were submitted by 63 different groups and 8 automatic web-servers. Fewer groups (15) participated in the newer scoring experiment, submitting a total of 305 models for the five targets of Rounds 9–12. As previously, no limits were set on the source or type of additional information (homologous proteins, biochemical data on interacting residues) that predictors could use to guide their docking calculations.

The described evaluation is based on a number of simple criteria, largely adopted by the CAPRI community. Results are presented for individual targets, as well as across predictor groups and across targets. Remaining challenges and new developments in docking procedures are highlighted and insights gained on scoring methods are discussed.

THE TARGETS

The seven evaluated targets, as well as the additional two of Round 8 that were subsequently canceled, were based on the X-ray structures represent a variety of biologically relevant complexes involved in gene expression, signal transduction, RNA and protein processing, and membrane maintenance. Information about these targets is summarized in Table I of Janin, 2007 (this issue[13]). These targets are denoted as T20–T28, reflecting the sequence at which they were made available to the successive CAPRI Rounds.

Prior to 2005, many CAPRI targets were antibody–antigen and enzyme-inhibitor complexes, mostly of the "unbound/bound" type where the unbound component is a structure of the free protein, and the bound component is taken from the complex. Docking procedures perform much better with the bound than unbound component because they need not model the conformational changes that can occur upon association. But

Table I
CAPRI Prediction Results for Individual Targets

	Target T20	Target T21	Target T24	Target T25	Target T26	Target T27.1	Target T27.2	Target T28
Predictor groups	29	37	37	37	37	38	38	38
Evaluated predictions	266	337	335	336	351	343	348	357
High accuracy(***)	0	0	0	1	0	0	0	0
Medium Accuracy(**)	0	4	0	13	22	0	2	0
Acceptable(*)	3	7	4	20	20	0	55	0
Incorrect	246	307	298	289	291	327	270	322
Predictions with clashes	17	19	33	13	18	16	21	35
Average no. of clashes (SD)	35.30 (39.15)	29.32 (38.17)	22 (29)	19 (24)	25 (29)	21 (26)	20 (26)	41 (51)

Model no. (category)	Predictors	f_{nat}	$f_{non\text{-}nat}$	$f_{IR\text{-}L}$	$f_{IR\text{-}R}$	N_{clash}	L_rms	l_rms	$\theta_L(^\circ)$	d_L(Å)
Target T20										
08(*)	Baker	0.360	0.559	0.729	0.805	15	12.092	3.129	31.8	6.53
Target T21										
05(**)	Ten Eyck	0.634	0.350	0.619	0.900	9	4.407	1.577	16.49	2.347
08(**)	Bonvin	0.537	0.569	0.714	0.850	7	5.262	1.941	11.40	4.449
03(**)	Gray	0.463	0.486	0.667	0.800	6	7.617	1.921	18.66	6.263
09(*)	Hirokawa	0.342	0.803	0.667	0.850	104	9.097	3.049	20.28	8.220
03(*)	Vajda	0.244	0.787	0.667	0.850	11	9.597	4.525	46.57	6.276
01(*)	Weng	0.317	0.764	0.714	0.800	6	10.043	2.803	25.76	8.436
	Unbound	*0.707*	*0.147*	*0.619*	*0.950*	*23*	*1.013*	*0.819*	*0.34*	*0.478*
Target T24										
06(*)	Camacho	0.333	0.839	0.391	0.920	9	8.94	2.841	42.8	6.62
09(*)	Weng	0.200	0.833	0.261	0.800	4	8.94	3.126	55.2	5.45
03(*)	Totrov	0.233	0.794	0.217	0.720	7	12.60	3.693	53.4	10.55
	Unbound	*0.700*	*0.222*	*0.522*	*0.640*	*14*	*0.588*	*0.410*	*0.95*	*0.133*
Target T25										
10(***)	Eisenstein	0.827	0.246	0.840	0.913	4	2.203	0.904	2.70	1.482
01(**)	Schomburg	0.808	0.276	0.920	0.913	5	1.829	1.062	3.69	0.407
10(**)	PATCHDOCK	0.692	0.357	0.960	0.913	19	2.334	1.166	6.28	1.139
01(**)	GRAMM-X	0.827	0.295	0.960	0.913	10	2.824	1.246	8.94	1.294
03(**)	Vajda	0.635	0.400	0.960	0.913	2	2.832	1.297	9.10	1.293
03(**)	Fernandez-Recio	0.788	0.281	0.880	0.913	35	2.844	1.234	9.36	1.563
02(**)	Totrov	0.692	0.122	0.960	0.783	2	3.018	1.246	10.70	1.643
01(**)	Facemyer	0.788	0.211	0.920	0.957	21	3.153	1.415	14.03	1.151
07(**)	SKE-DOCK	0.788	0.281	0.920	0.913	9	3.307	1.330	11.07	2.029
07(**)	Takeda-Shitaka	0.808	0.323	1.000	0.913	8	3.324	1.335	11.08	2.059
07(**)	Weng	0.808	0.364	0.960	0.913	6	3.804	1.506	13.01	2.450
02(**)	Smith	0.673	0.426	0.960	0.913	23	3.929	1.510	9.27	3.085
01(**)	SMOOTHDOCK	0.385	0.592	0.760	0.739	15	4.849	2.192	23.98	1.966
05(**)	Negi	0.673	0.470	0.920	0.913	55	5.745	1.862	17.36	4.029
01(*)	Bonvin	0.365	0.708	0.960	0.913	10	6.644	3.522	21.05	4.809
04(*)	Camacho	0.308	0.765	0.840	0.826	7	7.533	4.088	34.68	4.202
06(*)	CLUSPRO	0.654	0.477	0.840	0.739	12	7.564	2.845	34.45	4.634
05(*)	Bates	0.442	0.681	0.880	0.957	17	9.825	3.178	24.42	8.412
	Unbound	*0.923*	*0.094*	*0.960*	*0.957*	*0*	*0.282*	*1.134*	*0.65*	*0.109*
Target T26										
01(**)	Baker	0.543	0.227	0.796	0.800	8	2.425	1.064	12.58	1.588
01(**)	Bonvin	0.532	0.359	0.886	0.833	16	2.621	1.932	15.43	1.386
06(**)	Gray	0.468	0.200	0.796	0.800	9	2.698	1.239	13.66	1.812
03(**)	Vajda	0.543	0.150	0.773	0.867	5	3.039	1.183	14.45	2.170
01(**)	Smith	0.532	0.206	0.796	0.800	3	3.605	1.326	16.94	2.619
07(**)	Vakser	0.308	0.603	0.750	0.733	25	3.833	1.942	21.26	2.469
01(**)	S_Liang	0.489	0.281	0.750	0.600	16	5.430	1.862	25.98	4.091
01(**)	Hsu	0.500	0.397	0.727	0.733	16	6.773	1.996	24.50	5.237
04(*)	Weng	0.308	0.628	0.750	0.767	12	5.638	2.252	28.93	3.987
02(*)	Zachrias	0.447	0.288	0.750	0.667	13	6.179	2.073	27.09	4.725
01(*)	DelCarpio	0.468	0.532	0.773	0.667	25	6.341	2.330	32.56	4.101
04(*)	Eisenstein	0.447	0.276	0.773	0.767	8	6.410	2.422	13.80	5.723
02(*)	Totrov	0.308	0.442	0.818	0.733	13	7.111	3.054	28.40	4.946
04(*)	PATCHDOCK	0.457	0.427	0.659	0.700	20	7.932	2.318	29.66	6.437

Continued

DOI 10.1002/prot

Table I
Continued

Model no. (category)	Predictors	f_{nat}	$f_{non\text{-}nat}$	$f_{IR\text{-}L}$	$f_{IR\text{-}R}$	N_{clash}	L_rms	I_rms	θ_L(°)	d_L(Å)
02(*)	Schomburg	0.415	0.371	0.704	0.633	15	8.087	2.315	31.92	6.253
09(*)	Ritchie	0.372	0.527	0.727	0.733	55	10.031	3.353	36.22	6.617
	Unbound	*0.638*	*0.178*	*0.864*	*0.800*	*63*	*0.680*	*0.596*	*1.98*	*0.062*
Target T27.2										
07(**)	Zhou	0.732	0.538	0.941	0.947	20	6.176	1.535	23.08	4.641
06(**)	Weng	0.415	0.553	0.882	1.000	4	7.990	1.855	19.20	7.408
04(*)	Smith	0.561	0.589	1.000	1.000	3	5.444	2.326	8.94	5.036
09(*)	S_Liang	0.463	0.694	1.000	0.947	31	5.644	2.927	15.35	4.327
04(*)	Lorenzen	0.268	0.633	0.765	0.842	18	5.876	2.616	20.36	3.760
01(*)	Camacho	0.317	0.740	1.000	0.947	5	5.967	2.931	19.49	4.123
09(*)	Baker	0.415	0.564	1.000	0.842	5	5.978	2.693	16.55	4.550
06(*)	Zhou	0.390	0.729	0.824	0.947	21	6.460	3.040	25.58	3.134
05(*)	Vakser	0.293	0.700	0.941	0.790	4	6.489	3.082	18.04	4.925
07(*)	Bonvin	0.537	0.560	0.941	1.000	11	6.505	2.601	18.68	5.085
02(*)	Wolfson	0.537	0.639	0.941	0.895	24	7.432	2.645	18.95	6.013
06(*)	Zacharias	0.146	0.842	1.000	0.842	8	8.148	3.800	8.60	7.900
01(*)	TenEyck	0.439	0.625	0.882	0.842	5	8.559	2.592	21.07	7.308
07(*)	CLUSPRO	0.293	0.829	0.882	1.000	10	10.320	3.205	14.30	9.968
01(*)	Eisenstein	0.342	0.731	0.824	0.737	10	10.873	3.345	30.42	8.461
	Unbound	*0.634*	*0.103*	*0.882*	*0.895*	*10*	*0.781*	*0.446*	*2.10*	*0.556*

(a) Target T20: *E-coli* peptidyl tRNA release factor 1 (RF1)—Methyl transferase complex.
(b) Target T21: Yeast Orc1 BAH-Sir1p OIP domains complex.
(c) Target T24: Arf1-Arf1 Binding Domain of ARHGap21.
(d) Target T25: Arf1-Arf1 Binding Domain of ARHGap21.
(e) Target T26: E-coli TolB-Pal complex.
(f) Target T27.1: UBC9-HIP2 complex (A.D interface).
(g) Target T27.2: UBC9-HIP2 complex (A/C' interface).
(h) Target T28: NEDD4L catalytic domain dimer.
The table is divided into two parts. The top part provides a general summary of the predictions and the bottom part lists the key parameters of the best predictions ranked as acceptable or higher submitted by each group for individual targets T24–T27. Results for T28 are not listed, as no correct predictions were submitted for this target.
The submitted predictions were divided into four categories as detailed in the text. Predictions with a number of clashes exceeding a defined threshold were not evaluated. Clashes were defined as those between two nonhydrogen atoms on each side of the interface whose distance was less than 3 Å. The threshold was taken as two standard deviations plus the average of the number of clashes in all the predictions submitted for a given target.
Detailed results for the best predictions for each participant, which were of acceptable quality or better (bottom), ranked as indicated in Table I. Column 1 lists the model number (1–5) and the rank of the prediction (high accuracy (***), medium accuracy (**), and acceptable (*). Column 2 lists the participant groups in order of decreasing native contact fraction f_{nat} (column 3). Column 3 lists the fraction of non-native contacts $f_{non\text{-}nat}$, defined as the number of non-native contacts over the total number of contacts in the predicted complex. $f_{IR\_L/R}$ is defined as the number of native residues in the predicted interface over the total number of interface residues in the target, computed for both the R (receptor) or L (ligand) molecules. Column 7 ($N_{clashes}$) lists the number of atomic clashes in the predicted complex. Columns 8 and 9 list the root mean square deviation (rmsd) values. L_rms is the backbone rmsd (Å) of the ligand molecules in the predicted versus the target complexes after the receptor moieties have been superimposed. The I_rms is the interface rmsd (Å) computed by superimposing only the backbone of the interface residues from the target complex onto their counterparts in the predicted complex. The last two columns list the residual rigid-body rotation (θ_L) and translation (d_L) of the ligand in the predicted versus the target complexes after the corresponding receptor molecules have been superimposed. For further details on how the various parameters were computed, see Refs. 10 and 11.

docking two bound components is not considered in CAPRI as it has little predictive value.

With one exception all the targets considered in this evaluation were of the "unbound/unbound" type: predictors were provided with components whose structures were determined independently or built by homology from the PDB entries corresponding to a related protein. The exception was target T25 of Round 9. Like target T24 is was derived from the complex between Arf1 and the Arf binding domain (ArfBD) of ARHGap21.[14] However, while for T24, predictors were given the unbound structure of Arf1, and asked to model that of ArfBD, using as template a related PH (plekstrin-homology) domain, In T25, predictors were given the same Arf1 model, but ArfBD was taken from the complex and was hence the bound structure.

The interface areas of the target complexes span a somewhat wider range than usually encountered in protein complexes[15,16] as seen in Table I of Janin (2007) (Ref. 13, this issue). Large interface areas are buried in the homodimer (target T28) and in the two hetero complexes, the one between the TolB and Pal proteins from *E-coli*[17] (target T26) and in the HemK-RF1complex also from *E.coli* (target T20).[18] The extensive interface area (~4100 Å^2) in the latter complex is due to a large conformational change involving a loop of RF1 that is part of the intermolecular interface. A smaller than average interface area is displayed by the UBC9-HIP2 complex

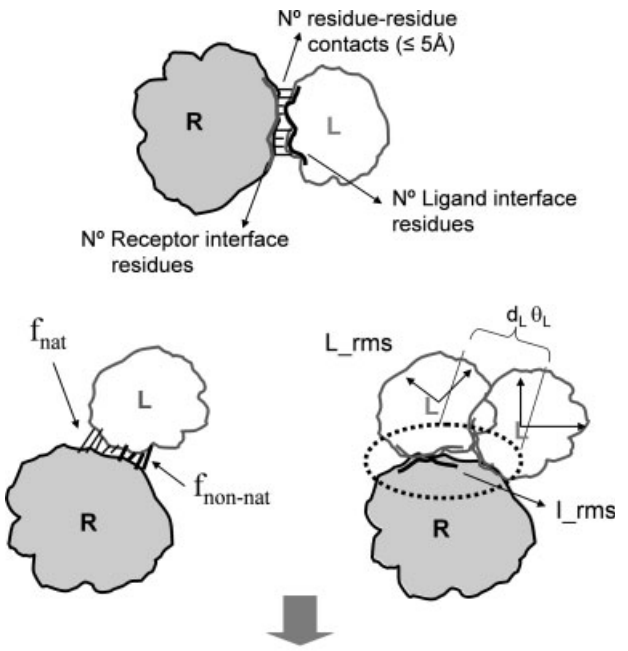
(Walker et al., unpublished data, 2007) of target T27. This is an interesting target that presents some ambiguity in identifying the biologically meaningful interface from the packing interactions in the crystallographic asymmetric unit, as will be discussed later.

Several factors determine the level of difficulty that a given target represents for the prediction programs as detailed previously.[19] For the seven targets evaluated here the main challenge has been to model the conformational changes that the unbound or homology-modeled structures undergo upon binding. Large changes had to be modeled in T20, the RF1-methyl transferase complex (backbone rmsd ~13.9 Å between the bound RF1 and its homolog RF2), and in the NEDD4L homodimer of T28 (backbone rmsd ~10 Å) (Walker et al., unpublished data, 2007). More limited changes but directly affecting the interface also occurred in targets T24 (Arf1-ArfBD).

For many targets, the difficulties described earlier were often offset by the use of biochemical data or information on sequence conservation in related proteins, enabling to identify residues that might be directly involved in the interaction. As already emphasized in previous CAPRI evaluation reports[10,11] such information continues to play a very important role in guiding docking calculations as well as filtering docking solutions, and rare are the predictors that do not exploit it.

THE EVALUATION PROTOCOL

The parameters and criteria used to evaluate the quality of the predicted complexes, summarized in Figure 1, are exactly the same as in previous CAPRI evaluations. The reader is referred to previous CAPRI reports[10,11] for a detailed description of these parameters and the corresponding thresholds used in classifying predictions as "high," "medium," and "acceptable" accuracy. Submitted models with a number inter atomic clashes exceeding a threshold defined for each target were not evaluated, as such models may retrieve a large number of native interactions simply because of the interpenetration of the corresponding structures.



Ranking	Conditions based on Capri computed parameters
High	$f_{nat} \geq 0.5$ AND (L_rms ≤ 1.0 OR I_rms ≤ 1.0)
Medium	($f_{nat} \geq 0.3$ AND $f_{nat} < 0.5$) AND (L_rms ≤ 5.0 OR I_rms ≤ 2.0) OR $f_{nat} \geq 0.5$ AND L_rms > 1.0 AND I_rms > 1.0
Acceptable	($f_{nat} \geq 0.1$ AND $f_{nat} < 0.3$) AND (L_rms ≤ 10.0 OR I_rms ≤ 4.0) OR $f_{nat} \geq 0.3$ AND L_rms > 5.0 AND I_rms > 2.0
Incorrect	$f_{nat} < 0.1$ OR (L_rms > 10.0 AND I_rms ≤ 4.0)

Figure 1

Schematic illustration of the quality measures used to evaluate the predicted models. The following quantities were computed for each target: (1) all the residue–residue contacts between the Receptor (R) and the Ligand (L), and (2) the residues contributing to the interface of each of the components of the complex. Interface residues were defined on the basis of their contribution to the interface area, as described in Refs. 10,11. For each predicted model the following quantities were computed: the fractions f_{nat} of native and $f_{non\text{-}nat}$ of non native contacts in the predicted interface; the root mean square displacement (rmsd) of the backbone atoms of the ligand (L_rms), the misorientation angle θ_L and the residual displacement d_L of the ligand center of mass after the receptor in the model and experimental structures were optimally superimposed.[20] In addition we computed I_rms, the rmsd of the backbone atoms of all interface residues after they have been optimally superimposed. Here the interface residues were defined less stringently on the basis of residue–residue contacts (see Refs. 10,11). As previously described,[10,11] models exhibiting a number of close atomic contacts (clashes) exceeding by at least two standard deviations the average number of such clashes in all the models submitted for a given target, were not evaluated. It should be noted that in the protocol for classifying predicted model into the four categories ("Incorrect," "Acceptable," "Medium," and "High"), the listed inequalities (bottom of Figure) were applied in the reverse order, starting with those defining incorrect predictions.

PREDICTION RESULTS AND DISCUSSION

This section is divided into two main parts. The first part presents the results of the docking experiment. The second part reports on the scoring experiment. In the docking section we present results for the seven targets from CAPRI rounds 6–12. We start by the prediction results obtained for individual targets. This is followed by an overview of the results across predictor groups and targets. The scoring section describes the prediction results for the 5 targets of Rounds 9–12 and contrasts them with those obtained in the docking experiment for the same targets.

Values of all the quality measures computed for all the submitted predictions for each target can be found on the CAPRI web site (http://capri.ebi.ac.uk/).

DOI 10.1002/prot

The docking experiment

Prediction results for individual targets

Target T20: E. coli peptidyl tRNA release factor 1(RF1)–methyl transferase complex. This target is a complex between the *E. coli* HemK methyl-transferase and peptidyl-tRNA release factor 1 (RF1).[18] The prediction results are summarized in Table I (a) and a pictorial summary is given in Figure 2. The top portion of the Table gives a general summary and the bottom portion lists the best of the acceptable or higher accuracy predictions obtained for this target by each group together with their quality measures. Twenty nine groups submitted a total of 283 models for this target. Seventeen of these were not evaluated as they had a larger than average number of clashes (see legend of Table I for details). Of the remaining 266 models a mere three, submitted by the group of Baker, were ranked "acceptable."

The limited success rate in predicting this target underscores yet again the major challenge that conformational changes represent for docking methods. The RF1 protein is a substrate of the HemK enzyme, which methylates the glutamine of a conserved GGQ motif. The predictors were given the X-ray structures of the free HemK and of two homologs of RF1. Knowing that the GGQ motif must be bound at the enzyme active site should have helped docking procedures to find the correct solution. But the motif is part of a mobile surface loop, which moves away from the body of the RF1 molecule and reaches deep into the active site pocket when the complex is formed. None of the submitted models reproduced this movement accurately. The Baker group, who used a Monte–Carlo refinement procedure, was however able to move the loop in the right direction thereby improving the geometry of the predicted complex. The best of their three acceptable models identified 36% of the native contacts, has an interface rmsd of 3.1 Å and a ligand rmsd of 12 Å.

Target T21: yeast Orc1 BAH-Sir1p OIR domains complex. T21 is part of the origin recognition complex, which prevents transcription of DNA at replication origins in yeast.[21] It is one of the unbound/unbound targets featuring limited conformation changes (largest backbone rmsd = 1.27 Å for the Orc1 domain). The fact that only seven "acceptable" and 4 "medium accuracy" models were obtained from a total of 337 evaluated submissions for this target [see Fig. 2 and Table I] is somewhat disappointing. It is probably due to the fact that little biochemical information was available to limit the search space, and suggests that unconstrained docking problems are still a challenge for most of the 37 predictor groups submitting models for this target. It was nevertheless reassuring to see [Table I] that six groups managed to produce models of "acceptable" (Hirokawa, Vajda, and Weng) and "medium" (Ten Eyck, Bonvin, Gray) accuracy. The best

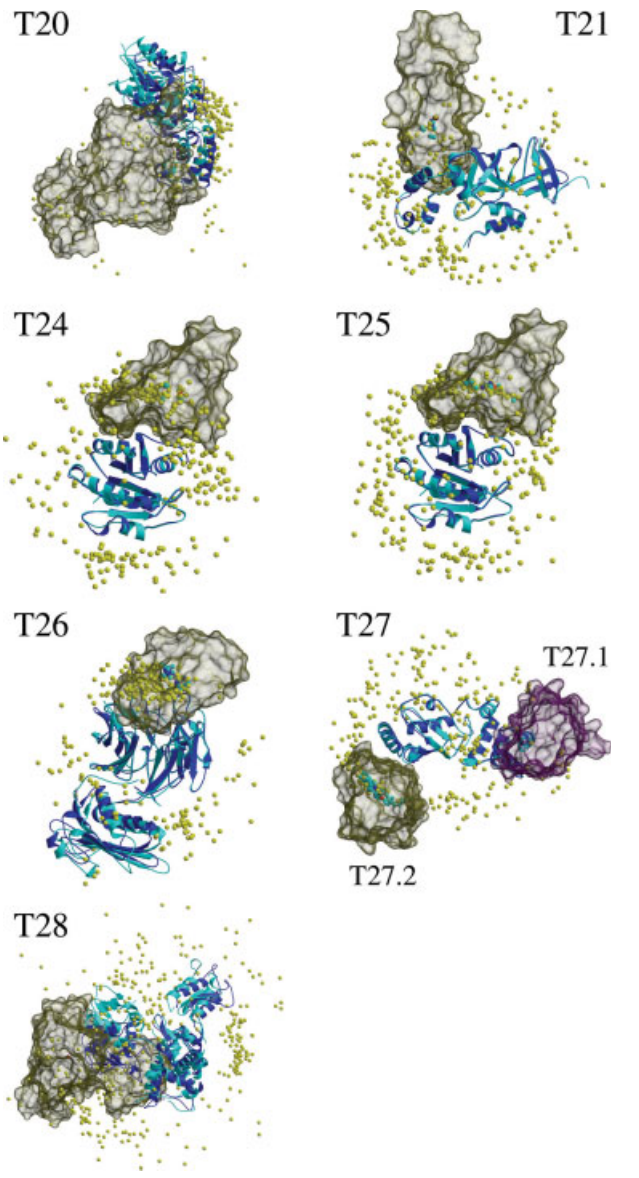


Figure 2

Pictorial views of CAPRI predictions for targets T20-T28. For each target complex (labeled by the target number), the ligand is represented by its molecular envelope and the receptor is shown using a ribbon representation. The shown targets are T20 (PDB code 2b3t),[17] T21 (PDB code 1zhi), (Walker et al., unpublished data, 2007) T24, and T25 (PDB code 2j59),[13] T26 (PDB code 2hqs),[16] T27 (PDB code 2o25),[18] and T28(PDB code 2oni).[19] Two version of the receptor structure are shown: the bound structure (light blue) and the unbound structure used by the predictors (dark blue). To optimally position the unbound structure within the complex structure superimpositions were performed using the same portions of the structure as those used in the CAPRI validation procedure (see text). The pictures display the positions of the geometric centre of the ligands relative to the each receptor molecule in submitted predictions for each target. To produce these pictures, the receptor molecules in each predicted complex were superimposed onto their counterpart in the target and the corresponding rigid-body transformation was applied to the entire complex (see Refs. 10,11). High accuracy predictions are represented by green sphere. Medium accuracy predictions are represented by blue spheres and acceptable predictions are represented by light blue sphere. Gold colored spheres correspond to incorrect solutions (for ranking criteria see Fig. 1). The red spheres represent the relative positions of the geometric centres of the ligands in the experimentally determined target structure.

model for this target, produced by the Ten Eyck group, had 63% of the native contacts correctly predicted, an I-rmsd of 1.6 Å, and a significant fraction of the interface residues (epitopes) predicted correctly (60% for the ligand and 90% for the receptor). It is noteworthy that this model differs little from the model generated by just superimposing the two unbound structures onto their counterpart in the complex [Unbound row, Table I].

Targets T24-T25: Arf1-Arf1 binding domain of ARHGap21. These two targets were derived from the same complex between Arf1 and the Arf binding domain (ArfBD) of ARHGap21.[14] Arf1 is a small G-protein and ArfBD is a PH (plekstrin-homology) domain. In T24 the ArfBD moiety had to be built by homology using as template the β–spectrin PH domain before docking it onto the unbound structure of Arf1. In T25, Arf1 was the same unbound structure, but ArfBD was taken from the complex and randomly oriented (bound version).

As expected, the prediction results were poorer for T24 than T25. This is mainly due to the fact that the C-terminal segment of the PH domain, a helix that participates in the interface with Arf1, was missing in the template and had to be modeled. But some predictors either chose not to do so or did not do it well enough.

Of the 335 evaluated models submitted by 37 groups for T24, only four models were of "acceptable" quality [Table I]. These were obtained by the groups of Camacho, Weng, and Totrov, with the best model featuring only 33% of the native contacts and an L-rms and I-rms of about 9 and 2.8 Å respectively. All the acceptable models had the two proteins misoriented by 40°–55°, with only a fraction of the interface residues correctly identified (30–40% for the Arf1 protein, and 70–90% for the PH domain). All the acceptable models were of significantly lower quality than that obtained by superimposing the homology built and unbound subunits onto those in the complex [unbound row in Table I].

Performing the prediction with the bound PH coordinates for T25 significantly improved the quality of the predicted models. Of the 336 evaluated models [Table I], 20 were ranked "acceptable," 13 "medium," and 1 "high." This "high accuracy" model, submitted by the group of Eisenstein, had 83% of the native contacts correctly predicted, between 84 and 91% of the interface residues correctly identified, an L-rms of 2.2 Å and an I-rms of 0.9 Å. This model came very close to the one built by superimposition of the independent subunits used in the docking calculations [Table I]. The 13 "medium accuracy" models were all submitted by different groups. Several of these, like the model submitted by the group of Schomburg, were very close to being high quality models, raising some debate on the criteria used by the assessors for discriminating between "medium" and "high quality" model categories (Fig. 1). Interestingly, the groups submitting medium quality models included four automatic Web servers (PATCHDOCK, GRAMM-X, SKE-DOCK, and SMOOTHDOCK), indicating that such servers are starting to successfully compete with expert-lead predictions.

Target T26: E. coli TolB-Pal complex. This target is derived from a complex involved in the maintenance of the *E. coli* outer membrane.[17] TolB has two domains whose relative orientation changes upon complex formation, but this change does not affect the interface with Pal. This interface buries a large surface area (2400 Å^2) and contains several partially buried glutamic acids as well as a number of buried and partially buried water molecules. It is therefore reassuring to see that docking procedures performed adequately on this target as well. A total of 20 "acceptable," and 22 "medium accuracy" predictions, but no "high accuracy" models were submitted, representing 12% of the 351 assessed models [Table I (e)]. These were contributed by 16 groups (including 1 server) out of the 37 that participated in the predictions. The medium accuracy models were however not as good as those generated for T25. The best T26 prediction, contributed by the Baker group, identified only 54% of the native contacts correctly, and had an L-rms of ∼2.4 Å, an I-rms of ∼1.1 Å, and featured a misorientation between the molecules of about 12°. The lower quality of these models relative to those obtained for T25, can be attributed in part to the residual conformational adjustments undergone by the several loops of the TolB domain that interacts with Pal (Fig. 2) and probably also to the role played by specific interactions with water molecules at the interface. The effect of failing to reproduce the conformational changes of the TolB domain can be estimated from the model generated by superimposing the individual unbound subunits onto their counterparts in the complex. This model retrieves only 60% of the native contacts, and only 80–86% of the interface residues [unbound row in Table I].

Target 27: UBC9–HIP2 complex. T27 is a particularly interesting case. It represents a complex between the SUMO-conjugating enzyme UBC9 and the ubiquitin-conjugating enzyme HIP2, a substrate of UBC9 (Walker et al., unpublished data, 2007). The asymmetric unit of the crystal contains two UBC9 and two HIP2 molecules that may be paired in at least two different ways, representing either the biological unit or crystal packing interactions. The corresponding binding modes are denoted T27.1 and T27.2 in Table I (top), and are illustrated in Figure 2. None of the 38 CAPRI predictor groups were able to predict the T27.1 mode, which was initially considered by the authors of the X-ray structure. Instead, many of the groups generated models, which approximate rather well the T27.2 interaction that buries more surface area (∼1000 Å^2) than T27.1 (860 Å^2). Interestingly T27.2 involves the lysine of HIP2 that reacts with

 DOI 10.1002/prot

SUMO, whereas T27.1 uses the opposite end of the elongated HIP2 molecule. Two medium-quality models and 55 acceptable models were submitted for T27.2. The medium accuracy predictions were by the groups of Zhang and Weng. One of these (by Zhang) captures a larger fraction of the native contacts (73%) than the model built by superimposition of the individual unbound subunits onto the complex. The latter captures only 63% of these interactions [unbound row in Table I]. The 55 acceptable models for T27.2 were submitted by 13 groups including one server (CLUSPRO). Laboratory experiments are currently being performed in order to establish which of the two binding modes is relevant to the biological role of the HIP2/UBC9 association. It is in fact quite likely that both modes are. Indeed, being components of a complex protein degradation system the UBC9 and HIP2 proteins probably interact not only with one another but also with other proteins. Several regions of either HIP2 or UBC9 may therefore be involved in these interactions.

Target 28: NEDD4L catalytic domain dimer. T28 is the HECT domain of the NEDD4L ubiquitin-ligase, observed to be a dimer in the crystal (Walker et al., unpublished data, 2007). Its homologs in the PDB are monomeric. CAPRI groups had to build the NEDD4L monomers by homology before assembling them into the dimer. All the 38 groups submitting predictions for this target produced incorrect models [Table I]. They were unable to model the large rotation, which one domain of the unbound NEDD4L monomer must undergo to form the observed dimer (Fig. 2). This confirms observations made in previous CAPRI rounds, namely that quaternary structure prediction is a difficult exercise in spite of the symmetry constraints, owing to the large changes that often take place within the subunits. Whenever predictor groups correctly predicted the direction of conformational changes, they seemed reluctant to move the protein domain as far as would be required.

Prediction results across predictor groups and targets

Table II summarizes the prediction results for all seven targets of CAPRI Rounds 6–12, obtained by all groups that submitted at least one prediction ranked as acceptable or better. The listed results represent only the best prediction obtained by each group for each target. Thus, if a group submitted two acceptable predictions and one high-accuracy prediction for a given target, only the high accuracy result is listed. For a full account of the results obtained by each group the reader is referred to the CAPRI web site (http://capri.ebi.ac.uk).

In total, 63 groups submitted predictions for at least one target in Rounds 6–12, up by 100% from the number of groups submitting in rounds 3–5. Of those, 31 have an entry in Table II. Inspection of this Table clearly confirms that the success rate very much depended on the particular target. Good models (mostly ranked as "medium accuracy") were submitted for the easier targets, T21, T25, T26, and T27. Only acceptable models were submitted for the more difficult targets T20, T24, and predictions failed altogether for NEDD4L homodimer of target T28.

Not too surprisingly, a record number of acceptable, or better, predictions, from as many as 18 groups were obtained for target T25, the Arf1-ArfBD complex. This was the only target from amongst the ones evaluated here to be of the bound/unbound category, where conformational changes are mainly confined to side chain reorientation, and can be readily modeled by the docking and refinement procedures. It is therefore encouraging that a more difficult target, such as the *E.coli* TolB-Pal complex of T26, which is of the unbound/unbound category, was also quite successfully predicted by 15 groups, despite the backbone adjustments that needed to be modeled and the presence of water molecules in the interface.

Table II also enables to assess the success rate of individual groups, although this is a difficult and possibly controversial undertaking, because the number of evaluated targets remains much too small for drawing conclusions on a statistically significant basis. The best performance in the CAPRI rounds evaluated here was by the Weng's group who successfully predicted five out of the seven targets evaluated here. For two of those, their best models were of medium accuracy, whereas for the remaining three they were ranked acceptable. The group of Bonvin followed closely with correct predictions for the same targets except for T24, which they missed. Interestingly this group, like several others, provides better quality predictions than the Weng group for targets T21 and T26. But for target T25 their best model was of acceptable quality, whereas as many as 13 other groups (including Weng's) submitted medium accuracy models, and one group (Eisenstein) predicted this target with high accuracy.

A further six groups made acceptable or better predictions for three targets, with five groups submitting as least one medium quality prediction. Eight groups made valid predictions for only two targets, of which six provided at least one medium quality model. The remaining 15 groups made valid predictions for only one target (mainly target T25 but also T26 and T27), with nine of these groups providing medium quality models. These latter lower prediction performances, as well as the failure to submit an acceptable model for even one target by the remaining 32 groups that have no entry in Table II, do not necessarily reflect shortcoming of the docking methods, as a many of these groups submitted predictions for only a few targets.

Lastly, it is noteworthy that five out of the eight automatic docking servers, have entries in Table II, mainly owing to their successful prediction of the bound–

Table II
Summary of Docking Predictions for CAPRI Rounds 6–12

Predictor group	T20	T21	T24	T25	T26	T27	T28	Predictor summary
Weng	0	*	*	**	*	**	0	5/2(**)
Bonvin	0	**	0	*	**	*	0	4/2(**)
Eisenstein	0	0	0	***	*	*	0	3/1(***)
Smith	0	0	0	**	**	*	0	3/2(**)
Vajda	0	*	0	**	**	0	0	3/2(**)
Baker	*	0	0	0	**	*	0	3/1(**)
Totrov	–	–	*	**	*	–	–	3/1(**)
Gray	0	**	0	0	0	**	0	2/2(**)
S_Liang	–	0	–	–	**	*	0	2/1(**)
PATCHDOCK	–	–	0	**	*	–	–	2/1(**)
Schomburg	–	0	0	**	*	–	–	2/1(**)
Ten Eyck	–	**	0	0	0	*	0	2/1(**)
Vakser	0	0	–	–	**	*	0	2/1(**)
GRAMM-X	–	0	0	**	–	0	0	1/1(**)
Hsu	–	–	–	–	**	0	0	1/1(**)
Negi	0	0	0	**	0	0	0	1/1(**)
SKE-DOCK	–	0	0	**	0	0	0	1/1(**)
SMOOTHDOCK	0	0	0	**	0	0	0	1/1(**)
Takeda-Shikata	–	0	0	**	0	0	0	1/1(**)
Zhou	–	–	–	–	–	**	0	1/1(**)
Facemyer	–	–	0	**	0	–	–	1/1(**)
Fernandez-Recio	–	–	0	**	0	0	0	1/1(**)
Camacho	0	0	*	*	0	*	0	3
CLUSPRO	0	0	0	*	0	*	0	2
Zacharias	0	0	0	0	*	*	0	2
Bates	0	0	0	*	0	0	0	1
Del Carpio	0	0	0	0	0	*	1	1
Hirokawa	–	*	–	–	–	–	–	1
Lorenzen	–	–	–	–	–	*	0	1
Ritchie	0	0	0	0	*	0	0	1
Wolfson	0	0	0	0	0	*	0	1
Target summary	1	6/3**	3	18/13**/1***	16/8**	14/2**	0	

This Table summarizes the results obtained by all the groups that submitted one or more predictions of acceptable quality or better for at least one target.
Column 1 lists the group's affiliation and the last name of the Principle Investigator. The next nine columns list the results obtained for each of the nine targets. The right-most column summarizes the results per predictor group, and the bottom row summarizes the results per target.
"0" indicates that none of the submitted predictions was of acceptable quality. "–" indicates that no predictions were submitted. "*" indicates that at least one of the submitted predictions was in the acceptable range, "**" indicates that at least one of the submitted predictions was of medium accuracy, and "***" indicates that at least one prediction was of high accuracy. See Refs. 10 and 11 for the definition of the parameters range used to rank the predictions.
The summary entries list the total number of acceptable predictions, followed by the number of predictions of medium and high accuracy denoted by a "**", and "***," respectively.

unbound complex of target T25. The best performance is by the PATCHDOCK and CLUSPRO servers. The former submitted predictions for only three targets, with one medium quality prediction for T25 and one acceptable prediction for the more difficult T26. CLUSPRO on the other hand submitted predictions for all seven targets, scoring two acceptable quality predictions for targets T25 and T27 respectively.

Taken together these results clearly show that most docking methods, including automatic servers, are capable of producing good quality models in cases, such as the bound/unbound target T25, where the individual subunits undergo little or no backbone readjustments upon complex formation. In the more common and realistic cases of unbound/unbound docking, where some level of conformational adjustment takes place, the performance of docking procedures is more diverse and unreliable. With large conformational changes usually occurring upon formation of homo-oligomers it is therefore not too surprising that docking procedures are still unable to deliver in cases such as the NEDD4L homodimer of Target T28. On the other hand, it is encouraging to see that each of the remaining targets have at least one acceptable model, and often higher quality models produced by at least one group, with different groups contributing models to different targets. Although this has been observed in previous CAPRI assessments[10,11] the fact that is occurs here with the majority of the targets being for the first time of the unbound/unbound category, is particularly noteworthy.

The scoring experiment

The scoring experiment was designed to be a blind test of scoring functions independently of the search for candidate docking solutions. It was conducted as follows: a

DOI 10.1002/prot

day after the announced submission deadline for a given target, predictors were invited to upload larger sets of 100 predicted models for this target. Those were shuffled in order to conceal their origins and made available to all registered CAPRI participants. The "scorers," which included some of the groups that submitted docking predictions as well as other groups that did not, then reranked the uploaded set using their preferred scoring function and submitted their own ten-models from that set for assessment in the usual way. The experiment was run for targets T22–28, but assessment results are reported only for targets 25–27, given that targets T22 and T23 were subsequently cancelled, and only incorrect models were uploaded for targets T24 and T28.

The quality of all the uploaded models and scorers' submissions were evaluated using exactly the same criteria as for the docking submissions (Fig. 1).

Overview

An overview of the scoring experiment is presented in the Supplementary Table S1. For target T25 seven groups each uploaded 100 models. Of these 700 models, only a small fraction (5%) was of acceptable quality or better (22 acceptable, 14 medium). The six scorer groups each submitting their 10 best models selected from among the uploaded set, produced 12 acceptable and six medium quality models altogether. A total of 1567 models were uploaded (by 16 groups) for target T26. Of those, 127 were of acceptable quality of better, including 65 medium quality predictions, but none of high quality. The eight scorer groups for this target submitted a total of 75 models. Those included seven acceptable and seven medium quality models. The 1489 models uploaded (by 15 groups) for target T27, included 124 acceptable and 24 medium quality models. The 150 models submitted by the 15 scorer groups included 46 acceptable and two medium quality models.

Because of an unfortunate combination of incorrect file format and software shortcoming, 396 models uploaded by four groups for both Targets T26 and T27, were not made available to the scorers. For T26, these missing models included 19 acceptable and 48 medium quality predictions, which could therefore not be picked up by scorers, thereby affecting the reported results. The 396 missing models from the uploaded set of Target T27 had little effect however, as these included only 1 acceptable model in total.

These early results of the scoring experiment confirm that for easy targets such as the bound/unbound T25, scoring functions are able to single out about half of the 5% correct models present in the dataset. The performance seems to drop for the more difficult targets T26 and T27. Although the fraction of correct models (acceptable of better) in the uploaded ensembles were higher for these targets (7% for T26 and ~10% for T27), only about 10% of these models were picked up by the scoring functions. While the observed drop for T26 may be due to the formatting error aforementioned, that for T27 is not.

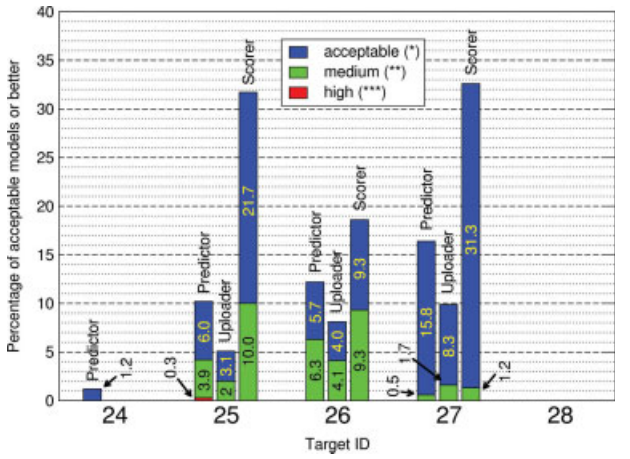


Figure 3

Relative success rates of CAPRI predictions for targets T25–T27. Fractions (%) of predicted models for targets T24–T27, rated as acceptable accuracy or better, in CAPRI docking (Predictor) and scoring experiments. In the latter experiment the "Scorer" bars represent the results of models submitted by the scorer groups, whereas the "Uploader" bars represent the fractions of correct predictions (acceptable accuracy or better) in the sets of 100 structures uploaded in the scoring experiment.

This notwithstanding, scorers tend to perform much better than uploaders or CAPRI predictors, when judged by the fraction of all submitted/uploaded models that are of acceptable quality or better (Fig. 3). The percentage of such models is significantly higher in scorers' submissions than in the classical CAPRI predictions, or in the uploaded ensembles. For instance, the consolidated percentage of acceptable and medium quality predictions by scorers for target T25 is 31.7%, compared to 9.9% in the CAPRI predictions and 5.1% in the uploaded set. A similar trend is observed for targets T26 and T27. Uploaded sets contain on average fewer correct models than the submissions to either the scoring or docking experiments, because they represent models of lower rank (100 best models) in the list of each uploading group, as compared to the 10 highest ranking models submitted for evaluation.

The challenge of singling out the better models

Since the scoring exercise was performed blindly, scorer groups needed not only to score models uploaded by other groups but to rescore their own models as well. Interestingly, we see that scorers tended to select models uploaded by other groups, rather their own. We see for example [Table III (a)] that Weng misses the medium ac-

Table III
Detailed Results of the Scoring Experiment for Target T25-T27

(a) **Target T25**

Uploaders	
Weng	1/1**
Bonvin	27/12**
Richie	3/1**
Bates	5
U06[b]	0
U07[b]	0
U14[b]	0

Scorers	
Wang	3/2**
Liang	4/1**
Fernandez-Recio	6/2**
Wolfson	2/1**
Bonvin	3
Weng	1

(b) **Target T26**

Uploaders	
Gray	49/41**
Bonvin	13/6**
Baker	5/3**
Wolfson	6/2**
S-Liang	32/9**
Nakamura	7/3**
Zhang	1/1**
PATCHDOCK	3
Del Carpio	1
Vajda	6
Richie	4
U02,05,08,20,24 [b]	0

Scorers	
Weng	4/3**
S-Liang	4/1**
Fernandez-Recio	4/2**
Wolfson	2/1**
J-Liang	0
Poupon	0
Takeda-Shitaka	0
Wang	0

(c) **Target T27**

Uploaders	
Bonvin	85/17**
Weng	9/7**
Smith	30
PATCHDOCK	21
SMOOTHDOCK	1
Bates	1
Baker	1
U03,08,10,12[b]	0
U17,24,28,36[b]	0

Scorers	
Camacho	9/1**
Mehio	2/1**
Bonvin	10
Zhou	8
Weng	7
Wolfson	6
Fernandez-Recio	2
Bates	2
J-Liang	1
Wang	1
S-Liang	1
Others[a]	0

The left and right hand Tables list the best models of acceptable accuracy or higher provided respectively, by each of the uploader and scorer groups (represented by the names of their PI's). Uploader groups each submitted 100 models, whereas each scorer group was allowed to submit only their top 10 models. The notation for the model quality (2nd column of each Table) is the same as in Tables II. In the first position is the total number of acceptable predictions or better; in the second and third positions are listed the numbers of medium accuracy (**) and high accuracy (***) predictions, when such predictions are submitted.
The arrows linking scorer to uploader rows in the two Tables indicate which of the uploader's set contributed to the predictions of a given scorer group. Dashed arrows indicate the 'traffic' of acceptable models, and full line arrows the 'traffic' of medium accuracy models.
The uploaded models and those submitted by scorer groups were evaluated using exactly the same criteria as the CAPRI docking predictions.
[a]Other scorer groups whose predictions were incorrect: Launay, Poupon, Sternberg, Takeda-Shitaka.
[b]Other uploader groups whose entire set of 100 models were incorrect.

curacy model that she uploaded for target T25 and selects one of Bates' acceptable models amongst her best 10 instead. Likewise, Bonvin misses all his 12 medium quality uploaded models for this target, submitting only three acceptable models, selected from his uploaded set and the set by Bates. A similar observation can be made for target T26 and T27 [Table III (b and c)]. Wolfson recognized none of his own acceptable or medium quality models for T26, but pick up two by S-Liang. S-Liang reciprocates by selecting one acceptable model from Wolfson's set. However he also recognizes 1 medium quality model and several acceptable models from his own uploaded set [Table III (b)].

The reason for this behavior is unclear. It might be due to the fact the ensembles of uploaded models from which scorers must select the best 10 is much smaller and highly enriched in acceptable or medium quality models in comparison to classical docking experiments. Hence, while the likelihood of picking up a correct model by chance is higher, singling out the better ones is challenging, as it requires finer discrimination between models.

As expected, the most "popular" uploaders, those whose correct models tend to be selected most often by scorers, are the groups contributing the largest number of such models to the uploaded set. The exception to this rule occurs for target T26, where the uploaded models of Gray (49 correct predictions) are picked up by none of the scorers, and those of Bonvin (13 correct predictions), are picked up only by the Weng group. This discrepant behavior is not due to any particularities in the scoring functions of these groups, but can be attributed to the fact that most of the correct models contributed by both uploader groups were not made available to scorers due to the formatting error aforementioned.

Are scorers improving the quality of the models they identify?

Since scoring procedures may involve some level of structure refinement prior to calculating the final score that is used to rank models, we compared the quality of the models submitted by scorers to the quality of the uploaded models from which they were derived. The comparison was based on the same criteria as those used to evaluate all CAPRI predictions, and the results are summarized in Supplementary Table S2. We find that in the majority of the cases scorers identified correct models without improving their quality. None of the best models submitted for target T25 represent improved versions of their uploaded counterpart. For the other two targets (T26 and T27), a few submitted models represented a slightly improved version of the uploaded complex, and only in one case (T27) a medium quality model was downgraded to acceptable.

DOI 10.1002/prot

HOW WELL ARE PREDICTORS RANKING THEIR MODELS?

The model number listed in the left-most column of Table I represent the position of the model in the ranked list of 10 predictions submitted by the predictors for each target. Many predictors produce this rank purely on the basis of the particular scoring scheme that they use. Some, however, trust less their scoring function and use ad hoc criteria. In either case the rank of the prediction reflects the degree of confidence that the predictors have in the submitted model, with high confidence models appearing at top of the list.

Figure 4 surveys the rankings of all the predictions submitted by each group for each of the seven targets evaluated here. Submissions for both the docking and scoring experiments are shown. Among the groups that most consistently rank their best predictions amongst the first five models are Weng, Bonvin, and Totrov. The groups of Wolfson and Fernandez-Recio achieve a similar performance in the scoring experiment. For most others, no clear trend could be detected, in either their docking or scoring submissions, and in general medium and high accuracy models (a single one in this evaluation) are not ranked higher than lesser quality models. This supports the conclusion reached from analyzing the results of the scoring experiment, namely, that scoring functions are presently not sensitive enough to discriminate between models of different accuracy levels.

DOCKING METHODS: HIGHLIGHTS OF RECENT TRENDS

CAPRI has been and still is a major testing ground for new docking methodology, and predictors often test different approaches from one target to the next.

With the rigid-body search algorithms, the core component of docking procedures, being well established,[22,23] recent efforts have concentrated on other aspects that are key to successful docking predictions. The most important of those are (1) developing better scoring function for singling out promising solutions, (2) modeling conformational flexibility, and (3) incorporation of nonstructural information.

Scoring functions

Recent developments in scoring functions have focused on the crude estimates of geometric and energetic complementarity during the rigid body search, as well as on the more sophisticated force fields used in subsequent refinement steps.

A new trend in scoring schemes used during the rigid body search is to rely on terms specifically designed to discriminate between correct and incorrect binding modes. Such terms, derived using statistical analyses of known protein interfaces, or Machine Learning techniques, were used by several groups who performed quite well in both the docking and scoring experiments (Weng, Jiang, and Vajda).

At the refinement step, which usually involves molecular mechanics or molecular dynamics techniques, force fields that are optimized to fit data on known interfaces (Rosetta, used by Baker's and Gray's groups), seem to perform better than generic classical empirical force fields. A major problem with both types of force fields seems to be their inability to account for specific interactions with water molecules that occur in some interfaces, like in target T26. To remedy this shortcoming several groups (Pal, Bonvin) are currently experimenting with the introduction of explicit solvent in the calculation.

Treatment of protein flexibility

Means of incorporating protein flexibility remains a major stumbling block that methods-developers are striving to surmount. Side chain rotations are relatively easy to handle, but modeling backbone flexibility is much harder. Two main strategies are being pursued by CAPRI groups. In one, championed by the groups of Bonvin, Bates, Zacharias, and Prévost, ensembles of conformers are generated for individual components of a complex by molecular dynamics or related methods. Pairs of conformers, one from each ensemble, are then systematically docked to one another. Groups such as those of Wang and Gray on the other hand, introduce backbone flexibility during the docking process. The latter is implemented using a Monte–Carlo sampling procedure, which includes moves representing rotations about selected backbone angles. The first of the two approaches is effective when the conformational changes are small, or involve low energy barriers. Rotation about selected bonds can produce large conformational changes, but is inefficient unless the relevant bonds, acting as hinges in the movement, can be located in advance. This is a difficult task for which new approaches relying on normal mode analysis and related methods have been proposed.[24]

Incorporation of nonstructural data

Most predictors try to optimally exploit biochemical information or any other nonstructural data that might be helpful. This is done either during the docking calculations by restricting the rigid body search to specific regions of the protein surface, or applying distance constraints, or by filtering candidate solutions output by the docking procedure. Automatic servers can be instructed to include or exclude certain residues from the computed interfaces, but introducing biases that incorporate a combination of features is still not readily feasible in an automatic fashion. A potentially useful source of information

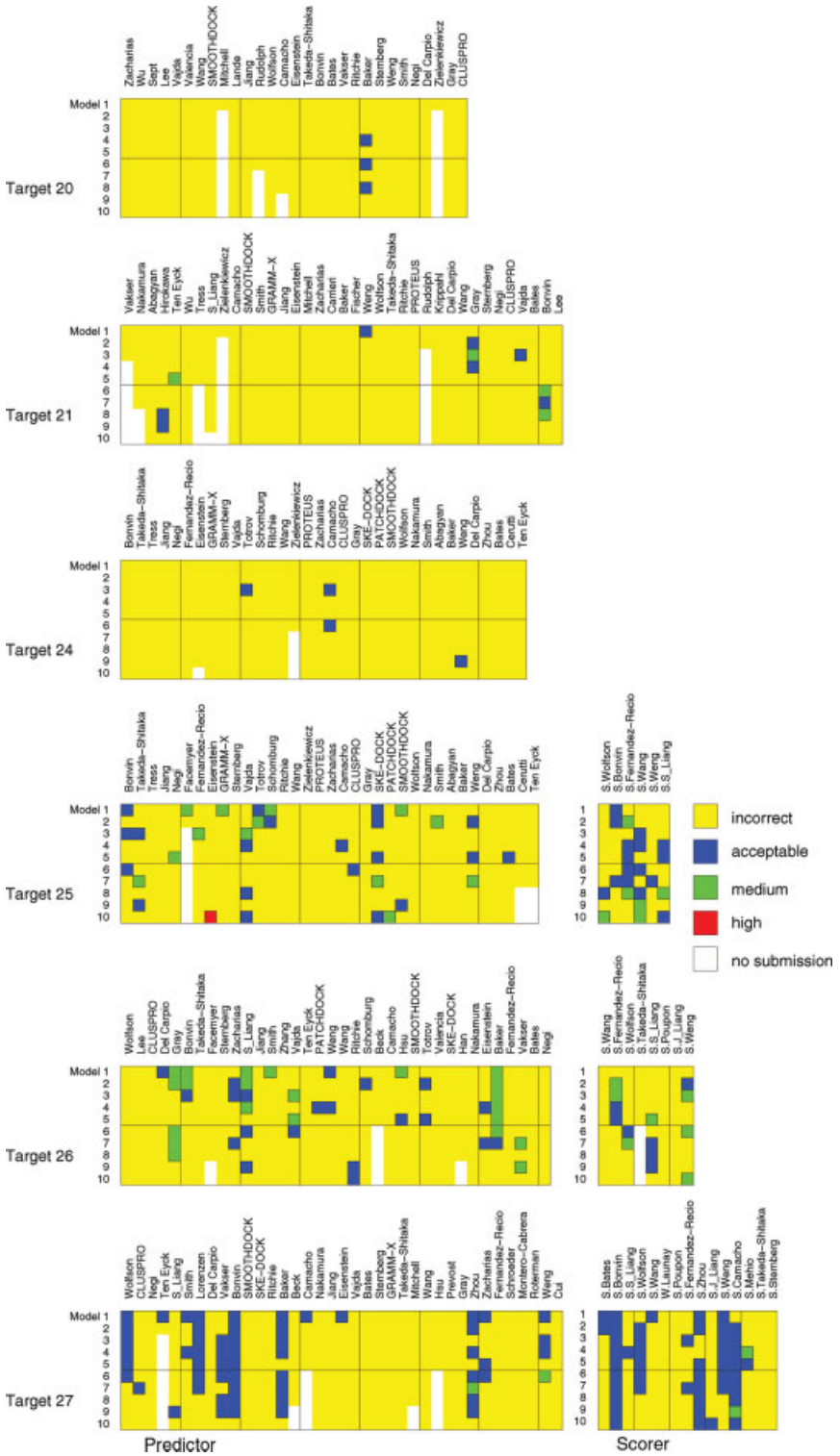


Figure 4

Participant's ranking of best predictions. The ranks given by each participant to their 10 submissions for each target is plotted. The participant groups (represented by the PI name) are listed along the abscissa. The rank of the submitted model (1–10) appears on the ordinate. Submitted models are color coded according to the model accuracy (shown on the right hand side), with the latter being defined as detailed in Figure 1. The left hand graphs show the rankings given to predictions for the classical docking experiment. The shorter graph on the right hand side shows rankings in the predictions for the scoring experiment.

DOI 10.1002/prot

is the sequences of related proteins. Multiple alignments of these sequences may reveal conserved positions, which when mapped onto the protein surface might represent regions involved in binding. This information was very instrumental in producing correct predictions for a number of targets in previous CAPRI rounds.[10,11] It was also used here by the groups of Nakamura, Zhou, and others for some of the targets. But with more of the targets in recent Rounds belonging to highly diverse protein families, or families with fewer members, the sequence signal was generally weak, and less helpful.

CONCLUDING REMARKS

What sets apart this latest set of CAPRI rounds from previous rounds, is that the majority of the targets were of the unbound–unbound category and involved types of proteins not often found in benchmark datasets used by docking methods developers. These targets therefore represent a collection of much more realistic and challenging docking problems than in previous reported assessments. Hence, although the results reported here do not reveal a striking breakthrough in docking performance, they should be considered as quite encouraging. Predictions were quite successful overall, when at least one of the subunits undergoes no backbone conformational changes upon association. In such cases interface residues are reliably identified, as is a good fraction of the native residue-residue contacts. The more accurate models also tend to have their side chains correctly positioned (see supplementary Fig. S1).

Two of the seven targets evaluated here are the products of structural genomics efforts (http://www.sgc.utoronto.ca/). With an increased focus of these efforts on protein assemblies, we expect this proportion to grow in future CAPRI rounds. This should supply more diverse and realistic docking problems, maintaining the role of CAPRI as the ideal testing ground for the performance of protein–protein docking procedures and of force-fields used to score predicted interactions. Recognizing this important role, new groups have been joining the CAPRI community, with their number doubling in the last 3 years, and many if not all of these groups are actively engaged in developing better methods. The thrust of recent developments is on handling protein flexibility, improving the discriminating power of scoring functions, and integration of nonstructural data. Although the challenges are significant, clear strides are being made, as recently reported in the 3rd CAPRI evaluation meeting held earlier in Toronto.[12]

In closing, we reiterate our call upon X-ray crystallographers and NMR experts to trust us with their structures. The submission of a target to CAPRI does not jeopardize the confidentiality of the work, since submitted atomic coordinates remain confidential until released by the author or by the PDB. On the other hand it enhances work done by the structural biologists, and in some instances, such as that of the UBC9-HIP2 complex of target T27[19] can even helps the interpretation of the structural data.

ACKNOWLEDGMENTS

The authors thank the University of Toronto for hosting the meeting where the results of this evaluation were presented. We also acknowledge support from Apple Canada, Cederlane Laboratories Ltd, the Canadian Institute of Health Research Institute of Genetics, Deloitte and Touche, Sun Microsystems, the SPINE2-Complexes EU program, Simulated Biomolecular Systems Inc., and Varian Medical Systems. We express the gratitude of all CAPRI groups to the crystallographers who provided the targets of Rounds 6–12: Marc Graille and Herman van Tilbeurgh (Université Paris-Sud, Orsay, France), James L. Keck (University of Wisconsin, Madison, USA), Tine K. Nielsen (Max Planck Institut, Göttingen, Germany), Louis Renault (LEBS-CNRS, Gif-sur-Yvette, France), Julie Ménétrey (Institut Curie, Paris, France), Colin Kleanthouse (University of York, UK), John Walker (Structural Genomics Consortium, Toronto, Canada). Finally our thanks also go to the CAPRI management team and all predictor groups for stimulating discussion, valuable input, and cooperation.

REFERENCES

1. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier MA, Hoffman V, Hoefert C, Klein K, Hudak M, Michon AM, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006;440:631–636.
2. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 2006;440:637–643.
3. Kim PM, Lu LJ, Xia Y, Gerstein MB. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 2006;314:1938–1941.
4. Dutta S, Berman HM. Large molecular complexes in the protein data bank: a status report. Structure 2005;13:381–388.
5. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The protein data bank. Acta Crystallogr D Biol Crystallogr 2002;58(Part 6 No 1):899–907.
6. Sali A, Glaeser R, Earnest T, Baumeister W. From words to literature in structural proteomics. Nature 2003;422:216–225.

7. Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T, Marinkovic N. High-throughput computational and experimental techniques in structural genomics. Genome Res 2004;14:2145–2154.
8. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round 6. Proteins 2005;61 (Suppl 7):3–7.
9. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of PRedicted Interactions. Proteins 2003;52:2–9.
10. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein–protein interactions: current status of docking methods. Proteins 2003;52:51–67.
11. Mendez R, Leplae R, Lensink MF, Wodak SJ. Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 2005;60:150–169.
12. Janin J, Wodak SJ. Meeting report: the third CAPRI assessment meeting Toronto, Canada, April 20–21, Structure 2007;15:755–759.
13. Janin J. The targets of CAPRI rounds 6–12, Proteins 2007;69:699–703.
14. Ménétrey J, Perderiset M, Cicolari J, Dubois T, Elkhatib N, El Khadali F, Franco M, Chavrier P, Houdusse A. Structural basis for ARF1-mediated recruitment of ARHGAP21 to golgi membranes. EMBO J 2007;26:1953–1962.
15. Lo Conte L, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol 1999;285:2177–2198.
16. Wodak SJ, Mendez R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14:242–249.
17. Bonsor DA, Grishkovskaya I, Dodson EJ, Kleanthous C. Molecular mimicry enables competitive recruitment by a natively disordered protein. J Am Chem Soc 2007;129:4800–4807.
18. Graille M, Heurgue-Hamard V, Champ S, Mora L, Scrima N, Ulryck N, van Tilbeurgh H, Buckingham RH. Molecular basis for bacterial class I release factor methylation by PrmC. Mol Cell 2005;20:917–927.
19. Vajda S. Classification of protein complexes based on docking difficulty. Proteins 2005;60:176–180.
20. McLachlan A. Rapid comparison of protein structures. Acta Crystallogr A 1982;38:871–873.
21. Hou Z, Bernstein DA, Fox CA, Keck JL. Structural basis of the Sir1-origin recognition complex interaction in transcriptional silencing. Proc Natl Acad Sci USA 2005;102:8489–8494.
22. Smith GR, Sternberg MJE. Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
23. Halperin I, Ma B, Wolfson H, Nussinov R. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 2002;47:409–443.
24. Bahar I, Rader AJ. Coarse-grained normal mode analysis in structural biology. Curr Opin Struct Biol 2005;15:586–592.

 DOI 10.1002/prot