Defining the limits of homology modeling in information-driven protein docking

J. P. G. L. M. Rodrigues,[1] A. S. J. Melquiond,[1] E. Karaca,[1] M. Trellet,[1,2] M. van Dijk,[1] G. C. P. van Zundert,[1] C. Schmitz,[1] S. J. de Vries,[1,3] A. Bordogna,[4] L. Bonati,[4] P. L. Kastritis,[1] and Alexandre M. J. J. Bonvin[1]*

[1] Faculty of Science/Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht 3584CH, The Netherlands

[2] Groupe VENISE, CNRS LIMSI, 91403 Orsay CEDEX, France

[3] Physik-Department (T38), Technische Universität München, James-Franck-Str. 1, 85748 Garching, Germany

[4] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy

ABSTRACT

Information-driven docking is currently one of the most successful approaches to obtain structural models of protein interactions as demonstrated in the latest round of CAPRI. While various experimental and computational techniques can be used to retrieve information about the binding mode, the availability of three-dimensional structures of the interacting partners remains a limiting factor. Fortunately, the wealth of structural information gathered by large-scale initiatives allows for homology-based modeling of a significant fraction of the protein universe. Defining the limits of information-driven docking based on such homology models is therefore highly relevant. Here we show, using previous CAPRI targets, that out of a variety of measures, the global sequence identity between template and target is a simple but reliable predictor of the achievable quality of the docking models. This indicates that a well-defined overall fold is critical for the interaction. Furthermore, the quality of the data at our disposal to characterize the interaction plays a determinant role in the success of the docking. Given reliable interface information we can obtain acceptable predictions even at low global sequence identity. These results, which define the boundaries between trustworthy and unreliable predictions, should guide both experts and nonexperts in defining the limits of what is achievable by docking. This is highly relevant considering that the fraction of the interactome amenable for docking is only bound to grow as the number of experimentally solved structures increases.

Proteins 2013; 81:2119–2128.

Key words: comparative modeling; structure quality; data-driven docking; biomolecular complexes; proteins; HADDOCK; CAPRI.

INTRODUCTION

Bruce Alberts referred to the cell as a "collection of protein machines".[1] This simple definition masks the complexity behind the machinery that indeed governs all processes essential to life. Much like most machines interface with others of their kind to collaboratively achieve a greater goal, proteins in the cell are organized in pathways, or networks, which are regulated with a formidable complexity.[2,3] These networks, collectively called the "interactome," are the fabric of life itself. Unfortunately, and despite decades of research, a large fraction of the interactome remains in the dark, unknown, and therefore beyond our reach.[4]

While cellular biology and molecular biology often answer the "what" and "where", knowledge of "how" a specific network operates begs for high-resolution structural information. Yet, experimental structural characterization of protein interactions is progressing slowly as compared to our increasing knowledge of the interactome. At the same time, we have access to a wealth of

Additional Supporting Information may be found in the online version of this article
Grant sponsor: VICI, Dutch Foundation for Scientific Research; Grant number: 700.56.442; Grant sponsor: Utrecht University; Grant sponsor: European FP7 e-Infrastructure project WeNMR; Grant number: 261572.
*Correspondence to: Alexandre M.J.J. Bonvin, Faculty of Science, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@uu.nl
Received 17 June 2013; Revised 16 July 2013; Accepted 25 July 2013
Published online 4 August 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24382

information that could potentially be used in computational structure prediction algorithms.[4,5]

Predicting the structure of a protein–protein complex *in silico* is not novel[6] and can nowadays be carried out using one of two major methods: comparative modeling or computational docking. Comparative modeling relies on the notion that a pair of interologs (conserved interaction between a pair of proteins, which have interacting homologs in another organism) often shares the same binding interface.[7] However, this approach can only reliably target a fraction of the interaction space: interactions for which no interologs can be found, or for which those found are below the threshold of reliability, cannot be modeled by comparative methods. Also, sequence similarity does not always convey interaction similarity,[8] nor even interaction specificity, as illustrated by a recent study on enzymes of the ubiquitination pathway.[9] In contrast to comparative modeling, computational docking predicts the structure of protein interactions from the structures of the unbound interacting partners by performing a search in the interaction space and assessing each model based on some scoring function. Community efforts on blind predictions (CAPRI)[10–13] have shown that explicit integration of experimental information during the docking calculations is valuable and increases their accuracy considerably.[13,14]

Regardless of the approach chosen, there is always the need for a three-dimensional (3D) structure of the interacting partners to start the modeling process. Large-scale structural genomics initiatives such as the Protein Structure Initiative[15] make it possible, to a certain degree, to find a sequence homolog with known structure, which can then be used to build a 3D model of the protein of interest. This, combined with the availability of homology modeling algorithms through web interfaces,[16,17] makes it rather simple for nonexperts to build models that can serve as input for docking predictions. However, simply put, a completely wrong model will never yield a good prediction. Akin to the notion of “twilight zone” of sequence alignment for homology modeling[18] that defines the sequence identity/similarity limit from which one can expect to build a reliable model, there must be an equally important “zone” where homology models are suitable for docking. The definition of this “twilight zone” for protein interaction modeling from homology models is therefore critical for single docking predictions and, perhaps more importantly, for high-throughput predictions of entire interactomes.

In this work, we address these concerns and identify the most suitable predictive metric for the reliability of homology-based information-driven docking. Using previous CAPRI targets for which information-driven docking has proven successful,[11–13] we generate by homology modeling structural models of varying sequence identities and perform docking with HADDOCK,[19–21] using the same information used in CAPRI. We analyze several sequence- and structure-based metrics, and discuss the impact of the quality of the homology model on the information-driven docking prediction. The influence of the quality of the interaction data on the final models is also analyzed. This allows us to define the limits of the achievable quality of a docking model for a given quality of a homology model. These are independently corroborated by our prediction results obtained in the last CAPRI rounds, which are also shortly discussed and summarized here.

MATERIAL AND METHODS

Dataset of protein-protein complexes

To assess the impact of the quality of a homology model in information-driven docking, we used protein–protein complexes from previous CAPRI targets (from rounds 4–19, Target 12 to Target 42) for which we had obtained a successful prediction. We only considered complexes for which at least one of the interacting partners has sequence homologs with an experimentally determined structure. These represent “real-life” scenarios. Protein–protein complexes presented in the last rounds of CAPRI were used as an independent validation set. The complexes that met the criteria and those used for validation are listed in Table I.

Homology modeling of interacting partners

Homology modeling was performed using a simple and straightforward protocol, detailed in the Supporting Information Material section. Thirty models per interacting partner/template pair (10/alignment method) were generated, resulting in a total of 870 different models across the entire dataset (for a total of 29 homologs for 6 targets—see Table I). Unaligned regions that resulted in long disordered loops or termini were removed.

Information-driven docking predictions using HADDOCK

We performed docking predictions for all targets using the various homology models of each chain and the reference bound structure of the other partner. In addition, for each target, bound–bound docking was performed to measure the best possible outcome. Homology model-homology model docking was not performed since it would be more difficult to isolate the impact of a given model on the docking results quality. This scenario is however present in the independent set from the current CAPRI rounds.

Two sets of restraints were used for each run: *CAPRI interface restraints and true interface restraints*. The CAPRI restraints comprise the information used during the corresponding CAPRI round, which is described in detail in previous publications[20,29] and is available upon

Table I
Dataset Collected for Measuring the Impact of Homology Models on the Docking Predictions

CAPRI target number	PDB ID	Protein name	Number of homologs found	Sequence identity of homologs (%)
Analysis set				
T12[22]	1OHZ	Cohesin	3	31–71
		Dockerin	2	37–46
T18[23]	1T6G	Xylanase	0	–
		TAXI	4	37–51
T26[24]	2HQS	TolB	0	–
		Pal	4	21–69
T27[a]	2O25	E2-25K	3	22–39
		Ubc9	4	29–55
T40[25]	3E8L	Serine proteinase inhibitor A	3	34–82
		Cationic Trypsin	0	–
T41[26]	2WPT	Im2 immunity protein	4	50–63
		Colicin-E9	2	66–67
Validation set				
T46[27]	3Q87	Trm112p-like protein	1	19
		Methyltransferase small domain	1	12
T50[28]	3R2X	Influenza Hemagglutinin	0	–
		HB36.3 designed protein	1	85
T53[b]	n/a	Rep4	1	67
		Rep2	0	–

[a] Walker JR, Avvakumov GV, Xue S, Newman EM, Mackenzie F, Weigelt J, Sundstrom M, Arrowsmith CH, Edwards AM, Bochkarev A, Dhe-Paganon SA. Novel and unexpected complex between the SUMO-1-conjugating enzyme UBC9 and the ubiquitin-conjugating enzyme E2-25 kDa (to be published).
[b] Designed Rep4/Rep2 α-repeat complex, by Minard P, Graille M. (Université Paris-Sud, France), in preparation.

request. True interface restraints were derived from the reference structure of the complex (all residues on each chain at a minimal atom distance cut-off of 5 Å from any residue in the other interacting partner), and included as ambiguous interaction restraints (AIRs).

All docking predictions were performed with HADDOCK[19,20] (beta version 2.2), using CNS[30] (version 1.3) for the structure calculations, with default settings, except for the number of models, which was set to match our previous CAPRI submissions. The HADDOCK score (explained in detail in the Supporting Information), used to rank the generated models after water refinement, consists of a weighted sum of physics-based energy terms (electrostatics and van der Waals) complemented by an empirical desolvation energy term (E_{desolv})[31] and a restraints energy term (E_{AIR}), as defined in Eq. (1):

$$\text{HADDOCK Score} = 0.2 \times E_{\mathrm{Elec}} + 1.0 \times E_{\mathrm{VdW}} + 1.0 \times E_{\mathrm{Desolv}} + 0.1 \times E_{\mathrm{AIR}} \qquad (1)$$

Structural quality and docking predictions assessment

To compare the structures of the homology models to those of the native proteins, individually or in the complex, two metrics based on the root mean square deviation of atomic coordinates (RMSD) were used:

- *backbone RMSD (bbRMSD)*, calculated between two chains and on the backbone atoms of the molecules (Cα, C, N, O).
- *interface RMSD (iRMSD)*, as defined by CAPRI,[10] calculated on the backbone atoms of residues within a minimal atom distance cut-off of 10 Å of any residues of a different molecule of the complex. When comparing single homology models to the crystal structure, this metric refers only to the interfacial backbone atoms of that partner.

To assess the quality of the restraints we calculated both precision and recall as defined in the equations below:

$$\text{Precision} = \frac{\text{correctly predicted residues}}{\text{predicted residues}}$$

$$\text{Recall} = \frac{\text{correctly predicted residues}}{\text{true interface residues}}$$

Predictive and measured indices for model quality

A previous study on the impact of homology models in protein-ligand docking[32] used several sequence- and structure-based indices to assess the quality of the models on the docking predictions. These are divided into two categories: "predictive" and "measured" indices. Predictive indices are those that can be calculated without prior knowledge of the structure of the complex, and can, therefore, be used to estimate the success of the docking prediction from the homology models. Measured indices are calculated knowing the structure of the complex and are used here to define the quality of a

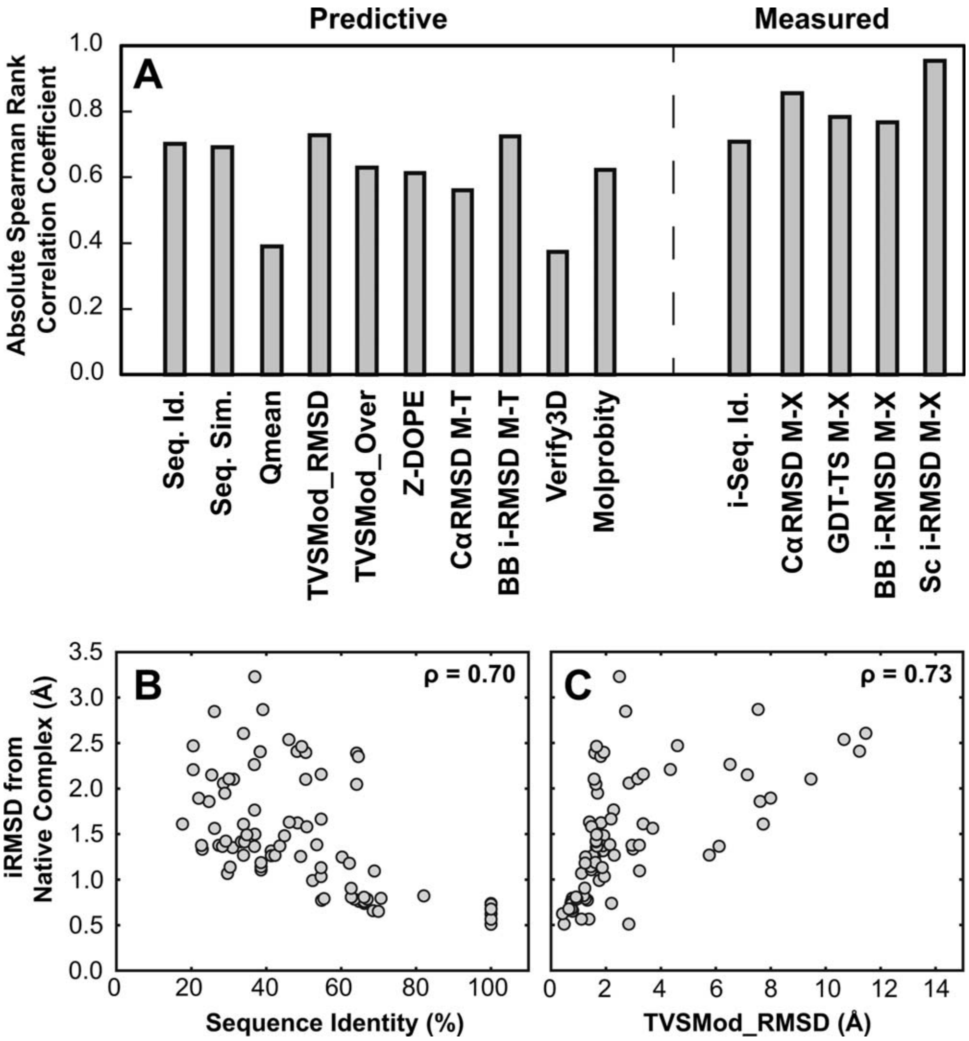

Figure 1

Correlation of sequence- and structure-based indices of homology models with docking model quality (see Supporting Information for a description and references of the methods used to calculated them). (**A**) Absolute Spearman correlation coefficients of the various indices with the backbone iRMSD from the native complex structure. (**B**) Correlation plot of sequence identity between target and template with the iRMSD of the best model produced using true interface restraints. (**C**) Correlation plot of TVSMod_RMSD score of the homology model and the iRMSD of the best model produced using true interface restraints. The corresponding Spearman rank correlation coefficients (ρ) are indicated in the figures.

docking prediction. See Supporting Information Table SI for details.

RESULTS

Correlation of predictive indices with docking model quality

We first performed a correlation analysis in order to assess if a particular index showed a predictive trend with respect to the quality of the final docking model [Fig. 1(A), Supporting Information Tables SI and SII]. All indices are described in detail in the Supporting Information. Sequence-based indices (e.g., sequence identity over the entire sequence or on the interface only) show relatively high and uniform correlations (∼0.7 absolute Spearman rank correlation coefficient). Structure-based indices, on the other hand, show a larger degree of heterogeneity, with coefficients ranging from ∼0.3 to ∼0.7. Overall, the highest coefficients are observed for the TVSMod_RMSD (0.73),[33] backbone iRMSD between model and template (0.73), and global sequence identity (0.70). The Qmean[34] and Verify3D[35] indices show the lowest correlation coefficients of all (0.39 and 0.37, respectively). The remaining indices have coefficients that fluctuate around 0.60.

We analyzed in more detail a representative of the sequence-based indices and another based on the structural properties of the model. For the sequence-based indices, we opted for sequence identity as it is easily derived from the alignment to be used in the modeling. As expected, as the sequence identity decreases, the quality of the model worsens [Fig. 1(B)]. Interestingly, even at very low identities, well inside the "twilight zone" of traditional homology modeling (~30% identity), our information-driven docking approach still produces near-native models (<3.5 Å iRMSD). The best correlating of the structure-based indices is TVSMod_RMSD.[33] It is based on support vector machine regression models and aims at predicting the RMSD and the fraction of Cα atoms of the model within 3.5 Å of those of the native structure after rigid superimposition. TVSMod_RMSD shows a similar trend as the sequence identity, although seemingly more discriminatory at lower iRMSD values [Fig. 1(C)]. Models that are predicted to be within 2 Å of the native structure produce, in general, docked models within 2.5 Å of the native complex. Beyond the 2 Å predictions, the correlation is less well defined, although there is still an observable trend.

Assessment of the impact of the quality of the data in the docking calculation

Although the quality of the homology model plays a large role in defining the success of the docking calculation, in information-driven docking the quality of the data is also quite relevant. To quantify this, we ran docking calculations with both perfect interface definition, derived from the crystal structure, and interface definitions as obtained from literature and/or bioinformatics predictions during the CAPRI round that produced the target (i.e., a reflection of what a researcher would have at hand in a real-case scenario). As expected, true interface restraints (meaning that the correct interface residues were used to define AIRs, which does not restrain the relative orientation of the molecules) produced very accurate results, with all models under 3.5 Å iRMSD of the native complex structure [Fig. 2(C,E)]. However, models produced using CAPRI restraints are highly dependent on the quality of the used information [Fig. 2(B,D)]. We calculated the precision and recall of the collected interface information with respect to the native interface [Fig. 2(A)]. The precision of the information across all targets was very high (above 80%), meaning that most of the data used to drive the docking were correct (these had been obtained from literature and bioinformatics predictions[20,29,36]. The recall was also reasonably high (in general above 50%), meaning that the fraction of the interface that was targeted was sufficient to avoid ambiguity during the calculations. The combination of these factors contributed to a good success rate and low iRMSD values for most targets. The exception was T18, in which the interface information for chain B was very narrow (recall 0.05), while for chain A the values of both precision and recall—0.38 and 0.31, respectively—were low. This led to docking solutions that were in general worse than for the other targets.

We also assessed if the quality of the homology model, measured by the sequence identity of the template to the target sequence, had a large impact on the ranking of the docking models at the last refinement stage in HADDOCK (Supporting Information Fig. S1). Again we observed that the quality of the data plays a more important role than the quality of the homology model, since the ranking of the final solutions is largely independent of the sequence identity.

Correlation of measured indices with docking model quality

Besides calculating predictive indices, which can be used to assess *a priori* whether a homology model will be able to produce a good docking model or not, we calculated a series of measured metrics that compare the individual models with the bound structure in the complex [Fig. 1(A)]. These, of course, cannot be used without *a priori* knowledge of the native complex. The structure-based metrics were: Cα RMSD, backbone iRMSD, side-chain iRMSD, and GDT-TS. All had high correlation coefficients with the quality of the docking models as measured by iRMSD, with the highest value found for the side-chain iRMSD of the model to the native complex (0.94). The correlation coefficient obtained for backbone iRMSD was the lowest of all four (0.69) but still reasonably high when compared with the predictive indices. We also calculated the sequence identity at the interface to assess its importance in defining the quality of the final models. Surprisingly, it shows only a slightly better correlation (0.71) as compared to global sequence identity, which can be calculated without knowledge of the native complex.

Impact of the flexible refinement on the quality of the docking predictions

The information used to drive the docking is translated generally into ambiguous distance restraints between residues on the surface of the proteins. This has particular importance during the flexible refinement stage, since interface residues are granted larger freedom. An analysis of the difference in iRMSD from the native structure between the initial rigid-body docking models and the final refined models reveals modest improvements up to about 1 Å RMSD changes [Supporting Information Fig. S2(A and B)]. A large fraction of models, however, does not improve or even deteriorates slightly, as it is typically observed in molecular dynamics simulations. Interestingly, the best improvements belong to cases with low template identity [Supporting Information Fig. S2(C and D)]. The quality of the restraints also plays a role in the extent of the improvement: using true interface restraints shifts the distribution of the

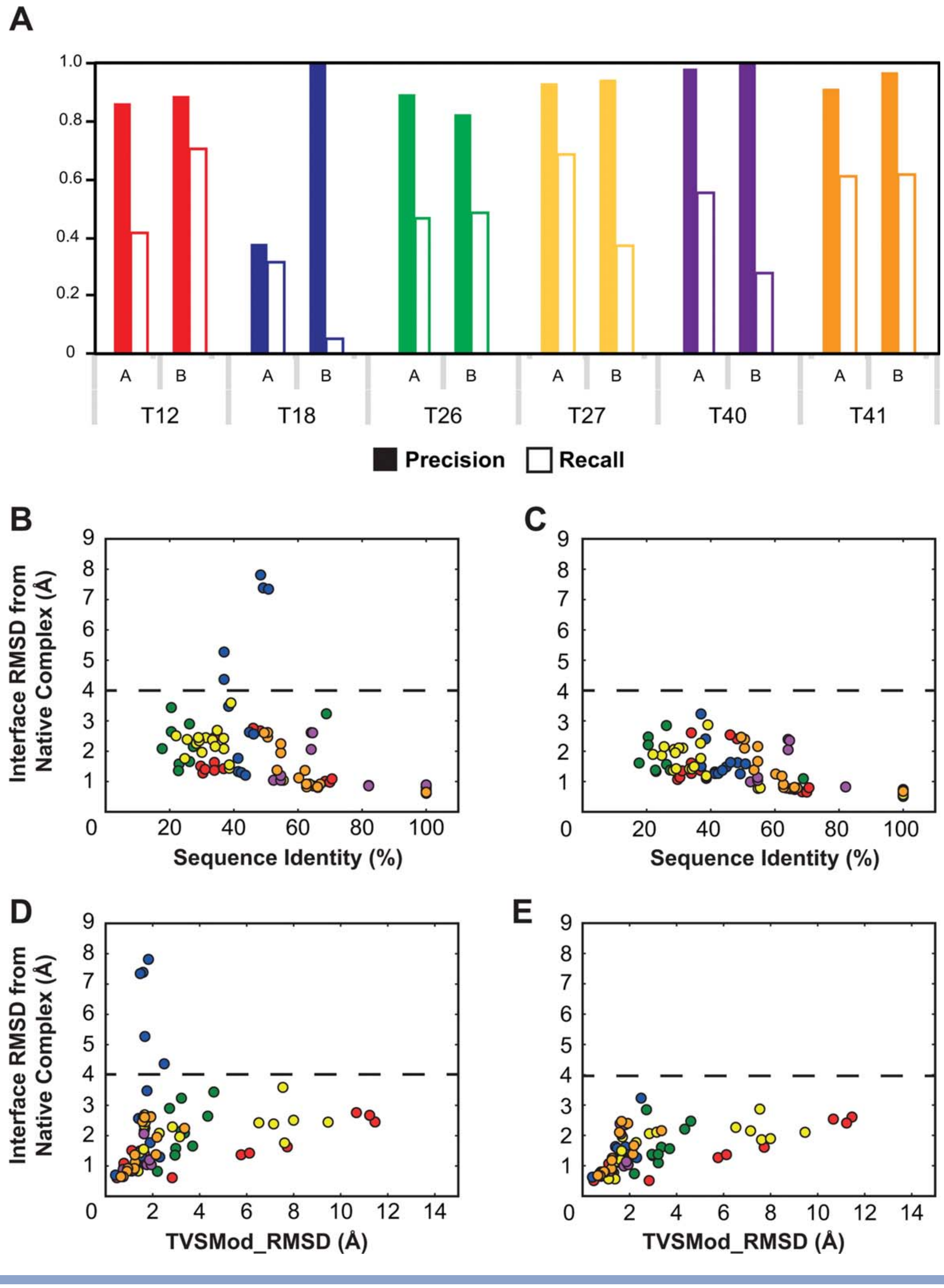

Figure 2

2Influence of the quality of the interaction data on the docking results. (**A**) Precision and recall metrics (see Methods section) on the interface information used as restraints to drive the docking process (color-coded by CAPRI target). (**B and D**) Correlation plot of sequence identity and TVSMod_RMSD of the homology model with the iRMSD from the native complex of the best docking solution based on CAPRI restraints. (**C** and **E**) Correlation plot of sequence identity and TVSMod_RMSD of the homology model and the iRMSD from the native complex of the best docking solution based on true interface restraints.

Table II
Performance of HADDOCK in Recent CAPRI Rounds

Target name	PDB Id	Target type	Manual submission performance										Manual submission ***/**/*	HADDOCK server ***/**/*	Uploaded structures ***/**/*	Scoring ***/**/*
T46[27]	3Q87	HH	0	1	0	1	0	1	0	0	0	0	0/0/3	0/0/10	0/0/22	0/0/2
T47[37]	3U43	UU	3	2	2	2	2	3	2	2	3	3	4/6/0/	9/1/0	112/88/0	9/1/0
T48[a]	N/A	UU	0	0	0	0	0	0	0	1	1	0	0/0/2	0/0/0	0/0/28	–[b]
T49[a]	N/A	UU	0	0	0	0	0	1	0	0	0	0	0/0/1	0/0/3	0/1/30	–[b]
T50[28]	3R2X	HU	0	0	0	0	0	0	2	2	0	0	0/2/0	0/0/0	0/15/10	0/0/2
T51[38]	N/A	UUHU(U)	0	0	0	0	0	0	1	1	0	0	0/2/0	0/0/0	0/1/7	0/0/0
T53[c]	N/A	UH	2	2	2	0	0	0	0	0	0	1	0/3/1	0/0/0	0/3/43	0/3/5
T54[d]	N/A	UH	0	0	0	0	0	0	0	0	0	0	0/0/0	0/0/0	0/0/1	0/0/0
T57[e]	4AK2	UU	0	1	0	1	0	0	2	0	0	1	0/1/3	0/1/1	Not assessed	Not assessed
T58[39]	4G9S	UU	0	0	0	0	0	0	0	0	1	0	0/0/1	0/0/0	Not assessed	Not assessed

We did not participate in the scoring round for targets T48/49. T52 was canceled and T55 and T56 were special scoring experiments, discussed in detail elsewhere.[40] The gray shading indicates the quality of the ranked predictions: high: black; medium: dark gray, and acceptable: light gray. The uploaded structures comprise 200 models corresponding to both sets of 100 models from the manual submission and from the server submission. The scoring column concerns the manual submission of the 10 best scored models according to HADDOCK from the pool of submissions from all research groups who contributed models to the scoring experiment.

[a] T48-49 T4moC/T4moH mono-oxygenase complex, by Fox B, Bailey L, Acheson J. (University of Wisconsin), in preparation.

[b] HADDOCK did not participate in this scoring round.

[c] Designed Rep4/Rep2 α-repeat complex, by Minard P, Graille M. (Université Paris-Sud, France), in preparation.

[d] Designed neocarzinostatin/Rep16 α-repeat complex, by Minard P., Graille M. (Université Paris-Sud, France), in preparation.

[e] A protein-polysaccharide complex, by Basle A, Lewis R. (Newcastle University, UK), in preparation.

differences in iRMSD between initial and docked model towards more negative (better) values and in general larger improvements. In these cases, the restraints were thus instrumental in improving the model or preventing it from deviating from the correct conformation. In general, these observations are in line with previous studies showing that the impact of flexible refinement is limited with typical maximal improvements in the order of 1.5 to 2.0 Å (see Fig. 2 in de Vries *et al.*[20].

Performance of information-driven docking with homology models in recent CAPRI rounds

The recent CAPRI rounds (22–27) provided a wealth of targets that required homology modeling with half of these requiring modeling of at least one of the binding partners. They can thus serve as an independent dataset. Table II reports the success rate of HADDOCK in these last rounds. Out of 10 targets, HADDOCK was successful in nine of them, corresponding to an unequaled 90% success rate when considering the manual submission entries. The HADDOCK server automated submission successfully predicted four out of 10 targets (40% success rate), but interestingly, in three of these targets (T46, T47, and T49) it outperformed the manual submission. In one, T48, the interaction information provided by the organizers (low-resolution SAXS data) did not improve the scoring of the models. In fact, the lowest fit to the SAXS data, with a χ^2 value close to that of the native crystal structure, belonged to a very different and wrong model. Instead, the application of a novel hydrophobicity potential[41] in the standard solvated docking algorithm of HADDOCK[42,43] yielded the only acceptable solution submitted for evaluation. Overall, these results rank HADDOCK as one of the best docking software to participate in CAPRI (and the best one in this round), reinforcing the idea that data-driven docking is a very successful approach to model biomolecular interactions.

Of the five targets requiring homology modeling, we could only analyze three, as the crystal structures are not yet all publicly available (Table II). The sequence identities of the models produced for CAPRI targets T46, T50, and T53 correlate nicely with the final quality of our best model (Fig. 3). The precision and recall rates for the information used to drive the docking were also extremely high (Fig. 3), in particular for target T46, in which we produced the only one-star models by CAPRI criteria (<4 Å iRMSD) despite the very low sequence identity to the templates used for the modeling: 12 and 18% (refer to the Supporting Information for a description of the restraints used in this target).

DISCUSSION

Sequence identity is a simple yet reasonably accurate predictor of docking success

The analysis of all sequence- and structure-based indices showed that none performs significantly better than the others. Sequence identity and similarity perform equally well (correlation coefficients of 0.70 and 0.69, respectively) and are trivial to calculate, requiring no further information than the pairwise alignment. Interestingly, the sequence identity at the interface is only a marginally better predictor (correlation coefficient of 0.71), which suggests that the overall fold of the molecule is relevant for a good arrangement of the interface and thus for the success of the docking.

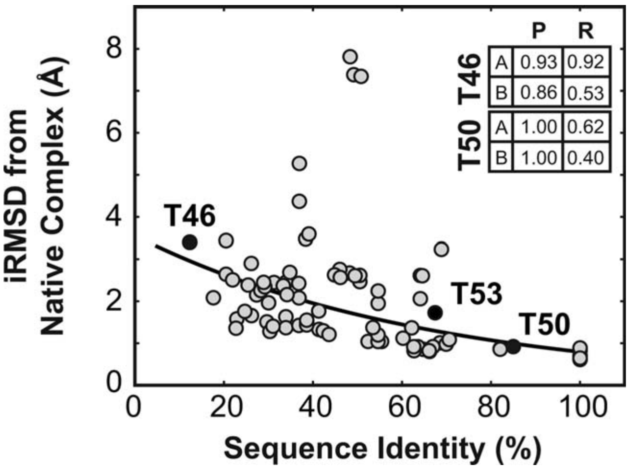

Figure 3

Relationship between sequence identity between target and template with iRMSD from the native complex of the docking models obtained using CAPRI restraints. Recent CAPRI targets (in black and labeled) nicely follow the predicted trend line from the homology-based docking study in this work. Inset Table Precision (P) and recall (R) of the information gathered for each target. Information for T53 is not yet publicly available.

Structure-based indices show a rather heterogeneous performance. The QMean,[34] Molprobity,[44] and Verify3D[35] metrics all evaluate the structural properties of the model, such as amino acid packing, distribution of torsion angles, etc. (Supporting Information Table S1). Since the homology models undergo a slight refinement, it is not expected that they have severe clashes or other deviant structural features. Nevertheless, Molprobity was very discriminative of native structures, attributing to these very low scores (almost always below 15 a.u.) in contrast to scores above 70 for the majority of the homology models. The scoring between the homology models was, however, heterogeneous and did not correlate with the docking results. Finally, the backbone iRMSD between model and template, a direct structural comparison measure, showed the highest correlation coefficient, on par with TVSMod_RMSD (0.73), and better than the overall structural similarity between the two structures (0.56).

The quality of the interaction restraints has a greater impact than the quality of the homology model

Information-driven docking narrows the conformational landscape of association of the molecules to the fraction that respects that information. Furthermore, if the information is integrated in the energy function used in refinement (i.e., not only for scoring), there is an added benefit of driving the interface refinement. Our results are in agreement with these assumptions, since docking calculations using literature-based information [CAPRI restraints, Fig. 2(B,D)] show worse results than those using true interface restraints [Fig. 2(C,E)]. The impact of the quality of the restraints is illustrated in the runs of T18, where precision and recall values were extremely low and the models were accordingly of bad quality (iRMSD over 4 Å). Overall however, despite starting the modeling process with templates as low as 20% sequence identity, the docked models are still quite reasonable (within 3 Å iRMSD), provided that the interaction information is reliable. This thus stresses the importance of the quality of the data over that of the model. The scoring of the models, helped by the interface information, is also robust enough to discriminate good quality models, regardless of the identity of the template used in the homology modeling. This again reinforces the notion that the quality of the data is more important than that of the model, since good data can refine a bad model and discriminate which solutions are closer to the native structure, while weak data pollute the docking protocol even when the model quality is reasonable.

Defining the limits of homology modeling in information-driven docking

On the basis of these observations, we can predict the quality of information-driven docking predictions given the sequence identity of the templates used to build the homology models (Fig. 3). Assuming reliable interface information, a homology model built with a template sharing 20% sequence identity can be expected to produce docking models within 4 Å iRMSD of the native complex. As the target-template identity increases, so does the expected quality of the final models. For example, most of the 60% identity models produced docking solutions around 2 Å iRMSD. This is likely to represent an overestimate of the achievable quality since one of the docking partners was taken in its bound form. Still, it is striking to see that the recent CAPRI targets, which were all homology–homology or homology–unbound docking cases, nicely follow the trend line of our model. This would indicate that the achievable docking quality is limited by the lowest sequence identity component of the interaction partners—in other words: the worse approximation defines the limits of your model.

The reliability of the information is of course hard to estimate. During a CAPRI round, most of the information is gathered from literature databases and bioinformatics predictions in the 24-hour period that comprises the server submission. All in all, this essentially means that reliable information is not so scarce as one might imagine. Finally, the homology modeling approach used in this study is standard, not using advanced refinement methods such as those available in structure prediction servers.[17,45] As such, the presented results can be considered a baseline, which can be further improved by expert knowledge of the system under study and/or more powerful structure prediction methods.

CONCLUSIONS

Information-driven docking is one of the most reliable and accurate prediction methods for modeling biomolecular complexes. Yet, it needs structural information of the interacting partners as starting point. Given the easy learning curve and widespread availability of homology modeling methods, experimentalists are bound to use them in docking in the absence of experimental alternatives. We have shown here that the global sequence identity between the target sequence and the template used for the modeling of the 3D structure is predictive of the achievable quality of the docking models. Nevertheless, the quality of the information used to drive the docking remains highly relevant and plays an important role in the outcome of the docking predictions. For templates well inside the so-called "twilight zone" of sequence identity (∼30%), good interface information is sufficient to produce models within 4 Å interface backbone RMSD. In contrast, bad quality information can severely diminish the success rate of the docking, even with models built with up to 60% sequence identity. These results allow assessing the suitability of a homology model in information-driven docking and set the stage for more confident predictions even in scenarios where the identity between the template and the sequence is remote, provided that the information on the interaction is reliable.

ACKNOWLEDGMENTS

Support of the European Grid Initiative (EGI) through the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, Spain, UK, South Africa, Malaysia, Taiwan and the Latin America GRID infrastructure via the Gisela project is acknowledged for the use of web portals, computing and storage facilities.

REFERENCES

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 1998;92:291–294.
2. Levy ED, Pereira-Leal JB. Evolution and dynamics of protein interactions and networks. Curr Opin Struct Biol 2008;18:349–357.
3. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. An in vivo map of the yeast protein interactome. Science 2008;320:1465–1470.
4. Melquiond AS, Karaca E, Kastritis PL, Bonvin AM. Next challenges in protein–protein docking: from proteome to interactome and beyond. WIREs: Comput Mol Sci 2012;2:642–651.
5. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A 2012;109:9438–9441.
6. Wodak SJ, Janin J. Computer analysis of protein-protein interaction. J Mol Biol 1978;124:323–342.
7. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003; 332:989–998.
8. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. J Struct Biol 2001;134:117–131.
9. van Wijk SJL, Melquiond ASJ, de Vries SJ, Timmers HTM, Bonvin AMJJ. Dynamic control of selectivity in the ubiquitination pathway revealed by an ASP to GLU substitution in an intra-molecular salt-bridge network. PLoS Comput Biol 2012;8:e1002754.
10. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. Critical Assessment of PRedicted Interactions. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.
11. Méndez 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. Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 2007;69:704–718.
13. Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. Proteins 2010;78:3085–3095.
14. Karaca E, Bonvin AMJJ. Advances in integrative modeling of biomolecular complexes. Methods 2013;59:372–381.
15. Montelione GT. The Protein structure initiative: achievements and visions for the future. F1000 Biol Rep 2012;4:7.
16. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009; 77(Suppl 9):128–132.
17. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–738.
18. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
19. 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.
20. de Vries SJ, van Dijk ADJ, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJJ. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 2007;69:726–733.
21. de Vries SJ, van Dijk M, Bonvin AMJJ. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 2010;5:883–897.
22. Carvalho AL, Dias FMV, Prates JAM, Nagy T, Gilbert HJ, Davies GJ, Ferreira LMA, Romão MJ, Fontes CMGA. Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc Natl Acad Sci U S A 2003;100:13809–13814.
23. 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.
24. 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.
25. Bao R, Zhou C-Z, Jiang C, Lin S-X, Chi C-W, Chen Y. The ternary structure of the double-headed arrowhead protease inhibitor API-A complexed with two trypsins reveals a novel reactive site conformation. J Biol Chem 2009;284:26676–26684.
26. Meenan NAG, Sharma A, Fleishman SJ, Macdonald CJ, Morel B, Boetzel R, Moore GR, Baker D, Kleanthous C. The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. Proc Natl Acad Sci U S A 2010;107:10080–10085.
27. Liger D, Mora L, Lazar N, Figaro S, Henri J, Scrima N, Buckingham RH, van Tilbeurgh H, Heurgué-Hamard V, Graille M. Mechanism of activation of methyltransferases involved in translation by the Trm112 "hub" protein. Nucleic Acids Res 2011;39:6249–6259.
28. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D. Computational design of proteins

targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816–821.

29. van Dijk ADJ, de Vries SJ, Dominguez C, Chen H, Zhou H-X, Bonvin AMJJ. Data-driven docking: HADDOCK's adventures in CAPRI. Proteins 2005;60:232–238.
30. Brunger AT. Version 1.2 of the Crystallography and NMR system. Nat Protoc 2007;2:2728–2733.
31. Fernández-Recio J, Totrov M, Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 2004;335:843–865.
32. Bordogna A, Pandini A, Bonati L. Predicting the accuracy of protein-ligand docking on homology models. J Comput Chem 2011;32:81–98.
33. Eramian D, Eswar N, Shen M-Y, Sali A. How well can the accuracy of comparative protein structure models be predicted? Protein Sci 2008;17:1881–1893.
34. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 2011;27:343–350.
35. Eisenberg D, Lüthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 1997;277: 396–404.
36. de Vries SJ, Bonvin AMJJ. CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS ONE 2011;6:e17695.
37. Wojdyla JA, Fleishman SJ, Baker D, Kleanthous C. Structure of the ultra-high-affinity colicin E2 DNase—Im2 complex. J Mol Biol 2012;417:79–94.
38. Najmudin S, Pinheiro BA, Prates JAM, Gilbert HJ, Romão MJ, Fontes CMGA. Putting an *N*-terminal end to the Clostridium thermocellum xylanase Xyn10B story: crystal structure of the CBM22-1-GH10 modules complexed with xylohexaose. J Struct Biol 2010;172:353–362.
39. Leysen S, Vanderkelen L, Weeks SD, Michiels CW, Strelkov SV. Structural basis of bacterial defense against g-type lysozyme-based innate immunity. Cell Mol Life Sci 2013;70:1113–1122.
40. Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, Zhou H-X, Mitchell JC, Demerdash ONA, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko J-S, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Pérez-Cano L, Pons C, Fernández-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AMJJ, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodríguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang S-Y, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 2011;414:289–302.
41. Kastritis PL, Visscher KM, van Dijk ADJ, Bonvin AMJJ. Solvated protein-protein docking using Kyte-Doolittle-based water preferences. Proteins 2013;81:510–518.
42. van Dijk ADJ, Bonvin AMJJ. Solvated docking: introducing water into the modelling of biomolecular complexes. Bioinformatics 2006; 22:2340–2347.
43. Kastritis PL, van Dijk ADJ, Bonvin AMJJ. Explicit treatment of water molecules in data-driven protein-protein docking: the solvated HADDOCKing approach. Methods Mol Biol 2012;819:355–374.
44. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 2010;66(Pt 1):12–21.
45. Rodrigues JPGLM, Levitt M, Chopra G. KoBaMIN: a knowledge-based minimization web server for protein structure refinement. Nucleic Acids Res 2012;40(Web Server issue):W323–W328.