PROTEINS: Structure, Function, and Bioinformatics 63:479–489 (2006)

WHISCY: What Information Does Surface Conservation Yield? Application to Data-Driven Docking

Sjoerd J. de Vries, Aalt D.J. van Dijk, and Alexandre M.J.J. Bonvin*
Bijvoet Center for Biomolecular Research, Utrecht University, Utrecht, The Netherlands

ABSTRACT **Protein–protein interactions play a key role in biological processes. Identifying the interacting residues is a first step toward understanding these interactions at a structural level. In this study, the interface prediction program WHISCY is presented. It combines surface conservation and structural information to predict protein–protein interfaces. The accuracy of the predictions is more than three times higher than a random prediction. These predictions have been combined with another interface prediction program, ProMate [Neuvirth et al. J Mol Biol 2004;338:181–199], resulting in an even more accurate predictor. The usefulness of the predictions was tested using the data-driven docking program HADDOCK [Dominguez et al. J Am Chem Soc 2003;125:1731–1737] in an unbound docking experiment, with the goal of generating as many near-native structures as possible. Unrefined rigid body docking solutions within 10 Å ligand RMSD from the true structure were generated for 22 out of 25 docked complexes. For 18 complexes, more than 100 of the 8000 generated models were correct. Our results demonstrates the potential of using interface predictions to drive protein–protein docking. Proteins 2006;63:479–489.** © 2006 Wiley-Liss, Inc.

Key words: protein complexes; interface prediction; conservation; docking; HADDOCK

INTRODUCTION

The number of known three-dimensional (3D) structures of proteins is growing faster than ever. A large majority of those corresponds, however, to structures of single proteins. In contrast, in the cell, proteins rarely carry out functions on their own, but usually by interacting with other proteins. Knowledge on interactions and biological function is growing quickly, while the number of 3D structures of protein–protein complexes is only slowly increasing. Considering the huge number of expected protein–protein interactions, conventional NMR and X-ray crystallography techniques will not be sufficient to tackle this problem. In particular, structures of weakly interacting or transient complexes are difficult to obtain. Therefore, there is a need for computational methods that can accurately predict the structure of a protein complex from the structures of its unbound components. This computational problem is known as the docking problem.

A variety of docking programs have been developed[1] (for review, see Halperin et al.[2]) To assess the state of the art in docking, the Critical Assessment of Predicted Interactions (CAPRI) experiment has been organized.[3,4] The results of the first CAPRI rounds have shown that there is currently no single method that can reliably dock each and every complex, although acceptable predictions are made for most complexes. Recently, we introduced the docking program HADDOCK.[5] Among the docking methods that have participated in CAPRI, HADDOCK is unusual because it is the only data-driven docking method. Many groups have made use of (putative) protein–protein interface residues to reduce the conformational search space or filter their solutions. In HADDOCK, however, these residues are translated into highly ambiguous intermolecular distance restraints used to directly drive the docking process. HADDOCK has been shown to reliably dock protein complexes provided proper information on the interface of the two proteins is supplied. HADDOCK results in CAPRI[6] have shown that the interface needs to be neither fully complete nor fully accurate to allow the generation of reasonable 3D models.[7] Incomplete experimental data can thus be sufficient for the docking, especially if multiple sources of data are employed, such as mutagenesis, truncation, or other biochemical data.[7] However, by far the fastest and most readily available source of restraints would be computational interface prediction from sequence and/or structural data.

The property that is most often associated with functional sites in general is sequence conservation.[8] Many residue mutations are not neutral, but negatively affect protein function, especially near functional sites. Such mutations are typically selected out by evolution, and as a result, functional residues are generally more conserved than others. This should be true for both residues within the protein core, important for folding and/or stability, and for residues on the surface, involved in biomolecular interactions. Therefore, as far as protein–protein docking

---

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: the Netherlands Organization for Scientific Research (N.W.O.) through a "Jonge Chemici" grant to AMJJB; Grant number: 700.50.512.

*Correspondence to: Alexandre Bonvin, Bijvoet Center for Biomolecular Research, Utrecht University, 3584CH, Utrecht, The Netherlands. E-mail: a.m.j.j.bonvin@chem.uu.nl

Received 4 April 2005; Revised 11 August 2005; Accepted 4 October 2005

Published online 31 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20842

is considered, it is essential to consider the conservation of surface residues only.

As reviewed in van Dijk et al.,[7] there are a few cases in which conservation data have been used as data in the docking of a protein–protein complex.[9–12] Their usefulness to filter docking solutions has recently been studied for a benchmark set[13] and the interface predictor ProMate,[14] which uses both conservation and biophysical properties, has recently been applied for the same purpose.[15] To the best of our knowledge, however, no systematic study has reported the direct use of conservation data to drive the docking process. Here, in the context of obtaining interface predictions for data-driven docking, we ask the following question: What information does surface conservation yield? We present a program named WHISCY, an acronym of this question, which tries to answer this question. In addition to sequence conservation, structural information in the form of surface neighbors used for smoothing and interface propensity is exploited as well. Our main goal is not to correctly predict the complete interface, but rather to obtain a few sufficiently reliable predictions for docking purposes. The feasibility of this approach is demonstrated by combining the WHISCY predictions with our docking approach HADDOCK. We also show that by combining various interface prediction methods, namely WHISCY and ProMate,[14] better results can be obtained.

THEORY

It has been previously recognized that multiple sequence alignments can be mined for conservation data. This has resulted in several different programs for interface and functional site prediction using conservation.[10,14,16–22] In addition, conservation is used in many predictors that combine multiple properties to predict protein–protein interfaces.[14,22–24]

To calculate conservation, a residue-specific matrix is often used to compare sequences in an alignment. This matrix can be either physicochemical or evolutionary. When comparing sequences, it is beneficial to also include the sequence distance, which is the number of mutations between sequences, because an alignment typically contains both sequences that are nearly identical and sequences that are very different. Evolutionary matrices offer the possibility to use a different matrix for each sequence distance. The direct use of evolutionary matrices induces, however, the following artifact: when a residue does not mutate (identity), it receives as a score the matrix diagonal element for that amino acid, which corresponds to the chance that the residue does not mutate. This chance is, however, the highest for amino acids that are known to mutate slowly; therefore, the comparison score will be the highest for identities of residues for which identities are most expected, violating the paradigm that conservation occurs when residues are expected to mutate but do not. This adverse scoring artefact is often overlooked.

In our approach, WHISCY, conservation is calculated by pairwise comparison of each sequence in a multiple sequence alignment to the master sequence only. All comparison scores for a surface residue are summed into a single conservation score for that residue, hence considering each surface residue independently. The matrix used for comparison is the Dayhoff matrix,[25] taking into account the sequence distance between each pair of compared sequences. The sequence distance is not determined by simply counting the number of mutations per hundred residues, because this does not take into account back-mutations and multiple mutations at a single site. Rather, the sequence distance is estimated using a maximum likelihood method implemented in the program PROTDIST from the PHYLIP package.[26] The appropriate matrix for each sequence distance is computed using eigenvector decomposition.

To avoid the adverse scoring artifact, a correction is applied to the Dayhoff matrix before calculating the comparison score. In general, if a residue has amino acid a in the master sequence, only one row of the mutation matrix is used: the row that describes all mutation probabilities $a \geq x$, where x is any amino acid. For each comparison, the sum of squares of all elements in this row is subtracted from each element in the row, a quantity that differs for each amino acid and each sequence distance:

$$S'_{ax} = S_{ax} - \sum_{k=1}^{N} S_{ak}^2$$

where S_{ax} indicates an element of the distance-dependent Dayhoff mutation matrix, S'_{ax} the corresponding element in the corrected matrix, and N the number of amino acids in the matrix. S_{ax} describes the probability of mutation of amino acid a into x, and can take values between 0 and 1.

Applying this correction makes sure that identities in fast mutating amino acids receive a higher score than identities in slow mutating amino acids. Moreover, it causes identities to receive a higher score at higher sequence distances, as long as the distance is not too large. This is opposite to the uncorrected matrix, which causes identity scores to decline with increasing sequence distance. The correction also results in the property that residues that behave as predicted by the matrix have, on average, an overall conservation score of zero. Only residues that mutate more slowly than predicted by the matrix will have an overall conservation score that is positive. Hence, WHISCY is robust for the presence of sequences in the alignment that have no functional relation to the master sequence and are similar in sequence only by chance, because their net effect will be zero. However, an alignment may also contain duplicate sequences, or certain organisms may be over-represented. To correct for this the sequences are weighted so that the whole range of sequence distances is represented equally. To accomplish this, the sequences are sorted from high to low sequence distance, and the attributed weight to each sequence is half the difference in sequence distance between the next higher sequence and the next lower sequence. This makes the score fully insensitive to the presence of duplicates.

The method described above was implemented and coded into a C++ program. The scoring scheme consis-

tently yields an average total score of zero for residues that behave exactly according to the matrix, as demonstrated by random (Monte Carlo) simulation (results not shown). However, for "real" residues, the total score was found to be consistently biased towards negative values. This is not surprising, because only surface residues are scored, which are generally more variable than core residues, while the whole sequence is used for the calculation of sequence distances. To correct this bias, the average total score over all residues is simply subtracted from each individual total score. Alternatively, only surface residues could be used in deriving the sequence distances. This would, however, be risky, because the Dayhoff matrix was not specifically designed for surface residues.

The translation of conservation into interface prediction can be refined by taking into account additional properties. Some amino acids are more likely to occur in interfaces than others, resulting in different interface propensities for each amino acid. Moreover, interface residues are not spread over the protein surface but often form one (or sometimes more) patches. Therefore, predicted residues that are surrounded by other predicted residues are more likely to be true predictions than isolated ones. To make use of this property, the 3D structure or a good model of the individual proteins must be available (which is anyway required when it comes to docking). These properties were each implemented separately to allow manual tuning of parameters. Interface propensity can be statistically derived as the frequency ratio of an amino acid to occur in interfaces and on the surface, divided by the frequency ratio of all interface and surface residues. Values for each amino acid have been established in several studies.[22,27–29] Considerable disagreement between these propensities exists, as there are large differences in dataset and in definitions of interface and surface. In WHISCY, the interface propensities as derived by Ma et al.[28] have been used. Because the conservation score is the sum of many comparison scores, it follows a normal distribution. This has been verified in a standard Q–Q plot (results not shown). Therefore, every score could be converted into a *p*-value. The *p*-value was divided by the interface propensity of the residue and converted back into a conservation score.

After adjusting for the interface propensity the scores were smoothed by considering surface neighbors. An optimal smoothing function was computed as a function of residue distance (see Supplementary Material). Crossvalidation showed that no overfitting had taken place. The exact shape of the curve does, however, not seem to be essential, as a simple Gaussian function with a single parameter (Fig. 1S) caused only a slight drop in performance (results not shown).

MATERIAL AND METHODS

Benchmark Sets

The docking benchmark assembled by Chen et al.[30] has been used as test set for the development of WHISCY. Antibody–antigen complexes were excluded. Fourteen chains were discarded due to lack of sequence data: 1ACBI, 1BTHP, 1CGII, 1CSEI, 1FQ1A, 1GOTB, 1KKLA, 1KKLH, 1L0YA, 1L0YB, 1PPEI, 1TABI, 1TGSI, 1UDII, 1UGHI, 2MTAA, and 2TECI. In addition, 2KAIAB and both chains of 2PTC were excluded because the numbering in those files prevented matching with the alignments. This resulted in a final dataset of 57 protein chains.

Six representative complexes were chosen from this set for testing in the context of prediction-driven docking: 1AVW, 1BRC, 1DFJ, 1WQ1, 2PCC, and 2SNI. These complexes were chosen blindly, checking if predictions were present for both chains but not if they were correct. In addition, we tested all enzyme–inhibitor/enzyme–substrate complexes taken from the new 2.0 benchmark set of Mintseris et al.[31] The complete docking set includes 25 complexes as listed in Table I. Unlike the development set, unbound structures were used for prediction.

The coordinates files of the proteins were downloaded from the Protein Data Bank[32] (http://www.rcsb.org/pdb), while the multiple sequence alignments were taken from the HSSP database (ftp://ftp.cmbi.kun.nl/pub/molbio/data/hssp).[21]

WHISCY Predictions

The first aligned sequence in the HSSP file was taken as master sequence. Multiple sequence alignments were used for WHISCY prediction: if there was any disagreement between structure and master sequences about a residue identity, the residue of the master sequence was used. The parts of a structure that were not present in the alignment were not predicted and ignored in the evaluation.

The protein surface was defined using NACCESS:[33] a residue was considered surface-exposed if its main chain or side chain was more than 15% accessible. The definition of the interface was performed with the program DIMPLOT, which is part of the LIGPLOT software[34] using default settings (3.9 Å heavy-atoms distance cutoff for nonbonded contacts).

Residues were predicted to be in the interface if the WHISCY score was higher than 0.180, corresponding to 29.4% sensitivity.

ProMate Predictions

ProMate predictions were obtained making use of the Web interface of ProMate (http://bioportal.weizmann.ac.il/promate/), using default settings. The file containing the quantitative, per-residue scores was parsed, and the top 10% scores were taken as prediction, as done in the original work of Gottschalk et al.[15]

Combining WHISCY and Promate

The WHISCY and ProMate predictions were combined in two different ways: (1) by addition (the Added score): a residue was predicted if it fulfilled either the WHISCY or the ProMate criterion; (2) by combination (the WHISCYMATE score): a residue was predicted if its ProMate score was higher than or equal to 98.520 or its WHISCY score higher than or equal to 0.371 or if its ProMate and WHISCY scores were both higher than or equal to 55.420 and 0.107, respectively. (See Supplementary Material for

TABLE I. Evaluation[a] of Interface Predictions for the Test Set Selected from the Docking Benchmarks[30,31]

	WHISCY		ProMate		WHISCYMATE		Added	
Complex	*correct*	*p* value	*correct*	*p* value	*correct*	*p* value	*correct*	*p* value
1ACB								
E (13/162)	6/16	4.0×10^{-04}	8/16	1.3×10^{-06}	7/17	4.6×10^{-05}	9/22	1.5×10^{-06}
I (10/61)	5/7	7.6×10^{-04}	3/6	0.050	8/15	7.6×10^{-05}	6/9	2.6×10^{-04}
1AVW								
A (19/144)	8/10	7.8×10^{-07}	10/14	9.3×10^{-08}	9/14	2.3×10^{-06}	12/18	6.6×10^{-09}
B (11/119)	0/3	1	5/12	1.5×10^{-03}	1/9	0.6	5/15	4.8×10^{-03}
1AVX								
A (18/143)	5/6	1.0×10^{-04}	12/14	1.6×10^{-11}	7/13	1.7×10^{-04}	12/14	1.6×10^{-11}
B (9/124)	0/2	1	6/12	1.5×10^{-05}	0/5	1	6/13	2.8×10^{-05}
1AY7								
A (10/78)	0/2	1	0/8	1	2/11	0.429	0/9	1
B (9/68)	6/7	5.1×10^{-06}	4/7	4.5×10^{-03}	7/8	2.9×10^{-07}	7/10	4.1×10^{-06}
1BRC								
E (14/140)	6/10	5.4×10^{-05}	6/10	5.4×10^{-05}	7/14	4.7×10^{-05}	7/14	4.7×10^{-05}
I (8/49)	2/3	0.065	1/7	0.738	4/5	1.5×10^{-03}	3/10	0.197
1BVN								
P (19/276)	6/17	3.6×10^{-04}	2/28	0.597	3/10	0.024	7/42	0.014
T (14/62)	4/4	1.9×10^{-03}	4/6	0.020	5/8	0.011	5/7	4.9×10^{-03}
1CGI								
E (22/161)	7/15	1.1×10^{-03}	12/16	2.2×10^{-09}	9/15	9.6×10^{-06}	12/21	2.5×10^{-07}
I (11/43)	1/6	0.851	3/4	0.046	4/11	0.284	4/10	0.214
1D6R								
A (12/145)	4/11	6.7×10^{-03}	7/15	1.7×10^{-05}	6/18	9.3×10^{-04}	7/19	1.2×10^{-04}
I (8/53)	0/1	1	0/5	1	2/31	0.994	0/6	1
1DFJ								
I (15/274)	3/9	9.2×10^{-03}	10/27	3.4×10^{-08}	3/5	1.3×10^{-03}	12/34	6.7×10^{-10}
E (16/98)	3/12	0.306	2/10	0.512	3/12	0.306	5/19	0.165
1E6E								
A (19/302)	13/52	3.6×10^{-07}	0/30	1	4/15	9.9×10^{-03}	13/70	1.8×10^{-05}
B (19/79)	11/20	4.7×10^{-04}	5/8	0.017	11/22	1.5×10^{-03}	11/20	4.7×10^{-04}
1EAW								
A (15/149)	7/15	9.8×10^{-05}	7/15	9.8×10^{-05}	8/21	1.4×10^{-04}	8/18	3.6×10^{-05}
B (7/48)	1/1	0.157	4/5	8.5×10^{-04}	2/2	0.019	4/5	8.5×10^{-04}
1EWY								
A (6/197)	5/28	2.3×10^{-04}	1/20	0.479	2/7	0.015	5/43	2.1×10^{-03}
C (7/73)	4/12	0.012	2/7	0.132	4/12	0.012	4/13	0.016
1EZU								
A (15/215)	8/19	3.5×10^{-06}	2/11	0.173	1/8	0.445	9/24	1.9×10^{-06}
C (23/144)	6/11	2.4×10^{-03}	6/14	0.011	6/20	0.071	8/18	2.0×10^{-03}
1F34								
A (20/204)	4/10	9.8×10^{-03}	3/20	0.31	4/5	3.2×10^{-04}	5/27	0.104
B (21/103)	0/0	—	5/10	0.028	0/1	1	5/10	0.028
1HIA								
A (15/140)	3/3	1.0×10^{-03}	8/14	4.8×10^{-06}	4/8	4.8×10^{-03}	9/15	4.4×10^{-07}
I (10/46)	1/2	0.395	1/5	0.725	5/18	0.33	2/7	0.48
1MAH								
A (12/296)	1/11	0.371	7/30	3.1×10^{-05}	1/5	0.188	7/39	2.0×10^{-04}
F (12/56)	0/3	1	0/6	1	0/4	1	0/8	1
1PPE								
E (15/146)	5/11	1.9×10^{-03}	10/15	8.6×10^{-09}	7/17	3.0×10^{-04}	10/19	2.3×10^{-07}
C (9/29)	1/1	0.376	0/3	1	5/19	0.88	1/4	0.796
1TMQ								
A (19/273)	4/12	5.9×10^{-03}	5/27	0.028	1/4	0.252	9/38	2.6×10^{-04}
B (14/87)	2/5	0.181	3/9	0.155	3/6	0.05	5/14	0.045
1UDI								
E (12/147)	4/10	4.2×10^{-03}	4/15	0.022	2/7	0.102	7/19	1.1×10^{-04}
I (13/66)	0/0	—	5/7	2.4×10^{-03}	2/6	0.337	5/7	2.4×10^{-03}
1WQ1								
G (13/213)	5/6	2.1×10^{-06}	0/21	1	0/0	—	5/27	0.014
R (14/113)	9/22	9.1×10^{-05}	5/11	4.3×10^{-03}	7/16	5.9×10^{-04}	9/23	1.4×10^{-04}

TABLE I. Continued

	WHISCY		ProMate		WHISCYMATE		Added	
Complex	*correct*	*p* value	*correct*	*p* value	*correct*	*p* value	*correct*	*p* value
2MTA								
L (11/280)	2/4	8.1×10^{-03}	7/28	1.2×10^{-05}	2/3	4.1×10^{-03}	7/30	2.0×10^{-05}
A (9/77)	3/5	0.01	4/8	5.2×10^{-03}	4/7	2.7×10^{-03}	4/10	0.014
2PCC								
A (4/192)	0/5	1	1/19	0.343	0/7	1	1/19	0.343
B (6/81)	0/10	1	0/8	1	0/14	1	0/12	1
2SIC								
E (17/161)	6/9	3.9×10^{-05}	3/16	0.229	4/5	4.1×10^{-04}	8/24	7.8×10^{-04}
I (12/85)	0/10	1	7/9	5.1×10^{-06}	2/14	0.63	7/19	3.9×10^{-03}
2SNI								
E (15/162)	7/11	3.5×10^{-06}	5/16	8.1×10^{-04}	5/6	2.0×10^{-05}	9/24	2.2×10^{-05}
I (6/53)	2/3	0.031	1/5	0.465	2/4	0.059	3/8	0.038
7CEI								
B (10/98)	0/4	1	0/10	1	0/2	1	0/13	1
A (12/71)	1/3	0.431	4/7	0.013	1/4	0.532	4/9	0.039

Predications were made on the unbound forms of the respective proteins.
[a]*Complex predictions:* list of all docked complexes with their respective chains with the number of interface and surface residues indicated between parentheses (see Material and Methods for the definition of interface residues). *Correct:* number of correct predictions and total number of predictions for each predictor. *p* value: chance of obtaining a prediction of at least the same quality by random selection of the same number of residues (hypergeometric distribution).

the optimization of the parameters for the combined score.)

Prediction-Driven Docking

The information content of our interface predictions in the context of docking was evaluated by rigid body docking only using a development version (2.0) of our data-driven docking program HADDOCK.[5] Predicted residues were defined as active. Surface residues (accessibility larger than 40%) within 6.5 Å of any active residue (distance between the closest nonhydrogen atoms) were defined as passive.

To account for the presence of false positives in our predictions, for each docking trial, 50% of the predictions were randomly discarded. The nonbonded intermolecular interactions were calculated with an 8.5 Å cutoff. The dielectric constant epsilon was set to 10. In each docking run, 2000 rigid body docking solutions were written to disk; for each, five trials were performed and only the best solution was kept according to the HADDOCK rigid body score, calculated as: $E_{\mathrm{vdw}} + 0.2 \times E_{\mathrm{elec}} + 0.01 \times E_{\mathrm{AIR}} + E_{\mathrm{desolv}} - 0.05 \times \mathrm{BSA}$; where E_{vdw} and E_{elec} are the van der Waals and electrostatic energies, E_{AIR} the ambiguous interaction restraint energy, E_{desolv} an empirical desolvation energy,[35] and BSA the buried surface area upon complex formation in Angstroms. For each set of interface predictions (WHISCY, ProMate, WHISCYMATE, and WHISCY + ProMate added predictions), a separate docking run was performed.

RESULTS AND DISCUSSION

WHISCY Performance on the Test Set

WHISCY was initially tested on the docking benchmark of Chen et al.[30] This benchmark consists of heterodimer complexes classified as enzyme–inhibitor, antibody–antigen, other complexes, and difficult complexes. All proteins were tested except the antibody–antigen complexes. Sequence alignments from the HSSP database were used. Because WHISCY is robust in respect to alignment errors, no further processing was necessary. However, some proteins were discarded because there were too few sequences available (see Materials and Methods).

To evaluate the performance of WHISCY, our predictions were compared with the "true" interface from the known 3D structures of the selected complexes. There are, however, multiple ways in which an interface can be defined. We have followed for this purpose a rather strict definition based on contact analysis using the program DIMPLOT.[34] This resulted, on average over all protein structures, in only 9.4% of all surface residues being defined as interface residues.

Our purpose has been to use WHISCY for restraint definition in HADDOCK. However, the output of WHISCY is a continuous range of scores for all surface residues supplied as input. This requires the definition of a cutoff for a binary classification of the scores into predicted and nonpredicted residues. The optimal cutoff depends not only on the reliability of the data, but also on the docking method and the protein complex. We therefore chose to evaluate the WHISCY performance in a cutoff-independent manner using a modified Receiver Operating Characteristic (ROC) plot.[36] We define the following quantities: (a) N_{TP} the number of correct interface predictions (true positives), (b) N_{FP} the number of wrong interface predictions (false positives), and (c) N_{I} total number of real interface residues as defined from the complex.

The pooled WHISCY performance over the test set is shown in Figure 1(A): the sensitivity S defined as $N_{\mathrm{TP}}/N_{\mathrm{I}}$ is plotted on the X axis, while the Y axis shows the normalized number of wrong predictions W, defined as

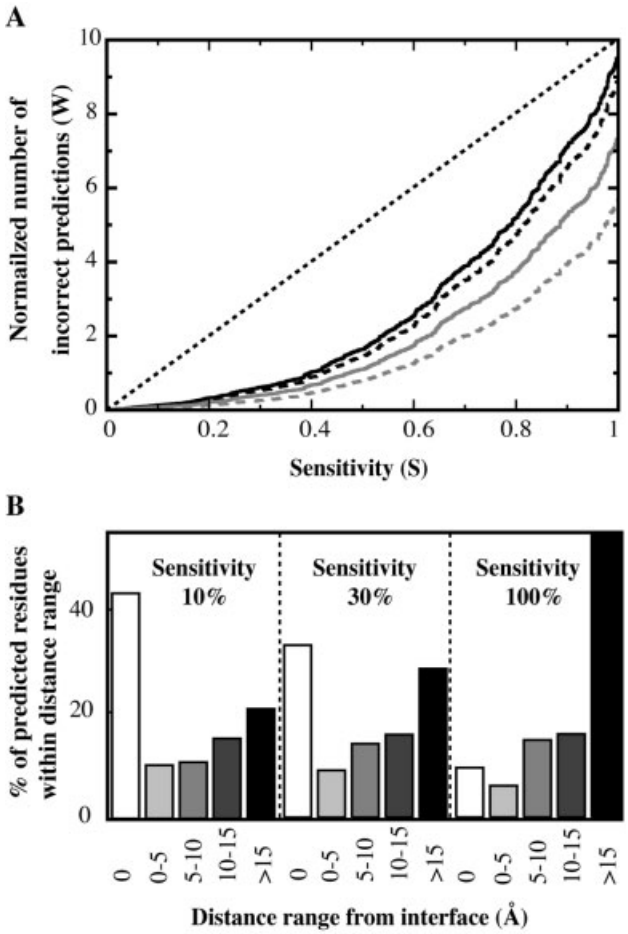

Fig. 1. WHISCY's performance on the test set. (**A**) Plot of the sensitivity S versus the normalized number of incorrect predictions *W*, considering all incorrect residues (black curve), considering only the incorrect residues >5 Å away (dashed black curve), >10 Å away (gray curve), and >15 Å away from the interface (dashed gray curve). *S* is defined as N_{TP}/N_I, where N_{TP} is the number of correct predictions and N_I the total number of real interface residues, and *W* is defined as N_{FP}/N_I, where N_{FP} is the number of incorrect predictions. Note that for a random predictor, this would be a straight line in the plot (dashed thin line). (**B**) Predicted interface residues classified according to their distance from the interface, for 15% sensitivity cutoff (left), 30% sensitivity cutoff (middle), and all residues (right). White bars: interface residues; light gray bars: residues within 5 Å of the interface; gray bars: residues between 5 and 10 Å; dark gray bars: residues between 10 and 15 Å; black: residues further away than 15 Å.

N_{FP}/N_I. This allows any chosen cutoff to be defined by its corresponding sensitivity *S*, and evaluated by its corresponding accuracy *A*, which is $N_{TP}/(N_{TP} + N_{FP})$ and can be evaluated from the plot as $S/(S + W)$.

Figure 1(A) clearly demonstrates that the WHISCY performance (black curve) is much better than a random predictor (dashed line). Using WHISCY with cutoff corresponding to a 10% sensitivity, 43% of the predictions are correct. For a cutoff of 30% sensitivity, one-third (33 %) of the predictions are correct according to our strict interface definition. The normalized number of wrong predictions made by WHISCY within a given minimum distance from the interface as a function of the sensitivity is presented in Figure 1(A) as well. When only noninterface residues further than 5 Å away from the protein partner are considered, the number of errors decreases somewhat. Note that the CAPRI[4,37] committee considers all residues within a 5 Å threshold as interface residues. Wrong predictions at longer distances can be analyzed by further increasing the distance threshold. Considering the entire surface, residues further than 15 Å away from the partner protein form the majority of the residues: for each interface residue, there are 9.6 noninterface residues, 5.5 of which are further than 15 Å away. These residues are, however, underrepresented in the WHISCY predictions, especially at strict score cutoffs. This is most clear from Figure 1(B), in which the predicted residues have been classified based on their distance from the partner protein; shown are the respective percentages for the 10 and 30% sensitivity cutoffs and for all residues. For both cutoffs, the interface fraction is the largest fraction.

The statistical significance of these results can be tested using the hypergeometric distribution. At 30% sensitivity, this test shows with high significance that the WHISCY predictions are better than a random selection (interface vs. noninterface residues: $p = 1.1 \times 10^{-74}$) Moreover, if only the noninterface residues are considered, the category of residues closest to the interface are overrepresented (residues <5 Å vs. residues >5 Å: $p = 4.6 \times 10^{-8}$). If only the residues further than 5 Å are considered, the category closest to the interface is again overrepresented (residues 5–10 Å vs. residues >10 Å: $p = 0.00028$), and this is also true for residues further than 10 Å (residues 10–15 Å vs. residues >15 Å: $p = 1.1 \times 10^{-8}$). This indicates that the WHISCY predictions do not only contain a large overrepresentation of the true interface, but also of residues close to the interface, which causes the large majority of the predictions to be at or near the interface. If all residues within 15 Å of the partner protein would be counted as correct, the accuracy would be 79 and 72%, even though those residues cover only 45% of the total surface.

This suggests that interface conservation is a low-resolution phenomenon: biomolecular interactions induce residue conservation up to quite a long distance from the interface. Alternatively, this could be an artefact of either our smoothing procedure or of our tight definition of the interface.

WHISCY Compared to ProMate

Figure 2 shows a comparison of prediction quality between WHISCY and ProMate[14] for the same test set. This figure shows that the performances of WHISCY and ProMate are comparable. This is remarkable, because WHISCY is based on conservation whereas ProMate is based on many properties, most of them biophysical, with only a simple form of conservation implemented. Below 15% sensitivity, both methods are highly accurate and show a similar performance. Between 15 and 55%, WHISCY is clearly superior. If a very high sensitivity is required, beyond 60% of the interface predicted, ProMate is the most accurate method. However, this is merely an average over the test set, and the two approaches differ widely in their performance for individual proteins. Visual

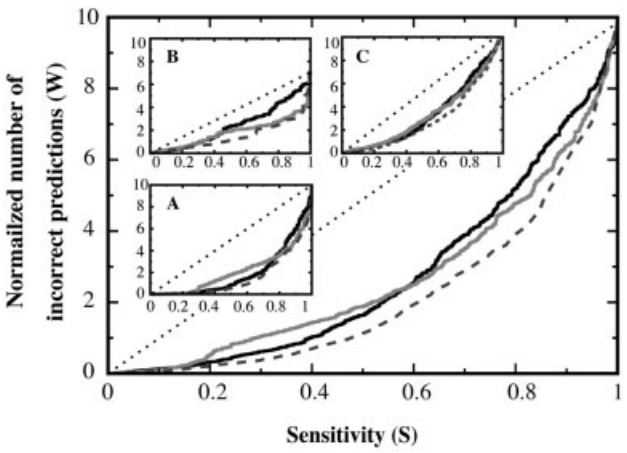

Fig. 2. Performance of interface prediction. Plot of the sensitivity *S* versus the normalized number of incorrect predictions *W* (see caption of Fig. 1 for definition). Main figure: comparison of WHISCY (black curve), ProMate (gray curve), and WHISCYMATE (dotted gray curve). Insets: WHISCY (black curve), ProMate (gray curve), and WHISCYMATE (dotted gray curve) for (**A**) enzyme interface prediction (21 test cases), (**B**) inhibitor interface prediction (11 test cases), and (**C**) nonenzyme/inhibitor interface prediction (25 test cases).

inspection of the predictions for each individual protein indicates that both ProMate and WHISCY outperform the other in about half of the cases. The better performance of WHISCY in Figure 2 merely indicates that WHISCY shows a better ability to determine the number of predictions that can be done at high confidence. Indeed, in using ProMate we did not use a cutoff but selected the top 10% ranking scores as was done in Gottschalk et al.[15] Note that in the original ProMate article[14] the authors also suggest to select instead the top ranking patch for prediction.

Overall, these results indicate that WHISCY and ProMate perform equally well in predicting protein interfaces. Both methods give accurate results if the required sensitivity is not too high. However, the large differences in performance for individual proteins indicate that a combination of the two may give even better results.

Combined Predictions with WHISCY and ProMate

We developed a procedure to combine the WHISCY and ProMate scores into a new predictor. The general model is that a residue score should be either very high for ProMate OR WHISCY; or moderately high for ProMate and WHISCY. Hence, this model is described by four threshold parameters: two that define "very high" for ProMate OR WHISCY, and two that define "moderately high." These parameters can be optimized for any desired sensitivity. The result of the procedure was 100 sets of these four parameters, roughly corresponding with cutoffs of 1 to 100% sensitivity (see Supplementary Material). To prevent overfitting of the data, crossvalidation was used. Together, these sets define the ProMate-WHISCY combined score, which we will refer to in the following as the WHISCYMATE score.

Figure 2 shows the performance of the WHISCYMATE score compared to WHISCY or ProMate alone. For a 10% sensitivity cutoff, the number of errors is halved, causing an accuracy increase from 43% to over 60%. For the 30 and 50% sensitivity cutoffs, the accuracy becomes 44 and 30%, respectively. This is comparable with the WHISCY accuracy for 10 and 30% sensitivity, respectively. Hence, by using the WHISCYMATE score, either the accuracy for a given number of predictions can be much increased, or a much larger number of predictions can be made at the same accuracy.

For the higher sensitivity region in the plot, the effect is substantial as well. Although ProMate performs better than WHISCY in this region, the performance difference between ProMate and the WHISCYMATE score is much larger. At 70% sensitivity, the WHISCYMATE score makes 16% less errors than ProMate, increasing the accuracy from 16.7 to 19.4%.

Figure 2 also shows the performance of the three methods for enzymes, inhibitors, and other proteins. Due to the low number of proteins in each class and the high homology between some of those proteins, a quantitative analysis cannot be performed. However, it is clear that in all three classes, the WHISCYMATE score performs as least as good as the best of ProMate and WHISCY. By far the best predictions are made for enzymes, and WHISCY is superior to ProMate in this class. Interfaces for inhibitors are much harder to predict, although the ratio of interface to noninterface is much better. WHISCY seems to be better at strict cutoffs, but ProMate is able to correctly eliminate a large percentage of the surface from the potential interface. The proteins that are not from an enzyme–inhibitor complex are the most difficult to predict. However, predictions are still much better than random. ProMate and WHISCY perform about equally here.

Using WHISCY and ProMate with HADDOCK

Our main purpose in developing WHISCY was to obtain interface predictions that could be used in data-driven docking.[7] Therefore, the predictions were used to define ambiguous interaction restraints for HADDOCK.[5] This requires the choice of a cutoff, which is not a trivial task. Choosing a cutoff that is too strict may cause too few restraints to be defined to drive the docking, whereas a more generous cutoff may introduce so many errors that only wrong solutions will be generated. An important criterion here is that at least one interface residue should be correctly predicted for the large majority of the proteins, whereas the number of proteins with at least one wrong prediction (>15 Å from the partner) should be minimized. A plot of these statistics as a function of the cutoff was generated (results not shown). The optimal cutoff was manually determined at 29.4% sensitivity: 75.4% of the proteins have at least one correct prediction, whereas 66.7% have one wrong prediction or more. The accuracy of the predictions at this cutoff is 33.3%. This is over three times the random score of 9.4%. The overall quality of the predictions is very similar to the 30% cutoff shown in Figure 1(B). The actual WHISCY score cutoff, corresponding to 29.4% of the true interface selected, is 0.180.

For ProMate, the best results are obtained if a top ranking percentage of the residues are predicted.[14] We selected the top 10% as interface prediction, as used by Gottschalk et al.[15] The cutoff for the WHISCYMATE score was chosen at 35.4% sensitivity, to obtain on average the same number of predictions as for WHISCY. Finally, a fourth set of predictions was obtained by simply adding the predictions for ProMate and WHISCY. This will be referred to as the Added predictions.

For testing the predictions in docking, we assembled a test set of 25 proteins, consisting of six representative complexes from the benchmark 1.0 used in WHISCY development,[30] and all enzyme–inhibitor/enzyme–substrate complexes from the recently published docking benchmark 2.0.[31] The latter set contains several complexes that have no equivalent in benchmark 1.0, and hence, have not been used in the development of WHISCY. For the selection of the six representative complexes from the 1.0 benchmark, the only criterion was that at least one WHISCY residue score passed the cutoff for each partner. It was not examined in advance to what extent the predictions were correct.

The performance of the interface residue predictions for the 25 selected complexes is evaluated in Table I. Each predictor performs better than random for the large majority of the proteins using the chosen cutoffs (see above). 62.5% of the predictions are significant at the 5% level. The Added prediction, which predicts the largest number of residues, performs best in terms of significance, followed by WHISCYMATE. Because the Added score combines WHISCY and ProMate in a nonoptimized way but usually predicts the largest number of residues, this suggests that the currently used cutoffs might be too conservative. The predictions mapped onto the surface of the proteins for the six representative complexes from the 1.0 benchmark are shown in Figure 3.

Note that neither WHISCY nor ProMate nor their combination is fully insensitive to the 3D structure used in the prediction: a comparison of predictions obtained from the bound and unbound 3D structures reveals small variations but no trend towards a performance gain or loss (data not shown). This is in agreement with previous results that have shown that interface prediction is robust for switching from bound to unbound predictions.[14,22] However, in individual cases, there might be small differences, especially if conformational changes are occurring between the unbound and bound forms.

Successful docking requires the tackling of two problems: the generation of correct structures and subsequently their identification by scoring. As an initial test of the inclusion of interface predictions to drive the docking and to limit computational costs, we limited ourselves to the first problem using only the rigid-body docking stage of HADDOCK. For each set of predictions (WHISCY, ProMate, the WHISCYMATE score, and the Added score) 2000 rigid-body docking solutions were generated for each of the 25 complexes. We also performed control runs in which random patches, different for each of the 2000 docking structures, were used to generate the restraints for docking. Interface prediction and docking were performed from the structures of the free proteins (unbound structures), meaning that (small) conformational changes might be required for proper docking.

The number of correct rigid-body docking solutions obtained for the various predictions is presented in Table II for the various complexes. A docking solution is defined as correct if its ligand root-mean-square deviation (ligand RMSD) from the experimentally determined complex is less than 10 Å (acceptable solutions in CAPRI terms). The ligand RMSD from the target is computed on the backbone atoms of the smaller protein after positional least-square fitting on the backbone of the largest component. For individual docking runs, large differences are found in the number of correct structures generated. However, for a majority of the runs (57 out of 97), prediction-driven docking performed significantly better than the control run at the 1% significance level. The best results were obtained with predictions made by the Added predictor, followed by WHISCYMATE.

If all the runs are pooled, correct structures are generated for 22 out of 25 complexes. Moreover, the number of correct structures is in the order of hundreds (out of a total of 8000 generated structures) for 18 out of 25 complexes. These 18 complexes include 1ACB and 1WQ1, which are considered (medium) difficult targets in the benchmark. These numbers can be considered an encouraging success for an initial test, especially when compared to the control runs with random interface patch definitions.

In general, docking was successful if the predictions for both protein chains were more than 20% accurate. However, it is surprising to see that many runs yielded correct structures even while the prediction for one or both of the partners was not very good. This could be due in part to the random removal of restraints that we implemented in HADDOCK: by discarding randomly half of the restraints for each docking trial, ambiguous interaction restraints involving false positives might be removed allowing for correct solutions to be generated, provided that at least one correct prediction was made for each chain.

There are two cases that escape this general pattern. For 1EZU, no correct solutions were generated even while the WHISCY predictions were excellent. The reason of this failure is not clear to us at this time. In contrast, for 2PCC, many good structures were obtained using ProMate and the Added predictor, even while there were few or no correct predictions at all. This can be explained by the fact that for 2PCC, most predicted residues are near the interface [see Fig. 3(F)], indicating that these residues can also be helpful in docking.

CONCLUSIONS AND PERSPECTIVES

In this study, we have described the interface prediction program WHISCY and its combination with another predictor, ProMate, in the context of prediction-driven docking. Our results strongly confirm the hypothesis that surface conservation yields useful information for data-driven docking. WHISCY predictions identifying 30% of those residues were 32.9% accurate, 3.5 times better than

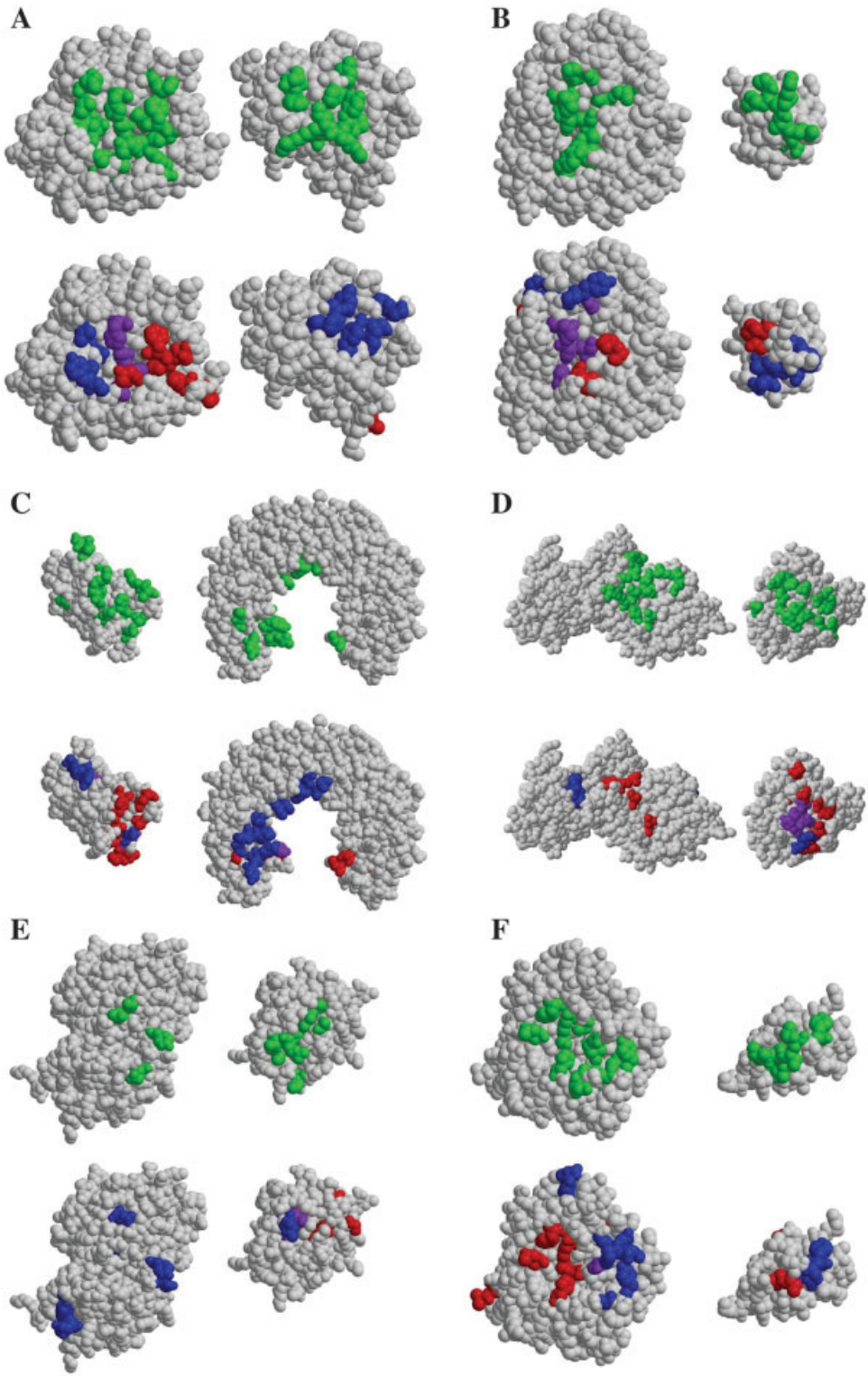

Fig. 3. Predicted versus true interface residues mapped onto the two chains of the six complexes selected from the 1.0 docking benchmark:[30] 1AVW (a), 1BRC (b), 1DFJ (c), 1WQ1 (d), 2PCC (e), and 2SNI (f). Green: interface as determined by DIMPLOT (see Material and Methods); red: WHISCY prediction; blue: ProMate prediction; purple: overlap between WHISCY and ProMate prediction.

TABLE II. Performance of Interface Prediction-Driven Docking with HADDOCK on the Test Set: Number of Correct Structures (1-RMSD <10 Å) out of 2000 Generated Structures

Complex	Control	WHISCY	ProMate	WHISCYMATE	Added	Total
1ACB	1	19*	51*	21*	30*	121
1AVW	0	0	57*	13*	66*	136
1AVX	0	1	69*	0	99*	169
1AY7	1	2	0	23*	1	26
1BRC	0	238*	29*	150*	143*	560
1BVN	5	41*	1	61*	11	114
1CGI	3	11	691*	171*	115*	988
1D6R	0	0	0	0	0	0
1DFJ	1	4	60*	11*	225*	300
1E6E	1	185*	3	380*	9	577
1EAW	0	45*	66*	21*	31*	163
1EWY	8	386*	4	138*	181*	709
1EZU	0	0	0	0	0	0
1F34	1	NP	1	0	7	8
1HIA	3	266*	114*	71*	167*	618
1MAH	0	1	0	1	1	3
1PPE	18	513*	579*	319*	647*	2058
1TMQ	0	265*	164*	71*	416*	916
1UDI	0	NP	117*	109*	136*	362
1WQ1	1	344*	2	NP	11*	357
2MTA	1	12*	59*	147*	26*	244
2PCC	1	1	92*	4	82*	179
2SIC	0	0	6	0	3	9
2SNI	4	69*	7	189*	103*	368
7CEI	3	0	0	0	0	0

An asterisk (*) indicates that the number of correct structures is significantly higher than the control run with random patch definition ($p < 0.01$; Fisher exact test). NP = not performed due to a lack of predictions for one of the chains.

random. By shifting the prediction cutoff, the accuracy can be increased even more at the expense of sensitivity, or vice versa. Some complexes clearly have more conserved interfaces than others. Predictions made by WHISCY were accurate for most enzymes, but they were less reliable for inhibitors and proteins not part of enzyme–inhibitor complexes. This is in agreement with Bradford et al.,[38] who determined interface conservation for enzymes and inhibitors using Rate4Site.[19] Our results are particularly impressive if one considers the fact that a protein may have multiple interfaces, so that predicted residues that are part of a different interface are scored as wrong predictions. For example, the Ras protein, chain 1WQ1R in our test set, is known to be involved in at least 11 distinct interactions[39] that might not be necessarily mediated by the same interface residues. Finally, one must bear in mind that the chosen interface criterion that marks 9.4% of the surface as interface is extremely strict. Other groups have used criteria resulting in 30% of the surface to be marked as interface;[22,40] all predictions in this large area of the surface would thus be considered correct. This explains some of the high accuracies that are sometimes reported in the literature.

For docking purposes, the interface predictions obtained from WHISCY, ProMate, and their combination were good enough to generate correct structures for 22 out of the 25 complexes used for testing purposes with only the rigid-body part of our data-driven docking program HADDOCK: for 18 complexes, more than 100 correct structures (out of 8000 generated) were obtained.

In general, the Added predictor that takes all predictions from both WHISCY and ProMate performed best; it correctly predicts a larger number of residues than the other methods described here. This suggests that the currently used cutoffs might be too conservative. In the future, improved translation of prediction scores into restraints may further enhance the docking. Our results provide a promising starting point for routine incorporation of interface prediction in data-driven docking.

Software Availability

The WHISCY source code is freely available from the authors upon request. Alternatively, WHISCY predictions can be made through our web server at http://www.nmr.chem.uu.nl/whiscy.

REFERENCES

1. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A. A structural perspective on protein–protein interactions. Curr Opin Struct Biol 2004;14:313–324.
2. 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.
3. Janin J. The targets of CAPRI rounds 3–5. Proteins 2005;60:170–175.
4. 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.

5. Dominguez C, Boelens R, Bonvin AM. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. J Am Chem Soc 2003;125:1731–1737.
6. van Dijk AD, de Vries SJ, Dominguez C, Chen H, Zhou HX, Bonvin AM. Data-driven docking: HADDOCK's adventures in CAPRI. Proteins 2005;60:232–238.
7. van Dijk ADJ, Boelens R, Bonvin AMJJ. Data-driven docking for the study of biomolecular complexes. FEBS J 2005;272:293–312.
8. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press; 1965. p 97–166.
9. Zhu S, Tytgat J. Evolutionary epitopes of Hsp90 and p23: implications for their interaction. FASEB J 2004;18:940–947.
10. Lichtarge O, Bourne HR, Cohen FE. Evolutionarily conserved G(alpha beta gamma) binding surfaces support a model of the G protein–receptor complex. Proc Natl Acad Sci USA 1996;93:7507–7511.
11. Pazos F, HelmerCitterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511–523.
12. Carettoni D, Gomez-Puertas P, Yim L, Mingorance J, Massidda O, Vicente M, Valencia A, Domenici E, Anderluzzi D. Phage-display and correlated mutations identify an essential region of subdomain 1C involved in homodimerization of *Escherichia coli* FtsA. Proteins Struct Funct Genet 2003;50:192–206.
13. Duan Y, Reddy BV, Kaznessis YN. Physicochemical and residue conservation calculations to improve the ranking of protein–protein docking solutions. Protein Sci 2005;14:316–328.
14. Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol 2004;338:181–199.
15. Gottschalk KE, Neuvirth H, Schreiber G. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. Protein Eng Des Sel 2004;17:183–189.
16. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
17. Aloy P, Querol E, Aviles FX, Sternberg MJ. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 2001;311:395–408.
18. Armon A, Graur D, Ben Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
19. Pupko T, Bell RE, Mayrose I, Glaser F, Ben Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 2002;18(Suppl 1):S71–S77.
20. Landgraf R, Xenarios I, Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 2001;307:1487–1502.
21. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
22. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001;44:336–343.
23. Bordner AJ, Abagyan R. Statistical analysis and prediction of protein–protein interfaces. Proteins 2005;60:353–366.
24. Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics 2005;21:1487–1494.
25. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure, vol 5, suppl 3. Washington, DC: National Biomedical Research Foundation; 1978. p 345–352.
26. Felsenstein J. PHYLIP—phylogeny inference package. Cladistics 1989;5:164—166.
27. Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol 1997;272:133–143.
28. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
29. Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. Protein Sci 2004;13:1043–1055.
30. Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. Proteins 2003;52:88–91.
31. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z. Protein–protein docking benchmark 2.0: an update. Proteins 2005;60:214–216.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
33. Hubbard SJ, Thornton JM. NACCESS: Department of Biochemistry and Molecular Biology, University College London; 1993.
34. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. Protein Eng 1995;8:127–134.
35. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein–protein interaction sites from docking energy landscapes. J Mol Biol 2004;335:843–865.
36. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–577.
37. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. Protein Sci 2005;14:278–283.
38. Bradford JR, Westhead DR. Asymmetric mutation rates at enzyme–inhibitor interfaces: implications for the protein–protein docking problem. Protein Sci 2003;12:2099–2103.
39. MINT Database of protein–protein interactions (http://mint.bio.uniroma2.it/mint), December 2004.
40. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. Eur J Biochem 2002;269:1356–1361.