PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein–protein docking benchmark version 4.0

Howook Hwang,[1] Thom Vreven,[1] Joël Janin,[2] and Zhiping Weng[1]*

[1] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605

[2] Yeast Structural Genomics, IBBMC Université Paris-Sud, CNRS UMR 8619, 91405-Orsay, France

ABSTRACT

We updated our protein–protein docking benchmark to include complexes that became available since our previous release. As before, we only considered high-resolution complex structures that are nonredundant at the family–family pair level, for which the X-ray or NMR unbound structures of the constituent proteins are also available. Benchmark 4.0 adds 52 new complexes to the 124 cases of Benchmark 3.0, representing an increase of 42%. Thus, benchmark 4.0 provides 176 unbound–unbound cases that can be used for protein–protein docking method development and assessment. Seventeen of the newly added cases are enzyme-inhibitor complexes, and we found no new antigen-antibody complexes. Classifying the new cases according to expected difficulty for protein–protein docking algorithms gives 33 rigid body cases, 11 cases of medium difficulty, and 8 cases that are difficult. Benchmark 4.0 listings and processed structure files are publicly accessible at http://zlab.umassmed.edu/benchmark/

Proteins 2010; 78:3111–3114.

Key words: protein–protein docking; protein complexes; protein–protein interactions, complex structure.

INTRODUCTION

During the last decade, the computational protein–protein docking field has advanced considerably. In part, this is due to the efforts of making algorithms available to the community through web servers and/or downloadable packages,[1–8] the community-wide CAPRI experiment,[9] and the development of publically available benchmarks of protein–protein complexes.[10,11]

A protein–protein docking benchmark provides the community with a set of non-redundant protein–protein complexes for which the complex structure and the constituent unbound structures are available. A benchmark forms a subset of the Protein Data Bank (PDB)[12] and provides a standard dataset that can be used for systematic comparison of docking algorithms. Quantity and diversity of interactions covered in a benchmark can be improved by tracking updates in PDB.

Eight years ago, we introduced the first protein–protein docking benchmark,[10] and we updated twice in 2005 (Benchmark 2.0) and 2008 (Benchmark 3.0).[13,14] Recently, Kastritis and Bonvin collected experimentally measured protein–protein binding affinities (K_ds) of 81 test cases in Benchmark 3.0.[15] Since the last release, the number of entries in the PDB has increased by more than 13,000. This enables us to release a new update to the Benchmark.

MATERIALS AND METHODS

Data collection

We collected candidate structures from the PDB in a semiautomatic way with the same resolution cutoffs for X-ray structures (3.25 Å) and chain length (minimum of 30 residues) as described earlier.[10,13,14] Unlike the previous release, we now also consider structures determined with nuclear magnetic resonance (NMR) for the unbound forms of the proteins. We still excluded NMR structures for complexes to preclude the possibility that they were generated with aid of docking algorithms. We used the biological assembly information from the PDB to distinguish crystal contacts from biological complexes. This initial pass yielded 47,767 unbound structures and 8654 complex structures that represent hetero complexes of at least two interacting chains. The unbound forms of both binding partners were available for 1667 complex structures, and we used the Structural Classification of Proteins (SCOP)[16] database (version 1.75) to check this set for redundancy

Additional Supporting Information may be found in the online version of this article.
The authors state no conflict of interest.
Grant sponsor: NIH; Grant number: R01 GM084884
*Correspondence to: Zhiping Weng, Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Room 1010, Lazare Research Building, 364 Plantation St, Worcester, MA 01605.
E-mail: zhiping.Weng@umassmed.edu
Received 20 May 2010; Revised 29 June 2010; Accepted 2 July 2010
Published online 23 July 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.22830

Table I
New Cases in the Protein–Protein Docking Benchmark 4.0

Complex	Cat.[a]	PDB ID 1	Protein 1	PDB ID 2[b]	Protein 2	RMSD (Å)	DASA (Å^2)[c]
Rigid body (33)							
1CLV_A:I	E	1JAE_A	α-Amylase	1QFD_A(1)	α-Amylase inhibitor	0.86	2086
1FLE_E:I	E	9EST_A	Elastase	2REL_A(4)	Elafin	1.02	1762
1GL1_A:I	E	1K2I_1	α-Chymotrypsin	1PMC_A(6)	Protease inhibitor LCMI II	1.21	1590
1GXD_A:C	E	1CK7_A	proMMP2 type IV collagenase	1BR9_A	Metalloproteinase inhibitor 2	1.39	2445
1JTG_B:A	E	3GMU_B	β-Lactamase inhibitory protein	1ZG4_A	β-lactamase TEM-1	0.49	2599
1OC0_A:B	E	1B3K_A	Plasminogen activator inhibitor-1	2JQ8_A(4)	Vitronectin Somatomedin B domain	1	1312
1OYV_A:I	E	1SCD_A	Subtilisin Carlsberg	1PJU_A	Two-headed tomato inhibitor-II	0.7	1929
1OYV_B:I	E	1SCD_A	Subtilisin Carlsberg	1PJU_A	Two-headed tomato inhibitor-II	0.5	1279
2ABZ_B:E	E	3I1U_A	Carboxypeptidase A1	1ZFI_A(1)	Leech carboxypeptidase inhibitor	0.9	1443
2J0T_A:D	E	966C_A	MMP1 Interstitial collagenase	1D2B_A(20)	Metalloproteinase inhibitor 1	1.23	1476
2OUL_A:B	E	3BPF_A	Falcipain 2	2NNR_A	Chagasin	0.53	1932
3SGQ_E:I	E	2QA9_E	Streptogrisin B	2OVO_A	Ovomucoid inhibitor third domain	0.39	1210
1FCC_AB:C	O	1FC1_AB	Fc domain of IgG1 MO6	2IGG_A(3)	Strep. protein G C2 fragment	0.93	1354
1FFW_A:B	O	3CHY_A	Chemotaxis protein CheY	1FWP_A	Chemotaxis protein CheA	1.43	1166
1H9D_A:B	O	1EAN_A	Runx1 domain of CBFα1	1ILF_A(1)	Dimerization domain of CBF-β	1.32	2121
1HCF_AB:X	O	1B98_AM	Neurotrophin-4	1WWB_X	TrkB-d5 growth factors receptor	0.88	2135
1JWH_CD:A	O	3EED_AB	Casein kinase II β chain	3C13_A	Casein kinase II α chain	1.27	1451
1OFU_XY:A	O	1OFT_AB	SulA (PA3008)	2VAW_A	Cell division protein FtsZ	1.1	1583
1PVH_A:B	O	1BQU_A	IL6 receptor βchain D2-D3 domains	1EMR_A	Leukemia inhibitory factor	0.34	1403
1RV6_VW:X	O	1FZV_AB	PIGF receptor binding domain	1QSZ_A	Flt1 protein domain 2	1.09	1625
1US7_A:B	O	2FXS_A	Heat shock protein 82 N-ter domain	2W0G_A	HSP 90 co-chaperone CDC 37 C-ter domain	1.06	1095
1WDW_BD:A	O	1V8Z_AB	Tryptophan synthase β chain 1	1GEQ_A	Tryptophan synthase α chain	1.29	3147
1XU1_ABD:T	O	1U5Y_ABD	TNF domain of APRIL	1XUT_A(11)	TNF receptor superfamily member 13B TACI CRD2 domain	1.3	1696
1ZHH_A:B	O	1JX6_A	Autoinducer 2-binding periplasmic protein LuxP	2HJE_A	Autoinducer 2 sensor kinase/ phosphatase LuxQ	1.31	2189
2A5T_A:B	O	1Y20_A	NMDA receptor R1–4A subunit ligand-binding core	2A5S_A	NMDA receptor R2A subunit ligand-binding core	1.28	1892
2A9K_A:B	O	1U90_A	Ras-related protein Ral-A	2C8B_X	Mono-ADP-ribosyltransferase C3	0.85	1750
2B4J_AB:C	O	1BIZ_AB	Integrase (HIV-1)	1Z9E_A(1)	PC4 and SFRS1 interacting protein	0.99	1273
2FJU_B:A	O	2ZKM_X	Phospholipase β 2	1MH1_A	Rac GTPase	1.04	1245
2G77_A:B	O	1FKM_A	GTPase-activating protein Gyp1	1Z06_A	Ras-related protein Rab-33B	1.75	2524
2OOR_AB:C	O	1L7E_AB	NAD(P) transhydrogenase subunit α part 1	1E3T_A	NAD(P) transhydrogenase subunit β	1.42	2065
2VDB_A:B	O	3CX9_A	Serum albumin	2J5Y_A	Peptostreptococcal albumin-binding protein GA module	0.47	1797
3BP8_AB:C	O	1Z6R_AB	Mlc transcription regulator	3BP3_A	PTS glucose-specific enzyme EIICB	0.45	1390
3D5S_A:C	O	1C3D_A	Complement C3d fragment	2GOM_A	Fibrinogen-binding protein C-ter domain	0.56	1620
Medium Difficult (11)							
1JIW_P:I	E	1AKL_A	Alkaline metalloproteinase	2RN4_A(1)	Proteinase inhibitor	2.07	1997
4CPA_A:I	E	8CPA_A	Carboxypeptidase A	1H20_A(9)	Potato carboxypeptidase inhibitor	1.97	1175
1LFD_B:A	O	5P21_A	Ras	1LXD_A	RalGDS Ras-interacting domain	1.79	1167
1MQ8_A:B	O	1IAM_A	ICAM-1 domains 1–2	1MQ9_A	Integrin α-L I domain	1.76	1252
1R6Q_A:C	O	1R6C_X	Clp protease subunit ClpA	2W9R_A	Clp protease adaptor protein ClpS	1.67	1651
1SYX_A:B	O	1QGV_A	Spliceosomal U5 15 kDa protein	1L2Z_A(1)	CD2 receptor binding protein 2 C-ter fragment	1.64	1292
2AYO_A:B	O	2AYN_A	Ubiquitin carboxyl-terminal hydrolase 14	2FCN_A	Ubiquitin	1.62	3026
2J7P_A:D	O	1NG1_A	SRP GTPase Ffh	2IYL_D	Cell division protein FtsY	1.93	3008
2OZA_B:A	O	3HEC_A	MAP kinase 14	3FYK_X	MAP kinase-activated protein kinase 2	1.89	6247
2Z0E_A:B	O	2D1I_A	Cysteine protease Atg4B	1V49_A(1)	Microtubule-associated proteins 1A/1B light chain 3B	2.15	2477
3CPH_G:A	O	3CPI_G	Ras-related protein Sec4	1G16_A	Rab GDP-dissociation inhibitor	2.12	1684
Difficult (8)							
1F6M_A:C	E	1CL0_A	Thioredoxin reductase	2TIR_A	Thioredoxin 1	4.9	1821
1ZLI_A:B	E	1KWM_A	Carboxypeptidase B	2JTO_A(6)	Tick carboxypeptidase inhibitor	2.53	2083
2O3B_A:B	E	1ZM8_A	NucA nuclease	1J57_A	NuiA nuclease inhibitor	3.13	1675
1JK9_B:A	O	1QUP_A	CCS metallochaperone	2JCW_A	SOD1 superoxide dismutase	4.87	2130
1JZD_AB:C	O	1JZO_AB	DsbC disulfide bond isomerase	1JPE_A	DsbD disulfide bond isomerase	2.71	2026
1ZM4_A:B	O	1N0V_C	Elongation factor 2	1XK9_A	Diphtheria toxin A catalytic domain	2.94	1554
2I9B_E:A	O	1YWH_A	Urokinase plasminogen activator surface receptor	2I9A_A	Urokinase-type plasminogen activator	3.79	2370
2IDO_A:B	O	1J54_A	DNA polymerase III ϵ exonuclease domain	1SE7_A(1)	HOT protein (P1 phage)	2.79	1953

[a]Complex category labels: E = Enzyme/Inhibitor or Enzyme/Substrate, O = Other.
[b]NMR model numbers from are shown in parenthesis.
[c]Change in accessible surface area (ΔASA) upon complex formation, defined as the ASA of Protein 1 plus the ASA of Protein 2 minus the ASA of the Complex. ASA is calculated using NACCESS.

Table II
Statistics of the Three Classes of Difficulty in the Entire Benchmark 4.0 and the New Cases (in Parentheses)

	I-RMSD	f_{nat}	$f_{non\text{-}nat}$	Number
Rigid body	0.90 (1.12)	0.79 (0.80)	0.21 (0.19)	121 (33)
Medium	1.76 (1.86)	0.63 (0.66)	0.35 (0.27)	30 (11)
Difficult	3.76 (3.45)	0.51 (0.60)	0.51 (0.41)	25 (8)

at the family level. Two complexes were deemed redundant if both proteins in one complex were in the same SCOP families as the two proteins in the other complex, respectively. This yielded 109 complexes that were non-redundant with the complexes in the previous release of the Benchmark and amongst themselves. (PDB entries without SCOP unique identifier sunid[17] were excluded from the bound candidate list to remove possible redundancy.) Finally, we used literature information to eliminate obligate complexes,[18] which further reduced the list to 52 complexes.

When we found multiple candidates for an unbound structure, we selected one structure based on a combination of several considerations: highest sequence similarity with the bound structure, highest resolution, and lowest number of missing residues in protein–protein interface area. For an ensemble of multiple candidate entries for NMR structures, we selected the model that had the lowest interface root-mean-square distance (RMSD) (I-RMSD; defined below) with the bound form. The final structure files that are on the benchmark website include cofactors that were present in the original PDB files, and in the case of an NMR structure, all the models that were provided in the original file.

Classification

As done for the previous releases of the Benchmark, we classify the new entries, according to expected difficulty for protein–protein docking algorithms, based on the structural difference between the bound and the unbound forms of the binding partners:[14]

Rigid body:

I-RMSD $\leq$ 1.5 Å and $f_{\text{non-nat}} \leq 0.4$

Medium difficulty:

[1.5 Å < I-RMSD $\leq$ 2.2 Å] or [I-RMSD $\leq$ 1.5 Å and $f_{\text{non-nat}} > 0.4$]

Difficult:

I-RMSD > 2.2 Å

We define I-RMSD as the RMSD between the unbound and the bound structures, superposed onto each other, calculated using the Cα atoms of the interface residues of both binding partners. In line with Mendez *et al.*,[19] f_{nat} and $f_{\text{non-nat}}$ are the fractions of native residue contacts and non-native residue contacts, respectively, of the superposed unbound structures.

RESULTS AND DISCUSSION

The 52 new cases are listed in Table 1. The entire updated Benchmark is reported in Supporting Information Table S1. 1OYV is a 1:2 complex of a two-headed inhibitor and subtilisin.[20] We split this complex into two cases for the Benchmark that represent the interaction between chain A of subtilisin and chain I (inhibitor) and the interaction between chain B of subtilisin and chain I, respectively. In addition to the aforementioned properties, the tables also report the change in accessible surface area (ASA) on complexation, which is a measure for the size of the interface between the binding partners.

Benchmark 4.0 includes 121 rigid body cases (33 new), 30 cases of medium difficulty (11 new), and 25 difficult cases (eight new). According to biochemical function, we have 52 enzyme-inhibitor (17 new), 25 antibody–antigen, and 99 complexes with other function (35 new). We did not find new antibody–antigen complexes. In this update of the Benchmark, we included 16 cases that involve NMR unbound structures. Among them, 11 cases are classified as rigid body, four cases of medium difficulty, and one case as difficult. Thus, the expected difficulty for docking algorithms using NMR structures in the benchmark is similar to the expected difficulty using X-ray structures. If we would consider NMR structures for the bound complexes, we would have included seven more cases (1GGR, 1J6T, 1O2F, 1P9D, 1UR6, 2ODG, and 3EZA). Although one can argue that exclusion of complex NMR structures from the Benchmark should be decided on a case-by-case basis, we decided to simply leave all out as inclusion would only lead to a small increase of the Benchmark.

Table 2 summarizes the average I-RMSD, f_{nat} and $f_{\text{non-nat}}$ for the different classes of docking difficulty. The numbers in Table 2 indicate that the new cases in Benchmark 4.0 (in parentheses) have generally higher I-RMSD for rigid body cases and cases of medium difficulty, which predicts the new test cases to be more challenging for computational docking. Also, the fraction of rigid body cases in the new cases is 0.63, somewhat lower than the 0.71 in Benchmark 3.0. Thus, the new cases are expected to be more difficult for protein–protein docking algorithms, and this must be taken into account when assessing docking algorithms, as performance will depend on the benchmark version utilized.

In summary, Benchmark 4.0 includes 52 new cases and a higher number of new rigid body and medium difficulty cases show larger conformational changes upon binding than cases in the previous release. This is especially useful for the development of protein–protein docking algorithms that incorporate protein flexibility, a problem that has recently received much attention but still remains a major challenge.[21]

REFERENCES

1. Vakser IA. Protein docking for low-resolution structures. Protein Eng 1995;8:371–377.

2. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004;20:45–50.
3. Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF. Protein docking using continuum electrostatics and geometric fit. Protein Eng 2001;14:105–113.
4. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52:80–87.
5. Ritchie DW, Kozakov D, Vajda S. Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. Bioinformatics 2008;24:1865–1873.
6. 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.
7. de Vries SJ, van Dijk M, Bonvin AM. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 5:883–897.
8. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res 2008;36(Web Server issue): W233–W238.
9. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ. CAPRI: a critical assessment of predicted interactions. Proteins 2003;52:2–9.
10. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91.
11. Gao Y, Douguet D, Tovchigrechko A, Vakser IA. DOCKGROUND system of databases for protein recognition studies: unbound structures for docking. Proteins 2007;69:845–851.
12. 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.
13. 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.
14. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. Proteins 2008;73:705–709.
15. Kastritis PL, Bonvin AM. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 2010;9:2216–2225.
16. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
17. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res 2002;30:264–267.
18. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci USA 2005;102:10930–10935.
19. Mendez R, Leplae R, De Maria L, Wodak SJ. Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 2003;52:51–67.
20. Barrette-Ng IH, Ng KK, Cherney MM, Pearce G, Ryan CA, James MN. Structural basis of inhibition revealed by a 1:2 complex of the two-headed tomato inhibitor-II and subtilisin Carlsberg. J Biol Chem 2003;278:24062–24071.
21. Zacharias M. Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 2010;20:180–186.