PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Protein–protein docking benchmark version 3.0

Howook Hwang,[1] Brian Pierce,[1] Julian Mintseris,[2] Joël Janin,[3] and Zhiping Weng[1,4,5]*

[1] Bioinformatics Program, Boston University, Boston, Massachusetts 02215
[2] Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115
[3] Yeast Structural Genomics, IBBMC Université Paris-Sud, CNRS UMR 8619, 91405-Orsay, France
[4] Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215
[5] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605

ABSTRACT

We present version 3.0 of our publicly available protein–protein docking benchmark. This update includes 40 new test cases, representing a 48% increase from Benchmark 2.0. For all of the new cases, the crystal structures of both binding partners are available. As with Benchmark 2.0, Structural Classification of Proteins (Murzin *et al.*, J Mol Biol 1995;247:536–540) was used to remove redundant test cases. The 124 unbound-unbound test cases in Benchmark 3.0 are classified into 88 rigid-body cases, 19 medium-difficulty cases, and 17 difficult cases, based on the degree of conformational change at the interface upon complex formation. In addition to providing the community with more test cases for evaluating docking methods, the expansion of Benchmark 3.0 will facilitate the development of new algorithms that require a large number of training examples. Benchmark 3.0 is available to the public at http://zlab.bu.edu/benchmark.

Proteins 2008; 73:705–709.

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

INTRODUCTION

In 2003 and 2005, we published two versions of a protein–protein docking benchmark.[2,3] It contains structures of proteins for which high-resolution crystal structures are available in both the unbound and bound states. Our goal is to provide a wide variety of test cases, so that the protein docking community can evaluate the progress of docking methods. Our benchmark, in its previous two editions,[2,3] has been widely used for training and testing protein docking algorithms,[4–9] developing reranking algorithms,[10] formulating energy functions,[11] and performing protein structure analysis.[12]

Since 2005, the number of protein structures in the Protein Data Bank[13] (PDB) has increased by more than 10,000, which allowed us to update the Benchmark to version 3.0. Although manual curation of the data during some steps of the benchmark construction was inevitable, we have constructed a semiautomated process to ensure that this update covers all available test cases in the PDB. The new test cases are exclusively unbound–unbound, in that three crystal structures are available, for the complex and each of the unbound proteins.

SEMIAUTOMATED DATASET RETRIEVAL AND CURATION

To collect unbound–unbound benchmark cases, we parsed all PDB entries as described previously.[2,3] We first identified multiprotein X-ray structures with individual sequence length longer than 30 amino acids and resolution better than 3.25 Å; these two cutoffs were used in the two previous editions of the benchmark. The biological unit information provided by the PDB was used to differentiate biologically relevant interactions from crystal contacts. We filtered out obligate complexes manually, after consulting the literature.

For the remaining protein complexes, we utilized Structural Classification of Proteins (SCOP)[1] to examine protein family–family pair redundancy within the new cases and against the existing cases from Benchmark 2.0. In addition to the latest version of SCOP (1.71), which was released in Oct. 2006, we used its preclassification version, Pre-SCOP (http://www.mrc-lmb.cam.ac.uk/agm/pre-scop/), for the structures deposited in PDB since the SCOP 1.71 release. Nonredundancy was set at the family level of

Grant sponsor: NSF; Grant numbers: DBI-0078194, DBI-0133834, DBI-0116574
*Correspondence to: Zhiping Weng, Program in Bioinformatics and Integrative Biology, The University of Massachusetts Medical School, Room 1010, Lazare Research Building, 364 Plantation St., Worcester, MA 01605.
E-mail: zhiping.weng@umassmed.edu
Received 27 November 2007; Revised 31 March 2008; Accepted 8 April 2008
Published online 19 May 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22106

Table I
Protein–Protein Docking Benchmark 3.0

Complex	Cat.[a]	PDB ID 1	Protein 1	PDB ID 2	Protein 2	RMSD[b] (Å)	ΔASA[c] (Å^2)
Rigid-body (88)							
1AVX_A:B	E	1QQU_A	Porcine trypsin	1BA7_B	Soybean trypsin inhibitor	0.47	1585
1AY7_A:B	E	1RGH_B	Barnase	1A19_B	Barstar	0.54	1237
1BVN_P:T	E	1PIG_	α-Amylase	1HOE_	Tendamistat	0.87	2222
1CGI_E:I	E	2CGA_B	Bovine chymotrypsinogen	1HPT_	PSTI	2.02	2053
1D6R_A:I	E	2TGT_	Bovine trypsin	1K9B_A	Bowman-Birk inhibitor	1.14	1408
1DFJ_E:I	E	9RSA_B	Ribonuclease A	2BNH_	Rnase inhibitor	1.02	2582
1E6E_A:B	E	1E1N_A	Adrenoxin reductase	1CJE_D	Adrenoxin	1.33	2315
1EAW_A:B	E	1EAX_A	Matriptase	9PTI_	BPTI	0.54	1866
1EWY_A:C	E	1GJR_A	Ferredoxin reductase	1CZP_A	Ferredoxin	0.8	1502
1EZU_C:AB*	E	1TRM_A	D102N Trypsin	1ECZ_AB	Ecotin	1.21	2751
1F34_A:B	E	4PEP_	Porcine pepsin	1F32_A	Ascaris inhibitor 3	0.93	3038
1HIA_AB:I	E	2PKA_XY	Kallikrein	1BX8_	Hirustatin	1.4	1737
1MAH_A:F	E	1J06_B	Acetylcholinesterase	1FSC_	Fasciculin	0.61	2145
1N8O_ABC:E*	E	8GCH_A	Chymotrypsin	1IFG_A	Ecotin	0.94	1851
1OPH_A:B	E	1Q1P_A	α-1-Antitrypsin	1UTQ_A	Trypsinogen	1.21	1360
1PPE_E:I	E	1BTP_	Bovine trypsin	1LU0_A	CMTI-1 squash inhibitor	0.44	1688
1R0R_E:I	E	1SCN_E	Subtilisin Carlsberg	2GKR_I	OMTKY	0.45	1409
1TMQ_A:B	E	1JAE_	α-Amylase	1B1U_A	RAGI inhibitor	0.86	2401
1UDI_E:I	E	1UDH_	Uracyl-DNA glycosylase	2UGI_B	Glycosylase inhibitor	0.9	2022
1YVB_A:I	E	2GHU_A	Falcipain 2	1CEW_I	Cystatin	0.51	1743
2B42_A:B	E	2DCY_A	Xylanase	1T6E_X	Xylanase inhibitor	0.72	2520
2MTA_HL:A	E	2BBK_JM	Methylamine dehydrogenase	2RAC_A	Amicyanin	0.41	1461
2O8V_A:B	E	1SUR_	PAPS reductase	2TRX_A	Thioredoxin	1.37	1619
2PCC_A:B	E	1CCP_	Cyt C peroxidase	1YCC_	Cytochrome C	0.39	1141
2SIC_E:I	E	1SUP_	Subtilisin	3SSI_	Streptomyces subtilisin inhibitor	0.36	1617
2SNI_E:I	E	1UBN_A	Subtilisin	2CI2_I	Chymotrypsin inhibitor 2	0.35	1628
2UUY_A:B	E	1HJ9_A	Trypsin	2UUX_A	Tryptase inhibitor from tick	0.43	1280
7CEI_A:B	E	1UNK_D	Colicin E7 nuclease	1M08_B	Im7 immunity protein	0.7	1384
1A2K_C:AB	O	1QG4_A	Ran GTPase	1OUN_AB	Nuclear transport factor 2	1.11	1603
1AK4_A:D	O	2CPL_	Cyclophilin	1E6J_P	HIV capsid	1.33	1029
1AKJ_AB:DE	O	2CLR_DE	MHC Class 1 HLA-A2	1CD8_AB	T-cell CD8 coreceptor	1.14	1995
1AZS_AB:C	O	1AB8_AB	Adenylyl cyclase	1AZT_A	AC activator Gs α complex	0.72	1911
1B6C_A:B	O	1D6O_A	FKBP binding protein	1IAS_A	TGFβ receptor	1.96	1752
1BUH_A:B	O	1HCL_	CDK2 kinase	1DKS_A	Ckshs1	0.75	1324
1E96_A:B	O	1MH1_	Rac GTPase	1HH8_A	p67 Phox	0.71	1179
1EFN_B:A	O	1AVV_A	HIV-1-NEF protein	1G83_A	SH3 domain	0.77	1254
1F51_AB:E	O	1IXM_AB	Sporulation response factor B	1SRR_C	Sporulation response factor F	0.74	2407
1FC2_C:D	O	1BDD_	Staphylococcus Protein A	1FC1_AB	Human Fc fragment	1.69	1307
1FQJ_A:B	O	1TND_C	Gt-α	1FQI_A	RGS9	0.91	1806
1GCQ_B:C	O	1GRI_B	GRB2 C-ter SH3 domain	1GCP_B	GRB2 N-ter SH3 domain	0.92	1208
1GHQ_A:B	O	1C3D_	Complement C3	1LY2_A	Epstein-Barr virus receptor CR2	0.34	800
1GLA_G:F	O	1BU6_0	Glycerol kinase	1F3Z_A	Glucose specific phosphocarrier	0.98	1304
1GPW_A:B	O	1THF_D	HISF protein	1K9V_F	Amidotransferase HISH	0.65	2097
1HE1_C:A	O	1MH1_	Rac GTPase	1HE9_A	Pseudomonas toxin GAP dom.	0.93	2113
1I4D_D:AB	O	1MH1_	Rac GTPase	1I49_AB	Arfaptin	1.41	1657
1J2J_A:B	O	1O3Y_A	Arf1 GTPase	1OXZ_A	GAT domain of GGA1	0.63	1209
1K74_AB:DE	O	1MZN_AB	RXR-α	1ZGY_AB	PPAR-γ	0.8	2200
1KAC_A:B	O	1NOB_F	Adenovirus fiber knob protein	1F5W_B	Adenovirus receptor	0.95	1456
1KLU_AB:D	O	1H15_AB	MHC class 2 HLA-DR1	1STE_	Staphylococcus enterotoxin C3	0.43	1254
1KTZ_A:B	O	1TGK_	TGF-β	1M9Z_A	TGF-β receptor	0.39	989
1KXP_A:D	O	1IJJ_B	Actin	1KW2_B	Vitamin D binding protein	1.12	3341
1ML0_AB:D	O	1MKF_AB	Viral chemokine binding p. M3	1DOL_	Chemokine Mcp1	1.02	2069
1QA9_A:B	O	1HNF_	CD2	1CCZ_A	CD58	0.73	1353
1RLB_ABCD:E	O	2PAB_ABCD	Transthyretin	1HBP_	Retinol binding protein	0.66	1439
1S1Q_A:B	O	2F0R_A	UEV domain	1YJ1_A	Ubiquitin	0.98	1288
1SBB_A:B	O	1BEC_	T-cell receptor β	1SE4_	Staphylococcus enterotoxin B	0.37	1064
1T6B_X:Y	O	1ACC_	Anthrax protective antigen	1SHU_X	Anthrax toxin receptor	0.62	1948
1XD3_A:B	O	1UCH	UCH-L3	1YJ1_A	Ubiquitin	1.24	2281
1Z0K_A:B	O	2BME_A	Rab4A GTPase	1YZM_A	RAB4 binding domain of Rabenosyn	0.53	1787
1Z5Y_D:E	O	1L6P	N-term of DsbD	2B1K_A	E.coli CCMG protein	1.23	1346
1ZHI_A:B	O	1M4Z_A	BAH domain of Orc1	1Z1A_A	Sir Orc-interaction domain	0.68	1322

(Continued)

Table I
(Continued)

Complex	Cat.[a]	PDB ID 1	Protein 1	PDB ID 2	Protein 2	RMSD[b] (Å)	ΔASA[c] (Å^2)
2AJF_A:E	0	1R42_A	ACE2	2GHV_E	SARS spike protein receptor binding domain	0.65	1704
2BTF_A:P	0	1IJJ_B	Actin	1PNE_	Profilin	0.75	2063
2HLE_A:B	0	2BBA_A	Ephrin B4 receptor	1IKO_P	Ephrin B2 ectodomain	1.4	2116
2HQS_A:H	0	1CRZ_A	TolB	1OAP_A	Pal	1.14	2333
2OOB_A:B	0	2OOA_A	Ubiquitin ligase	1YJ1_A	Ubiquitin	0.85	808
1AHW_AB:C	A	1FGN_LH	Fab 5g9	1TFH_A	Tissue factor	0.69	1899
1BVK_DE:F	A	1BVL_BA	Fv Hulys11	3LZT_	HEW lysozyme	1.24	1321
1DQJ_AB:C	A	1DQQ_CD	Fab Hyhel63	3LZT_	HEW lysozyme	0.75	1765
1E6J_HL:P	A	1E6O_HL	Fab	1A43_	HIV-1 capsid protein p24	1.05	1245
1JPS_HL:T	A	1JPT_HL	Fab D3H44	1TFH_B	Tissue factor	0.51	1852
1MLC_AB:E	A	1MLB_AB	Fab44.1	3LZT_	HEW lysozyme	0.6	1392
1VFB_AB:C	A	1VFA_AB	Fv D1.3	8LYZ_	HEW lysozyme	1.02	1383
1WEJ_HL:F	A	1QBL_HL	Fab E8	1HRC_	Cytochrome C	0.31	1177
2FD6_HL:U	A	2FAT_HL	Plasminogen receptor antibody	1YWH_A	Plasminogen activator receptor	1.07	1139
2i25_N:L	A	2I24_N	Shark single domain antigen receptor	3LZT	Lysozyme	1.21	1425
2VIS_AB:C	A	1GIG_LH	Fab	2VIU_ACE	Flu virus hemagglutinin	0.8	1296
1BJ1_HL:VW	AB	1BJ1_HL	Fab	2VPF_GH	vEGF	0.5	1731
1FSK_BC:A	AB	1FSK_BC	Fab	1BV1_	Birch pollen antigen Bet V1	0.45	1623
1I9R_HL:ABC	AB	1I9R_HL	Fab	1ALY_ABC	Cd40 ligand	1.3	1498
1IQD_AB:C	AB	1IQD_AB	Fab	1D7P_M	Factor VIII domain C2	0.48	1976
1K4C_AB:C	AB	1K4C_AB	Fab	1JVM_ABCD	Potassium Channel Kcsa	0.53	1601
1KXQ_H:A	AB	1KXQ_H	Camel VHH	1PPI_	Pancreatic α-amylase	0.72	2172
1NCA_HL:N	AB	1NCA_HL	Fab	7NN9_	Flu virus neuraminidase N9	0.24	1953
1NSN_HL:S	AB	1NSN_HL	Fab N10	1KDC_	Staphylococcal nuclease	0.35	1776
1QFW_HL:AB	AB	1QFW_HL	Fv	1HRP_AB	Human chorionic gonadotropin	1.31	1580
1QFW_IM:AB	AB	1QFW_IM	Fv	1HRP_AB	Human chorionic gonadotropin	0.73	1637
2JEL_HL:P	AB	2JEL_HL	Fab Jel42	1POH_	HPr	0.17	1501
Medium difficulty (19)							
1ACB_E:I	E	2CGA_B	Chymotrypsin	1EGL_	Eglin C	2.26	1544
1IJK_A:BC	E	1AUQ_	Von Willebrand Factor dom. A1	1FVU_AB	Botrocetin	0.68	1648
1KKL_ABC:H	E	1JB1_ABC	HPr kinase C-ter domain	2HPR_	HPr	2.2	1641
1M10_A:B	E	1AUQ_	Von Willebrand Factor dom. A1	1M0Z_B	Glycoprotein IB-α	2.1	2097
1NW9_B:A	E	1JXQ_A	Capase-9	1IFG_A	Ecotin	1.97	2112
1GP2_A:BG	0	1GIA_	Gi-α	1TBG_DH	Gi-β,γ	1.65	2287
1GRN_A:B*	0	1A4R_A	CDC42 GTPase	1RGP_	CDC42 GAP	1.22	2332
1HE8_B:A	0	821P_	Ras GTPase	1E8Z_A	PIP3 kinase	0.92	1305
1I2M_A:B	0	1QG4_A	Ran GTPase	1A12_A	RCC1	2.12	2779
1IB1_AB:E	0	1QJB_AB	14-3-3 protein	1KUY_A	Serotonin N-*acteylase*	2.09	2808
1K5D_AB:C	0	1RRP_AB	Ran GTPase	1YRG_B	Ran GAP	1.19	2527
1N2C_ABCD:EF	0	3MIN_ABCD	Nitrogenase Mo-Fe protein	2NIP_AB	Nitrogenase Fe protein	2.13	3635
1WQ1_R:G*	0	6Q21_D	Ras GTPase	1WER_	Ras GAP	1.16	2913
1XQS_A:C	0	1XQR_A	HspBP1	1S3X_A	Hsp70 ATPase domain	1.77	2350
2CFH_A:C	0	1SZ7_A	BET3	2BJN_A	TPC6	1.55	2384
2H7V_A:C	0	1MH1_	Rac GTPase	2H7O_A	YpkA	1.63	1574
2HRK_A:B	0	2HRA_A	Glutamyl-t-RNA synthetase	2HQT_A	GU-4 nucleic binding protein	2.03	1595
2NZ8_A:B	0	1MH1_	Rac GTPase	1NTY_A	DH/PH domain of TRIO	2.13	2599
1BGX_HL:T	A	1AY1_HL	Fab	1CMW_A	Taq polymerase	1.48	5814
Difficult (17)							
1FQ1_A:B	E	1B39_A	CDK2 kinase	1FPZ_F	CDK inhibitor 3	3.41	1832
1PXV_A:C	E	1X9Y_A	Cystein protease	1NYC_A	Cystein protease inhibitor	2.63	2336
1ATN_A:D	0	1IJJ_B	Actin	3DNI_	Dnase I	3.28	1774
1BKD_R:S	0	1CTQ_A	Ras GTPase	2II0_B	Son of Sevenless	2.86	3163
1DE4_AB:CF	0	1A6Z_AB	β2-microglobulin	1CX8_AB	Transferrin receptor ectodom.	2.59	2066
1EER_A:BC	0	1BUY_A	Erythropoietin	1ERN_AB	EPO receptor	2.44	3347
1FAK_HL:T	0	1QFK_HL	Coagulation factor VIIa	1TFH_B	Soluble tissue factor	6.18	3363
1H1V_A:G	0	1IJJ_B	Actin	1D0N_B	Gelsolin	6.62	2071
1IBR_A:B	0	1QG4_A	Ran GTPase	1F59_A	Importin β	2.54	2270
1IRA_Y:X	0	1G0Y_R	Interleukin-1 receptor	1ILR_1	Interleukin-1 receptor antagonist protein	8.38	3367
1JMO_A:HL	0	1JMJ_A	Heparin cofactor	2CN0_HL	Thrombin	3.21	3461
1R8S_A:E	0	1HUR_A	Arf1 GTPase	1R8M_E	Sec 7 domain	3.73	2986

(Continued)

Table I
(Continued)

Complex	Cat.[a]	PDB ID 1	Protein 1	PDB ID 2	Protein 2	RMSD[b] (Å)	ΔASA[c] (Å^2)
1Y64_A:B	O	2FXU_A	Actin	1UX5_A	BNI1 protein	4.69	2745
2C0L_A:B	O	1FCH_A	PTS1 and TRP region of PEX5	1C44_A	SCP2	2.62	2013
2OT3_B:A	O	1YZU_A	Rab21 GTPase	1TXU_A	Rabex-5 VPS9 domain	2.79	2306
1E4K_AB:C	A	2DTQ_AB	FC fragment of human IgG 1	1FNL_A	Human FCGR III	2.59	1634
2HMI_CD:AB	AB	2HMI_CD	Fab 28	1S6P_AB	HIV1 reverse transcriptase	2.26	1234

[a]Complex category labels: E, enzyme/inhibitor or enzyme/substrate; A, antibody/antigen; O, others; AB, antigen/bound antibody.
[b]RMSD of Cα atoms of interface residues calculated as described previously[14] after finding the best superposition of bound and unbound interfaces.
[c]Change in accessible surface area 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.[15]
*The receptors and ligands of 1EZU and 1N8O belong to the same SCOP superfamily. Similarly, the receptors and ligands of 1GRN and 1WQ1 belong to the same SCOP superfamily.
The test cases identical or homologous to CAPRI targets are: 1KKL (Target 1 or T1; identical), 1KXQ (T6; identical), 1SBB (T7; homologous), 2B42 (T18; homologous), 1ZHI (T21; identical), 2HQS (T26; identical).
The new cases in Benchmark 3.0 are shaded.

SCOP, that is, no two test cases in Benchmark 3.0 are allowed to belong to the same family–family pair. The users who are interested in developing statistical potentials with our benchmark may also want to exclude test cases that belong to the same superfamily–superfamily pairs. This would affect two pairs of test cases: 1EZU/1N8O and 1GRN/1WQ1 (labeled with "*" in Table I). To avoid this level of redundancy, one test case from each of these pairs can be removed. We then eliminated the test cases for which the unbound structures had less than 96% sequence identity to the corresponding bound structures, as defined by BLAST.[16] For the remaining test cases with multiple crystal structures of the unbound proteins, we chose the unbound structure with the highest sequence similarity, highest structure resolution, and fewest missing residues. Finally, we discarded test cases that present unusual difficulties for docking algorithms, for example, three or more residues in the binding site were missing in the unbound structure, or the bound and the unbound structures have different cofactors at the binding site. The cofactors included in structures are listed in the table at the benchmark website (http://zlab.bu.edu/benchmark).

BENCHMARK TEST CASES AND CLASSIFICATION

There are a total of 40 new test cases. They are listed in Table I, along with the existing cases from Benchmark 2.0. Six of these test cases are identical or homologous to CAPRI targets, indicated in the legend of Table I. To assign difficulty levels of the test cases, we used the degree of conformational changes, as measured by Interface Cα-RMSD (I-RMSD[17]) and fraction of non-native residue contacts ($f_{\text{non-nat}}$[18]), of the unbound structures fitted onto the bound structures. Specifically, the rigid-body cases are cases with I-RMSD $\leq$ 1.5 Å and $f_{\text{non-nat}} \leq$ 0.4, the difficult cases are cases with I-RMSD $>$ 2.2 Å, and the medium cases are all remaining cases (i.e., with 1.5 Å $<$ I-RMSD $\leq$ 2.2 Å, or I-RMSD $<$ 1.5 Å and $f_{\text{non-nat}} > 0.4$). We used Cα-RMSD instead of backbone RMSD (the latter is used in the CAPRI evaluation), because we have been using Cα-RMSD since the creation of the Benchmark, which predates CAPRI.

We use this difficulty classification to quantify the extent of conformational change around the binding interface, which broadly affects most docking methods. For Benchmark 2.0, we assigned difficulty level based on the number of possible high-quality hit predictions (as measured by the CAPRI criteria[18]) attainable using rigid-body docking on a grid. To remove possible bias due to this method and to simplify the classification, we opted to utilize the I-RMSD and $f_{\text{non-nat}}$ metrics for the new cases, selecting cutoffs to maintain consistency among the new cases and those from Benchmark 2.0. Besides conformational changes, other factors such as the size and hydrophobic/electrostatic composition of the interface, as well as the available experimental data on the complex, can also affect the difficulty of a test case.[14,17]

In total, Benchmark 3.0 has 88 rigid cases, 19 medium cases, and 17 difficult cases. There are two difficult cases with large hinge movement (1E4K and 1IRA). Table II provides the average values of the three classes in terms of I-RMSD and $f_{\text{non-nat}}$. We have also included the statistics

Table II
Statistics on the Three Difficulty Groups in Benchmark 3.0

	I-RMSD[a]	f_{nat}[b]	$f_{\text{non-nat}}$[c]	Number
Rigid-body	0.83	0.79	0.22	88
Medium	1.70	0.62	0.41	19
Difficult	3.91	0.48	0.56	17

[a]I-RMSD of Cα atoms of interface residues (in Å) calculated as described previously[14] after finding the best superposition of bound and unbound interfaces.
[b,c]f_{nat}, the fraction of native residue contacts in a predicted complex and $f_{\text{non-nat}}$, the fraction of non-native residue contacts in a predicted complex, were calculated following Mendez et al.[18] with the predicted complex obtained by minimizing the interface RMSD between bound and unbound structures.

of the fraction of native residue contacts (f_{nat}[18]) even though it does not provide additional values to the above two metrics, because it is used in CAPRI evaluation.

In addition to the difficulty assessment, we have classified the new test cases into three biochemical categories: enzyme–inhibitor (E; 9 cases), antigen–antibody (A; 3 cases), and others (O; 28 cases), as with previous Benchmark versions.[2,3] This information is provided in Table I. We corrected the category assignments for two Benchmark 2.0 cases (1IJK and 1FQ1) from O to E.

COMPARISON WITH DOCKGROUND

DOCKGROUND is a relational database of X-ray and simulated protein–protein complexes. Its second release[19] contains 99 test cases for which the X-ray structures of the complex and the individual proteins are available. Among these, 30 cases are included in our Benchmark 3.0, based on the PDB IDs of the complexes. For an additional 20 cases, the unbound proteins fall within the same SCOP family pairs as test cases in our Benchmark 3.0. The remaining 40 cases were rejected by our annotation pipeline because of redundancy or complications at the interface (e.g., one or combinations of the following criterions: three or more missing contact residues at binding site, cofactors at the binding site of the complex structure but not in the unbound structure or vice versa, different numbers of protein chains at the interface between the bound and unbound states, or dimerization of receptor or ligand or both in the complex but no corresponding unbound structures). Note that antigen–antibody cases were kept although they have multiple chains in interface. One difference between our curated benchmark and automatically generated databases such as DOCKGROUND is that we provide the residue-aligned and superposed structures of the unbound proteins, which greatly facilitates evaluation of the RMSDs of docked structures. Because the bound and unbound molecules are often not identical, this step requires nontrivial manual effort. The sequence alignments are accessible by following the "Sequence Alignment" column of each test case and the cleaned-up PDB files of the superposed structures can be downloaded as a single gzipped file (http://zlab.bu.edu/benchmark). We suggest using randomly rotated configurations of the superposed structures as the starting structures for docking, so that the results are not biased because a near-native conformation is sampled by default.

SUMMARY

Benchmark 3.0 includes all possible test cases from the structures deposited in the PDB up to May 2007, and represents a significant increase in cases over the previous versions. With 127 nonredundant test cases, this benchmark should enable the development and testing of algorithms that require a large training set, in addition to those developed for a particular biochemical category or difficulty level.

ACKNOWLEDGMENTS

The authors are grateful to the Scientific Computing Facilities at Boston University and the Advanced Biomedical Computing Center at NCI, NIH for computing support.

REFERENCES

1. 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.
2. Chen R, Mintseris J, Janin J, Weng Z. A protein-protein docking benchmark. Proteins 2003;52:88–91.
3. 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.
4. Bordner AJ, Gorin AA. Protein docking using surface matching and supervised machine learning. Proteins 2007;68:488–502.
5. Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. Proteins 2007;69:139–159.
6. Li CH, Ma XH, Shen LZ, Chang S, Chen WZ, Wang CX. Complex-type-dependent scoring functions in protein-protein docking. Biophys Chem 2007;129:1–10.
7. Liang S, Liu S, Zhang C, Zhou Y. A simple reference state makes a significant improvement in near-native selections from structurally refined docking decoys. Proteins 2007;69:244–253.
8. Lorenzen S, Zhang Y. Identification of near-native structures by clustering protein docking conformations. Proteins 2007;68:187–194.
9. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 2006;34:W310–W314 (Web Server issue).
10. Pierce B, Weng Z. ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 2007;67:1078–1086.
11. Audie J, Scarlata S. A novel empirical free energy function that explains and predicts protein-protein binding affinities. Biophys Chem 2007;129:198–211.
12. Headd JJ, Ban YE, Brown P, Edelsbrunner H, Vaidya M, Rudolph J. Protein-protein interfaces: properties, preferences, and projections. J Proteome Res 2007;6:2576–2586.
13. 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.
14. Chen R, Weng Z. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 2002;47:281–294.
15. Hubbard SJ, Thornton JM. NACCESS. 2.1.1: Department of Biochemistry and Molecular Biology, University College London; 1993.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
17. Vajda S. Classification of protein complexes based on docking difficulty. Proteins 2005;60:176–180.
18. 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.
19. 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.