Protein–Protein Docking Benchmark 2.0: An Update

Julian Mintseris,[1] Kevin Wiehe,[1] Brian Pierce,[1] Robert Anderson,[1] Rong Chen,[1] Joël Janin,[2] and Zhiping Weng[1,3]*
[1]*Boston University Bioinformatics Program, Boston, Massachusetts*
[2]*Laboratoire d'Enzymologie et Biochimie Structurales, Gif-sur-Yvette, France*
[3]*Boston University Biomedical Engineering Department, Boston, Massachusetts*

ABSTRACT **We present a new version of the Protein–Protein Docking Benchmark, reconstructed from the bottom up to include more complexes, particularly focusing on more unbound–unbound test cases. SCOP (Structural Classification of Proteins) was used to assess redundancy between the complexes in this version. The new benchmark consists of 72 unbound–unbound cases, with 52 rigid-body cases, 13 medium-difficulty cases, and 7 high-difficulty cases with substantial conformational change. In addition, we retained 12 antibody–antigen test cases with the antibody structure in the bound form. The new benchmark provides a platform for evaluating the progress of docking methods on a wide variety of targets. The new version of the benchmark is available to the public at http://zlab.bu.edu/benchmark2. Proteins 2005;60:214–216.** © **2005 Wiley-Liss, Inc.**

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

INTRODUCTION

Protein–protein docking continues to be an active area of research, and as with any active field, it is important to create and maintain standards and benchmarks that help follow the progress of method development as well as keep the community on the same page. Since the first official release of the Protein–Protein Docking Benchmark,[1] the size of the Protein Data Bank[2] (PDB) has continued to increase both in the number of cocrystallized complex structures and in the number of independently crystallized components of these and already existing complexes. Here we present an update of the benchmark. We have tried to make the data processing involved in compiling the benchmark more automatic to ensure good coverage of the existing protein complex space. In addition, we have sought to make the benchmark more realistic by focusing more on unbound–unbound cases and also modified our redundancy criteria to standardize the process and provide even coverage. As a result, this new version of the benchmark is more of a reconstruction than an update. While almost all of the cases in the old benchmark have similar equivalents in the new version, we have made no effort to keep the same exact structures, choosing instead the best quality structures in all cases.

SEMIAUTOMATED DATA SET RETRIEVAL AND CURATION

To reconstruct the benchmark from the bottom up, we parsed the PDB using methods similar to those described previously.[3] We selected all multichain heteromeric X-ray crystal structures with chain lengths > 30 amino acids and root-mean-square deviation (RMSD) better than 3.25 Å. We also excluded large molecular assemblies, as these are currently unrealistic for docking. Methods based on atomic contact vectors (ACVs)[3] were used to distinguish biological from crystal contacts and to choose a representative biological contact from those asymmetric units that contained multiple copies of a complex. The separation of the remaining complexes into obligate and transient was done mostly by hand, and the obligate complexes were discarded.

For the remaining complexes, the residues involved in protein–protein interaction were mapped onto Structural Classification of Proteins[4] (SCOP) domains. We considered the SCOP family–family pair, as well as the superfamily–superfamily pair as the nonredundant unit. While the SCOP superfamily is commonly used as the cutoff point for protein structure comparison, we felt that a superfamily–superfamily pair nonredundancy unit would be too restrictive. Indeed, in some cases the interactions may be quite different from a structural or physicochemical point of view, even with the same family–family pair. In the context of the data presented here, it turns out that the difference between the 2 redundancy criteria affects only 1 test case, and we decided to keep it. In the case of antibody–antigen complexes, nonredundant test cases were selected manually. We used BLAST[5] to find individually crystallized structures that best matched the cocrystallized interactors. For each family–family pair, we chose the unbound structures that (1) had highest sequence identity to the bound interactors and (2) the highest quality crystal structures, specifically, lowest resolution and fewest residues with missing electron density. Those cases for which we could not find high-quality unbound structures were excluded with the exception of antibody–antigen complexes, which were retained as (unbound antigen)–(bound antibody) cases. These targets offer the possibility of testing epitope recognition methods. The conformational changes in antibody complementarity determining region (CDR) loops could be simulated by deleting them and modeling on the bound antibody framework.

Grant sponsor: U.S. Department of Energy Computational Science Graduate Fellowship (support to J. Mintseris). Grant sponsor: National Science Foundation; Grant numbers: DBI-0078194, DBI-0133834, and DBI-0116574.

*Correspondence to: Zhiping Weng, Department of Biomedical Engineering and Bioinformatics Program, Boston University, 44 Cummington Street, Boston, MA 02215. E-mail: zhiping@bu.edu

Received 21 January 2005; Accepted Accepted 21 February 2005
DOI: 10.1002/prot.20560

TABLE I. Protein–Protein Docking Benchmark 2.0

Complex	Cat.	PDBid 1	Protein 1	PDBid 2	Protein 2	RMSD[b] (Å)	DASA[c] (Å^2)
Rigid-body (63)							
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	0.80	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	1BXB_	Hirustatin	1.40	1737
1MAH_A : F	E	1J06_B	Acetylcholinesterase	1FSC_	Fasciculin	0.61	2145
1PPE_E : I	E	1BTP_	Bovine trypsin	1LU0_A	CMTI-1 squash inhibitor	0.44	1688
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	0.90	2022
2MTA_HL : A	E	2BBK_JM	Methylamine dehydrogenase	2RAC_A	Amicyanin	0.41	1461
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
7CEI_A : B	E	1UNK_D	Colicin E7 nuclease	1M08_B	Im7 immunity protein	0.70	1384
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	0.60	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
2VIS_AB : C	A	1GIG_LH	FAB	2VIU_ACE	Flu virus hemagglutinin	0.80	1296
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
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 GTApase	1HH8_A	p67 Phox	0.71	1179
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_	Epstein–Barr virus receptor CR2	1LY2_A	Complement C3	0.34	800
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
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
1SBB_A : B	O	1BEC_	T-cell receptor β	1SE4_	Staphylococcus enterotoxin B	0.37	1064
2BTF_A : P	O	1IJJ_B	Actin	1PNE_	Profilin	0.75	2063
1BJ1_HL : VW	AB	1BJ1_HL	FAB	2VPF_GH	vEGF	0.50	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	1.30	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 (13)							
1ACB_E : I	E	2CGA_B	Chymotrypsin	1EGL_	Eglin C	2.26	1544
1KKL_ABC : H	E	1JB1_ABC	HPr kinase C-ter domain	2HPR_	HPr	2.20	1641
1BGX_HL : T	A	1AY1_HL	FAB	1CMW_A	Taq polymerase	1.48	5814
1GP2_A : BG	O	1GIA_	Gi-α	1TBG_DH	Gi-βγ	1.65	2287
1GRN_A : B	O	1A4R_A	CDC42 GTPase	1RGP_	CDC42 GAP	1.22	2332
1HE8_B : A	O	821P_	Ras GTPase	1E8Z_A	PIP3 kinase	0.92	1305
1I2M_A : B	O	1QG4_A	Ran GTPase	1A12_A	RCC1	2.12	2779

TABLE I. (Continued)

Complex	Cat.	PDBid 1	Protein 1	PDBid 2	Protein 2	RMSD[b] (Å)	DASA[c] (Å^2)
1IB1_AB : E	O	1QJB_AB	14-3-3 protein	1KUY_A	Serotonin *N*-acteylase	2.09	2808
1IJK_BC : A	O	1FVU_AB	Botrocetin	1AUQ_	Von Willebrand factor dom. A1	0.68	1648
1K5D_AB : C	O	1RRP_AB	Ran GTPase	1YRG_B	Ran GAP	1.19	2527
1M10_A : B	O	1AUQ_	Von Willebrand factor dom. A1	1MOZ_B	Glycoprotein IB-α	2.10	2097
1N2C_ABCD : EF	O	3MIN_ABCD	Nitrogenase Mo-Fe protein	2NIP_AB	Nitrogenase Fe protein	2.13	3635
1WQ1_R : G	O	6Q21_D	Ras GTPase	1WER_	Ras GAP	1.16	2913
Difficult (8)							
1ATN_A : D	O	1IJJ_B	Actin	3DNI_	Dnase I	3.28	1774
1DE4_AB : CF	O	1A6Z_AB	β2-microglobulin	1CX8_AB	Transferrin receptor ectodom	2.59	2066
1EER_A : BC	O	1BUY_A	Erythropoietin	1ERN_AB	EPO receptor	2.44	3347
1FAK_HL:T	O	1QFK_HL	Coagulation factor VIIa	1TFH_B	Soluble tissue factor	6.18	3363
1FQ1_A : B	O	1FPZ_F	CDK inhibitor 3	1B39_A	CDK2 kinase	3.41	1832
1H1V_A : G	O	1IJJ_B	Actin	1D0N_B	Gelsolin	6.62	2071
1IBR_A : B	O	1QG4_A	Ran GTPase	1F59_A	Importin β	2.54	3370
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,[7] after finding the best superposition of bound and unbound interfaces.
[c] Change in accessible surface area upon complex formation calculated using NACCESS.[8]

A NONREDUNDANT BENCHMARK

The above sifting process yielded 84 benchmark cases summarized in Table I, which is organized into 3 groups, meant to provide a rough estimate of the expected difficulty for most docking methods. To calculate the difficulty level for each complex, we mapped the structures onto a 1.2 Å grid and used a 6° Euler angle set[6] to perform a Fast Fourier Transform (FFT) search for conformations that would represent high-quality hits. High-quality hits were defined using CAPRI evaluation criteria, specifically, interface RMSD, fraction of native residue contacts, f_{nat}, and fraction of non-native residue contacts, $f_{non\text{-}nat}$.[7] Three difficulty categories—rigid-body (52), medium difficulty (13), and high difficulty (8)—were formed according to the number of hits attainable using rigid-body transformations of the unbound interactors. This difficulty is primarily related to the degree of conformational change at the protein–protein interface and should be independent of the specific docking methods used. Table II provides a summary analysis of the 3 groups in terms of the CAPRI evaluation parameters mentioned above. The parameters were calculated for "predicted complexes" obtained by finding the superposition of unbound onto bound interactors that minimizes interface RMSD. Therefore, the values in Table II should be close to the best possible results attainable using strictly rigid-body docking.

Following the original benchmark, the complexes in Table I are classified into broad biochemical categories: Enzyme–Inhibitor (23), Antibody–Antigen (10 unbound–unbound + 12 bound–unbound), and Others (39). There are 2 complexes—1GRN and 1WQ1—that would be considered redundant using the SCOP superfamily pair criterion but were left in the benchmark because they were deemed sufficiently different. While the Others category in Table I contains a wide variety of complexes, 12 of them contain a small G-protein domain, involved in signal transduction with different partners. More information, including details on cofactors and other ligands, as well as case-specific information, is provided on the website http://zlab.bu.edu/benchmark2.

TABLE II. Average Statistics for 3 Difficulty Groups in Benchmark 2.0

	I_RMSD[a]	f_{NAT}[b]	$f_{NON\text{-}NAT}$[c]	Number
Rigid Body	0.82	0.75	0.24	63
Medium	1.63	0.58	0.47	13
Difficult	3.67	0.43	0.62	8

[a] RMSD of Cα atoms of interface residues calculated as described previously,[7] 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 $f_{NON\text{-}NAT}$, the fraction of non-native residue contacts in a predicted complex, were calculated following Méndez et al.,[7] with the predicted complex obtained by minimizing the interface RMSD.

ACKNOWLEDGMENTS

Our thanks to the many crystallographers who worked on the structures used in this article and made these structures publicly available.

REFERENCES

1. Chen R, Mintseris J, Janin J, Weng Z. A protein–protein docking benchmark. Proteins 2003;52:88–91.
2. 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.
3. Mintseris J, Weng Z. Atomic contact vectors in protein–protein recognition. Proteins 2003;53:629–639.
4. 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.
5. 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.
6. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52:80–87.
7. Méndez 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.
8. Hubbard SJ, Thornton JM. NACCESS 2.1.1. Department of Biochemistry and Molecular Biology, University College, London; 1993.