doi:10.1016/j.jmb.2011.12.026

J. Mol. Biol. (2012) 415, 443–453

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Modulating Protein–Protein Interactions with Small Molecules: The Importance of Binding Hotspots

Ratna Rajesh Thangudu, Stephen H. Bryant, Anna R. Panchenko* and Thomas Madej*

National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894, USA

Received 22 June 2011; received in revised form 23 November 2011; accepted 12 December 2011
Available online 16 December 2011

Edited by B. Honig

Keywords:
protein structures;
molecular interactions;
protein–protein interactions;
drug development hotspots

The modulation of protein–protein interactions (PPIs) by small drug-like molecules is a relatively new area of research and has opened up new opportunities in drug discovery. However, the progress made in this area is limited to a handful of known cases of small molecules that target specific diseases. With the increasing availability of protein structure complexes, it is highly important to devise strategies exploiting homologous structure space on a large scale for discovering putative PPIs that could be attractive drug targets. Here, we propose a scheme that allows performing large-scale screening of all protein complexes and finding putative small-molecule and/or peptide binding sites overlapping with protein–protein binding sites (so-called "multibinding sites"). We find more than 600 nonredundant proteins from 60 protein families with multibinding sites. Moreover, we show that the multibinding sites are mostly observed in transient complexes, largely overlap with the binding hotspots and are more evolutionarily conserved than other interface sites. We investigate possible mechanisms of how small molecules may modulate protein–protein binding and discuss examples of new candidates for drug design.

Introduction

Protein–protein interactions (PPIs) play a key role in numerous biological processes such as cell proliferation, growth, differentiation, signal transduction and apoptosis; moreover, it has been shown that PPIs are disrupted in many diseases including cancer.[1,2] This suggests the attractive possibility of manipulating PPIs for therapeutic intervention. However, targeting PPIs is more challenging than traditional drug discovery that, for example, designs small molecules to bind to enzyme active sites. The complications arise from the fact that PPI interfaces are relatively large, less conserved, often flat or more shallow and featureless in contrast with ligand binding pockets.[3–7] Presently, there are small-molecule drugs known to affect about 1% of human proteins,[8] and 10–15% of all human proteins are considered "druggable".[9] The historical record of drug design and discovery has given rise to the idea that PPIs are much more intractable with respect to small-molecule drug discovery.[8] Indeed, for therapeutic use, the chemicals or drugs should be small enough to get inside the cell and also be able to affect the large and often shallow PPI interaction sites.

Nevertheless, there have been a number of studies that suggest targeting PPIs for treatments for some human diseases,[10–21] showing that protein–protein interfaces or regions near interfaces might be inherently flexible or intrinsically disordered

*Corresponding authors. E-mail addresses: panch@ncbi.nlm.nih.gov; madej@ncbi.nlm.nih.gov.

Abbreviations used: PPI, protein–protein interaction; PDB, Protein Data Bank; CDD, Conserved Domain Database; MMDB, Molecular Modeling Database; GO, gene ontology.

allowing a small molecule to penetrate these complexes and displace the protein interaction partner.[22–24] Several papers review the progress made in this research area.[8,25–29] Additional evidence that small molecules do not have to cover the entire protein–protein binding interface but rather target only a small number of interface residues, the so-called "binding hotspot" sites, which contribute the most to the binding energy, has been obtained.[30]

Many approaches so far have focused on discovering druggable PPIs by *in silico* screening of small-molecule libraries and searching the chemical space of PPI inhibitors. It was found that PPI inhibitors usually represent relatively large rigid small molecules, containing hydrophobic and aromatic groups.[4–7] A recent study showed that SVM (*s*upport *v*ector *m*achine) kernels can be successfully used to select molecular descriptors for PPI inhibitors, which characterize specific molecular shapes and the presence of a privileged number of aromatic groups.[9] Molecular dynamics simulations, drug design and protein docking studies tried to uncover dynamic and physicochemical properties of protein–protein complexes to find those regions and pockets that can be targeted in small-molecule library screenings.[31] Recently, a model was proposed to predict the druggability of pharmaceutically important proteins based on the crystal structures of the binding pockets.[32] Despite these efforts, the most comprehensive list of known structure complexes with small molecules disrupting protein–protein interfaces is still very limited [27 Protein Data Bank (PDB) complexes in total][26] and is represented by only eight protein families. More recently, a database with an extensive analysis of these known protein–protein interfaces was made available.[33]

It is clear that the systematic and large-scale analysis of experimentally observed protein–small-molecule and protein–protein complexes is needed to discover potential protein–protein interfaces that, at the same time, have tendencies to bind small molecules. Such an approach has been undertaken recently by looking for homologous complexes in a protein structure database with overlapping protein–protein and protein–small-molecule binding sites.[34,35] This study demonstrated that sampling of the space of homologs is an extremely useful and encouraging approach that not only allowed the recovery of known interaction modulators but also provided a list of potential drug targets. However, a large source of error might come from including complexes that are the result of crystal packing interactions. In this regard, specific methods have been developed to predict and confirm the biological relevance of specific interfaces in crystal structures.[36,37] Another source of annotation errors comes from inferring that functions and interactions from distant homologs are common descents and do not necessarily imply similarity in function or interactions and that annotations transferred from one protein to a homolog may result in incorrect functional or interolog assignment at larger evolutionary distances.[38] In the current study, we used the homology inference approach and used our recently developed IBIS (*I*nferred *B*iomolecular *I*nteraction *S*erver) method[39,40] to find those protein complexes that can potentially be targeted by small molecules. To ensure biological relevance of binding sites to the query, IBIS clusters similar binding sites found in homologous proteins based on their sequence and structure conservation, further validates them using various approaches and finally ranks binding sites to assess how well they match the query.

We look for those sites that bind both proteins and small molecules and define them as "multibinding sites". We ask what thermodynamic and structural properties of protein complexes make them more targetable by small molecules. We find that small molecules have a tendency to bind to hotspot residues and preferentially target weaker and more transient protein–protein interfaces. Moreover, we show that multibinding sites are more conserved than the rest of the interface. From the most recent update of the protein structure databank (Research Collaboratory for Structural Bioinformatics PDB[41]), we compile a nonredundant set of potential PPI interfaces from 642 proteins representing 60 protein families, with strong evidence of multibinding and potential properties of small-molecule PPI inhibitors.

Results

Currently, a total of 239,395 protein chains/domains from 61,413 protein structures are present in IBIS with at least one type of interaction either observed in their structural complexes or inferred from their homologs[†]. Our method allows analyzing the mechanisms of how a small molecule competes with a natural protein partner.

Observed cases of small molecules modulating PPIs

First, we focused on those cases where protein–protein and protein–small-molecule complexes are available as separate structures in the structure databases. Some of the known examples are shown in Fig. 1 with the structure superpositions of the two

[†] All multibinding sites observed in our study for proteins in the PDB are accessible from http://www.ncbi.nlm.nih.gov/Structure/ibis/P-D/multibinding.html.

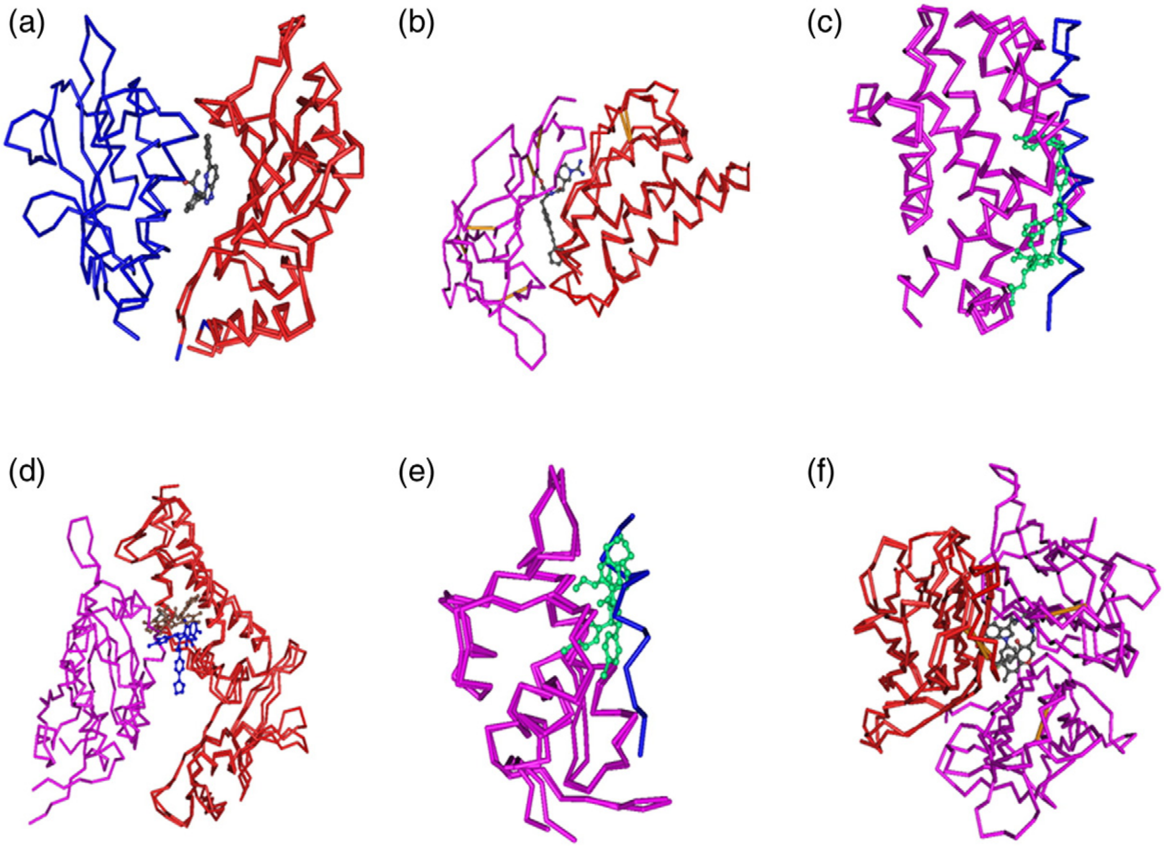

Fig. 1. Known cases of binding of small molecules to protein–protein interfaces with a comparison of the crystal structures of their protein–protein and protein–small-molecule complexes. (a) Small-molecule inhibitor of the ZipA–FtsZ PPI (PDB IDs: 1Y2G and 1F46). (b) A small molecule bound to adaptive IL2 and IL2R-alpha interface (PDB IDs: 1Z92 and 1M48). (c) Small molecule triggered Bax/Bak-mediated apoptosis in Bcl-2 proteins [Bcl-2, Bcl-x(L) and Bcl-w] (PDB IDs: 2BZW and 2YXJ). (d) A small molecule bound to the N-terminal transactivation domain of human papillomavirus type 11 E2 and inhibits its interaction with E1 (PDB IDs: 1TUE and 1R6N). (e) Small molecule inhibiting the p53–MDM2 to activate the p53 pathway in cancer cells (PDB IDs: 1RV1 and 1YCR). (f) A small-molecule inhibitor of tumor necrosis factor-alpha that promotes disassembly of this trimeric cytokine (PDB IDs: 1TNF and 2AZ5).

different complexes of the same protein. We first tested if our method can recover known examples from the literature of PPIs that are modulated by small molecules. The method successfully recovered six out of eight known cases (Table SM1 and Fig. 1). The interaction between IL2 and IL2-alpha receptor was not recovered because of a partial Conserved Domain Database (CDD) domain mapping (see Materials and Methods). The case of tumor necrosis factor-alpha trimer dissociation mediated by a small molecule has been missed due to the stringent overlap threshold used in our method. In a survey of all the observed PPI and small-molecule interactions in the current PDB, 3223 domains/chains were found to have their observed PPI interfaces overlap with observed or inferred small-molecule binding sites; on the other hand, 4532 protein domains/chains have their observed small-molecule binding sites overlapping with observed or inferred PPI interfaces.

The same protein may be represented in multiple structures solved under different conditions or with mutations and/or in complex with several different small molecules. To account for this, we inferred PPI from close homologs with more than 90% sequence identity and found 6255 chains/domains in PDB with multibinding interfaces. A few examples are presented in Table 1, and the complete list can be accessed from the Web page provided in Supplementary Information.

Binding sites inferred from homologs

For each protein chain in PDB/Molecular Modeling Database (MMDB), we assembled a comprehensive list of inferred protein–protein and protein–small-molecule binding sites using IBIS. Since inferred binding sites represent the consensus of binding sites from close homologs, their location reflects the conformational diversity and variability

Table 1. Protein–protein interfaces overlapping with small-molecule binding sites from very close homologs

PDB	Interacting chains	Interacting domains	Homolog with bound small molecule	Small molecule
1A14	N:H	Sialidase–IgV_H	2C4AA	*O*-Sialic acid
1AY7	A:B	RNase_Sa–barstar	1RGEA	2′-Guanylic acid
1EAW	A:B	Tryp_SPc–KU	1EAXA	Benzamidine
1EYM	B:A	FKBP_C–FKBP_C	1FKIA	AC1L9IC3
1F46	B:A	ZipA–ZipA	1S1JA	CHEMBL1233656
1G5J	A:B	Bcl-2_like–Bcl-2_BAD	1YSGA	4-(4-Fluorophenyl)benzoic acid
1JSU	B:C	CYCLIN–CDI	2UUEB	GVC
1L6X	A:A	Ig–Ig	3D6GB	CHEMBL1236746
1M9C	A:D	Cyclophilin_ABH_like–Gag_p24	3CYHA	L-Proline
1TGS	Z:I	Tryp_SPc–KAZAL_PSTI	3LJOA	CHEMBL1229618
1VLB	A:A	Fer2–Ald_Xan_dh_C2	1VLBA	2Fe–2S cluster
1XEV	A:B	alpha_CA_I_II_III_XIII–alpha_CA_I_II_III_XIII	1ZFQA	CHEMBL6685
1XFP	L:A	LYZ1–Ig	1H6MA	Acetylglucosamine
1XX9	B:D	Tryp_SPc–ecotin	3BG8A	Benzamidine
2E1A	D:C	WHTH_GntR–WHTH_GntR	2E1AD	Selenomethionine
2G2U	A:B	beta-Lactamase–BLIP	2ZD8A	MER
2GPV	D:B	NR_LBD_ERR–NR_LBD_ERR	1S9QA	Cholic acid
2HS1	A:B	HIV_retropepsin_like–HIV_retropepsin_like	9HVPB	AC1L9IBF
2KNE	A:B	EFh–ATP_Ca_trans_C	1QIWA	AC1L9LMM
2VU8	E:I	Tryp_SPc–Pacifastin_I	1PQ7A	AC1L9LDG
2W0D	A:C	ZnMc_MMP–ZnMc_MMP	1UTZA	AC1L9MIX
3BIM	C:F	BTB–BTB	3LBZB	Z89
3BX7	A:C	Lipocalin–IgV_CTLA-4	3BX8B	PE5
3CKI	A:B	ZnMc_TACE_like–NTR_like	2I47A	AC1L9GK4
3DAW	A:B	ACTIN–ADF	2Q36A	KAB
3E1Z	B:A	Peptidase_C1A–Chagasin_I42	1BP4A	AC1L9GMA
3FDL	A:B	Bcl-2_like–Bclx_interact	1YSGA	4-(4-Fluorophenyl)benzoic acid
3FP7	J:E	KU–Tryp_SPc	3LDJB	Sucrose octasulfate
3FXI	C:B	ML–LRR_RI	3FXIC	AC1L9IAU
3IG6	B:D	Tryp_SPc–Tryp_SPc	1SQTA	AC1L9M5D
4CPA	A:I	M14_CPA–CarbpepA_inh	3CPAA	Glycinamide
7GSP	A:B	Fungal_RNase–fungal_RNase	1RGAA	Guanosine

of homologous complexes and at the same time differences in the sizes of small molecules. A total of 27,340 protein chains from 501 CDD families have been determined to contain inferred multibinding sites, and a nonredundant set of 642 chains (culled at 50% sequence identity) from 60 families was compiled with multibinding interfaces that are biologically relevant according to the PISA (*P*rotein *I*nterfaces, *S*urfaces, and *A*ssemblies server) algorithm.[36]

Although establishing the biological validity of each of these multibinding sites still requires experimental verification, these sites might be used as starting points to target small-molecule PPI inhibitors. We provide additional annotation for the multibinding sites including their PISA status, biological relevance of small molecules or verification of small molecules using DrugBank[42] in Supplementary Materials. For example, of the 642 protein chains, about 400 have at least one multibinding site in which the bound small molecule is also biological.

It should be mentioned that protein structures in PDB often contain additives, detergents and other types of substances used for crystallization. These are not true biological ligands, but they can sometimes be difficult to distinguish from the biologically relevant ones. In the current study, we distinguish the most common biological ligands (see Supplementary Materials). Indeed, the crystallization agents, if present, may sometimes provide additional insights into the binding interfaces on a protein. For example, the nonbiological small molecule in the interface between the E2 protein and E1 helicase of human papillomaviruses probably defines additional regions of the binding pocket that could be exploited to design more potent inhibitors[21] (Fig. 1d).

We also note that the PISA program[36] is regarded as a state-of-the-art method for the annotation of biological assemblies, with an accuracy estimated at 80–90%. One might expect that the author-determined biological units in the PDB files would be the most reliable. However, it is not difficult to find examples where the annotations in the PDB file are not consistent with the authors' own paper; some authors are quite diligent about this, whereas others may not be so careful. Moreover, an estimate for the accuracy of the author-determined annotations is lacking. Therefore, a highly dependable method such as PISA provides the most consistent results, although researchers should be careful and should

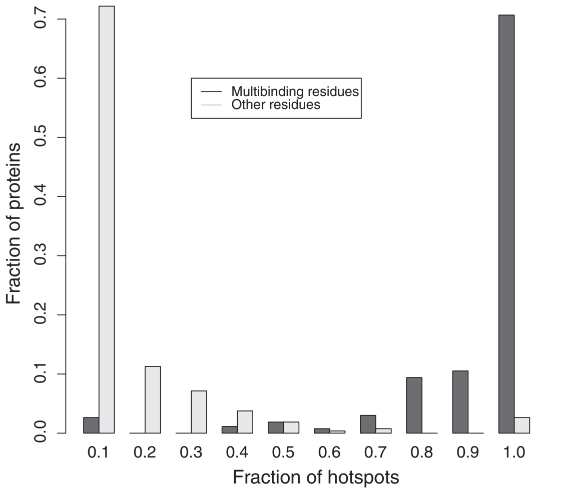

Fig. 2. Distribution of fraction of binding hotspots that overlap multibinding residues (dark gray) and that do not overlap multibinding residues (light gray).

consult the relevant literature when accuracy is critical.

Multibinding interfaces have a tendency to include binding hotspots

We calculated a conservation score using the Shannon entropy with the Henikoff–Henikoff sequence weights for each position in a multibinding site based on the binding site cluster alignments. The highly conserved residues are most likely critical for binding and may well be the binding hotspot residues, namely, those that contribute the most to the binding energy of the protein complex. We analyzed the nonredundant set of protein chains in our data set to check how often the multibinding residues are predicted to be binding hotspots. Hotspot residues were predicted using the PCRPi (*P*resaging *C*ritical *R*esidues in *P*rotein *i*nterfaces) method, which integrates a number of different metrics involving sequence and structure, sequence conservation, "topographical index", computational alanine scanning and others into a probabilistic measure by using Bayesian networks.[30] A residue in an inferred binding site is considered as a hotspot if the corresponding residue in the homolog contributing to the site is annotated as a hotpot by PCRPi.

Among the 642 nonredundant protein sequences with biological interfaces and multibinding sites, 259 chains had at least one hotspot on their multibinding inferred site. We found that the association between multibinding sites and hotspots is statistically significant (χ^2 *p*-value $\ll 0.01$) (Fig. 2), which points to the critical role of binding hotspots in modulating PPIs by small molecules.

At the same time, the analysis of residues in the protein–protein interfaces showed that multibinding residues are more evolutionarily conserved than the rest of the interface (χ^2 *p*-value $\ll 0.01$) (Fig. 3), and hotspots on multibinding interfaces are more conserved than the rest of multibinding interface (χ^2 *p*-value $\ll 0.01$). Previously, it was shown that binding hotspots are more evolutionarily conserved than the rest of the interface.[43]

We mention that small-molecule binding sites can overlap protein–protein binding sites for different functional reasons. First, small molecules can represent natural substrates of enzymes that, in turn, can be inhibited through the mechanism of competitive binding by other proteins. A second group consists of all other cases where small molecules modulate PPIs. Automatically, classifying a small molecule as a native substrate is a challenging task, and there is no such information provided in the PDB files. Therefore, we used a data source from the previous study,[44] which compared the PDB ligands to small molecules from the ENZYME and KEGG databases using graph matching algorithms to assess the chemical similarity between small molecules. We mapped these native/cognate ligand annotations to inferred binding sites in IBIS. A total of 108 proteins from the nonredundant set of 642 proteins were found to have at least one multibinding site for a cognate ligand/native substrate. The analysis of these multibinding sites with native substrates in these 108 proteins showed a similar trend of multibinding residues to be more conserved and to have a higher tendency to include binding hotspots (χ^2 *p*-value $\ll 0.01$). However, the first group of enzyme–substrate complexes showed a lower hotspot frequency but a higher evolutionary conservation of

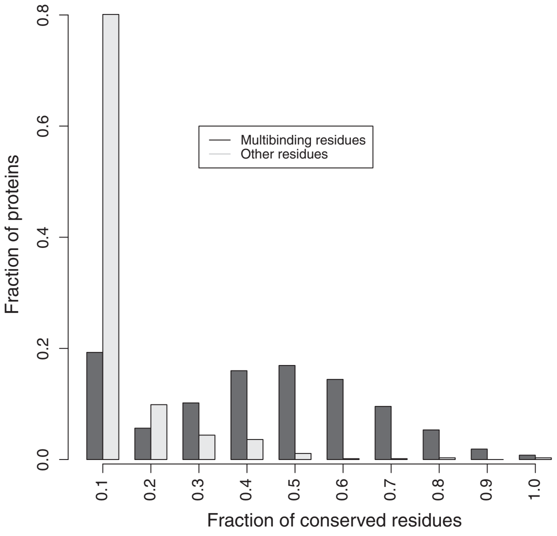

Fig. 3. Distribution of fraction of conserved multibinding residues (dark gray) and of conserved nonmultibinding residues (light gray). Conserved residues are defined as those with normalized entropy scores between 0.6 and 1.039.

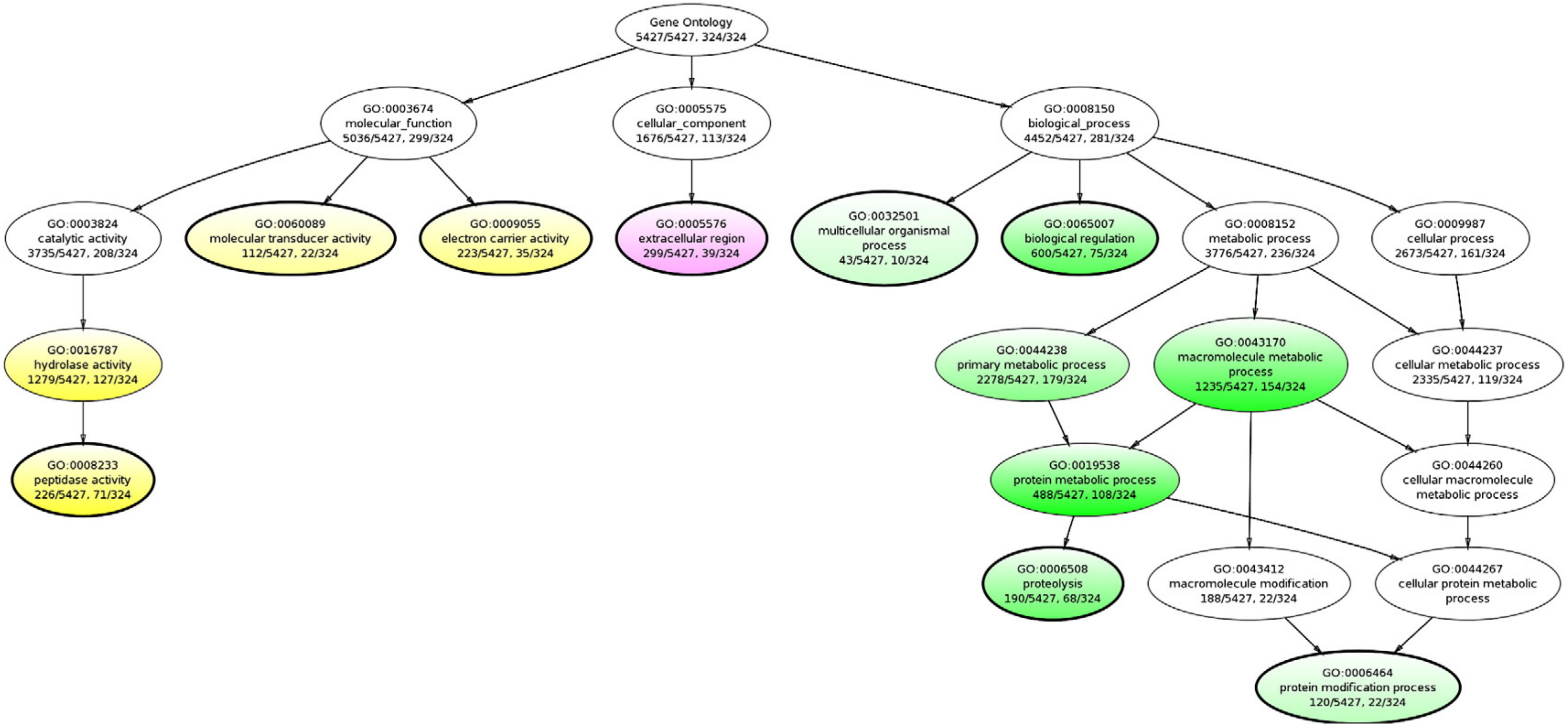

Fig. 4. GO function distribution of 891 nonredundant protein chains (culled at 50% sequence identity) based on GO slim terms.

multibinding residues compared to the second group of multibinding interfaces.

Functional analysis of proteins with multibinding sites

The functional enrichment of the nonredundant set of 642 multibinding proteins has been assessed using a selected list (GO slim) of gene ontology (GO)[45] functional terms. We define this as a study group—a set of multibinding proteins found in PDB from our study—and a population group—all of the proteins in PDB. The frequency of annotation to a GO term for the study group is then compared to the overall population, which compensates for the functional bias in the PDB. We used Ontologizer,[46] which performs a modified Fisher's exact test with correction for multiple testing and also takes the parent–child relationship[47] in the GO hierarchy into consideration. GO assignments to PDB entries have been derived from the "gene_association.goa_pdb" gene association file provided by the UniProtKB-GOA group. We found that metabolic, regulation, transducer activity, electron carrier activity and multicellular organismal processes are significantly enriched with multibinding proteins (χ^2 p-value $\ll 0.01$) (Fig. 4). An analysis based on the assigned Enzyme Commission[48] numbers of the nonredundant set of multibinding proteins and the rest of the proteins in PDB showed that multibinding sites are significantly overrepresented in enzymes (χ^2 p-value $\ll 0.01$), that is, protein chains with assigned Enzyme Commission numbers.

Small molecules modulate transient PPIs

We checked the hypothesis whether the ability of small molecules to modulate and inhibit PPIs will depend on the stability of protein–protein complexes. We estimated the free energy of dissociation ΔG^{diss} by using the PISA algorithm. We compared the stability of dimeric complexes containing multibinding sites against all nonredundant dimers in the PDB as determined by PISA. As can be seen in Fig. 5, dimers with multibinding sites are less stable compared to the other dimers (t-test p-value = 0.016). Indeed, proteins with multibinding sites constitute a significantly smaller fraction among permanent protein complexes with the dissociation constant in the nanomolar-to-picomolar range and ΔG^{diss} of 20–30 kcal/mol or higher.

Mechanisms of action in the proteins with multibinding interfaces

A majority of multibinding proteins observed in structure databases[41] include synthetic small molecules targeting protein inhibitor binding sites. The modulation of PPIs is carried out through disruption, inhibition, stabilization or allosteric regulation.

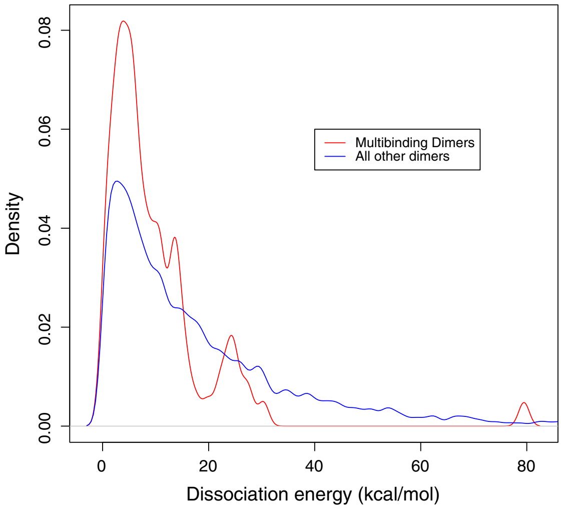

Fig. 5. Distribution of dissociation energies of protein dimers with and without multibinding sites at the protein–protein interfaces (red, dimers with multibinding sites; blue, all other dimers). Bandwidth (h) = 1.

For example, kirromycin antibiotics such as aurodox lock EF-Tu (*e*longation *f*actor-*t*hermo *u*nstable) in its EF-Tu/GTP conformation, preventing its release from the ribosome, which illustrates the mediation of a PPI by a small molecule. Similarly, the fungal phytotoxin fusicoccin stabilizes the interaction of 14-3-3 with PMA2, a plant proton pump.[49] Two more small molecules, pyrrolidone 1 and epibestatin, have recently been found to stabilize the 14-3-3–PMA2 interaction.[50] Rapamycin, a potent immunosuppressive drug, mediates the interaction between the human signaling proteins FKBP12 and FRAP that do not normally interact with one another.[51] The overlap of the rapamycin binding site (1FAP_A: RAP) with type I TGF-beta receptor in complex with FKBP12 (1B6C_A:B) suggests a possible role of rapamycin in TGF-beta signaling. Indeed, FKBP12 binding to the TGF-beta receptor shields a regulatory segment of the receptor from phosphorylation, which maintains the receptor in its inactive state. Therefore, rapamycin bound to FKBP12 should permit easier activation of the TGF-beta receptor.[52]

We found that the GTPase (Cdc42) and cell polarity protein (Par6) PDZ domain interface (1NF3_A:C) overlaps with the AMP binding site of a small GTPase Rab1b (3NKV:AMP), suggesting a possible regulation of Par6 PDZ binding to Cdc42 through AMPylation. It has also been shown that mutational disruption of Cdc42–Par6 PDZ coupling leads to inactivation of Par6 in a certain type of epithelial cells.[53] Another example is the cresidine binding site of Cu–Zn superoxide dismutase (SOD1) (2WZ0_F:ZZT), which overlaps with the dimer interface of Cu–Zn superoxide dismutase (2NNX_D:A). Recently, this binding site has been annotated as "druggable" for therapeutic purposes for SOD1-associated motor neuron disease.[54] A complete list of all the protein chains with observed and inferred multibinding sites is available in Supplementary Materials.

Discussion

The rapid increase in data on protein sequences and structures is posing new challenges to interpret and use these data productively. While many studies have presented novel methods for functional annotation of these sequences and structures, the annotations in public databases are still error prone or often hypothetical. Even when structural data are available for a particular protein, it can be inconclusive or hard to interpret. Manual curation by an expert, which is the most rigorous and reliable means to legitimize a functional site, is limited by the sheer volume of the data.

Toward addressing this problem, we recently developed an algorithm (IBIS) to analyze and conservatively annotate binding sites in proteins based on knowledge gained via homologous complexes. One important advantage of IBIS-derived sites is that they are weighted based on their recurrence in homologous proteins and ranked using binding site sequence and structure conservation. In the current study, we used the IBIS database to discover many protein–protein interfaces that potentially also bind small molecules. We have found that about 33% of all protein chains/domains with both PPI and small-molecule sites inferred from their homologs have multibinding sites. The likelihood that these sites are biologically relevant is increased by the conservative thresholds built into our method. The GO analysis suggests that multibinding sites are often enriched in metabolic processes. This is also reflected by the significant number of multibinding sites in GO enzyme annotations.

An earlier study showed that those positions, which may bind both small molecules and other proteins, are less conserved compared to monofunctional sites and also exhibit different amino acid propensities.[34] However, we found that multibinding sites are more evolutionarily conserved and more likely to contain binding hotspots than other interface positions that are potentially involved in binding of one partner. Moreover, it has been shown previously that hotspots are more conserved than other interface residues,[43] and we confirmed this result here with regard to hotspots on multibinding interfaces. Since hotspot residues contribute the most to the binding energy of PPIs, we suggest that binding of small molecules to these positions will have the most disruptive effect on those interactions. It is especially true if these interactions are not very strong. Indeed, as was shown in our study, small molecules mostly bind to transient protein–protein complexes.

It should be mentioned that the apparent conflict between our conservation results for multibinding residues and the results presented previously[34] are undoubtedly due to important differences in methodology between the two studies. For instance, the earlier work used SCOP and CATH domain definitions and classifications, which allows for considerably more remotely related proteins and may affect the reliability of homology inference. IBIS uses a conservative threshold for homology inference (at least 30% sequence identity), and we observe that multibinding sites are under stronger evolutionary constraints than even the fairly conserved family background.

Binding of small molecules to proteins and protein complexes could cause a shift of equilibrium in favor of a subset of conformations that has higher or lower preferences to binding another partner. A small molecule may bind at a site far from the protein–protein interface and regulate protein binding through allosteric mechanisms or might bind at or near the protein–protein interface and directly influence the binding. In this study, we focused on the latter case. We showed that small molecules can bind to hotspots and through competitive binding prevent PPI. At the same time, we show examples of small molecules mediating PPIs. Other mechanisms of small molecule binding and their functional roles might be elucidated in future studies when more structural complexes are available.

Materials and Methods

Defining interactions

Protein complexes from the current release of the MMDB,[55] an automatically parsed and validated derivative of the Research Collaboratory for Structural Bioinformatics PDB,[41] are used in this study. PPIs are identified and analyzed at the domain level. The domain assignment is performed by searching the protein sequence against a comprehensive collection of domain models in the CDD.[56] PPIs are recorded between different functional domains in the same chain or between different chains from a protein complex.

Protein–small-molecule and protein–peptide interactions are defined for a complete protein chain regardless of its domain annotations. Peptide is defined as a segment of polypeptide chain of 20 amino acids or fewer. Both protein–protein and protein–peptide interactions are considered as PPIs in this study. For small molecules bound to multiple chains, the interaction is assigned to the chain with dominant contacts (>75% of the contacts); otherwise, each protein interaction with the small molecule is recorded separately. An interaction is defined if a protein domain/chain has at least five residues in contact with another protein, small molecule or peptide, and two residues are said to be in contact if any of the heavy-atom interatomic distances is smaller than 4 Å. The "binding site" refers to a group of residues that make a contact with an interaction partner.

Inferring binding sites using IBIS

We have used the IBIS method described earlier[39,40] to analyze experimentally observed complexes and at the same time infer interaction partners and binding sites in proteins without known complexes by inspecting homologs with known interactions. For a given query protein, IBIS collects all its homologs with known structures of complexes from MMDB that have significant structural similarity and at least 30% sequence identity to the query as calculated from the structure–structure VAST[57] alignment. IBIS then clusters binding sites using a complete-linkage clustering algorithm. Binding site similarity is assessed based on the structural alignment using similarity scores. At the end of this step, a list of all inferred binding site clusters and binding partners (chain/domains, small molecules and peptides) is compiled, which is derived from homologous structural complexes. We refer to these inferred binding site clusters as inferred binding sites. For each interaction type (protein–protein, protein–small molecule and protein–peptide), all binding site clusters are ranked in terms of their biological relevance and similarity to the query. The components of the ranking score include the sequence PSSM score, the average sequence identity between the query and cluster members calculated over the whole structure–structure alignment and the number of interfacial contacts and the average sequence conservation of binding site alignment columns. The binding site clusters that contain the observed interactions of the query are regarded as observed binding sites, and the rest are defined as inferred binding sites coming from the homologs.

In addition to the ranking scheme, which aims to rank the inferred binding sites using evolutionary relatedness with respect to the query protein, we used other sources. In the case of protein–small-molecule interactions, small molecules were all validated and standardized by the PubChem database,[58] which often provides extensive information on their known biological activities. Small molecules with less than five heavy atoms and/or having a molecular mass outside the range of 70–800 Da were ignored in this study. We also excluded nonbiological small molecules based on the list used in our previous study.[40] A small-molecule-inferred binding site is deemed nonbiological if all the bound small molecules in an inferred binding site are nonbiological.

Likewise for PPIs, the oligomeric states and binding interfaces were verified using PISA algorithm,[36] which identifies biologically relevant interfaces present in crystal structures. If all interfaces in an inferred PPI binding site are invalid according to PISA, the site is deemed nonbiological.

Finding sites that bind proteins and small molecules

To detect those sites that can bind to both proteins/peptides and small molecules, for a given protein with known structure, we extract all IBIS-inferred sites interacting with other proteins, peptides and small molecules. An inferred binding site is a union of the observed binding site residues of the members of an inferred binding site cluster. The overlap score between protein–protein binding sites and protein–small-molecule binding sites is calculated as:

$$SC = \frac{N_{ab}}{N_a + N_b - N_{ab}}$$

where N_a is the number of residues in protein–protein binding site "a", N_b is the number of residues in the protein–small-molecule binding site "b" and N_{ab} is the number of residues in the intersection of the binding sites "a" and "b", that is, the number of multibinding residues (as indexed with respect to the query). The multibinding sites are defined as those having an SC score greater than 0.5.

Supplementary materials related to this article can be found online at doi:10.1016/j.jmb.2011.12.026

Acknowledgements

This work was supported by National Institutes of Health/Department of Health and Human Services (Intramural Research Program of the National Library of Medicine).

References

1. Schuster-Bockler, B. & Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biol.* **9**, R9.
2. Teng, S., Madej, T., Panchenko, A. & Alexov, E. (2009). Modeling effects of human single nucleotide polymorphisms on protein–protein interactions. *Biophys. J.* **96**, 2178–2188.
3. Fletcher, S. & Hamilton, A. D. (2007). Protein–protein interaction inhibitors: small molecules from screening techniques. *Curr. Top. Med. Chem.* **7**, 922–927.
4. Pagliaro, L., Felding, J., Audouze, K., Nielsen, S. J., Terry, R. B., Krog-Jensen, C. & Butcher, S. (2004). Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Curr. Opin. Chem. Biol.* **8**, 442–449.
5. Fry, D. C. (2006). Protein–protein interactions as targets for small molecule drug discovery. *Biopolymers*, **84**, 535–552.
6. Fry, D. C. (2008). Drug-like inhibitors of protein–protein interactions: a structural examination of effective protein mimicry. *Curr. Protein Pept. Sci.* **9**, 240–247.
7. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. (2008). Principles of protein–protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.* **108**, 1225–1244.
8. Whitty, A. & Kumaravel, G. (2006). Between a rock and a hard place? *Nat. Chem. Biol.* **2**, 112–118.

9. Reynes, C., Host, H., Camproux, A. C., Laconde, G., Leroux, F., Mazars, A. *et al.* (2010). Designing focused chemical libraries enriched in protein–protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.* **6**, e1000695.
10. Asada, S., Choi, Y. & Uesugi, M. (2003). A gene-expression inhibitor that targets an α-helix-mediated protein interaction. *J. Am. Chem. Soc.* **125**, 4992–4993.
11. Bruncko, M., Oost, T. K., Belli, B. A., Ding, H., Joseph, M. K., Kunzer, A. *et al.* (2007). Studies leading to potent, dual inhibitors of Bcl-2 and Bcl-xL. *J. Med. Chem.* **50**, 641–662.
12. Charpentier, T. H., Wilder, P. T., Liriano, M. A., Varney, K. M., Zhong, S., Coop, A. *et al.* (2009). Small molecules bound to unique sites in the target protein binding cleft of calcium-bound S100B as characterized by nuclear magnetic resonance and X-ray crystallography. *Biochemistry*, **48**, 6202–6212.
13. Daelemans, D. (2002). A synthetic HIV-1 Rev inhibitor interfering with the CRM1-mediated nuclear export. *Proc. Natl Acad. Sci. USA*, **99**, 14440–14445.
14. Lepourcelet, M., Chen, Y. N. P., France, D. S., Wang, H., Crews, P., Petersen, F. *et al.* (2004). Small-molecule antagonists of the oncogenic Tcf/β-catenin protein complex. *Cancer Cell*, **5**, 91–102.
15. Murray, J. K. & Gellman, S. H. (2007). Targeting protein–protein interactions: lessons from p53/MDM2. *Biopolymers*, **88**, 657–686.
16. Oltersdorf, T., Elmore, S. W., Shoemaker, A. R., Armstrong, R. C., Augeri, D. J., Belli, B. A. *et al.* (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature*, **435**, 677–681.
17. Oost, T. K., Sun, C., Armstrong, R. C., Al-Assaad, A. S., Betz, S. F., Deckwerth, T. L. *et al.* (2004). Discovery of potent antagonists of the antiapoptotic protein XIAP for the treatment of cancer. *J. Med. Chem.* **47**, 4417–4426.
18. Sutherland, A. G., Alvarez, J., Ding, W., Foreman, K. W., Kenny, C. H., Labthavikul, P. *et al.* (2003). Structure-based design of carboxybiphenylindole inhibitors of the ZipA–FtsZ interaction. *Org. Biomol. Chem.* **1**, 4138.
19. Tilley, J. W., Chen, L., Fry, D. C., Emerson, S. D., Powers, G. D., Biondi, D. *et al.* (1997). Identification of a small molecule inhibitor of the IL-2/IL-2Rα receptor interaction which binds to IL-2. *J. Am. Chem. Soc.* **119**, 7589–7590.
20. Tsao, D. H. H., Sutherland, A. G., Jennings, L. D., Li, Y., Rush Iii, T. S., Alvarez, J. C. *et al.* (2006). Discovery of novel inhibitors of the ZipA/FtsZ complex by NMR fragment screening coupled with structure-based design. *Bioorg. Med. Chem.* **14**, 7953–7961.
21. Wang, Y. (2003). Crystal structure of the E2 transactivation domain of human papillomavirus type 11 bound to a protein interaction inhibitor. *J. Biol. Chem.* **279**, 6976–6985.
22. Volkov, A. N., Worrall, J. A., Holtzmann, E. & Ubbink, M. (2006). Solution structure and dynamics of the complex between cytochrome *c* and cytochrome *c* peroxidase determined by paramagnetic NMR. *Proc. Natl Acad. Sci. USA*, **103**, 18945–18950.
23. Eyrisch, S. & Helms, V. (2007). Transient pockets on protein surfaces involved in protein–protein interaction. *J. Med. Chem.* **50**, 3457–3464.
24. Fong, J. H., Shoemaker, B. A., Garbuzynskiy, S. O., Lobanov, M. Y., Galzitskaya, O. V. & Panchenko, A. R. (2009). Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS Comput. Biol.* **5**, e1000316.
25. Arkin, M. R. & Wells, J. A. (2004). Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev., Drug Discov.* **3**, 301–317.
26. Higueruelo, A. P., Schreyer, A., Bickerton, G. R., Pitt, W. R., Groom, C. R. & Blundell, T. L. (2009). Atomic interactions and profile of small molecules disrupting protein–protein interfaces: the TIMBAL database. *Chem. Biol. Drug Des.* **74**, 457–467.
27. Morell, M., Aviles, F. X. & Ventura, S. (2009). Detecting and interfering protein interactions: towards the control of biochemical pathways. *Curr. Med. Chem.* **16**, 362–379.
28. Toogood, P. L. (2002). Inhibition of protein–protein association by small molecules: approaches and progress. *J. Med. Chem.* **45**, 1543–1558.
29. Wells, J. A. & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, **450**, 1001–1009.
30. Assi, S. A., Tanaka, T., Rabbitts, T. H. & Fernandez-Fuentes, N. (2010). PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res.* **38**, e86.
31. Arkin, M. R., Randal, M., DeLano, W. L., Hyde, J., Luong, T. N., Oslob, J. D. *et al.* (2003). Binding of small molecules to an adaptive protein–protein interface. *Proc. Natl Acad. Sci. USA*, **100**, 1603–1608.
32. Cheng, A. C., Coleman, R. G., Smyth, K. T., Cao, Q., Soulard, P., Caffrey, D. R. *et al.* (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.* **25**, 71–75.
33. Bourgeas, R., Basse, M. J., Morelli, X. & Roche, P. (2010). Atomic analysis of protein–protein interfaces with known inhibitors: the 2P2I database. *PLoS One*, **5**, e9598.
34. Davis, F. P. & Sali, A. (2010). The overlap of small molecule and protein binding sites within families of protein structures. *PLoS Comput. Biol.* **6**, e1000668.
35. Davis, F. P. (2011). Proteome-wide prediction of overlapping small molecule and protein binding sites using structure. *Mol. Biosyst.* **7**, 545–557.
36. Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797.
37. Xu, Q. & Dunbrack, R. L., Jr (2011). The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.* **39**, D761–D770.
38. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.* **5**, e1000605.
39. Shoemaker, B. A., Zhang, D., Thangudu, R. R., Tyagi, M., Fong, J. H., Marchler-Bauer, A. *et al.* (2010). Inferred Biomolecular Interaction Server—a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.* **38**, D518–D524.
40. Thangudu, R. R., Tyagi, M., Shoemaker, B. A.,

Bryant, S. H., Panchenko, A. R. & Madej, T. (2010). Knowledge-based annotation of small molecule binding sites in proteins. *BMC Bioinformatics*, **11**, 365.

41. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, D301–D303.
42. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A. *et al.* (2011). DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* **39**, D1035–D1041.
43. Ofran, Y. & Rost, B. (2007). Protein–protein interaction hotspots carved into sequences. *PLoS Comput. Biol.* **3**, e119.
44. Bashton, M., Nobeli, I. & Thornton, J. M. (2008). PROCOGNATE: a cognate ligand domain mapping for enzymes. *Nucleic Acids Res.* **36**, D618–D622.
45. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
46. Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. (2008). Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.
47. Grossmann, S., Bauer, S., Robinson, P. N. & Vingron, M. (2007). Improved detection of overrepresentation of Gene-Ontology annotations with parent–child analysis. *Bioinformatics*, **23**, 3024–3031.
48. Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305.
49. Wurtele, M., Jelich-Ottmann, C., Wittinghofer, A. & Oecking, C. (2003). Structural view of a fungal toxin acting on a 14-3-3 regulatory complex. *EMBO J.* **22**, 987–994.
50. Rose, R., Erdmann, S., Bovens, S., Wolf, A., Rose, M., Hennig, S. *et al.* (2010). Identification and structure of small-molecule stabilizers of 14-3-3 protein–protein interactions. *Angew. Chem., Int. Ed. Engl.* **49**, 4129–4132.
51. Choi, J., Chen, J., Schreiber, S. L. & Clardy, J. (1996). Structure of the FKBP12–rapamycin complex interacting with the binding domain of human FRAP. *Science*, **273**, 239–242.
52. Huse, M., Chen, Y. G., Massague, J. & Kuriyan, J. (1999). Crystal structure of the cytoplasmic domain of the type I TGF beta receptor in complex with FKBP12. *Cell*, **96**, 425–436.
53. Peterson, F. C., Penkert, R. R., Volkman, B. F. & Prehoda, K. E. (2004). Cdc42 regulates the Par-6 PDZ domain through an allosteric CRIB–PDZ transition. *Mol. Cell*, **13**, 665–676.
54. Antonyuk, S., Strange, R. W. & Hasnain, S. S. (2010). Structural discovery of small molecule binding sites in Cu–Zn human superoxide dismutase familial amyotrophic lateral sclerosis mutants provides insights for lead optimization. *J. Med. Chem.* **53**, 1402–1406.
55. Chen, J., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S. *et al.* (2003). MMDB: Entrez's 3D structure database. *Nucleic Acids Res.* **31**, 474–477.
56. Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H. *et al.* (2009). CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.* **37**, D205–D210.
57. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385.
58. Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J. & Bryant, S. H. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37**, W623–W633.