RESOURCE

ProtCID: A tool for hypothesis generation of the structures of protein interactions

Qifang Xu, Roland L. Dunbrack, Jr.*

Fox Chase Cancer Center

333 Cottman Avenue

Philadelphia PA 19111

* Roland.Dunbrack@fccc.edu

Abstract

Interaction of proteins with other molecules is central to their ability to carry out biological functions. These interactions include those in homo- and heterooligomeric protein complexes as well as those with nucleic acids, lipids, ions, and small molecules. Structural information on the interactions of proteins with other molecules is very plentiful, and for some proteins and protein domain families, there may be 100s or even 1000s of available structures. While it is possible for any biological scientist to search the Protein Data Bank and investigate individual structures, it is virtually impossible for a scientist who is not trained in structural bioinformatics to access this information across all of the structures that are available of any one extensively studied protein or protein family. Furthermore, the annotation of biological assemblies in the PDB is only correct for about 80% of entries in the PDB, and there is therefore a great deal of biological information that is present in protein crystals but not annotated as such. We present ProtCID (the Protein Common Interface Database) as a webserver and database that makes comprehensive, PDB-wide structural information on the interactions of proteins and individual protein domains with other molecules accessible to scientists at all levels. In particular, in the paper we describe the utility of ProtCID in generating structural hypotheses from available crystal structures that may not be readily apparent even to experienced structural biologists. One example is a homodimer of HRAS, KRAS, and NRAS (the α4-α5 dimer) that has recently been experimentally validated for all three proteins, and for which we are able to present structural data from 16 different crystal forms and 108 PDB entries, including a structure of the NRAS dimer for the first time. We also present a provocative hypothesis for the dimer structure of the second bromodomain of BET histone-binding proteins (including human BRD2, BRD3, BRD4, BRDT) that is present in every structure of these proteins available in the PDB. We also show how ProtCID can be utilized to extend existing experimental data on some proteins to many other proteins in the same family or even much larger superfamilies and to identify structural information for multi-protein complexes and hub proteins.

All proteins function via interactions with other molecules, including nucleic acids, small molecular ligands, ions, and other proteins in the form of both homo- and heterooligomers. How such interactions occur and defining their role in protein function are the central goals of structural biology. For many proteins and protein families, there is abundant structural information available in the form of structures determined by X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy and deposited in the Protein Data Bank (PDB). The number of structures for a protein and its homologues can reach into the hundreds or thousands. Some proteins occur in different structural forms during their functional lifetimes, depending on conditions such as pH, phosphorylation state, interaction partners, and cofactors. For any protein system of interest, it is most valuable to understand the structure in any and all of its structural and functional states, including different homooligomeric states and functional interactions with nucleic acids, ligands, and other proteins. To accomplish this, it is often necessary to examine many available structures of a protein and even structures of its homologues. However, the process is very challenging and time consuming even for scientists trained in bioinformatics. It is virtually impossible when there are dozens or hundreds of available structures.

To address this issue, we have previously developed PDBfam[1], which assigns protein domains defined by Pfam[2] to all structures in the PDB. It enables a user to browse a page of PDB entries for any particular domain family. Each PDB entry listed on a PDBfam page includes the Uniprot identifier, species, and Pfam architecture of the chain that contains the query domain as well as the same features of other protein partners in the same entries. It also includes information on other bound partners such as peptides, nucleic acids, and ligands. In this way, it is possible to rapidly identify structures within a protein family that will provide information on interactions that are critical to biological function.

Of central importance to the utility of experimental structures is the accuracy of annotations. Authors of crystal structures are required to deposit a ‘biological assembly’ into the

PDB, which is what they believe to be the biologically relevant oligomeric form present in the crystal. This is in contrast to the asymmetric unit, which is the set of coordinates used to model the unit cell and the crystal lattice when copied and placed with rotational and translational symmetry operators. The author-deposited biological assembly is different from the asymmetric unit (ASU) for about 40% of crystal structures in the PDB[3]. When it is different, roughly half the time the biological assembly is larger than the ASU (i.e., made from parts or all of multiple copies of the ASU), and half the time it is smaller than the ASU (i.e., a sub-assembly of the ASU). Various authors have estimated the accuracy of the biological assemblies in the PDB in the range of 80-90%[4, 5].

For some entries, the PDB provides additional biological assemblies derived from the Protein Interfaces, Surfaces and Assemblies (PISA)[4] server from the EBI that are predicted to be stable in solution based on calculations of thermodynamic stability of the complexes. Several servers analyze interfaces in either the asymmetric units and/or the deposited biological assemblies of PDB entries[6-11], and some are intended to predict which interfaces may be biologically relevant by measuring conservation scores and physical features and using machine learning predictors[12-14]. Interactions with peptides, nucleic acids, and ligands have also been analyzed and presented in several webservers and databases[9, 15-21].

A common approach for identifying biological interactions of molecules within protein crystals is to compare multiple crystal forms of the same or related proteins. Each crystal form will have different symmetry and different non-biological interfaces between proteins and with crystallization ligands (sulfate ion, glycerol, etc.), while in most cases the biological interactions will be shared between them. We have shown that if a homodimeric or heterodimeric interface is present in a number of crystal forms, especially when the proteins in the different crystals are homologous but not identical, then such interfaces are very likely to be part of biologically relevant assemblies[5]. To enable this form of analysis, we previously developed the Protein Common Interface Database (ProtCID) which compares and clusters the interfaces of pairs of

full-length protein chains with defined domain architectures in different entries and crystal forms in the PDB[22] (e.g., homodimers of one Pfam architecture such as (SH2)_(Pkinase), or heterodimers of two different Pfam architectures such as (C1_set) and (MHC_I)_(Ig_3) present in Class I MHC proteins). ProtCID has been very useful in identifying biologically relevant interfaces and assemblies within crystals[23-26], including those that may not have been annotated in the biological assemblies provided to the PDB by the authors of each structure. ProtCID allows users to download coordinates the PyMol scripts for visualizing all available interfaces

In this paper, we extend the ProtCID approach to clustering the interactions of individual domains within and between multi-domain proteins and between protein domains and peptides, nucleic acids, and small-molecular ligands. We focus in particular on how ProtCID can be used to generate hypotheses of which molecular interactions within crystals provide biologically relevant information that can be tested experimentally. The inclusion of interactions between individual domains greatly extends our ability to generate hypotheses about the functional interactions of proteins. We show examples of both chain-level and domain-level ProtCID clusters for some experimentally validated, biologically relevant protein-protein interactions that were in some way challenging to identify in the biological literature. This is especially true of weaker interactions within homooligomers, which are very difficult to distinguish from crystallization-induced interactions. There have been several cases where structures containing homodimers of proteins had existed in the Protein Data Bank (PDB) for many years before the homodimers were recognized as biologically relevant structures. Examples of this include the asymmetric homodimerization of the EGFR kinase domain involved in kinase activation[27], the homodimer of cytosolic sulfotransferases involved in half-sites reactivity[28, 29], and very recently a homodimer common to both H-RAS[30] and K-RAS[31]. We show cases where at least one protein in a ProtCID cluster has been well validated as a homodimer but the cluster contains other proteins in the same family which have not been described as such. In this way, we generate hypotheses of how these proteins function as oligomers.

Many proteins interact with peptide segments from other proteins, often when these segments are from intrinsically disordered regions[32]. Thus, when attempting to define how two proteins interact, it is important to identify domains in one or both partners that might bind short segments of the other protein via a known peptide-binding domain. We have clustered peptide/Pfam-domain interactions in the PDB to identify a subset of domain families that we refer to as "professional peptide-binding domains" (PPBDs). The function of PPBDs in most contexts is to bind peptide segments from other proteins. Peptides bind to PPBDs in a similar manner across each family, often with a specific binding motif. Examples include SH2[33], BRCT[34], and PDZ domains[35]. These structurally characterized PPBDs are explicitly identified in ProtCID during searches of how two or more proteins might interact structurally.

Many protein domain families bind to nucleic acids, usually in ways that are well conserved across each family. Analysis of the available structures in large families has enabled the identification of how DNA or RNA sequence specificity is encoded in the protein sequence and structure[36, 37]. This can even be extended to domain families within larger superfamilies, where some member families do not have any structures bound to DNA. Like other classifications of protein domains, Pfam is organized into superfamilies (called clans in Pfam), which sometimes contain very remotely related protein families of similar structure. Thus, the existence in ProtCID of domain/nucleic-acid interactions for one protein family can be used to develop hypotheses for the structures of other protein families within the same superfamily. We show some examples of this approach.

Binding sites of biological ligands are usually conserved within protein families, and the bioinformatics analysis of these sites with bound ligands has enabled the development of specific inhibitors for many proteins[38-40]. We have clustered ligands bound to each Pfam domain in the PDB using a metric based on volume overlap of the ligands. This enables the rapid identification of common binding sites of biological ligands and inhibitors as well as pockets in proteins that bind molecular fragments present in the solutions used to crystallize proteins.

Many proteins act as hub proteins and it can be difficult to develop hypotheses of how such proteins might function. We have enabled a new search feature in ProtCID with which a user can identify possible domain-domain and PPBD/peptide interactions that are possible between a hub protein and its partners. Interacting partners can be obtained from several databases, such as IntAct and String, and input to ProtCID. Many proteins participate in large complexes of many different proteins. ProtCID identifies the Pfam domains in a list of input Uniprot identifiers and then provides a list of potential domain-domain and PPBD/peptide interactions among these proteins that are represented in homologous proteins in the PDB. These utilities might enable the identification of direct protein-protein interactions that might occur in large multi-subunit protein complexes.

RESULTS

Generating hypotheses for oligomeric protein assemblies with ProtCID

The ProtCID database contains information on four types of interactions: protein-protein interactions at the chain level, domain-domain interactions, domain-peptide interactions, and the interactions of domains with nucleic acids and ligands. Domain interactions can be between domains of the same Pfam or different Pfams and can be interchain or intrachain. Generating hypotheses for protein interactions by observing them in multiple crystals of homologous proteins requires grouping proteins in the PDB into homologous families at the domain and chain levels and clustering them based on an appropriate metric. To accomplish this, we utilize the Pfam database of multiple sequence alignments of known protein domain families (e.g., SH2, Pkinase) and assign Pfam domains to all sequences in the PDB. Currently, 8,636 Pfams are represented within the PDB in our database called PDBfam[1]. Each PDB chain is annotated by a Pfam architecture as the ordered sequence of Pfams along the chain, e.g. '(SH3)_(SH2)_(Pkinase)'. Statistics on the number of PDB entries, Pfam domains, the interactions and clusters are provided in Supplementary Tables 1 and 2. We compare interfaces

of homologous proteins via a Jaccard-index-like metric[41] that compares the contacts in one interface with homologous contacts in the other interface.

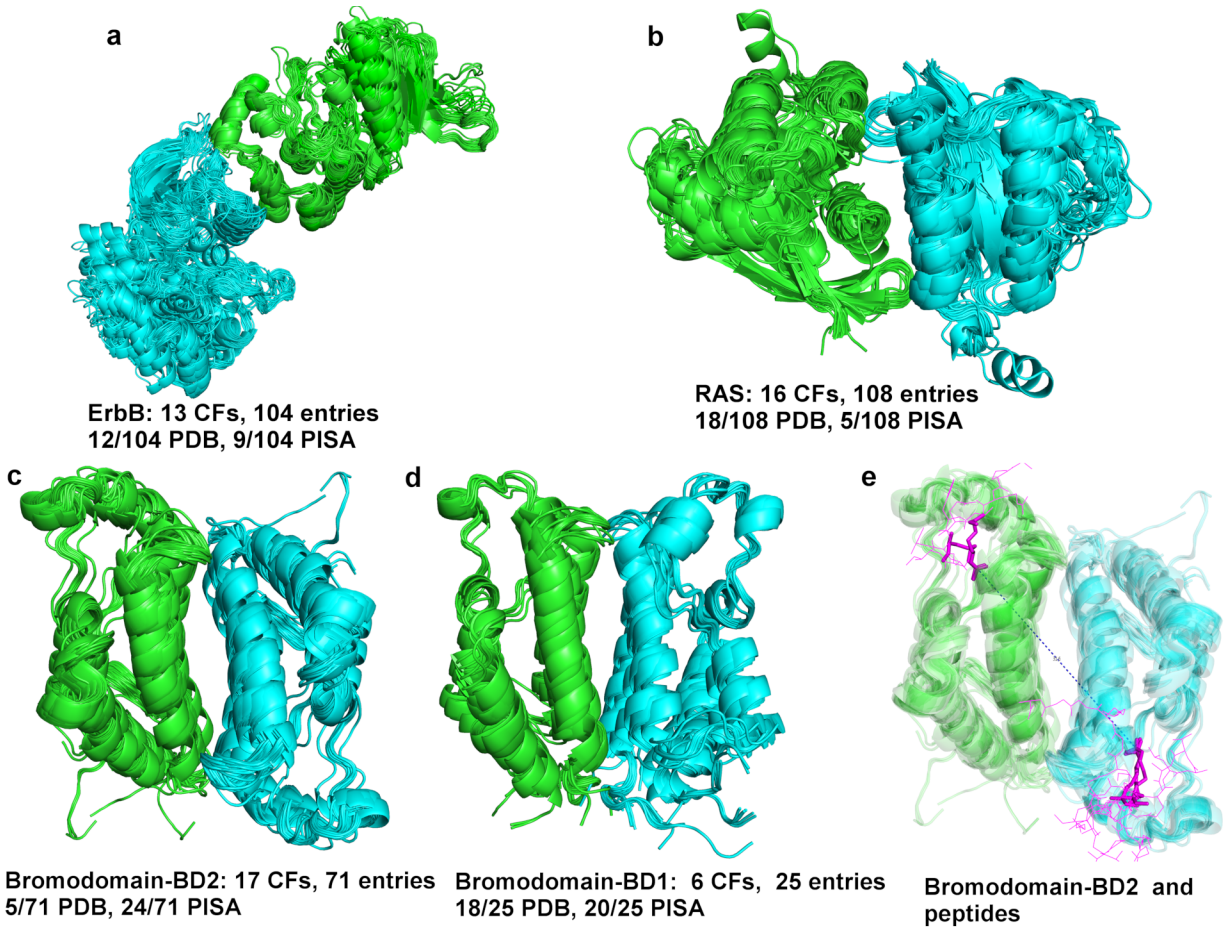

Figure 1 | Hypothesis generation for homodimers. (**a**) The asymmetric dimer of ErbB proteins is evident in a large ProtCID cluster of 13 crystal forms (CFs) and 104 PDB entries, comprising homodimers of EGFR (10 of 30 CFs, 98/140 entries), ErbB2 (1/1 CF, 1/1 entry), ErbB4 (2 CFs, 2 entries), and heterodimers of ErbB3 (in green) and EGFR (in cyan) (1 CF, 3 entries). The chain in green is the activating monomer, the chain in cyan is the activated monomer. The caption in the image indicates that the PDB biological assembly contains the dimer in 12 of the 104 crystal structures with the dimer and 9 of 104 PISA assemblies. (**b**) A proposed Ras α4-α5 dimer occurs in crystals structures of human HRAS (13/35 CFs, 89/145 entries), rat HRAS (1/1 CF, 1/1 entry), human KRAS (3/32 CFs, 13/82 entries), human NRAS (1/3 CF, 1/3 entry), human RAB11B (1/3 CF, 1/3 entry) and mouse RND3 (RhoE) (1/1 CF, 1/1 entry). (**c**) A proposed homodimer found in all structures of the BD2 bromodomain of BET proteins.. This BD2 head-tail (H-T) dimer occurs in 17 crystal forms and 71 PDB entries, and in all four BET proteins: human BRD2 (11/11 CFs, 35/35 entries), BRD3 (3/3 CFs, 5/5 entries), and BRD4 (7/7 CFs, 27/27 entries), and mouse BRD4 (2/2 CFs, 3/3 entries) and BRDT (1/1 CF, 1/1 entry). (**d**) BET BD1-bromodomain head-to-head (H-H) homodimer. This dimer occurs in 6/28 CFs and 25/76 PDB entries: human BRD2 (3/18 CFs, 20/60 entries), BRD3 (1/6 CF, 1/9 entry), and BRDT (2/4 CFs, 3/5 entries) and mouse BRDT (1/2 CF, 1/2 entry). (**e**) Peptide substrates of BD2 H-T homodimers (PDB: 2E3K, 2WP1, 4KV4). The peptides are shown in lines and colored in magenta. The acetyl-lysine (ALY) residues are shown in sticks and colored in magenta. The average distance between ALY residues of two peptides is about 31 Å.

A primary goal of ProtCID is to generate hypotheses of the structures of oligomeric protein assemblies that may not be readily obvious to authors of crystal structures. Many such structures are due to weakly interacting dimers that are facilitated by attachment to the membrane or by scaffolding by other proteins or nucleic acids. One verified case that inspired the development of ProtCID is the small dimer interface of cytosolic sulfotransferases involved in half-sites reactivity of these enzymes[29], and initially identified by cross-linking, protease digestion, and mass spectrometry[42]. It is currently observed in 24 crystal forms, 14 different proteins, and 50 PDB entries[28]. To demonstrate the utility of ProtCID, we present several examples of this phenomenon, both confirmed experimentally in the literature and new but provocative hypotheses (Figure 1).

In 2006, Kuriyan and colleagues discovered the biological relevance of an asymmetric dimer of the EGFR kinase domain[27] that was already present in the PDB at that time. Extensive experimentation indicated that this interface between the C-terminal domain of one monomer and the N-terminal domain of another monomer served to activate the latter. Later structures of ErbB2[43], ErbB4[44] and a heterodimer of EGFR and ErbB3 were also noted to contain the same dimer[45]. This dimer was unexpected because most protein homodimers are symmetric or "isologous", and asymmetric or "heterologous" dimers have the risk of forming polymer chains as true polymeric chains do such as actin; EGFR probably cannot do this because of crowding at the membrane.

The biological effect of this dimer serves as an example of how ProtCID might lead to such a hypothesis. In ProtCID, there is a domain- and chain-level cluster of the (Pkinase_Tyr) Pfam that contains the asymmetric dimer from 10 crystal forms (CFs) and 98 PDB entries of EGFR, one CF and entry of ErbB2, one CF and 3 entries of an ErbB3/EGFR heterodimer, and 2 CFs and 2 entries of ErbB4 (Figure 1a; Supplementary Figure 1; a list of ErbB PDB entries that do and do not contain the dimer is provided in Supplementary Data 1). A total of 100 of 104 structures that contain the asymmetric ErbB dimer in Figure 1a contain an active kinase in both

positions (91 cases) or only the activated domain (9 cases). The remaining 4 structures contain only inactive kinase domains. Fifty out of 51 structures of EGFR that do not contain the asymmetric dimer consist only of inactive kinase domains (Supplementary Table 3). As a testament to how difficult it is for biophysical calculations and sequence conservation in interfaces to determine biological assemblies, PISA predicts the asymmetric dimer of these structures as biological in only 6 of the 98 EGFR entries and all 3 of the heterodimer entries. EPPIC predicts only the heterodimers as biological assemblies; it predicts that all of the homodimers in this ProtCID cluster are monomers[46]. These and other methods typically do not support the prediction of asymmetric homooligomeric assemblies (those that are not Cn or Dn symmetric)[3]. ProtCID may therefore be a suitable tool for identifying such assemblies.

RAS proteins form oligomeric structures at the plasma membrane, where they are anchored by palmitolyation and farnesylation[47]. In ProtCID, there is a large domain-based cluster of interfaces for the Ras Pfam comprising 16 crystal forms and 108 PDB entries. The cluster includes structures of HRAS (14 CFs, 92 entries), KRAS (3 CFs, 13 entries), NRAS (1 CF, 1 entry), RAB-11B (1 CF, 1 entry), and mouse Rnd3 (1 CF, 1 entry) (Figure 1b, Supplementary Figure 2a; a list of HRAS, KRAS, and NRAS PDB entries that do and do not contain the dimer is provided in Supplementary Data 1). The structures are symmetric dimers involving helices α4 and α5, with an average surface area of 797 $Å^2$, which is a moderately sized interface for a homodimer. PISA recognizes the dimer as biological in only 5 of 108 entries; EPPIC does not recognize the α4-α5 dimer as part of a biological assembly for any of these PDB entries. We also find a smaller cluster (5 crystal forms) of a beta dimer, which has been studied by Muratcioglu et al.[48] By contrast, we do not find the α3-α4 dimer implicated in the same study in any PDB entry.

The α4-α5 dimer in our ProtCID cluster has been implicated as a biologically relevant assembly for NRAS, KRAS, and HRAS. Spencer-Smith et al. found that a nanobody to the α4-α5 surface, as determined by a co-crystal structure of the nanobody with HRAS (PDB: 5E95),

disrupted HRAS nanoclustering and signaling through RAF[49]. Via manual search, they identified the α4-α5 dimer in 74 of 80 active structures of HRAS but in none of 33 inactive structures of HRAS. Our α4-α5 cluster contains 92 structures of HRAS, 81 of them (88%) with GTP or guanine triphosphate analogs, 9 of them with GDP, and two without a ligand (Supplementary Table 4). By contrast, 27 of 54 (50%) HRAS structures that do not contain the α4-α5 dimer are bound with GTP or a triphosphate analog. Similarly, Ambrogio et al. very recently identified the α4-α5 dimer in their crystal of KRAS[31], and used mutagenesis to show that disruption of this dimer abolished the ability of mutant KRAS to drive tumor growth and the ability of wild-type KRAS to inhibit mutant KRAS. In the 13 KRAS structures in our cluster, 5 are bound with GTP or a triphosphate analogue and 8 are GDP bound. Sixty-nine KRAS structures do not contain the dimer.

Güldenhaupt et al. performed ATR-FTIR (attenuated total reflection Fourier transform infrared) experiments to determine that the α helical content of NRAS (primarily the parallel helices α3, α4, and α5) is oriented perpendicular to a POPC membrane[50]. They identified the α4-α5 dimer from multiple crystal forms of HRAS, and performed MD simulations of an NRAS monomer and a modeled NRAS α4-α5 dimer, and found that only the dimer maintained an angle between the helices and the membrane normal close to the experimentally determined value of 23°. FRET experiments demonstrated a distance of 46 Å between donor (MANT) and acceptor (TNP) derivatives of GDP bound to NRAS. With ProtCID, we have identified an α4-α5 dimer in a crystal of GNP-bound NRAS (PDB: 5UHV), that was unrecognized by the authors of this structure[51]. In this structure, the GNP-GNP distance is about 36 Å, which is consistent with the MANT/TNP distance which would be somewhat larger. Whether the Rnd3 and RAB-11B structures are biologically relevant remains a hypothesis in need of testing.

Finally, it has been noted that the α4-α5 Ras homodimer is consistent with binding of Ras effector domains to the surface on the opposite side of the protein from the homodimer interface[52]. Indeed, several of the crystals that contain the α4-α5 dimer also contain Ras binding

partners, and therefore consist of Ras/Ras-effector heterotetramers, including the RBD domain of RAF1 kinase, the RasGEF domain of RAS guanyl releasing protein and the RA domains of Phospholipase C epsilon and Rassf5 (Supplementary Figure 2d). Other Ras crystals that do not contain the α4-α5 dimer but do contain heterodimeric partners, show that the partners could bind to the α4-α5 dimer. These include the RA domains of GRB14 and RALGDS, the RBD domain of APOA1, the PI3Krbd domain of PK3CG, the RasGAP domain of RasGAP and the RasGEF domains of Son-of-Sevenless (Supplementary Figure 2e).

Bromodomains are modules that bind acetylated lysine residues, primarily but not exclusively in histones[53]. The domain is a four-helix bundle consisting (in sequence order) of αZ, αA, αB, and αC. We examined both the domain- and chain-based clusters of the "Bromodomain" Pfam in ProtCID. The largest chain-based cluster contains 17 crystal forms and 71 PDB entries out of a total of 71 crystal forms and 835 entries containing proteins with the (Bromodomain) Pfam architecture. This head-to-tail symmetric dimer has an interface consisting of the αB and αC helices, with an average surface area of 741 $Å^2$ (a list of PDB entries that contain the dimer is provided in Supplementary Data 1). The Pfam for bromodomains is shorter than the observed domains in structures of these proteins by about 28 amino acids in the C-terminal αC helix. Since the αC helix makes up a substantial portion of the interface in this cluster, the domain-based interfaces fall below our cutoff of 150 $Å^2$ in most of the structures, leaving a related domain-based cluster of only 6 CFs and 32 PDB entries. The distinction highlights the utility of clustering full-length chains as well as Pfam-defined domains to compensate for shortcomings in Pfam's definitions of some domains.

All of the proteins in this chain-based cluster are members of the BET (Bromodomain and Extra-Terminal domain) family of bromodomain proteins. These proteins, BRD2, BRD3, BRD4, and BRDT, contain two tandem bromodomains, BD1 and BD2, followed by a small Extra-Terminal (ET) domain towards the C-terminus. In between BD2 and the ET domain there is a coiled-coil "Motif B" that is associated with homo- and heterodimerization[54]. BET proteins

bind to acetylated sites within histones via their bromodomains and serve as scaffolds for transcriptional machinery proteins via their ET domains[53]. The chain-based cluster we have identified consists solely of BD2 domains from BET proteins (Figure 1c, Supplementary Figure 3a), including human BRD2, BRD3, and BRD4 as well as mouse BRD4 and BRDT. It is an important observation that *every* crystal of a BD2 domain from a BET protein in the PDB contains this head-to-tail dimer.

In addition to motif B, the first bromodomain of BET proteins has also been shown to dimerize in solution through dynamic light scattering, cross-linking, and co-immunoprecipitation experiments of tagged BD1 constructs[55]. The same authors determined the crystal structure of BD1 of human BRD2, and identified a head-to-head symmetric dimer as the likely biological assembly, which was verified by mutation of residues in the interface. Mutants unable to form the dimer bound to H4K12ac with drastically reduced affinity compared to the wildtype. This dimer is distinct from the BD2 dimer we found in the largest chain-based cluster in ProtCID. Instead, the BD1 head-to-head cluster is the fifth largest chain-based cluster, comprising 6 crystal forms and 25 PDB entries (Figure 1d, Supplementary Figure 3b), including BD1 of human BRD2 and BRD3 and both mouse and human BRDT. The average surface area is 1015 $Å^2$. The biological assemblies in the PDB contain this assembly for 18 of the 25 entries in the cluster, while PISA recognizes 20. While the head-to-tail BD2 dimer contains all structures of BD2 proteins in the PDB, the head-to-head dimer cluster contains only 25 of 235 PDB entries with BD1 domains of BET proteins.

While there is strong evidence for dimerization of full-length BET proteins, there is not yet specific evidence that the BD2 domains homodimerize in vitro or in vivo. Huang et al found that the BD2 of BRD2 expressed as a single-domain protein did not form stable dimers in solution by gel-filtration and GST-pulldown analysis[56]. However, there is a possibility that if the BD1 domains and motif B of a BET protein homodimerize, then the BD2 domain that sits in between them in the sequence may form homodimers given the increased concentration

induced by the dimerization of the flanking domains. The dimer we have identified with ProtCID is a strong candidate for the form of a BD2 dimer, if it exists, and serves as an example of how ProtCID is able to generate credible and testable hypotheses for the formation of weak or transient interactions of proteins. Several of these structures contain bound acetylated histone peptides. The two peptides bound to the dimer are about 30 Å apart, indicating that if this dimer is biologically active, then the dimer binds peptides from different histone chains either within the same nucleosome or in adjacent nucleosomes (Figure 1e).

Extending dimers with established biological activity to other family members whose crystals contain the same dimer

The observation of similar interfaces in crystals of homologous proteins can be used to utilize experimental data available on one protein to generate hypotheses for other members of the same family. ProtCID enables this kind of inference in an easily accessible way. One intriguing example is observed for the Pfam domain (Pkinase_Tyr). This Pfam includes tyrosine kinases and most of the tyrosine-kinase-like (TKL) family of kinases. The catalytic activity of TKL kinases BRAF and RAF1 is regulated by side-to-side homodimers of their kinase domains[57, 58], both of which are well represented in the largest ProtCID cluster for Pkinase_Tyr. However, the cluster also includes 5 other kinases: TKL-family kinases RIPK2, MLKL, and *Arabidopsis* CTR1, and Tyrosine kinases CSK and ITK (Figure 2a). In ProtCID, this cluster contains 99 structures in 31 crystal forms with minimum sequence identity 17% and average surface area ~1,000 $Å^2$ (a list of PDB entries that contain the dimer is presented in Supplementary Data 1). The structures of RAF-like kinase CTR1 (PDB: 3PPZ and 3P86) have been described as similar to the BRAF kinase[59].

Newly available experimental data on RIPK2 confirms the significance of this ProtCID cluster. Pellegrini et al. have very recently shown by sedimentation velocity analytical ultracentrifugation that RIPK2 (Receptor-interacting serine/threonine-protein kinase 2) forms a

homodimer in solution[60]. Activation and trans-autophosphorylation were shown by mass spectrometry to be tightly coupled with formation of the back-to-back dimer similar to that of BRAF, found in crystal structures solved by the same authors. Mutations at the dimer interface verified that this dimer, found in four different crystal forms of RIPK2 in the ProtCID cluster, is similar to the homodimer of BRAF.

The presence of other kinases in the cluster, however, allows us to form hypotheses about their dimerization structures for which there is not yet experimental evidence. Human mixed lineage kinase domain-like protein (MLKL) is a TKL-family pseudokinase that functions as a substrate of Receptor-interacting serine/threonine kinase 3 (RIPK3) in necroptosis[61]. The cluster shows that MLKL can be dimerized in the same way as BRAF, CRAF, and RIPK2. The structure of human MLKL (PDB: 4M67) is annotated by the authors as a monomer[62] but the crystal contains an MLKL homodimer very similar to BRAF. The asymmetric unit of 4M69 is a mouse RIPK3-MLKL heterodimer[62]. After we applied symmetry operators on the 4M69 asymmetric unit, there is one RIPK3 dimer and one MLKL dimer which are integrated by the RIPK3-MLKL heterodimer interface to build an A2B2 heterotetramer (Supplementary Figure 4). Both the mouse MLKL dimer and the RIPK3 dimer in 4M69 are similar to the BRAF dimer. The MLKL dimer is part of the cluster, but the RIPK3 dimer in 4M69 does not appear in the cluster because the sequence of RIPK3 is closer to Pfam domain (Pkinase) than (Pkinase_Tyr). The RIPK3 dimer occurs in 10 CFs and 65 PDB entries in a (Pkinase) domain-based cluster. The structure of mouse RIPK3 alone (PDB: 4M66) also contains the same dimer. RIPK3 dimerization is essential for recruitment of MLKL[63]. After phosphorylation by RIPK3, MLKL forms an oligomer that allows MLKL to move to intracellular membranes, disrupting membrane integrity resulting in necrotic death[64]. Very recently, Raju et al. found that mutations in the BRAF-like interface of RIPK3 impaired dimerization and necroptosis[65]. The BRAF-like homodimers of all these structures are not annotated as such in the PDB or described by the authors[62].

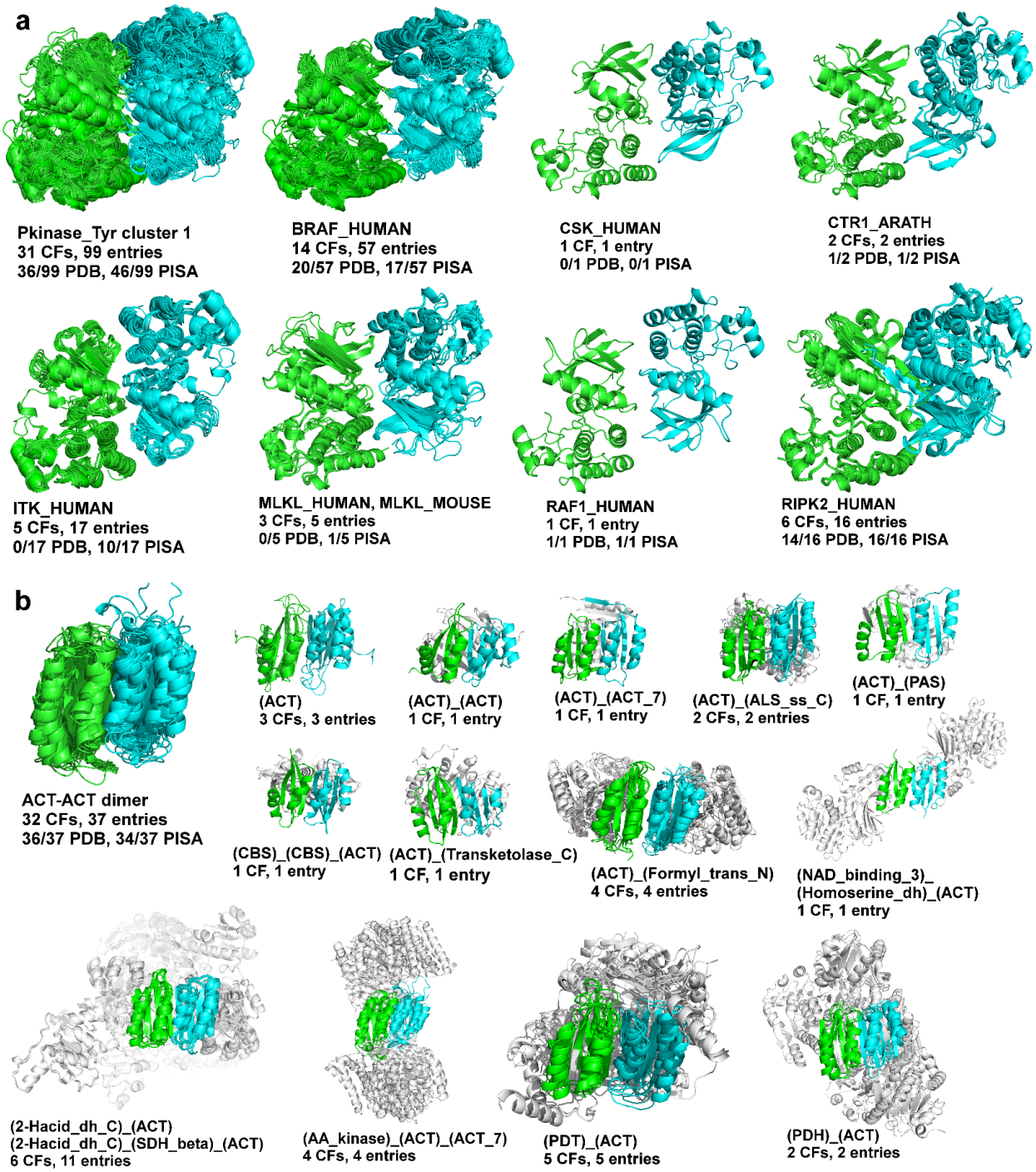

Figure 2 | Generating hypotheses for homodimers via structural similarity and homology. (**a**) Common domain interfaces in Pkinase_Tyr (PF07714) cluster 1 in ProtCID. The cluster contains 31 CFs and 99 entries, and exists in 36 PDB biological assemblies and 46 PISA assemblies. It contains 8 different kinases: human BRAF, CSK, ITK, MLKL, RIPK2, and RAF1; mouse MLKL; and *Arapidopsis* CTR1. (**b**) The ACT domain occurs in proteins with many different Pfam architectures from 32 crystal forms and 37 PDB entries. ACT domains are colored in green and cyan. The other domains are colored in light gray. The cluster was used to generate a hypothesis of how the ACT domain in Phenylalanine hydroxylase functions prior to the determination of its structure, which confirmed the hypothesis (PDB: 5FII).

Interleukin-2-inducible kinase (ITK) plays an important role in the activation of T-cells in

the immune response. ITK has been found to form homodimers at the plasma membrane, which does not require the PH, proline-rich, SH2, or SH3 domains[66]. Although catalytic activity is not required for dimerization[66], the only remaining domain that may be involved in dimerization is the kinase domain. The ITK kinase dimers, found in five different crystal forms in the BRAF ProtCID cluster, are a reasonable hypothesis for the mechanism of homodimerization of ITK at the membrane. These ITK dimers are not discussed or annotated by the authors of these structures[67-74].

Many enzymes have regulatory domains that are involved in dimerization or higher-order oligomerization. ACT domains (***A***spartate kinase, ***C***horismate mutase, ***T***yrA domains) are present in the sequences of a number of enzymes, and typically bind single amino acids at their dimer interface. The largest cluster of (ACT)/(ACT) domain-level interfaces (Figure 2b) comprises 14 different multi-domain Pfam architectures, present in 37 PDB entries and 32 crystal forms. The domain dimer is well annotated as part of the biological assemblies of these proteins. We used this cluster to generate a hypothesis that the ACT domain of human phenylalanine hydroxylase (PAH) would form the same dimer in response to binding of phenylalanine and act as a mechanism of enzyme activation[24]. While the structure of full-length activated PAH has not been determined, a recent structure (PDB: 5FII) of the ACT domain of human PAH with bound Phe contains the same ACT dimer of the ProtCID cluster. Inherited mutations at the domain-domain interface of the PAH ACT domain are associated with the autosomal recessive disease phenylketonuria[75]. We can hypothesize that the same ACT dimer is associated with activation in human tryptophan 5-hydroxylases 1 and 2 and tyrosine 3-hydroxylase, which are homologous to phenylalanine hydroxylase and contain similar domain architectures. This ACT dimer commonly occurs in other ACT Pfams in the same Pfam clan, including ACT_3, ACT_4, ACT_5, ACT_6 and ACT_7 (Supplementary Figure 5 and Supplementary Table 5).

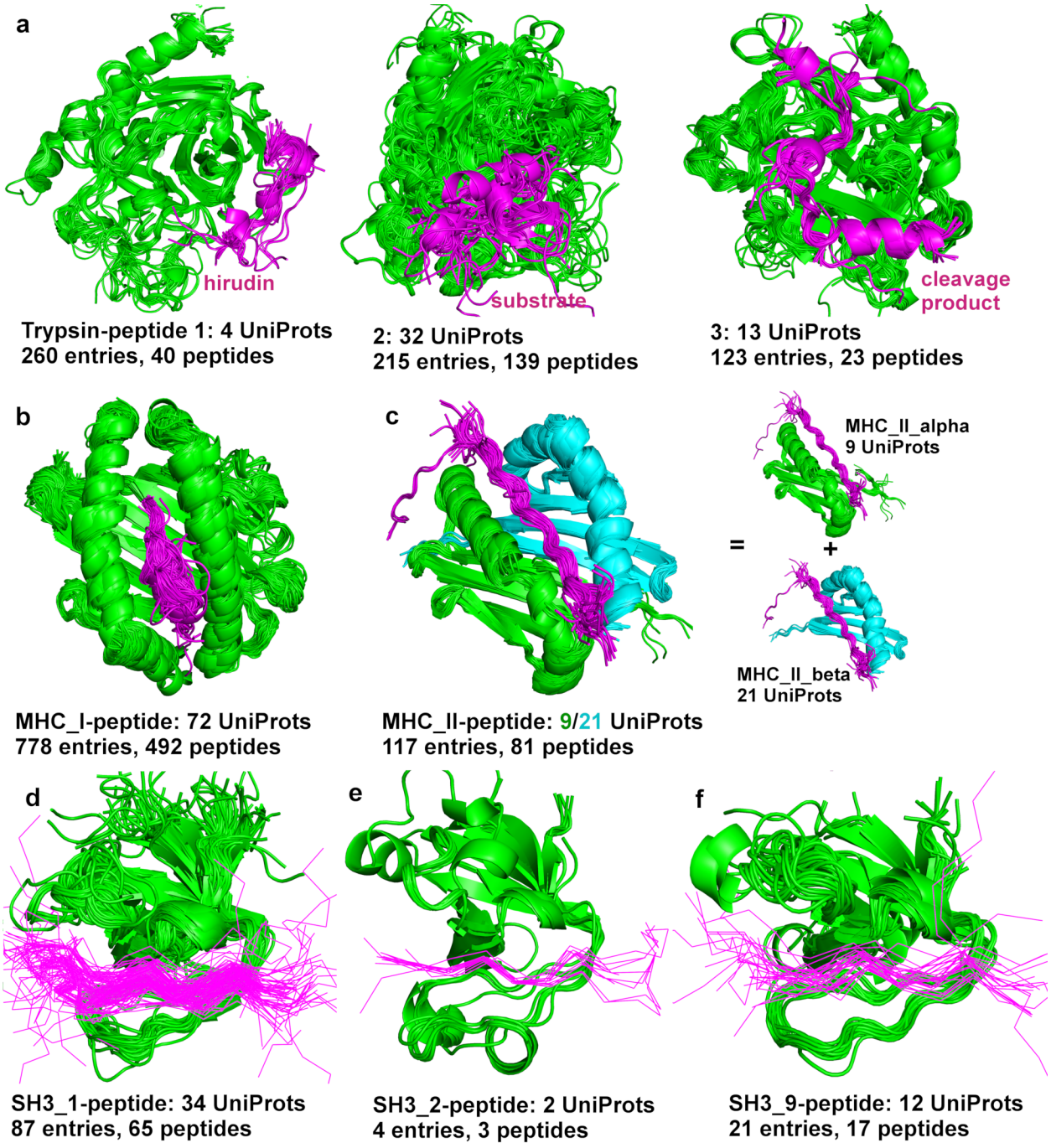

Figure 3 | Peptide-binding Pfam domains in ProtCID. (**a**) Three major Trypsin-peptide interface clusters. The number of peptides indicates the number of unique peptide sequences; the sequences of each cluster can be found on each cluster page on ProtCID. The first cluster mostly constists of human thrombin bound to peptides from hirudin from leeches. The peptides of the second cluster is a large set of substrates bound to the active sites of trypsin-like proteases. The peptides of the third cluster are cleavage protducts, created by internal cleavage by activating enzymes. (**b)** Major histocompatibility complex (MHC) class I peptide interface cluster. MHC_I is present in 72 proteins and 778 entries in the PDB. (**c)** MHC class II consists of two Pfams (MHC_II_alpha and MHC_II_beta), each with its own peptide cluster. (**d)** SH3_1-peptide interface cluster 1. (**e)** SH3_2-peptide cluster 1. (**f)** SH3_9-peptide cluster 1. These clusters in the SH3 clan show a similar peptide-binding groove in different Pfams.

Peptide-binding domains in ProtCID

The function of many protein domains is to bind peptides from other proteins. Some of these peptide-binding domains are catalytic including proteolysis and amino acid modifications such as phosphorylation and methylation. Others serve to bring other protein domains into contact or to regulate the activity of the bound partner. We define a peptide as a protein chain with less than 30 residues. ProtCID provides data on 1084 Pfam-domains with peptide interactions in the PDB. For each Pfam, we cluster peptides by the number of shared Pfam HMM positions and Root-mean-square deviations (RMSDs) of backbone atoms of the peptides. As an example, Trypsin domains have three large peptide-interface clusters in ProtCID (**Figure 3a**). The first is a set of 260 entries mostly consisting of human thrombin bound to a peptide from hirudin from leeches[76]. The cluster also includes a few structures of complexes of thrombin with hirudin-like peptides from mouse or human proteinase activated receptor 1 and 3 (PAR1, PAR3)[77]. The PAR peptides and hirudins are molecular mimics and bind to thrombin in very similar ways. The second large cluster is a large set of inhibitor peptides bound to the active sites of more than a dozen trypsin-like protease family members. The third cluster is entirely made of complexes of the heavy and light chains of some trypsin-like proteases, including thrombin, chymotrypsin, urokinase and tissue-type plasminogen activators, acrosin, and matriptase. These complexes are created by internal cleavage by activating enzymes. Some Pfams bind peptides in a similar manner across Pfam clans or superfamilies. Examples include the homologous Class I Histocompatibility antigen (MHC_I) proteins (Figure 3b) and Class II Histocompatibility antigens (MHC_II) (Figure 3c).

Many protein domain families primarily function as peptide-binding modules, often within larger proteins with other domains. If two proteins can be demonstrated to interact, for instance through high-throughput protein-protein interaction studies, and one of them contains a peptide-binding domain, then in many cases a reasonable hypothesis is that the peptide-binding domain of one protein may bind to an intrinsically disordered region of the other. We have therefore

compiled a list of the more common peptide-binding domain families within the human proteome with available structural information in ProtCID and the PDB. We included only domains that primarily function as peptide binders and are well represented in the PDB and the human proteome (see Online Methods).

There are currently a total of 42 non-catalytic PPBD Pfam families with peptide-bound structures in the PDB (Supplementary Table 6). These domains are present in 1,051 human protein sequences, or about 5% of the proteome. Some of the Pfams belong to clans, including SH3 domains of the SH3_1 (Figure 3d), SH3_2 families (Figure 3e), and SH3_9 families (Figure 3f). These clusters indicate that Pfams in the same clan, or superfamily, typically bind peptides in the same way. This observation can be used to derive hypotheses of how structures in other Pfams in the same clan with no available peptide-bound structures may bind peptides, e.g. SH3_5, hSH3 and Phn, which have the same groove as SH3_1 in experimental structures, although there are no peptide-bound structures of proteins containing these domains in the PDB (Supplementary Figure 6).

Ligand/protein and nucleic-acid/protein interactions in ProtCID

We define all non-polymer molecules except water in the PDB as ligands. A total of 6,514 Pfams have contacts with 23,156 ligand types in the PDB (Supplementary Table 2). The ligands are clustered based on the extent to which they share Pfam HMM positions that they contact. ProtCID provides coordinates for two different views of Pfam-ligands interactions: 1) one ligand and its interacting Pfams; 2). one Pfam and its interacting ligands. Figure 4a displays the interactions of heme (HEM) binding to two different Pfams, Peroxidase domains and Heme oxygenase domains. Heme binds to 134 different Pfams in 4,319 PDB entries. Figure 4b shows the major clusters of the interactions of pyruvate kinases (PK) and ligands, including sites for metal ions, substrates, allosteric activators such as fructose-1,6-bisphosphate (FBP), and allosteric inhibitors such as ATP and alanine. Pfam-ligand interactions can be queried in

ProtCID by Pfam ID, or browsing Pfams and ligands in the Browse page. Pfam-ligand interactions might be used to model ligand interactions for proteins for which there are no ligand-bound structures (Supplementary Table 7).

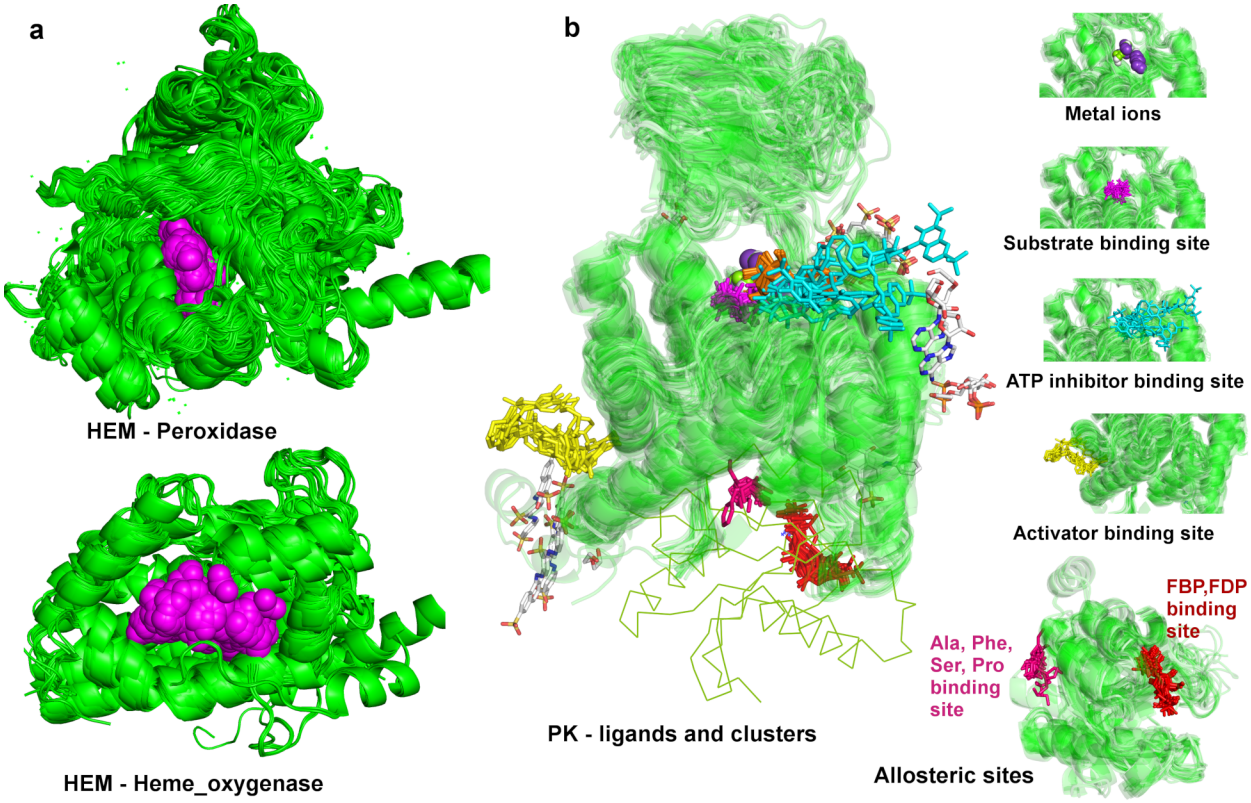

Figure 4 | Pfam-ligand interactions. (**a**) The interactions of Heme (PDB: HEM) and Pfams. The coordinates of HEM-Pfams interactions can be downloaded from the ligand web page in file HEM.tar.gz file. The file provides the domain coordinates of all Pfams which interact with heme (HEM) ligand and the PyMOL scripts for each Pfam. HEM–Peroxidase interactions (top) and HEM–Heme_oxygenase interactions (bottom) are generated by peroxidase_HEM_pairFitDomain.pml and Heme_oxygenase_HEM_pairFitDomain.pml respectively. Hemes are shown in spheres and colored in magenta. (**b**) The interactions of clusters of Pyruvate Kinase (PK) ligands are generated from PK_pdb.tar.gz file which can be downloaded from PK web page. The ligands and clusters are represented as selection objects in PyMOL. Different ligand clusters are shown in different colors. The C-terminal domain (PK_C) is added in lines to show the full-length pyruvate kinase.

In ProtCID, nucleic acids are treated in the same way as other ligands. They are clustered by the extent of overlap of Pfam HMM positions in a protein domain in contact with the nucleic acid. This makes it very easy to find all structures of proteins from a nucleic-acid binding domain family with bound DNA or RNA. A total of 1,260 Pfams interact with DNA or RNA in

6,504 PDB entries.

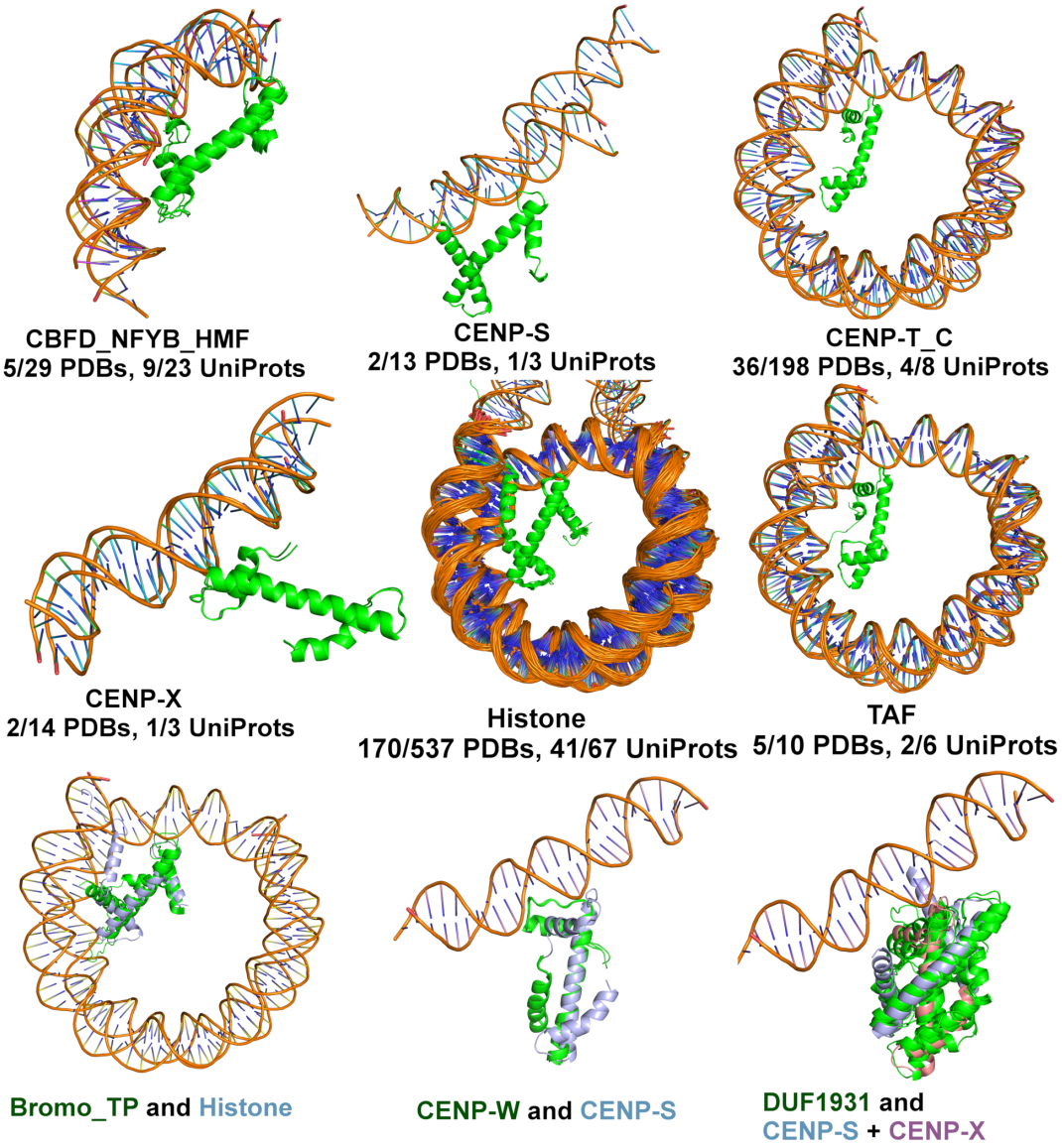

Figure 5 | Protein-domain/DNA interactions can be extended across in Pfam clans. The Histone clan (CL0012) comprises 15 Pfams with similar folds, 6 of which are in the PDB with bound-DNA (shown in the first and second rows). The 9 Pfams in the clan without bound DNA structures are: Bromo_TP (3 PDBs, 2 UniProts), Bromo_Tp_Like (no PDBs), CENP-W (4 PDBs, 1 UniProt), DUF1931 (4 PDBs, 3 UniProts), TAF4 (1 PDB, 1 UniProt), TAFII28 (2 PDBs, 1 UniProt), TFIID_18kDa (2 PDBs, 1 UniProt), TFIID_20kDa (1 PDB, 1 UniProt), and TFIID_31kDa (1 PDB, 1 UniProt). Three of these 9 Pfams (Bromo_TP, CENP-W, and DUF1931) are shown in the bottom row structurally superimposed on closely related clan members with DNA-bound structures (structure without DNA shown in green; DNA-bound structure shown in light blue). Bromo_TP domains are superimposed to Histone domains well (third row, first column). The Pfam DUF1931 is 138 amino acids long, which is close to the sum of the length of CENP-S Pfam (76) and the length of CENP-X Pfam (75). The structure of a member of the DUF1931 Pfam (PDB: 1R4V) can be superimposed to PDB entry 4NDY with contains both CENP-S and CENP-X domains. CENP-X domain is colored in light pink; CENP-S is in light blue.

We can use ProtCID to develop structural hypotheses for Pfams which do not yet have

nucleic-acid bound structures in the PDB. It is a reasonable hypothesis that each Pfam domain in a Pfam clan may bind a nucleic acid to homologous sites. For example, the Histone clan contains 15 Pfams, only 6 of which are in the PDB with bound nucleic acids. However, another 8 Pfams in the clan are represented by experimental structures without bound nucleic acids. The similarity of the nucleic acid binding in 6 members of the Histone clan (Figure 5) indicates that other Pfam members such as Bromo_TP, CENP-W, and DUF1931 also bind nucleic acid in the same manner. Superposition of structures of members of this Pfam clan without bound nucleic acid onto structures with bound nucleic acid leads to models of how these proteins may interact with DNA and RNA (Supplementary Tables 8 and 9). A similar procedure can be used to generate hypotheses for how members of the double-strand RNA binding clan (DSRM) bind RNA molecules (Supplementary Figure 7).

Modeling of protein complexes in ProtCID

Protein-protein interaction studies have identified protein complexes that consist of subunits expressed by many different protein-coding genes. Other proteins function in homooligomeric complexes larger than dimers. ProtCID can be used to develop hypotheses of the structures of these large complexes. We provide several examples.

4HBT proteins have a 'hot dog' fold and form different oligomers. In ProtCID, the 4HBT Pfam has 7 common clusters with at least 10 crystal forms (Figure 6a). The largest cluster has 144 entries in 83 distinct crystal forms. After identifying common PDB entries in these 7 clusters, there are three distinct combinations with no overlapping structures: clusters 1 and 2, clusters 1 and 5, and clusters 1 and 6, which form a β-sheet tetramer, an α-helix tetramer, and a hexamer respectively. Clusters 3, 4, and 7 are smaller interfaces that are present in most of the 1-2, 1-5, and 1-6 structures respectively. A phylogenetic tree shows that 4HBT proteins that are more closely related also tend to form the same oligomers (Supplementary Figure 8). We added the remaining four human proteins (BCHL, ACO11, ACO12 and ACOT9) with 4HBT domains.

These proteins are located in the branch of the hexamers, and we can therefore hypothesize that the correct structure of their oligomers resembles the hexamers in Figure 6a. Currently, ProtCID provides 1,030 combinations of clusters that can be used in a similar fashion to construct homooligomers larger than dimers. These cluster combinations can be used to validate the structures of the biological assemblies of the relevant entries in the PDB[3].

Ionotropic glutamate receptors (iGluRs) form tetramers to permit ion flux, and can be grouped into four subtypes: AMPA, delta, kainate (KA), and NMDA receptors. A typical iGluR contains three Pfam domains: (ANF_receptor), (Lig_chan-Glu_bd) and (Lig_chan) domains. Of the 436 crystal structures containing these domains, the only full-length tetrameric structures are those of rat glutamate receptor 2 and *Xenopus* NMDA receptor (Figure 6c, middle). However, ProtCID contains large clusters derived from single-domain and two-domain structures of many other proteins that can be used to produce full-length, tetrameric models of several human glutamate receptors. These clusters include homodimeric interfaces of the three Pfam domains (two for Pfam (ANF_receptor) and one each for (Lig_chan-Glu_bd) and (Lig_chan)) and two heterodimeric interface clusters that connect these domains to each other: (ANF_receptor)/(Lig_chan-Glu_bd), and (Lig_chan-Glu_bd)/(Lig_chan). Figure 6c shows these six common clusters of (ANF_receptor), (Lig_chan-Glu_bd) and (Lig_chan) from different species which can be used to assemble the tetramer, even if the full-length structures were not known.

TECPR1 binds to ATG5-ATG12 heterodimers and mediates the fusion of the autophagosome-lysosome[78]. Currently there is no crystal structure for the ATG5-ATG12-TECPR1 complex, which can be modeled from an ATG5-ATG12 cluster (4 CFs and 4 entries) and an ATG5-TECPR1 structure (PDB: 4TQ1) by superposing the ATG5 chains (Figure 6b top). ATG16 exists as a homodimer in solution and crystals, which further assembles the autophagosome into higher-order hetero-oligomers[79]. The ATG5-ATG12-ATG16-ATG3 complex is essential in autophagosome formation, in which a disordered region of ATG3 binds to

ATG12[80]. The ProtCID cluster of ATG16 homodimers and two ProtCID clusters of ATG16 interacting with ATG5 can be used to build a heterooctameric structure of two copies each of ATG16, ATG5, ATG12, and ATG3 (Figure 6b bottom). This octameric structure is present in the crystal of ATG16-ATG5-ATG12-ATG3 (PDB: 4NAW) but was not identified by the authors as the biological assembly (they only deposited and showed the structure of the ATG16-ATG5-ATG12-ATG3 heterotetramer).

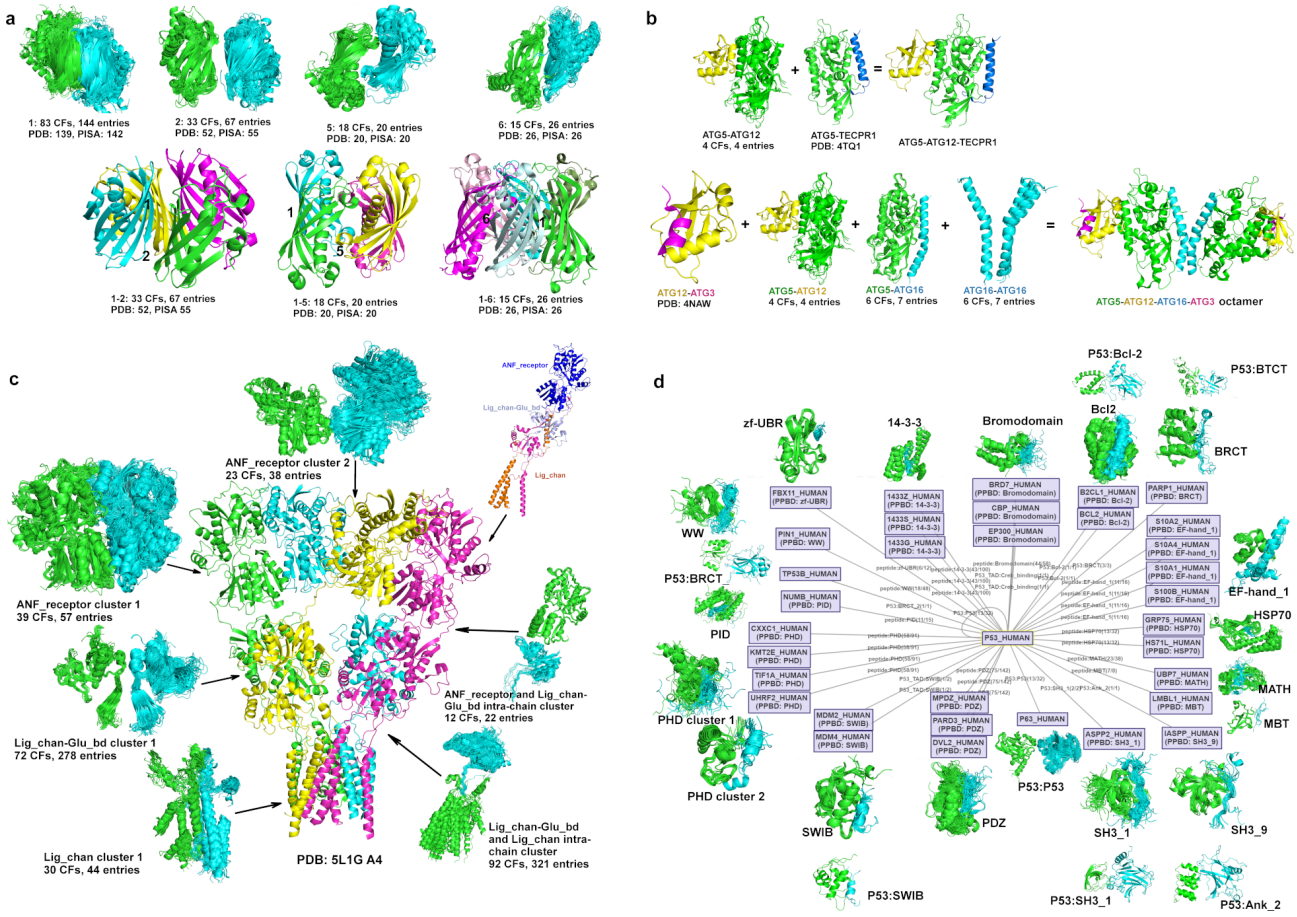

Figure 6 | Hypotheses for the structures of biological complexes and hub protein complexes can be generated from interface clusters in ProtCID. (**a**) Three 4HBT complexes (two tetramers and a hexamer) are assembled from 4 different clusters of Pfam:4HBT: (1) a tetramer with D2 symmetry, assembled from two copies of cluster 1 interfaces and two copies of cluster 2 interfaces. Most of these structures also contain the cluster 3 interface with small surface area; (2) a tetramer with D2 symmetry, assembled from two copies of cluster 1 interfaces and two copies of cluster 5 interfaces; (3) a D3 hexamer with three copies of cluster 1 interfaces and three copies of cluster 6 interfaces. The interfaces in each complex are labeled by their cluster numbers. (**b**) Autophagosome complexes. Top: ATG5-ATG12-TCPR1 trimer from Complex Portal (https://www.ebi.ac.uk/complexportal/complex/CPX-358) is assembled from an ATG5-ATG12 interface cluster and one PDB entry (4TQ1: ATG5-TCPR1 dimer) by

superposing ATG5. Bottom: ATG5-ATG12-ATG16-ATG3 octamer can be assembled from three ProtCID clusters: the ATG5-ATG16 cluster, one ATG16 homodimer cluster, and ATG5-ATG12 cluster. The ATG12-ATG3 dimer is from PDB: 4NAW. The same chains of these interface clusters are superposed to assemble an autophagosome octamer. This octamer occurs in the crystal of PDB: 4NAW. (**c**) Multi-domain Ionotropic glutamate receptor complex can be assembled from domain clusters. The tetramer is composed of six common domain clusters in ProtCID, including four inter-chain domain clusters: ANF_receptor domain cluster 1 and cluster 2, Lig_chan-Glu_bd cluster 1 and Lig_chan cluster 1, two intra-chain domain clusters: ANF_receptor : Lig_chan-Glu_bd domain cluster and Lig_chan-Glu_bd : Lig_chan domain cluster. (**d**) The interactions of hub protein P53_HUMAN. PPBD indicates a professional peptide binding domain. An edge is labeled by Pfam:Pfam if there are crystal structures in PDB, or by peptide:Pfam if the interactor contains a PPBD. If there are no PDB entries containing the interfaces of two UniProt nodes, but one of them has a PPBD, an interaction is predicted and a peptide:PPBD edge is added. The label shows the number of crystal forms and the number of entries in the largest cluster. For peptide:PPBD, the number of crystal forms and the number of entries are counted from the peptide interface cluster. Users click an edge to retrieve the domain interactions between two node proteins and their clusters. Each node is also clickable to query the structures of the node protein. The Pfam assignments and the complete list of interactions for the input UniProts are provided in the table format by the links above the Cytoscape picture.

Many proteins are part of large protein complexes or molecular machines with many different components and thus interact both directly with some partners and indirectly with many others. We have enabled a search function in ProtCID for the structural analysis of large protein complexes. Large protein complexes consisting of many types of proteins can be examined in ProtCID with queries consisting of a list of the Uniprot sequences in a complex. ProtCID identifies all the Pfams in the proteins, and presents all Pfam pairs among the input proteins that directly interact in structures present in the PDB. For example, the Cascade complex is the center of CRISPR-Cas system, which is composed of 5 proteins. After inputting these proteins, ProtCID provides the interactions of UniProts and their Pfams (Supplementary Figure 9). Furthermore, the “Browse” → “Pfam Bio Networks” also provides the Pfam-Pfam networks where Pfams are connected by a common interface in at least two crystal forms (Supplementary Figure 9b) which might provide hypotheses of direct protein interactions within the query complex.

Modeling of hub proteins in ProtCID

Large-scale studies of protein-protein interactions have identified “hub” proteins that participate in a large number of interactions with other proteins[81-83]. Structural analysis of hub

proteins is often hampered by the complexity of structural information in the PDB that might be utilized to develop hypotheses of how these proteins may interact with a large number of other proteins. We have enabled searches on ProtCID designed to provide hypothetical interactions of hub proteins and their partners and among the subunits in protein complexes. The hub protein search page in ProtCID allows a user to upload a list of proteins that are likely to interact with a specific hub protein; the server then returns a list of potential domain-domain and PPBD-peptide interactions that may explain how the hub protein interacts with each of its partners. Figure 6d shows the interactions in the PDB between P53 and its interactors based on Pfams and PPBDs. ProtCID identifies PPBDs (defined in Supplementary Table 4) in both hub and partner proteins so that this mode of binding is also presented to the user as a viable hypothesis. For example, studies show that a peptide segment of the p53 C-terminus binds to 14-3-3 proteins[84, 85]; however there is no structure of this interaction in the PDB. A potential interaction between human a peptide of P53 and and the PPBD 14-3-3 proteins is identified by the edge in the network in Figure 6d.

ProtCID web site

ProtCID is composed of data pipelines, databases, and a web site. On the ProtCID web site, there are four types of inputs: (i) a PDB ID; (ii) one or two Pfam IDs or accession codes; (iii) one or two protein sequences; (iv) one or more UniProt IDs. A user can browse Pfam IDs, Clan IDs, Pfam-Pfam pairs, peptide-interacting Pfams, ligands and Pfam-Pfam networks (Supplementary Figure 10). Pfams are assigned to user-input sequences and UniProt IDs by HMMER3 (http://hmmer.janelia.org/). The procedure of a web query of UniProts is given in Online Methods. Except for PDB ID input, all inputs result in a list of PDB structures containing the Pfams. This gives a user a comprehensive overview of homologous structures in the PDB for any given query. From one structure, a user can check the clusters of all interactions for chain-chain, domain-domain and domain-peptide. Coordinates and sequences are

downloadable for each cluster including PyMOL scripts which can be double clicked to visualize the interactions in PyMOL.

Discussion

Correctly identifying biologically relevant interfaces and assemblies in protein crystals remains a challenging task in structural biology. Automated methods for doing this depend on the biophysical properties[4] and sequence conservation of interfaces[46], but their accuracy is only about 85% on large benchmarks[86]. An alternative approach is to take advantage of multiple crystal forms of the same protein or related proteins[87, 88], and identifying the presence of similar interfaces and assemblies in the crystals. We previously demonstrated that this approach performs very well on benchmarks[5], and developed a database of clusters of similar interfaces of homologous protein chains[22]. We have extended this resource by clustering protein structures at the level of individual domains, which greatly increases the number of independent crystal forms for a large fraction of the protein families in the PDB. We further extended ProtCID by clustering domain-peptide and domain-ligand interactions.

We showed the power of this approach for generating hypotheses for the structures of biologically relevant interfaces and assemblies in cases where identification of the correct assembly is particularly challenging. This occurs for low-affinity dimers whose ability to form complexes is due to the concentration effect of confinement. This is true of proteins in the membrane and of small protein domains adjacent to other dimerization motifs and domains. We showed several examples of this, including the asymmetric dimer of ErbB proteins, a symmetric dimer of the BD2 bromodomains of BET proteins, and a symmetric homodimer of HRAS, KRAS, and NRAS. Notably, programs which aim to identify biological assemblies within crystals from biophysical properties or sequence conservation usually assume that assemblies must be symmetric. The extent of asymmetric assemblies such as the ErbB dimer that are not filamentous (like actin) is unknown. The identification of such complexes is difficult to perform

manually for a large number of structures. ProtCID provides a very useful resource to systematically analyze structural data from the PDB and generate hypotheses with strong structural evidence. The BD2 dimer of BET proteins is found in all crystal forms that contain BD2 domains, and a dimer of these domains may be formed due to the increased concentration caused by the dimerization of the neighboring BD1 domains and the ET dimerization motif. There is significant evidence in favor of the α4-α5 dimer of RAS proteins, particularly for HRAS. The dimer, found in 16 crystal forms of HRAS, KRAS, and NRAS, is consistent with the simultaneous binding of a pair of RAS effector domains to the RAS homodimers, as demonstrated by the presence of the α4-α5 dimer in co-crystal structures of HRAS with RA and RBD domains.

ProtCID can also be used to generate hypotheses about the structures of biological assemblies by extending experimental data from one system to another through the presence of similar domain-domain or chain-chain interfaces in crystals of homologous proteins. We demonstrated this for tyrosine kinases where a ProtCID cluster that contains the well-studied and validated BRAF homodimer also includes MLKL, RIPK2, and ITK. The power of the domain-level interactions was shown by a very large cluster of ACT domain homodimers present in several different multi-domain protein families in the PDB. We used this cluster to generate a hypothesis about the structure and function of ACT-domain interactions in phenylalanine hydroxylase, the gene mutated in different forms of phenylketonuria, a disorder of metabolism. ProtCID can be used to obtain define larger oligomeric assemblies, which we demonstrated for tetrameric ionotropic glutamate receptors and for three different assemblies of 'hotdog' fold proteins, which enabled the generation of a hypothesis for the correct hexameric assembly of four human proteins.

ProtCID compiles and clusters structures with peptides, nucleic acids and ligands together for each domain family in the PDB. Statistical analysis of these complete data sets enables bioinformatics studies to determine rules for binding specificity in these domain types.

The clusters can further be used to model interactions of protein domains with peptides, ligands, and nucleic acids across wide evolutionary distances through superfamily or clan relationships in Pfam.

Many proteins participate in large protein complexes or as hub proteins, interacting with large numbers of other proteins. It can be very time consuming to manually search for structural information on a long list of proteins involved in large complexes and hub-protein interactions. Typically one would have to perform sequence searches for each protein in the list individually and then compile and correlate the data to determine if there are experimental structures that contain homologous domains of each pair of proteins in the list. We have implemented a search function on the ProtCID website that takes a list of Uniprot protein sequence identifiers and searches for possible structure-based interactions among the domains present in any of the proteins in the list. The search function also identifies which proteins in the list contain domains whose primary function is to bind peptides from other proteins ("professional peptide binding domains"). A large fraction of protein-protein interactions in the human proteome may be formed by binding of a disordered protein region of one protein to the peptide-binding domain of another protein. We showed the utility of these search functions in building the structure of a heterooctamer involved in autophagy and to identify the interaction networks of CRISPR proteins and the hub protein p53.

All downloadable structure files in ProtCID come with different PyMOL scripts which can be just double clicked to open in PyMOL. We hope the visualization of interactions at different levels in ProtCID provides straightforward and informative interpretation and clues for further study. The common interfaces of ProtCID are a useful data source for training and testing data sets for prediction methods, and can be used in different fields, including drug discovery, protein-protein Interaction (PPI) prediction on both molecular and proteome levels, mutation predictions and homology modeling.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGEMENTS

This work was supported by NIH grants R01 GM084453 and R35 GM122517.

AUTHOR CONTRIBUTIONS

Q.X. conceived and designed the work, developed the project, analyzed the data, implemented the ProtCID web site, and wrote the manuscript. R.D. conceived and designed the work, developed the project, analyzed the data, and wrote the manuscript.

REFERENCES

1. Xu, Q. & Dunbrack, R.L., Jr. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. *Bioinformatics* **28**, 2763-2772 (2012).
2. Finn, R.D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222-230 (2014).
3. Xu, Q. & Dunbrack Jr, R.L. Principles and characteristics of biological assemblies in experimentally determined protein structures. *Curr. Opin. Struct. Biol.* **in press** (2019).
4. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, 774-797 (2007).
5. Xu, Q. et al. Statistical analysis of interface similarity in crystals of homologous proteins. *J. Mol. Biol.* **381**, 487-507 (2008).
6. Teyra, J., Samsonov, S.A., Schreiber, S. & Pisabarro, M.T. SCOWLP update: 3D classification of protein-protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds. *BMC Bioinformatics* **12**, 398 (2011).
7. Faure, G., Andreani, J. & Guerois, R. InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res.* **40**, D847-856 (2012).
8. Shoemaker, B.A. et al. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.* **40**, D834-840 (2012).
9. Finn, R.D., Miller, B.L., Clements, J. & Bateman, A. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res.* **42**, D364-373 (2014).
10. Mosca, R., Ceol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.* **42**, D374-379 (2014).
11. Krishna, A.A.D.R. PICI: A web server with a multi-parametric algorithm for identifying interaction sites within protein complexes. *Bioinformatics* **12** (2016).
12. Zhu, H., Domingues, F.S., Sommer, I. & Lengauer, T. NOXclass: prediction of protein-

protein interaction types. *BMC Bioinformatics* **7**, 27 (2006).
13. Bernauer, J., Bahadur, R.P., Rodier, F., Janin, J. & Poupon, A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics* (2008).
14. Baskaran, K., Duarte, J.M., Biyani, N., Bliven, S. & Capitani, G. A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Struct. Biol.* **14**, 22 (2014).
15. London, N., Movshovitz-Attias, D. & Schueler-Furman, O. The structural basis of peptide-protein binding strategies. *Structure* **18**, 188-199 (2010).
16. Vanhee, P. et al. PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res.* **38**, D545-551 (2010).
17. Das, A.A., Sharma, O.P., Kumar, M.S., Krishna, R. & Mathur, P.P. PepBind: a comprehensive database and computational tool for analysis of protein-peptide interactions. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute* **11**, 241-246 (2013).
18. Reddy, A.S., Amarnath, H.S., Bapi, R.S., Sastry, G.M. & Sastry, G.N. Protein ligand interaction database (PLID). *Comput. Biol. Chem.* **32**, 387-390 (2008).
19. Anand, P., Nagarajan, D., Mukherjee, S. & Chandra, N. PLIC: protein-ligand interaction clusters. *Database (Oxford)* **2014**, bau029 (2014).
20. Yang, J., Roy, A. & Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res* **41**, D1096-1103 (2013).
21. Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405-412 (2015).
22. Xu, Q. & Dunbrack, R.L., Jr. The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.* **39**, D761-770 (2011).
23. Bojja, R.S. et al. Architecture and assembly of HIV integrase multimers in the absence of DNA substrates. *J. Biol. Chem.* **288**, 7373-7386 (2013).
24. Jaffe, E.K., Stith, L., Lawrence, S.H., Andrake, M. & Dunbrack, R.L. A new model for allosteric regulation of phenylalanine hydroxylase: implications for disease and therapeutics. *Arch. Biochem. Biophys.* **530**, 73-82 (2013).
25. Xu, Q. et al. Identifying three-dimensional structures of autophosphorylation complexes in crystals of protein kinases. *Sci Signal* **8**, rs13 (2015).
26. Arthur, J.W. & Reichardt, J.K. Modeling single nucleotide polymorphisms in the human AKR1C1 and AKR1C2 genes: implications for functional and genotyping analyses. *PloS one* **5**, e15604 (2010).
27. Zhang, X., Gureasko, J., Shen, K., Cole, P.A. & Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **125**, 1137-1149 (2006).
28. Weitzner, B., Meehan, T., Xu, Q. & Dunbrack, R.L., Jr. An unusually small dimer interface is observed in all available crystal structures of cytosolic sulfotransferases. *Proteins* **75**, 289-295 (2009).
29. Sun, M. & Leyh, T.S. The human estrogen sulfotransferase: a half-site reactive enzyme. *Biochemistry* **49**, 4779-4785 (2010).
30. Spencer-Smith, R. et al. Inhibition of RAS function through targeting an allosteric regulatory site. *Nat. Chem. Biol.* **13**, 62 (2017).
31. Ambrogio, C. et al. KRAS Dimerization Impacts MEK Inhibitor Sensitivity and Oncogenic Activity of Mutant KRAS. *Cell* (2018).
32. Stein, A. & Aloy, P. Contextual specificity in peptide-mediated protein interactions. *PLOS ONE* **3**, e2524 (2008).
33. Zhou, S. et al. SH2 domains recognize specific phosphopeptide sequences. *Cell* **72**, 767-778 (1993).

34. Yu, X., Chini, C.C.S., He, M., Mer, G. & Chen, J. The BRCT domain is a phospho-protein binding domain. *Science* **302**, 639-642 (2003).
35. Songyang, Z. et al. Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. *Science* **275**, 73-77 (1997).
36. Burd, C.G. & Dreyfuss, G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615-621 (1994).
37. Boch, J. et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509-1512 (2009).
38. Müller, S., Chaikuad, A., Gray, N.S. & Knapp, S. The ins and outs of selective kinase inhibitor development. *Nat. Chem. Biol.* **11**, 818 (2015).
39. Engelman, A. & Cherepanov, P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nature Reviews Microbiology* **10**, 279 (2012).
40. Isberg, V. et al. GPCRdb: an information system for G protein-coupled receptors. *Nucleic Acids Res.* **45**, 2936 (2017).
41. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaud. Sci. Nat.* **37**, 547-549 (1901).
42. Petrotchenko, E.V., Pedersen, L.C., Borchers, C.H., Tomer, K.B. & Negishi, M. The dimerization motif of cytosolic sulfotransferases. *FEBS Lett.* **490**, 39-43 (2001).
43. Aertgeerts, K. et al. Structural analysis of the mechanism of inhibition and allosteric activation of the kinase domain of HER2 protein. *J. Biol. Chem.* **286**, 18756-18765 (2011).
44. Qiu, C. et al. Mechanism of activation and inhibition of the HER4/ErbB4 kinase. *Structure* **16**, 460-467 (2008).
45. Littlefield, P. et al. Structural analysis of the EGFR/HER3 heterodimer reveals the molecular basis for activating HER3 mutations. *Sci Signal* **7**, ra114 (2014).
46. Bliven, S., Lafita, A., Parker, A., Capitani, G. & Duarte, J.M. Automated evaluation of quaternary structures from protein crystals. *PLOS Comput. Biol.* **14**, e1006104 (2018).
47. Hancock, J.F., Magee, A.I., Childs, J.E. & Marshall, C.J. All ras proteins are polyisoprenylated but only some are palmitoylated. *Cell* **57**, 1167-1177 (1989).
48. Muratcioglu, S. et al. GTP-dependent K-Ras dimerization. *Structure* **23**, 1325-1335 (2015).
49. Spencer-Smith, R. et al. Inhibition of RAS function through targeting an allosteric regulatory site. *Nat. Chem. Biol.* **13**, 62 (2017).
50. Güldenhaupt, J. et al. N-Ras forms dimers at POPC membranes. *Biophys. J.* **103**, 1585-1593 (2012).
51. Johnson, C.W. et al. The small GTPases K-Ras, N-Ras and H-Ras have distinct biochemical properties determined by allosteric effects. *J. Biol. Chem.*, jbc. M117. 778886 (2017).
52. Vetter, I.R. Interface analysis of small GTP binding protein complexes suggests preferred membrane orientations. *Biol. Chem.* **398**, 637-651 (2017).
53. Fujisawa, T. & Filippakopoulos, P. Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nature Reviews Molecular Cell Biology* **18**, 246 (2017).
54. Garcia-Gutierrez, P., Mundi, M. & Garcia-Dominguez, M. Association of bromodomain BET proteins with chromatin requires dimerization through the conserved motif B. *J. Cell Sci.* **125**, 3671-3680 (2012).
55. Nakamura, Y. et al. Crystal structure of the human BRD2 bromodomain: insights into dimerization and recognition of acetylated histone H4. *J. Biol. Chem.* **282**, 4193-4201 (2007).
56. Huang, H. et al. Solution structure of the second bromodomain of Brd2 and its specific interaction with acetylated histone tails. *BMC Struct. Biol.* **7**, 57 (2007).

57. Rajakulendran, T., Sahmi, M., Lefrancois, M., Sicheri, F. & Therrien, M. A dimerization-dependent mechanism drives RAF catalytic activation. *Nature* **461**, 542-545 (2009).
58. Hatzivassiliou, G. et al. RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth. *Nature* **464**, 431-435 (2010).
59. Mayerhofer, H., Panneerselvam, S. & Mueller-Dieckmann, J. Protein kinase domain of CTR1 from Arabidopsis thaliana promotes ethylene receptor cross talk. *J. Mol. Biol.* **415**, 768-779 (2012).
60. Pellegrini, E., Signor, L., Singh, S., Boeri Erba, E. & Cusack, S. Structures of the inactive and active states of RIP2 kinase inform on the mechanism of activation. *PLoS One* **12**, e0177161 (2017).
61. Sun, L. et al. Mixed lineage kinase domain-like protein mediates necrosis signaling downstream of RIP3 kinase. *Cell* **148**, 213-227 (2012).
62. Xie, T. et al. Structural insights into RIP3-mediated necroptotic signaling. *Cell reports* **5**, 70-78 (2013).
63. Wu, X.N. et al. Distinct roles of RIP1-RIP3 hetero- and RIP3-RIP3 homo-interaction in mediating necroptosis. *Cell Death Differ.* **21**, 1709-1720 (2014).
64. Wang, H. et al. Mixed lineage kinase domain-like protein MLKL causes necrotic membrane disruption upon phosphorylation by RIP3. *Mol. Cell* **54**, 133-146 (2014).
65. Raju, S. et al. Kinase domain dimerization drives RIPK3-dependent necroptosis. *Sci. Signal.* **11**, eaar2188 (2018).
66. Qi, Q., Sahu, N. & August, A. Tec kinase Itk forms membrane clusters specifically in the vicinity of recruiting receptors. *J. Biol. Chem.* **281**, 38529-38534 (2006).
67. Brown, K. et al. Crystal structures of interleukin-2 tyrosine kinase and their implications for the design of selective inhibitors. *J. Biol. Chem.* **279**, 18727-18732 (2004).
68. Kutach, A.K. et al. Crystal structures of IL-2-inducible T cell kinase complexed with inhibitors: insights into rational drug design and activity regulation. *Chem. Biol. Drug Des.* **76**, 154-163 (2010).
69. Charrier, J.D. et al. Discovery and structure-activity relationship of 3-aminopyrid-2-ones as potent and selective interleukin-2 inducible T-cell kinase (Itk) inhibitors. *J. Med. Chem.* **54**, 2341-2350 (2011).
70. McLean, L.R. et al. X-ray crystallographic structure-based design of selective thienopyrazole inhibitors for interleukin-2-inducible tyrosine kinase. *Bioorg. Med. Chem. Lett.* **22**, 3296-3300 (2012).
71. Harling, J.D. et al. Discovery of novel irreversible inhibitors of interleukin (IL)-2-inducible tyrosine kinase (Itk) by targeting cysteine 442 in the ATP pocket. *J. Biol. Chem.* **288**, 28195-28206 (2013).
72. Alder, C.M. et al. Identification of a Novel and Selective Series of Itk Inhibitors via a Template-Hopping Strategy. *ACS Med. Chem. Lett.* **4**, 948-952 (2013).
73. MacKinnon, C.H. et al. Structure-based design and synthesis of potent benzothiazole inhibitors of interleukin-2 inducible T cell kinase (ITK). *Bioorg. Med. Chem. Lett.* **23**, 6331-6335 (2013).
74. Pastor, R.M. et al. Discovery and optimization of indazoles as potent and selective interleukin-2 inducible T cell kinase (ITK) inhibitors. *Bioorg. Med. Chem. Lett.* **24**, 2448-2452 (2014).
75. Gersting, S.W. et al. Loss of function in phenylketonuria is caused by impaired molecular motions and conformational instability. *The American Journal of Human Genetics* **83**, 5-17 (2008).
76. Grütter, M. et al. Crystal structure of the thrombin-hirudin complex: a novel mode of serine protease inhibition. *The EMBO journal* **9**, 2361-2365 (1990).
77. Mathews, II et al. Crystallographic structures of thrombin complexed with thrombin receptor peptides: existence of expected and novel binding modes. *Biochemistry* **33**,

3266-3279 (1994).
78. Chen, D. et al. A mammalian autophagosome maturation mechanism mediated by TECPR1 and the Atg12-Atg5 conjugate. *Mol. Cell* **45**, 629-641 (2012).
79. Kuma, A., Mizushima, N., Ishihara, N. & Ohsumi, Y. Formation of the approximately 350-kDa Apg12-Apg5.Apg16 multimeric complex, mediated by Apg16 oligomerization, is essential for autophagy in yeast. *J. Biol. Chem.* **277**, 18619-18625 (2002).
80. Metlagel, Z., Otomo, C., Takaesu, G. & Otomo, T. Structural basis of ATG3 recognition by the autophagic ubiquitin-like protein ATG12. *Proc Natl Acad Sci U S A* **110**, 18844-18849 (2013).
81. Huttlin, E.L. et al. The BioPlex network: a systematic exploration of the human interactome. *Cell* **162**, 425-440 (2015).
82. Havugimana, P.C. et al. A census of human soluble protein complexes. *Cell* **150**, 1068-1081 (2012).
83. Hein, M.Y. et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell* **163**, 712-723 (2015).
84. Schumacher, B., Mondry, J., Thiel, P., Weyand, M. & Ottmann, C. Structure of the p53 C-terminus bound to 14-3-3: implications for stabilization of the p53 tetramer. *FEBS Lett.* **584**, 1443-1448 (2010).
85. Rajagopalan, S., Sade, R.S., Townsley, F.M. & Fersht, A.R. Mechanistic differences in the transcriptional activation of p53 by 14-3-3 isoforms. *Nucleic Acids Res.* **38**, 893-906 (2010).
86. Korkmaz, S. et al. Investigation of protein quaternary structure via stoichiometry and symmetry ınformation. *PloS one* **13**, e0197176 (2018).
87. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Scientific reports* **7**, 10480 (2017).
88. Dey, S., Ritchie, D.W. & Levy, E.D. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nat. Methods* **15**, 67 (2018).

ONLINE METHODS

ProtCID databases. To create ProtCID, we first assign Pfam domains to all unique sequences in the Protein Data Bank with a procedure we developed for this purpose[1]. Our protocol employs alignment of hidden Markov models (HMMs) of PDB protein sequences to Pfam HMMs as well as structure alignments for validating the assignments of weakly scoring hits. We carefully annotate "split domains," which are protein domains with other domains inserted within them, and correctly identify a larger number of repeats in repeat-containing proteins.

With these Pfam assignments, we define a Pfam architecture for each protein in the PDB, which is defined as the ordered list of Pfams along the chain along with their positions in the sequence. For example, a protein with one kinase domain and one SH2 domain in that order is annotated as '(Pkinase)_(SH2)'. We build all possible chain-chain dimers from crystals that contain proteins with the same single architecture or pair of architectures using the crystallographic symmetry operators provided by the PDB, and then cluster all of these interfaces (e.g., all (Pkinase)_(SH2):(Pkinase)_(SH2) interfaces, and all (SH3):(Pkinase)_(SH2) interfaces)) with a Jaccard-index[41] metric based on shared residue-residue contacts across each interface, thus producing "chain-based" clusters.

Separately, we also cluster interfaces of all pairs of individual Pfam domains from single and multi-domain proteins. This includes both same-Pfam pairs and different Pfam pairs. Same-Pfam pairs are mostly homodimers, but they also include some interchain homologous heterodimers and intrachain domain-domain interactions with the same Pfam. Different-Pfam pairs are heterodimers (domains with different sequences from two different protein chains) but they also include some intrachain domain-domain interactions. For example, we cluster all (Pkinase):(Pkinase), (SH2):(SH2), and (Pkinase):(SH2) interfaces, regardless of what single-domain or multi-domain architectures these domains come from. For each homodimeric or heterodimeric pair of Pfam chain architectures or Pfam domains, ProtCID reports clusters that contain the interface (sorted by the number of crystal forms they are observed in), as well as the

PDB entries that contain them and the interface surface areas, and whether or not the interface is present in the author or PISA-defined biological assemblies for these entries. Links are provided to download the coordinates of the domain and chain dimers and PyMol scripts to visualize them.

Interface clustering. Supplementary Figure 11 shows the procedure we use to generate the interface clusters. Each PDB structure is represented by its chain Pfam architectures. For instance, the heterodimeric Pfam architecture of PDB:1A22 is (Hormone_1);(EpoR_lig-bind)_(fn3). There are two types of interfaces in ProtCID: chain interfaces and domain interfaces. For PDB:1A22, there are 3 different chain interfaces: (Hormone_1) homodimers, (EpoR_lig-bind)_(fn3) homodimers, and (Hormone_1):(EpoR_lig-bind)_(fn3) heterodimers, while there are 6 types of domain interfaces: (Hormone_1) homodimers, (EpoR_lig-bind) homodimers, (fn3) homodimers, (Hormone_1):(EpoR_lig-bind) heterodimers, (Hormone_1):(fn3) heterodimers, and (EpoR_lig-bind):(fn3) heterodimers. Each of these chain interfaces is referred to as a "Pfam architecture interaction." Each of these domain interfaces is referred to as a "Pfam-Pfam interaction." One entry may belong to several Pfam-Pfam interactions if the entry contains more than one type of Pfam domain. In Supplementary Figure 11, the Sulfotransferase interaction contains 70 entries in 37 crystal forms. Entries are assigned to the same crystal form if: 1) they have the same entry Pfam architectures; 2) the same space group; 3) the same asymmetric unit stoichiometry of protein chains; 4) crystal cell dimensions and angles within 1% variance. We then compare the interfaces of crystals with the same Pfam entry architectures and if at least two thirds of their interfaces are highly similar interfaces, then we merge the crystal forms into one crystal form. This can happen when two different structures are essentially the same crystal and contain all of the same interfaces, but one is solved as an asymmetric unit monomer and a space group with N symmetry operators and the other is solved as an asymmetric unit dimer and a related space group with N/2 symmetry operators.

Chain interfaces are generated from a collection of 27 unit cells arranged in a 3x3x3 lattice. The domain interfaces are generated from chain interfaces by the sequence boundary of domains. Two domains are considered to be interacting if and only if they have at least 10 pairs of C_β atoms with distance ≤ 12 Å and at least one atomic contact ≤ 5 Å, or at least 5 atomic contacts ≤ 5 Å.

To measure the similarity of two interfaces, we must have a correspondence of the residues in one interface with homologous positions in the other, especially when the proteins are homologous and not identical sequences. Since we have an alignment of every structure to Pfams, we can use the Pfam HMM positions to identify homologous positions in two homologous proteins or protein domains (as long as the HMM covers 80% of the shorter domain). The similarity of interface pairs was calculated by the Q function described by Xu et al[5], which is equal to a weighted count of the common contacting residue pairs in two interfaces divided by the total number of unique pairs. A value of Q of 1 means two interfaces are interacting in an identical way. A value of Q of 0 means there are no common contacts. We cluster domain interfaces with surface area > 100 $Å^2$ by a hierarchical average-linkage clustering algorithm. Initially each interface is initialized to be in its own cluster. At each step, the two clusters with the highest average Q-score between them are merged as long as their Q_{avg} ≥ 0.30. When no two clusters can be merged with Q_{avg} ≥ 0.30, then the algorithm is stopped.

Pfam-Peptide interfaces. A peptides is defined as any protein chain with length less than 30 amino acids in the PDB. The chain type is based on the attribute "Polymer Type" defined in the PDB mmCIF files. A protein chain has the type defined as "polypeptide". A domain-peptide interface is defined as an interaction with ≥ 10 C_β-C_β contacts with distance ≤ 12 Å, or ≥ 5 atomic contacts with distance ≤ 5 Å. If a peptide is contacting several chains in a biological assembly, the interface with ≥ 75% atomic contacts is used as the peptide-interacting interface; otherwise,

we keep all the interfaces. For any two Pfam-peptide interfaces in the same Pfam, the number of same-Pfam HMM positions interacting with peptides are counted as N_{hmm}. RMSDs of peptides ($RMSD_{pep}$) are calculated after superposing the domain structures. We used PyMOL to generate the coordinates by using the "pair_fit" command to align the domain structures via their Pfam HMM positions, then calculated the minimum RMSD by applying linear least-squares fit on the domain interacting regions of two peptides, that is, the region between the first residue interacting with the domain and the last residue interacting with the domain. We clustered Pfam-peptide interfaces using a hierarchical average-linkage clustering algorithm by N_{hmm} and $RMSD_{\mathrm{pep}}$. In this method, each interface is initialized to be a cluster. At each step, the two clusters with the highest common interacting Pfam HMM sites are merged, as long as the $N_{\mathrm{hmm}} \geq 3$ and $RMSD_{\mathrm{pep}} \leq 10$ Å.

We used several criteria to define Professional Peptide Binding Domains (PPBDs):

1. The primary function of the domain must be peptide-binding in most proteins that contain the domain. Some domains primarily perform other functions such as binding DNA or other folded protein domains, and their functions are modified by peptide binding; these are excluded from PPBDs. In most cases, there is a common motif, often confined to one amino acid position, that demonstrates that peptide binding was a function of the common ancestor of proteins that contain the domain.
2. There must be structures of at least three different peptide-bound complexes of the domain in the PDB (i.e. different domain Uniprots), and the peptides must all bind in the same location and orientation on the surface of the domain.
3. There must be at least 3 human proteins that contain the domain
4. We exclude repeat proteins (e.g. TPR repeats) that have evolved the ability to bind peptide multiple times in a manner consistent with convergent evolution.

We exclude domains for which peptide-binding includes catalytic modification of the peptide, which includes proteases and enzymes that add or remove post-translational modifications.

Pfam-ligand interactions. Domain-ligand interactions are calculated from the asymmetric units in the PDB. We define all non-polymer molecules except water in the PDB as ligands. A ligand interaction refers to at least one atomic contact with distance ≤ 4.5 Å between a protein domain and a ligand. Nucleic acids are treated in the same way as other ligands. Domain-DNA/RNA interfaces are calculated from the biological assemblies in the PDB. The interface with the largest number of Pfam HMM positions in contact with DNA or RNA is selected.

For each Pfam, we superpose domain structures based on their Pfam HMM positions and provide a PyMOL script file to display the interactions in PyMOL. The domain structures and PyMOL script files are compressed as a single file, named after the Pfam ID (e.g. PK.tar.gz). The ligands are clustered based on the number of common Pfam HMM positions that they contact. One PyMOL session contains the coordinates of domains of a Pfam, and all interacting ligands. Each ligand and each group of ligands are defined as selection objects with the names of ligands, and cluster numbers; users can turn on or off each selection by clicking it in the object list. There are three types of coordinates provided by ProtCID: (i) the best domain with least missing coordinates of a PDB entry, marked by "pdb" (e.g. PK_pdb.tar.gz.); (ii) the best domain of a unique sequence, marked by "unp" (e.g. PK_unp.tar.gz); (iii) the best domain of a crystal form, marked by "cryst" (e.g. PK_cryst.tar.gz). The "unp" and "cryst" files contain fewer domains, and these may be easier to open in PyMOL for big Pfam families. A user can download the coordinates files of domain-ligand interactions along with PyMOL scripts from each Pfam page on the ProtCID web site (e.g., (http://dunbrack2.fccc.edu/ProtCiD/Results/EntityPfamArchWithPfam.aspx?PfamId=PK).

Besides the data for a Pfam and its interacting ligands, ProtCID also provide data sets for a ligand and its interacting Pfams. For each ligand, ProtCID provides domain coordinates and PyMOL scripts for each Pfam and this ligand. All PyMOL scripts and structures are compressed as a single file named after this ligand (e.g. HEM.tar.gz) and can be downloaded

from the page for each ligand (e.g., The coordinates of HEM-Pfams interactions can be downloaded from the web page (http://dunbrack2.fccc.edu/ProtCiD/IPdbfam/PfamLigands.aspx?Ligand=HEM). In this way, users can have all homologous proteins interacting with a specific ligand at one click, and open a PyMOL script to study the interaction between a Pfam and this ligand.

Interactions of user sequences. A user can input one or two sequences on the ProtCID web site to retrieve common interfaces between them or between homologues with the same Pfams (http://dunbrack2.fccc.edu/ProtCiD/Search/sequence.aspx and http://dunbrack2.fccc.edu/ProtCiD/Search/sequences.aspx). Sequences are searched against the Pfam HMM database by running HMMER 3.1 (http://hmmer.org/). A simple greedy algorithm is used to assign Pfams to the sequences. The hit with the best E-value (smallest value) and at least ≥ 80% coverage of the Pfam HMM is selected first. For each additional Pfam hit in order of E-value, if it does not overlap with any existing Pfam assignments by > 10 residues on either end, then an assignment is made. After Pfam assignments, ProtCID returns chain interactions of Pfam architectures and Pfam domain interactions if input is one sequence. For the input of two sequences, chain and domain interactions are returned between the chain architectures and Pfams of the first sequence and the chain architectures and Pfams of the second sequence.

Protein-protein interactions. A user can input a list of UniProt codes to identify interactions among them (http://dunbrack2.fccc.edu/ProtCiD/Search/Uniprots.aspx). There are two ways to identify interactions: “First to All” and “All to All”. “First to All” refers to the interactions between the first protein in a list and all other proteins in the list (proteins 2 to *N*). These results may be of interest if the first protein is a hub protein with many protein interactors. “All to All” means the pairwise interactions among a list of input proteins, which is more useful for large protein complexes or complicated pathways with uncertain connections.

A user can choose interface types to search for: either “Interfaces on Pfams” or

“Interfaces on Structures”. “Interfaces on Structures” means ProtCID only returns structures that contain the actual user-input Uniprot sequences (i.e. not their homologues as defined by Pfam). This may be useful as a first search to determine whether the query sequences are actually in the PDB in heterocomplex structures. The “Interfaces on Pfams” search means ProtCID will return interfaces from the PDB between any pairs of homologous proteins of the query sequences. The UniProts are assigned to Pfams either by the ProtCID database, or by running HMMER (http://hmmer.org/) on Pfam HMM database if the UniProts are not in our ProtCID database. In this case, sequence files are downloaded from the UniProt web site (http://www.uniprot.org/), Pfams are assigned to these UniProt sequences. Interactions are written into a graph XML file (graphml format) and displayed in Cytoscape (http://www.cytoscape.org/) web, where input proteins are represented as nodes and interactions as edges. An edge is labeled by the number of crystal forms (numerator) and the number of PDB entries (denominator) of the biggest interface cluster. If there are no interface clusters, the label is the number of crystal forms and the number of structures containing the interaction. If there are no Pfam-Pfam interactions identified from crystal structures and if one of the proteins has a professional peptide binding domain (PPBD), then an interaction is predicted and an edge is drawn. A user can click a node to display all PDB entries of the UniProt if the “Interfaces on Structures” option is selected, or entries containing any Pfams of the UniProt protein if the “Interfaces on Pfams” is selected. Clicking an edge will return a summary table of all Pfam pairs and their biggest interface clusters between two node proteins. A user can click a Pfam pair ID to retrieve the interface clusters. Pfam assignments and interactions can be downloaded as text files. Like other queries, all coordinate files are downloadable for further study. Supplementary Figure 12 shows the flowchart of the query of P53_HUMAN and its interactors. The list was collected from UniProt web page (https://www.uniprot.org/uniprot/P04637#interaction).

Pfam-Pfam interaction networks. ProtCID provides two types of Pfam and Pfam interaction

networks: First, one Pfam to other Pfams (e.g., http://dunbrack2.fccc.edu/ProtCiD/IPDBfam/PfamNetwork.htm?Pfam=Pkinase) is a centralized network with the input Pfam at the center, containing the interactions with other Pfams including itself. An interaction is represented as an edge, with the number of structures as the weight. Second, in Pfam interaction networks (http://dunbrack2.fccc.edu/ProtCiD/Browse/PfamBioNet.aspx) a Pfam interacts with at least one other Pfam, where an interaction refers to a domain interface cluster with at least two crystal forms and minimum sequence identity < 90%. A graph XML file is generated, where a node is a Pfam, an edge is the interaction between two Pfams. An edge weight is the number of crystal forms of the biggest interface cluster. The label of an edge is the number of crystal forms of the biggest cluster (numerator), the number crystal forms of an interaction (denominator) and the minimum sequence identity in parentheses.