Article

Structural Perspectives on the Evolutionary Expansion of Unique Protein-Protein Binding Sites

Alexander Goncearenco,[1] Alexey K. Shaytan,[1] Benjamin A. Shoemaker,[1] and Anna R. Panchenko[1,*]
[1]Computational Biology Branch of the National Center for Biotechnology Information, Bethesda, Maryland

ABSTRACT Structures of protein complexes provide atomistic insights into protein interactions. Human proteins represent a quarter of all structures in the Protein Data Bank; however, available protein complexes cover less than 10% of the human proteome. Although it is theoretically possible to infer interactions in human proteins based on structures of homologous protein complexes, it is still unclear to what extent protein interactions and binding sites are conserved, and whether protein complexes from remotely related species can be used to infer interactions and binding sites. We considered biological units of protein complexes and clustered protein-protein binding sites into similarity groups based on their structure and sequence, which allowed us to identify unique binding sites. We showed that the growth rate of the number of unique binding sites in the Protein Data Bank was much slower than the growth rate of the number of structural complexes. Next, we investigated the evolutionary roots of unique binding sites and identified the major phyletic branches with the largest expansion in the number of novel binding sites. We found that many binding sites could be traced to the universal common ancestor of all cellular organisms, whereas relatively few binding sites emerged at the major evolutionary branching points. We analyzed the physicochemical properties of unique binding sites and found that the most ancient sites were the largest in size, involved many salt bridges, and were the most compact and least planar. In contrast, binding sites that appeared more recently in the evolution of eukaryotes were characterized by a larger fraction of polar and aromatic residues, and were less compact and more planar, possibly due to their more transient nature and roles in signaling processes.

INTRODUCTION

Large resources are being devoted to understanding the mechanisms of protein function and analyzing protein interactions. There are several major experimental techniques that can be used to identify protein-protein interactions. Two-hybrid and affinity-purification assays, for example, provide data on binary and nonbinary interaction partners, respectively, whereas structural biology methods, mainly x-ray and NMR, offer the atomistic detail of binding-site locations. The most comprehensive two-hybrid study to date resulted in 14,000 interactions between human proteins (1), and a recent affinity-purification-based census of the human proteome detected ~600 protein complexes, resulting in a network of 14,000 interactions (2). The Protein Data Bank (PDB) (3) contains more than 26,000 human protein structures, approximately half of which represent protein complexes. However, the structural database is quite redundant, as summarized in Table 1, with only ~6000 and 2000 nonredundant human protein structures and structural complexes, respectively. Progress in structural biology is often evaluated by analyzing the structural coverage of protein domain families, since proteins have evolved through the shuffling of functional domains. As a result, in the course of evolution, domains have developed specific interaction interfaces. Recently, we surveyed the structural coverage of protein interactions in the protein domain families and superfamilies defined in the Conserved Domain Database (CDD) and Pfam databases, and identified families with multiple protein-protein binding sites that could be potential targets of future structural studies (4).

The trends of PDB growth have been periodically reviewed (5); however, the census of protein-protein complexes has not been updated recently. Here, we considered the biologically relevant interactions represented by the biological units of protein complexes and analyzed the growth of the number of protein complexes in the PDB throughout the last 30 years. The wealth of structural data suggests that most human proteins may be involved in protein-protein interactions; however, many of these proteins have not yet been structurally characterized. We compared the binding sites of domains among different proteins using the method implemented in the Inferred Biomolecular Interaction Server (IBIS) (6,7), which superimposes the structures of protein domains and clusters binding sites into similarity groups. These clusters allowed us to identify a nonredundant set of unique binding sites.

A large number of databases and methods are available for structural analysis of protein-protein interactions and interfaces (8–11). Previous comparative structural analyses of

Submitted March 24, 2015, and accepted for publication June 25, 2015.
*Correspondence: panch@ncbi.nlm.nih.gov
Editor: H. Jane Dyson.

0006-3495/15/09/1295/12
http://dx.doi.org/10.1016/j.bpj.2015.06.056

TABLE 1 Current census of the number of human protein structures and protein complexes in the PDB

	Redundant Structures in the PDB	Nonredundant Structures (40% Sequence Identity)
Monomer	15,000	4000
Homooligomer	6500	1500
Heterooligomer	5000	500
Total	26,500	6000

Data were taken from the PDB (3). The numbers are rounded to thousands.

different protein complexes revealed the recurrence of sequence motifs and binding arrangements/modes on protein-protein interfaces (12–14). Although binding arrangements evolve quite rapidly as proteins diverge (15–17), certain characteristic binding modes are conserved among homologs and in some cases even among nonhomologous proteins (18,19). Over the last decade, several attempts have been made to estimate the number of all possible types of protein interactions; however, these estimates largely varied depending on the definitions of interfaces and similarity measures employed in each study (14,20–24).

It may appear that the space of all possible quaternary protein architectures and binding interfaces has been thoroughly explored already (24), and one way to shed some light on this problem is to examine the growth dynamics of the number of unique protein-protein binding sites. As a result of our analysis, we found that the coverage of the human proteome with protein structural complexes remains very low (<10%), and the growth rate of the number of unique binding sites deposited in the PDB is much slower than the deposition rate for structures of protein complexes. Numerous computational methods have been designed to close the gap in the experimental structural coverage of proteomes and interactomes, particularly for human (1,17,21,25). These methods rely on two major strategies: comparative modeling/threading of protein complexes and docking (26). Questions arise, however, regarding the suitability of many structural complexes for modeling the human interactome (27,28) and the use of structural templates from other organisms for modeling human protein interfaces.

To understand the evolutionary patterns in the conservation of protein-binding sites, we considered protein-protein binding-site clusters (unique binding sites) and identified their most recent common ancestors (MRCAs). We analyzed the major expansion points in the number of unique binding sites along the taxonomic tree in the human lineage. We then identified protein domain families that provided the most novel protein binding sites, revealing their different and sometimes opposite evolutionary trends. The distribution of the number of unique binding sites per taxonomic rank revealed that the majority of conserved sites date back to the origin of cellular life, and that other binding sites were conserved within the major domains of life (Eukaryota, Bacteria, Archaea, and viruses). We compared the physicochemical properties of unique sites among different taxonomic groups—a markedly different task compared with previous analyses of the interfacial properties of nonredundant proteins or complexes. We observed that the most ancient unique binding sites were larger, more hydrophobic, more compact, and less planar than more recent ones, suggesting strong and perhaps obligatory associations in ancient complexes. On the other hand, binding sites that appeared more recently in evolution, particularly in Eukaryota, were less compact and more planar, and utilized more polar and aromatic residues, possibly due to their transient nature required by signaling pathways. Interestingly, binding sites that were conserved in viruses differed significantly from others and were larger, less compact, and generally more polar. Archaeal complexes, possibly as a result of their specific habitat requiring a high protein-complex stability, evolved the most compact binding sites. These sites were rich in hydrophobic and charged residues, with the largest number of salt bridges per charged interfacial residue.

MATERIALS AND METHODS

Searching for human homologs in the PDB

We used a snapshot of the PDB and National Center for Biotechnology Information (NCBI) Molecular Modeling Database (MMDB) available as of December 2014 (29,30). Protein-protein binding sites were annotated according to the IBIS database, October 2014 update (6). We considered a protein domain (with at least 20 residues and three secondary structure elements) as a unit of interaction and defined a binding site as a set of residues on one side of the protein interface that were in contact (within a distance of 4 Å) with any heavy atom on another protein chain in the biological assembly. We downloaded human proteome sequences from the Uniprot database (20,100 proteins) and used the longest isoform in cases having multiple isoforms (31). The entire human proteome was then searched against protein sequences from the PDB using DELTA-BLAST (32). Proteins from the PDB with at least 25% sequence identity to human protein sequences were considered homologous, and those that passed an E-value threshold of 0.001 were considered remotely similar to human proteins. In principle, due to domain recombination events, homology can be partial; thus, we excluded alignments with <50 amino acid residues covered.

Comparison of protein-protein binding sites

Binding sites were compared by the NCBI IBIS (http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi) (6). Biological assemblies (so-called biounits) corresponded to the first biounit chosen by the PDB (3). Protein domains with structurally observed protein-protein interactions were superimposed using the VAST structure alignment method (33) and the corresponding binding-site residues were clustered based on their similarity of structural and sequence features by a procedure described previously (6). Hereafter, binding-site representatives from each cluster are called unique binding sites. It should be mentioned that we only clustered together binding sites of domains belonging to the same CDD superfamily release 3.1.12 (34). Therefore, when protein domains from different CDD superfamilies had similar binding sites, we counted them as different binding-site clusters. In this way, we obtained a nonredundant set of protein-protein binding sites in which the number of unique binding sites was equal to the number of binding-site clusters. A schematic representation of the binding-site clusters is given in Fig. 1 *a*.

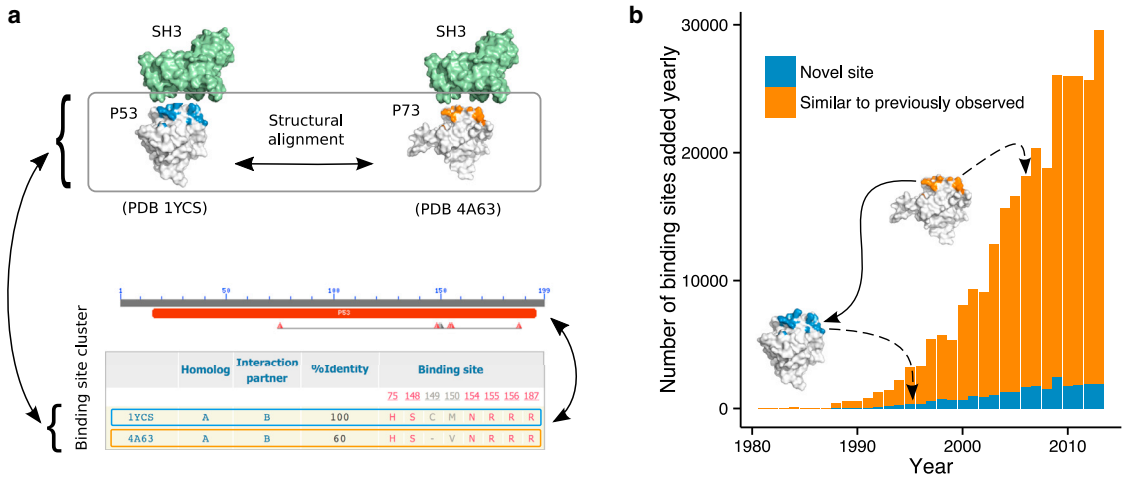

FIGURE 1 Protein-protein binding sites observed in structures of protein complexes. (*a*) An example of two pairs of interacting domains: P53 with SH3, and P73 with SH3. Proteins containing P53 and P73 domains are homologous and some of their binding-site residues (*blue* and *orange*) that are involved in interaction with SH3 are conserved, as shown in red in the lower panel. These domains were structurally superimposed using the VAST algorithm and their binding sites were clustered using the IBIS method. (*b*) Yearly growth of the number of protein-protein binding sites observed in protein complexes in the PDB for unique sites (*blue*) and for sites that are similar to previously observed sites from previously deposited structures (redundant, *orange*). To see this figure in color, go online.

Cladistic reconstruction

We reconstructed the major expansion points in the evolution of protein-protein interactions using the NCBI taxonomy (35). Not all taxa have ranks in this taxonomic classification. For instance, the taxonomic ranks above superkingdoms are not defined; thus, we show the taxon cellular organisms as a distinct rank. We defined the MRCA for a cluster of protein-protein binding sites by identifying the taxa of all proteins from a given cluster. The taxon from which all taxa in the binding-site cluster descended was considered the MRCA (illustrated in Fig. 2). In order to attribute protein-protein interactions to particular species, we excluded host-virus complexes, engineered chimeras, and interactions between proteins from different species.

The cladogram in Fig. 3 was generated with python scripts and visualized with Cytoscape 3.2 (36). The cladogram has two free parameters: the number of nonredundant structures of protein complexes in each unique binding-site cluster (r), and the number of unique clusters in a given taxon (n). We identified nonredundant PDB structures using a 95% sequence identity threshold. Everywhere in our analysis we required at least two nonredundant structures ($r > 1$) representing the binding-site cluster; thus, singletons were excluded. The resulting cladogram is provided in Cytoscape format in File S1 (see also Fig. S4 in the Supporting Material). For visualization clarity, we pruned the cladogram by showing only the taxa with at least five unique sites ($n > 4$).

Additionally, to assess the robustness of our cladistics reconstruction, we removed several species with many structures dominating the PDB database (see the list of species in Table S2) and recalculated the cladogram with these species excluded (Fig. S5).

Calculation of the physicochemical properties of binding sites

A set of unique binding sites was attributed to each taxon based on the MRCA of proteins in the corresponding binding-site cluster, as described in the "Cladistic reconstruction" section below. We calculated the physicochemical properties of unique binding sites in a given taxon by taking a random representative of each binding-site cluster assigned to the taxon. We then calculated the total number of residues in the binding sites and the proportions of charged (RKDE), polar (NQSTYGH), hydrophobic

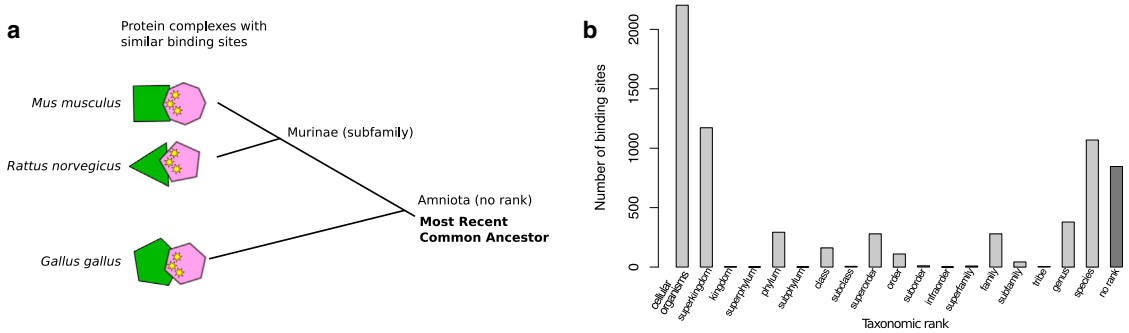

FIGURE 2 (*a*) Schematic representation of the MRCA (according to NCBI taxonomy) defined for three structures belonging to the same cluster of protein-protein binding sites in homologous proteins from mouse, rat, and chicken. (*b*) Distribution of the number of unique binding sites (each site is represented by at least two structures) plotted versus the taxonomic rank of their MRCA. Taxa without ranks (e.g., *Amniota*) are denoted as No Rank. To see this figure in color, go online.

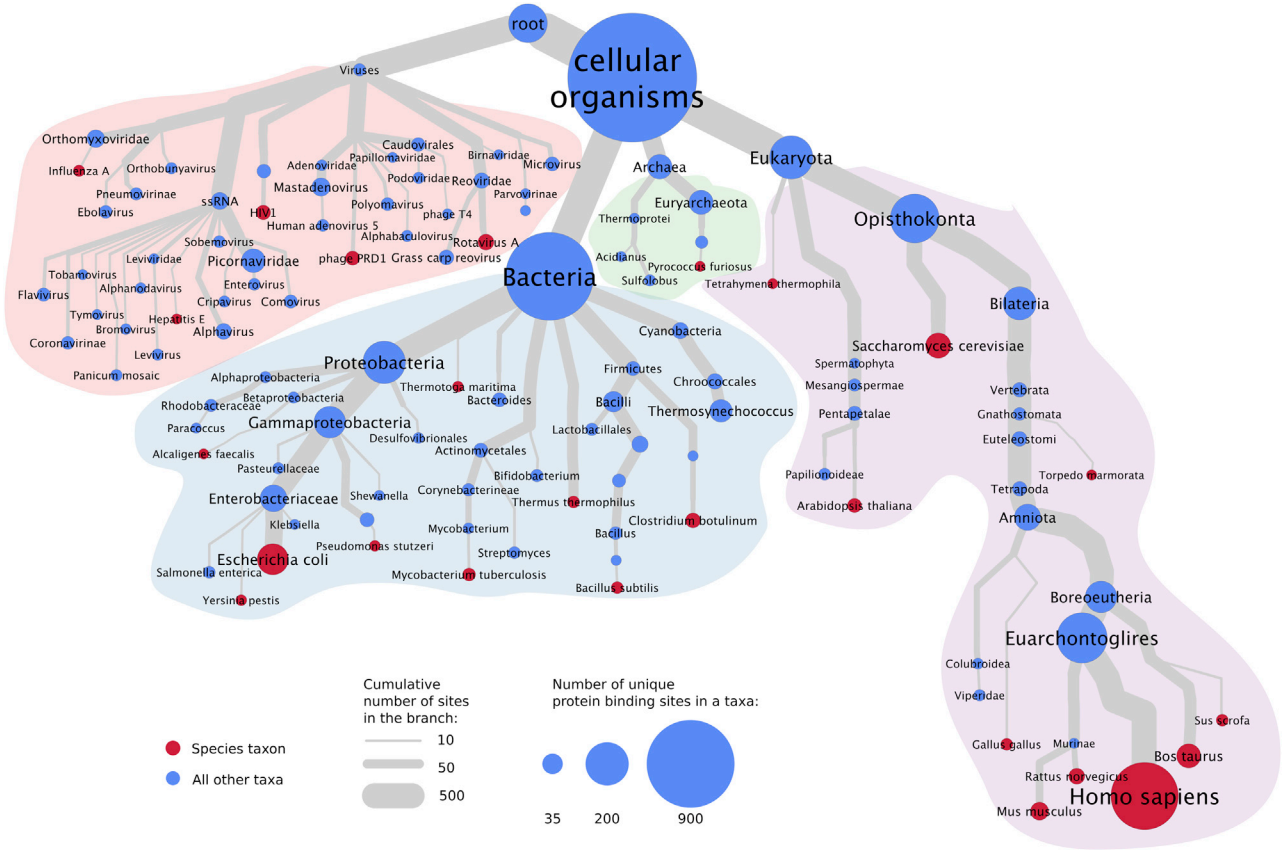

FIGURE 3 Cladogram of the evolutionary expansion of protein-protein binding sites based on a structural analysis of protein complexes. The cladogram is based on NCBI taxonomy. Each node denotes a taxon, which is the MRCA for the cluster of protein-protein binding sites (see Fig. 2). Red nodes denote species. We counted the number of unique binding sites (each represented by at least two nonredundant complexes) belonging to each taxon. The area of each node/circle is proportional to the number of unique sites. Edge lengths are arbitrary and do not represent evolutionary distances; however, the edge thickness scales logarithmically with the cumulative number of unique sites added in all taxa in the underlying branch. The branches of viruses, Bacteria, Archaea, and Eukaryota are colored, and the same colors are used in the other figures. To see this figure in color, go online.

(AVLIPFVMW), and aromatic (FYWH) amino acid residues in the binding sites.

We used VMD (37) and Python scripts to identify the contacts with a distance threshold of 4 Å between heavy atoms in representative structures from each binding-site cluster. For charged residues on the interface, we found all possible salt bridges as contacts between oppositely charged atoms. Fig. S6 shows the distribution of atomic protein-protein contacts per residue in the binding sites. Each representative binding site was characterized by the number of charged residues that formed salt bridges.

To assess the compactness of the binding sites, we calculated the radius of gyration, R_g, of each site: $R_g^2 = (1/N)\sum_i (r_i - \langle r \rangle)^2$, where r represents the coordinates of each heavy atom belonging to the binding-site residues. More compact sites will have a smaller radius of gyration. We fitted the scaling factor (a) and Flory exponent (ν) in $R_g = aN^{\nu}$ with nonlinear least squares for binding sites in different taxonomic groups. Additionally, we calculated the planarity of the binding sites, which Nooren and Thornton (38) described as the root mean-square deviation (RMSD) of C-α atoms in a given binding site from the plane fitted by least squares to the atomic coordinates of these atoms. See also "Planarity calculated for all heavy atoms" in Fig. S9.

Statistical analysis and plots were prepared using R (http://R-project.org/). All figures with physicochemical properties and amino acid content show mean values with 95% bootstrap confidence intervals. We compared all distributions using the Wilcoxon rank sum test with Bonferroni correction, and report the p values in Tables S3–S5.

RESULTS AND DISCUSSION

Coverage of proteomes by structural complexes: growth dynamics and beyond

We comprehensively assessed the diversity of structures and protein interfaces in the context of data deposition dates. The trend in Fig. 4 *a* shows that the number of structures deposited yearly is increasing, and soon will reach the mark of 10,000 structures per year. The biological units of approximately half of all structures represent protein-protein complexes. The most common type of complex is the dimer, and the number of dimers in the PDB keeps growing at persistently higher rates than the number of tetramers, trimers, and higher-order oligomeric states (Fig. 4 *a*, *inset*). However, as can be seen from this figure, since the early 2000s, significant progress has been made in resolving biological complexes with more than 20 subunits.

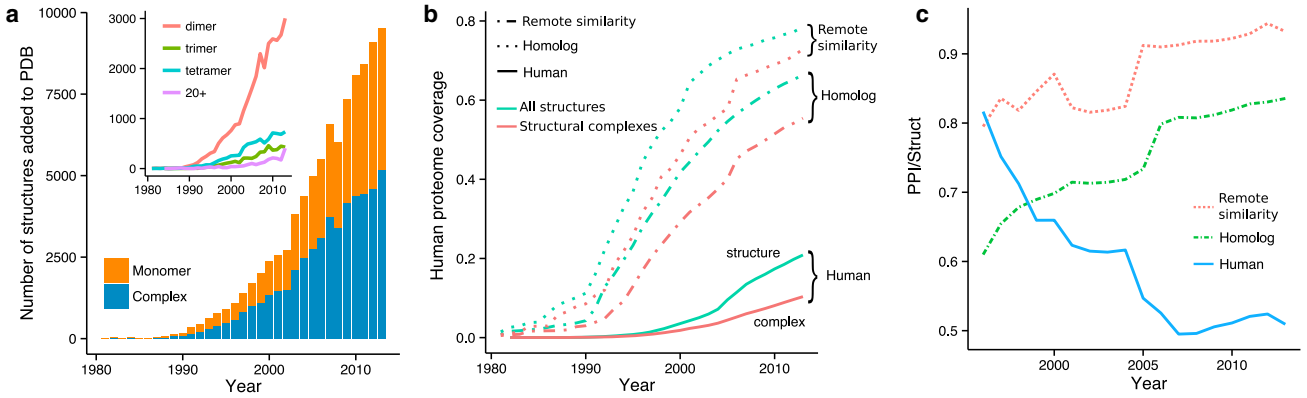

FIGURE 4 (*a*) Number of structures deposited in the PDB every year. All structures are split into biological assemblies with more than one protein chain (*blue*) and monomeric biological units (*orange*). Inset: number of biological assemblies with different oligomeric states. (*b*) The fraction of human proteins (20,100 total proteins were considered) that can be covered by PDB structures. We distinguish three levels of structural evidence: actual human protein structures (*solid line*), protein structures from other organisms that are homologous to human proteins (>25% sequence identity, *dashed line*), and protein structures from other organisms that are remotely similar to human proteins, with a sequence similarity detectable by Delta-BLAST with E-value < 0.001 (*dotted line*). Cases in which structures represent protein complexes are counted separately (*teal*) from the overall number of structures (*red*). (*c*) The ratio of the number of protein complexes to the total number of structures is shown for three groups of structural evidence. To see this figure in color, go online.

As was pointed out in the reports of the Protein Structure Initiative (PSI) and other structural-genomics endeavors, instead of trying to cover as many structures as possible, investigators should focus on resolving protein complexes to cover biochemical pathways that are relevant to human health (39). For this purpose, we addressed the coverage of human proteins with structures and structural protein-protein complexes, and investigated how this coverage has changed in the last several years. As is evident from Fig. 4 *b*, coverage of the human proteome with homologs of known structures has shown rapid growth since 1990, with signs of saturation since 2007. However, direct structural evidence for human proteins has lagged the first and final drafts of the human genome in 2000 and 2003, respectively, with a modest structural coverage of 20%.

Currently, structural complexes are available for <10% of human protein-coding genes (Fig. 4 *b*). We obtained Fig. 4 *c* by calculating the ratio of the number of complexes to the number of all structures, including monomers (values shown in red divided by the values shown in teal color in Fig. 4 *b*). Fig. 4 *c* suggests that based on the analysis of homologous or remotely similar protein complexes, ~80–90% of all human proteins may be involved in protein complexes, and half of the crystallized human proteins are currently characterized as monomers (as shown in Fig. 4 *c*).

Growth dynamics of unique binding sites and the emergence of novel interfaces

Fig. 1 *a* shows two examples of protein complexes with similar binding sites between tumor-suppressor proteins (P53 and P73 from the p53/p63/p73 family of genes) and the SH3 domain of proapoptotic factors from the ASPP2 family (PDB codes 1YCS and 4A63). Despite the structural variation in the interface and the insertion of two residues in the P73 binding site (*gray residues* in the Fig. 1 *a* alignment), these two sites can be considered to be similar and belong to the same binding-site cluster, according to the structure and sequence similarity of binding sites calculated by the IBIS algorithm.

The growth dynamics of unique binding sites is shown in Fig. 1 *b*. A binding site is considered novel if no similar binding sites were available in the PDB before its deposition date. One can see an astonishing redundancy in the terms of protein-binding sites in the PDB: only ~10% of all binding sites deposited each year are novel, and overall, only 27,000 unique binding sites are found in the whole PDB. The peak in the number of novel binding sites around the year 2009 signifies the success of structural genomics initiatives. However, since 2009, the number of novel binding sites has slightly dropped, reflecting the saturation in the number of protein complexes deposited each year (see previous section and Fig. 4). This indicates that the coverage of distinct protein-binding interfaces may not be a priority in many structural studies. The structural characterization of complexes should thus become one of the most important challenges and future targets in the field of structural genomics (39). Many computational methods (1,17,21,25) are being used in an attempt to fill the gap in experimental coverage of protein complexes, and these methods rely on comparative modeling of structures, threading, or docking (26). It is unclear, however, to what extent protein interactions are conserved and whether protein complexes from remotely related species can be used to infer interactions in humans. The reliability of computationally derived protein-protein interactomes depends on

our understanding of evolutionary constraints and the completeness of experimental structural characterizations of diverse interfaces (4).

Tracing back the evolution of protein-binding sites to the origin of life

Evolutionary constraints for protein-protein binding sites may be obtained from the conservation patterns of similar binding sites within the binding-site clusters. Therefore, we identified the evolutionary roots of each site by finding the MRCA (see the corresponding section in Materials and Methods) of the corresponding binding-site cluster. Fig. 2 *a* shows domains with similar binding sites (*violet circles with yellow stars*) belonging to proteins from different species: mouse, rat, and chicken. Based on the NCBI taxonomy, we identified the Amniota taxon as the MRCA for this cluster. Due to the highly redundant nature of the PDB, we only considered cases in which a binding-site cluster is represented by at least two protein complexes with at most 95% sequence identity between the proteins. In Fig. 2 *b* we show the distribution of unique binding sites by the rank of the corresponding MRCA taxon (Fig. S1 shows this distribution without a nonredundancy requirement). The distribution of the number of unique binding sites per taxonomic rank reveals an interesting pattern, namely, the majority of unique sites date back to the origin of cellular life. At the same time, phylum, class, and superorder nodes on the taxonomic tree show a relatively moderate expansion of protein-binding sites, whereas genus and species ranks have a large number of unique sites. The latter may represent truly species-specific interactions or it could be a result of incomplete and biased structure sampling in the PDB. On the other hand, these clusters may indicate the presence of conserved binding sites in paralogous proteins from the same organism. In this study we did not distinguish between these two cases; therefore, we did not analyze the properties of binding-site clusters attributed to species and subspecies levels.

Fig. 3 shows a cladogram in which the hierarchy of taxa corresponds to the NCBI taxonomy and the Tree of Life. The area of each node is proportional to the number of unique binding sites that originate from the corresponding taxon, and the width of edges represents the cumulative number of sites from the underlying branch in the logarithmic scale. Taxa with the largest nodes (*circles*) represent evolutionary points where a major expansion of protein-protein binding sites occurred. Fig. 3 shows only the taxa with five or more unique binding sites (see Fig. S4 and File S1 for a complete cladogram). To prove the robustness of our calculations, we rebuilt the cladogram after excluding the most represented species in the PDB (Fig. S5), but the taxon distribution of the unique sites above the genus level was not affected, and major evolutionary expansion points were preserved. It should be mentioned that due to the nature of molecular evolution (40) and our structural evidence-based analysis, we could only observe binding sites that endured natural selection and are represented in modern-day species. Therefore, the estimates of binding-site expansions presented here can be considered as low boundary estimates. We also excluded intrachain binding sites in our analysis, and considered only domain-domain binding sites involved in interchain interactions. Therefore, the expansion of domain architectures or domain promiscuities should not affect our estimates of the expansion of the number of binding sites.

Cellular organisms are the largest node on the cladogram rooting Bacteria, Archaea, and Eukaryota, and it has the largest number of unique binding sites. These sites mainly represent homooligomeric interactions between highly conserved proteins and describe interactions within the essential molecular complexes shared between all superkingdoms, such as amino-acyl synthetases and ribosomal protein subunits. Archaeal organisms share interactions between the components of multisubunit enzymes that are unique to Archaea, such as methanogenic methyl-coenzyme M reductase (MCR), whereas Bacteria share many more interactions related to specific bacterial enzymes, such as those with a hot-dog fold, which are involved in thioester hydrolysis and fatty acid metabolism, and various metalloproteins from the Glo_EDI_BRP-like superfamily. In addition, Bacteria have many specific transcription regulators that are involved in the xenobiotics-resistance response, as in the case of the HTH_XRE superfamily. Consistent with our findings, it was previously shown that the node degree in interaction networks was higher for more ancient taxonomic nodes (41) and ancient proteins had high connectivity and centrality in protein interactomes (42). Moreover, it was observed that protein interaction networks partitioned into two subnetworks: one corresponding to the most ancient interactions and one related to the recently emerged interactions in animal evolution that are involved in cell division and cell communication (41). A high level of horizontal gene transfer between species, a mechanism by which new interactions may emerge, is also characteristic for early stages of evolution (43).

We then analyzed the functions of interacting protein domains by identifying the CDD domain superfamilies that contributed the largest number of binding sites (Fig. 3). For example, the nitric oxide (NO)-synthase family represented in Bacteria and Eukaryota plays important roles in cell communication, immune response, and oxidative stress defense, and is characterized by many unique binding sites (Fig. S2; Table S1). The NO molecule is heavily used in cell signaling, and typically homodimeric NO-synthase consists of several specific domains and is coupled to various cofactors. Its interactions with other proteins, such as calcium sensors, are diverse and critical for its function (44). Another superfamily represents ATP-binding cassette (ABC) transporters, which are ubiquitous proteins with

essential functions. The diversity of ABC proteins results in a large number of unique protein binding sites in this superfamily. Several other superfamilies, such as SDR, TIM, Thioredoxin-like, Ferritin-like, NTN-hydrolase, and immunoglobulins, are notable examples of functionally diverse domains with highly designable structural folds (45). The ubiquitous nature of these domains is not surprising; however, it is remarkable that many of their protein-protein binding modes and interfaces are conserved dating back to the origin of cellular life.

Tracing back the evolution of binding sites in Eukaryota and the human lineage

We traced the most prominent taxa in terms of protein-binding-site expansion in the lineage leading toward human, which included Opisthokonta, Bilateria, Amniota, Boreoeutheria, and Euarchontoglires (Fig. 3). Opisthokonta (originating 1.5 Gya (46)) and Euarchontoglires (<100 Mya) nodes are characterized by the largest expansion of the number of unique binding sites. Some protein-binding sites that are conserved in Eukaryota, and particularly in Opisthokonta, include C-terminal extensions of proteins of S60 ribosomal subunit. These extensions specifically interact with each other in ribosomal proteins and may also interact with specific initiation factors (47). At the same time, Euarchontoglires developed a number of interactions between regulatory and signaling proteins such as chemokines, TNF cytokines, RNase A, kinases, and calcium-binding EF-hand domains. Additionally, the superfamilies that contributed to the expansion of protein-binding sites in these taxa, such as immunoglobulins and MHC class I and II antigens, are involved in the immune system.

We compared the evolutionary trends of several CDD domain superfamilies in the context of expansion of protein-binding sites in taxa leading to the human lineage. Fig. 5 shows four superfamilies and the number of unique binding sites they provide in each taxon. Some domain families, such as the calcium-binding EF-hand motif, have always been employed as modules in signal transduction, and their binding-site expansion shows constant rates at different taxonomic nodes. On the other hand, ABC-related ATPases show one large binding-site expansion at the level of cellular organisms followed by a second expansion in mammals, demonstrating a continued growth in the number of distinct binding sites and the evolution of novel, specific interactions. The immunoglobulin superfamily shows the opposite trend. This fold is employed in several basic enzymatic functions, and novel binding modes did not appear before the emergence of the immune system in animals. The protein-binding sites of the fibronectin type 3 (FN3) domain, which is involved in multiple processes in the extracellular matrix, including blood clot formation, have shown expansion since the emergence of placental mammals.

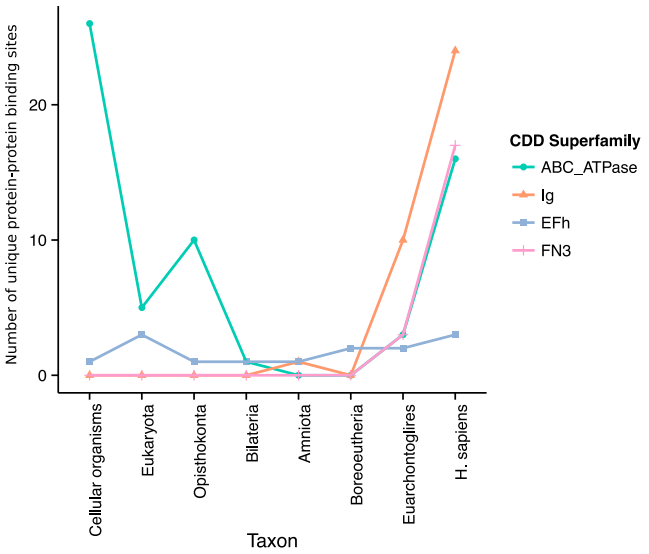

FIGURE 5 Evolutionary trends in the expansion of protein-protein binding sites of representative CDD domain superfamilies across the taxa leading to the human lineage. The taxa are ordered from left to right from the most ancient ones to human. Here, Eukaryota refers to the binding sites in a particular taxon (node in Fig. 3) rather than to the whole domain of life and the taxonomic branch. The numbers of protein-binding sites in other superfamilies are provided in Table S1. ABC_ATPase, ATP-binding cassette transporter nucleotide-binding domain; Ig, immunoglobulin domain; EFh, EF-hand (calcium binding motif, a diverse superfamily of calcium sensors and calcium signal modulators); FN3, fibronectin type 3 domain (one of three types of internal repeats found in the plasma protein fibronectin). To see this figure in color, go online.

Physicochemical properties of binding sites and their taxonomic distribution

In this section, we address the question of whether there are any physicochemical properties that would distinguish ancient binding sites from sites that emerged more recently in evolution. Previous studies reported that the physicochemical properties of binding sites may depend on the type of protein complex involved, e.g., homo- or heterooligomers (16,48), obligate or transient (38,49–51). These studies, however, did not consider redundancy and possible similarities between binding sites from different types of complexes. Here, we considered differences in physicochemical properties among unique binding sites in major branches on the Tree of Life (Archaea, Bacteria, and Eukaryota); in viruses, the latest ancestor of all cellular organisms; and among taxa in the eukaryotic lineage leading toward human.

First, we compared the sizes of binding sites measured as the number of residues. As can be seen in Fig. 6, Eukaryota have the smallest sites compared with Bacteria, Archaea, and viruses, whereas viruses have the largest binding sites on average (p values are given in the Supporting Material). The first observation can be attributed to the larger number of transient interactions that are involved in regulatory pathways, whereas the increased sizes of viral binding sites stem

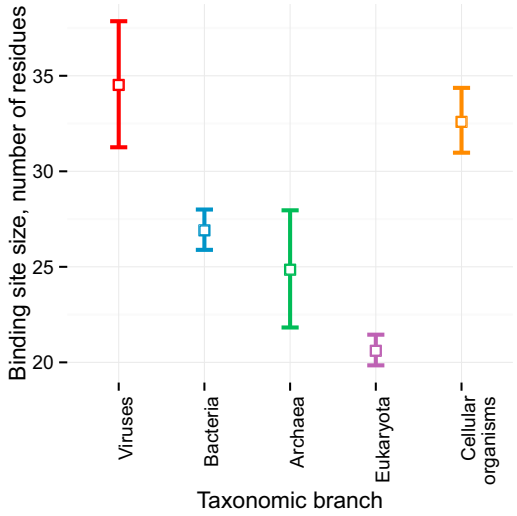

FIGURE 6 Sizes of the protein-protein binding sites (measured in number of residues) for all unique binding sites that have been attributed to ancestral taxa (Fig. 3). Squares denote mean values; error bars show 95% confidence intervals; *p* values for the corresponding pairwise tests are listed in Table S3. To see this figure in color, go online.

from large interfaces in viral capsids. Consistent with these results, it was previously shown for homooligomeric protein families that the most ancient binding modes tended to involve symmetrical binding arrangements with larger interfaces, whereas recently evolved binding modes more often exhibited asymmetrical arrangements and smaller interfaces (16). Moreover, many novel interactions were shown to be transient (52) and involve multibinding interfaces (53).

It is well known that archaeal and bacterial genomes and proteomes differ in their nucleotide and amino acid compositions (54), and that the composition of protein-protein binding sites may also differ among species (55). However, the evolutionary trends in the composition and physicochemical properties of unique protein-binding sites have not been analyzed in detail previously. We found that archaeal binding sites showed the compositional trends (Fig. 7) common to all Archaea in general. They had the smallest fraction of polar and aromatic residues, and the largest fraction of charged and hydrophobic residues in binding sites compared with other superkingdoms. This trend could be explained by the demand for stability and selection against aggregation at higher temperatures (56), which are typical for the habitat of many archaea.

Salt bridges between charged residues on protein interfaces play important roles in molecular recognition and the specificity of interactions, particularly at higher temperatures (56). Salt bridges on interfaces were also found to have very distinct evolutionary conservation patterns (57). We considered all charged residues in unique binding sites and calculated the ratio of the number of charged residues involved in salt bridges to the overall number of charged residues in the binding site. As can be seen in Fig. 8, archaeal

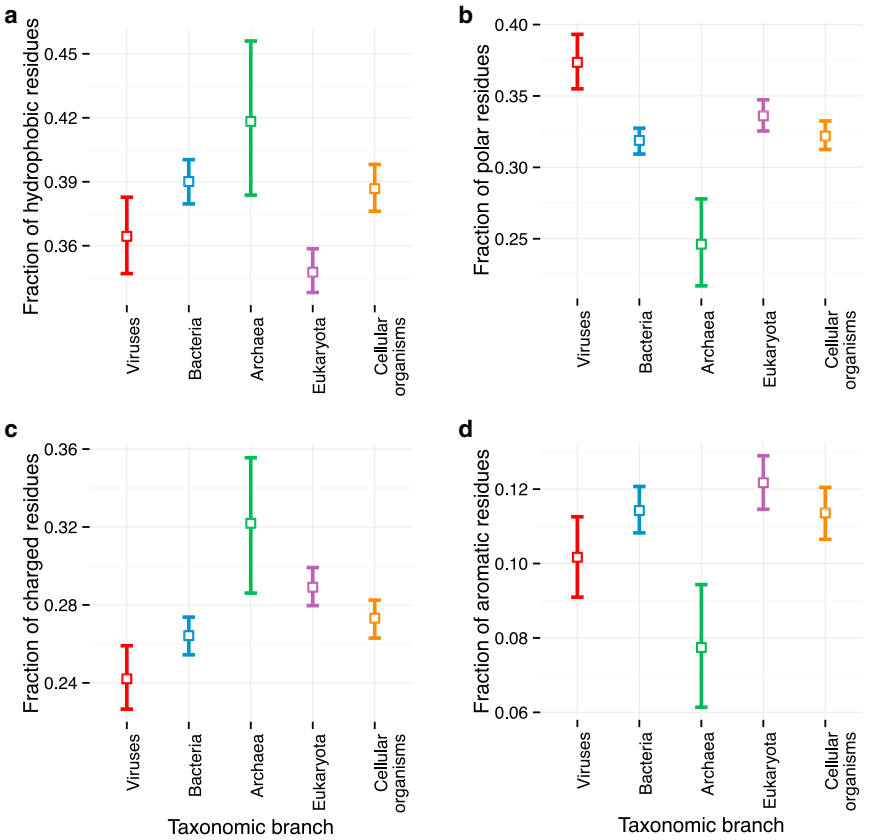

FIGURE 7 Fractions of binding-site residues with different physicochemical properties for major taxonomic branches. Squares denote mean values; error bars show 95% confidence intervals. (*a–d*) Amino acid residues: (*a*) hydrophobic (AVLIPFVMW), (*b*) polar (NQSTGYH), (*c*) charged (RKDE), and (*d*) aromatic (FYWH). The *p* values for the corresponding pairwise tests are listed in Table S3. To see this figure in color, go online.

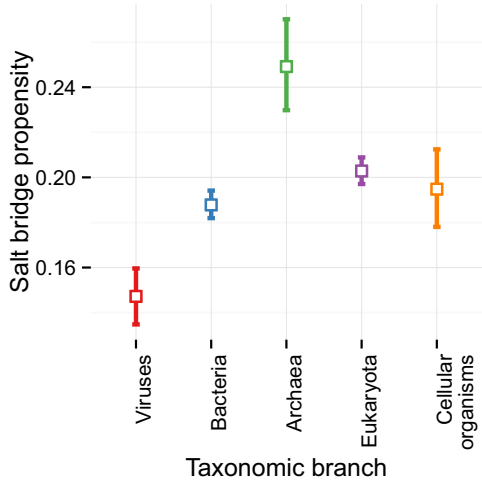

FIGURE 8 Proportion of charged residues that form salt bridges relative to the overall number of charged residues in binding sites. Squares denote mean values; error bars show 95% confidence intervals; p values for pairwise tests are listed in Table S4. Fig. S7 shows the corresponding probability densities. To see this figure in color, go online.

binding sites have the highest recruitment of charged residues into salt bridges compared with other branches of life. This observation may be explained in part by the specific requirements of the archaeal habitat to enhance protein and protein complex stability. Additionally, the universally conserved binding sites of all cellular organisms have a slightly higher proportion of charged residues involved in the formation of salt bridges compared with Bacteria and Eukaryota (all p values are given in Tables S3–S5).

Compactness is another important physical attribute of protein-binding sites. We used the radius of gyration, R_g, as a measure of compactness. We analyzed how the radius of gyration scales with the binding-site size for different taxonomic branches, quantitatively expressed by different Flory exponents (ν), as described in Materials and Methods (Fig. 9 *a*). A Flory exponent of 1/3 is characteristic for natural proteins, $\nu = 1/2$ describes a result of random walk of an ideal polymer chain, and denatured proteins have $\nu = \sim 3/5$. Although the binding sites represent only the interacting parts of the domain globules, the values of ν that we obtained are within the above-described ranges. Viruses and Eukaryota had the lowest compactness of binding sites, with ν-values close to that of a random coil (0.5 and 0.42, respectively; Fig. 9 *a*). The binding sites in the most ancient cellular-organisms group were the most compact and had $\nu = 0.37$, which is close to that of a fully compact natural protein. Archaeal binding sites were shorter than sites in other taxonomic branches; therefore, we could only approximate the values for sites with <75 residues. Overall, the compactness of bacterial and archaeal binding sites is close to that of the most conserved sites in cellular organisms, with $\nu = 0.38$ and 0.42, respectively.

Additionally, we calculated the planarity of binding sites as defined previously by Nooren and Thornton (38), and measured it as the RMSD of binding-site coordinates from the plane, which was obtained by least-squares fitting to all C-α atoms in the binding site (Fig. 8 *b*). Planarity is an important geometric characteristic of the interface and is linked to its compactness. Planarity and increased polar content were previously shown to be a characteristic of weak interfaces (38). It was previously argued that packing of flat interacting surfaces formed by hydrogen-bonded secondary structure elements is responsible in large part for the degeneracy of the structural space of protein-protein interfaces (24). Analysis of the deviation of binding-site residues from planarity showed that unique binding sites characteristic for Eukaryota were more planar than those of other taxonomic groups, except for viruses (Fig. 9 *b*), reflecting the fact that many eukaryotic proteins interact through weak transient interactions. This could possibly be explained by a large number of interactions in signaling that are modulated by weak planar (and in many cases linear-in-sequence) interfaces. The universally conserved binding sites in all cellular organisms, on the other hand, were characterized by the smallest planarity (Fig. 9 *b*), suggesting strong and perhaps obligatory associations in ancient complexes. Thus, ancient binding sites were the largest in size,

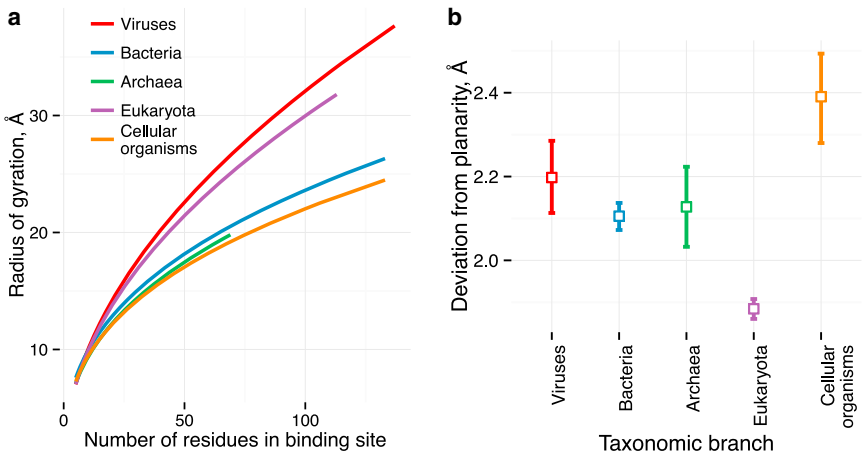

FIGURE 9 (*a*) Compactness of binding sites measured as the radius of gyration (R_g). The lines show the nonlinear fitting of dependency of R_g on the number of residues in the binding site. Corresponding Flory exponents (ν): 0.5 in viruses, 0.49 in Eukaryota, 0.42 in Archaea, 0.38 in Bacteria, and 0.37 in cellular organisms. (*b*) Deviation from planarity of binding sites measured in Å, RMSD (see Materials and Methods). Squares denote mean values; error bars show 95% confidence intervals; p values for pairwise tests are listed in Table S5. To see this figure in color, go online.

involved many salt bridges, and were the most compact and least planar, whereas binding sites in viruses were the least compact (Fig. 8 *a*) and generally more polar (Fig. 7).

The mechanisms of the evolution of protein-protein interactions and their oligomeric states were previously studied in the context of interactomes (58,59). It was found, for example, that the evolutionary rate of protein-protein interactions is three orders of magnitude lower than the rate of protein sequence evolution (60). In addition, more ancient protein complexes were suggested to be less flexible than more evolutionarily recent proteins (61). Interestingly, more evolutionarily recent protein domains are more likely to be disordered (62); however, in our analysis we did not account for binding sites formed through disordered regions because these regions may lack coordinates.

CONCLUSIONS

In this work, we studied the growth of structural data from the PDB, with a particular focus on protein complexes. Despite the fact that a quarter of all protein structures are from human, the coverage of the human proteome with protein complexes remains very low (<10%). It is possible, however, to increase this coverage by inferring protein interactions based on homologous protein structures and complexes from other species. The number of unique protein-protein binding sites defined at the level of protein domains is relatively low (~27,000) and its growth rate in the PDB is much slower than that of structural complexes. Thus, although the majority of protein folds have already been exemplified by experimentally determined structures, and the majority of sequences can be at least partially structurally modeled, the focus in structural biology has to be shifted to the characterization of largely undersampled structures such as protein assemblies. Novel complexes and interfaces are especially important in this respect for the construction of structure-validated interactomes, therapeutic modulation of protein-protein interactions, and rational protein design to create protein complexes with the desired specificities (63–67).

Here, we traced back the evolution of binding sites and their taxonomical patterns. We found that a large number of binding sites are conserved universally among all cellular organisms and some sites are even conserved between cellular organisms and viruses. We explored the major expansion points in the evolution of protein-binding sites and their functional characteristics. These trends included the latest expansions driven mainly by the acquisition of protein interactions in cellular signaling and immune response pathways. This provides a way to infer interactions from other species, to build structure-based interactomes and improve the structural coverage of human interactions identified by high-throughput experimental methods (25,68).

In addition, we thoroughly explored the biophysical properties of binding sites with respect to their phyletic origin. Apparently, the most ancient binding sites differ from the most recently emerged sites in terms of size, amino acid composition, compactness, and planarity. The latter constitute presumably transient interactions with smaller, less compact, and more planar binding sites consisting of more polar and aromatic residues, and typically are involved in cell signaling and regulation. We also found that viral and archaeal protein-binding sites are significantly different from the other sites. Viral proteins have the largest and least compact sites, and the binding sites in archaeal proteins are enriched in charged residues that form salt bridges.

SUPPORTING MATERIAL

Nine figures, five tables, and one cladogram are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)00667-0.

AUTHOR CONTRIBUTIONS

A.R.P. and A.G. designed the research. A.G., A.K.S., and B.A.S. contributed analytic tools. A.G. and A.K.S. analyzed data. All authors wrote and approved the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

REFERENCES

1. Rolland, T., M. Taşan, …, M. Vidal. 2014. A proteome-scale map of the human interactome network. *Cell.* 159:1212–1226.
2. Havugimana, P. C., G. T. Hart, …, A. Emili. 2012. A census of human soluble protein complexes. *Cell.* 150:1068–1081.
3. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
4. Goncearenco, A., B. A. Shoemaker, …, A. R. Panchenko. 2014. Coverage of protein domain families with structural protein-protein interactions: current progress and future trends. *Prog. Biophys. Mol. Biol.* 116:187–193.
5. Berman, H. M., B. Coimbatore Narayanan, …, C. Zardecki. 2013. Trendspotting in the Protein Data Bank. *FEBS Lett.* 587:1036–1045.
6. Shoemaker, B. A., D. Zhang, …, A. R. Panchenko. 2012. IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.* 40:D834–D840.
7. Shoemaker, B. A., D. Zhang, …, A. R. Panchenko. 2010. Inferred Biomolecular Interaction Server–a web server to analyze and predict protein interacting partners and binding sites. *Nucleic Acids Res.* 38:D518–D524.
8. Winter, C., A. Henschel, …, M. Schroeder. 2012. Protein interactions in 3D: from interface evolution to drug discovery. *J. Struct. Biol.* 179:347–358.
9. Sudha, G., R. Nussinov, and N. Srinivasan. 2014. An overview of recent advances in structural bioinformatics of protein-protein interactions and a guide to their principles. *Prog. Biophys. Mol. Biol.* 116:141–150.
10. Xu, Q., and R. L. Dunbrack. 2011. The protein common interface database (ProtCID)—a comprehensive database of interactions of homologous proteins in multiple crystal forms. *Nucleic Acids Res.* 39:D761–D770.

11. Finn, R. D., B. L. Miller, …, A. Bateman. 2014. iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Res.* 42:D364–D373.
12. Janin, J., and F. Rodier. 1995. Protein-protein interaction at crystal contacts. *Proteins.* 23:580–587.
13. Jones, S., A. Marin, and J. M. Thornton. 2000. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* 13:77–82.
14. Shoemaker, B. A., A. R. Panchenko, and S. H. Bryant. 2006. Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci.* 15:352–361.
15. Aloy, P., H. Ceulemans, …, R. B. Russell. 2003. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* 332:989–998.
16. Dayhoff, J. E., B. A. Shoemaker, …, A. R. Panchenko. 2010. Evolution of protein binding modes in homooligomers. *J. Mol. Biol.* 395: 860–870.
17. Cukuroglu, E., A. Gursoy, …, O. Keskin. 2014. Non-redundant unique interface structures as templates for modeling protein interactions. *PLoS One.* 9:e86738.
18. Zhang, Q. C., D. Petrey, …, B. H. Honig. 2010. Protein interface conservation across structure space. *Proc. Natl. Acad. Sci. USA.* 107:10896–10901.
19. Henschel, A., W. K. Kim, and M. Schroeder. 2006. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics.* 22:550–555.
20. Aloy, P., and R. B. Russell. 2004. Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.* 22:1317–1321.
21. Petrey, D., and B. Honig. 2014. Structural bioinformatics of the interactome. *Annu. Rev. Biophys.* 43:193–210.
22. Tuncbag, N., A. Gursoy, …, O. Keskin. 2008. Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.* 381: 785–802.
23. Garma, L., S. Mukherjee, …, Y. Zhang. 2012. How many protein-protein interactions types exist in nature? *PLoS One.* 7:e38913.
24. Gao, M., and J. Skolnick. 2010. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. USA.* 107:22517–22522.
25. Tyagi, M., K. Hashimoto, …, A. R. Panchenko. 2012. Large-scale mapping of human protein interactome using structural complexes. *EMBO Rep.* 13:266–271.
26. Vakser, I. A. 2013. Low-resolution structural modeling of protein interactome. *Curr. Opin. Struct. Biol.* 23:198–205.
27. Mika, S., and B. Rost. 2006. Protein-protein interactions more conserved within species than across species. *PLOS Comput. Biol.* 2:e79.
28. Lewis, A. C., N. S. Jones, …, C. M. Deane. 2012. What evidence is there for the homology of protein-protein interactions? *PLOS Comput. Biol.* 8:e1002645.
29. Madej, T., K. J. Addess, …, S. H. Bryant. 2012. MMDB: 3D structures and macromolecular interactions. *Nucleic Acids Res.* 40:D461–D464.
30. Dutta, S., and H. M. Berman. 2005. Large macromolecular complexes in the Protein Data Bank: a status report. *Structure.* 13:381–388.
31. Consortium, T. U. 2015. UniProt: a hub for protein information. *Nucleic Acids Res.* 43:D204–D212.
32. Boratyn, G. M., A. A. Schäffer, …, T. L. Madden. 2012. Domain enhanced lookup time accelerated BLAST. *Biol. Direct.* 7:12.
33. Gibrat, J. F., T. Madej, and S. H. Bryant. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 6:377–385.
34. Marchler-Bauer, A., C. Zheng, …, S. H. Bryant. 2013. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* 41:D348–D352.
35. NCBI Resource Coordinators. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 43:D6–D17.
36. Shannon, P., A. Markiel, …, T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498–2504.
37. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–38.
38. Nooren, I. M., and J. M. Thornton. 2003. Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* 325:991–1018.
39. Montelione, G. T. 2012. The Protein Structure Initiative: achievements and visions for the future. *F1000 Biol. Rep.* 4:7.
40. Wolf, Y. I., and E. V. Koonin. 2013. Genome reduction as the dominant mode of evolution. *BioEssays.* 35:829–837.
41. Bezginov, A., G. W. Clark, …, E. R. Tillier. 2013. Coevolution reveals a network of human proteins originating with multicellularity. *Mol. Biol. Evol.* 30:332–346.
42. Peterson, G. J., S. Pressé, …, K. A. Dill. 2012. Simulated evolution of protein-protein interaction networks with realistic topology. *PLoS One.* 7:e39052.
43. Cohen, O., U. Gophna, and T. Pupko. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol. Biol. Evol.* 28:1481–1489.
44. Kone, B. C., T. Kuncewicz, …, Z.-Y. Yu. 2003. Protein interactions with nitric oxide synthases: controlling the right time, the right place, and the right amount of nitric oxide. *Am. J. Physiol. Renal Physiol.* 285:F178–F190.
45. Thornton, J. M., C. A. Orengo, …, F. M. Pearl. 1999. Protein folds, functions and evolution. *J. Mol. Biol.* 293:333–342.
46. Hedges, S. B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* 3:838–849.
47. Klinge, S., F. Voigts-Hoffmann, …, N. Ban. 2011. Crystal structure of the eukaryotic 60S ribosomal subunit in complex with initiation factor 6. *Science.* 334:941–948.
48. Hashimoto, K., H. Nishi, …, A. R. Panchenko. 2011. Caught in self-interaction: evolutionary and functional mechanisms of protein homooligomerization. *Phys. Biol.* 8:035007.
49. Ofran, Y., and B. Rost. 2003. Analysing six types of protein-protein interfaces. *J. Mol. Biol.* 325:377–387.
50. Bahadur, R. P., P. Chakrabarti, …, J. Janin. 2004. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.* 336: 943–955.
51. Dey, S., A. Pal, …, J. Janin. 2010. The subunit interfaces of weakly associated homodimeric proteins. *J. Mol. Biol.* 398:146–160.
52. Ozbabacan, S. E. A., H. B. Engin, …, O. Keskin. 2011. Transient protein-protein interactions. *Protein Eng. Des. Sel.* 24:635–648.
53. Tyagi, M., B. A. Shoemaker, …, A. R. Panchenko. 2009. Exploring functional roles of multibinding protein interfaces. *Protein Sci.* 18:1674–1683.
54. Goncearenco, A., B.-G. Ma, and I. N. Berezovsky. 2014. Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. *Nucleic Acids Res.* 42:2879–2892.
55. Levy, E. D. 2010. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.* 403:660–670.
56. Ma, B.-G., A. Goncearenco, and I. N. Berezovsky. 2010. Thermophilic adaptation of protein complexes inferred from proteomic homology modeling. *Structure.* 18:819–828.
57. Zhao, N., B. Pang, …, D. Korkin. 2011. Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. *Protein Sci.* 20:1275–1284.
58. Andreani, J., and R. Guerois. 2014. Evolution of protein interactions: from interactomes to interfaces. *Arch. Biochem. Biophys.* 554:65–75.
59. Levy, E. D., and S. Teichmann. 2013. Structural, evolutionary, and assembly principles of protein oligomerization. *Prog. Mol. Biol. Transl. Sci.* 117:25–51.

60. Qian, W., X. He, …, J. Zhang. 2011. Measuring the evolutionary rate of protein-protein interaction. *Proc. Natl. Acad. Sci. USA.* 108:8725–8730.

61. Marsh, J. A., and S. A. Teichmann. 2014. Protein flexibility facilitates quaternary structure assembly and evolution. *PLoS Biol.* 12:e1001870.

62. Moore, A. D., and E. Bornberg-Bauer. 2012. The dynamics and evolutionary potential of domain loss and emergence. *Mol. Biol. Evol.* 29:787–796.

63. Huang, P. S., J. J. Love, and S. L. Mayo. 2007. A de novo designed protein protein interface. *Protein Sci.* 16:2770–2774.

64. Khare, S. D., and S. J. Fleishman. 2013. Emerging themes in the computational design of novel enzymes and protein-protein interfaces. *FEBS Lett.* 587:1147–1154.

65. Levin, K. B., O. Dym, …, D. S. Tawfik. 2009. Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nat. Struct. Mol. Biol.* 16:1049–1055.

66. Melero, C., N. Ollikainen, …, T. Kortemme. 2014. Quantification of the transferability of a designed protein specificity switch reveals extensive epistasis in molecular recognition. *Proc. Natl. Acad. Sci. USA.* 111:15426–15431.

67. Lua, R. C., D. C. Marciano, …, O. Lichtarge. 2014. Prediction and redesign of protein-protein interactions. *Prog. Biophys. Mol. Biol.* 116:194–202.

68. Aloy, P., and R. B. Russell. 2002. Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. USA.* 99:5896–5901.

Structural perspectives on evolutionary expansion of unique protein-protein binding sites

Alexander Goncearenco, Alexey K. Shaytan, Benjamin A. Shoemaker and Anna R. Panchenko

Supplementary Materials

Figures: **page**

Tables:

Additional File 1:
Cladogram.cys (in Cytoscape 3.2.2 session format)

Figures

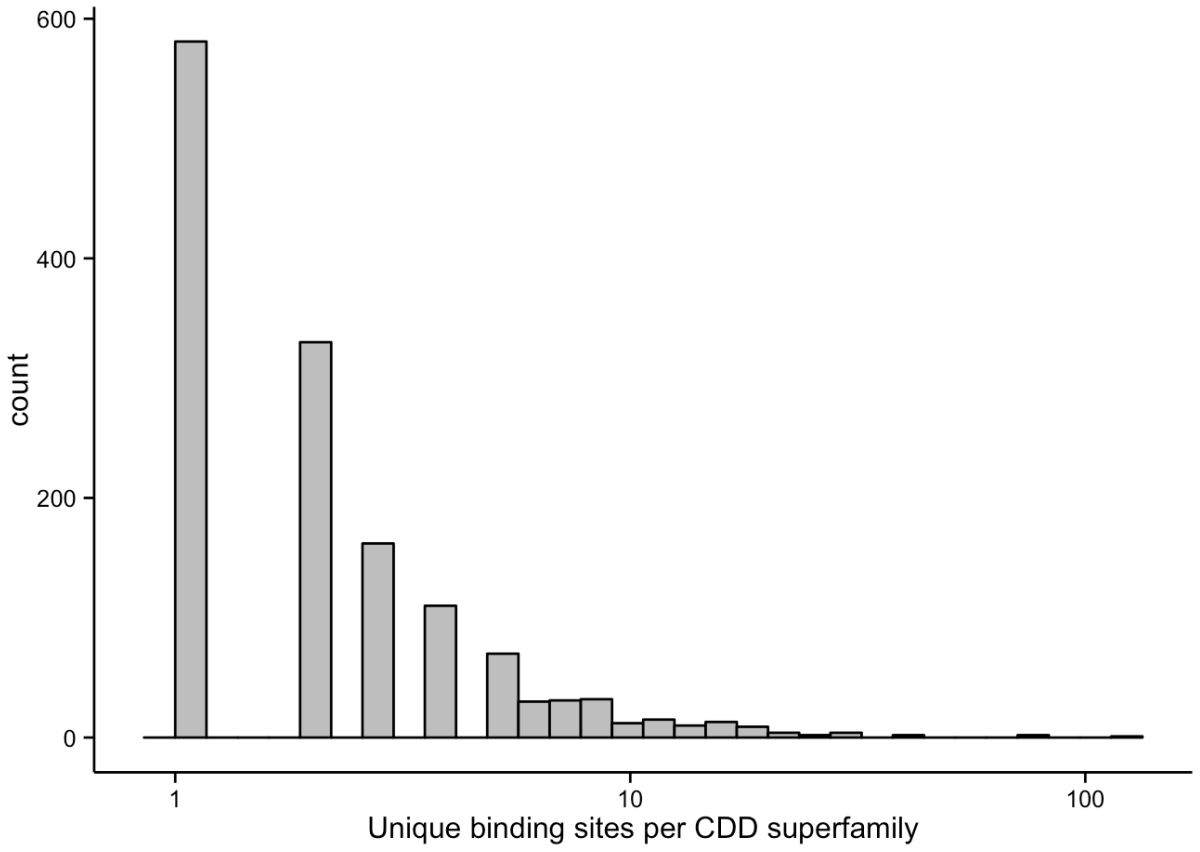

FIGURE S1 Distribution of unique binding sites per CDD superfamily of protein domains. Top 20 superfamilies contributing the most unique binding sites are listed in Table S1.

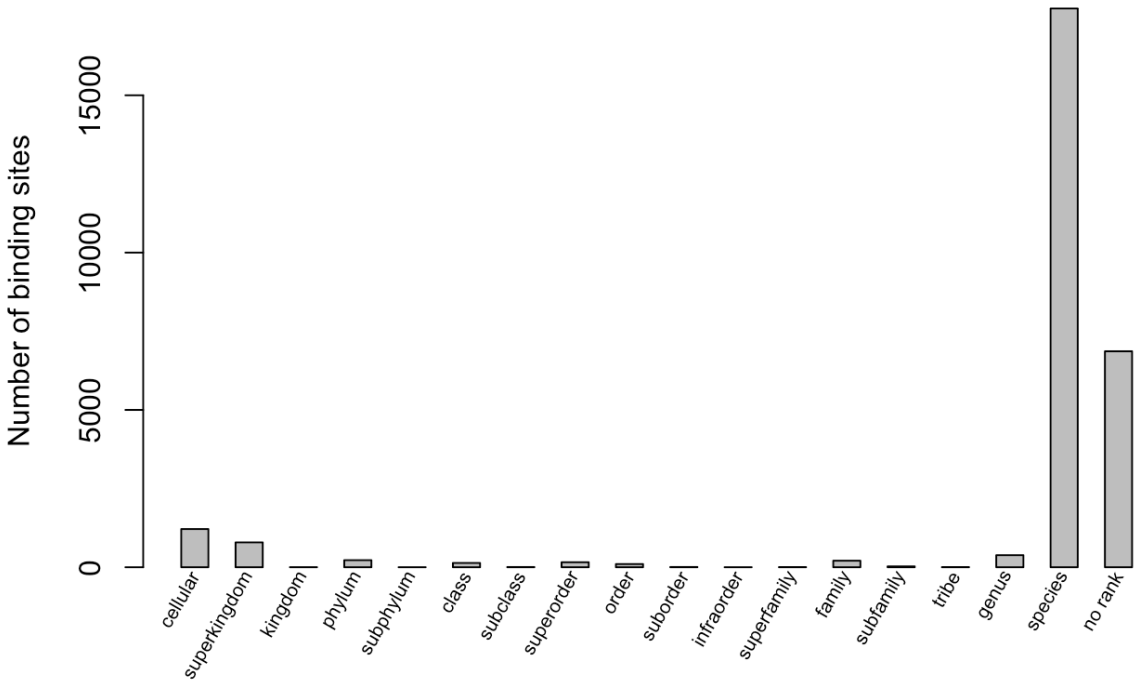

FIGURE S2 The number of unique binding sites distributed by their taxonomic rank, including the structures that are represented by one structure or by several redundant structures. These cases are obviously attributed to the species rank, thus there is no evidence about their evolutionary conservation.

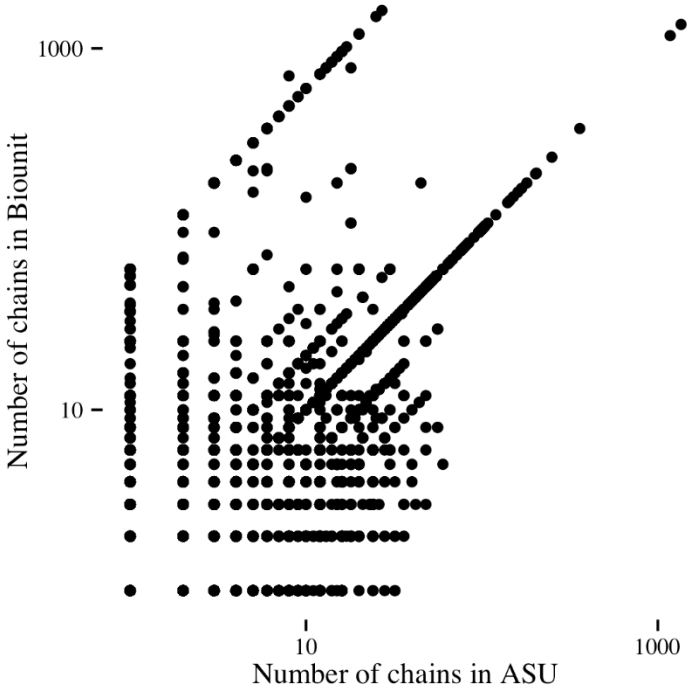

FIGURE S3 Number of chains in the asymmetric unit in the PDB X-ray crystal structures (ASU) versus the number of chains in the Biounit suggested by authors and/or software. Note the logarithmic scales on both axes. The main diagonal shows the structures where the asymmetric unit represents the biological assembly.

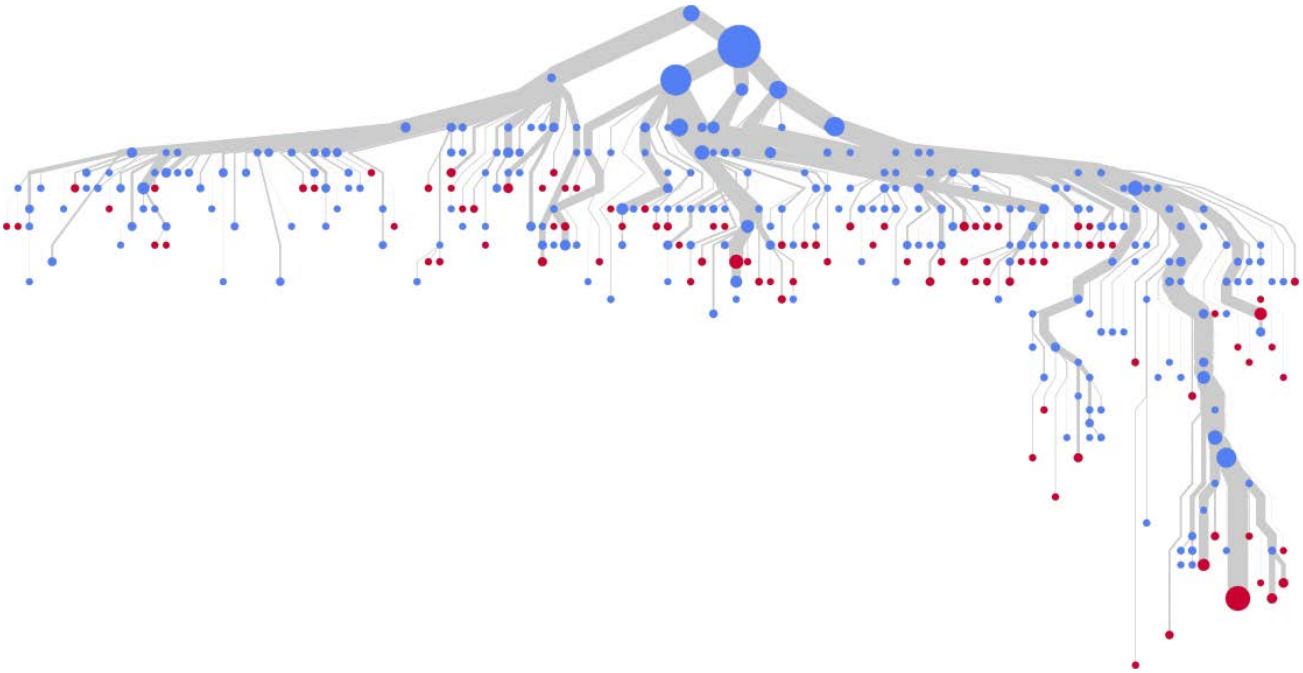

FIGURE S4 The cladogram constructed for all taxa which have at least one unique binding site cluster represented by two or more non-redundant structures. Labels omitted for clarity. The cladogram is accessible as a Cytoscape session in Supplementary File 1.

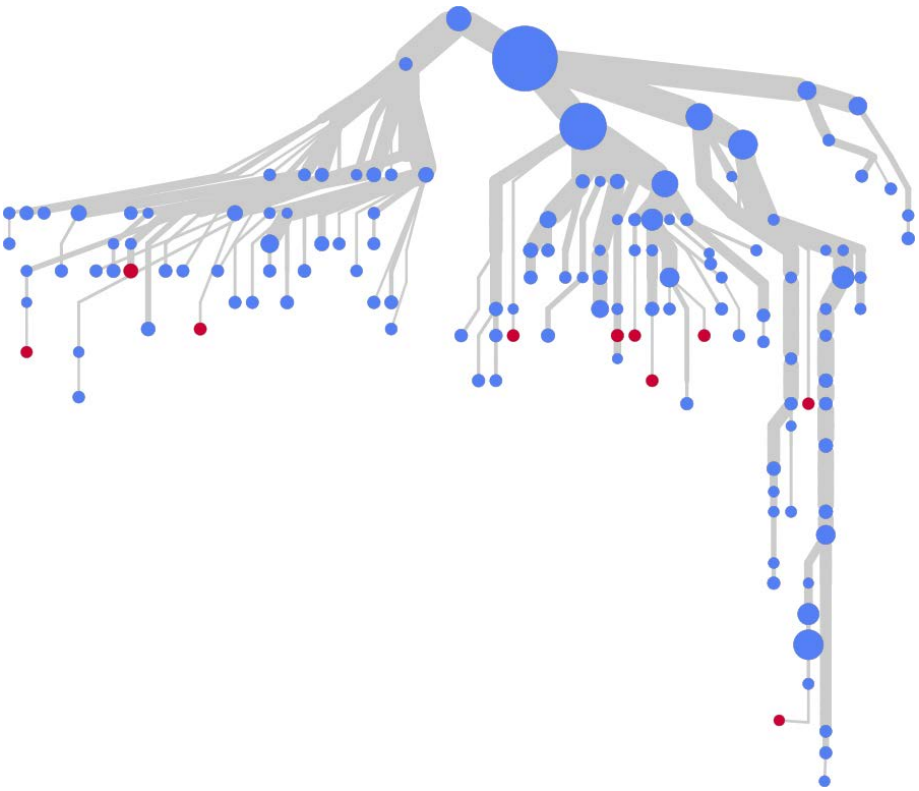

FIGURE S5 Cladogram constructed without the structural evidence from the most abundant and model organisms, including human, yeast, mouse, rat, cow, cress, and major bacterial, viral and archaeal species that dominate the PDB, including E.coli, HIV1 virus (listed in Table S2). Only the taxa represented by at least five unique protein binding sites are shown. Labels omitted for clarity. The cladogram is accessible as a Cytoscape session in Supplementary File 1.

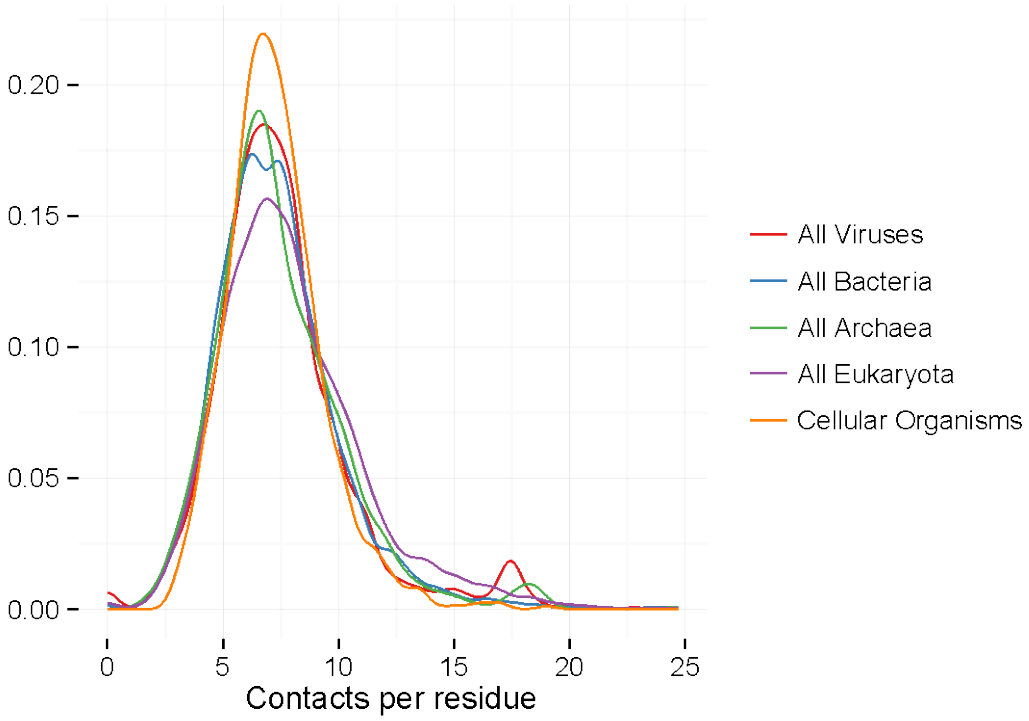

FIGURE S6 Distribution of the number of atomic protein-protein contacts (4 Å threshold) per binding site residue in different taxonomic branches.

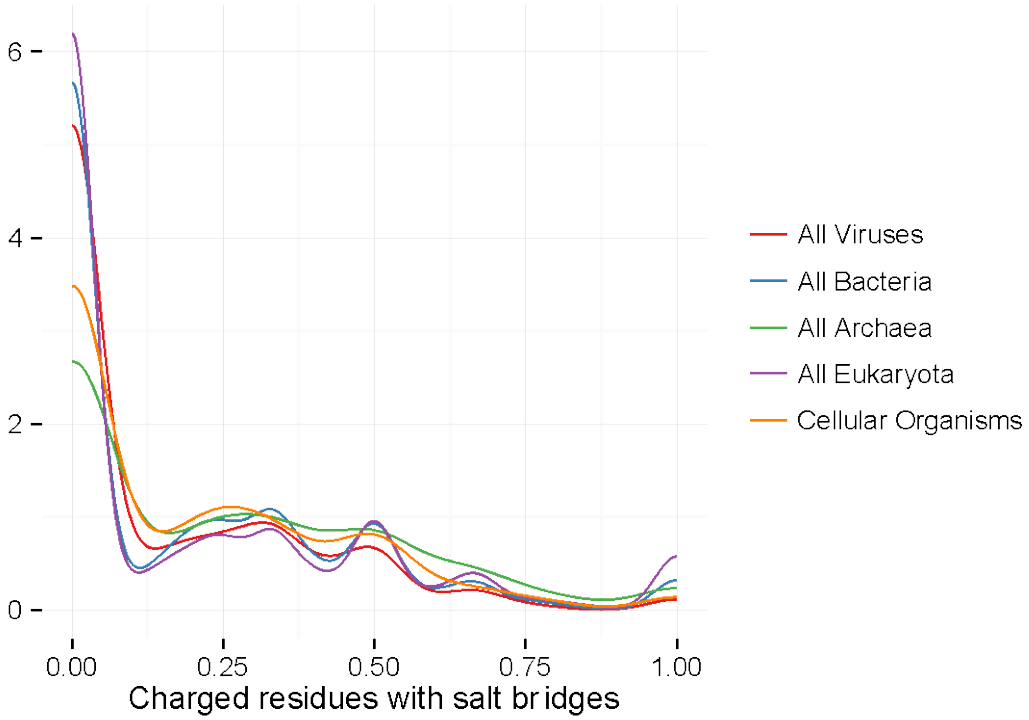

FIGURE S7 Distribution of proportion of charged residues forming salt bridges in different taxonomic branches.

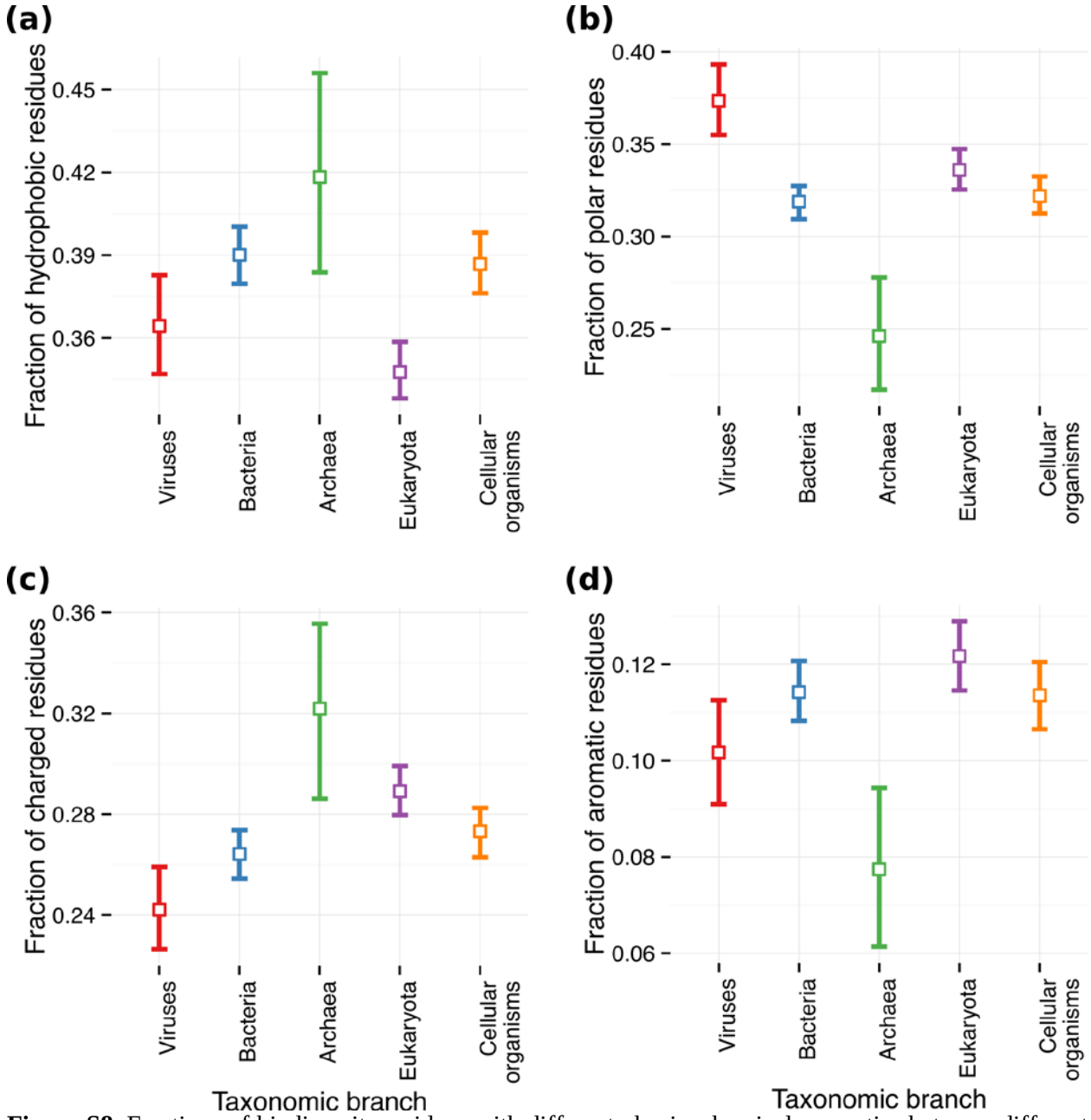

Figure S8. Fractions of binding site residues with different physicochemical properties between different taxa in human lineage. The taxa are ordered from left to right, from the most ancient to the most recent. The median value is shown as a dot on the violin plot. **(a)** Hydrophobic amino acids: AVLIPFVMW; **(b)** Polar: NQGSTY; **(c)** Charged: RKDE; **(d)**: Aromatic: FYWH.

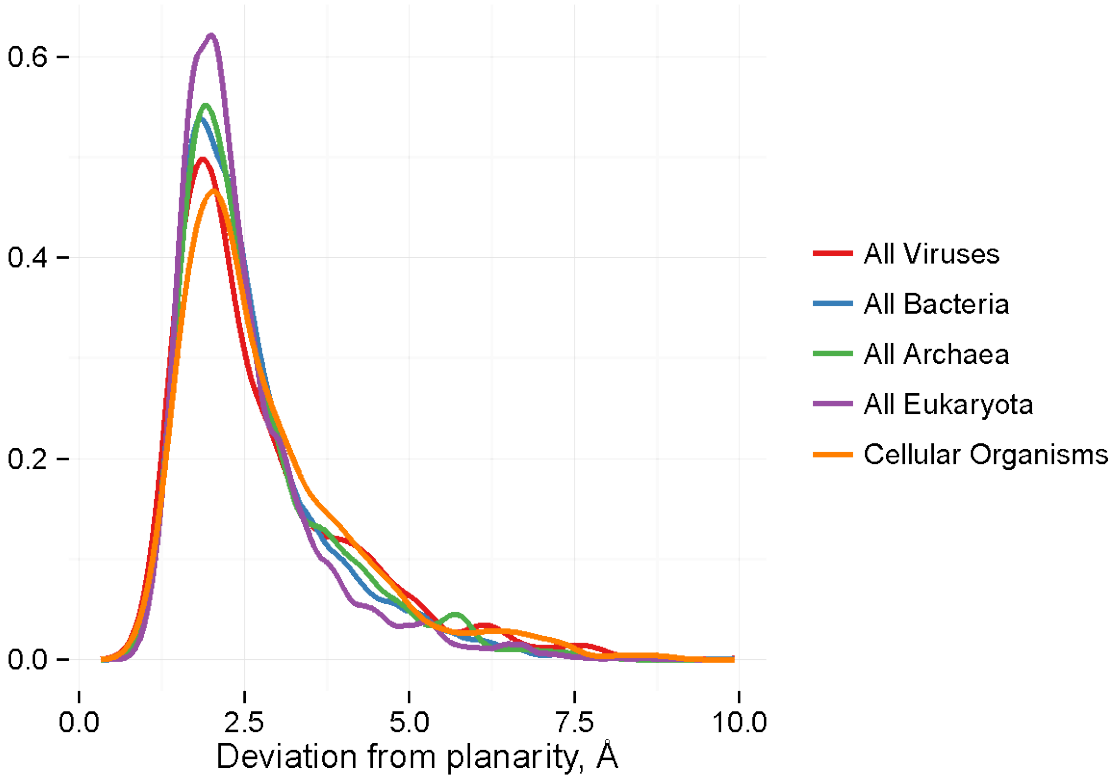

FIGURE S9. Planarity calculated for all heavy atoms on the binding sites

Tables

TABLE S1 Top 20 CDD superfamilies contributing the largest number of unique binding sites that are evolutionary conserved and represented by at least two non-redundant structures of protein complexes. The short CDD superfamily names can be looked up at the NCBI website: http://www.ncbi.nlm.nih.gov/Structure/cdd/

CDD Superfamily	Number of unique binding sites
ABC_ATPase	114
TIM_phosphate_binding	83
SDR	73
AAT_I	43
Ig	39
UBQ	32
Ferritin_like	31
Thioredoxin_like	30
rhv_like	28
metallo-dependent_hydrolases	27
NTF2_like	25
Photo_RC	22
NBD_sugar-kinase_HSP70_actin	22
Cytochrom_C	22
EFh	22
FN3	20
HTH_XRE	20
SIS	20
class_II_aaRS-like_core	19
Esterase_lipase	19

TABLE S2 The list of NCBI taxa excluded from the analysis in order to check the cladogram in Fig. 3 for robustness. These species have the largest number of structures in the PDB.

NCBI Taxon ID	NCBI Taxon name
9606	Homo sapiens
10090	Mus musculus
10116	Rattus norvegicus
9913	Bos taurus
4932	Saccharomyces cerevisiae
562	Escherichia coli
83333	Escherichia coli K-12
4932	Saccharomyces cerevisiae
559292	Saccharomyces cerevisiae S288c
1423	Bacillus subtilis
2261	Pyrococcus furiosus
28875	Rotavirus A
11676	Human immunodeficiency virus 1
10658	Enterobacteria phage PRD1
11320	Influenza A virus
12461	Hepatitis E virus
1773	Mycobacterium tuberculosis
1772	Mycobacterium smegmatis
1491	Clostridium botulinum
300852	Thermus thermophilus HB8
274	Thermus thermophilus
727	Haemophilus influenzae
9823	Sus scrofa
9031	Gallus gallus
7227	Drosophila melanogaster
7788	Torpedo marmorata
6239	Caenorhabditis elegans
3702	Arabidopsis thaliana

TABLE S3 Pairwise tests for binding site sizes and amino acid content. Wilcoxon rank sum test, Bonferroni-adjusted p-values are reported for each pairwise comparison. Only p-values < 0.1 are reported.

Binding site size:				
-	**Viruses**	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	0.013			
Archaea	0.016			
Eukaryota	< 2e-16	< 2e-16	0.140	
Cellular Organisms	-	5.1e-05	0.011	< 2e-16

Charged residues:				
-	**Viruses**	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	-			
Archaea	4.9e-05	0.00187		
Eukaryota	0.00031	0.00161	-	
Cellular Organisms	0.00676	-	0.03284	-

Polar residues:				
-	**Viruses**	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	1.7e-08			
Archaea	1.1e-12	1.4e-05		
Eukaryota	0.00023	-	9.7e-07	
Cellular Organisms	1.3e-07	-	2.0e-06	-

Hydrophobic residues:				
-	**Viruses**	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	-			
Archaea	0.00915	-		
Eukaryota	-	3.4e-08	0.00041	
Cellular Organisms	0.08757	-	-	4.9e-07

Aromatic residues:				
-	**Viruses**	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	-			
Archaea	0.01443	0.00070		
Eukaryota	-	-	0.00034	
Cellular Organisms	-	-	9e-05	-

TABLE S4 Pairwise tests for salt bridge propensity. Wilcoxon rank sum test, Bonferroni-adjusted p-values are reported for each pairwise comparison. Only p-values < 0.1 are reported.

	Viruses	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	4.0e-05			
Archaea	6.0e-16	2.8e-09		
Eukaryota	4.3e-05	-	2.1e-08	
Cellular organisms	3.5e-05	-	0.0037	-

TABLE S5 Pairwise tests for planarity. Wilcoxon rank sum test with Bonferroni correction. Only p-values < 0.1 are reported.

	Viruses	**Bacteria**	**Archaea**	**Eukaryota**
Bacteria	-			
Archaea	-	-		
Eukaryota	2.5e-05	< 2e-16	1.5e-06	
Cellular organisms	0.00027	4.1e-06	0.00680	< 2e-16