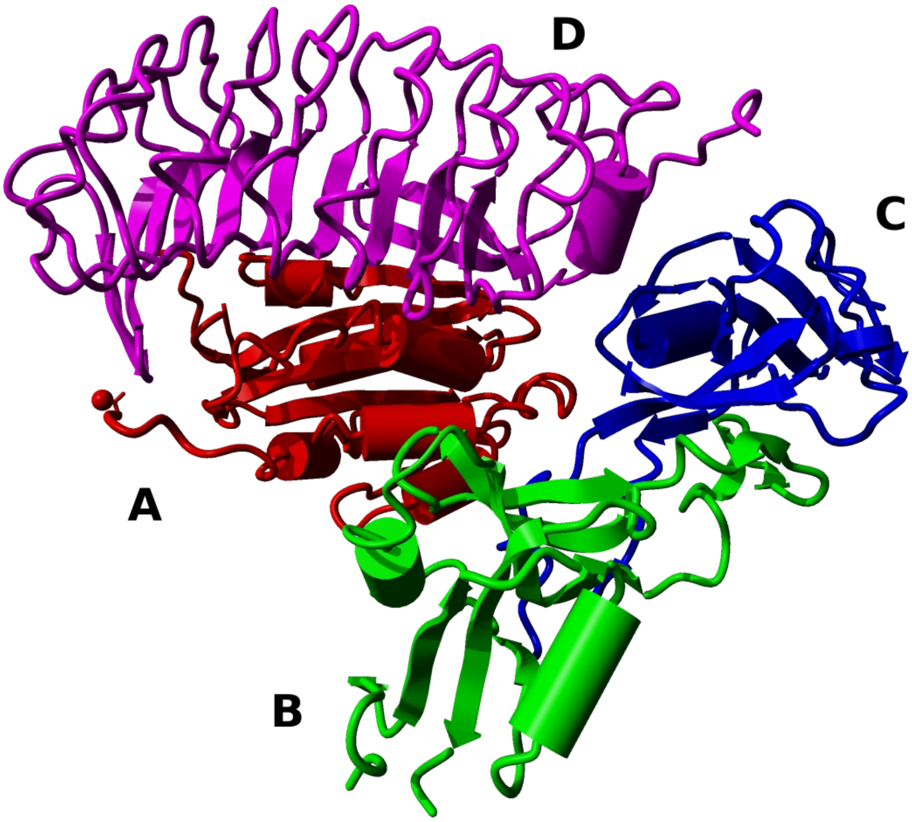
Supplementary Materials



Supplementary Figure 1. Ternary von Willebrand Factor A1-glycoprotein Ibalpha-botrocetin complex (PDB code: 1U0N) as an example of a non-interacting protein pair. Red – Von Willebrand factor (1U0N:A), green – botrocetin alpha chain (1U0N:B), blue – botrocetin beta chain (1U0N:C), magenta – platelet glycoprotein Ib (1U0N:D). Interacting pairs are: A-B, A-C, A-D, B-C, C-D. Chains B and D do not interact.

Comparing non-interacting pairs with predictions from STRING

The STRING database (1) aggregates vast amounts of data and predictions of protein-protein associations and interactions including the evidence based on genetic and functional context, experimental data, and text-mining results. Annotation is also transferred between different organisms. STRING predicts functional relations among proteins, some of which reflect physical binding. We mapped our non-interacting protein pairs against the STRING protein-protein interaction data using a 100% sequence identity threshold. The mapping between proteins and orthologous groups was used as provided by STRING. Consequently, we observed that only a small fraction (13.8%, 9.3%, 8.9%, and 8.3% for PDB, PDB-stringent, Manual, and Manual-stringent, respectively) of our non-interacting protein pairs were functionally associated by STRING. We speculate that the higher percentage of functional associations between non-interacting protein pairs derived from the PDB compared with those derived manually can be explained by the fact that non-interacting pairs from the PDB belong to the same complex and are thus more likely to be functionally coupled (e.g., co-regulated). We list the fraction of protein pairs from each of our non-interaction datasets which are predicted to be functionally related by individual evidence types aggregated by STRING (**Supplementary Table 1**). A similar listing is available for orthologous pairs (**Supplementary Table 2**). Most commonly non-interacting pairs are functionally associated through the "text-mining" evidence (**Supplementary Table 1**). The names of non-interacting protein pairs derived from manual annotation appear in the same text by definition. Consequently, we see a high percentage (85% and 86.9%) of the non-interacting protein pairs from the Manual and Manual-stringent datasets as functionally associated by text mining in STRING. Likewise, proteins co-occurring in the same structural complex have a very high chance to be described together in publications, and we also observe a high percentage of PDB-derived non-interacting pairs found in STRING functionally associated by text-mining (75.4% and 81.3% for the total and stringent PDB set, respectively). Thus, it is reasonable to assume that large numbers of protein pairs co-occurring in the same text are not directly interacting.

Overall, because the fraction of our non-interacting pairs found in STRING is small (8.3 - 13.8%), the fractions of non-interacting datasets that have been functionally associated by text mining remain small. For example only 7.6% of our non-interaction pairs from the Manual dataset, have been functionally associated by text mining (85% from 8.9% of the total number of pairs associated by STRING).

Two methods measuring shared evolutionary pressure ("neighborhood", the evolutionary conservation of gene order, and "co-occurrence", correlated occurrence or absence of proteins in complete genomes) of proteins in pairs support association of less than 27% and 7% of non-interacting protein pairs found in STRING derived from the PDB and manual annotation, respectively. The higher amount of functional association support for PDB-derived non-interacting pairs by these methods can be explained by tighter evolutionary constraint between protein pairs in PDB complexes compared with those in the manual data.

"Experimental" is another type of evidence which associates many non-interacting proteins (52-71% of the pairs found in STRING). It consists mostly of high throughput experimental results and may contain substantial amounts of false positive entries. The "Database" evidence type consists of protein associations from BIND (2), KEGG (3) and MIPS (4). There is relatively low support for functional association between manually derived non-interacting proteins by this type of evidence (10.3% and 13.3% of pairs found in STRING). In contrast, there is much higher support for association of PDB-derived non-interacting proteins (65% of PDB-stringent and 72% of PDB non-interacting pairs found in STRING), in the range comparable to "Experimental" data (56% and 71% for Manual-stringent and Manual, respectively). This observation is probably due to the fact that each PDB protein pair is found in a common protein complex. Thus, these pairs should more likely be members of at least one KEGG pathway compared with manually derived data and possibly have a membership in at least one BIND or MIPS protein complex. Interestingly, there is higher support for functional association of PDB non-interacting proteins by the "Co-expression" evidence compared with manual data (~60% vs. 4%) which contrasts to the generally poor levels of co-expression of complex members found in yeast (5).

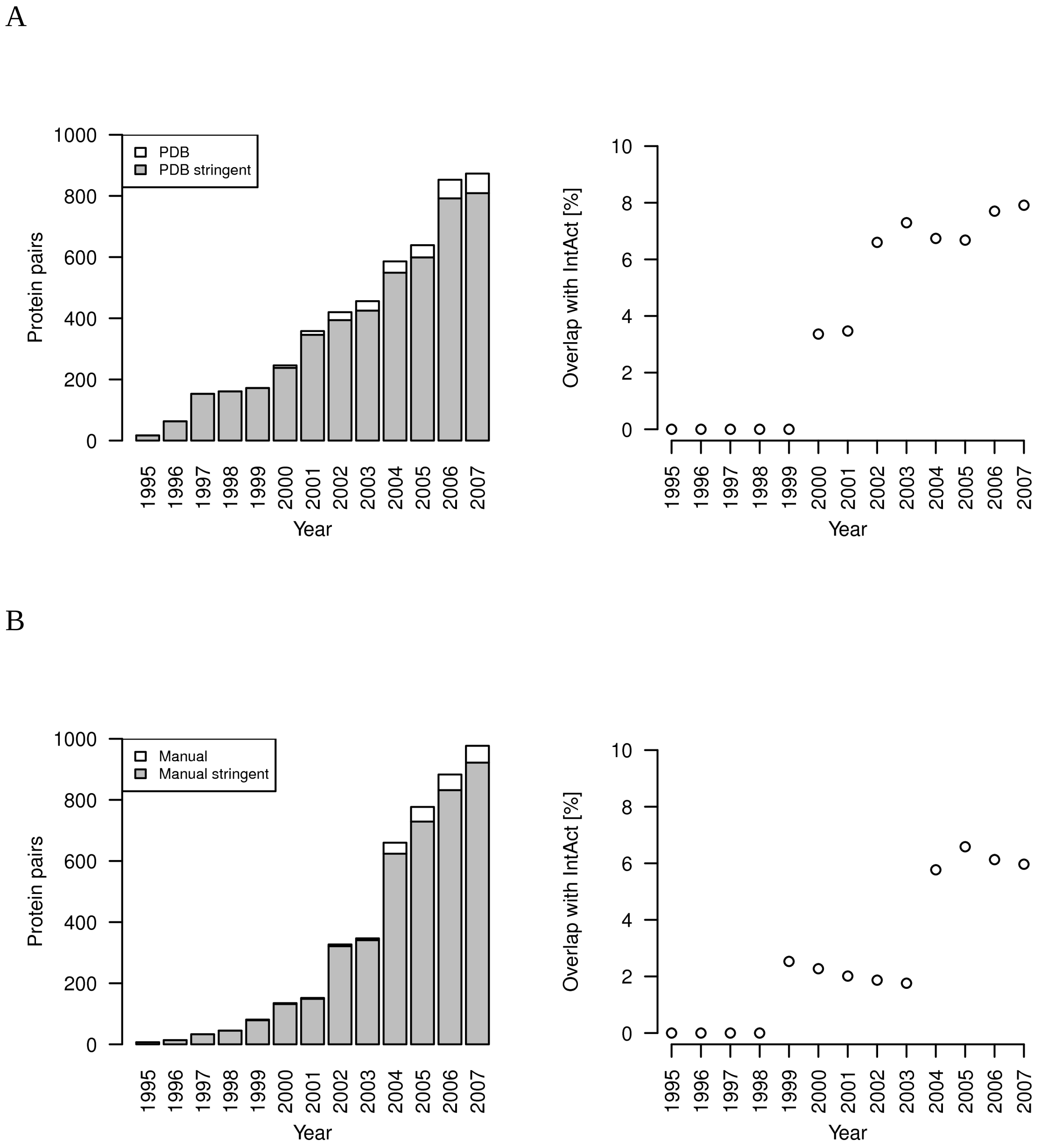
Summarizing, only a small fraction (8-14%) of our non-interacting pairs from both the PDB and the Manual datasets are functionally associated by STRING. Most of these associations are by "Text-mining" and "Experimental" evidences. The fact that two protein names co-occur in text does not necessarily mean that these proteins physically interact. The "Experimental" evidence may contain a significant number of false positive interactions due to its high-throughput nature. Pairs from the PDB dataset are additionally associated by "Database" evidence as they are frequently part of the same functional complex and commonly share the same metabolic pathway.

	Coverage of non-interacting datasets [%]			
Evidence Type	PDB	PDB-stringent	Manual	Manual-stringent
Neighborhood	9.8	14.5	0.0	0
Fusion	0.0	0.0	0.0	0
Co-occurence	26.8	20.3	6.7	6.5
Co-expression	62.5	59.4	4.2	3.7
Experimental	71.4	56.5	57.5	52.3
Database	72.3	65.2	13.3	10.3
Text mining	81.3	75.4	85.0	86.9
Overall coverage	13.8	9.3	8.9	8.3

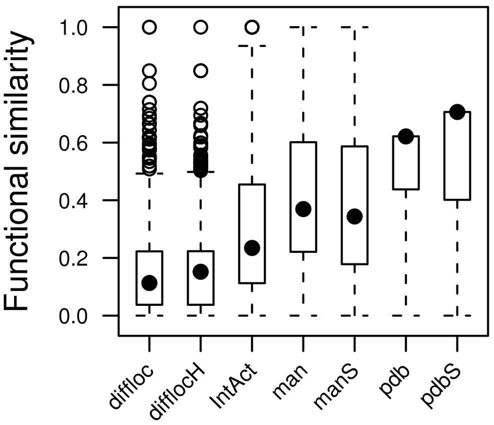
Supplementary Table 1. Coverage of non-interacting protein pairs by the STRING database. For each evidence type available in STRING we provide the percentage of non-interacting pairs from different Negatome subsets that were assigned as functionally linked by STRING.

	% coverage of orthologs of our non-interacting datasets			
Evidence Type	Pdb	Pdb-stringent	Manual	Manual-stringent
Neighborhood	13.1	15.5	5.4	5.5
Fusion	0.4	0.0	2.6	2.7
Co-occurence	21.6	26.8	10.2	9.5
Co-expression	54.6	52.6	35.3	31.4
Experimental	60.3	52.1	70.4	70.7
Database	68.1	64.8	24.1	21.1
Text-mining	75.5	71.4	85.5	85.6

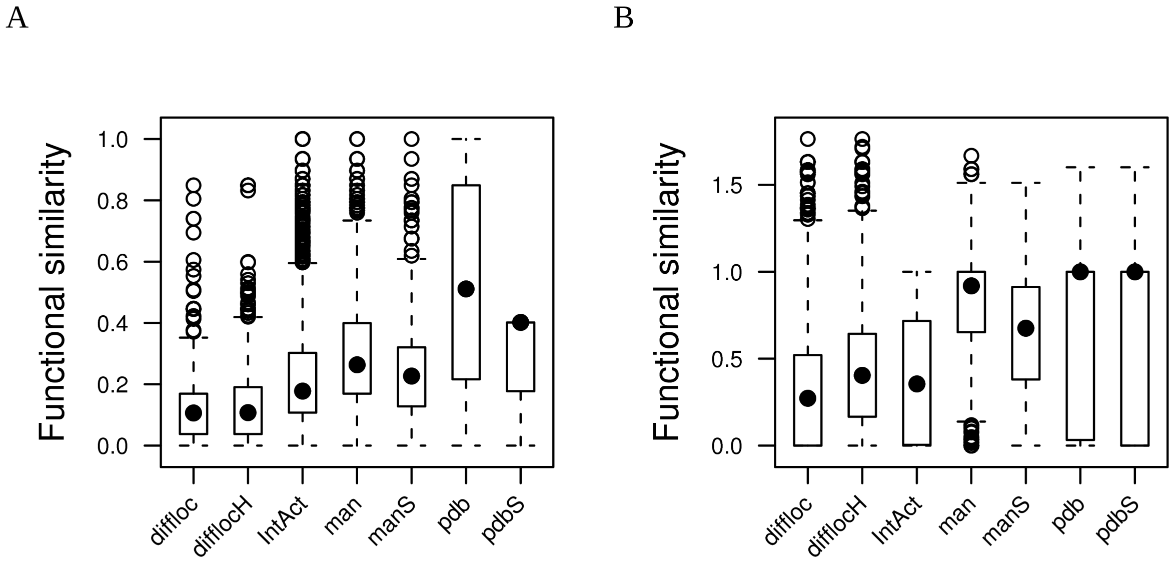
Supplementary Table 2. The percentages of non-interacting protein pairs which have an othologous pair in STRING supported by different evidence types. Example: 75.5% of pairs in the PDB dataset with orthologs in STRING are supported by text-mining evidence.



Supplementary Figure 2. Time trends in the number of non-interacting protein pairs in PDB (A) and the Manual (B) dataset. Left panels: Absolute numbers of non-interacting pairs found in the PDB (top) and Manual (bottom) datasets are depicted by white bars; grey bars show the number of non-interacting pairs not found in the IntAct database. Right panels: relative percentage of non-interacting pairs contradicted by IntAct.



Supplementary Figure 3. Functional similarity of interacting and non-interacting protein pairs. Similarity scores ranging from 0 (no similarity) to 1 (identical function) computed from the biological process subset of the Gene Ontology for interacting proteins obtained from IntAct (IntAct), random protein pairs from different cellular locations for arbitrary species (diffloc) and human-only(difflocH), our manually and PDB derived non-interacting pairs (man, pdb), and our manual and PDB-derived data filtered by IntAct (manS, pdbS) using the Resnik method(6)



Supplementary Figure 4. . Functional similarity of interacting and non-interacting protein pairs. Similarity scores ranging from 0 (no similarity) to 1 (identical function) computed from the biological process subset of the Gene Ontology for interacting proteins obtained from IntAct (IntAct), random protein pairs from different cellular locations for arbitrary species (diffloc) and human-only(difflocH), our manually and PDB derived non-interacting pairs (man, pdb), and our manual and PDB-derived data filtered by IntAct (manS, pdbS) using the GraSM/Resnik(7) (A) and Jiang-Conrath(8) (B) methods.

Analysis of domain and protein coevolution

The bit difference between profiles of domains A and B were computed by the Hamming distance $\sum_{i=1}^{n} |A_i - B_i|$ where $A_i = 1$ if domain A is present in the organism i ∊ n, where n is the number of organisms. If domain A is not detected (9,10) in the organism i, then $A_i = 0$. B_i is similarly assigned for the organism i. The level of conservation for an individual domain A is computed by the formula $\sum_{i=1}^{n} A_i$. The conservation score for a pair of domains A and B is computed by the formula $\sum_{i=1}^{n} \left(A_i \wedge B_i \right)$, where $A_i \wedge B_i = 1$ if $A_i =1$ and $B_i =1$, 0 otherwise.

At the protein level, we considered the sequence identity between a protein and its aligned homolog instead of simply their presence or absence. Differences and conservation of sequence identity profiles for proteins were computed with methods analogous to those used for domains. The percent identity between aligned protein homologs was retrieved from SIMAP (11).

We measure the 'profile identity difference' of two proteins (A_0, B_0) by $\sum_{i=1}^{n} |A_i - B_i|$, where n is the number of organisms and A_i and B_i are the % identities of the most similar homolog in organism i ∊ n when aligned to A_0 and B_0, respectively. If a homolog for a species does not exist, the % identity for that organism is set to 0. The 'conservation sum' is computed by the formula $\sum_{i=1}^{n} A_i + B_i$.

Calculations were based on 10 organisms: List of 10 organisms used for calculations in analysis of domain and protein coevolution: *Pyrococcus abyssi* GE5, *Thermoanaerobacter tengcongensis* MB4, *Trypanosoma brucei* TREU927, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Danio rerio*, *Homo sapiens*, *Escherichia coli K12*, *Bacillus subtilis subsp. subtilis* str. 168, and *Synechocystis* sp. PCC 6803.

Coevolution of interacting and non-interacting protein and domain pairs

We analyzed domains detected in our non-interacting protein pairs derived by manual annotation and from PDB using phylogenetic profiling. The DIMA2 resource (9,10) provides phylogenetic profile strings of 1's and 0's representing the presence or absence, respectively, of particular domains across 460 genomes. Differences between a pair of domain phylogenetic profiles can be measured in terms of bit differences between the profile strings (see *above formulas*). We find that the bit difference scores between the phylogenetic profiles corresponding to non-interacting domain pairs derived from PDB-stringent and Manual-stringent were significantly higher compared with interacting domains from 3DID and iPFAM (mean bit differences between pairs: 57 for 3DID/iPFAM, 102 for PDB-stringent and 130 for Manual-stringent domain pairs; Mann-Whitney-U-test (MW-test) p-value $< 3.7 \times 10^{-27}$) (**Supplementary Table 3**). These differences remained significant even when we compared our non-interacting domain pairs against a subset of our 3DID/iPFAM domain pairs composed of domains found only in our non-interaction pairs. We also found that interacting domains co-occur significantly more often than non-interacting domains even when we profiled the domains using a variety of subsets of the 460 genomes. These observations indicate that our assignments of the non-interaction status to domain pairs are corroborated by the tendency to lower levels of co-occurrence over large evolutionary time scales.

Conservation of non-interacting domains in 460 genomes

Many known domain-domain interactions tend to be conserved (12). Considering domains individually, we find that interacting domains from iPFAM and 3DID tend to be more conserved than those in our non-interacting protein datasets (**Supplementary Table 4**) across the 460 genomes from DIMA2. 3DID/iPFAM interacting domains were found to be more conserved in pairs than those in our non-interacting datasets (**Supplementary Table 5**). Individual partners of non-interacting pairs are more readily lost over time than those found in interacting pairs considering these genomes.

Sequence divergence of non-interacting proteins across 10 genomes

At the protein level, one can generate sequence identity profiles for a given protein by finding the closest homologs to that protein in a number of organisms and creating a vector of sequence identities computed after aligning the original protein to each of those homologs. Analogous to the bit difference between profiles for two domains, the 'profile identity difference' between two proteins is computed by summing the sequence identity differences between each pair of homologs of the original pair of proteins (see *p. 6 formulas*). The profile identity difference is a simple measure of co-evolution between two proteins based on the level of sequence divergence between their homologs across a number of organisms.

We find that protein pairs involved in binary protein-protein interactions according to IntAct have significantly smaller profile identity differences than our sets of non-interacting protein pairs derived manually and from the PDB (mean identity difference: 136.45 for IntAct, 155.84 for Manual-stringent and 159.85 for PDB-stringent dataset; MW-test: p-value $< 6.6 \times 10^{-13}$) (Supplementary Table 6). These differences remained significant even when we compared our non-interacting protein pairs against IntAct binary interactions involving proteins found in our non-interacting datasets. Unlike domain interactions, interacting proteins in IntAct were not found to be more conserved (in terms of the 'conservation sum': see Methods) than protein pairs in our manual non-interaction data (Mean conservation sum: 517.73 for IntAct, 559.28 for the Manual and 482.83 for the PDB dataset) (Supplementary Table 7). We conclude that interacting protein pairs showed higher uniformity in sequence divergence across the 10 examined species compared to our datasets of non-interacting proteins.

Comparison between known domain-domain interaction pairs and our non-interacting domain pair data according to phylogenetic profiles for domains and domain conservation.

The datasets are denoted as follows:
DDI = domain-domain interaction pairs in 3DID and IPFAM
MNDDI = manual stringent non-interacting domain-domain pair set
PDBNDDI = stringent PDB-derived non-interacting domain set
DDIM = DDI subset in which both domains in each pair are found in MNDDI
DDIP = DDI subset in which both domains in each pair are found in PDBNDDI
Domain phylogenetic profiles using 460 genomes were retrieved from DIMA2.

For the first two columns, the numbers of domain pairs are displayed in parentheses.

Average values are displayed as Mean +/- Standard Deviation (Median).

Domain Set1	Domain Set2	Average Bit Difference for Domain Set 1	Average Bit Difference for Domain Set 2	Estimated P-value (Mann-Whitney U-test)
DDI (3113)	MNDDI (1142)	57.03 +/- 110.09 (0.00)	151.16 +/- 154.81 (91.00)	$< 3.7 \times 10^{-56}$
DDIM (276)	MNDDI (1142)	47.11 +/- 99.16 (0.00)	151.16 +/- 154.81 (91.00)	$< 5.1 \times 10^{-35}$
DDI (3113)	PDBNDDI (315)	57.03 +/- 110.09 (0.00)	101.81 +/- 144.52 (17.00)	$< 3.7 \times 10^{-27}$
DDIP (270)	PDBNDDI (315)	85.39 +/- 122.15 (8.00)	101.81 +/- 144.52 (17.00)	$< 3.5 \times 10^{-20}$
MNDDI (1142)	PDBNDDI (315)	151.16 +/- 154.81 (91.00)	101.81 +/- 144.52 (17.00)	$<7.8 \times 10^{-10}$

Supplementary Table 3. Bit differences between domain pairs. Interacting domain pairs have significantly smaller profile bit differences than non-interacting domain pairs we selected. The manual non-interaction domain pairs have the most bit differences between domains.

Domain Set1	Domain Set2	Average Conservation Score for Domain Set 1	Average Conservation Score for Domain Set 2	Estimated P-value (Mann-Whitney U-test)
DDI (3113)	MNDDI (1142)	400.77 +/- 319.28 (368.00)	242.92 +/- 233.69 (150.50)	$<8.1 \times 10^{-29}$
DDIM (276)	MNDDI (1142)	306.51 +/- 326.52 (130.50)	242.92 +/- 233.69 (150.50)	< 0.40 (not significant)
DDI (3113)	PDBNDDI (315)	400.77 +/- 319.28 (368.00)	193.79 +/- 243.96 (41.00)	$< 2.7 \times 10^{-24}$
DDIP (270)	PDBNDDI (315)	322.39 +/- 303.47 (256.00)	193.79 +/- 243.96 (41.00)	$< 2.5 \times 10^{-05}$
MNDDI (1142)	PDBNDDI (315)	242.92 +/- 233.69 (150.50)	193.79 +/- 243.96 (41.00)	$<4.3 \times 10^{-08}$

Supplementary Table 4. Conservation of individual domains. The Conservation Score for a domain is measured by the sum of the 1's in the profile. Considering domains individually, those found in our non-interacting sets (MNDDI and PDBNDDI) are significantly less conserved than those in the DDI set. No significant difference was found between MNDDI and DDIM but a significant difference was found between PDBNDDI and DDIP.

Domain Set1	Domain Set2	Average Pairwise Conservation Score for Domain Set 1	Average Pairwise Conservation Score for Domain Set 2	Estimated P-value (Mann-Whitney U-tes)
DDI (3113)	MNDDI (1142)	171.87 +/- 165.38 (106.00)	45.88 +/- 91.51 (17.00)	$< 2.4 \times 10^{-50}$
DDIM (276)	MNDDI (1142)	129.70 +/- 164.73 (25.00)	45.88 +/- 91.51 (17.00)	$< 9.5 \times 10^{-16}$
DDI (3113)	PDBNDDI (315)	171.87 +/- 165.38 (106.00)	45.99 +/- 103.03 (14.00)	$<5.5 \times 10^{-36}$
DDIP (270)	PDBNDDI (315)	118.50 +/- 159.00 (20.00)	45.99 +/- 103.03 (14.00)	$< 4.1 \times 10^{-08}$
MNDDI (1142)	PDBNDDI (315)	45.88 +/- 91.51 (17.00)	45.99 +/- 103.03 (14.00)	Not significant

Supplementary Table 5. Conservation of domains pairs. The conservation score for a domain-pair is measured by the number of organisms which have both domains in the pair. Considering such domains pairs, those found in our non-interacting sets (MNDDI and PDBNDDI) are significantly less conserved than those in the DDI set. No significant difference was found between MNDDI and PDBNDDI.

Comparison between known domain-domain interaction pairs and our non-interacting domain pair data using subsets of the 460 genomes

Because results from Supplementary Tables 3, 4, and 5 could be specific to our use of the 460 genomes from DIMA2, we repeated the comparisons using subsets of these genomes. We did the following:

1) We repeated the comparisons 100 times, each time selecting randomly 25 / 460 genomes.
2) We repeated the comparisons using only Archaea
3) We repeated the comparisons using only Eukaryota
4) We repeated the comparisons using only Bacteria
5) We repeated the comparisons using a divergent set of 10 genomes:
(Pyrococcus abyssi GE5, Thermoanaerobacter tengcongensis MB4, Arabidopsis thaliana, Saccharomyces cerevisiae, Bacillus subtilis subsp. subtilis str. 168, Synechocystis sp. PCC 6803, encephalitozoon cuniculi, Streptococcus_pneumoniae_TIGR4, Xylella_fastidiosa_9a5c, Archaeoglobus_fulgidus_DSM_4304).

In all comparisons, we find that interacting domains (from 3DID, iPFAM) co-occur significantly more often than our non-interacting manual or PDB-derived domains (Mann-Whitney U-test, $P < 0.01$). These differences remained significant even when we compared our non-interacting domain pairs against a subset of our 3DID/iPFAM domain pairs composed of domains found only in our non-interaction pairs.

While our manual non-interaction data consists of mammalian proteins, our PDB non-interaction data consists mostly of proteins from prokaryotes. The interacting domain pairs from 3DID and iPFAM may also exhibit certain phylogenetic biases.

6) We repeated all comparisons 100 times, each time selecting randomly 25 genomes out of the 460 genomes. For this step, we also randomly selected subsets of our mammalian, PDB and interacting domain pairs randomly, using only 10-15% of the original data for each comparison.

In all comparisons, we find that interacting domains co-occur significantly more often than noninteracting domains (MW U-test: $P < 0.05$). Note that with such reduced numbers of domains for comparison, when we compared our non-interacting domain pairs against a subset of our 3DID/iPFAM domain pairs composed of domains found only in our non-interaction pairs, significant differences could not be established in many cases.

Nevertheless, these additional results suggest that non-interaction status between domains is associated with their level of co-occurrence across a diverse set of genomes.

For steps, 1, 5, and 6, we also found that non-interacting domains were less conserved individually and as pairs across examined genomes, compared with interacting domains (MW U-test: $P < 0.05$). However, significant differences could not be established if we restricted the profiling to Archaea, or Eukaryota.

Comparison between known protein-protein interaction pairs and our non-interacting protein pair data according to protein sequence conservation

The datasets are denoted as follows:
PPI = protein-protein interactions from IntAct
MNPPI = manual non-interacting protein pair set
PDBPPI = more stringent PDB-derived non-interacting protein pair set
PPIM = Interaction protein pairs (from PPI) in which both proteins in each pair are found in MNPPI
PPIP = Interaction protein pairs (from PPI) in which both proteins in each pair are found in PDBNPPI
For the first two columns, the number of protein pairs is displayed in parentheses.

Best hit homologs of a given protein were searched for in 10 different species (*Pyrococcus abyssi* GE5, *Thermoanaerobacter tengcongensis* MB4, *Trypanosoma brucei* TREU927, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Danio rerio*, *Homo sapiens*, *Escherichia coli K12*, *Bacillus subtilis subsp. subtilis str. 168* and *Synechocystis sp*. PCC 6803) using SIMAP (E-value threshold: 0.0001). The % identity for each best hit is recorded. For each protein pair (A, B) we compute the difference between % identities in each of the 10 species. If a homolog for a species does not exist, the % identity for that organism is set to 0. The "Profile Identity Difference" for a protein pair is measured as the sum of the identity differences in the 10 species. For example if protein A had homologs with identities (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%) and protein B had homologs with identities (11%, 21%, 31%, 41%, 51%, 61%, 71%, 81%, 91%, 99%) in the 10 species, then the conservation difference for the pair (A,B) would be 10%.

Table S6 shows the average profile identity differences for protein pairs in the considered datasets (column 1 and 2) as Mean +/- Standard Deviation (Median) (in columns 3 and 4).

Protein Set1	Protein Set2	Average Profile Identity Difference for Protein Set 1 (%)	Average Profile Identity Difference for Protein Set 2 (%)	Estimated P-value (Mann-Whitney U-tes)
PPI (127514)	MNPPI (1162)	136.45 +/- 89.83 (117.50)	155.84 +/- 92.45 (147.75)	$< 6.6 \times 10^{-13}$
PPIM (385)	MNPPI (1162)	141.87 +/- 105.90 (132.80)	155.84 +/- 92.45 (147.75)	$<8.70 \times 10^{-5}$
PPI (127514)	PDBNPPI (759)	136.45 +/- 89.83 (117.50)	159.85 +/- 83.96 (151.90)	$<6.9 \times 10^{-16}$
PPIP (178)	PDBNPPI (759)	120.10 +/- 103.34 (86.05)	159.85 +/- 83.96 (151.90)	$<2.2 \times 10^{-12}$
MNPPI (1162)	PDBNPPI (759)	155.84 +/- 92.45 (147.75)	159.85 +/- 83.96 (151.90)	<0.18 (not significant)

Supplementary Table 6. Sequence identity differences of protein pairs. Interacting protein pairs have significantly fewer sequence differences than non-interacting protein pairs. Significant differences between non-interacting PPI derived manually and from the PDB could not be established.

The "conservation" of the proteins for each pair can also be measured by summing the identities of the homologs. For example if protein A had homologs with identities (10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%) and protein B had homologs with identities (11%, 21%, 31%, 41%, 51%, 61%, 71%, 81%, 91%, 99%) in the 10 species, then the conservation score for the pair (A,B) would be 10% + 20% + 30% +40% +50% +60%+ 70%+ 80%+ 90%+100% +11%+ 21%+ 31% +41%+ 51%+ 61%+ 71%+ 81%+ 91%+ 99% = 1108%.

Table S7 shows the average conservation sum for protein pairs in the considered datasets (column 1 and 2) as Mean +/- Standard Deviation (Median) (in columns 3 and 4).

Protein Set1	Protein Set2	Average Conservation Sum for Protein Set 1 (%)	Average Conservation Sum for Protein Set 2 (%)	Estimated P-value (Mann-Whitney U-tes)
PPI (127514)	MNPPI (1162)	517.73 +/- 144.34 (510.10)	559.28 +/- 154.14 (576.80)	<2.9 x10^{-20}
PPIM (385)	MNPPI (1162)	573.78 +/- 169.33 (557.90)	559.28 +/- 154.14 (576.80	<0.9 (not significant)
PPI (127514)	PDBNPPI (759)	517.73 +/- 144.34 (510.10)	482.83 +/- 188.37 (472.20)	<0.003
PPIP (178)	PDBNPPI (759)	590.86 +/- 150.23 (616.00)	482.83 +/- 188.37 (472.20)	<5.3 x 10^{-7}
MNPPI (1162)	PDBNPPI (759)	559.28 +/- 154.14 (576.80	482.83 +/- 188.37 (472.20)	<6.9 x 10^{-14}

Supplementary Table 7. Sequence conservation of individual proteins. Proteins in the manual non-interaction data tend to be more conserved than those in IntAct and non-interacting protein pairs from the PDB. Interacting proteins found in PPI were more conserved than those in PDBNPPI.

REFERENCES:

1. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. and Bork, P. (2007) STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, **35**, D358-362.
2. Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*, **31**, 248-250.
3. Kanehisa, M. (2002) The KEGG database. *Novartis Found Symp*, **247**, 91-101; discussion 101-103, 119-128, 244-152.
4. Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W. and Stumpflen, V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, **34**, D436-441.
5. Liu, C.T., Yuan, S. and Li, K.C. (2009) Patterns of co-expression for protein complexes by size in Saccharomyces cerevisiae. *Nucleic Acids Res*, **37**, 526-532.
6. Resnik, P. (1995), *14th International Conference Research on Computational Linguistics*. IJCAI-95, Montreal, Canada, Vol. 1, pp. 448-453.
7. Couto, F.M., Silva, M.J. and Coutinho, P.M. (2005) Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, **6 Suppl 1**, S21.
8. Jiang, J. and Conrath, D. (1998), *International Conference Research on Computational Linguistics*. ROCLING X, Taiwan, Vol. 1.
9. Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stumpflen, V. and Frishman, D.

(2008) DIMA 2.0--predicted and known domain interactions. *Nucleic Acids Res*, **36**, D651-655.

10. Pagel, P., Wong, P. and Frishman, D. (2004) A domain interaction map based on phylogenetic profiling. *J Mol Biol*, **344**, 1331-1346.

11. Rattei, T., Tischler, P., Arnold, R., Hamberger, F., Krebs, J., Krumsiek, J., Wachinger, B., Stumpflen, V. and Mewes, W. (2008) SIMAP--structuring the network of protein similarities. *Nucleic Acids Res*, **36**, D289-292.

12. Itzhaki, Z., Akiva, E., Altuvia, Y. and Margalit, H. (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol*, **7**, R125.