Chapter 2

Distribution and evolution of interacting domains

2.1 Introduction

I have mentioned in the introduction the importance of evolutionary relationships for the understanding of protein function. Families of related sequence regions, collected in the Pfam database (Finn *et al.*, 2008), usually constitute structurally and functionally conserved modules. Categorising proteins according to their sequence similarity vastly reduces the size and complexity of protein space. It is assumed that binding interfaces, too, are conserved evolutionary modules that are reused between proteins of different functions and retained during evolution (Aloy and Russell, 2004; Itzhaki *et al.*, 2006). Accordingly, it would be desirable to understand the relationships between interacting proteins from a point of view of their sequence genealogy.

In recognising this, several groups have attempted to derive a set of domain–domain pairs that are likely to comprise evolutionarily conserved modules for protein interaction. Ng *et al.* (2003) described an approach to predict domain–domain interactions using literature curation, evolutionary history and the distribution of domains in protein

interactions. More recently, other groups have come up with sophisticated statistical methods to estimate putatively interacting domain pairs, based on the assumption of domain reusability (Jothi *et al.*, 2006; Lee *et al.*, 2006; Nye *et al.*, 2005; Pagel *et al.*, 2004; Riley *et al.*, 2005). However, none of these approaches offers structural evidence that the predicted domain pairs are able to form an interaction. As described in the introduction, the *i*Pfam database (Finn *et al.*, 2005) provides this missing link between sequence family membership in the form of Pfam domain annotations and protein interactions, as derived from crystal structures of molecular complexes (Littler and Hubbard, 2005; Park *et al.*, 2001) deposited in the PDB (Kouranov *et al.*, 2006).

Theoretically, the *i*Pfam database should thus provide a structural explanation for most protein interactions. Unfortunately, the selection of complexes in the PDB is rather small[1] and biased (Peng *et al.*, 2004). There is often only a single structure that shows a certain protein pair to interact, while other complexes like the haemoglobin tetramer have been crystalized dozens of times. This makes it difficult to assess whether some domain pairs act as reusable modules in protein interactions from PDB data alone.

One of the aims of the work presented in this chapter was therefore to understand the possibilities and limitations of *i*Pfam when applied to protein interaction networks. To achieve this, I investigated how pairs of protein families taken from *i*Pfam are distributed in protein interaction networks of five major model species. I specifically addressed the question what proportion of each organism's protein interaction network, its *interactome*, can be attributed to a known domain–domain interaction, and conversely, how many interacting domain pairs are still unknown. These insights, together with the tools and data-sources compiled for this analysis, lay the foundation for the following chapters.

The other aim of this chapter is to shed some light on the conservation of domain–

[1] Out of a total of 31522 PDB entries, comprising 11338 distinct sequences, 12790 entries contain a protein complex, corresponding to only 5938 proteins. In comparison, there were $3.17 \cdot 10^6$ sequences in UniProt at the time of analysis.

domain interactions between species. Despite the continuing growth of protein interaction databases, even the best studied protein interaction network of *S. cerevisiae* is thought to be incomplete (Cusick *et al.*, 2005; Grigoriev, 2003; von Mering *et al.*, 2002). Given that this network already comprises around 60000 interactions, questions arise as to how such networks have evolved and how they are organised. By comparing the sets of interacting domain pairs found in the investigated model organisms, I can make inferences about the evolution of protein interactions.

2.2 Methods

2.2.1 Protein interaction data

The complete interaction sets from BioGRID (Breitkreutz *et al.*, 2008), DIP (Salwinski *et al.*, 2004), HPRD (Mishra *et al.*, 2006), IntAct (Kerrien *et al.*, 2007), MINT (Chatraryamontri *et al.*, 2007) and MPact (Guldener *et al.*, 2006) were downloaded on the 24th January 2008. A wide range of databases were used to cover as many distinct experimental data sets as possible. Taken together, these databases represent most of the protein interactions currently stored in machine-accessible form.

Despite great efforts to unify access to protein interaction data (Hermjakob *et al.*, 2004), acquiring large data sets from diverse sources is still far from trivial and error prone. The PSI-MI XML data exchange format version 2.5 (Hermjakob *et al.*, 2004) provided by the aforementioned databases was used to generate a local relational database of protein interactions. For each protein participant, it was attempted to assign a sequence, either from data provided by the source database or by mapping the entry to UniProt *via* secondary annotations provided in the source file. A schematic flow-chart of the database creation process is shown in Figure 2.1.

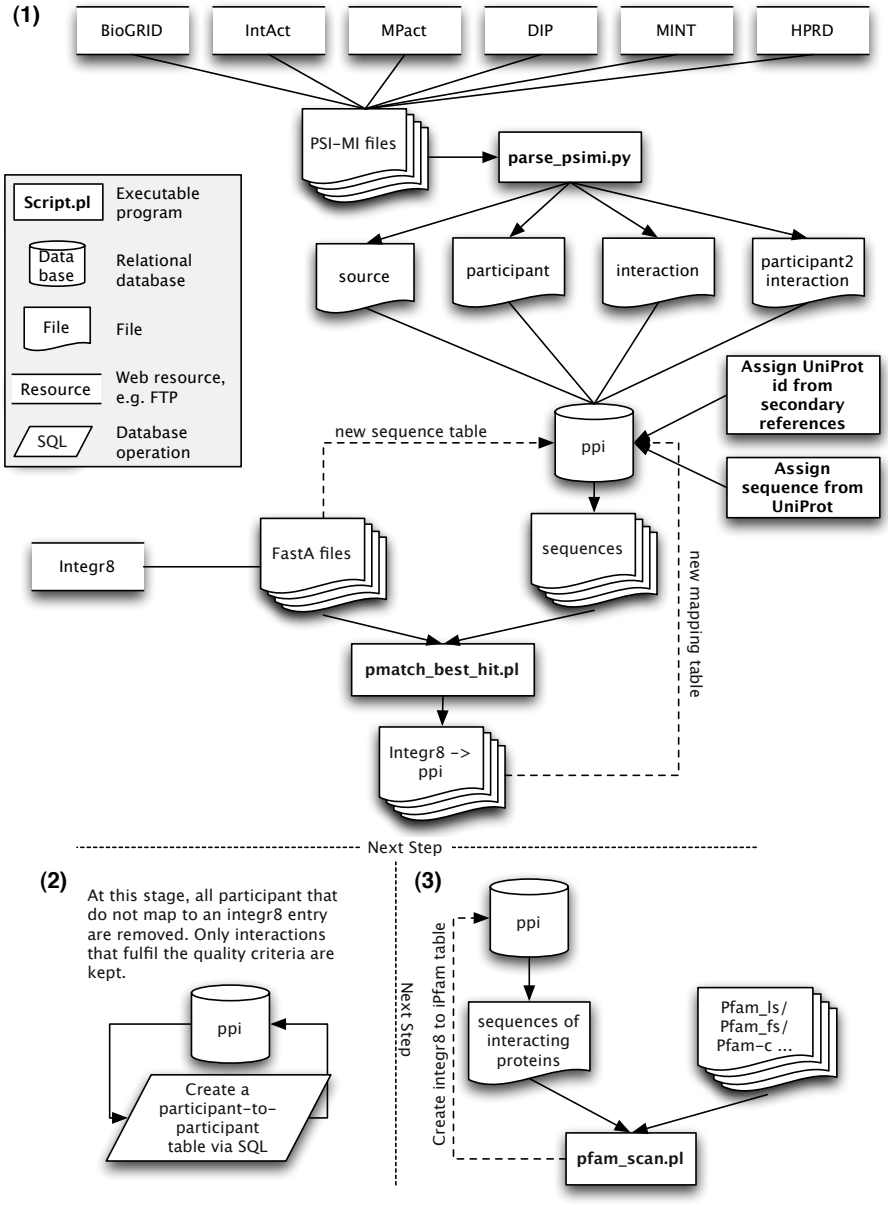

Figure 2.1: Flow-chart of protein-interaction database creation process. **(1)** Interaction information is loaded from numerous online resources by parsing flat-files in PSI-MI XML 2.5 format and subsequently stored in a database as 4 distinct tables. UniProt identifiers are assigned to each protein if secondary references are available. For proteins with no sequence information, the corresponding sequence in UniProt is assigned if possible. Sequence files for model species are downloaded from Integr8 and stored in the database. Integr8 sequences are then matched to interacting proteins of the same species using pmatch. The resulting mapping is loaded back into the database. **(2)** A new participant2participant table is created *via* a sequence of SQL queries. **(3)** Pfam domain annotations for each interacting protein (after mapping to integr8) are identified directly from the sequence using Pfam HMMs.

2.2.2 Filtering

There are many types of experiments used to derive protein interactions, with different properties and error rates. For this analysis, solely the properties of physically interacting proteins are of interest. Therefore, only interactions between exactly two proteins per experiment were considered. This is desirable because the real combination of interactions cannot be inferred from the data: Assuming a complex of 3 proteins A, B and C, several combinations are possible:

- $A \leftrightarrow B$ and $A \leftrightarrow C$
- $A \leftrightarrow B$ and $B \leftrightarrow C$
- $A \leftrightarrow B$, $A \leftrightarrow C$ and $B \leftrightarrow C$

Any one of these three combinations could reflect the biological condition, whereas the remaining two would introduce an error into the analysis. As a consequence, all protein complex data that were derived by co-purification methods were removed, unless a particular experiment had identified exactly two binding partners. All genetic interactions were also removed. For a list of the experimental method identifiers that were excluded see Table 2.1. This filtering step is applied at stage 2 in Figure 2.1.

2.2.3 Species

To allow cross-species comparisons, the data were split into five distinct species sets: *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*. It should be noted that the proportion of proteins for which an interaction is known varies from 13% in *C. elegans* to 92% in *S. cerevisiae*, see Table 2.2. This might affect the results if there is a systematic bias on the composition of a protein interaction set.

To prevent bias from multiple alternative versions of the same protein, all interacting proteins were mapped to reference proteomes as defined by Integr8 (Kersey *et al.*, 2005)

Table 2.1: List of experimental method identifiers that were excluded from the analysis. The controlled vocabulary for the PSI-MI terms can be found at `http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=MI`. The BioGRID terms are only available as part of the complete interaction database download. The term definition is shown in the *Description* column.

Method ID	Method DB	Description
MI:0001	PSIMI	"Interaction Detection Method" - data source unclear
MI:0045	PSIMI	"experimental interaction detection" - contains many data of unclear origin
10	BioGRID	Synthetic Lethality
11	BioGRID	Synthetic Growth Defect
12	BioGRID	Synthetic Rescue
13	BioGRID	Dosage Lethality
14	BioGRID	Dosage Growth Defect
15	BioGRID	Dosage Rescue
16	BioGRID	Phenotypic Enhancement
17	BioGRID	Phenotypic Suppression

using `pmatch`[1] (see Figure 2.1), a very fast pairwise sequence comparison algorithm developed by Richard Durbin. Approximately 12% of original sequence identifiers were lost in the mapping process, either if no sequence was provided with the original entry or if no significant matching sequence could be found in Integr8. The total number of missing unique proteins will be lower, as there are, on average, two original sequence identifiers for each Integr8 identifier.

2.2.4 *i*Pfam

The *i*Pfam database is derived from protein structures deposited in the PDB. Regions in every protein structure that match a Pfam domain are scanned for atomic contacts with residues in another Pfam domain. All such interacting domain pairs are stored in a database together with detailed information on the residues involved (Finn *et al.*,

[1]Unpublished, however it forms part of the Ensembl pipeline. The source-code is available from the Sanger Institute CVS repository: `http://cvs.sanger.ac.uk/cgi-bin/viewcvs.cgi/rd-utils/`

2005). Every *pair* of Pfam families that are found to interact in a PDB structure are called an *iPfam domain pair* throughout the text. Single Pfam families that are part of an *i*Pfam domain pair are then called *iPfam domains*. For example, in PDB entry 1k9a the two *i*Pfam domains SH2 (Pfam accession PF00017) and Pkinase_Tyr (PF07714) interact, therefore they form an *i*Pfam domain pair. In this study, *i*Pfam version 21 was employed, containing 2837 *i*Pfam domains, forming 4030 *i*Pfam domain pairs. Some *i*Pfam domain pairs are seen to form interactions between distinct peptide chains in the structure (*interchain*), while others form an interaction between two distinct domains within the same chain (*intrachain*). Out of the 4030 domain pairs, 2859 are found exclusively on two different chains (interchain), 623 are found exclusively within the same chain (intrachain) and 548 domain pairs are found both as inter- and intrachain pairs. It has been assumed that intrachain interactions can become interchain interactions and *vice-versa* as a result of a gene-fission/fusion events (Enright *et al.*, 1999). In this analysis, both inter- and intrachain interactions were used and compared where appropriate.

Figure 2.2 shows the species distribution of *i*Pfam domain pairs. *H. sapiens*, *E. coli* and *S. cerevisiae* are clearly over-represented compared to the other 1113 species with less than 179 complex structures. It is therefore expected to observe more matches to these species compared to the worse represented ones.

2.2.5 Prediction of crystal contacts

Not all interaction interfaces observed in crystal structures also occur *in vivo*. As I described in Section 1.1.1.4, non-biological interactions, here referred to as *crystal contacts*, are artefacts induced by the crystallisation process. I employed the NOXclass predictor to discriminate between biological interfaces and crystal contacts (Zhu *et al.*, 2006). NOXclass uses a range of sequence and structure based properties as feature vectors in a support-vector machine to classify interaction interfaces:

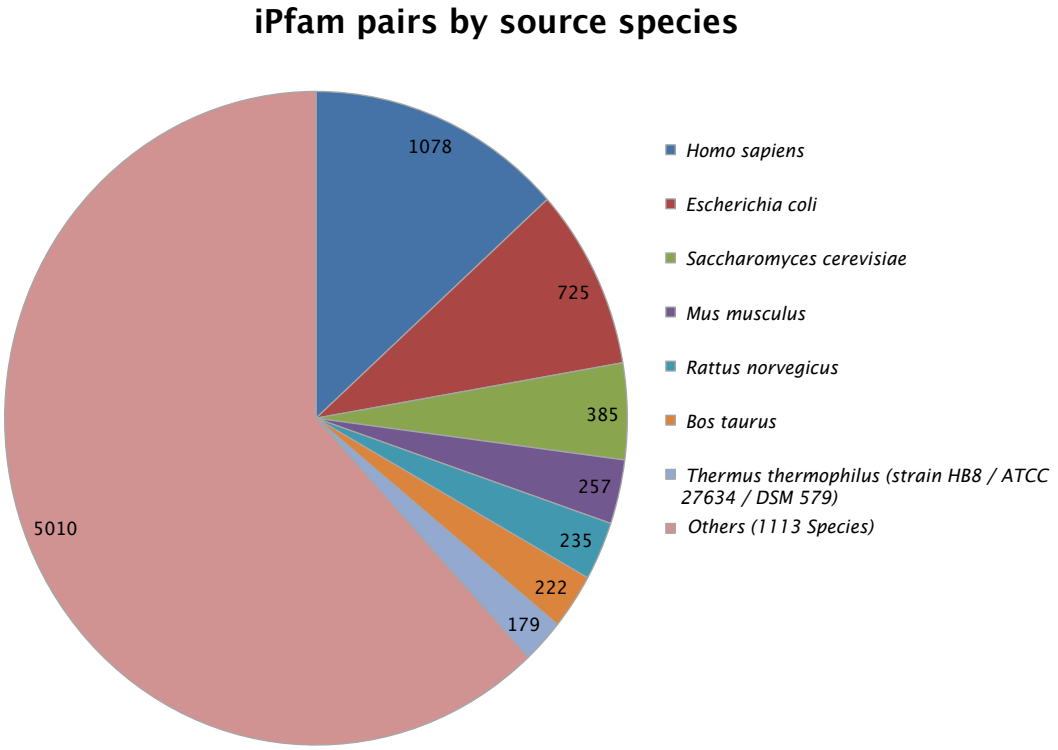

Figure 2.2: This pie chart shows how many *i*Pfam domain pairs were found in PDB structures from each species. The total number is larger than the 4030 unique *i*Pfam pairs in the database because an *i*Pfam pair can be found in structures from several species.

- Amino-acid (AA) composition of the interface
- Correlation between AA compositions of interface and the rest of the surface
- Distance between the AA compositions of the interfaces
- Conservation of interface residues
- Gap volume
- Interface area
- Solvent accessible surface

Reference values for these features were calculated on a set of 182 manually compiled biological and 106 crystal contact interfaces. According to the developers, NOXclass achieved 91.8% accuracy in a leave-one-out cross validation.

2.2.6 Random Networks

Randomised protein interaction networks with identical degree distributions were generated from the original filtered experimental interaction data for each species using two different methods. The first method will be referred to as *node sampling* (NS): In each randomisation step, a mapping is created that assigns every node a randomly chosen replacement node. In this way the edges of the network remain in place, while the nodes are shuffled randomly. It should be noted that the degree distribution per node is not maintained. Instead, this behaviour simulates a network with a high false positive rate, where random new connections between two proteins occur. The second method is referred to as *edge swapping* (ES). The methods implements the algorithm described by Maslov and Sneppen (2002). For a pair of randomly selected non-overlapping edges, the start and end nodes are swapped, unless the resulting edge already exists. This step is repeated $2 \cdot n$ times, where n is the total number of edges in the network. This

algorithm maintains the degree per node. This corresponds to the assumption that the observed number of interactions per protein reflects the real number of interactions the protein can form.

2.2.7 P-values

Unless otherwise specified, P-values for observations x were calculated as $P(X \geq x) = f(x; \mu, \sigma)$, where $f(x; \mu, \sigma)$ is the probability density function of the normal distribution with mean μ and standard deviation σ, where μ and σ are estimated through randomisation experiments. The density function thus provides the probability that a value less than or equal to x is observed by chance, given the distribution estimated by a random resampling method. Where appropriate, the inverse probability $P(X < x) = 1 - f(x; \mu, \sigma)$ was applied.

2.3 Results

2.3.1 Coverage of *i*Pfam domain pairs on different interactomes

I analysed the distribution of Pfam families known to interact from a PDB structure (*iPfam domain pairs*) in experimentally derived protein interactions (*experimental interactions*). The experimental interactions were filtered to only include interactions with exactly two partners (see Methods). The fraction of experimental interactions that contain at least one *i*Pfam domain pair is referred to as the *iPfam coverage*. Accordingly, the fraction of experimental interactions that contains any pair of Pfam domains (excluding the *i*Pfam domain pairs) is called the *Pfam coverage*.

Figure 2.3 shows the Pfam and *i*Pfam coverage for the analysed species as a column chart. The number of resolved protein interactions varies greatly between species, as does the size of the underlying proteome (see Table 2.2). The Pfam coverage lies between 51.74% and 82.38%. Given that almost 74% of all UniProt proteins contain

Table 2.2: For each species, I list the size of the proteome as defined in Integr8 and the fraction of this proteome that is represented in the protein interaction sets, followed by the total number of binary protein interactions and the fraction of those that contain an *i*Pfam domain pair. The last columns show the results of the network shuffling experiments (both NS and ES): The mean of interactions with an *i*Pfam domain pair in the randomised networks and the corresponding standard deviations were used to compute the likelihood of observing the original results by chance.

Species	Proteins in proteome	% proteome in interaction set	Total number of interactions	Interactions with *i*Pfam domain pair	Randomised mean		Standard deviation		P-Value	
					NS	ES	NS	ES	NS	ES
E. coli	4346	47.26%	7185	960	712	37	13	6	$4.69 \cdot 10^{-82}$	$< 10^{-100}$
S. cerevisiae	5834	92.12%	45804	2524	679	465	23	23	$< 10^{-100}$	$< 10^{-100}$
C. elegans	23491	13.24%	5403	275	80	46	8	7	$< 10^{-100}$	$< 10^{-100}$
D. melanogaster	23693	36.15%	31137	1002	295	255	19	15	$< 10^{-100}$	$< 10^{-100}$
H. sapiens	54035	18.61%	36040	5521	1391	852	42	46	$< 10^{-100}$	$< 10^{-100}$
Results excluding interchain *i*Pfam domain pairs										
E. coli				930	682	33	12	6	$1.06 \cdot 10^{-88}$	$< 10^{-100}$
S. cerevisiae				2457	646	452	24	23	$< 10^{-100}$	$< 10^{-100}$
C. elegans				267	76	43	8	7	$< 10^{-100}$	$< 10^{-100}$
D. melanogaster				964	271	230	19	15	$< 10^{-100}$	$< 10^{-100}$
H. sapiens				5350	1,295	746	38	49	$< 10^{-100}$	$< 10^{-100}$
Results only on non-crystal contact *i*Pfam domain pairs										
E. coli				845	615	31	13	6	$9.61 \cdot 10^{-73}$	$< 10^{-100}$
S. cerevisiae				2010	528	368	21	19	$< 10^{-100}$	$< 10^{-100}$
C. elegans				233	66	37	8	6	$< 10^{-100}$	$< 10^{-100}$
D. melanogaster				855	226	195	17	14	$< 10^{-100}$	$< 10^{-100}$
H. sapiens				4840	1,123	663	36	44	$< 10^{-100}$	$< 10^{-100}$

at least one Pfam match[1], this is not by itself surprising. The *i*Pfam coverage, shown in light blue in Figure 2.3, is much smaller, ranging from 3.22% in *D. melanogaster* to 15.32% in *H. sapiens*. In *S. cerevisiae* the species with the most comprehensively studied interactome, the *i*Pfam coverage is 5.51%, while the average between the five species is 8.50%.

The fact that only a small fraction of protein interactions contain known domain pairs could be a result of the scarcity of available structures of protein complexes. Therefore, I asked whether the observed *i*Pfam coverage is larger than would be expected by chance. To test this, I created 1000 random networks per species using the algorithms described in Methods. I then calculated the *i*Pfam coverage on the protein interactions in each randomised network. The green bars in Figure 2.3 show the random distribution calculated using the node-sampling algorithm. Results of the edge-swapping randomisation are similar and therefore not plotted. Mean and standard deviations of both randomisation experiments are however listed in Table 2.2. No P-value (see Methods) was greater than $1.84 \cdot 10^{-06}$. This proves that the observed *i*Pfam coverage is significantly higher than expected and *i*Pfam domain pairs are enriched in real experimental protein interactions.

2.3.2 Domain pair frequency within interaction networks

To understand why *i*Pfam domain pairs occur more often in experimental interactions than expected by chance, I analysed the distribution of *i*Pfam domain pairs relative to the number of covered experimental interactions. Figure 2.4 shows a plot of the frequency of *i*Pfam domain pairs over the number of interactions they occur in, reflecting how many *i*Pfam domain pairs cover how many experimental interactions. Domain pairs to the left of the plot can be called *specific* domain pairs, as they only occur in very few covered experimental interactions. Conversely, domain pairs to the right of

[1]For Pfam version 21, 2343026 out of 3169275 sequences had at least one significant Pfam hit, corresponding to 73.92%.

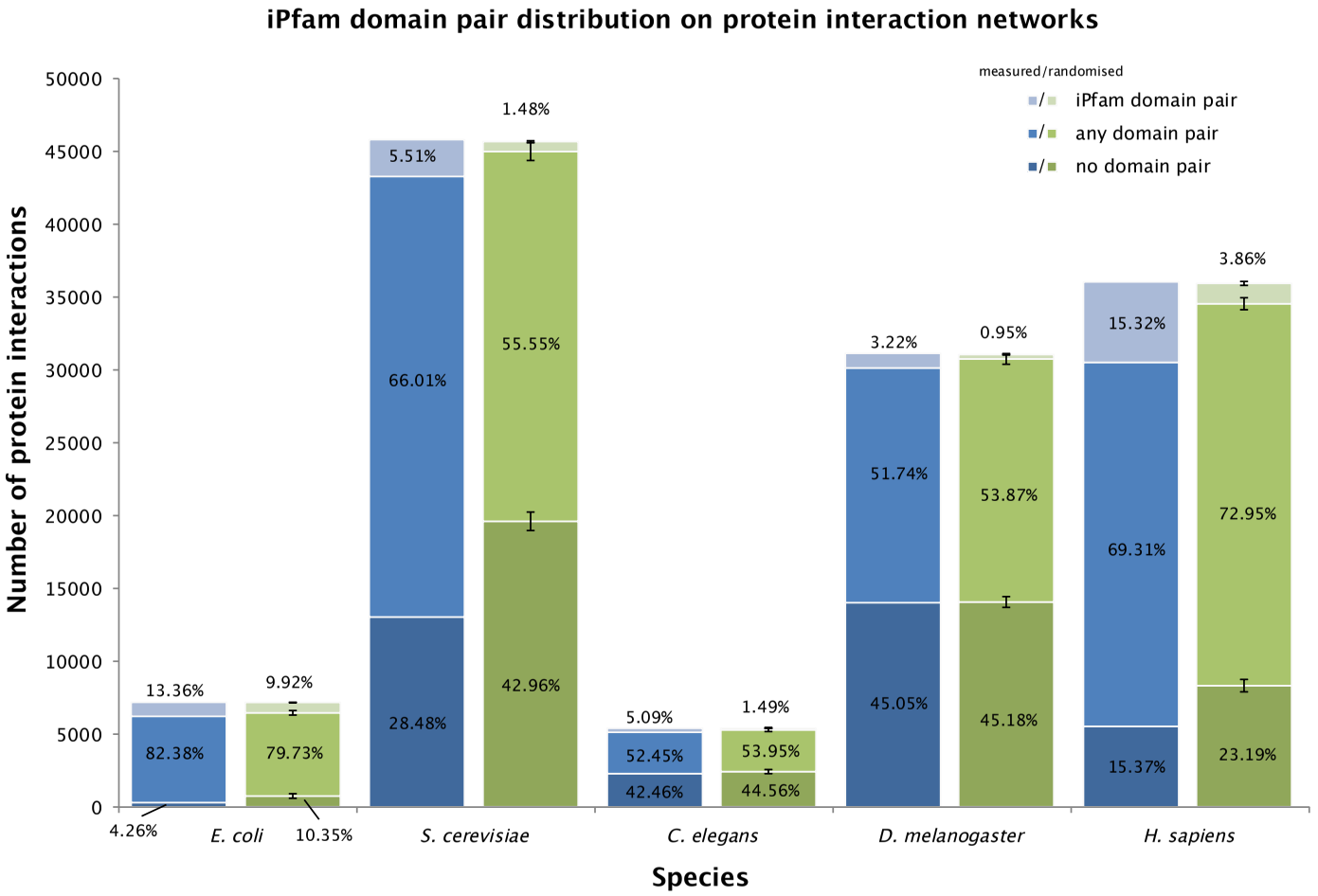

Figure 2.3: Pfam and *i*Pfam coverage on real (blue) and randomised (green) interaction networks. For each species, the height of the columns reflects the number of known protein–protein interactions in the data set. The columns are split according to the proportion of interactions that contain an *i*Pfam domain pair (top), that contain any other Pfam domains on both proteins (middle), and those that contain no Pfam domain pair (bottom).

the plot occur in a large number of different covered experimental interactions and can be called *promiscuous* domain pairs.

All five distributions in Figure 2.4 resemble a power law distribution, according to the good fit of log-linear functions ($\log(f(x)) = k \log x + \log a$) shown as dotted lines. The slopes k of the eukaryotic distributions are very similar (between -1.31 and -1.61), while *E. coli* has a markedly smaller slope (-2.13). If I assume *E. coli* to be an exemplary prokaryote, this suggests that the ratio of specific to promiscuous *i*Pfam domain pairs differs between eukaryotes and prokaryotes, whereby *E. coli* features fewer multiply reoccurring *i*Pfam domain pairs.

The power law distribution of *i*Pfam frequencies implies that the majority of covered protein interactions can be attributed to a minority of *i*Pfam domain pairs: 88.1% of *S. cerevisiae* and 95.0% of *H. sapiens* covered experimental interactions contain an *i*Pfam domain pair that occurs more than once. This explains the highly significant P-values listed in Table 2.2. Conversely, 46.0% of the *i*Pfam domain pairs in *S. cerevisiae* and 37.3% in *H. sapiens* are seen in just one experimental interaction.

2.3.3 Promiscuous domain pairs

As I showed above, the distribution of *i*Pfam domain pairs is composed of both very promiscuous pairs which are seen in many interactions and specific domain pairs which occur in only very few distinct interactions. Appendix A lists the 20 most frequent *i*Pfam domain pairs in the experimental protein interactions of all 5 model organisms. Similarly, Appendix B lists the 20 most frequent *i*Pfam domains alone.

As expected, more frequent domains are also more likely to be found as pairs in interacting proteins. The network randomisation experiments described earlier assert that this relationship between frequency of the individual domains and the frequency of the domain pairs is not the underlying reason for the observed *i*Pfam coverage, otherwise one would expect to observe a similar coverage in randomly reshuffled networks.

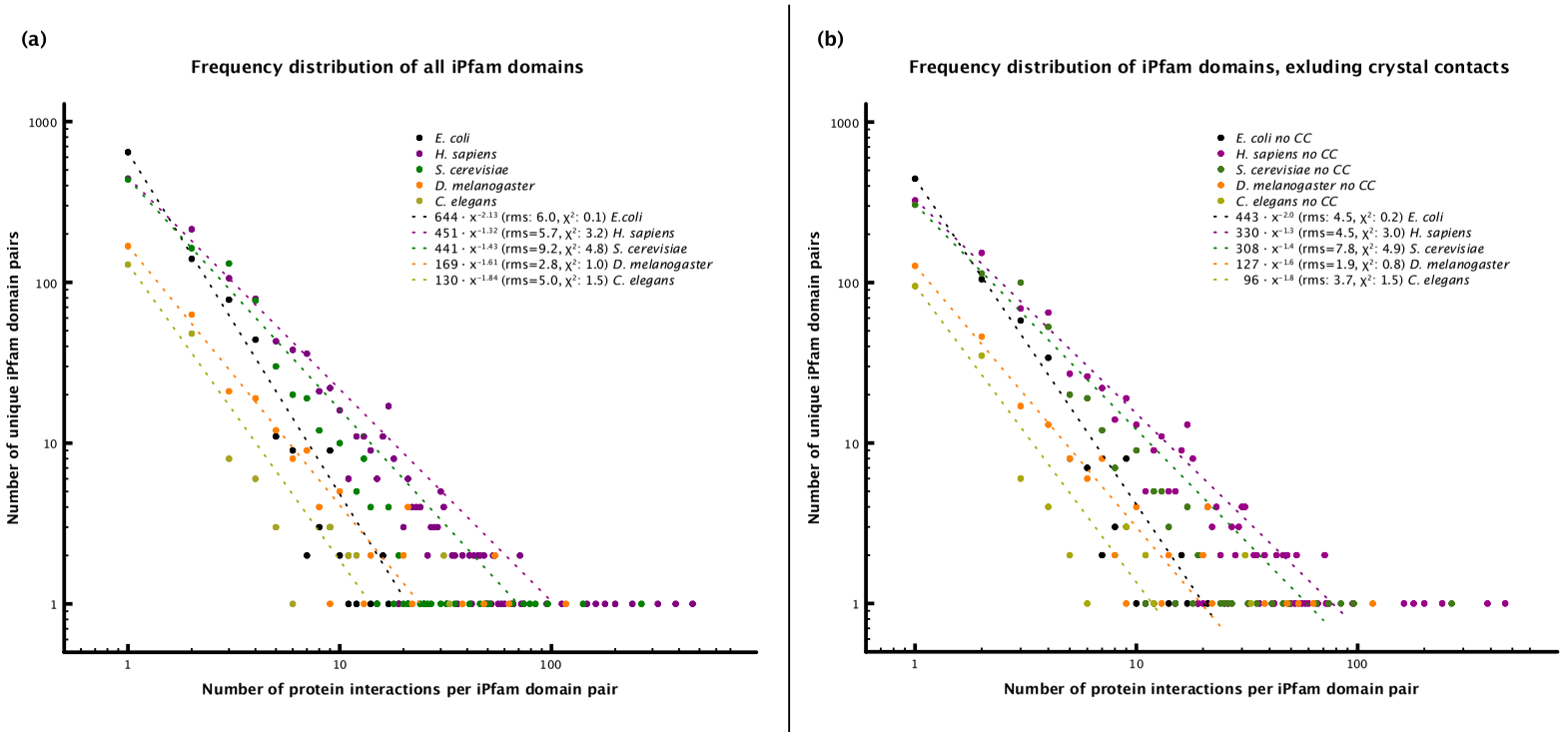

Figure 2.4: Scatter plot illustrating how many *i*Pfam domain pairs occur in how many proteins interactions per species. First, I counted the number of protein interactions each *i*Pfam domain pair occurs in. The x-axis represents the *occurrence frequency*. Then, I counted the number of *i*Pfam domain pairs with the same occurrence frequency and plotted that along the y-axis. Points to the left show how many *i*Pfam domains occur in only a few different interactions, whereas points to the right show how many *i*Pfam domain pairs are found in a wide variety of experimental interactions. Logarithmic axes were used to stress the log-linear distribution. For each group of points, a power law curve was fitted. The parameters and goodness-of-fit statistics are listed in the figure legend. Curve fitting was performed in Plot (`http://plot.micw.eu/`). **(a)** All *i*Pfam domain pairs were counted, summing to 2169 *i*Pfam domain pairs on 10282 experimental interactions. **(b)** Only *i*Pfam domain pairs from structures with $< 90\%$ NOXClass crystal contact P value were counted, summing to 1524 *i*Pfam domain pairs on 8784 experimental interactions.

The only prokaryote in this comparative analysis, *E. coli* features many transcription factor activity related *i*Pfam domain pairs amongst the 20 most frequent pairs. Examples include the HTH_1 domain (PF00126, Helix-Turn-Helix domain, a component of transcription factors) or Helicase_C (PF00271, a component of DNA unwinding proteins) with numerous binding partners, alongside some domains which are particular to prokaryotes, such as the Response_reg domain (PF00072), the signal receiver of the bacterial two-component system.

The DNA-regulation related *i*Pfam domains are also frequently observed in interactions of eukaryotes. However, the most frequent pairs involve protein kinase domains as well as recognition domains such as SH2 or SH3. This is likely to be a result of the large number of signalling pathways that underpin the biology of complex multi-cellular organisms.

It should be noted that in the PDB structures, some of the observed domain pairs (Helicase_C $\leftrightarrow$ DEAD, Pkinase_C $\leftrightarrow$ SH3_1 and others) are only seen to interact within one protein (intrachain interactions) as opposed to interactions between two distinct proteins (interchain interaction). Out of 2169 *i*Pfam domain pairs that are observed in any of the 5 species, 307 ($\approx 15\%$) are exclusively interchain. Table A.2 in Appendix A lists the 20 most frequent *i*Pfam domain pairs, excluding those which are only observed to interact within a chain. The key findings do not change: DNA-regulation and signal transduction related domain pairs are still prevalent. Similarly, excluding the 10%[1] of *i*Pfam domain pairs which are only observed in structures which are likely to be crystal contacts does not fundamentally alter the composition of the promiscuous domain pairs.

2.3.4 Domain co-ocurrences

A basic assumption of this study is that interacting proteins that contain an *i*Pfam domain pair actually interact through these domains. This, of course, is not necessarily

[1]Out of the 2169 *i*Pfam domain pairs which are observed in at least one interactome, 1690 pairs could be checked for their crystal-contact status. Out of these 1690, 167 ($\approx 10\%$) were removed.

the case. Although it has been shown that sequence similarity is linked to the mode of interaction (Aloy *et al.*, 2003), not every protein interaction that contains an *i*Pfam domain pair is necessarily mediated by exactly this domain pair. In fact, the observed high frequency of certain signalling domains such as SH2, SH3_1 or Pkinase_tyr can partially be attributed to the fact that they often reside in succession on the same protein. Table C.1 in Appendix C contains a list of the 30 most frequent *i*Pfam domain architectures in the analysed interacting sequences.

While I cannot assign the correct interacting domains with certainty, I attempted to ascertain that domain co-ocurrence is not causative for the observed enrichment of *i*Pfam domain pairs in interacting proteins. To do so, I analysed the distribution of single-domain proteins only. These are proteins which contain only a single *i*Pfam domain, and this domain stretches over at least 70% of the length of the sequence. In the same way as before, I counted the number of interacting single-domain proteins with an *i*Pfam domain pair and compared this to 1000 randomly reshuffled networks.

Table 2.3: Frequency of *i*Pfam domain pairs on single-domain proteins. Real observed number of *i*Pfam domain pairs in interaction between single domain proteins is listed in column two. Results of random resampling by node sampling (NS) or edge swapping (ES) and associated P-values are also shown.

Species	**Real observed**	**Resampling mean**		**Resampling SD**		**P-value**	
		NS	**ES**	**NS**	**ES**	**NS**	**ES**
E. coli	361	260	6	10	2	$2.8 \cdot 10^{-25}$	$< 10^{-100}$
S. cerevisiae	324	116	12	9	3	$< 10^{-100}$	$< 10^{-100}$
C. elegans	43	10	1	3	1	$9.9 \cdot 10^{-30}$	$< 10^{-100}$
D. melanogaster	53	22	4	5	2	$8.6 \cdot 10^{-12}$	$< 10^{-100}$
H. sapiens	513	143	19	11	4	$< 10^{-100}$	$< 10^{-100}$

The results summarised in Table 2.3 clearly show that real protein interactions are enriched for *i*Pfam domains even if only single-domain proteins are considered.

2.3.5 *i*Pfam domain pairs in stable complexes of *S. cerevisiae*

I tested whether *i*Pfam domain pairs are enriched in known protein complexes from *S. cerevisiae*, using the collection of complexes described by Gavin *et al.* (2006) as the reference. This is interesting because domain–domain interactions are thought to be particularly important for strong, obligate interactions between subunits of protein complexes, as opposed to weaker transient interaction which are thought to be also often mediated by smaller *linear motifs* as described by *e.g.* Neduva and Russell (2005).

While the data of Gavin *et al.* provides a very systematic analysis of complexes in *S. cerevisiae*, it was unfortunately derived by affinity purification, only containing very few binary interactions (see Methods on "Filtering"). I therefore counted the number of *complexes* with at least one *i*Pfam domain pair between any two members of the complex, rather than analysing binary interactions. Out of 491 complexes described by Gavin *et al.*, 472 contained at least one pair of proteins with an *i*Pfam domain pair (96.13%). Testing the significance of this result can not easily be done by network resampling: Shuffling the existing nodes will not change the network substantially when all proteins within one complex are assumed to be connected. Instead, I replaced all proteins in all complexes with randomly sampled proteins from the *S. cerevisiae* proteome. This tests whether the observed *i*Pfam coverage on the complexes is related to the composition of the complexes. After 1000 resamplings, an average of 447 complexes of randomly chosen proteins contained an iPfam domain pair, with a standard deviation of 6, giving a P-Value of $5.7 \cdot 10^{-5}$ to observe 472 complexes with an *i*Pfam domain pair purely by chance. This indicated that yeast complexes are slightly enriched for *i*Pfam domain pairs.

Are the *i*Pfam domain pairs that occur in *S. cerevisiae* complexes evenly spread over all complexes, or do some complexes contain more *i*Pfam domain pairs than others? In other words: If protein pairs were chosen by chance from all complexes, would I observe the same distribution of pairs per complex? Employing a χ^2-test, I verified

that the observed distribution of protein pairs with an *i*Pfam domain pair per complex deviates significantly from expectation, given the total number of protein pairs per complex ($P = 4.9 \cdot 10^{-4}$). Some complexes contain a greater number of *i*Pfam domain pairs, while other complexes do not contain any at all. This suggests that some sets of domain pairs are specific to certain complexes or pathways. A typical example is the RNA polymerase II complex (IntAct id: EBI-815049) which contains numerous *i*Pfam domain pairs that are specific to this complex.

2.3.6 *i*Pfam domain pair conservation between species

Within the 3 to 15% of experimental interactions covered by *i*Pfam, I analysed the conservation of *i*Pfam domain pairs between species. I call an *i*Pfam domain pair *conserved* when the same pair is observed in experimental interactions of two different species. The matrix in Table 2.4 shows the pair-wise conservation of *i*Pfam domain pairs. The prokaryote *E. coli* shares fewer *i*Pfam domain pairs (an average of 31.8%) with the eukaryotic species, compared to the overlap between the eukaryotes (an average of 69.3%).

I performed pair-wise Fisher-Exact-Tests to evaluate whether the overlap between the sets of *i*Pfam domain pairs is statistically significant, denoted as up- or down pointing arrows in Table 2.4. The significance of the overlap between *E. coli* and the eukaryotic species gradually gets smaller towards *H. sapiens*, where I in fact observe a smaller than expected overlap.

Figure 2.5 shows a Venn diagram of the mutual overlaps between the two eukaryotes *S. cerevisiae* and *H. sapiens* and the prokaryote *E. coli*. This figure outlines the results in Table 2.4: While the two eukaryotes share 522 domain pairs, only 375 *i*Pfam domain pairs are shared between *S. cerevisiae* and *E. coli*, and only 245 between *E. coli* and *H. sapiens*. However, it should be noted that 43.9% of the observed *i*Pfam domain pairs in *E. coli* are also observed in one of the two eukaryotes, and 202 *i*Pfam domain

Table 2.4: The Table shows the number of co-occurences of *i*Pfam domain pairs between two species. The right-most column lists the total number of unique *i*Pfam pairs found in each species' experimental interactions. The lower triangle of the table show the fraction of all *i*Pfam domain pairs that is shared between the two species (relative to the smaller set). Arrows denote significant enrichment (↑) or depetion (↓) for shared domain pairs as determined by a Fisher exact test. If not explicitly stated, P-values were below 10^{-16}.

	E. coli	*S. cerevisiae*	*C. elegans*	*D. melanogaster*	*H. sapiens*	*i*Pfam domain pairs in total
E. coli		375	63	64	245	952
S. cerevisiae	39.5% ↑		138	193	522	949
C. elegans	30.7% ↑ ($P = 0.01$)	67.3% ↑		116	183	205
D. melanogaster	31.2% ↓ ($P = 0.03$)	58.8% ↑	56.6% ↑		291	328
H. sapiens	25.7% ↓ ($P = 0.002$)	55.0% ↑	89.3% ↑	88.7% ↑		1183

pairs are even conserved amongst all three species. Appendix D contains a list of these most conserved *i*Pfam domain pairs. The *i*Pfam domains in these conserved pairs are predominantly related to housekeeping activities such as translation, replication or basic energy metabolism, suggesting that the shared *i*Pfam domain pairs could trace back as far as the last universal common ancestor. A list of GO annotation for the overlapping *i*Pfam domain pairs can be found in Appendix E.

Given that there are great differences between *i*Pfam domain pairs regarding their frequency in interacting proteins, I wondered whether this "promiscuity" is also conserved between different species. I compared the *i*Pfam domain pair frequencies between *H. sapiens* and *S. cerevisiae* directly, as shown in Figure 2.6.

I measured a Spearman correlation coefficient of 0.43 between the coverages of *S. cerevisiae* and *H. sapiens* conserved *i*Pfam domain pairs. To test the significance of this correlation, I recalculated the correlation 1000 times after shuffling the values in one species. From these random results, I derive a P value of $1.8 \cdot 10^{-20}$. Evidently,

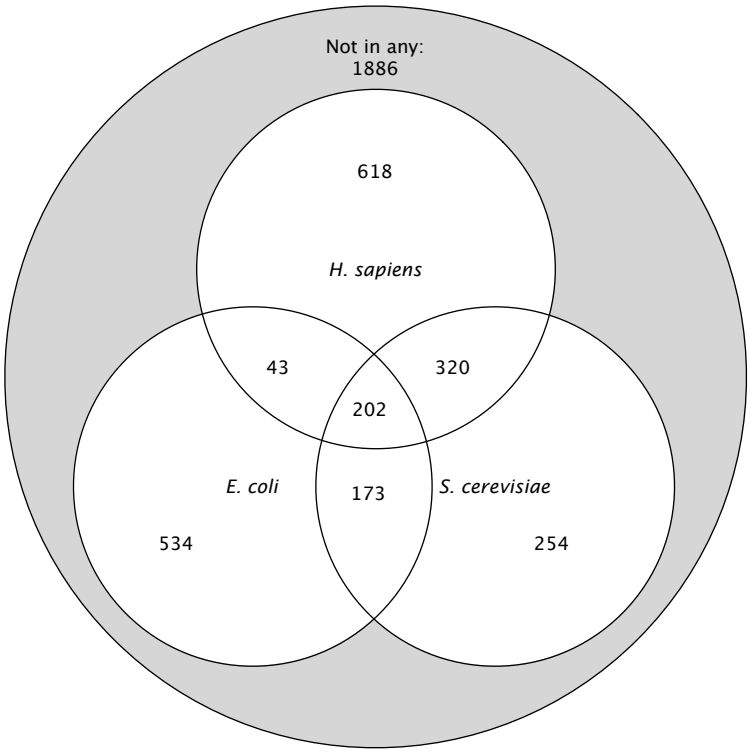

Figure 2.5: The three circles represent the *i*Pfam domain pairs observed in the respective species. The overlaps denote co-observed *i*Pfam domain pairs. The grey set in the background represents *i*Pfam domain pairs not found in the three species.

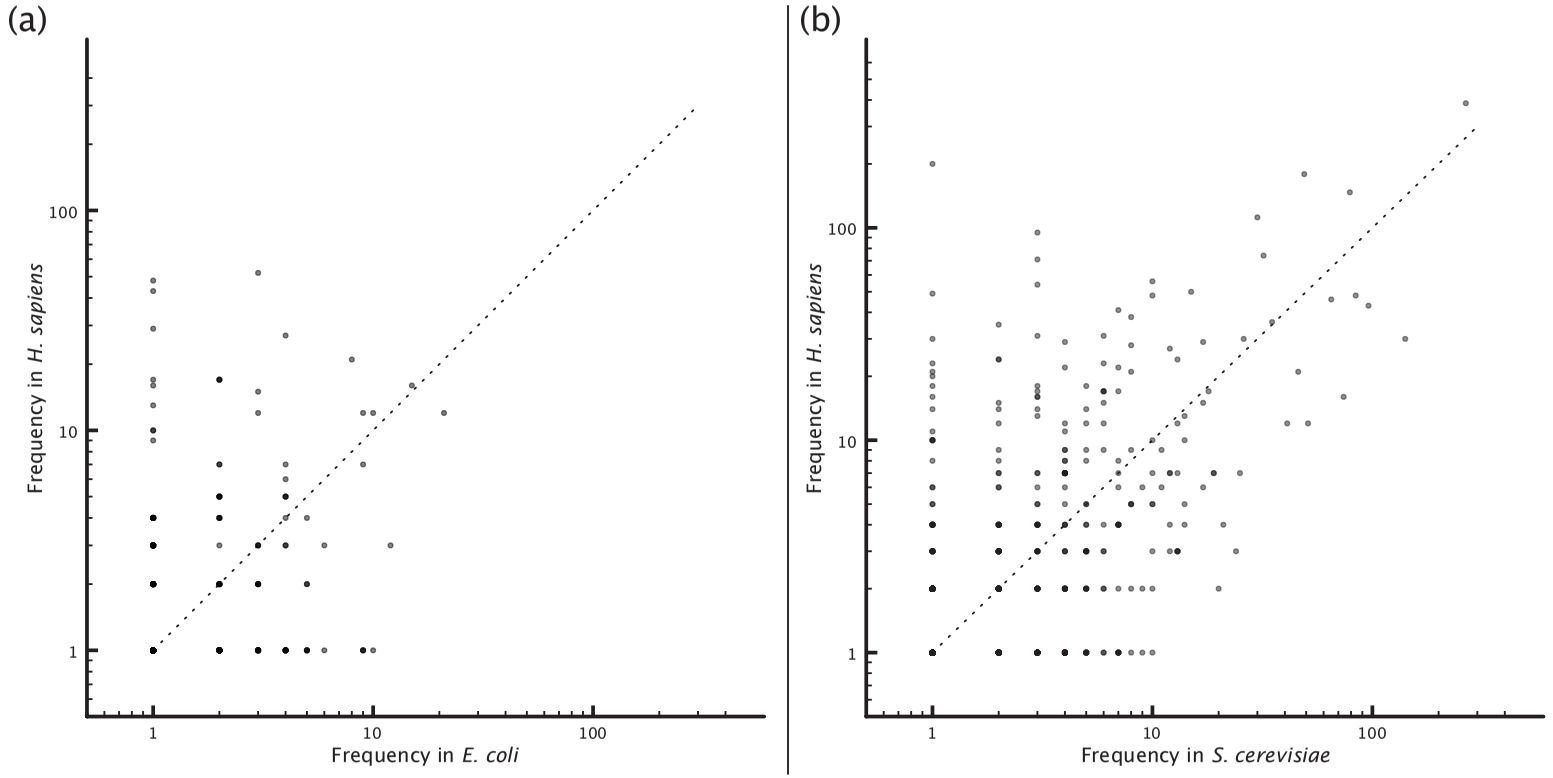

Figure 2.6: Comparison of domain pair frequency between species. **(a)** *E. coli* compared to *H. sapiens*: There is almost no visible correlation between the frequencies. **(b)** *S. cerevisiae* compared to *H. sapiens*: The correlation is much more pronounced, particularly for *i*Pfam domain pairs observed in more than 10 interactions. This is confirmed by a Spearman correlation of 0.42 with a high significance when tested against random re-orderings, see main text. Points are drawn 80% transparent, so darker points denote multiple *i*Pfam domain pairs with the same frequencies in both species. Dotted line denotes the intersect $y = x$.

*i*Pfam domain pairs with a large number of occurrences in *S. cerevisiae* tend also to be more frequent in *H. sapiens*. In comparison, the correlation between *E. coli* and *H. sapiens* is relatively weak (Spearman correlation: 0.13). Again, this difference is most likely a result of the expansion of signalling-related interacting domains in the eukaryotic lineage.

2.3.7 Predicting the total number of *i*Pfam domain pairs in nature

How many *i*Pfam domain pairs would be required to eventually cover all protein interactions? Aloy and Russell (2004) attempted to predict this parameter, estimating that ≈ 10000 domain pairs would cover all protein interactions. Similar to their approach, I make a linear estimation with the following factors:

χ_S The number of *i*Pfam domain pairs observed in species S

θ_S The number of observed interactions in species S that contain an *i*Pfam domain pair

Θ_S The total number of observed interactions in species S

ψ_S The number of proteins from species S that are seen in an interaction screen

Ψ_S The proteome size for species S

ξ_S The number of Pfam domains observed in all protein of species S

Ξ The total number of known Pfam domains

I denote the estimated number of *i*Pfam domain pairs in species S with $\hat{x}_S$. The formula I apply is

$$\hat{x}_S = \chi_S \cdot \frac{\Theta S}{\theta_S} \cdot \frac{\Psi_S}{\psi_S} \tag{2.1}$$

This means I scale the observed number of *i*Pfam domain pairs to cover all observed interactions. I then use the relative proteome coverage to estimate the total number

Table 2.5: Parameters for the prediction of the number of interacting domain pairs in nature. Prediction results are shown in **bold font**.

Species	χ_S [a]	Θ_S [b]	θ_S [c]	Ψ_S [d]	ψ_S [e]	$\hat{x}_S$ [f]	ξ_S [g]	$\hat{x}$ [h]
E. coli	952	7185	960	4346	2054	**15075**	2070	**65234**
S. cerevisiae	949	45804	2524	5834	5374	**18696**	2119	**79027**
C. elegans	205	5403	275	23491	3110	**30422**	2612	**104324**
D. melanogaster	328	31137	1002	23693	8564	**28198**	2777	**90952**
H. sapiens	1183	36040	5521	54035	10055	**41499**	3476	**106936**

[a] The number of *i*Pfam domain pairs observed in species S
[b] The total number of observed interactions in species S
[c] The number of observed interactions in species S that contain an *i*Pfam domain pair
[d] The proteome size for species S
[e] The number of proteins from species S that are seen in an interaction screen
[f] The predicted total number of *i*Pfam domain pairs in species S
[g] The number of Pfam domains observed in all protein of species S
[h] The estimated total number of *i*Pfam domains in all species

of *i*Pfam domain pairs in all proteins. Finally, I follow the argument of Aloy and Russell that the number of Pfam families seen in species S indicates the fraction of the protein universe represented in the species. I therefore predict the total number of *i*Pfam domain pairs $\hat{x}$ as

$$\hat{x} = \hat{x}_S \cdot \frac{\Xi}{\xi_S} \tag{2.2}$$

Both parameters and results of the calculation are shown in Table 2.5. Depending on the species the calculations were based on, the estimates for the total number of *i*Pfam domain pairs range from 65234 to 106936, with an average of 89295.

2.4 Discussion

2.4.1 Many domain–domain interfaces remain to be resolved

*i*Pfam in its current form covers only a small portion of the interactome of various species. For *S. cerevisiae*, the species with the largest fraction of known interactions, only 5.51% of the protein interactions contain an *i*Pfam domain pair. Even in *H. sapiens*, where I suspect slight ascertainment bias due to the overrepresentation of disease-related proteins in both the PDB and protein interaction databases, 85% of protein interactions do not contain an *i*Pfam domain pair (see Figure 2.3). This reveals the limits of our current understanding of the molecular structure of protein interactions.

In contrast, Figure 2.3 also shows that a majority of protein interactions contain at least one pair of Pfam domains. While there is no structural information about putative interactions between these pairs, this fraction can already be analysed using statistical methods to identify putative domain interactions (Jothi *et al.*, 2006; Lee *et al.*, 2006; Riley *et al.*, 2005). This in turn creates new targets for future structural genomics projects (Bravo and Aloy, 2006). Prioritising these targets according to the number of covered experimental interactions could increase the coverage of databases like *i*Pfam quickly.

I thus tried to estimate how many *i*Pfam domain pairs exists in all interactomes. My prediction is that there are approximately 90000 interacting domain pairs in nature, almost an order of magnitude more than the 10000 domain interaction types proposed by Aloy and Russell (2004) whose analysis was based on fewer data. While all such estimates should be taken with caution, my results imply that only about 5% of all structural domain pairs are represented in *i*Pfam. The aforementioned statistical methods can currently only cover a small fraction of this domain interaction space. For example, Riley *et al.* report only 3005 interacting domain pairs which could be inferred from protein interactions. It thus seems that the majority of domain–domain

interactions remain unknown.

I maintain, nevertheless, that analysing the structures of more interacting proteins is worthwhile. Solving protein structures is still a time-consuming task, so a call for time and resources to be spent on solving domain–domain interaction examples requires sufficient justification. I find that *i*Pfam domain pairs occur significantly more often in experimental interactions than would be expected by chance. This requires that at least a subset of the *i*Pfam domain pairs are reused in several experimental interactions. Also, there is substantial conservation between the sets of interacting domain pairs in different species. That means that a structural model for the interactions of numerous proteins can be derived from a single structure. These models can for example be used to investigate human disease genes, as I will demonstrate in the next chapter.

2.4.2 *i*Pfam domain pairs can act as modules

Despite the low overall coverage, *i*Pfam domain pairs are found in more protein interactions than would be expected by chance (see Table 2.2). This statistical overrepresentation suggests that certain *i*Pfam domain pairs constitute modules of molecular recognition which are reused in different protein interactions (Aloy and Russell, 2004). In fact, the characteristic power law distribution seen in Figure 2.4 hints at the fact that a minority of *i*Pfam domain pairs cover a large portion of the protein interactions. I find the most frequent *i*Pfam domain pairs in eukaryotes to be recognition domains in signal transduction. This suggests that the most promiscuous domain pairs actually function as reusable modules of molecular recognition. In a related study, Basu *et al.* (2008) noticed that domains that co-occur with a large number of diverse other domains often form protein interactions. They also note that signalling-related domains are the most frequently co-occuring domains in eukaryotes, which agrees well with my findings.

Conversely, a large number of *i*Pfam domain pairs are specific to a small number

of protein interactions. This implies that recognition specificity amongst proteins is often achieved by maintaining an exclusive interacting domain pair. This could pose a problem for purely statistical approaches to infer domain interactions that rely on the frequency with which domain pairs are observed in interacting proteins: if for many interfaces the real interacting domain pair will only occur in a single pair of proteins, elucidating the corresponding domain pair will not be detected.

In my analysis, I addressed several potential sources of error that could introduce a bias. Firstly, the collection of domain pairs in *i*Pfam consists of both inter- and intrachain interaction pairs. Also, there is a potential for false positive *i*Pfam domain pairs due to crystal contacts that are mistaken for biological interfaces. I analysed the distribution of *i*Pfam domain pair frequency excluding both intrachain interaction- and potential crystal contact derived *i*Pfam domain pairs, respectively. Neither restriction affected the basic finding that *i*Pfam domains are enriched in real protein interactions and that the most common *i*Pfam domain pairs are recognition modules.

2.4.3 *i*Pfam domain pairs are conserved during evolution

*i*Pfam domain pairs are not only recurrent within the protein interaction network of one species. They also appear to be conserved between species. In a small set of protein structures from *S. cerevisiae*, it has been shown that interacting domain pairs are more conserved than non-interacting domain pairs (Jothi *et al.*, 2006). In another study, Gandhi *et al.* (2006) have assessed the conservation of protein interactions by counting the number of interacting proteins in various species that are orthologous to each other (often called *interologs*). They found only 16 interologs that were conserved in *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*.

Conversely, I find that 83 *i*Pfam domain pairs are conserved in the experimental interactions of these four eukaryotic species. Even between a prokaryote like *E. coli* and the two eukaryotes *S. cerevisiae* and *H. sapiens* there are 202 conserved *i*Pfam

domain pairs. These domains are predominantly related to transcription, translation and other essential cellular activities, which is in congruence with the findings of Gandhi *et al.*. However, conservation at the domain level appears to be stronger than at the level of orthologous proteins. This not only supports the call for more structures of domain–domain interactions to be resolved, but also raises the question of whether one could establish a comprehensive set of domain interactions that were present in the last universal common ancestor.

Although the low overall *i*Pfam coverage somewhat hampers the interpretation of my results, it looks as if there has been a diversification of domain interactions from *E. coli* to *H. sapiens*. While more than half of the *i*Pfam domain pairs in *E. coli* have been retained throughout evolution, numerous new ones seem to have emerged in eukaryotic development. The significant positive correlation in the frequency of *i*Pfam domain pairs conserved between *S. cerevisiae* and *H. sapiens* also suggests that the binding interfaces are more often kept or even reused rather than lost in the course of evolution.