Chapter 1

Introduction

The interactions between proteins are an important component of organismal complexity. As a result, there has been rising interest in protein interactions, bringing about developments to automate their detection. This growing flood of molecular interaction data has been compared to the development of genome sequencing in the past decade, where the number of sequences deposited in public databases grew rapidly over the years (Sharan and Ideker, 2006). For example, more than 20000 human and 45000 *S. cerevisiae* protein interactions have been deposited in protein interaction databases (Gandhi *et al.*, 2006) and many more can be inferred from other model organisms, but it is assumed that this only constitutes a fraction of the full protein interaction network in a human cell (Hart *et al.*, 2006).

One of the key findings that has helped to tackle the data avalanche in genomics is that genes, or at least parts of a gene, fall into evolutionarily related families with homologous sequence. This means that it is possible to summarise thousands of individual sequences into a single group which is likely to share similar structural and often also functional properties. For coding genes, protein family databases such as Pfam (Finn *et al.*, 2008) collect these data and allow to quickly search new sequences for homology against known families.

The evolutionary relationships that can be inferred in this way hold great potential for the analysis of interaction networks. They can both assist in understanding the evolution of observed connections, as well as allow us to make predictions on the behaviour of proteins which belong to a family but have not themselves been thoroughly studied.

In this introduction, I will first give an overview of the field of protein interaction research, describing known structural properties of interactions, followed by an overview of the most important experimental techniques used to infer protein interactions. I will then discuss several previous finding relating to networks of protein interactions, before introducing the Pfam and iPfam databases.

1.1 Protein Interactions

The combination of protein subunits into large multimeric complexes was first described by Theodor Svedberg in 1929 (Svedberg, 1929). He observed that in a density ultracentrifuge, large proteins would separate into subunits of smaller molecular weight. His findings did not meet a wider audience until, 30 years later, Gerhart *et al.* first described allosteric regulation between proteins (Gerhart and Schachman, 1965; Gerhart and Pardee, 1962). This discovery revealed the importance of interactions between proteins and spawned a multitude of investigations into the quarternary structure of proteins. In their excellent review, Klotz *et al.* (1970) outline the importance of subunit stoichiometry, geometry, energetics and cooperativity for the function of protein complexes.

Quarternary structure Figure 1.1 shows the structure of the bacterial HslUV protein. On different levels of granularity, this complex can be described by merely listing the composition of subunits, reflecting stoichiometry. On this level, we can distinguish between homo- and heteromeric complexes as well as combinations thereof. The

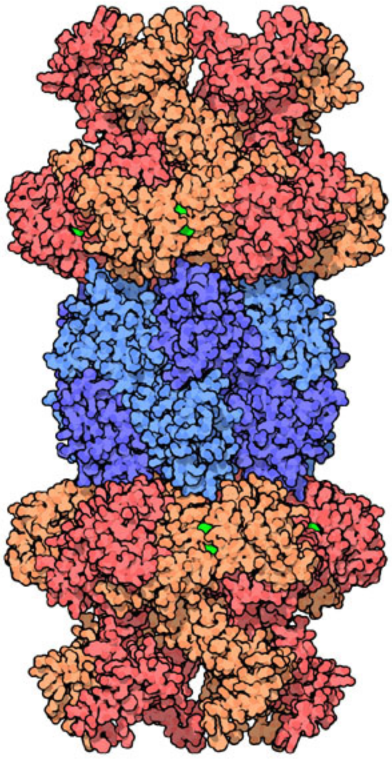


Figure 1.1: Structure of bacterial AAA+ Protease (PDB 1yyf). This chaperone consists of three homo-oligomeric subcomplexes which form a hetero-oligomeric complex. Illustration taken from the "PDB molecule of the month", courtesy of David S. Goodsell: `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb80_1.html`.

structure in Figure 1.1 for example is composed of two homo-oligomeric components of hslU and one homo-oligomeric hslV protease, which assemble into a hetero-oligomeric complex. Several technological advances, reviewed in brief further below, have greatly accelerated the detection of interactions between proteins without requiring crystal structures. However, these methods cannot determine the molecular details of the interaction, such as the region of the protein which contains the binding site or even the exact atoms which mediate the contact between the bound proteins.

Interaction interfaces Beyond stoichiometry, it is important to identify the interfaces through which the individual subunits of a protein interact. This information can usually only be acquired by crystallography or, in some cases, by nuclear magnetic resonance imaging (NMR), and is therefore only available for a small number of complexes. Even more difficult to elucidate are mechanisms of information transfer between protein subunits. Thus, it is often not clear how the stoichiometry and geometry contribute to the function of the complex as a whole.

Duration of interaction Finally, it is important to differentiate between protein complexes which are permanent, or even necessary for the correct folding of the subunit proteins (*obligate* complexes) and interactions which only occur under certain physiological conditions and are usually time-limited (*transient* interactions). The complex shown in Figure 1.1 is obligate, *i.e.* it stays permanently assembled, whereas Figure 1.2 shows the G-protein coupled receptor signalling cascade where information is transmitted between proteins through transient interactions.

Properties of binding interfaces A range of investigations have attempted to describe the properties of interaction interfaces in terms of geometry and residue composition. In their comprehensive review, Jones and Thornton (1996) noted that interfaces of both homo- and heteromeric complexes vary substantially in size and shape. They

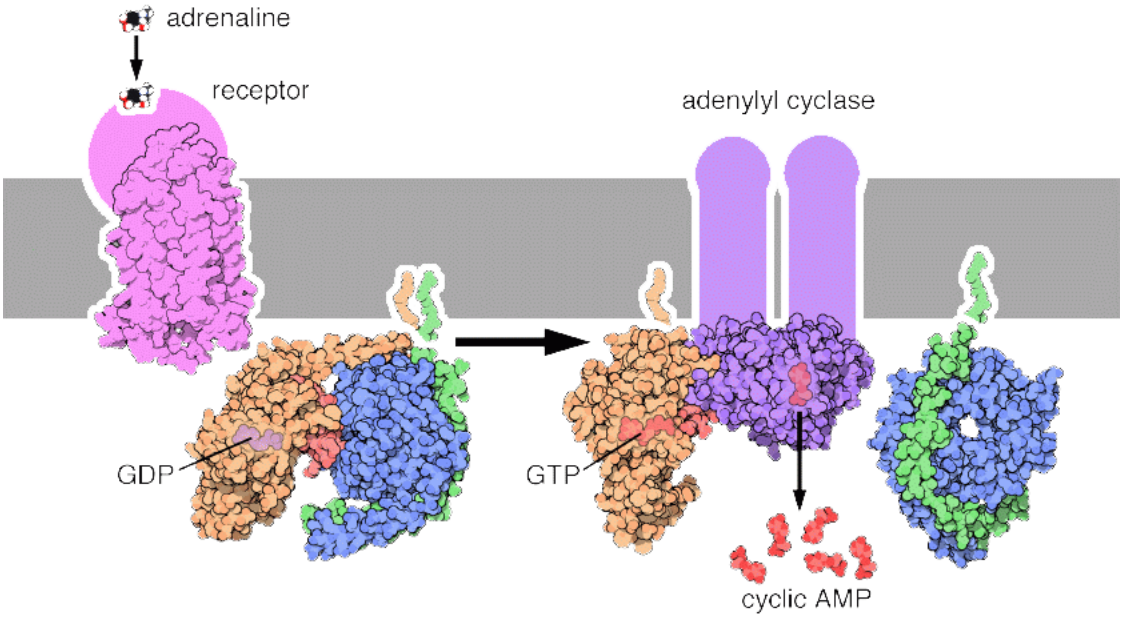


Figure 1.2: Schematic view of the G-Protein coupled receptor signalling pathway. Illustration taken from the “PDB molecule of the month”, courtesy of David S. Goodsell: `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb58_2.html`. Structures in this picture were taken from PDB entries 1f88, 1got, 1cul and 1tbg. Colour-filled areas denote regions for which no structure is available.

also found that large hydrophobic and uncharged polar residues were more frequent in the interfaces compared to the rest of the surface. It has furthermore been established that transient interactions generally employ smaller interfaces compared to obligate interactions (Janin *et al.*, 2007).

Another important discovery regarding protein interaction interfaces was the existence of so-called *hot-spots* within the interface which contribute over-proportionally to the free energy upon binding (Cunningham and Wells, 1989). Measuring the individual contribution of a residue to the overall binding energy through targeted mutagenesis is a laborious process. Thorn and Bogan (2001) have created a repository for the results of such *alanine-scanning* experiments called ASEdb which I will describe in more detail later in this thesis. However, even though progress has been made, the current knowledge about protein interfaces is not sufficient to reliably predict the position of such interfaces in monomeric structures, let alone from sequence alone.

1.1.1 Methods to detect protein interactions

There have been several attempts to identify all interactions between all proteins in an organism by means of automated high-throughput approaches. Two techniques have proven most suitable for this purpose: *Affinity Purification* and *Yeast-Two-Hybrid*. Each of these methods has its own advantages and drawbacks, which have to be taken into consideration when handling the resulting data. It is therefore instructive to review the fundamental principles of the most common techniques.

1.1.1.1 Affinity purification based methods

Several methods for the detection of protein interactions are based on *affinity purification* (AP) (Berggård *et al.*, 2007). In all AP methods, a bait protein is fused to a retrievable tag. The tag should be alien to the host cell into which the construct is transfected, and not interfere with the function of the tagged protein. The cells are eventually lysed and the tagged protein is retrieved using column chromatography against the tag. Interactors bound to the bait protein will be eluted with the bait. After washing, all purified components are identified by *e.g.* mass-spectrometry.

Figure 1.3 outlines the popular Tandem-Affinity-Purification (TAP)-tagging method (Rigaut *et al.*, 1999). In this protocol, the bait protein is fused to a construct of two affinity tags, spaced by a short sequence that can be cleaved by tobacco etch virus (TEV) protease. The TEV protease recognition sequence is very rare in mammalian cells, which minimises the risk of cleaving the bait or a target protein. The advantage of TAP-tagging is the use of two subsequent chromatography steps which substantially reduces the false positive rate. After expression of the bait-tag construct in a suitable cell line, the bait will associate with its target proteins in the cell. After lysis, the first chromatography extracts the entire bait-target complex *via* the first part of the construct, *e.g.* Protein A. After rinsing, TEV protease is added to release the bait-target complex from the beads. In a subsequent purification step, the second part of the construct, commonly calmodulin binding peptide, is recognised by calmodulin-coated beads. After elution, the components bound to the bait protein are usually identified *via* mass-spectrometry. The combination of two purification steps greatly reduces the number of false-positive results, at the slight expense of sensitivity. Weak transient interactions and interactions involving low abundance proteins are particularly prone to be lost during the consecutive washes. Therefore, new techniques have been devised which improve the sensitivity and concentration requirements of AP methods in mammalian cells (GS-TAP, strep-tag III and others) (Burckstummer *et al.*, 2006; Junttila *et al.*, 2005).

AP methods can be sensitive and specific and provide a robust system to detect protein interactions. Nevertheless, there are a number of inherent problems with certain types of interactions (Berggård *et al.*, 2007). Firstly, weak and transient interactions with low binding affinity are prone to be lost during the washing stages. Therefore, AP

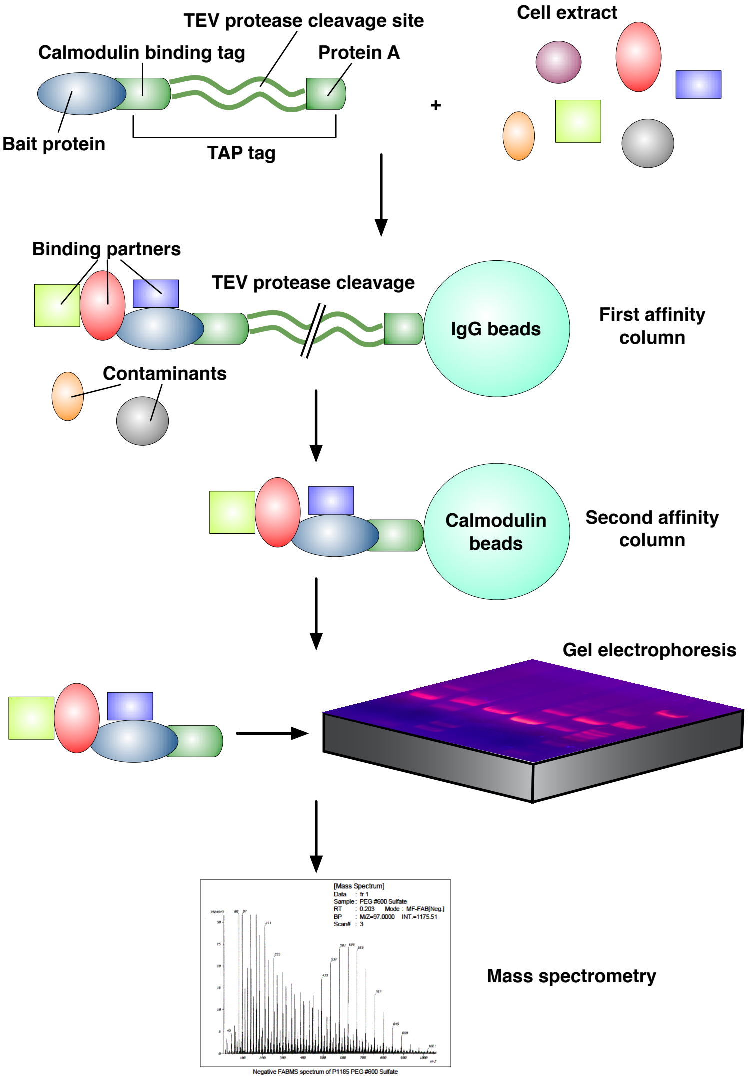


Figure 1.3: Tandem affinity purification with mass spectrometry: A bait protein is fused to calmodulin binding protein, which is in turn connected to a protein anchor (originally *Staphylococcus aureus* Protein A) with a TEV cleavable linker. Complex formation occurs *in vivo*. The first purification step involves a column of IgG beads against the protein A anchor. Subsequently, the protein anchor is removed by TEV protease cleavage and the bait-target complex is recovered in a second column of calmodulin beads. Identification of complex components is performed *via* mass spectrometry, after fractions were separated with electrophoresis. Illustration adapted from Huber (2003)

methods are biased towards stable, high-affinity interactions. Secondly, AP methods are biased towards proteins with high abundance. This is mainly a result of the detection stage: low concentrations of a protein are likely to be missed in the electrophoresis step, and might not yield enough peptide to be confidently detected with a mass spectrometer. Other issues can also arise by introducing a foreign peptide into the host cell, as well as through unwanted interactions between the bait protein and the tag.

1.1.1.2 The yeast-two-hybrid approach

The yeast-two-hybrid analysis was first described by Fields and Song (1989). It has since become one of the most widely used methods to detect protein interactions. Due to its simplicity and cost-effectiveness, it was also the method of choice for the first whole-genome interaction assays.

The method is based on the fact that some transcription factors, such as the yeast enhancer *Gal4*, are composed of two independent domains: a promoter domain, which binds a promoter region upstream of the transcription start site, and a separate activator domain which is required for the assembly of the transcriptional machinery. Neither of the two domains can act independently, as the activator domain needs to be directed to the correct transcription site by the promoter domain. Therefore, transcription of the downstream gene is disrupted if the two domains are physically separated.

Figure 1.4 shows an outline of the yeast-two-hybrid method. The promoter domain (BD) and activator domain (AD) are separated into two plasmids and each fused to a bait and a target protein, respectively. In case the bait and target proteins interact, the BD and AD domain are brought into sufficient spacial proximity to initiate transcription of the reporter gene. Initially, *lacZ* was used as a reporter, but today nutritional selectors such as *HIS3* are often used because they accelerate the screening of large libraries on fewer plates (Bartel and Fields, 1997).

Intuitively, the Y2H method was first applied to study interactions between yeast

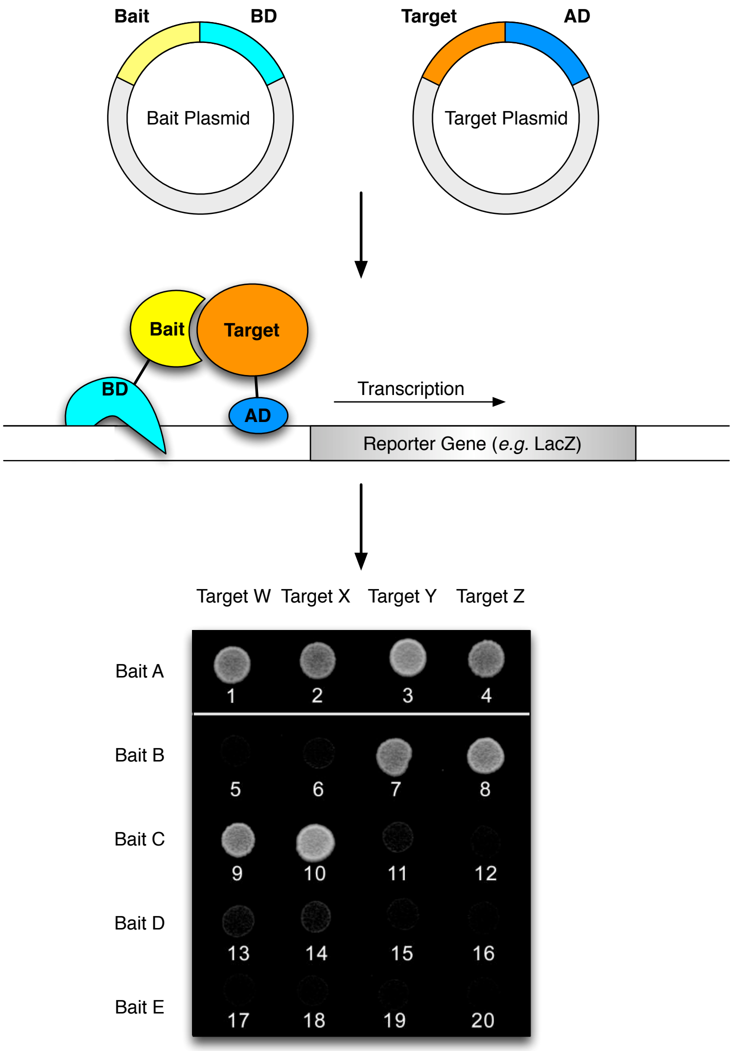


Figure 1.4: Schematic outline of Yeast-two-Hybrid analysis. Two proteins (bait and target) are fused to two separated components of a *S. cerevisiae* transcription factor, *e.g. Gal4*. Both components, the activator domain (AD) as well as the promoter domain (BD) are required in close spacial proximity to activate transcription of the reporter gene. When a library of target vectors is screened against a collection of baits, a matrix is derived where the presence of colonies denotes the successful binding of bait and target.

proteins. However, the system can also be applied to identify interactions between proteins of other species. Viral and prokaryotic genes are more easily cloned and inserted into the yeast system. For higher eukaryotes, un-spliced open-reading frames (ORFs) are required to generate the hybrid constructs. Since large cDNA libraries for several eukaryotic model organisms have been created, it is possible to use recombination cloning technology to create the required hybrid constructs for Y2H screening (Koegl and Uetz, 2007).

The Y2H system allows detection of interactions at lower concentrations than AP. Another advantage (as well as a disadvantage) of the system is that it resolves binary interactions. On the one hand, this allows the exact identification of physical interactors, but on the other hand renders it difficult to define which proteins belong to complexes. On the downside, the Y2H system cannot deal with proteins which require post-translational modifications, or interactions which depend on certain host-specific physiological conditions. This is the case, for example, with extracellular proteins or integral membrane proteins, both of which will not fold correctly in the yeast nucleus. Some proteins, such as active tyrosine kinases, can actually be toxic to yeast if expressed at too high concentrations, and are therefore unsuitable to be used as baits (Berggård *et al.*, 2007).

1.1.1.3 Literature Curation

Scanning the existing literature for reports of interactions between proteins is not, in a literal sense, a method to detect protein interactions. Nevertheless, a large fraction of the known protein interaction networks have been extracted from thousands of individual publications, rather than being identified by high-throughput methods. Literature curation has the advantage that obvious annotation errors can be detected and removed by human curators. Furthermore, a number of literature curation efforts are based on publications which are focused on a small number of genes and as such are likely to adhere to higher standards of positive and negative controls than high-throughput methods can do (Mewes *et al.*, 2008; Reguly *et al.*, 2006). As a consequence, curated protein interaction datasets are generally thought to be more reliable than data from single high-throughput experiments. This increase in quality requires a large number of human annotators and is therefore slow and costly. Furthermore, human annotators will almost inevitably introduce a bias, depending on their understanding of the subject matter. Several groups[1] have tried to address these issues by

- distributing the annotation of new publications between different groups to reduce redundancy
- agreeing to strict guidelines for annotators in order to harmonise rules for acceptance of identified interactions.

To my knowledge, there has been no comparative assessment of the quality of literature curated data, so the reputation of literature-curated data to be a “gold-standard” for protein-interaction data cannot be verified. However, in this thesis I do follow the notion that literature curated data is of high quality and contains few false positive interactions.

1.1.1.4 X-ray crystallography

The determination of protein structure has a long history, dating back to the pioneering work of Kendrew and Perutz in the 1950s and 60s (Kendrew *et al.*, 1958; Perutz *et al.*, 1960). Since then, more than 50000 structures have been deposited in the Protein Data Bank (PDB) (Kouranov *et al.*, 2006), see Figure 1.5. It cannot be the aim of this section to give a comprehensive overview of the field of structural biology. Rather, I want to introduce basic facts about protein structures of interacting proteins that are relevant to various parts of this thesis.

[1] Currently, the IMEx consortium consists of the IntAct, DIP, MINT, MPact and MatrixDB databases. Details can be found in Section 1.1.3.

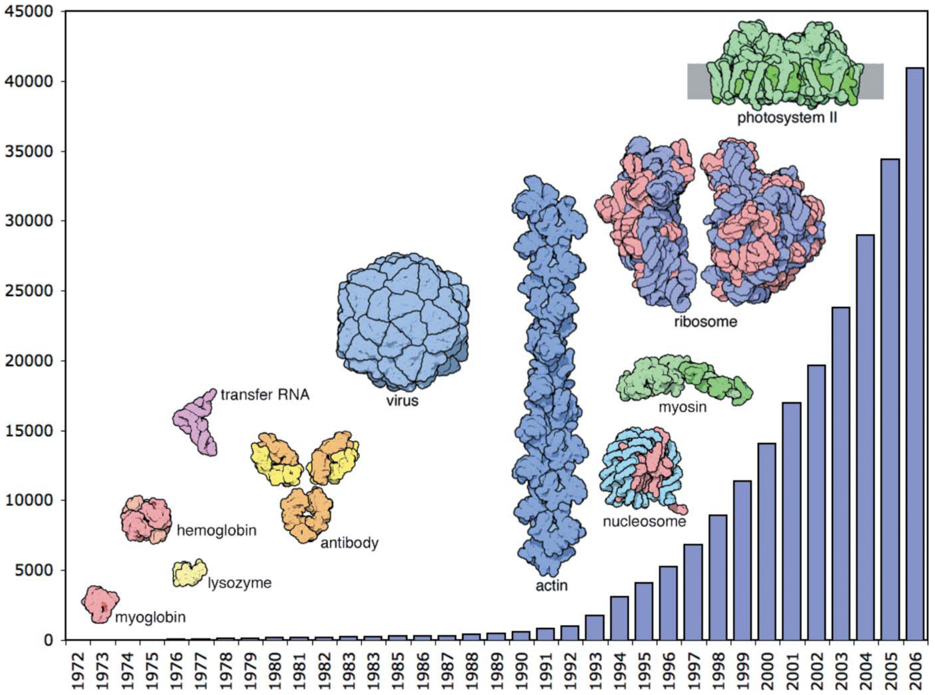


Figure 1.5: Growth of the PDB from its inception in 1972 to 2006. Several landmark structures are shown above the year they were deposited. Figure reproduced with permission from Berman (2008).

X-ray crystallography requires that the protein under investigation can be grown into crystals of sufficient size and purity to diffract X-rays. This is a difficult and time-consuming process which usually requires many attempts to determine the optimal crystallisation conditions. This is the reason why the PDB contains a biased representation of the protein universe: some proteins are significantly easier to crystallise, especially if suitable parameters have already been determined for a similar molecule, whereas other proteins, most notably membrane-associated proteins, are difficult, and sometimes impossible, to grow into a crystal without substantially interfering with their natural structure (Branden and Tooze, 1991).

Once a suitable crystal has been grown, it can be used to create diffraction patterns which are characteristic of the arrangement and properties of the atoms in the structure. Without going into too much detail, it should be noted here that the object of observation in a crystallisation experiment is not necessarily a single molecule, but rather the smallest unit that, when repeated in all three dimensions, forms the crystal. This is called the *asymmetric unit* (ASU) and is a fundamental property of the crystal. The ASU does not necessarily correspond to a biological unit: it might contain a single protein, which is nevertheless biologically able to bind to itself. It can also show two proteins in contact, however the contact is a non-physiological interaction which only occurs under the conditions of crystal formation. The latter case is often referred to as *crystal packing* or *crystal contacts* and is the major potential source of error when inferring protein interactions from crystal structures (Krissinel and Henrick, 2007).

The desired result of a crystallisation experiment is an electron density map which reflects the three-dimensional landscape of the molecule. While the intensities and the diffractions of the X-rays by the crystal can be immediately observed, a third parameter, the *phase* of the rays, is lost in the experiment. However, phase information is needed in order to perform a Fourier-transformation and calculate the electron density map. Several methods exist to infer the phase for larger molecules: Isomorphous replacement, pioneered by Kendrew *et al.* (1958), uses heavy atoms which are introduced into the crystal through soaking as a marker to infer the phase from the differences between the diffraction patterns of the original and multiple “soaked” crystals. Today, the most popular method is multi-wavelength anomalous diffraction (MAD) which requires synchrotron radiation and the presence of metal ions or sulphur atoms which cause anomalous scattering (Jhoti, 2001). If sulphur is not naturally present in the protein, methionine can be replaced by selenomethionine to artificially introduce sulphur atoms into the structure.

After an electron density map has been mathematically derived from the observed diffraction patterns using Fourier transformation, a structure model is fitted into the map. This step usually relies on previous knowledge about the molecule under investigation, such as its amino-acid sequence. Model-building and refinement are not absolutely deterministic steps, so errors can be introduced by the crystallographer, even though nowadays there are many computer programs which attempt to detect badly fitted regions or non-biological arrangements in a structure model (Kleywegt, 2000).

The great utility of protein structures stems from the fact that sequence similarity almost always implies structural similarity. This means that a single structure can provide valuable information not only for the particular protein and species the crystallised proteins were derived from, but also for many other related proteins within the same species and, importantly, also for proteins in other evolutionarily distant species (Chothia and Lesk, 1986). There is now evidence that this conservation of structure also extends to the geometry of binding sites (Aloy *et al.*, 2003). As I will discuss in subsequent chapters, protein structures of molecular complexes therefore provide a template for the mode of interaction of other related proteins.

1.1.1.5 Other methods

AP and Y2H are without doubt the most widely used methods for high-throughput interaction detection. There are, however, a range of other methods which are used either individually on a small scale or in order to validate interactions derived in a high-throughput fashion. These methods encompass co-immunoprecipitation (Markham *et al.*, 2007), protein arrays (MacBeath and Schreiber, 2000), phage display (Sidhu *et al.*, 2003), surface plasmon resonance (Smith and Corn, 2003) and others. Some methods are also specifically designed to deal with certain types of proteins: For example, I was involved in evaluating the performance of a technique specifically targeted towards extracellular interactions which are not typically well detected with other methods (Bushell *et al.*, 2008). Many publications which were collected by literature curation efforts are based on such slower and less easily automated methods.

Furthermore, there are methods that detect *genetic interactions* rather than physical interaction between proteins. A genetic interaction is a functional relationship, stating that two proteins have a combined phenotypic effect (*epistasis*) (Mani *et al.*, 2008). Genetic interaction between proteins can sometimes be detected from indirect evidence, for example correlated gene expression. It is intuitive and could also be shown experimentally that interacting proteins have to be expressed at similar times and appropriate rates in order to be able to interact. Therefore, gene expression profiles derived under different physiological conditions allow the identification of sets of genes whose expression changes are correlated, hinting towards a functional relationship. Similarly, co-localization is a requirement for an interaction to occur, allowing for the verification of a suspected interaction by means of *e.g.* confocal microscopy.

A direct way to detect genetic interactions are so-called *synthetic lethal* screens which have so far been performed systematically in *S. cerevisiae* and *C. elegans* (Lehner *et al.*, 2006; Tong *et al.*, 2004). A synthetic lethal denotes a combined deletion of two genes which is fatal, whereas each individual deletion is viable. Screening genetic interactions with synthetic lethals is a powerful way to identify genes that act in related processes, but it cannot be inferred that they also physically interact.

1.1.2 Error rate and coverage

After the first large automated screens for protein interaction in yeast had been published (Ito *et al.*, 2001; Uetz *et al.*, 2000), criticism was voiced regarding what seemed to be a soaring error rate of the high-throuhput methods (Deane *et al.*, 2002; von Mering *et al.*, 2002; Sprinzak *et al.*, 2003). Some estimates of the false positive rate are as high as 50% for the early Y2H experiments. The error rate of interaction detection methods has since become both a hotly debated issue in the protein interaction community and an intensely investigated area of research.

As a response to the criticism surrounding both AP and Y2H sceens, the methods were improved to include more positive and negative controls as well as repeat experiments in order to reduce noise. In modern screens, the error rate is usually evaluated as part of the experiment and a reliability index is provided with the resulting data. For example, in the yeast proteome survey performed by Gavin *et al.* (2006), the error rate was estimated by repeat experiments and a confidence score for all detected interactions was derived. Similarly, Rual *et al.* (2005) performed a Y2H screen where they tested both reproducibility of the Y2H experiments themselves and the reproducibility of the interactions in a separate AP screen, while also taking into account several other sources of error such as auto-activating constructs.

The other important question that was raised shortly after the first high-throughput experiments were published is: how large are the interactomes of different species? This is relevant because it defines the search space for future experiments. It was noted that many experimental screens for protein interactions show low overlap (von Mering *et al.*, 2002), but without knowledge of the expected size of the interactome, it is impossible to say whether this lack of overlap is due to the vast number of interactions or a result of the large error rate of the experimental method.

Estimates for the size of the interactomes of different species vary substantially. Sprinzak *et al.* (2003) estimated no more than ≈ 16000 interactions make up the entire *S. cerevisiae* interactome. In contrast, Hart *et al.* (2006) predict up to 75500 interactions for *S. cerevisiae*. For human, the numbers range from 154000 to 650000 (Stumpf *et al.*, 2008).

1.1.3 Protein Interaction Databases

The large volume of interaction data generated by high-throughput experiments and literature curation efforts has necessitated the inception of public databases for storage and accessibility. Several groups around the world have created resources for this purpose:

IntAct The interaction database provided by the European Bioinformatics Institute has a broad focus and contains both actively curated data as well as high-throughput datasets. IntAct is not restricted to model organisms but tries to capture all available interaction data. Recently, a small number of negative data have been added to the database (Kerrien *et al.*, 2007).

The BioGRID BioGRID focuses on a selection of model organisms and human. They have performed a thorough manual evaluation of the literature to identify interactions in both budding (*S. cerevisiae*) and fission yeast (*S. pombe*). The data also comprise genetic interactions, *i.e.* interactions inferred from synthetic lethal screens (Breitkreutz *et al.*, 2008).

MPact The MIPS protein interaction resource on yeast is a collection of interactions of high confidence, including the widely used set of complexes usually referred to as the "MIPS complexes". (Mewes *et al.*, 2008).

DIP The Database of Interacting Proteins has been one of the earliest efforts to catalog protein interactions from various sources in a single database. It contains interaction data of varying quality for numerous organisms (Salwinski *et al.*, 2004).

Mint The Molecular INTeraction database, hosted by the University of Rome, focuses on manually searching the scientific literature to find reports of interactions between proteins (Chatr-aryamontri *et al.*, 2007).

HPRD The Human Protein Reference Database aims to collect annotations for all human proteins, including an extensive collection of literature derived interactions (Mishra *et al.*, 2006).

Table 1.1: Overlap between different interaction databases. The numbers in the upper right part of the table denote the number of protein pairs (excluding self-interactions) that are shared between two databases. The lower left part of the matrix lists the fraction of protein pairs of the smaller of the two databases that are shared. The "matrix model" was applied to convert complexes into pairwise interactions. The last row of the table lists the fraction of the respective database that is shared with any other database.

	MPact	**IntAct**	**DIP**	**BioGRID**	**MINT**	**HPRD**	**Total**
MPact		29283	16515	8771	8101	0	51455
IntAct	56.9%		39260	38021	51782	9523	797431
DIP	32.1%	36.6%		24610	22137	316	107396
BioGRID	17.0%	47.5%	30.8%		32113	6194	79999
MINT	15.7%	62.5%	26.7%	40.1%		6708	82800
HPRD	0.0%	24.1%	0.8%	15.7%	17.0%		39545
Total	61.9%	11.9%	45.4%	67.6%	73.2%	41.6%	968084

Table 1.1 lists the size and overlap between the different databases. It clearly shows that no single resource is comprehensive. Even between a small database like MPact and IntAct, the largest resource, there is only a 56.9% overlap (relative to the size of MPact). In the bottom row of Table 1.1, the total fraction of shared interactions is

listed. Again, it emerges that all databases contain a substantial number of unique interactions that are not found in any other database.

In order to gradually overcome these inconsistencies, a number of the listed databases (IntAct, MINT, DIP and MPact) have recently agreed to collaborate in curating and sharing the data. The IMEx initiative (`http://imex.sourceforge.net/`) aims to distribute the curation effort by assigning specific journals to just one group, and then exchange the extracted data. However, at the time of writing, the exchange of records was still in progress and thus incomplete. It is therefore still necessary to merge the data acquired from several databases in order to create the most complete available interaction network for any one species.

1.1.4 Interactomics - The science of networks

The technological advances described in the previous section have resulted in a deluge of molecular interaction information. In the same way that genome-related science was referred to as *genomics*, the term *interactomics* was coined (Sanchez *et al.*, 1999). The *interactome* is the sum of all physical protein interactions in an organism. The first attempts to elucidate the complete interactome of an organism were performed by Uetz *et al.* (2000) and Ito *et al.* (2001). Using a systematic, automated Y2H approach, they were able to identify several thousand protein interactions in *S. cerevisiae.*

As more and more interaction network information became available, the structure and global properties of these networks became the subject of great interest. Barabasi and Albert (1999) suggested that a wide variety of systems, from social interactions to the world-wide web, had similar topological properties and were governed by the same principles. It was observed that most nodes are only sparsely connected, while a small number of nodes accumulates the majority of connections (often called *hub* proteins). This so-called "scale-free" distribution of edges per node (the *degree distribution*) follows a Power law of the form $P(k) \sim c \cdot k^{-\gamma}$, where c and γ are constants.

The “Power-law” and “scale-free network” concepts attracted a lot of interest by the scientific community (Luscombe *et al.*, 2002), because they were thought to lead to several corollaries. It was noted that the overall low number of connections per node leads to greater robustness towards random node deletions (Albert and Barabási, 2002). Robustness in this context is defined as the impact of node deletions on the connectedness of the network. The other important inference that was made from the network topology concerns the mechanism by which the network evolved. Power laws are thought to emerge through a process called *preferential attachment*, whereby whenever a node is added to the network, it is likely to connect to a node that already has many connections. Translated into biology, preferential attachment was argued to be a result of evolution through gene duplication. Under the assumption that there is no bias as to which gene is duplicated and the rate of gene loss is low, older genes will gradually accumulate connections. Karev *et al.* (2002) extended this concept and described how a simple model of domain duplication, loss and *de-novo* creation can explain the observed size distribution of protein domain families. They argue that the same model should also be applicable to other evolving networks.

Jeong *et al.* (2001) applied the principles of network analysis to protein interaction networks. They did not only show that the yeast interactome, to the extent it was available at the time, is a scale free network, they also claimed that there is a correlation between the degree of a protein and its essentiality. This was remarkable as it seemed to prove that the network-theoretical concept of robustness could be extrapolated to biological systems. Subsequently, it was also claimed that the principle of preferential attachment underlies the evolution of protein interaction networks (Barabasi and Oltvai, 2004; Eisenberg and Levanon, 2003).

The interpretation of protein interaction networks under the paradigm of scale-free networks has since attracted criticism. It was shown by Khanin and Wit (2006) that other distributions than power-laws better fit the observed degree distributions in various protein interaction and metabolic networks. It is also important to consider that the available protein interaction data is just a sampling from the actual biological network. Stumpf *et al.* (2005) and Han *et al.* (2005) showed both theoretically and by examples that subnetworks sampled from a larger scale-free network are not themselves scale free, and that the degree distribution of a sampled subnetwork does not reliably predict the distribution of the global network. The real mechanisms by which interaction networks have evolved are thus still not satisfactorily explained.

1.2 Genetic variation

A simple but fundamental principle of Darwin's theory of natural selection is that there is no evolution without variation. In the plant and animal kingdom, such variation can be observed in abundance. Darwin himself was inspired by the variability in birds that he witnessed during his journey on board H.M.S. Beagle (Darwin, 1859). Similarly, differences in shape and colour of flowers and seeds of pea plants lead Mendel to deduce the first systematic description of a link between observable phenotypes and a then-unknown genetic substance that induces such phenotypes (Mendel, 1865). Today, we know that the main carrier of genetic information is DNA. The consequential next questions are: what are the sources of variation, and how is phenotypic diversity related to genetic variation?

1.2.1 Types and causes of mutations

In sexually reproducing organisms, individuals carry two versions of the genetic information that is passed on from the parent generation[1], each version called an *allele*, grouped together on two homologous chromosomes. Variation between individuals is to a large degree the result of the combinatorial shuffling of alleles, where for every

[1] Notwithstanding exceptions such as *e.g.* sex chromosomes or mitochondrial DNA, where only one copy is inherited from one parent.

corresponding gene there are four possible allele pairs an individual can inherit. This alone does not explain the existence of differing alleles itself. Variation between alleles is a result of *mutations* that change their genetic sequence. There are four broad types of mutations: Point mutations, insertions/deletions, translocations and inversions[1]. In this thesis, I consider only the first two types of mutations.

For each type of mutation, there can be numerous causes. Point mutations are the most frequent mutation event to occur. They are randomly introduced in the genetic code mostly *via* mistakes during replication and as a result of mutagens. It is often assumed that point mutations occur by chance with a constant frequency uniformly across the genome, which makes it possible to use the mutation rate as a kind of molecular clock (Zuckerkandl and Pauling, 1962).

Not all mutations lead to a phenotypic effect. This is partly a result of the fact that the majority of eukaryotic genomes are composed of long regions of non-coding DNA which is insensitive to mutations. Furthermore, even point mutations inside coding regions do not necessarily alter the encoded protein. The genetic code is degenerate, *i.e.* some nucleotide changes will not affect the encoded protein sequence because there are multiple codons encoding for the same amino-acid. This redundancy in the genetic code can be used to quantify the selective pressure on a gene. This is done by calculating the ratio of active (non-synonymous) to silent mutations (synonymous mutations) for a gene, where a mutation is defined by comparing the DNA sequence to the sequence of an orthologous gene from another species (Kafatos *et al.*, 1977). The resulting measure is referred to as the *dN/dS ratio*[2]. dN/dS values below 1 indicate negative selection, whereas values above 1 are taken as a sign of positive selection (Hughes and Nei, 1988).

Apart from point mutations, larger chromosomal rearrangements can be caused by errors during *homologous recombination*. Usually, homologous recombination is a

[1] For the sake of simplicity, I subsume chromosomal deletions and duplications into the "insertion/deletion" category.

[2] Sometimes also denoted as K_a/K_s ratio.

controlled process which allows the swapping of genetic information between the two homologous chromosomes during meiosis. However, there are numerous errors that can occur. Most notably, non-allelic homologous recombination is a process in which recombination occurs not between the corresponding allelic regions on the chromosomes, but between homologous regions within the same chromosome, causing a deletion. Such regions can be low copy repeats (LCRs) or segmental duplications. Beyond that, there are numerous other less frequent causes of mutations such as viruses or transposable elements, *e.g.* Alu repeats, which can cause insertions, deletions and other genomic rearrangements (Batzer and Deininger, 2002).

1.2.2 Human variation

H. sapiens is subject to mutations, natural selection and thus evolution the same as any other species. However, history has shown that this fact is easily misinterpreted or even deliberately misused to justify arbitrary discrimination[1] . It is for these ethical reasons that it is difficult to discuss variation in humans in quite the same way as we discuss variations in animals: concepts such as race or ethnicity predate modern population genetics and are as such hard to define for a scientific purpose (Feldman *et al.*, 2003; Sankar and Cho, 2002). In fact, it has been suggested that variation on the DNA level is larger amongst individuals thought to belong to the same "race" as between different "races" (Barbujani *et al.*, 1997; Disotell, 2000). The sequencing of genomes of individuals which is currently underway (Siva, 2008) will hopefully shed new light on the question whether "race" has a clearly detectable genetic footprint or whether we have to redefine our concepts of "race". For the remainder of this thesis, I will try to focus not on differences between populations but on differences between individuals.

[1] As an example, I refer to the insightful documentary on biology and medicine in fascist Germany provided by the United States Holocaust Memorial Museum: `http://www.ushmm.org/museum/exhibit/online/deadlymedicine/`

1.2.3 Variation in healthy individuals

One of the first types of human variation that were used to study genetics in entire populations were the blood groups. Since Landsteiner's initial description of the AB0 system at the beginning of the 20th century, numerous other blood type systems have been defined. The key property of blood types is that they constitute distinct classes with a simple Mendelian pattern of inheritance, hence they must be determined by individual genetic loci. In the 1950s and 60s, studies on haemoglobin variants offered a first glimpse at the molecular mechanisms as well as the distribution of genetic variation in humans (Boyd, 1963; Livingstone, 1958). Together, these data allowed a first assessment of genetic diversity between individuals and populations (Lewontin, 1972).

DNA technology has since greatly accelerated the identification of genomic variants. The human genome is now known to contain millions of single-nucleotide polymorphisms (SNPs). Understanding the distribution, frequency and linkage between these variants holds great promise for the analysis of human evolution as well as for the understanding of complex diseases. Therefore, a concerted effort was undertaken to identify up to one million *tagSNPs* across the entire human genome of individuals of European, Asian and African descent (The International HapMap Consortium, 2003). The key property of tagSNPs is that they occur at a frequency of $> 0.1\%$ in the population and they are linked to a haplotype block, *i.e.* a region of the chromosome which is relatively stable to recombination.

Recently, it has also been discovered that there are frequent insertion and deletion polymorphisms, so-called copy-number variations (CNVs) that are abundant in the human genome. They are defined as regions of $>$ 1kb which are deleted or duplicated in the genome of an individual (Freeman *et al.*, 2006). They seem to be closely related to *segmental duplications*, *i.e.* regions larger than 1kb and $> 90\%$ sequence identity which occur multiple times in the genome. The main distinction between CNVs and segmental duplications is that a region which is duplicated in all members of a population is called a segmental duplications but not a CNV. There have been numerous reports of CNVs in individuals sampled from different populations (Conrad *et al.*, 2006; Iafrate *et al.*, 2004; Redon *et al.*, 2006; Sebat *et al.*, 2004). Interestingly, these initial results were derived from seemingly healthy individuals, even though many CNVs seem to overlap protein coding genes. This indicates that many genes are robust against changes in copy number. In Chapter 4, I will discuss the issue of dosage sensitivity in the context of protein interactions in more detail.

Many studies regarding CNVs were performed using a technique called array-based comparative genomic hybridisation (array CGH) (Shinawi and Cheung, 2008). Samples of genomic DNA of two individuals, one reference and one target, are labelled with different fluorescent dyes. Upon hybridisation to an array containing > 25000 large insert clones reflecting most of the human genome as probes, regions with uneven hybridisation can be detected by the shift in colour. The start and end position of putative CNVs are then calculated from the overlaps between the clones. Given the length of the clones ($\approx$ 200kb), the resolution of the CNV coordinates is coarse, but new methodologies with substantially higher resolution are currently being developed.

1.2.3.1 Genetic diseases

Another form of variation that has been studied extensively are genetic diseases. A wealth of investigations have been undertaken to identify loci responsible for Mendelian diseases. Botstein and Risch (2003) give an insightful historical perspective into the development of the field. Since the late 1980s, the prevalent method to identify genes responsible for a disease phenotype has been *positional cloning*, preceded by linkage analysis of affected individuals and their families. This approach works best if the phenotype is unambiguous and the genotype-to-phenotype relationship is simple. Before a physical map of the human genome was available, positional cloning relied on the genetic map, often using polymorphic repeats as a marker. The effectiveness of this method is evident from the fact that almost all known Mendelian disease loci were mapped in this way.

Today, the Online Mendelian Inheritance In Man (OMIM) database (Hamosh *et al.*, 2005) contains over 14000 disease associated genetic variants in more than 1800 genes. Studying these variants, it could be shown that genes carrying dominant mutations are slower evolving than recessive genes (Blekhman *et al.*, 2008). Interestingly, the same study also found that only 45% of genes in OMIM carry recessive mutations. According to the classic explanation of dominance provided by Wright (1934), most mutations were expected to be recessive: Wright argued that dominance of the wild type allele is a result of the fact that most metabolic pathways can maintain their function even if one step has reduced capacity. In other words, not all components of a metabolic pathway are rate-limiting steps, hence the pathway is robust against a reduction in the amount of one particular catalyst. However, it is emerging now that this theory does not in the same way apply to proteins other than enzymes. Kondrashov and Koonin (2004) described that recessive mutations are in fact most common in enzymes, but mutations in transcription factors or structural proteins are more often dominant. This shows that the genetics of diseases and their underlying molecular mechanisms are tightly linked. Currently, there are few mechanistic explanations for the disease-causing effects of the majority of mutations. Identifying such molecular mechanisms hence presents an interesting field for further development.

This becomes even more striking if one considers that Mendelian diseases only reflect a subset of human genetic disorders. Many disease, from diabetes over schizophrenia to susceptibility to infectious diseases such as tuberculosis, have been shown to have a genetic component, however unlike Mendelian diseases, the contribution of individual loci is small, *i.e.* an unknown number of individual mutations contribute to the disease. Genome-wide association studies have been used to identify such loci which are significantly but weakly associated with a disease (Risch and Merikangas, 1996). In such a study, large cohorts of case and control individuals are tested for the presence of one or several diseases, before each individual is genotyped. Recent studies used array-based methods to query known SNPs along the entire genome (Wellcome Trust Case Control Consortium, 2007). In the future, it will likely be possible to re-sequence entire genomes in order to detect all sequence variants. Finally, statistical analyses of the data provide putative associations between certain SNPs and the disease status of an individual. The problem is that the identified SNPs only point towards genes that are likely to be relevant for a disease, however little is known about the mechanism by which a polymorphism induces disease susceptibility. In such cases, using information on biochemical pathways and protein interactions can help to uncover connections between target genes or provide a ranking which SNPs are most worthwhile to be studied in more depth.

1.3 Protein Domains and the Pfam database

In structural biology, it has long been known that proteins are to a large extent composed of conserved modular building blocks commonly called *domains*. It was also quickly noted that structures with even just remotely related sequences usually shared stronger structural similarity (Chothia, 1992). As a consequence, methods for detecting remote sequence homology were being developed. Initially, most methods employed scoring functions that incorporated manually defined weights, in an attempt to capture "expert knowledge" about a particular family of proteins.

A major leap towards a more generalised concept of homology detection was the use of a probabilistic framework called *Hidden Markov Models* (HMMs) (Krogh *et al.*, 1994). HMMs are a way to model stochastic processes. Their great advantage is the fact that efficient algorithms exist to calculate the probability that an observed phenomenon was produced by the stochastic model. In the case of sequence homology, the model describes the composition of the representative parts of a sequence family. A hypothesis test can then be performed on a query sequence, comparing the chance that the query was created by the predefined model. The model itself does not have to be manually created, but can be automatically generated from a multiple sequence alignment containing typical members of the family. This short description cannot do justice to the complexity and power of HMMs and their applications. More detail can however be found elsewhere (Durbin *et al.*, 1998; Schuster-Böckler and Bateman, 2007a).

One of the key features of HMMs is that *any* sequence family is modelled using a *common framework*. It is hence possible to create a collection of many sequence families and search a new sequence against a range of such family descriptions in order to identify putative evolutionary relationships. The Pfam database (Finn *et al.*, 2008) is one of the largest resources for domain annotation. In the Pfam terminology, a *domain* denotes any conserved sequence region, rather than just referring to an independent structural element in a protein. The Pfam database today contains over 10000 protein families and is still constantly growing (Sammut *et al.*, 2008). For every release, the entire UniProt database (Wu *et al.*, 2006) is searched for occurrences of any domain in Pfam. The Pfam database to date covers $\approx$ 75% of all sequences, *i.e.* 75% of all sequences in UniProt contain at least one region that matches an HMM listed in Pfam. For proteins in the PDB, the coverage is substantially higher (currently $\approx$ 95%).

Thus, by projecting the protein universe, *i.e.* all known protein sequences[1], down to the domain universe, one can achieve a reduction in complexity of several orders of magnitude. At the level of conserved domains, the traces of evolutionary history can be observed more clearly. This has been exploited *e.g.* in inferring the evolutionary history of nematodes with respect to chordates and insects, see Wolf *et al.* (2004). In this thesis, Pfam was used extensively to investigate the function and evolution of

[1] Currently, UniProt contains over 3 million sequences, not including the expected deluge of metagenomics derived sequences

interacting proteins.

1.3.1 *i*Pfam

I have so far described how protein interactions can be identified biochemically as well as by crystallography. I have also introduced the relationship between sequence and structure conservation. As the function of a protein, including its interaction preference, is dependent on its three-dimensional structure, it is an obvious next step to describe the interactions between proteins in terms of conserved sequence regions such as Pfam families. Several recent studies have indeed found that protein domains can mediate protein interactions. There seems to be a limited set of domain interactions that is being reused in proteins of different backgrounds (Aloy and Russell, 2004).

Figure 1.6 shows a typical example of a protein structure of an interacting protein, in this case the *E. coli* Oxidoreductase, where a specific domain mediates the interaction. The asymmetric unit of the structure only contains two of the four subunits that make up the functional macromolecule. The two subunits bind each other through a large interface (shown as a surface representation in the figure) which matches the Pfam family 2-Hacid_dh [Pfam-id: PF00389]. The interface exhibits structural complementarity, thus excluding solvent and creating the necessary binding energy to maintain a stable interaction.

Pfam domains are defined solely through sequence, but a conserved structure is very often associated with them. In order to find structures that match a certain Pfam domain, one could search the raw sequences stored in the PDB entries against the library of Pfam HMMs. However, a complete search of the UniProt database is performed at every release of Pfam. Rather than searching the complete Pfam database again, it is more efficient to map every residue in the PDB structures to a residue in a UniProt sequence. Such a mapping is conveniently provided by the Molecular Structure Database (MSD) at the EBI (Velankar *et al.*, 2005). Identifying regions in

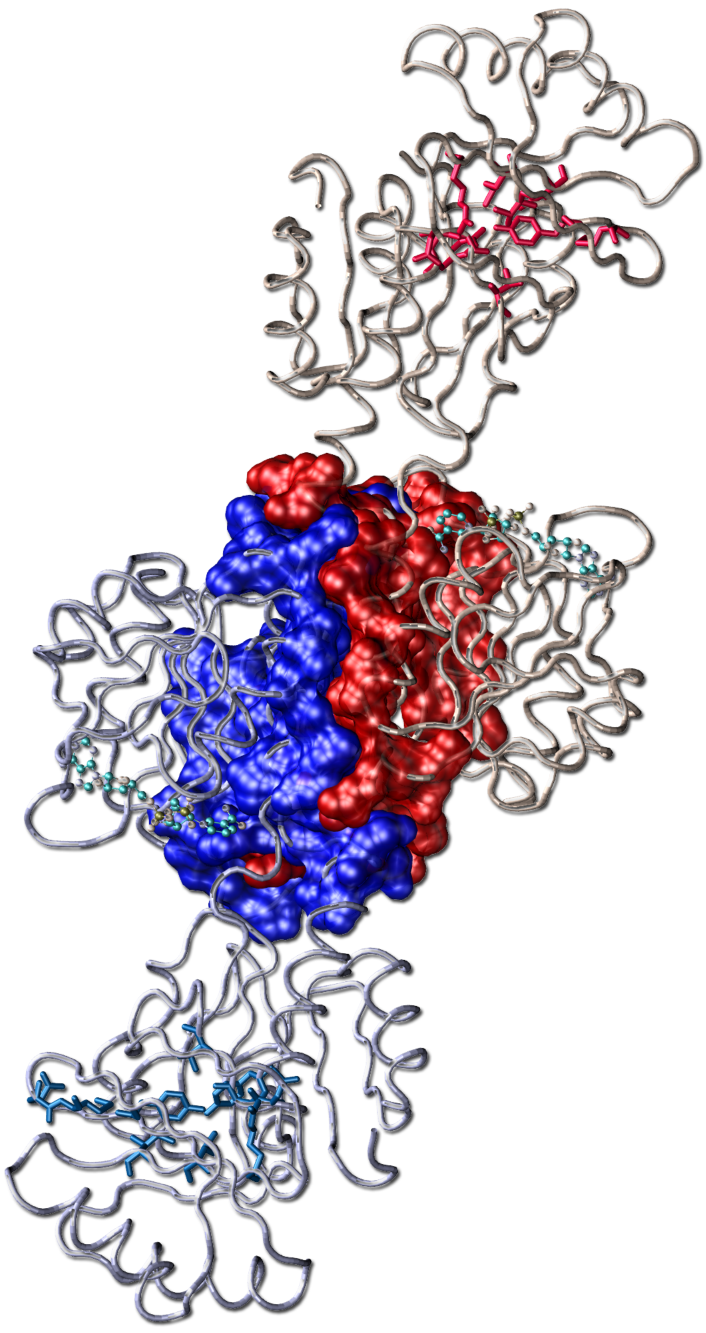
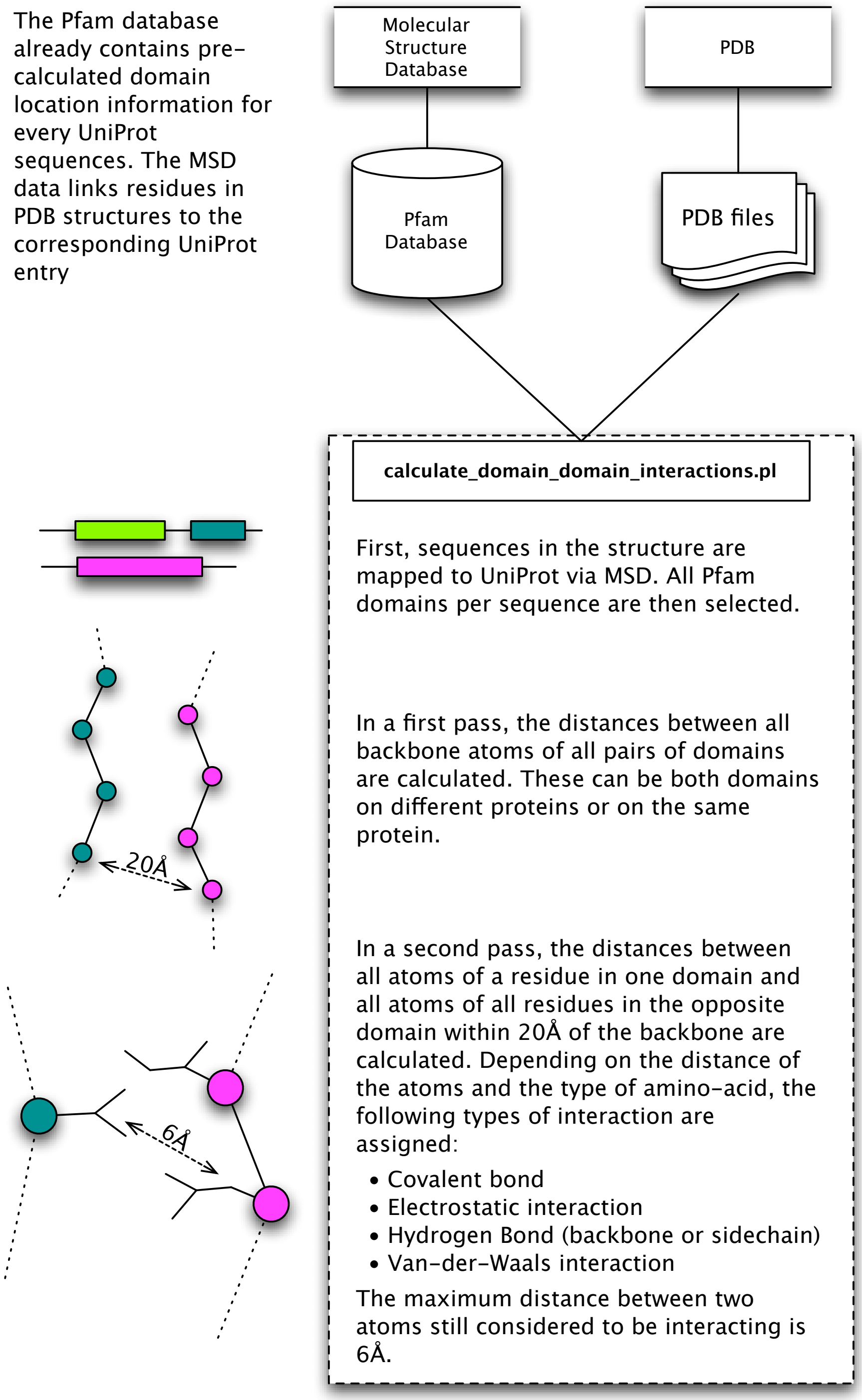


Figure 1.6: Structure of *E. coli* Oxidoreductase dimer [PDB-id: 1psd] with interacting residues as defined in *i*Pfam highlighted. The structure shows the asymmetric unit, the biological molecule is a tetramer, employing additional interaction interfaces which are not identified by *i*Pfam. The interchain interactions between the two distinct subunits are shown as a continuous surface. Intrachain interactions between two distinct domains (ACT interacting with 2-Hacid_dh) of each subunit are shown as sticks.

PDB structures that match a Pfam domain thus becomes a simple database query which joins the two co-ordinate systems.

*i*Pfam is a database of physically interacting protein domains that was derived by gathering all interactions between distinct Pfam domains in asymmetric units as deposited in the PDB (Finn *et al.*, 2005). Figure 1.7 illustrates the steps that comprise the generation of *i*Pfam. For each pair of regions that match a domain within a sequence, it is evaluated whether the backbone atoms are in sufficient proximity ($<$ 20 Å) to each other to allow a contact between the sidechains. This initial filtering step substantially reduces the search space. Subsequently, all atoms in one domain are tested for their exact distance to all other atoms in the adjacent domain. Depending on the observed distance, geometry and type of atoms, a bond type is assigned to the pair. The maximum distance between any two atoms still considered as a contact is 6 Å. There is currently no lower limit to how many atom contacts are required for a domain pair to be recorded. It is also important to note that the version if *i*Pfam used throughout this thesis is based solely on interactions in the asymmetric units of PDB entries. Therefore, interfaces involved in the assembly of large repetitive structures such as virus capsids as well as other interactions between repeated individual units are missing from *i*Pfam.

As illustrated in Figure 1.6, not only interactions between two distinct proteins are considered, but also the residue contacts between two domains within one protein. The rationale behind this is that many domains are structurally independent units which can, over the course of evolution, be combined with other protein sequences. In such cases, an intrachain interface can become a potential new interchain recognition site, as described by Enright *et al.* (1999).


Molecular Structure Database
PDB
Pfam Database
PDB files
calculate_domain_domain_interactions.pl
20Å
6Å

The Pfam database already contains pre-calculated domain location information for every UniProt sequences. The MSD data links residues in PDB structures to the corresponding UniProt entry

First, sequences in the structure are mapped to UniProt via MSD. All Pfam domains per sequence are then selected.

In a first pass, the distances between all backbone atoms of all pairs of domains are calculated. These can be both domains on different proteins or on the same protein.

In a second pass, the distances between all atoms of a residue in one domain and all atoms of all residues in the opposite domain within 20Å of the backbone are calculated. Depending on the distance of the atoms and the type of amino-acid, the following types of interaction are assigned:

- Covalent bond
- Electrostatic interaction
- Hydrogen Bond (backbone or sidechain)
- Van-der-Waals interaction

The maximum distance between two atoms still considered to be interacting is 6Å.

Figure 1.7: Outline of *i*Pfam creation process. Structure data and PDB to UniProt mappings are downloaded from the MSD and PDB, respectively. A single script (calculate_domain_domain_interactions.pl) then performs a sequence of calculations on each structure to identify all atoms in every pair of Pfam domains in the structure that are in contact.

1.4 Outline of this thesis

The remaining chapters of this thesis consist of three separate investigations. I first analyse the coverage of *i*Pfam in order to assess the power of the structural domain annotations to explain existing protein interactions. This also allows me to make inferences on the level of conservation and reusability of domain interactions amongst different proteins and between species. This work lays the foundations for applying domain interaction information to human disease data. In the second chapter, I estimate the impact of protein interaction defects on human genetic diseases and show how the structural information can be practically applied to gain insights into the function of a related protein complex. Finally, I follow up on an interesting observation related to the evolution of protein interactions, namely the tendency of interacting proteins to be more dosage sensitive. I use the newly available human population copy-number variation data to investigate whether protein complexes are under stronger selective pressure to maintain their abundance in the cell.

Parts of the results described in this thesis have been published (Schuster-Böckler and Bateman, 2007b, 2008). The respective articles can be found in the Appendix. In addition to that, I have published a paper on the visualisation of profile–profile comparisons (Schuster-Böckler and Bateman, 2005) which is outside the focus of this thesis. I was also involved in several collaborations which resulted in two publications (Bushell *et al.*, 2008; Finn *et al.*, 2006).