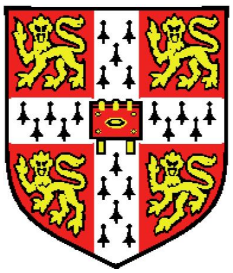
The role of protein interactions in evolution and disease



Benjamin Schuster-Böckler

Selwyn College

University of Cambridge

A dissertation submitted for the degree of

Doctor of Philosophy

September 2008

In loving memory of Margarethe and Erwin Gregor.

"Humility and knowledge are the origins of wisdom."

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between March 2005 and October 2008. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this dissertation nor anything substantially the same has been or is being submitted for any qualification at any other university.

Summary

The network of interactions between proteins is the scaffold that shapes the properties of every living cell. Whether it is enzymatic pathways or cascades of signal transduction, most processes rely on the ability of proteins to recognise and bind each other. New experimental techniques have fuelled interest in these networks, leading to a rapid increase in available data on protein interactions from various species.

In the first part of this thesis, I investigate to what extent networks of protein interactions are mediated by conserved regions in proteins, generally called domains. I make use of a set of domain pairs which have been shown to interact in 3-dimensional structures. By analysing the frequency of co-occurrence of these domain pairs in networks of protein interactions from five different species, I show that some domain pairs form reusable recognition modules, while others are confined to a specific protein pair. Overall, the number of known protein interactions that contain a domain pair with known structure is small. This underlines the necessity to resolve more structures of interacting proteins. Finally, I observe a large overlap in the domain pairs present in different species, suggesting many recognition modules are ancient in origin.

In the second part of my thesis, I combine sequence analysis techniques to investigate the impact of protein interactions on human diseases. I make use of the detailed information provided by 3-dimensional structures to

identify interacting residues within known protein domains. I then use hidden Markov models to search for structurally corresponding residues in proteins that cause genetic diseases. I identify cases where these structurally corresponding residues have been reported to cause Mendelian disorders, such as an Ile to Val substitution in the dimerisation interface of the H-Twist transcription factor leading to Baller-Gerold syndrome. I report 1428 mutations which potentially affect a protein interaction. This corresponds to $\approx 4\%$ of all known single-residue mutations.

I found that mutations in interaction interfaces frequently cause dominant phenotypes. I subsequently discovered that many dosage sensitive genes related to human disease are members of protein complexes. From the analysis of recently published data of gene expression and structural variation between individuals it emerges that members of protein complexes exhibit lower expressional noise than the rest of the genome and that variation of gene copy-number between individuals has a measurable effect on dosage. I show that this effect causes negative selection against large scale copy-number variations in dosage sensitive genes, such as members of protein complexes.

Acknowledgements

First and foremost, I am greatly indebted to my supervisor Alex Bateman for patience, guidance and ongoing support over the last three years. Secondly, I want to thank Robert Finn not only for creating the databases and tools which form the basis of most of this work but also for countless fruitful discussions that greatly helped me to put a sense of direction into this work.

I also wish to thank all current and previous members of the Pfam, Rfam and Merops teams. I am grateful for the many helpful suggestions by my Thesis Committee (Sarah Teichmann and Anton Enright). Many people have contributed in various ways to this work: Donald Conrad, Gavin Wright, Ben Lehner, Manolis Dermitzakis, Matt Hurles, Christine Bird to name just a few. Cara Woodwark, Marija Buljan and Rafael Najmanovich kindly read a first draft of this manuscript and provided many helpful comments.

Finally, I want to thank Roberta for putting up with me, and for supporting me over the years, especially in the last few months.

This work was made possible by generous financial support from the Wellcome Trust.

Contents

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

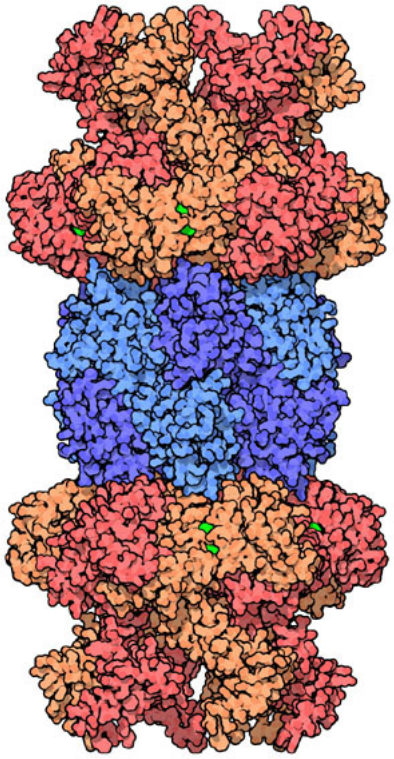


Figure 1.1: Structure of bacterial AAA+ Protease (PDB 1yyf). This chaperone consists of three homo-oligomeric subcomplexes which form a hetero-oligomeric complex. Illustration taken from the "PDB molecule of the month", courtesy of David S. Goodsell: `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb80_1.html`.

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

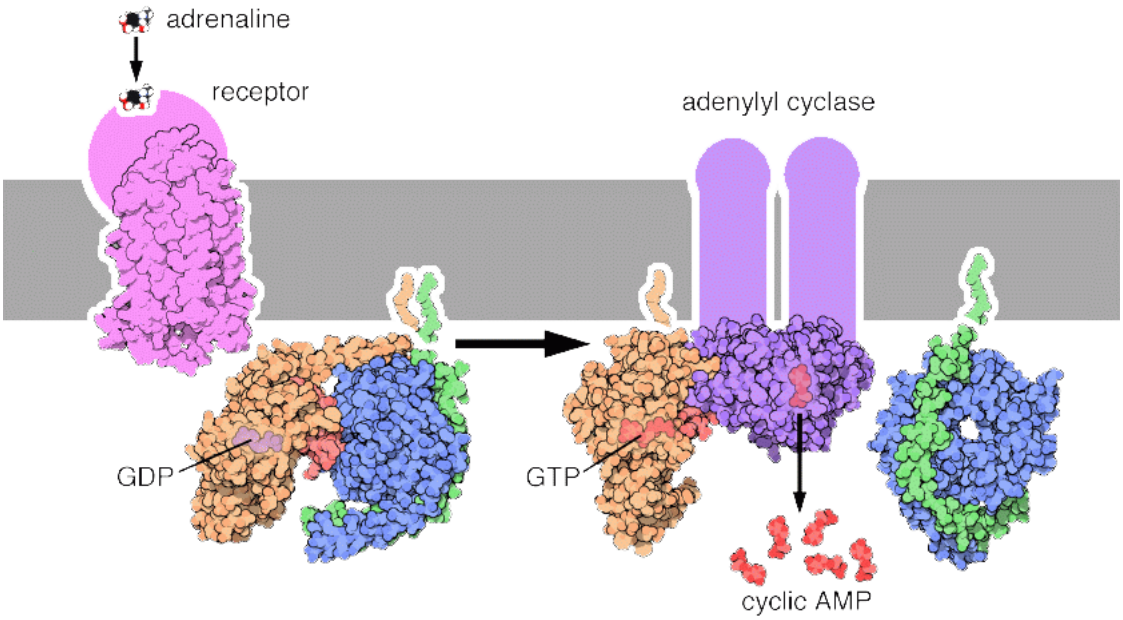


Figure 1.2: Schematic view of the G-Protein coupled receptor signalling pathway. Illustration taken from the "PDB molecule of the month", courtesy of David S. Goodsell: `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/pdb58_2.html`. Structures in this picture were taken 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

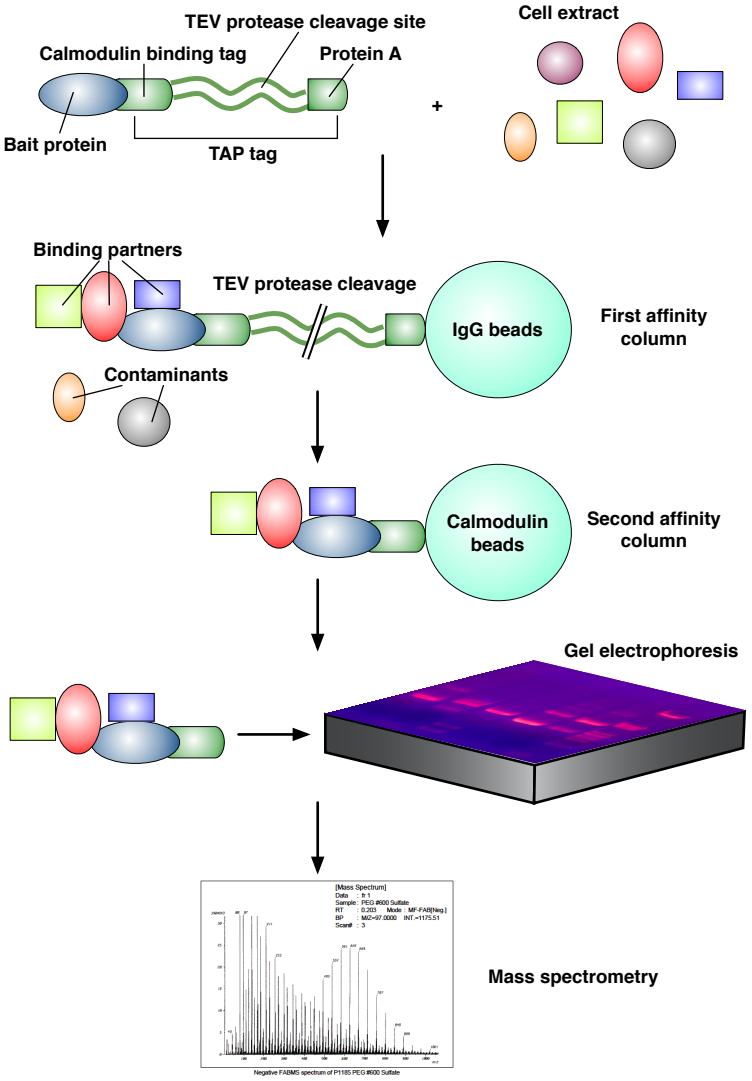


Figure 1.3: Tandem affinity purification with mass spectrometry: A bait protein is fused to calmodulin binding protein, which is in turn connected to a protein anchor (originally *Staphylococcus aureus* Protein A) with a TEV cleavable linker. Complex formation occurs *in vivo*. The first purification 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

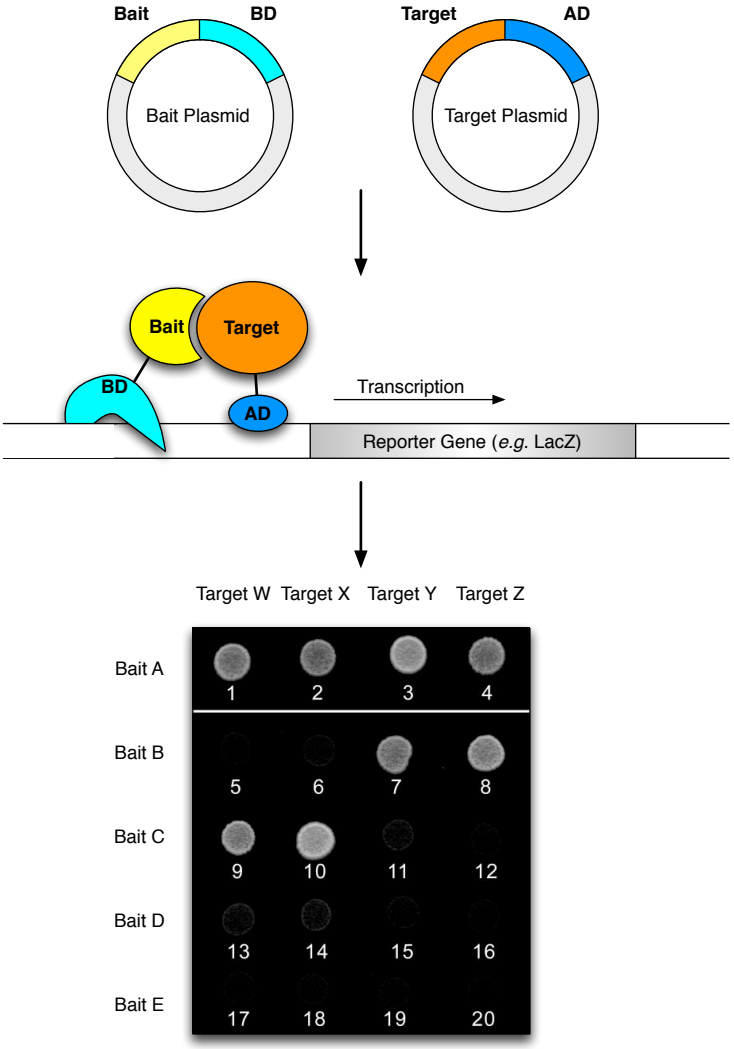


Figure 1.4: Schematic outline of Yeast-two-Hybrid analysis. Two proteins (bait and target) are fused to two separated components of a *S. cerevisiae* transcription factor, *e.g. Gal4*. Both components, the activator domain (AD) as well as the promoter domain (BD) are required in close spacial proximity 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.

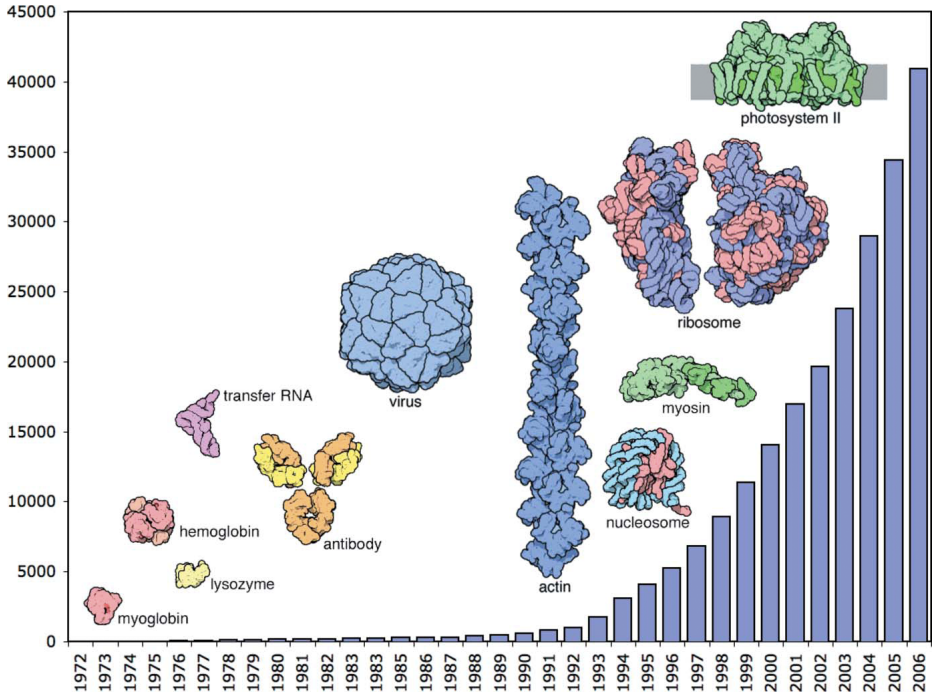


Figure 1.5: Growth of the PDB from its inception in 1972 to 2006. Several landmark structures are shown above the year they were deposited. Figure reproduced with permission from Berman (2008).

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).

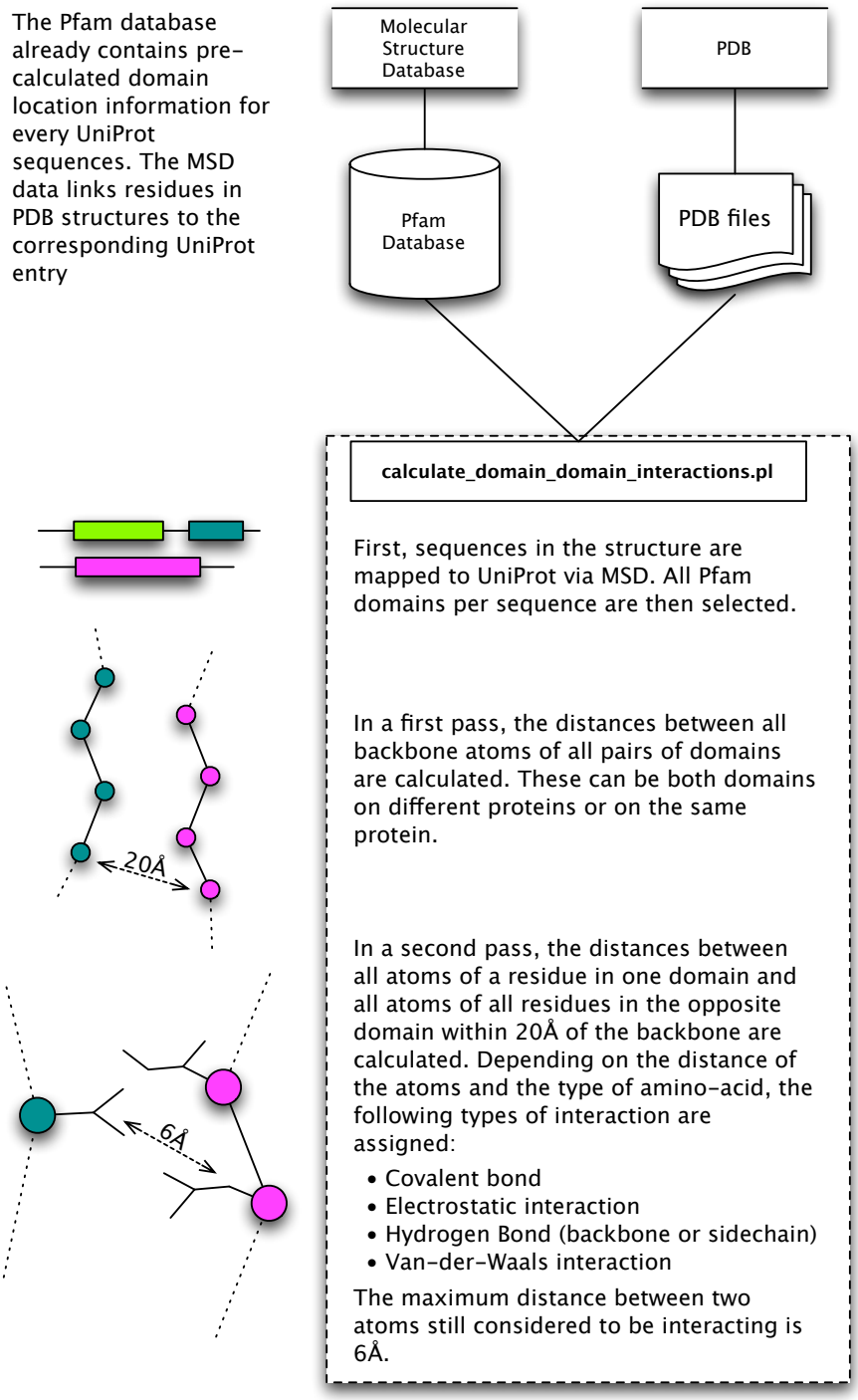


Figure 1.7: Outline of *i*Pfam creation process. Structure data and PDB to UniProt mappings are downloaded from the MSD and PDB, respectively. A single script (calculate_domain_domain_interactions.pl) then performs a sequence of calculations on each structure to identify all atoms in every pair of Pfam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

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.

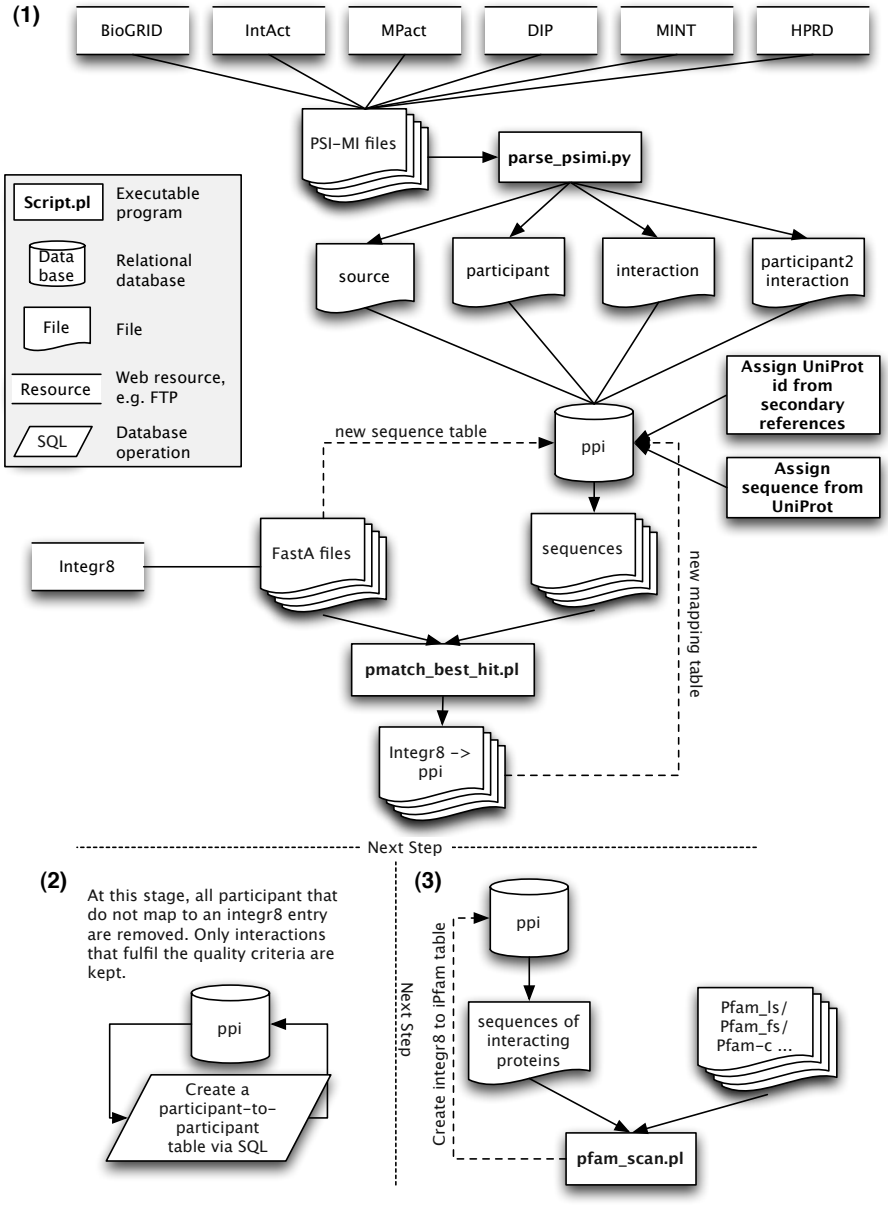


Figure 2.1: Flow-chart of protein-interaction database creation process. **(1)** Interaction information is loaded from numerous online resources by parsing flat-files in PSI-MI XML 2.5 format and subsequently stored in a database as 4 distinct tables. UniProt identifiers are assigned to each protein 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:

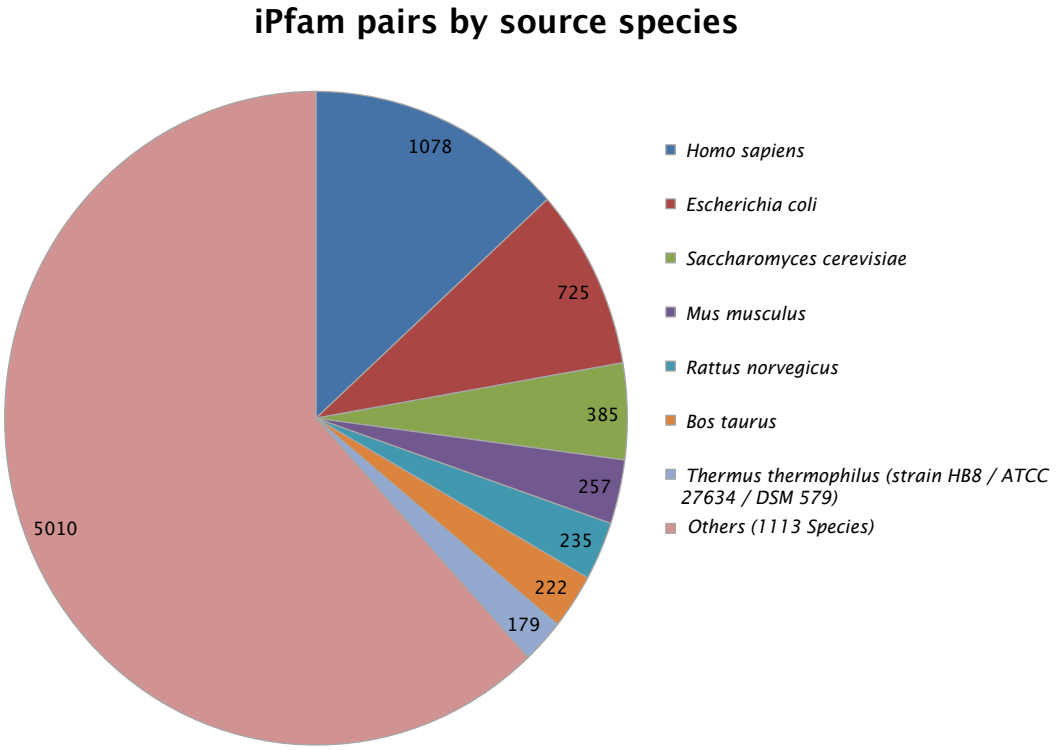


Figure 2.2: This pie chart shows how many *i*Pfam domain pairs were found in PDB structures from each species. The total number is larger than the 4030 unique *i*Pfam pairs in the database because an *i*Pfam pair can be found in structures from several species.

- 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%.

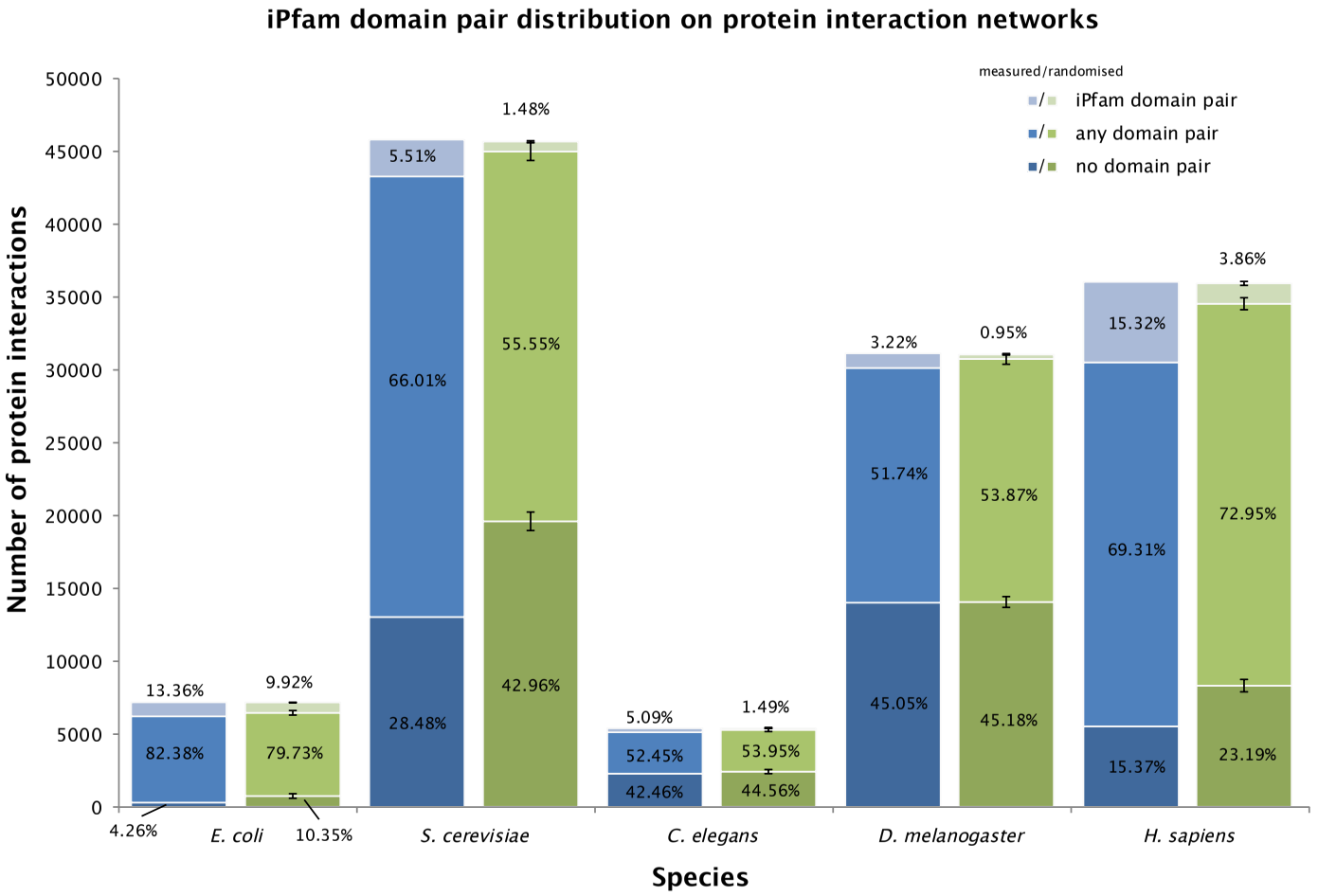


Figure 2.3: Pfam and *i*Pfam coverage on real (blue) and randomised (green) interaction networks. For each species, the height of the columns reflects the number of known protein–protein interactions in the data set. The columns are split according to the proportion of interactions that contain an *i*Pfam 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.

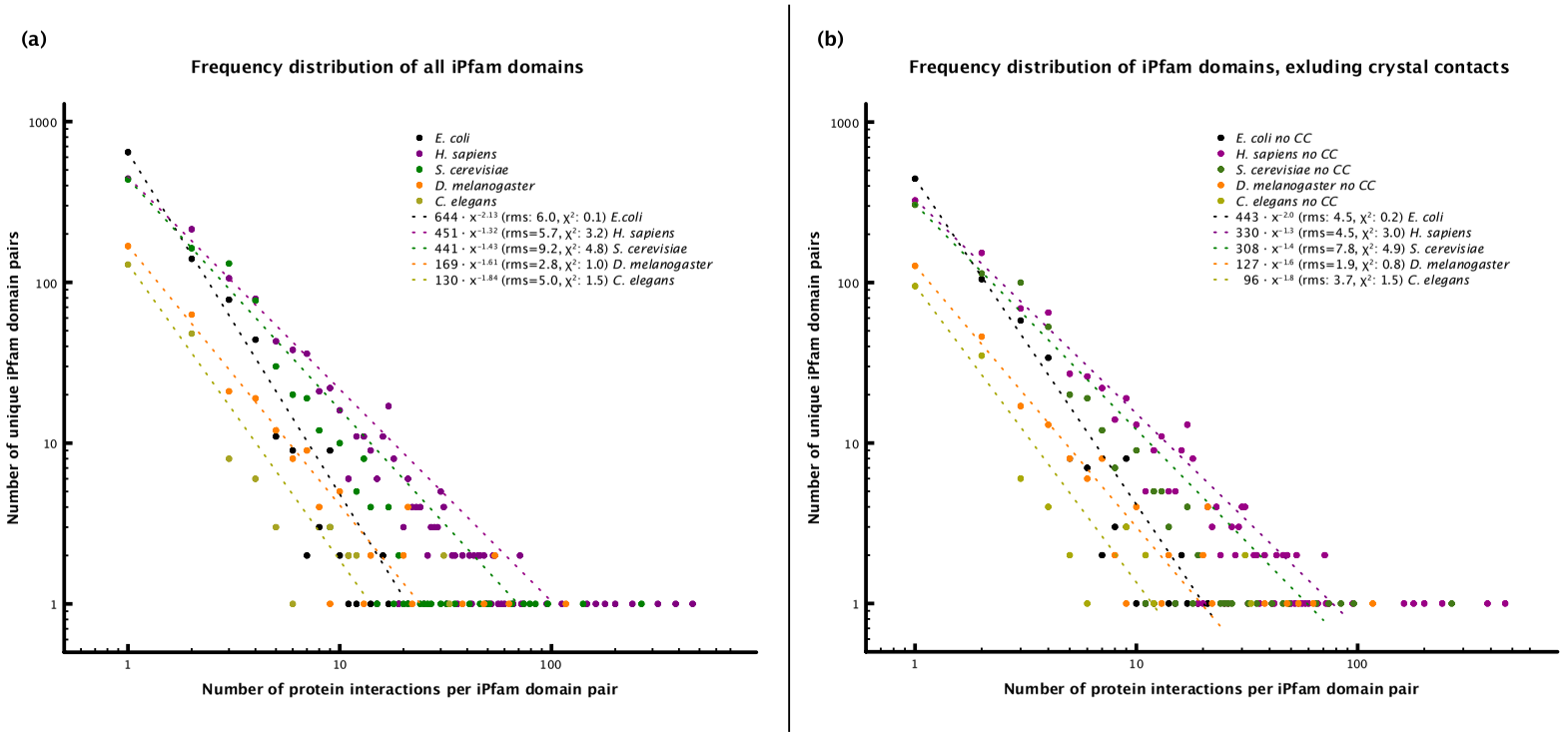


Figure 2.4: Scatter plot illustrating how many *i*Pfam domain pairs occur in how many proteins interactions per species. First, I counted the number of protein interactions each *i*Pfam domain pair occurs in. The x-axis represents the *occurrence frequency*. Then, I counted the number of *i*Pfam domain 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,

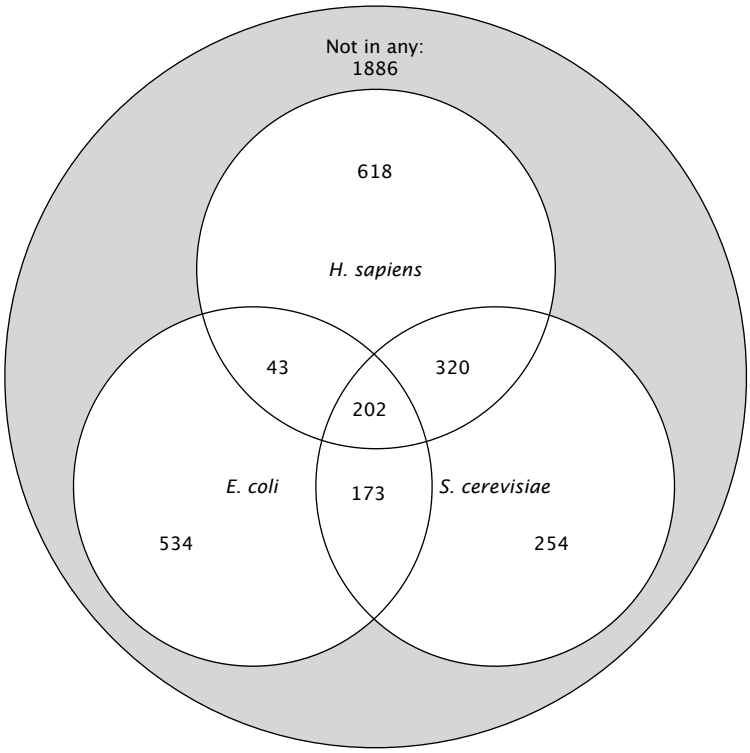


Figure 2.5: The three circles represent the *i*Pfam domain pairs observed in the respective species. The overlaps denote co-observed *i*Pfam domain pairs. The grey set in the background represents *i*Pfam domain pairs not found in the three species.

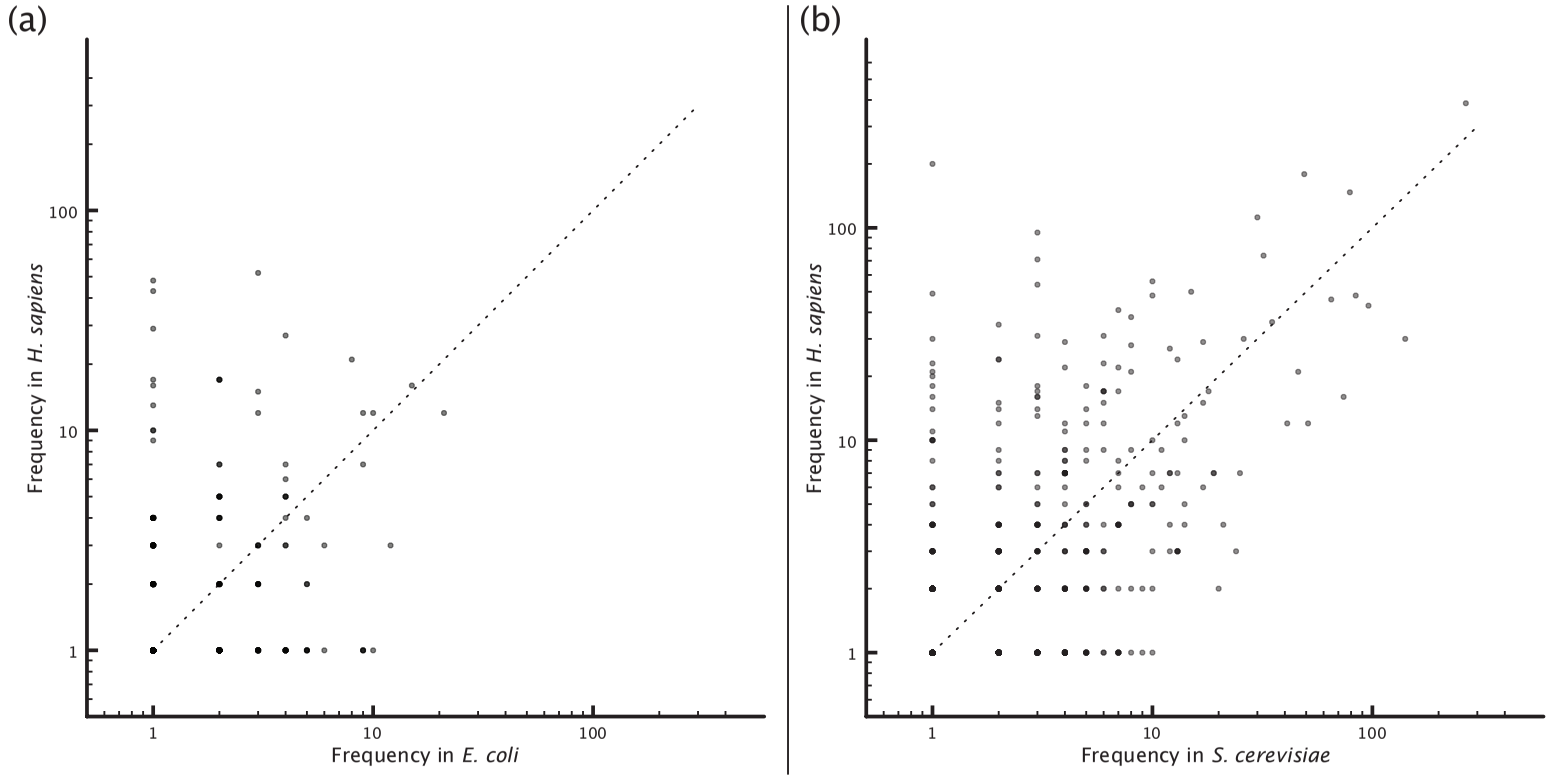


Figure 2.6: Comparison of domain pair frequency between species. **(a)** *E. coli* compared to *H. sapiens*: There is almost no visible correlation between the frequencies. **(b)** *S. cerevisiae* compared to *H. sapiens*: The correlation is much more pronounced, particularly for *i*Pfam domain pairs 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.

Chapter 3

Disease mutations in interaction interfaces

3.1 Introduction

In the previous chapter, I described how *i*Pfam and protein interaction data can be combined to investigate the conservation of interaction interfaces within and between species. Now I will focus on the effects of mutations in interaction interfaces, extending the previously applied methods to the investigation of human disease.

I have mentioned in Chapter 1.2.2 that human genetic diseases with mendelian inheritance have been extensively studied since the 1980s. As a result, databases such as the "Online Mendelian Inheritance In Man database" (OMIM) (Hamosh *et al.*, 2005) and UniProt (Wu *et al.*, 2006) together contain almost 30000 experimentally verified mutations in over 3000 genes. Nevertheless, the exact mechanisms by which mutations alter a protein's function are in many cases poorly understood. Collins *et al.* (1997) estimated that 90% of the variation between individuals can be attributed to single-nucleotide polymorphisms (SNPs). While recent studies (Lu *et al.*, 2007; Redon *et al.*, 2006) have pointed out the importance of large-scale chromosomal structural

variations, most of the known disease-related mutations are non-synonymous single nucleotide polymorphisms in the coding regions of a gene (nsSNPs). It has been suggested that up to 80% of disease-associated nsSNPs destabilize the protein through steric or electrostatic effects (Wang and Moult, 2001; Yue *et al.*, 2005), while a small subset of disease-associated SNPs affect splicing and post-translational modifications (Buratti *et al.*, 2006) or cause stop or nonsense mutations (Savas *et al.*, 2006).

Here, I focus on those diseases that are caused by mutations in protein interaction interfaces. Ferrer-Costa *et al.* (2002) compared disease-associated and neutral nsSNPs in 73 proteins and estimated that 10% of disease-associated nsSNPs may affect the quaternary structure of the protein, thereby changing protein interactions. However, compared to the over 3000 genes for which a mutation is known, 73 proteins reflect only a very limited sample. In recent years, some interaction-related diseases such as Alzheimer's and Creutzfeldt-Jacob disease have received much attention (Chiti and Dobson, 2006; Giorgini and Muchowski, 2005; Ross *et al.*, 2005). These conditions feature an induced aggregation of proteins, often called *amyloidoses*. Figure 3.1 outlines the process of amyloid fibril formation from a native monomer.

Diseases can also be caused by the disruption of protein binding. A typical example is Charcot-Marie-Tooth disease, which can be triggered by the loss of interaction between myelin protein zero monomers which link adjacent membranes of the myelin sheath (Shy *et al.*, 2004). In other cases, protein binding is a means of allosteric regulation. To give an example, mutations in the binding interface of pantothenate kinase lead to inherited pantothenate kinase associated neurodegeneration (PKAN): Enzymatic function critically relies on dimerisation (Hong *et al.*, 2007). Finally, there is also the possibility for mutations to change the binding specificity of a protein and thus lead to new and potentially disruptive interactions. For mutations in the family of human crystallin genes it has been shown that they alter the affinity for the binding partners (Fu, 2003). These erroneous interactions lead to congenital cataract.

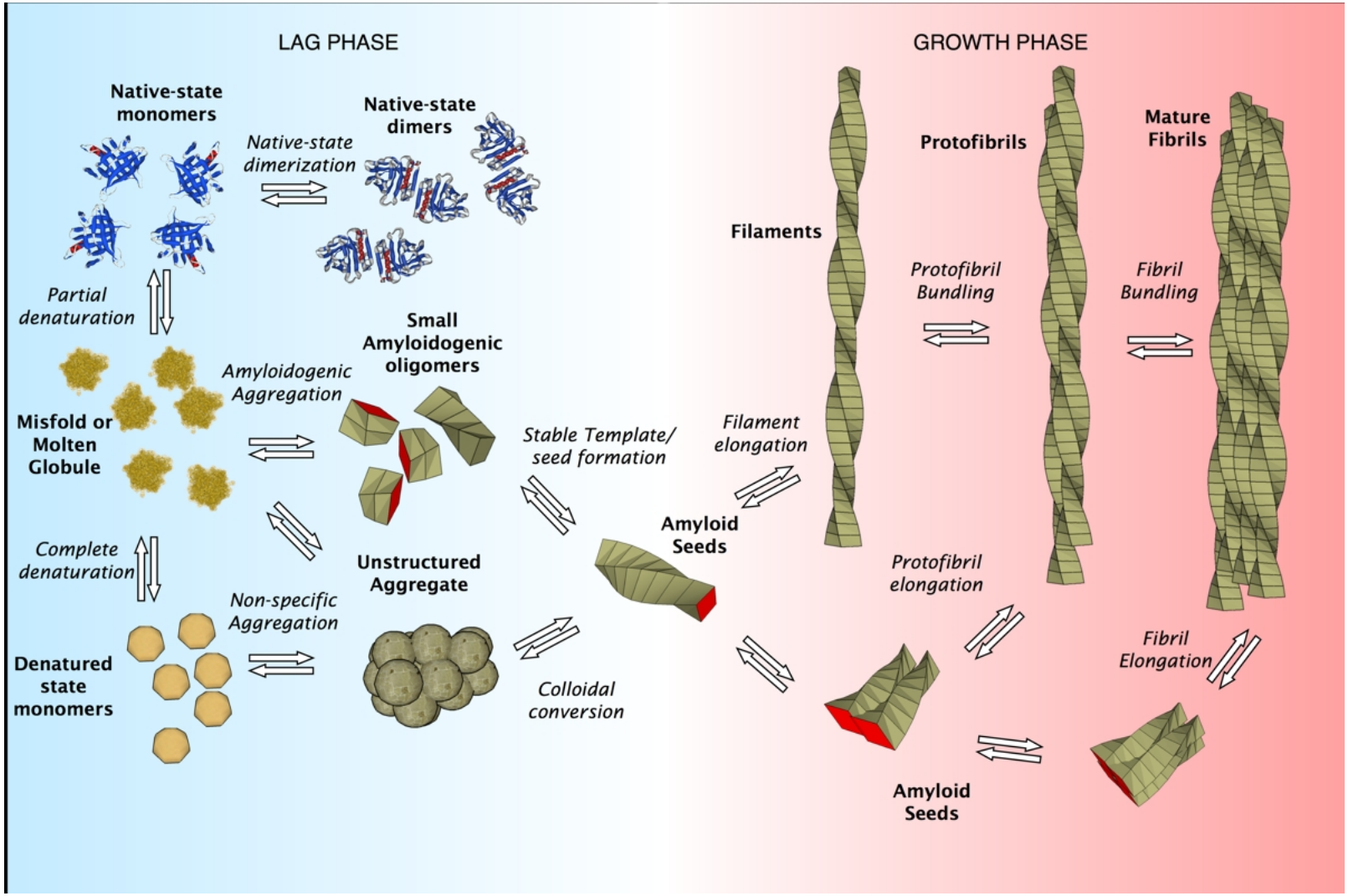


Figure 3.1: Disease pathways of amyloid disorders. In the native state, amyloid precursor proteins maintain an equlibrium between monomeric and dimeric state. Misfolding of a fraction of the protein leads to aggregation of denaturated monomers which subsequently clump into increasingly larger structures which can form into fibrils. Certain mutations in the monomeric subunits can increase the propensity for aggregation. Reproduced with permission from `http://talaga.rutgers.edu/research/amyloid.php`.

While there are numerous topical reports of such interaction related disease, there is to my knowledge no systematic study which investigates the impact of mutations in protein interactions on human disease. Extending the approach outlined in Chapter 2, I describe a method that combines protein structure with experimental protein interaction data in order to computationally identify residues which form part of a binding interface. I apply this algorithm to mutations from OMIM and UniProt, identifying 1428 mutations that are likely to affect protein interactions. Subsequently, I collected numerous topical reports of changes in protein interaction that result in disease. I present a list of 119 interaction-related mutations causing 65 different diseases that was derived manually from the scientific literature. On the basis of these sets I discuss general properties of interaction-related mutations.

3.2 Materials and Methods

3.2.1 Disease Mutations

Mutation data was collected from UniProt (Wu *et al.*, 2006) and OMIM (Hamosh *et al.*, 2005). For UniProt, human sequences with variation information were acquired using SRS (Zdobnov *et al.*, 2002). The analysis was restricted to disease-related single residue mutations by regular expression matching on the variant description line in UniProt entries. Only lines in the form of the following example were parsed:

```
FT   VARIANT     264    264       N -> Y (in CPX).
FT                                /FTId=VAR_021830.
```

OMIM (omim.txt.Z, genemap) and Entrez gene mappings (mim2gene, gene2refseq.gz) were downloaded from the NCBI FTP server (`ftp://ftp.ncbi.nih.gov/`) as flat files. All files were acquired in December 2006. Mapping OMIM entries to a reference sequence is not trivial. Historically, OMIM does not use a well-defined reference database for protein sequences. The curators of OMIM rather refer to the co-ordinates provided

in the original publication for each mutation. Especially old publications frequently refer to the processed protein product rather than the translated gene, which leads to difficulties in assigning the correct locations to the annotated mutations. To accomplish this, protein sequences for every gene id reference in the OMIM entry were acquired from NCBI and UniProt through SRS. To identify the correct co-ordinate system that fits an OMIM entry, removal of combinations of signal peptide and other post-translationally cleaved regions were considered. If the amino-acid annotations in the OMIM entries for a gene matched the residues at the respective position in the reference sequence, that co-ordinate system was used. Figure 3.2 outlines the combination of scripts and data involved in this process.

3.2.2 *i*Pfam

*i*Pfam version 20 was employed, containing 3020 interacting domain pairs composed of 2147 individual domains (Finn *et al.*, 2005). A detailed description of *i*Pfam can be found in the introduction (Section 1.3.1).

3.2.3 Predicting crystal contacts

As described in detail in the Methods for Chapter 2, the *NOXclass* classifier (Zhu *et al.*, 2006) was applied to the structures from which *i*Pfam was derived. NOXclass requires ConSurf conservation scores. The last release of pre-calculated ConSurf data (ConSurf-HSSP, see Glaser *et al.* (2005)) has not been updated since March 2005. Hence, only 7588 out of the 9263 structures with two distinct protein chains in *i*Pfam v20 could be passed through NOXclass. 2592 structures contained a putative crystal contact with greater than 90% probability.

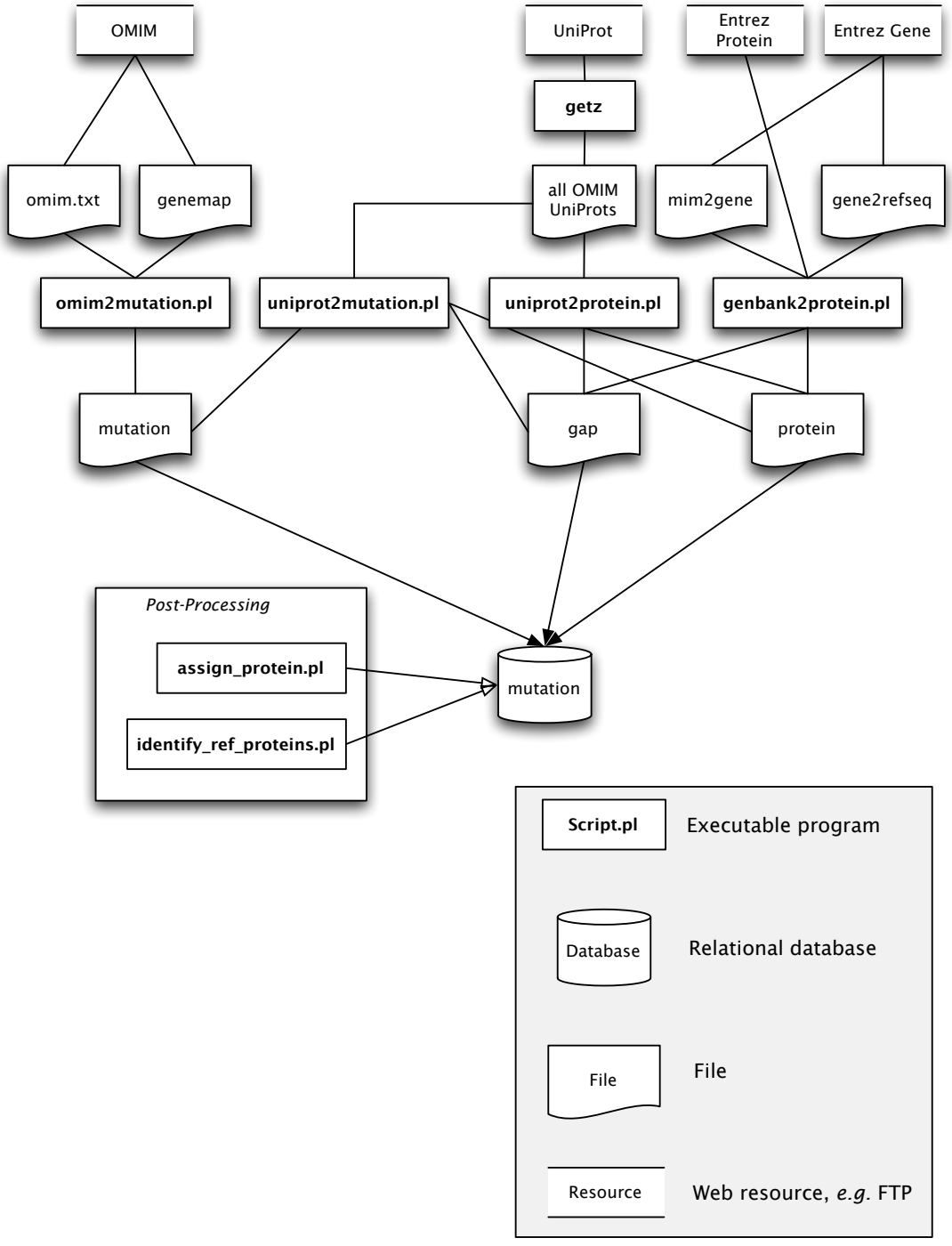


Figure 3.2: Workflow for generation the mutation database from OMIM and UniProt. Several Perl scripts merge and format the data to be imported into a relational database. The post-processing scripts then identify the sequence/post-translational modification combination that best matches the observed mutations.

3.2.4 Homology Detection and Alignment

Protein sequences were screened for *i*Pfam families using hidden Markov models with the `pfam_scan.pl` script which can be downloaded from `ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/`. This script searches a collection of sequences in a FASTA file against Pfam family definitions in the form of HMM files. It uses the `hmmpfam` program which is part of the HMMer package (Eddy, 2001). It automatically applies significance thresholds and clan overlap definitions before returning a tab-delimited output of significant matches of families per sequence in the input file.

Here, a custom HMM library was employed which only contained *i*Pfam HMMs. For each identified family, matching regions in query protein were aligned to the sequences for which an interacting structure is known. Alignments were performed using `hmmalign` from the HMMER package. The percentage sequence identity between all pairs of aligned regions was calculated using the exact (non-heuristic) implementation in the `Bio::SimpleAlign` BioPerl module. A flow-chart outlining the steps involved is shown in Figure 3.3.

3.2.5 Residue prevalence

Residue prevalence denotes the frequency with which a certain amino-acid occurs at a given position in a domain when numerous homologous sequence regions are compared. Residue prevalence was extracted directly from the Pfam HMM that matched a sequence region. Each emitting state in an HMM, *i.e.* Match and Insert states, contain a distribution of observation probabilities (usually called *emission probabilities*) for each amino-acid. This distribution is learned from the training files, involving the application of elaborate prior models to account for possible biases due to small training sets. In addition to that, the HMM file also contains a background distribution (the *null-model*) which is fixed and represents the global frequency of amino-acids. Columns in the alignment were mapped back to states in the HMM *via* the RF line

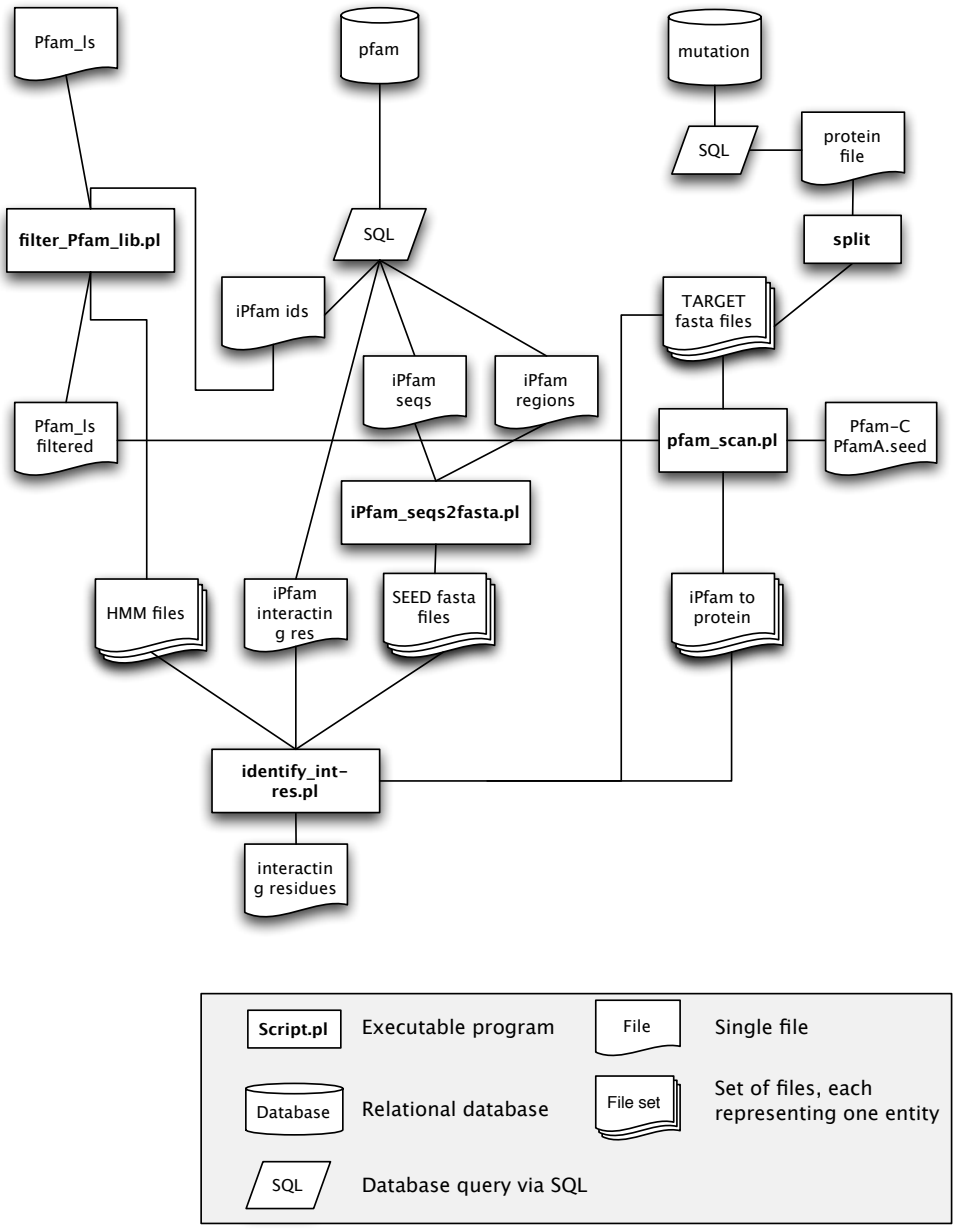


Figure 3.3: Outline of the computational steps leading to the mapping of interacting residues to known disease mutations. The central script is called `identify_int-res.pl` and takes an HMM library file and two sets of fasta files corresponding to domain regions, one containing the structural seeds and another the target sequences, in this case disease genes. It then aligns the target sequences to the structural template regions using `hmmalign` which is part of the HMMER package. For each column in the resulting multiple sequence alignment, the script then outputs all predicted interacting residues and the originating template residues, as well as the percentage sequence identity between the target and query sequences.

in the Stockholm-format output of `hmmalign`. The `HMM` Perl library (Schuster-Böckler *et al.*, 2004) was employed to extract all data from the HMM file. For every column in the alignment, the log-odds scores $\log_2(P_{emission}/P_{null-model})$ were calculated and used as prevalence scores.

3.2.6 Alanine Scanning Database

The ASEdb database (Thorn and Bogan, 2001) containes data from 101 alanine scanning experiments extracted from 74 publications (`http://www.asedb.org`). 81 mutations extracted from five recent publications were added manually for this analysis (Grace *et al.*, 2007; James *et al.*, 2007; Logsdon *et al.*, 2004; Walsh and Kossiakoff, 2006; Williams *et al.*, 2006). In such an alanine scan, residues in the binding interface of a protein are mutated to alanine by site-directed mutagenesis (Cunningham and Wells, 1989). The difference in binding free energy ($\Delta\Delta G$) between wild-type (ΔG_0) and mutated protein (ΔG_A) describes the contribution of a particular residue at position i to the total binding free energy: $\Delta\Delta G_i = \Delta G_O - \Delta G_{A,i}$. 3010 residue mutations are recorded in ASEdb. Mutations leading to incorrectly folded proteins or premature degradation were excluded from ASEdb if this information was available in the source publication. In order to use hidden Markov models to search for *i*Pfam domains, protein sequences corresponding to the gene name annotated in ASEdb were retrieved from UniProt. Only proteins for which all amino acid annotations in ASEdb matched the sequence were included. For 858 residue mutations, a UniProt sequence could be identified.

109 mutations came from experiments that involved an antibody as the binding partner. In this investigation, I am interested in evolutionarily conserved interactions between molecules in living cells. Conversely, the interactions between antibodies and antigens are not representative for normal biological interactions and were therefore removed from ASEdb.

3.2.7 Compiling the curated set of interaction-related mutations

In order to identify known interaction-related mutations, all OMIM "Description" fields were searched for keywords such as "interaction", "binding" or "complex". For all matching mutations, the available literature was manually evaluated. Subsequently, PubMed was searched for the same keywords. Lastly, cases that were identified by the prediction method were added if they were found to be known in the literature. If a mutation was shown to be causative and described to directly affect a protein interaction, it was added to the list. Mutations that lead to folding errors were excluded from the data set. The complete list can be found in Table F in the Appendix.

3.2.8 Statistical Analysis

All statistical calculations were performed in R (R Development Core Team, 2006). In particular, the test of difference in proportions was performed *via* the R function `prop.test` with default settings.

3.2.9 Graphics

Three-dimensional protein images were prepared using VMD (Humphrey *et al.*, 1996) and rendered with PovRay (`http://www.povray.org/`).

3.3 Results

3.3.1 Prediction algorithm

In order to identify residues in a protein that are involved in a protein interaction, I devised a method that combines structural and experimental information. Using the *i*Pfam (Finn *et al.*, 2005) database of known interacting domains, I first select domain regions on all target proteins that have a homologous structure including interaction partners in the PDB (Kouranov *et al.*, 2006) (see Section 3.2.4). I then select positions

which form residue-to-residue contacts between distinct polypeptide chains in these *structural templates* and record the corresponding positions in the target proteins as potentially interacting residues, see Figure 3.4.

3.3.2 Prediction accuracy

To estimate the accuracy of my prediction approach, I undertook two independent benchmarking experiments. First, I performed a cross validation experiment where for each *i*Pfam family, I attempted to identify the correct interacting residues in a PDB structure not used for prediction. This process was repeated 5 times for different combinations of training and target sequences. In a second experiment, I used the ASEdb database of alanine scanning energetics experiments in protein binding (Thorn and Bogan, 2001) as a "gold-standard" test set (see Section 3.2.6).

In order to apply an accuracy threshold, I needed to choose a scoring function that discriminates between residues that are really involved and crucial for an interaction and those that are not. For this purpose, I tested the effect of two different variables on prediction accuracy:

3.3.2.1 Percent sequence identity with structural template

There is a well known correlation between sequence similarity and structural similarity (Chothia and Lesk, 1986) which also extends to interacting domains (Aloy *et al.*, 2003). An interaction is more likely to be conserved and to display similar topology when sequence similarity is high. For many target proteins, there are several structural templates that could be applied to predict the interacting residues. I hypothesised that the sequence similarity as measured by percentage sequence identity could discriminate between trustworthy and less convincing predictions. Accordingly, percentage sequence identity was tested as a threshold parameter in the following benchmark experiments.

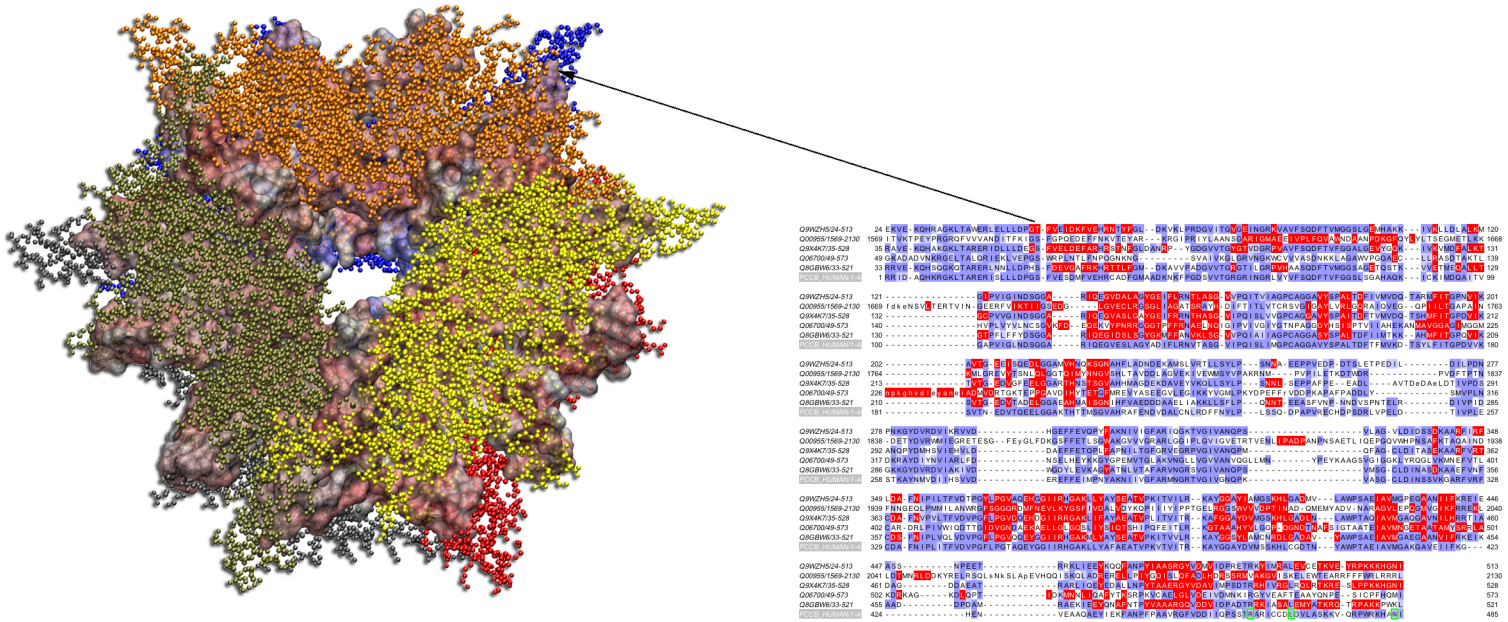


Figure 3.4: Predicting potentially interacting residues from structure. To the left: Structure of Propionyl-CoA carboxylase beta chain with the interaction interface shown as a surface (PDB 1vrg). To the right: Alignment of all sequences matching Pfam domain *Carboxyl_trans* (PF01039) for which a structure of a multimer is available. Residues which are part of the interaction interface in at least one structure are shown in red in the alignment. The three residues framed in green are known to inhibit multimerization.

3.3.2.2 Prevalence of mutated residues

For all predicted interaction-related residues, I calculated a prevalence score (see Section 3.2.5). This score reflects the frequency with which an amino-acid occurs at a given position in a protein family, relative to a universal background distribution. If I look at the frequency of prevalence scores over all wild type compared to all mutated alleles (Figure 3.5), I find that the scores for both wild-type as well as mutated alleles seem to follow a normal distribution, see Figure 3.5). The exceptionally large number of original residues with log-odds scores around 3 can be attributed to the fact that mutations are more likely to be severe in functionally important residues, which in turn are more likely to be conserved. The mutated residues exhibit markedly smaller average prevalence scores (2.4 vs. -2.2 than the original residues. Thus, a residue that is found in the wild type of a protein will usually be more conserved than the residue found in the mutated version (Ng and Henikoff, 2003). I therefore tested whether residue prevalence could be used as an indicator of the functional importance of a residue, even for surface exposed residues like the ones under investigation here.

3.3.2.3 Cross validation results

I performed a random sub-sampling cross validation experiment to determine if my algorithm is capable of identifying interacting residues in proteins for which a similar interacting structure is known. The cross-validation procedure included the following steps:

1. Collect all structures with an interaction containing iPfam family P.
2. If there are less than 5 distinct sequences amongst all structures, skip the family.
3. If possible, check for each distinct chain pair in the structure if it is a potential crystal contact by applying the NOXclass classifier (see Methods).

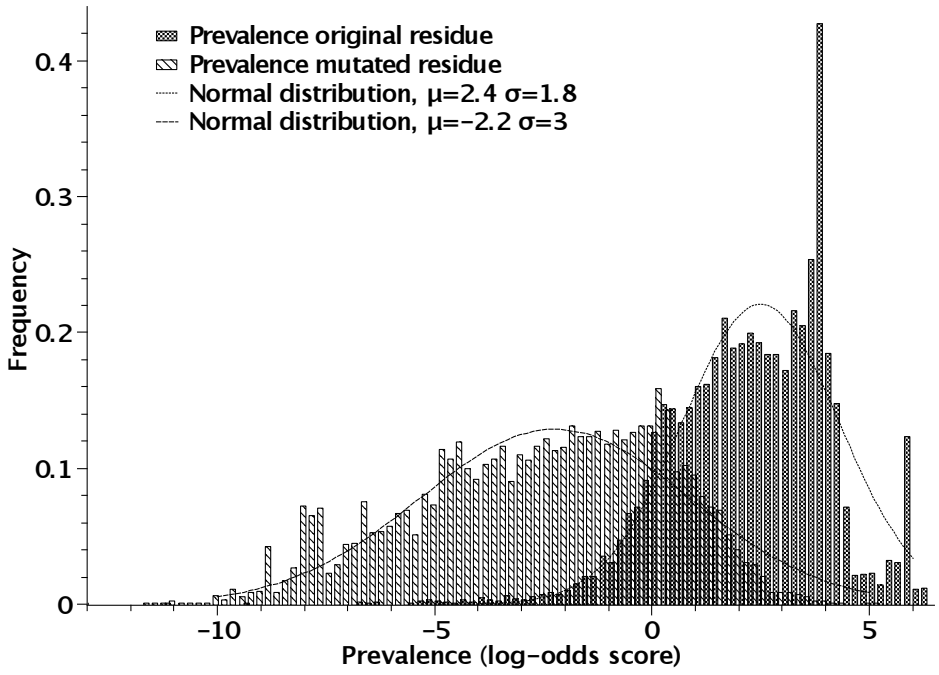


Figure 3.5: Histogram of prevalence of wild-type and mutated residues. The prevalence score distributions of mutated and wild-type residues are clearly separated. They intersect around 0, suggesting that residues whose frequency is similar to the background distribution are as common in mutations as in wild-type alleles. Trendlines are added to delineate that both distributions are approximating a normal distribution.

4. Select one target sequence at random out of the set of all sequences with at least one interacting structure including family P

5. Apply the interacting residue prediction as described above, using all structures except the ones including the target sequence.

6. Compare the predicted interacting residues to the residues actually observed in any structure of domain P in the target sequence.

7. repeat for all iPfam families. Then concatenate results and calculate performance.

Figure 3.6 shows the resulting receiver operator characteristic (ROC) curves (Fawcett, 2006), a plot of the frequency of true positive over the frequency of false positive predictions for a given algorithm. From left to right, points mark decreasing score thresholds, until no thresholds are applied any more and both true positive as well as false positive rates reach 100% in the upper right corner. The different plots reflect combinations of different thresholds and testing data. Notably, percentage sequence identity between seed and target sequence is a good discriminator between true and false positive predictions, as seen in Figure 3.6a. Removing crystal contacts and excluding residues involved in intra-chain interactions also slightly improves prediction accuracy. Residue prevalence (Figure 3.6b) performs very similarily. In comparison, a combination of prevalence and percentage identity where all predictions from seeds with $\leq 30\%$ sequence identity were removed (Figure 3.6c) performs significantly worse. This indicates that the most important step in the prediction algorithm is the assignment of interacting residues itself, whereas the subsequent filtering of residue according to percentage identity or residue prevalence has only a small effect on accuracy.

3.3.2.4 ASEdb results

The cross validation experiments verify that the algorithm can retrieve residues which are involved in interaction interfaces from homologous sequences. In order to determine

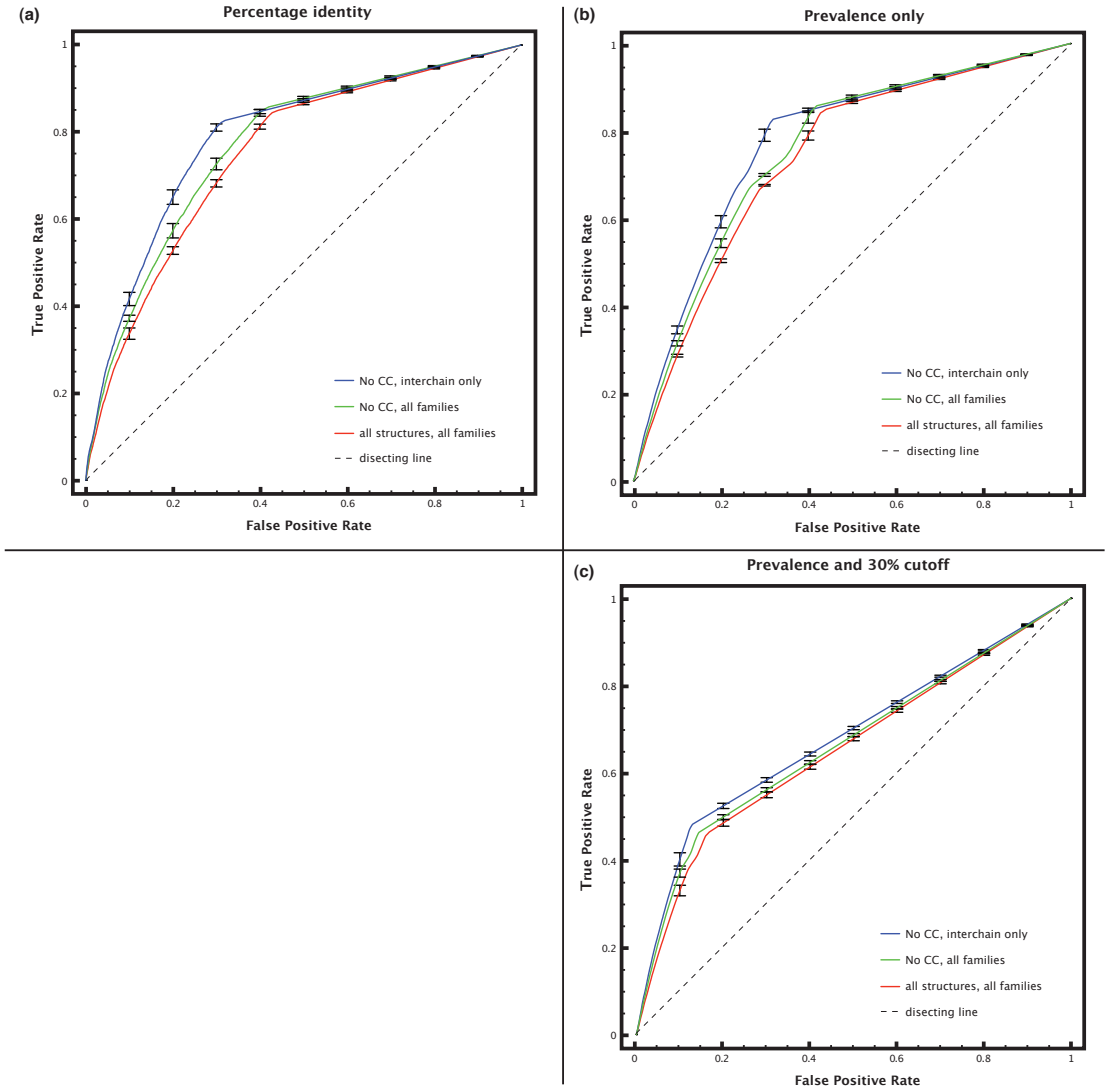


Figure 3.6: Receiver Operator Characteristic (ROC) curves calculated on cross-validation results. Each curve is the combined classification result of all predictions made on the sum of all the individual *i*Pfam families. Bars reflecting standard deviation between repetitions with different training/target sets are shown. Red lines denote benchmarks on all structures for all *i*Pfam families (red). Green lines were calculated on data excluding chain pairs with $\geq 90\%$ probability of being a crystal contact. For blue lines, all interacting residues derived from intra-chain interactions were excluded from the training data in addition to the crystal contacts. **(a)** Percentage sequence identity between seed and target sequence as a threshold. **(b)** Only residue prevalence as a threshold. **(c)** Mixture of percentage identity and residue prevalence as threshold: Residues with $\leq 30\%$ identity to the seed sequence were set to minimum prevalence. ROC curves were computed using the ROCr package for R (Sing *et al.*, 2005).

the impact of a mutation in a protein interaction interface, I also want to assess how well I can predict the functional importance of individual interacting residues.

I assessed how well my method could predict residues with a large change in ΔG upon mutation as recorded in the ASEdb database (see Methods). Randles *et al.* (2006) showed that for two model proteins, $\Delta\Delta G$ was correlated with the severity of disease. They show that even changes < 2 kcal/mol could cause disruption of protein binding. Here, I defined a residue as correctly identified (true positive) if $\Delta\Delta G > 2.5$kcal/mol. This threshold is also used in another recent publication (Ofran and Rost, 2007). Residues below this threshold were considered neutral (false positive). This criterion might in itself cause some "false-negatives", *i.e.* some residues might be crucial for the function of the protein despite a measured $\Delta\Delta G$ less than 2.5kcal/mol, but I considered a conservative threshold to be preferable.

Figure 3.7 shows ROC curves for the ASEdb benchmark. The green and red lines represent the performance of my algorithm using either percentage sequence identity (green) or residue prevalence (red) to score the predictions. With both scoring methods, my method retrieves more true positives than would be expected by chance. The prevalence threshold however is far superior in distinguishing true from false positives. At a false positive rate of $\approx 20\%$, I can achieve a true positive rate of almost 60%. These benchmark results underline that the algorithm is able to identify interaction disruptive mutations with reasonable confidence.

I again tested a combination of the two measures, represented by a blue line in Figure 3.7. In this case, only structural templates with at lease 30% sequence of the interacting domain were selected before applying the prevalence threshold. The performance improves slightly in the low false-positive region, yielding a true positive rate of 40% at a false positive rate of only 7%. More importantly, a minimum sequence identity threshold increases the confidence in the structural similarity between seed and target proteins. Hence, I decided on a residue prevalence threshold of > 2 in

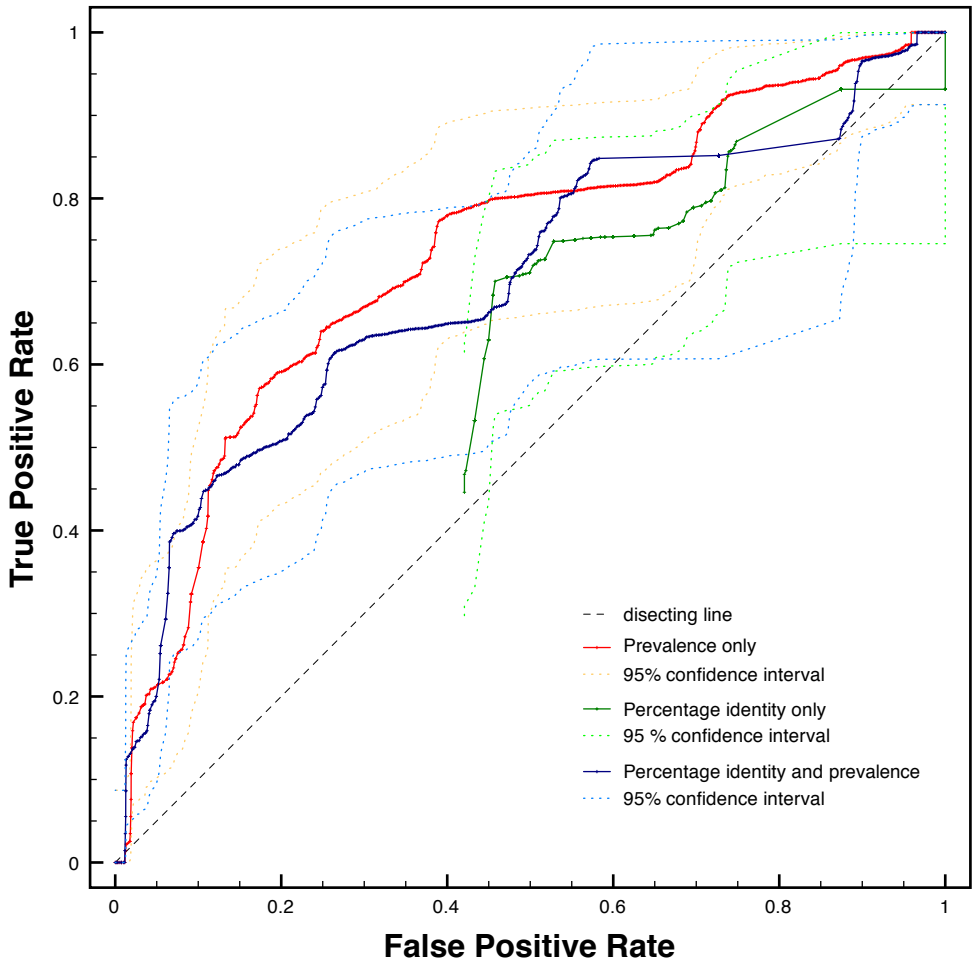


Figure 3.7: Receiver Operator Characteristic (ROC) curves calculated on a set of alanine scanning experiments. The red line represents the performance of my algorithm when changing only the residue prevalence threshold, applying no percentage identity cutoff. The green line shows the performance using only percentage identity as a threshold. The blue line reflects performance using prevalence as threshold after applying a 30% sequence identity cutoff. Confidence intervals where calculated using the `Statistics::ROC` Perl module (Kestler, 2001).

combination with a 30% sequence identity cutoff for all subsequent analyses.

3.3.3 Application to Disease Mutations

I applied the prediction algorithm as described above to all single-residue disease mutations extracted from OMIM and UniProt (see Methods). In the case of disease mutations, the disruptive nature of a residue mutation is already known. It is unclear, however, whether an interaction is in fact taking place and is likely to be mediated by the domain in question. Mutations were therefore only reported if the disease associated protein has a close homolog which has been proven experimentally to interact with a protein that contains the same binding partner domain as seen in the PDB structure the interaction was modelled from: Target proteins had to have a homologous sequence (BLAST e-value of less than 10^{-6}) in one of five major repositories for protein interaction information (IntAct (Kerrien *et al.*, 2007), BioGRID (Breitkreutz *et al.*, 2008), MPact (Guldener *et al.*, 2006) or HPRD (Mishra *et al.*, 2006)) and DIP (Salwinski *et al.*, 2004) [1]. Subsequently, target proteins were excluded if no homologous experimental interaction involved both interacting *i*Pfam domains that were seen in the structural template. For example, [OMIM: +264900.0011] is a Ser576Arg mutation of the human coagulation factor IX (PTA). The residue is part of a Trypsin domain and seen to interact with Ecotin in several structures [*e.g.* PDB: 1xx9]. However, the interaction between PTA and Ecotin is not yet recorded in any interaction database, therefore the mutation cannot be included in my predictions.

Using these criteria, 1428 mutations from 264 proteins were predicted to be interaction-related (see Figure 3.8). The full list is attached in Appendix G. In total, I collected 25322 mutations from OMIM and UniProt. This means that approximately 5.6% of all mutations could be linked to a protein interaction.

Amongst these mutations, 454 mapped to a structure that exhibits an interac-

[1] MINT was temporarily unavailable when the analysis was performed and could thus not be included.

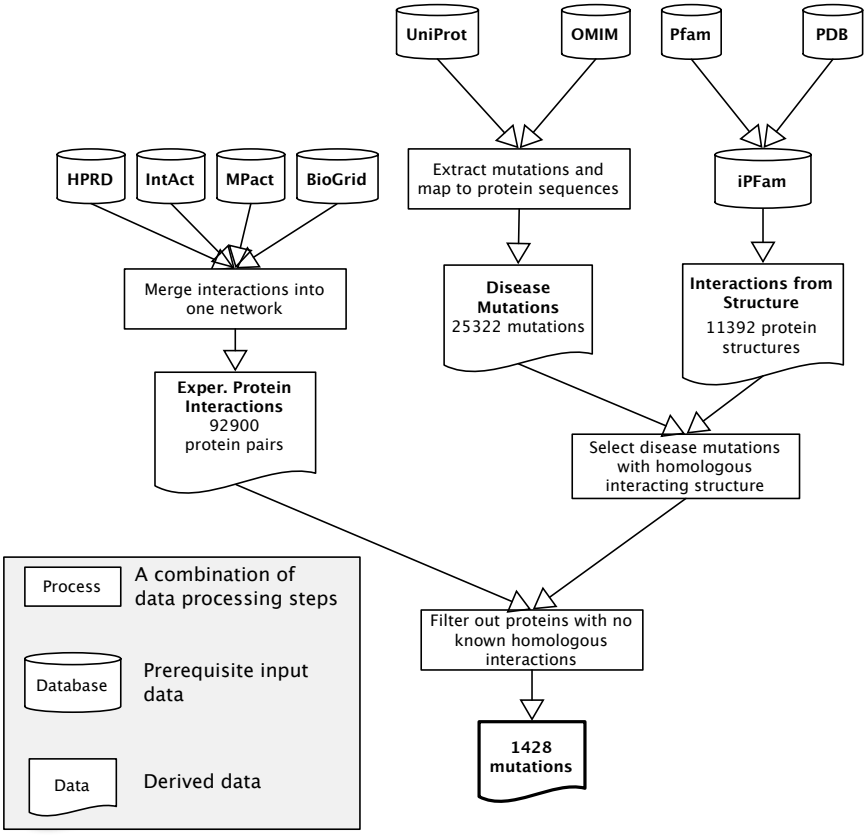


Figure 3.8: Schematic outline of data integration for the prediction of interacting residues. Mutations from OMIM and UniProt for which a residue in a homologous structure is involved in an interaction are selected. This set is restricted further by searching for homologous proteins with known interactions, taken from a range of protein interaction databases. I require that the the homologous interacting proteins contain the same pair of Pfam domains that was observed in the structural template. This results in a set of 1428 interaction related mutations.

tion between different proteins (hetero-interaction), while 1094 mutations mapped to a structure with an interaction between two identical proteins (homo-interaction). This means that 120 mutations are found in structures of both homo- and hetero-interactions. The large proportion of homo-interactions can be explained by the over-representation of homo-interactions in the structural templates set: 70% of all distinct protein pairs in *i*Pfam are homo-interactions, which is in accordance with recent findings that homo-interactions are more common than hetero-interactions (Ispolatov *et al.*, 2005).

Finally, I test if some of the predictions are based on structures which are likely to be a crystal contact. 309 interacting residues were predicted from a chain pair with NOXclass P-values > 0.9, slightly reducing the fraction of interaction related mutations to 4.4%.

3.3.4 Properties of mutations in interaction interfaces

Below, I explore differences between interaction-related mutations and non-interaction-related mutations. I focus on the mechanism of the mutation, the mode of inheritance and residue composition. For most of the 1428 mutations from the automatically generated set, no information about their mode of inheritance or functional mechanism was instantly available. I therefore randomly sampled 100 mutations out of those 1428 and conducted a manual search of the literature in order to annotate their properties.

3.3.4.1 Curated set of interaction-related mutations

In addition to the automatically derived data, I collected 119 mutations in 65 distinct diseases from the scientific literature for which there is evidence that they change the interactions of the protein they occur in (see Methods). I call this the *curated set* of interaction-related mutations (see Appendix F). To my knowledge, it represents the biggest dedicated collection of high confidence interaction-related mutations to date.

3.3.4.2 Classification according to function

I suggest a classification that groups mutations according to their effects into loss of function (LOF) and gain of function (GOF). Below this broad distinction, the GOF mutations can be further divided into two groups: Pathological aggregation and aberrant recognition. Similarly, LOF mutations can be split into one class that disrupts obligate interactions between protein subunits and another class which interferes with transient interactions.

From the curated set of interaction-related mutations, 95 mutations result in LOF, 17 in GOF, four mutations were reported to change the interaction preference of the protein and three could not be determined. The class of GOF mutations that result in protein aggregation contains 12 cases, comprising amyloid diseases like Alzheimer or Creutzfeldt-Jacob, but also for example sickle cell anaemia [OMIM: +141900.0243]. Five cases result in aberrant recognition, for example a Gly233Val mutation in glycoprotein Ib that leads to von Willebrand disease [OMIM: *606672.0003] by increasing the affinity for von Willebrand factor.

Amongst the LOF mutations, 61 affect transient interactions and 34 affect obligate interactions. The latter usually render proteins dysfunctional, for example in the case of lipoamide dehydrogenase deficiency caused by impaired dimerization (Shany *et al.*, 1999). LOF mutations in transient interactions cause changes in localization or transmission of information, exemplified by a mutation in the BRCA2 gene that predisposes women to early onset breast cancer: a Tyr42Cys mutation in BRCA2 inhibits the interaction of BRCA2 with replication protein A (RPA), a protein essential for DNA repair, replication and recombination (Wong *et al.*, 2003). Lack of this interaction inhibits the recruitment of double stranded break repair proteins and eventually leads to an accumulation of carcinogenic DNA changes.

3.3.4.3 Mode of inheritance

I investigated the mode of inheritance for all mutations in the curated set, if information was available in the literature. All GOF mutations showed dominant inheritance (the two hemoglobin mutations exhibit incomplete dominance). Out of 61 LOF mutations for which inheritance information was available, 24 were autosomal dominant and 37 were recessive. Jimenez-Sanchez *et al.* (2001) studied the mode of inheritance of human disease genes. According to them, mutations in enzymes are predominantly recessive, while mutations in receptors, transcription factors and structural proteins are often dominant. Overall, they find a ratio of 188 : 335 of dominant to recessive diseases. In my data set, the ratio of dominant to recessive mutations is 41 : 37[1]. This enrichment for dominant mutations, compared to Jimenez-Sanchez *et al.*, is statistically significant, as determined by a two-sided test for equality of proportions (P-value < 0.014). In the 100 randomly chosen mutations from the predicted set, I found a ratio of dominant to recessive mutations of 38 : 41, which is very similar to the ratio observed in the curated set (P-value > 0.68, *i.e.* no significant difference between the predicted and the curated set).

3.3.4.4 Residue frequency

The residue frequency of the predicted interaction-related mutations was compared to the frequencies of residues over all mutation in OMIM and UniProt (Vitkup *et al.*, 2003). I find that the frequency distribution of wild-type residues in interaction-related mutations is mostly similar to the overall mutational spectrum, with the exceptions of a significant enrichment in Gly and, to a lesser extent, a higher frequency of Trp and Gln and a reduced frequency of Ala, Ser and Val (see Figure 3.9). The enrichment in Gly can not be readily explained by the composition of residues on the protein surface

[1] Jimenez-Sanchez *et al.* counted diseases, not individual mutations. In terms of diseases, I observe a ratio of 31 : 29

or in interaction interfaces (Chakrabarti and Janin, 2002; Ofran and Rost, 2003) but might be due to the disruptive nature of the residues Gly is most likely to mutate to, namely Arg, Ser and Asp (Vitkup *et al.*, 2003).

3.3.5 Examples of putative interaction-related mutations

In the following section I describe four diseases identified by my method which appear likely to be related to changes in protein interaction.

3.3.6 2-Methyl-3-Hydroxybutyryl-CoA Dehydrogenase Deficiency [OMIM: #300438]

Ofman *et al.* (2003) identified a Leu to Val mutation at position 122 in the short-chain 3-hydroxyacyl-CoA dehydrogenase (HADH2) that causes a defect in isoleucine metabolism. The clinical effect was psychomotor retardation and non-progressive loss of mental and motor skills. Ofman *et al.* investigated the molecular effects of the Leu122Val mutation. Immunoblotting showed almost no reduction in the amount of enzyme, but enzyme activity was greatly reduced.

Powell *et al.* (2000) resolved the crystal structure of the homologous protein for HADH2 in rat [PDB: 1e3s, 1e3w, 1e6w], see Figure 3.10. The rat protein shares 84% sequence identity with the human homolog. Like other members of the short-chain dehydrogenase (SDR) family, HADH2 forms a homotetramer. The mutated Leu122 is part of the αD helix adjacent to the NAD binding pocket, as shown in Figure 3.10. NAD binding does not seem to affect the conformation of the αD helix, according to the three crystal structures of the complex at different stages of the enzymatic reaction. Kissinger *et al.* (2004) crystallised the human form of HADH2. Their investigation focused on the effect of HADH2 on Alzheimer's disease, specifically on the binding of HADH2 to amyloid β precursor protein. They did not mention the effect of mutations in the dimerization domain on protein function. The human structure shows the same

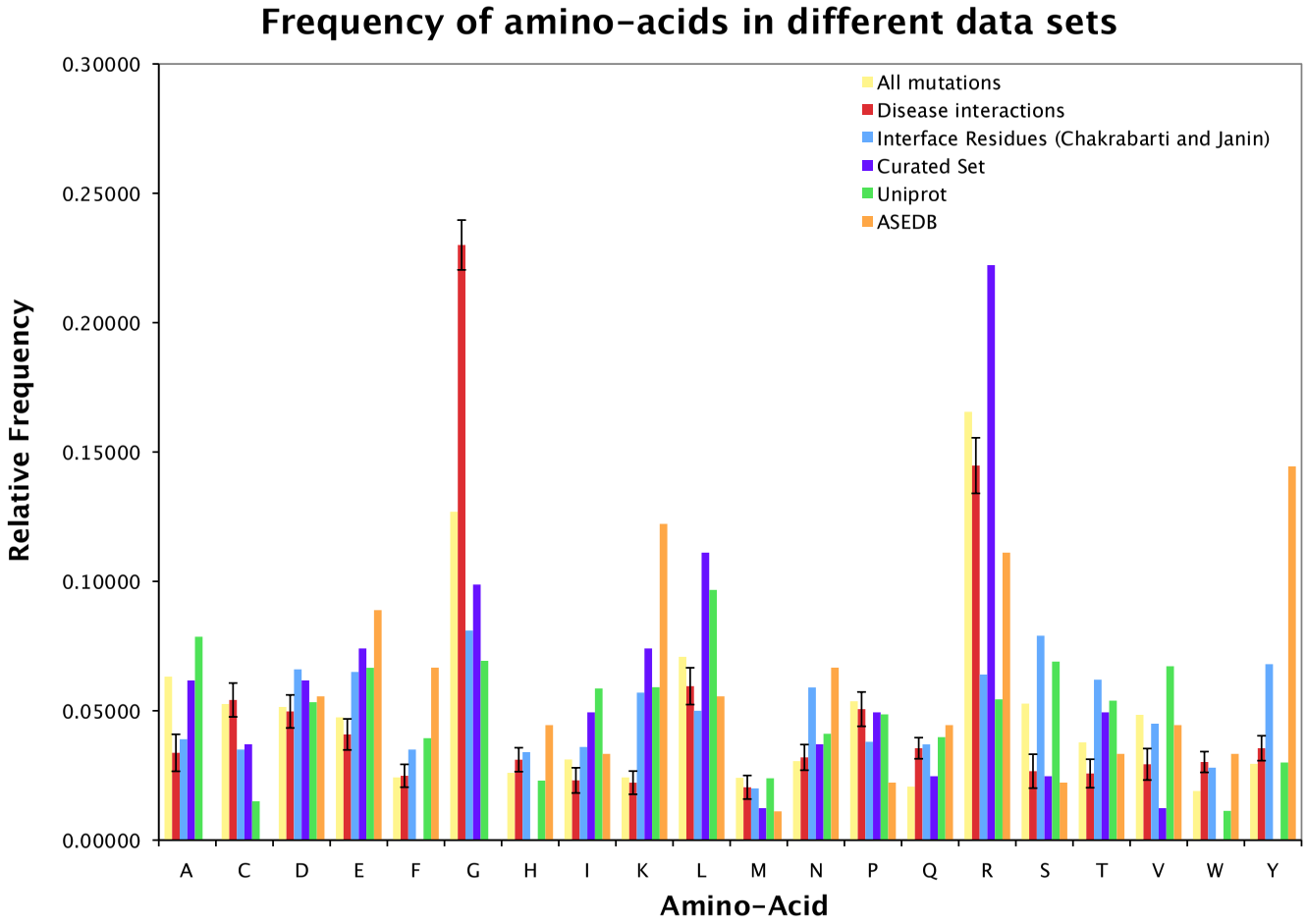


Figure 3.9: Distributions of residue frequencies for all mutations in OMIM and Uniprot (wild type), the predicted set (wild type), the curated set, for interface residues as described by Chakrabarti and Janin (2002), the whole of UniProt and for residues from ASEdb with $\Delta\Delta G > 2$kcal/mol. Error bars for the predicted set were calculated by randomly resampling 1428 residues from all mutations 1000 times and calculating the standard deviation.

characteristics as the previously described rat structure.

The Leu122 residue forms part of the obligate interaction interface between the two monomeric subunits. Each Leu122 forms non-covalent bonds with Phe114, Ile118, Ala170 and Leu122 from the opposite chain. The amino acids change from leucine to valine does not change the physico-chemical properties of the residue significantly. In fact, the conservation scores show that the two amino acids are similarly frequent at position 122 (Leu: 1.64, Val: 1.54). The likely reason for the severe effect of this mutation is a steric clash of the valine sidechain with serine at position 171 of the same chain. Even a small conformational change will affect the residue contacts Leu122 is involved in.

3.3.6.1 Griscelli syndrome, type 2 [OMIM: #607624]

Griscelli syndrome is a disease which features abnormal skin and hair pigmentation as well as, in some cases, immunodeficiency due to a lack of gammaglobulin and insufficient lymphocyte stimulation. Without bone marrow transplantation, the disease is usually fatal within the first years of life (Klein *et al.*, 1994). The type 2 form of Griscelli syndrome usually maps to the Rab-27A gene (Menasche *et al.*, 2000). The RAS domain of Rab-27A shares 46.8% sequence identity with the same domain in Ras-related protein Rab-3A from *Rattus norvegicus*. The crystal structure of Rab-3A interacting with Rabphilin-3A was solved by Ostermeier and Brunger (1999) [PDB: 1ZBD], see Figure 3.11. I found that a Trp73Gly mutation in Rab-27A affects a residue that is both highly conserved (Scores of 5.62 for Trp and -1.84 for Gly) and in the center of the interaction interface. There is strong evidence that Rab-27A interacts with Myophillin (Strom *et al.*, 2002). For these reasons the Trp73Gly mutation seems likely to affect vesicle transport by reducing affinity of Rab-27A to Myophilin.

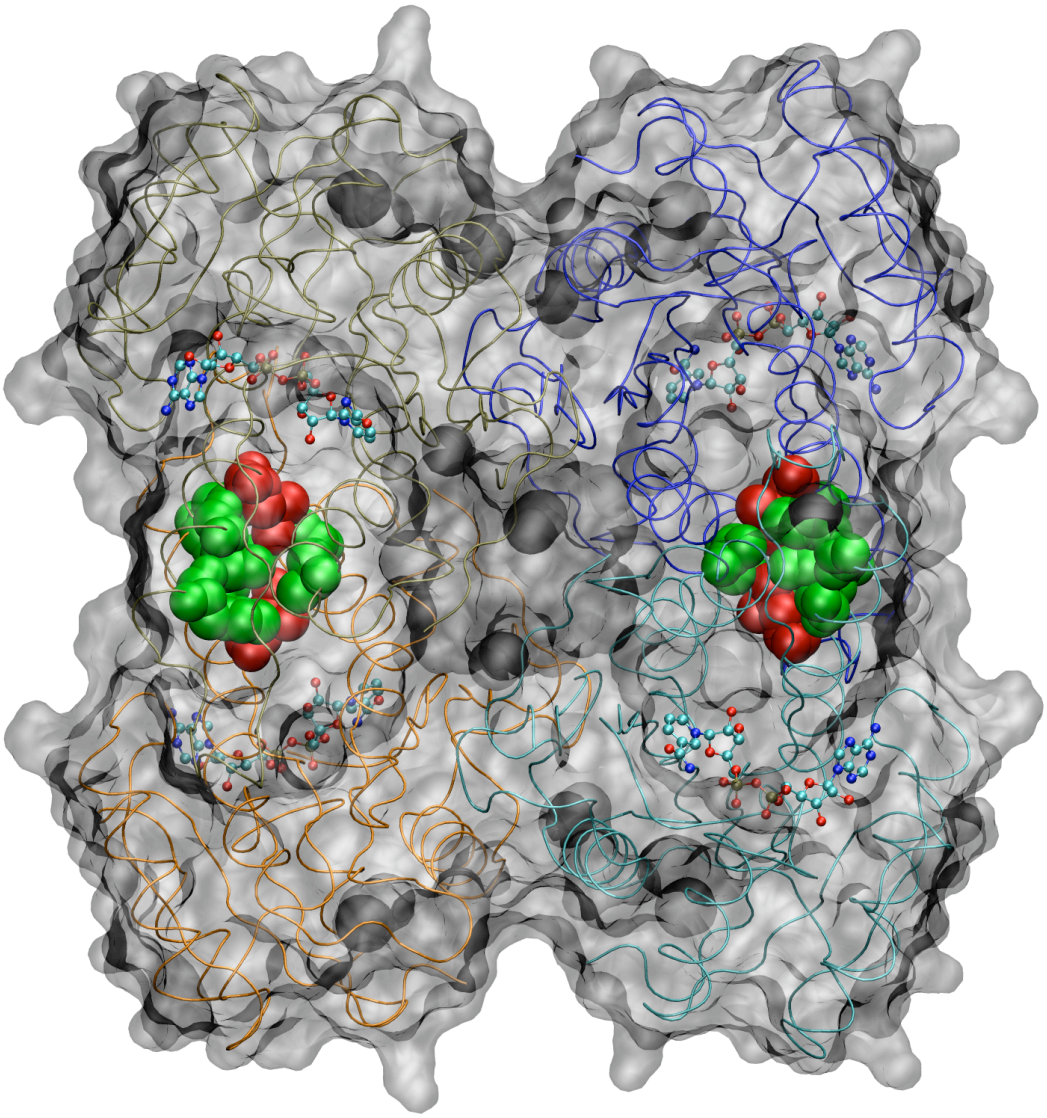


Figure 3.10: Structure of Rat brain 3-hydroxyacyl-CoA dehydrogenase with bound NADH [PDB: 1e3s]. The molecule is composed of 4 monomers, shown as different coloured ribbons. The Leu122 residue is highlighted in red with its binding partners shown in green. As Leu122 also interacts with the Leu122 of the other bound monomer, it is intuitive to assume that a mutation at this residue will affect binding.

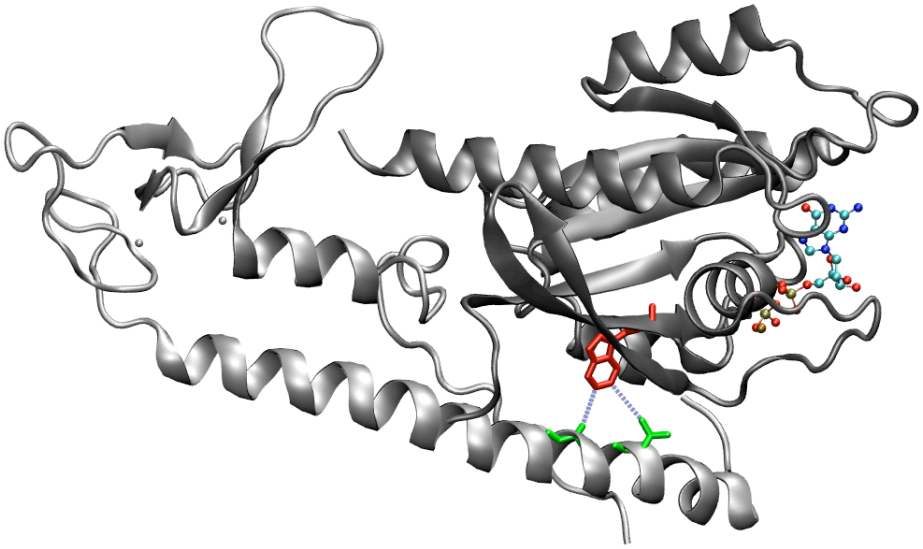


Figure 3.11: The small G protein Rab3A with bound GTP interacting with the effector domain of rabphilin-3A. The residue corresponding to the mutated Trp73 from human RAB27A, is highlighted in red, while the two residues in contact with it are coloured green.

3.3.6.2 ACTH deficiency [OMIM: #201400]

Adrenocorticotropin hormone (ACTH) deficiency is characterized by a marked decrease of the pituitary hormone ACTH and other steroids. Its symptoms include amongst others weight loss, anorexia and low blood pressure. Lamolet *et al.* (2001) identified a Ser128Phe mutation in the T-box transcription factor TBX19 that leads to a dominant loss of function phenotype [UniProt: O60806, VAR_018387]. The crystal structure of the homologous T-Box domain from the *Xenopus laevis* Brachyury transcription factor (Müller and Herrmann, 1997) (81% sequence identity to the human TBX19 protein; [PDB: 1XBR]) shows that this particular residue is at the core of the dimerization interface, see Figure 3.12. The mutation substitutes a small polar with a large aromatic side-chain. Accordingly, the residue features strong conservation, while Phe is very rare at this position (Scores of 3.31 and -1.78 for Ser and Phe respectively). Pulichino *et al.* (2003) report that the Ser128Phe mutation shows virtually no DNA binding affinity. I predict that this loss of affinity is due to a drop in binding free energy between monomer and DNA, as compared to the dimer.

3.3.6.3 Baller-Gerold Syndrome [OMIM: #218600]

Baller-Gerold syndrome is a rare congenital disease characterized by distinctive malformations of the skull and facial area as well as bones of the forearms and hands. The disease phenotypically overlaps with other disorders like Rothmund-Thomson syndrome or Saethre-Chotzen syndrome. Seto *et al.* (2001) reported a case of Baller-Gerold syndrome that also included features of Saethre-Chotzen syndrome. They identified an Ile to Val substitution at position 156 of the H-Twist protein as the causative mutation. Experimental studies using yeast-two-hybrid have reported the loss of H-Twist/E12 dimerization ability as a possible cause of Saethre-Chotzen syndrome (El Ghouzzi *et al.*, 2000).

The basic helix-loop-helix domain of H-Twist shares 45% sequence identity with

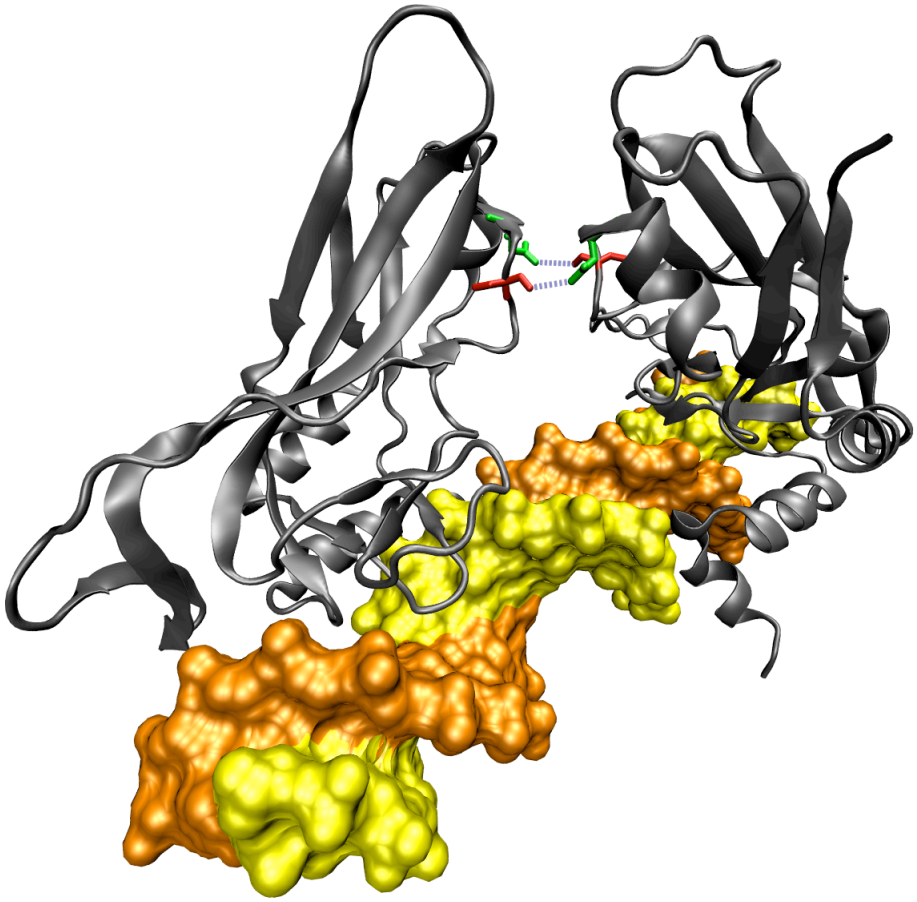


Figure 3.12: The crystal structure of a T-domain from *Xenopus laevis* bound to DNA. The residues highlighted in red are the mutated Ser128, with green residues representing the contact residues in the partner protein. Blue dashed lines show residue contacts.

the c-Myc transcription factor that was crystalized by Nair and Burley (2003), see Figure 3.13. The structure shows a dimer of c-Myc and Max bound to DNA. The c-Myc/Max dimerization is essential for the transcriptional regulation. The Ile156Val mutation is located at the core of the interaction interface. Although the Ile156Val mutation constitutes a biochemically similar substitution, reflected by the relatively high frequency of Val at this position in other helix-loop-helix proteins (prevalence scores 2.76 for Ile and 1.23 for Val), the change in volume could slightly change the interaction propensity. Correspondingly, the Ile156Val mutation causes a mild form of Baller-Gerold Syndrome.

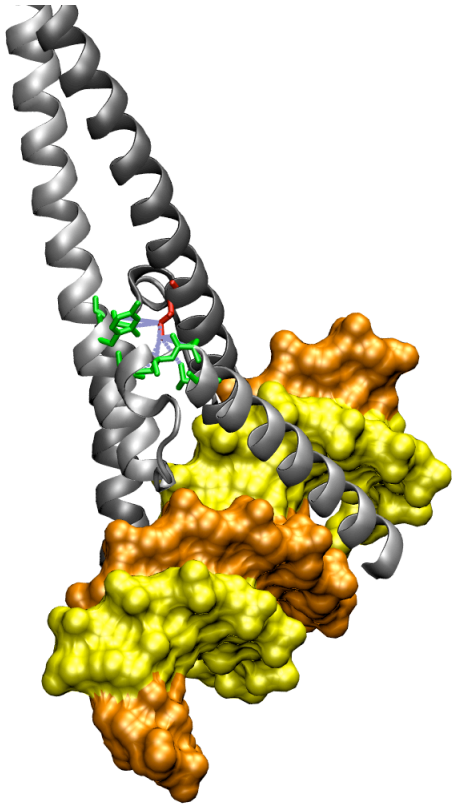


Figure 3.13: Both Myc-c and Max form a basic helix-loop-helix motif. They dimerize mainly through their extended helix II regions. The residue that corresponds to Ile156 in H-Twist is Ile550, shown in red. The residue sits at a key position of the interface, forming bonds with seven residues in Max, shown in green.

3.4 Discussion

3.4.1 Accuracy of interacting residue prediction

The wealth of information provided by protein structures of interacting proteins can be applied to evolutionary related sequences (Aloy and Russell, 2002). I developed an algorithm that identifies structurally corresponding residues in sequences that contain a domain which is homologous to a known structural interaction. Two distinct benchmarks provide evidence that the algorithm can identify interacting residues with reasonable accuracy. A cross-validation experiment showed that percentage identity between the predictions source and the target sequence is the best determinant for prediction quality. This finding fits the relationship between sequence similarity and similarity of interaction geometry described by Aloy *et al.* (2003).

A benchmark against a database of alanine scanning energetics experiments (ASEdb) reveals that the residue prevalence threshold is particularly suitable for identifying residues with a large change of binding energy upon mutation. The percentage identity threshold does not perform as favourably in the ASEdb benchmark as in the cross-validation experiments. It has to be considered in this context that alanine scanning experiments are often guided by homologous structures in order to restrict the number of mutated residues. Therefore, the true positive to true negative ratio decreases and the performance decreases. Conversely, the residue prevalence score improves because fewer false positives can be detected. As a consequence, I decided to employ a threshold that combines percentage identity and residue prevalence. In this way, any prediction should have be sufficiently likely to represent a real interaction, while the results are also enriched for structurally important residues.

3.4.2 Disease causing interacting residues occur frequently

Protein interactions can be the root cause of genetic pathologies, yet their significance for health and disease remained to be quantified. When I apply the prediction algorithm to all disease causing mutations from OMIM and UniProt, I retrieve a set of 1428 interaction-related mutations. This suggests that approximately 5% of mutations could have an effect on protein interactions. On the one hand, low structural coverage of *i*Pfam domains on protein interactions described in Chapter 2 could mean that this is a large underestimate. On the other hand, there are a number of potentially false positive predictions due to crystal packing which could result in an overestimation of the importance of interaction related mutations. Taking into account previous work on this matter (Ferrer-Costa *et al.*, 2002), I believe that an estimated fraction of 4 to 5% of interaction related mutations is well justified given the presented observations.

My curated list of interaction-related diseases further underlines that a variety of proteins are susceptible to mutations that alter protein interaction. The list provides examples to categorise mutations according to their functional and molecular properties. Namely, many interaction related mutations can lead to a gain of function, usually by losing the interface for an inhibitory protein or by aggregating uncontrollably and causing various forms of amyloidosis. Analysis of the amino-acid spectrum of residues in interaction-related diseases reveals marginal deviations from the distribution of amino-acids in all mutations. These properties could in the future be combined with other features to improve the accuracy of prediction algorithms.

Further mutagenesis and protein interaction experiments on selected examples from my predicted set could shed new light on the molecular mechanisms behind human genetic diseases. In turn, knowledge of more cases of interaction-related disease will help to improve the accuracy of prediction algorithms.

3.4.3 Enrichment for dominant mutations

In comparison to non-interaction related mutations, I observed an enrichment for dominant or co-dominant mutations in both the curated as well as in the predicted set. In GOF mutations, dominant inheritance is not surprising, but the high proportion (39%) of dominant LOF mutations is noteworthy. Dominant inheritance in LOF mutations can be explained by either *haploinsufficency* or dominant negative effects (Veitia, 2002).

Inhibiting one of the two alleles of a gene is likely to reduce the overall dosage level of functional protein. If this leads to a visible phenotype, the effect would be labelled as haploinsufficiency, *i.e.* a phenotype is caused by a shortage of functional protein.

Conversely, "dominant negative" refers to cases where a mutated allele actively inhibits other proteins which are otherwise functional. This effect is also often referred to as *interallelic complementation* in cases were the combination of two slightly differing alleles of a gene causes a change in the overall function of the protein.

For example, mutations of phenylalanine hydroxylase can lead to phenylketonuria (Leandro *et al.*, 2006) by inhibiting necessary conformational changes between monomers. In such cases where the protein function relies on the dynamic interactions between subunits, a mutation in one of the binding interfaces can actively inhibit the function of the other bound members of the complex. From my results, it is not clear whether hapoinsufficiency or interallelic complementation are the driving force behind the enrichment for dominant mutations amongst mutations in interaction interfaces. Detailed experimental analysis of dominant LOF mutations could reveal the relative importance of dominant negative effects compared to haploinsufficency.

In summary, however, the observation remains that interaction related mutations are more often dominant than expected by chance. Previous results also confirm that there is a relationship between dosage sensitivity and the protein interactions (Papp *et al.*, 2003). In the next chapter, I will further investigate this issue using a more global, genome-wide approach.

Chapter 4

Protein complexes, dosage sensitivity and copy-number variations

4.1 Introduction

In the previous chapter, I described the bias towards dominant mutations amongst mutations in protein interaction interfaces. As I mentioned there, dominance can be explained by haploinsufficiency or dominant negative effects. In either case, a 0.5 fold change in gene dosage of the functional (or mis-functional) protein causes a visible phenotype. It has been estimated that at least 20% of the entries in the OMIM database cause a phenotype as a *heterozygous* mutation (Kondrashov and Koonin, 2004). In contrast, the popular hypothesis explaining gene dominance formulated by Wright (1934) states that dominance is caused by "bottlenecks" in metabolic pathways and should generally be rare (Orr, 1991). Apparently, there are far more proteins that are *dosage sensitive* than can be explained by perturbations of biochemical pathways alone.

Papp *et al.* (2003) attempted to explain a similar observation made by Steinmetz *et al.* (2002) in *S. cerevisiae*. The latter had systematically created heterozygous deletion mutants for a range of genes orthologous to human disease-related genes. Papp *et al.* found that many haploinsufficient genes were members of protein complexes. They postulated that multi-protein complexes need to maintain the stoichiometry of their subunits to perform their biological function (the *balance hypothesis*). If this balance is disturbed, the function of the entire complex is disrupted. This conveniently explains the enrichment of haploinsufficiency amongst members of protein complexes. A range of other experiments also lend support to the balance hypothesis. It has been noted that expression levels of interacting proteins are highly co-ordinated (Jansen *et al.*, 2002), hinting that proportionality of subunit abundances is important. It has also been argued that tolerance towards polyploidization, compared to the sometimes severe effects of smaller duplications can be explained by conservation of stoichiometry (Aury *et al.*, 2006). The proposition in this case is that single gene duplications or deletions will cause a stronger negative fitness effect than copying all components of the complex, maintaining stoichiometric balance. Finally, it has been noted that highly-interacting proteins in higher organisms belong to small gene families (Yang *et al.*, 2003), which could be conveniently explained by a bias against duplication acting on multi-protein complexes.

There have been, however, several conflicting reports. Deutschbauer *et al.* (2005) performed a heterozygous deletion screen in *S. cerevisiae* that incorporated all open reading frames (ORFs) available for cloning at the time. They reported only 3% of genes to be haploinsufficient. While these genes were enriched for members of protein complexes, their subsequent overexpression did not cause a similar phenotype as their deletion. Unfortunately, it is not clear from the publication how the well described whole genome duplication that is characteristic for the *S. cerevisiae* lineage (Kellis *et al.*, 2004) affects these results. Subsequently, Sopko *et al.* (2006) systematically

induced gene overexpression for all ORFs in *S. cerevisiae*. The genes found to be toxic when overexpressed did not overlap with the haploinsufficient genes described by Deutschbauer *et al.*, and were not significantly enriched for protein complexes. This is in conflict with the dosage hypothesis in so far as it shows that deletion and duplication of the same gene do not usually lead to loss-of-function of the entire complex, as was initially suggested by Papp *et al.*. One important issue that has to be noted about the study by Sopko *et al.* is related to their experimental set-up. To assure that overexpression of the gene is controllable, they used an inducible promoter. They found that duplication sensitive genes were highly enriched for cell cycle proteins. A likely explanation for this bias is that the untimely expression of the proteins due to the non-physiological promoter is responsible for the negative fitness effect, rather than the actual dosage. The second important fact to consider is that single-cellular eukaryotes such as *S. cerevisiae* which are able to sustain both a haploid and diploid life-cycle, are likely to have different regulatory and dosage-compensatory mechanisms than multi-cellular organisms. One hint towards this difference is the increasing constraint on the number of paralogs of highly-interacting proteins in higher organisms, as described by Yang *et al.* (2003).

In light of the above points, Birchler *et al.* (2007) argued for a more elaborate concept to explain dosage sensitivity that they refer to as *regulatory balance*. Experiments in plants and later in *D. melanogaster* showed that duplications or deletions of some chromosomal regions cause no change in gene expression (Birchler, 1981; Devlin *et al.*, 1982), while variations of other genes causes up- or downregulation of various distal genes (Birchler *et al.*, 2001). One example referred to by Birchler *et al.* is *D. melanogaster* white eye colour controlled by the single gene *white*. Over the years, duplications of some and deletions of other genes (47 in total so far) have all been found to affect the expression of *white*. The majority of modulators of *white* act as negative regulators, *i.e.* a duplication of the regulator leads to lower expression of *white*. Birchler

et al. suggest that these regulators form a complex regulatory network where information transfer happens mostly through protein interactions, see for example Figure 4.1.

Considering these findings, it appears that there are multiple possible causes of dosage sensitivity, whereby deletion and duplication of the same gene do not necessarily lead to the same outcome:

- A limited number of enzymes are sensitive to low dosage because they are the rate limiting factor in a biochemical reaction.
- A range of proteins are likely to cause non-physiological binding or even agglomeration as a result of overexpression, as exemplified by susceptibility to early-onset Alzheimer's disease as a result of duplication of the APP locus (Lee and Lupski, 2006).
- Haploinsufficiency as well as duplication sensitivity are likely to affect the regulators controlling the balanced expression of a range of other proteins. As I described above, these proteins are in fact often complexes.

Dosage sensitivity and the concept of regulatory balance have important implications for gene duplicability and thus for the understanding of gene evolution. The widely accepted paradigm states that gene duplications can either create a non-functional pseudogene (*nonfunctionalization*) or relax selection constraints on one of the paralogous sequences, allowing it to diverge into related (*subfunctionalization*) or, in rare cases, new functions (*neofunctionalization*) (Prince and Pickett, 2002). Historically, it was assumed in this context that most genes can be duplicated without substantial negative fitness effects. It has been shown, however, that there are distinct differences between genes as to their duplicability (Veitia, 2005; Yang *et al.*, 2003) and that duplicated genes are in many cases still under negative selection (Kondrashov *et al.*, 2002; Lynch and Conery, 2000). How exactly these pressures on gene evolution are linked to dosage sensitivity and thereby to protein complexes is the focus of this chapter.

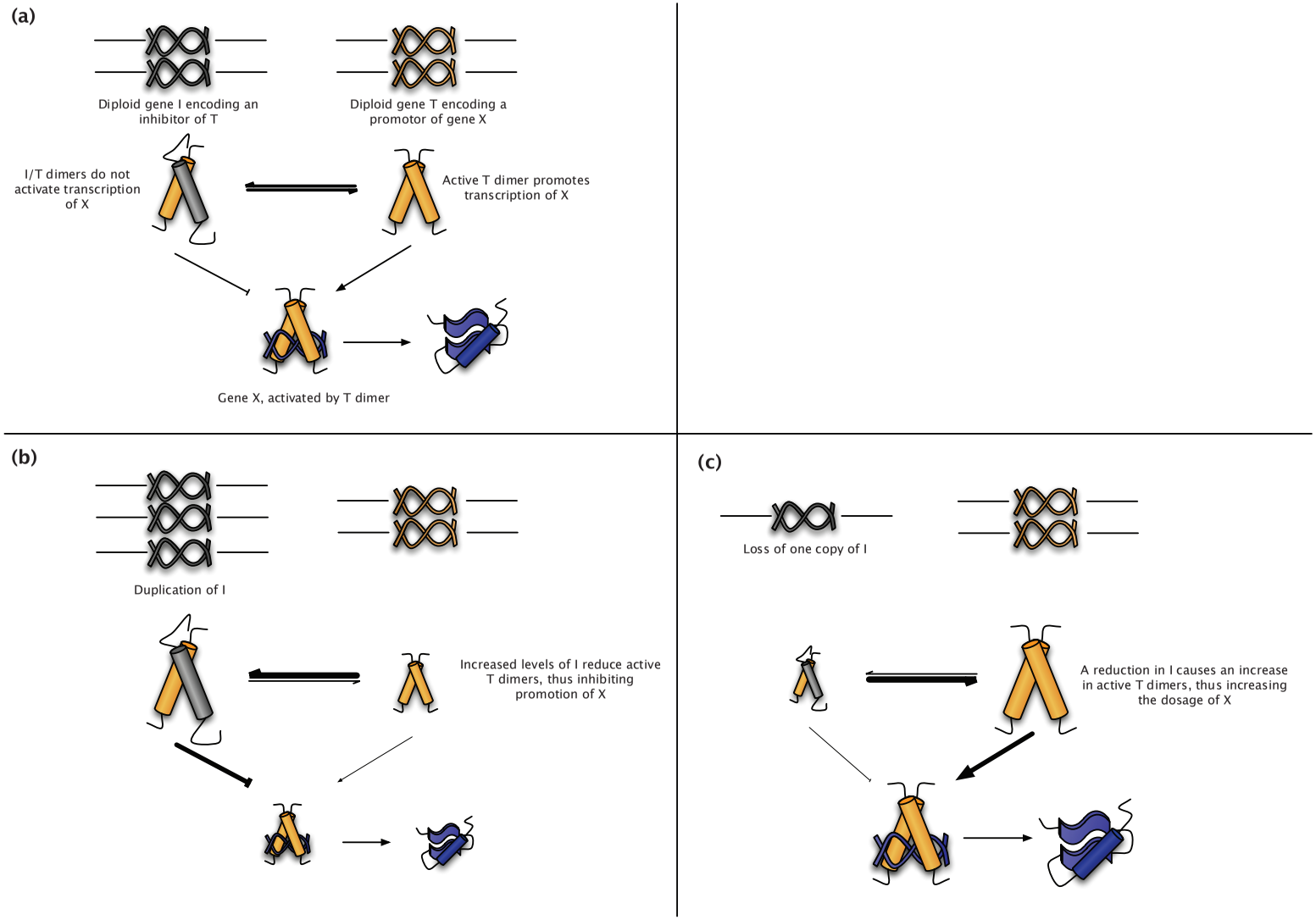


Figure 4.1: Schematic representation of a regulatory network controlling expression of gene X. Gene I encodes an inhibitor of transcription factor T. **(a)** In the normal, diploid state, I and T are in balance. **(b)** Duplication of I will cause lower than normal levels of X, denoted by smaller symbols and arrows. **(c)** Heterozygous deletion of I will lead to overactivation of T and thus to excess amounts of X, represented as large symbols. Figure adapted from Birchler *et al.* (2005).

It has been estimated that at least 2% of the human genome is affected by structural variations (Cooper *et al.*, 2007), such as inversions, small insertions/deletions or large copy-number variants (CNVs) (Conrad and Hurles, 2007). These sometimes large rearrangements may be seen as an important driving force of genome evolution. As a consequence, theories on gene evolution have to be re-evaluated in the context of such rapid and widespread large scale variation. Previous studies have already shown that the locations of CNVs and the function of genes inside CNV regions are biased (Cooper *et al.*, 2007; Nguyen *et al.*, 2006). CNVs are found more often in pericentromeric and subtelomeric regions, they overlap significantly with regions of segmental duplications and are more gene dense than the average for the genome. Genes within CNV regions are frequently involved in sensory perception and immune system activity, to a lesser extent in cell adhesion and in a number of cases signal transduction (Cooper *et al.*, 2007). Two theories have been postulated to explain this non-random distribution of CNVs. The *mutational hypothesis* states that most CNVs are in effect phenotypically neutral, but are carried by flanking genomic elements like ALU repeats which cause the bias in CNV distribution. The opposing theory could be called the *selection hypothesis*, stating that negative and positive selection shape the distribution of CNVs through the functional elements they encompass.

In this work, I use gene expression and copy-number variation data to study the relationship between protein complexes, dosage sensitivity and recent gene evolution in the human population. Firstly, I show that changes in gene copy number have a weak but measurable effect on gene expression. Next, I describe how genes involved in protein complexes are enriched for known dosage sensitive genes and exhibit substantially lower expressional noise than other genes. Consequentially, I observe that dosage sensitive genes tend to be underrepresented in CNV regions. Given these functional and positional biases on genes in CNV regions, I hypothesise that the regulatory balance of dosage sensitive genes exerts negative selective pressure on chromosomal structural

variations.

4.2 Methods

A wide range of diverse sources of data were combined in order to perform the analyses in this chapter. In the following paragraphs, I describe the provenance and composition of these different datasets. When no web URL is given, the data was extracted from supplementary materials files provided with the referenced publication.

4.2.1 Gene identifiers

A common problem when combining several independent data sets is inconsistencies in naming conventions. To assure that all gene identifiers were consistent, all data sets were mapped to the most recent HUGO Gene Nomenclature Committee (HGNC) identifiers in March 2008 (Bruford *et al.*, 2008). In case a gene name did not correspond to a primary gene symbol in HGNC, the HGNC *previous symbols* column was searched for an exact match, followed by a search in *aliases*. If no exact match could be found, the gene was removed from the set and not included in any further analysis.

4.2.2 Mammalian protein complexes

The CORUM database (Ruepp *et al.*, 2008) is a manually annotated resource, containing, at the time of writing, 1679 protein complexes from 10 mammalian species, with a strong focus on human. Entries are based on individual publications, not including high-throughput experiments. Table 4.1 lists Gene Ontology annotations for which CORUM deviates significantly from the rest of the genome. CORUM is enriched for nuclear proteins and contains a large number of transcriptional regulators. Conversely, extracellular and membrane proteins are underrepresented in the dataset. Figure 4.2 visually conveys an idea of the size distribution of this network of human complexes, as well

as reflecting its highly interconnected nature. Relationships for 2080 proteins in 1109 human complexes were downloaded from `http://mips.gsf.de/genre/proj/corum` on the 29th January 2008. 2028 proteins could be mapped to 1975 HGNC identifiers. Genomic coordinates for these gene identifiers were retrieved from Ensembl (v49) (`http://www.ensembl.org`) *via* BioMart.

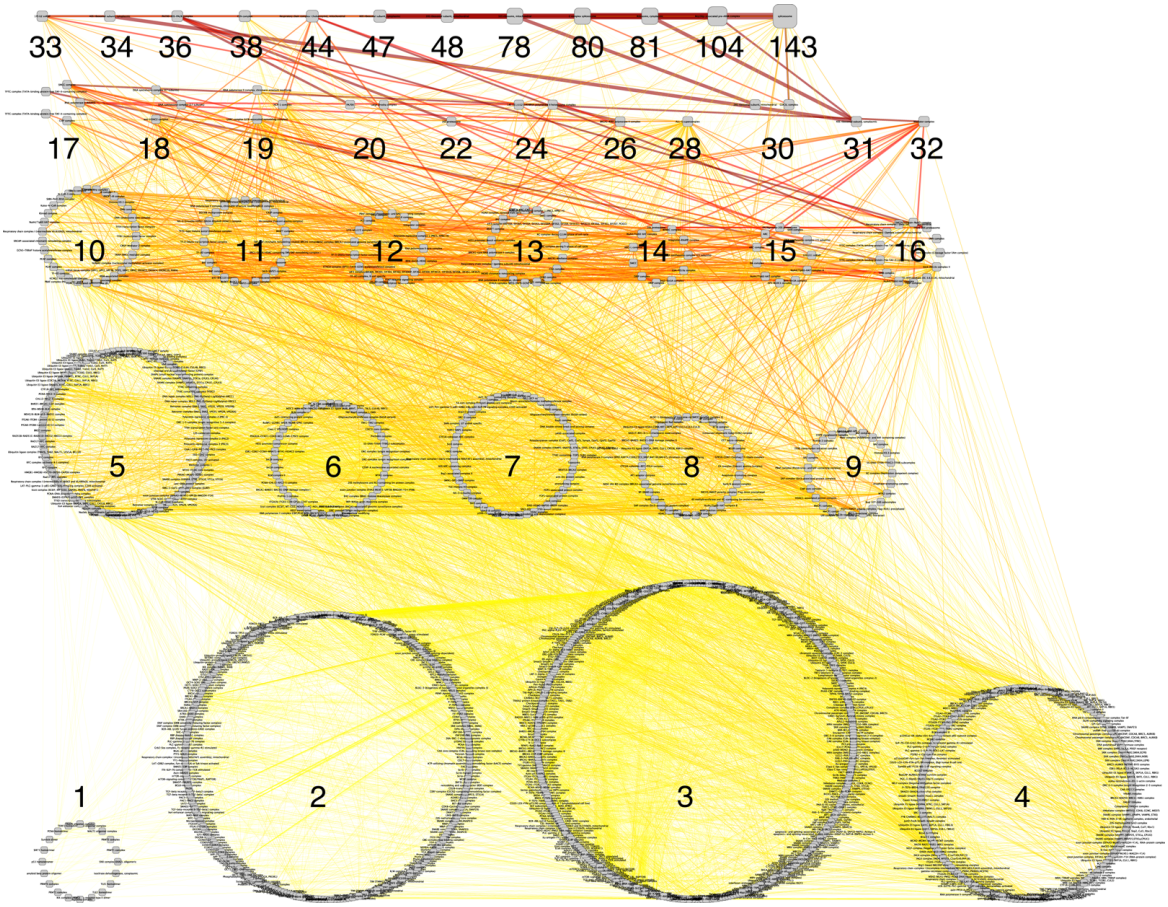


Figure 4.2: A network representation of the CORUM database. Nodes represent complexes and are ordered by number of unique components (shown as number next to groups). Edges denote shared components between complexes. The number of shared components is reflected in the colour (from yellow (few) to red (many) shared components) as well as in the line width. The large, highly overlapping complexes in the first row are mainly modules of the ribosome (6 out of 12) and spliceosome (3 out of 12). Other large complexes include RNA polymerase, respiratory chain complex and the proteasome. The group of complexes with only one member are homo-multimers.

Table 4.1: Composition of the CORUM database. Underrepresented terms are set in **bold font**. P-Values were calculated using Fisher's Exact Test, see Methods.

GO-Slim Term	Number of CORUM genes	P-Value
protein binding	1348	$1.78 \cdot 10^{-210}$
nucleus	1058	$3.73 \cdot 10^{-207}$
macromolecule metabolic process	1321	$1.59 \cdot 10^{-205}$
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	852	$4.52 \cdot 10^{-148}$
nucleic acid binding	708	$5.73 \cdot 10^{-86}$
cytoplasm	933	$2.72 \cdot 10^{-62}$
regulation of biological process	722	$1.24 \cdot 10^{-51}$
chromosome	168	$7.95 \cdot 10^{-46}$
structural molecule activity	227	$5.51 \cdot 10^{-38}$
transcription regulator activity	301	$1.63 \cdot 10^{-30}$
biosynthetic process	279	$5.37 \cdot 10^{-26}$
helicase activity	53	$1.14 \cdot 10^{-15}$
cell death	146	$1.12 \cdot 10^{-12}$
protein transporter activity	45	$3.32 \cdot 10^{-11}$
response to stimulus	378	$3.42 \cdot 10^{-08}$
translation regulator activity	34	$2.29 \cdot 10^{-06}$
cell differentiation	232	$1.54 \cdot 10^{-05}$
extracellular region	77	$\mathbf{1.94 \cdot 10^{-06}}$
membrane	532	$\mathbf{3.35 \cdot 10^{-15}}$

4.2.3 Interaction and complex data

As an alternative to the manually compiled set of complexes in CORUM, an independent set of putative complexes was computationally derived from high-throughput protein interaction experiments by identifying highly connected clusters of proteins in an extended network of human protein interactions (Krogan *et al.*, 2006). Interaction data for three recent high-throughput studies (Ewing *et al.*, 2007; Rual *et al.*, 2005; Stelzl *et al.*, 2005) were retrieved from IntAct (Kerrien *et al.*, 2007) and subsequently

merged into a single network. As for CORUM, UniProt identifiers were mapped to HGNC identifiers to ensure consistency. This was achieved by extracting the HGNC annotations in the "cross-references" section of the UniProt flat-files. Clustering analysis was performed using the Markov clustering tool `mcl` (van Dongen, 2000) (parameter $I = 3.0$). The "alternative complex set" was defined as containing all clusters with more than three components (2325 unique genes).

4.2.4 Set of dosage sensitive genes

Dosage sensitive genes were extracted from the annotations of the Baylor College of Medicine Medical Genetics Laboratory 105k diagnostic Chromosomal Microarray (version 7), available at `http://www.bcm.edu/geneticlabs/cma/`. This post-natal screening tool comprises a manually compiled set of 146 genes (after mapping to HGNC) known to be sensitive to chromosomal imbalances (Cheung *et al.*, 2005). A complete list of the genes and the associated diseases can be found in Table H.1.

A separate set of genes overexpressed in cancer tissue was also used (Axelsen *et al.*, 2007). The dataset contains 2362 genes which are at least 4-fold overexpressed in brain (astrocytoma and glioblastoma), breast, colon, endometrium, kidney, liver, lung, ovary, prostate, skin, and thyroid cancers compared to healthy tissue of the same type.

4.2.5 Expression profiles

Gene expression can be measured on a large scale using expression arrays. Stranger *et al.* (2007) performed gene expression analysis on Eppstein-Barr virus transformed lymphoblast cell lines from each of the HapMap individuals. Gene expression was quantified using high-throughput human whole-genome expression arrays designed by Illumina (Kuhn *et al.*, 2004). These arrays consist of $\approx$ 48000 bead types, where each bead consists of several hundred thousand copies of a gene specific oligonucleotide probe. After RNA was extracted from the cell lines, it was carefully amplified and

labelled with Biotin-16-UTP. After hybridisation to the array, Cy3-streptavidin was applied to the array which binds to Biotin and subsequently allows the measurement of luminescence intensities for each bead type in a specially designed scanner. Kuhn *et al.* showed in a benchmark experiment that luminescence intensities are directly proportional to the expression strength within a defined dynamic range (Limit of Detection: ≈ 0.13pM, dynamic range: ≈ 3.2-fold). Each bead type is also replicated several times on the array, thus providing robustness and redundancy for quality control. Subsequent to data readout, the raw intensities for each redundant bead type were summarised by proprietary software provided by Illumina. Stranger *et al.* performed 4 replicate hybridisations per cell line, the results of which were summarised on a log scale using a quantile normalisation method across replicates of a single individual, followed by a median normalisation method across all 270 individuals. The resulting data, consisting of a matrix of gene expression values of 47293 probes over 270 individuals, were downloaded from `http://www.sanger.ac.uk/humgen/genevar/`.

Due to the sensitivity and dynamic range limitations of the Illumina WG6 expression arrays used by Stranger *et al.*, there is a correlation between detectable expression variation and total expression strength for genes with low overall expression, or no expression at all. Notably, there is a cluster of genes with both low detected expression and markedly lower coefficients of variation (CV, defined as the standard deviation of expression between individuals per gene, normalised to the mean absolute expression level) than the majority of genes, plotted in grey in Figure 4.3. These genes may be distinguished from the remaining genes by their lower absolute variation, that is the standard deviation between individuals before normalisation to the expression mean. In total, 6440 genes with an absolute population standard deviation ≤ 7 were removed from the dataset, as they are likely to be expressed below the confident detection threshold or not to be expressed at all.

A second set of expression data for 44760 probes applied to samples from 79 different

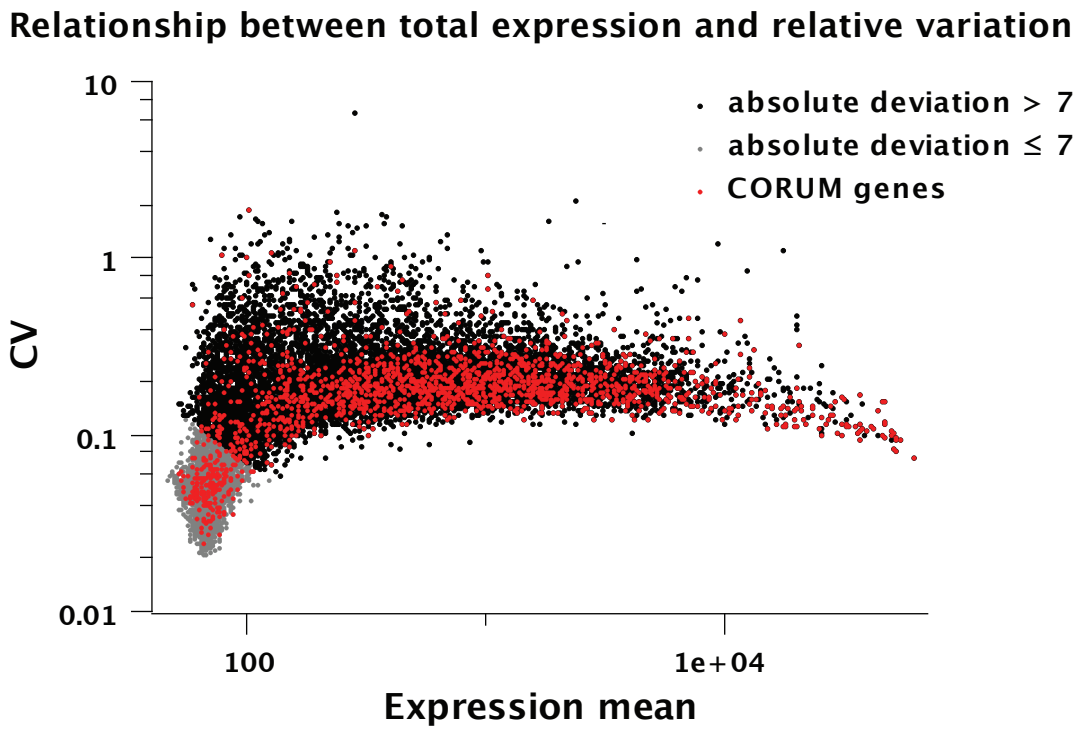


Figure 4.3: Coefficients of gene expression variation (CV) relative to absolute expression level. The measurable variation in gene expression is limited by the sensitivity of the employed array technology. Genes which are expressed at extremely low levels, or not expressed at all, cluster in the low expression/low CV region. Shown in grey are genes which were excluded from further calculations (standard deviation ≤ 7).

tissue types were provided by GNF SymAtlas (Su *et al.*, 2004) (`http://symatlas.gnf.org`). For the latter, different Affymetrix expression arrays were employed, raw results of which were normalised using global median scaling.

Probe identifiers for both data sets were mapped to HGNC gene names through Ensembl BioMart. Probes which could not be mapped to a gene name were exluded from further analysis. The resulting matrices contained expression data for 17122 genes (HapMap set) and 15012 genes (tissue set), respectively.

4.2.6 Correlation computation

As a measure of correlation between expression levels of two genes in different tissues/individuals, the Pearson product-moment correlation coefficient was employed. For two vectors x and y representing genes with n expression levels, the correlation r_{xy} is given by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{4.1}$$

where $\overline{x}$ and $\overline{y}$ are the means and s_x and s_y are the standard deviations of x and y, respectively. For complexes with more than 2 components, correlations for all $n(n-1)/2$ combinations of gene pairs were averaged.

4.2.7 Copy-number variations

Chromosomal locations of variations relative to the NCBI36 human genome assembly were downloaded from the Database of Genomic Variants (DGV) (Iafrate *et al.*, 2004): `http://projects.tcag.ca/variation/`. This data also contains information on number of individuals and gain/loss annotation per CNV. CNV locations and whole genome tiling-path (WGTP) array hybridisation values for each HapMap individual were downloaded from `http://www.sanger.ac.uk/humgen/cnv/data`. The distribution of CNVs on selected human chromosomes is shown in Figure 4.4.

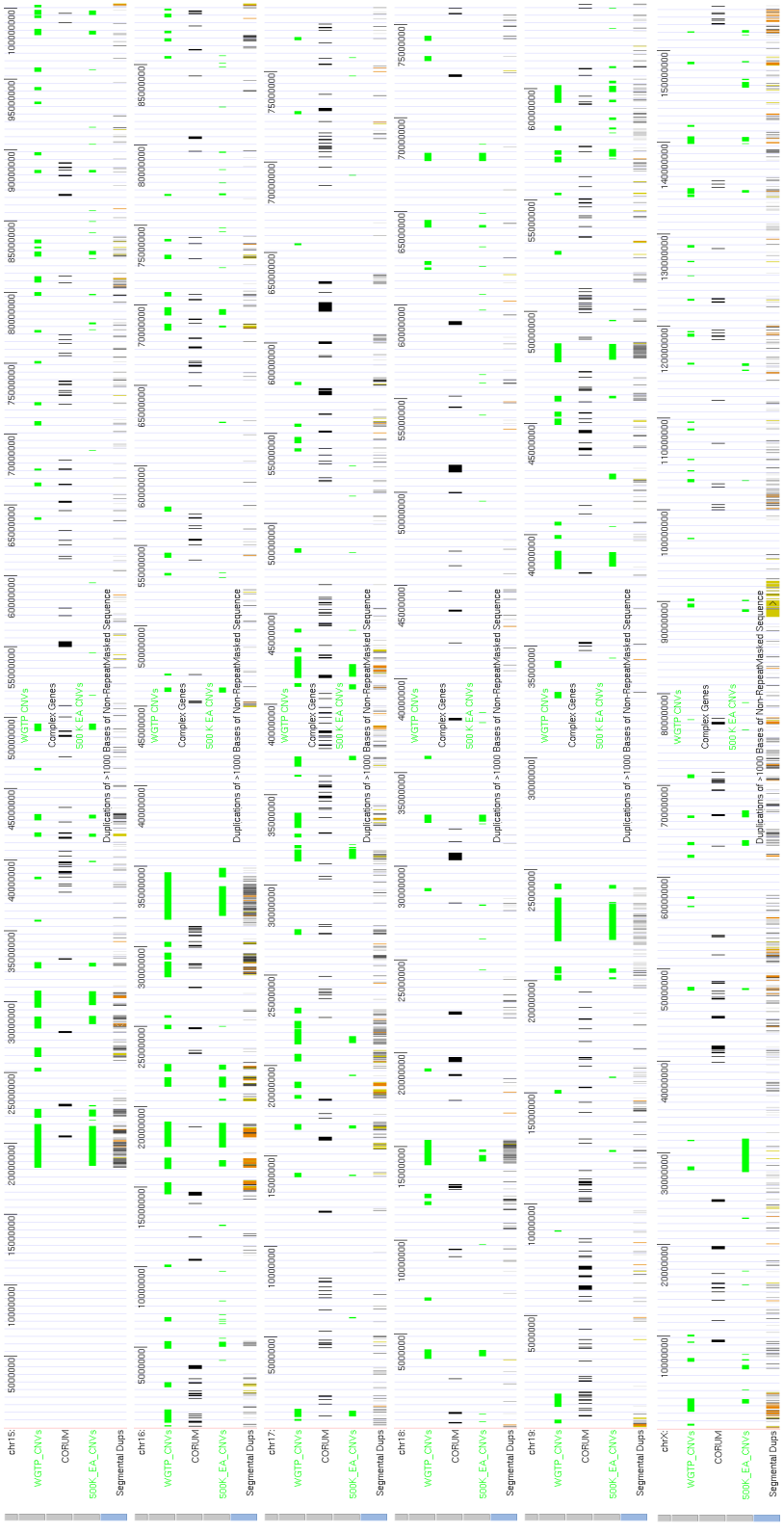


Figure 4.4: Position of CORUM genes (black), copy-number variants (green) and segmental duplications (shades of orange) on 5 autosomes and the X chromosome from human. CNVs from Redon *et al.* were derived by two different methods: WGTP and 500k array, which are shown separately. Graphics generated with the UCSC Genome Browser (Kent *et al.*, 2002)

4.2.8 Segmental duplications

Human segmental duplications of $\geq 90\%$ sequence identity and ≥ 1 kilobase length were provided by the segmental duplication database (She *et al.*, 2004) (`http://humanparalogy.gs.washington.edu`).

4.2.9 Gene Ontology analysis

181651 Gene Ontology (GO) annotations for 34591 human UniProt entries were provided by the GOA project (Camon *et al.*, 2004), available at `http://www.ebi.ac.uk/GOA/`. UniProt enries were mapped to HGNC identifiers through BioMart, resulting in 16213 annotated HGNC gene identifiers. There were 6775 unique GO terms in the full GOA dataset. The complexity of this hierarchical data structure was reduced by mapping GO terms to 64 GO-slim categories as defined by the GOA project themselves (`ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/`).

4.2.10 Identification of paralogs

In-species paralogs for 10755 HGNC gene identifiers were downloaded from Ensembl Compara via BioMart. The paralog prediction uses automatically generated phylogenetic trees of all species in the Ensembl database. According to the Ensembl compara help website (`http://www.ensembl.org/info/about/docs/compara/homology_method.html`), the algorithm to identify orthologs comprises the following steps:

1. Align all pairs of full-length protein sequences of the longest transcript of two genes from two species using WUBlastp and subsequent Smith-Waterman.

2. Cluster genes by single-linkage clustering according to Best Reciprocal Hits and Best Score Ratio.

3. Create a multiple sequence alignment (MSA) for each cluster using MUSCLE (Edgar, 2004).

4. For each MSA, calculate a phylogenetic tree using TreeBeST (`http://treesoft.sourceforge.net/treebest.shtml`) and infer orthology and paralogy. TreeBeST in this case combines 5 tree building methods (maximum likelihood on protein and codon sequences *via* `phyml` (Guindon and Gascuel, 2003) and neighbour-joining on p-distance as well as dN and dS distances) and calculates a consensus tree.

4.2.11 Analysis of selection pressure

dN/dS values for human genes relative to mouse orthologs were acquired from Ensembl via BioMart. The calculation of dN/dS values is part of the automatic gene tree generation described above: dN/dS values are generated by `codeml` (model=0, NSsites=0) from the PAML package (Yang, 1997) for all genes from closely related species after the initial tree generation. In this analysis, only genes with a single unique ortholog in mouse were used in the analyses.

4.2.12 P-Values

Statistical significance of overlaps between gene sets was computed with Fisher's exact test (FET). The Mann-Whitney-U test (MWU) was employed to determine significance of differences between two distributions. In cases of multiple testing, Bonferroni correction was applied. All calculations were performed in R (R Development Core Team, 2006). Significance of differences in dN/dS ratios was calculated by random resampling: For the null hypothesis, 1000 sets of genes with identical size as the test set were each created by randomly drawing without replacement from the complete gene set. P-Values were calculated as the probability of observing a result at least as extreme, given the normally distributed null model derived from the resampling.

4.3 Results

In order to investigate the relationship between copy-number state, protein complexes and dosage, I need to assert several preconditions. Firstly, I investigate the impact of copy-number change on gene expression. Secondly, I analyse the relationship between protein interactions and dosage sensitivity. Finally, I combine these points to describe the effects of dosage sensitivity of protein complexes on the evolution of chromosomal structural variations.

4.3.1 Effects of CNVs on gene expression

Association studies (Stranger *et al.*, 2007) have shown both *cis* and *trans* effects of copy-number variations (CNVs) on genes. Stranger *et al.* also measured the relative contribution of single-nucleotide polymorphisms (SNPs) and CNVs on the observed variation in gene expression. They report that 83.6% of variation can be attributed to SNPs, whereas 17.7% of variation is associated with CNVs. However, the study was designed to identify associations between all genes and CNVs within a 2 million base-pair (MB) window simultaneously and thus had to use stringent multiple-testing correction. While Stranger *et al.* report 238 genes to be associated with a CNV within a 2MB window, it is not immediately clear what immediate effects CNVs have on contained genes, and whether there is a distinguishable effect between deletion and duplication polymorphisms.

I therefore focused my attention on the relationship between copy-number variations and gene dosage. I combined gene expression data derived from lymphoblast cell lines of 270 HapMap individuals (Stranger *et al.*, 2007) with the CNV dataset of Redon *et al.* (2006) on the same individuals.

I find that duplications and deletions have distinguishable profiles of expression ratios, see Figure 4.5. The expression ratio is defined as the average expression of a gene

in individuals with a CNV phenotype, divided by the average expression in unaffected individuals. Assuming a simple linear relationship between copy-number and expression level, one would expect a distribution with peaks at 0.5, 1 and 1.5, corresponding to a heterozygous deletion, balanced expression and heterozygous duplication, respectively. The observed distribution shown in Figure 4.5 reflects a more complex relationship.

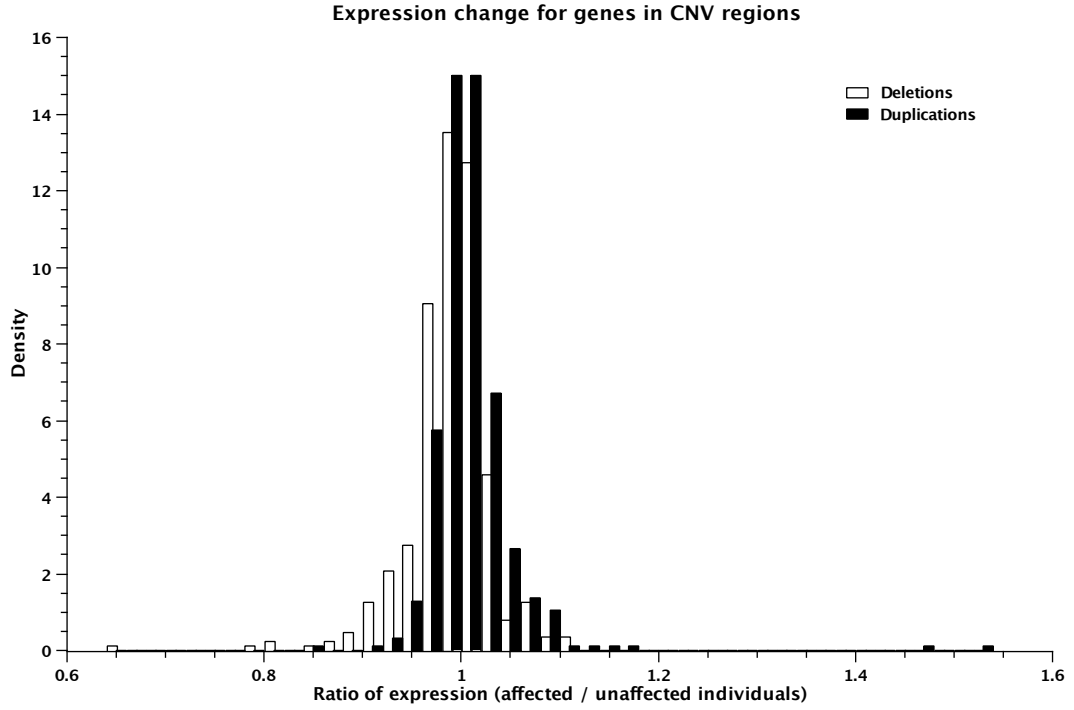


Figure 4.5: Difference between deletion (white) and duplication (black) variations in HapMap individuals. The histograms show the ratio of average expression levels between affected and unaffected individuals for all genes inside a copy number varied region. The shift between the two distributions is significantly larger than would be expected by chance (MWU: $P = 1.22 \cdot 10^{-11}$).

The magnitude of the expression difference between CNV and wild type individuals is smaller and more continuous than expected. However, the location shift between the two distributions is highly significant (MWU: $P = 1.22 \cdot 10^{-11}$). This indicates that deletions reduce gene expression, while duplications tend to increase expression.

As mentioned in the Methods, sensitivity and dynamic range of the expression arrays could partly account for the observed noise, but I did not find a correlation between absolute gene expression level and ratio of expression difference for genes overlapping CNV regions (Figure 4.6).

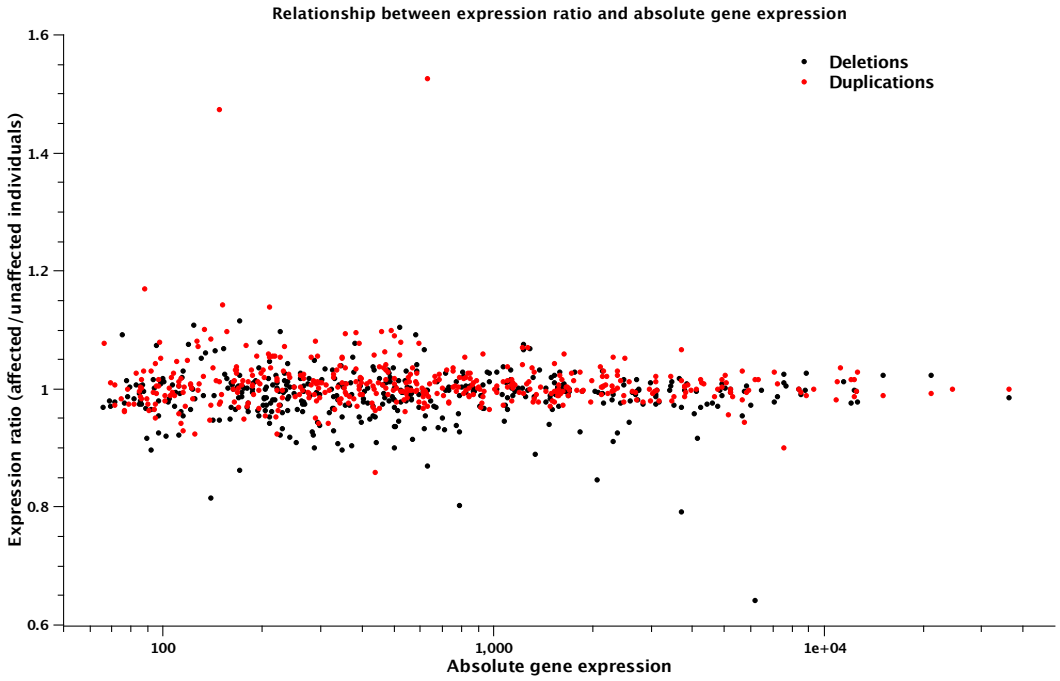


Figure 4.6: Relationship between effect of CNV on gene expression and absolute expression levels. The horizontal distribution suggests that there is no discernible correlation between absolute gene expression and expression ratio. A positive or negative correlation between absolute detection level and the fold expression change between affected and unaffected individuals could indicate a measurement-sensitivity induced bias, but within the analysed data no such relationship is detected.

The expression ratio distribution reflects a summary over a wide range of individuals. To elucidate the effects of CNVs on gene expression on a per-individual basis, I plotted the logarithm of hybridisation strength on the genomic hybridization arrays relative to the reference individual ($\log_2^H$) against the logarithm of expression, relative to the reference individual ($\log_2^E$). As a positive control, I compared two X-chromosomal

genes, one being inactivated (L1CAM, Figure 4.7a), the other being known to escape X-inactivation (UTX, Figure 4.7b). The latter exhibits a marked increase in expression in female individuals relative to the (male) reference individual. In contrast, L1CAM maintains equivalent expression in males and females levels due to inactivation of one gene copy in females.

I found 94 gene duplications and 98 gene deletions where the average log_2^H and log_2^E are at least one standard deviation below (deletions) or above (duplications) the mean of the unaffected individuals. Figures 4.7c and 4.7d show two examples of genes inside frequent CNVs exhibiting induced dosage effects. Deletions and duplications have clearly distinguishable expression levels. Notably, though, the expression ratios of the deletion/duplication individuals overlap with the expression ratios of unaffected individuals. In other words, CNVs only partly account for the differences in expression between individuals, while a large portion of the variance must stem from other sources. Figures 4.7e and 4.7f show two examples of rarer CNVs which also show a clear deviation of log_2^H and log_2^E relative to the majority of unaffected individuals.

Notably, several individuals were not called as CNVs, despite similar log_2^H and log_2^E ratios in the analysed region as the identified CNV individuals. These putative false negatives will reduce the magnitude of expression ratios between CNV and unaffected individuals. Summarising these individual effects leads to the conclusion that duplications and deletions have a measurable effect on gene expression, even though they are just one source of expression variation amongst others.

4.3.2 Limited expressional noise of protein-complex genes

It has previously been reported that expression levels of proteins within a complex are significantly more correlated across tissue types than would be expected by chance (Hahn *et al.*, 2005; Jansen *et al.*, 2002). Using both the expression from HapMap individuals mentioned above as well as a tissue-specific gene expression dataset, I verify

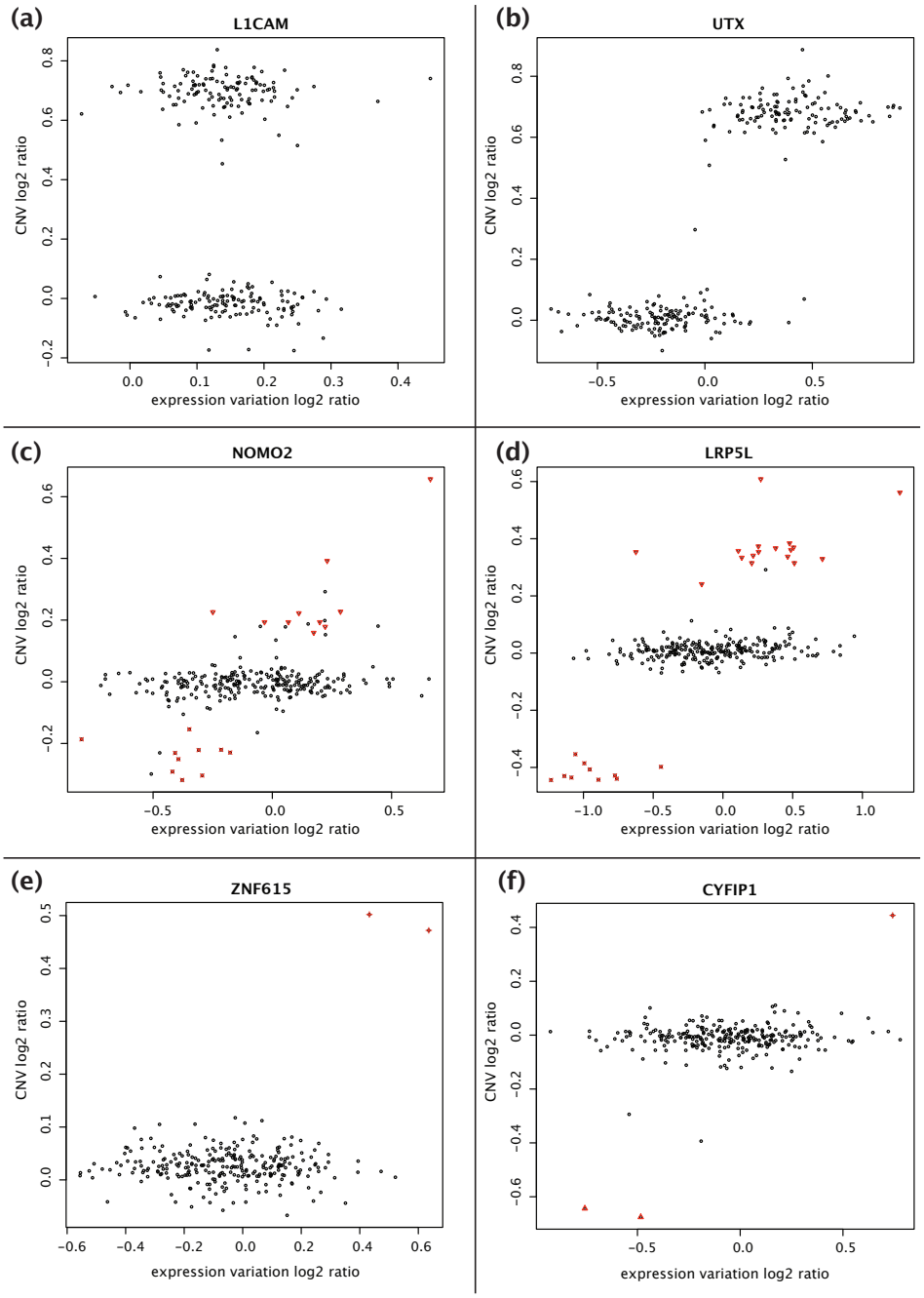


Figure 4.7: Ratio of WGTP array hybridisation intensity over relative expression level for four example genes. **(a)** L1CAM and **(b)** UTX. The increase in expression as a result of the copy-number increase in females is clearly visible for UTX which is known to escape X-inactivation. **(c)** and **(d)** Examples of autosomal genes with common CNV polymorphisms. Red crosses denote individuals in which a deletion phenotype has been called by Redon *et al.*, red triangles denote duplications. The plot highlights several potential false negatives with similar expression and hybridisation strength as the called deletions/duplications. Non-CNV related expression variation is substantial. **(e)** and **(f)** Examples of rare CNV genotypes with significant expression change.

that members of complexes from the CORUM database exhibit increased expression correlation (Figure 4.8).

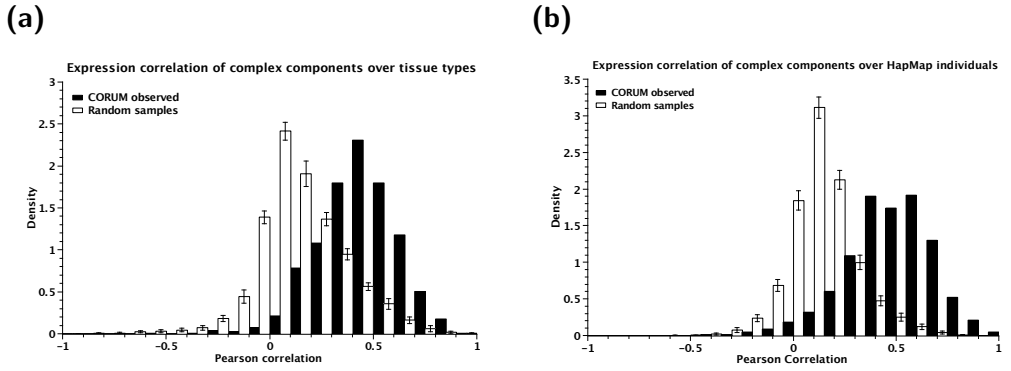


Figure 4.8: Distribution of average Pearson correlation coefficients between all members of known protein complexes as defined in CORUM (black), and randomly sampled proteins (white, N=100). **(a)** Expression intensities from 79 tissue types of different individuals. **(b)** Expression intensities from lymphoblast cell lines of 270 HapMap individuals.

In addition to that, the HapMap expression data allow me to perform a direct comparison of expression levels between individuals. I calculated coefficients of variation (CV), defined as the standard deviation of expression between individuals per gene, normalised to the mean absolute expression level. These values represent a dimensionless magnitude of variation for each gene. The CVs are significantly lower for CORUM genes than for the rest of the genome (MWU: $P = 2.67 \cdot 10^{-10}$), see Figure 4.9a/b. Interestingly, the average CV of genes within one complex decreases with the size of the complex, as shown in Figure 4.9c. This is independent of the mean absolute expression per gene, as shown in Figure 4.9d. I asserted that this effect is not a sampling artefact: When splitting all CORUM genes into sets with complexes of size ≥ 10 and size < 10 and comparing the distribution of CVs, it emerges that small complexes possess higher CVs (MWU: $P < 2.2 \cdot 10{-}16$). These results indicate that members of protein complexes are not just more likely to maintain relative expression levels between tissue types, but they are also more restricted as to their expression variation between

individuals within the same tissue.

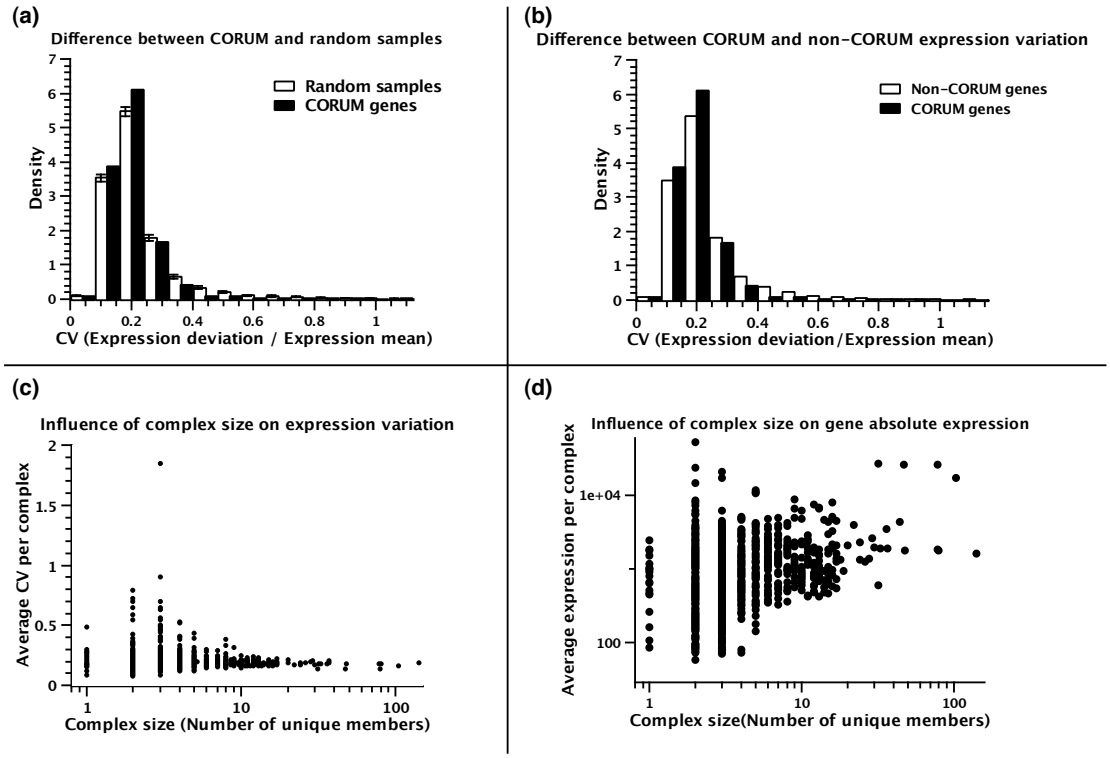


Figure 4.9: Coefficients of gene expression variation (CV) vary between CORUM and non-CORUM genes. **(a)** CORUM genes have significantly lower CVs than random sets of genes. **(b)** CORUM genes have significantly lower CVs than non-CORUM genes. Outliers beyond 1.4 are not shown. **(c)** Large CORUM complexes exhibit lower average CVs of their members. **(d)** Low absolute expression is not the reason for the lower noise in large complexes: mean absolute expression of large complexes is above average.

CORUM is a manually curated data source and thus prone to ascertainment bias. To ensure that these results are not biased by the composition of CORUM, I generated a separate dataset of putative protein complexes extracted from several high-throughput protein interaction detection experiments (see Section 4.2.3). The clusters represent an alternative set of "complexes" composed of 2325 proteins, 505 of which are also contained in CORUM. The CV distribution difference between these highly interacting proteins and the rest of the genome is also skewed towards lower CVs ($P = 7.0 \cdot 10^{-3}$).

This suggests that highly connected proteins in general avoid imbalances in protein expression.

Is there evidence that tight control of gene expression is actually relevant for human disease? Axelsen *et al.* (2007) compiled a list of 2362 genes which are overexpressed in various cancer tissues (see Section 4.2.4). I tested whether these cancer related genes are enriched for dosage sensitive genes, under the assumption that dosage sensitive genes are more likely to be causal in these diseases. In fact, I find that CORUM genes are overrepresented in these cancer related genes (356 genes, FET: $P = 6.56 \cdot 10^{-13}$). The fact that the tight regulation of expression of CORUM genes is disturbed in cancer tissue provides an interesting link between cancer, protein complexes and dosage sensitivity.

4.3.3 Dosage sensitive genes and CNVs

I have so far assembled evidence that protein complexes seem to be under constraint to maintain their relative expression levels and show limited expression variability between individuals. For the further analysis of dosage sensitivity, I also used an independently assembled set of 146 genes with known dosage-related disease phenotypes (see Section 4.2.4). There is a significant overlap between CORUM and this set of dosage sensitive genes (32 genes, FET: $P = 1.2 \cdot 10^{-5}$), further supporting the link between dosage sensitivity and protein complexes.

As previously stated, I found that CNVs can affect the expression levels of genes they contain. I therefore hypothesised that a CNV that encompasses a gene which is part of a protein complex will be more likely to have a negative effect on fitness. As the Redon *et al.* CNV data were derived from healthy individuals, I expect that genes encoding protein complexes will be underrepresented in CNV regions.

Out of 18534 protein coding genes for which both genomic locations and a unique gene name could be retrieved, 2311 genes are fully inside a CNV region. From 1975 proteins in the CORUM database, only 165 are found in a CNV region, significantly

fewer than one would expect by chance (FET: $P = 3.5 \cdot 10^{-10}$). The set of automatically clustered complexes were also underrepresented in CNV regions (256 out of 2325 genes, $P = 0.012$). Lastly, both the set of 146 dosage sensitive genes (8 genes inside CNV, $P = 4.7 \cdot 10^{-3}$) as well as the 2362 genes overexpressed in cancer (246 genes inside CNV, $P = 5.82 \cdot 10^{-4}$) are unlikely to be contained in CNV regions.

Nguyen *et al.* as well as Cooper *et al.* reported a highly significant depletion of genes with the Gene Ontology (GO) category "binding" within CNV regions, but they do not comment further on this fact. I verified independently that "binding" is the second most underrepresented GO category after "intracellular" amongst genes in CNV regions. This lends further support to the hypothesis that dosage sensitivity due to protein complex membership has an influence of the composition of CNV regions.

I speculated that a negative fitness effect due to a copy-number variation will increase the likelihood of subsequent removal of that CNV from the gene pool. The CNVs that contain CORUM genes occur in significantly fewer individuals (MWU: $P = 1.6 \cdot 10^{-4}$) than non-CORUM genes, indicating that purifying selection may have acted on some of the genes.

I also tested whether CORUM genes are underrepresented in gains compared to losses. Out of the 167 CORUM genes that overlap a CNV, 18.5% occur in a gain, compared to 29.8% of non-CORUM genes. This significant difference in ratios (FET: $P = 9.6 \cdot 10^{-4}$) suggests that amongst copy-number varied genes, there is indeed a bias against duplications for genes in protein complexes, supporting the notion that stoichiometric imbalance has a negative effect on protein complexes.

4.3.4 Compositional bias of copy-number varied genes

Various compositional biases on genes in CNV regions have been described (Cooper *et al.*, 2007; Nguyen *et al.*, 2006). Most notably, it has been reported that genes within CNV regions exhibit higher dN/dS than the rest of the genome. Is the observed low

frequency of CORUM and other dosage sensitive genes in CNV regions merely a result of a bias against slower evolving genes? I verified that dN/dS ratios of genes within CNV regions were elevated compared to their mouse orthologs (Median: 0.131, P-Value by resampling: $P = 3.2 \cdot 10^{-7}$). Conversely, CORUM genes exhibit lower than expected dN/dS (Median: 0.070, $P < 10^{-40}$). In contrast to non-complex genes, there is no significant difference in dN/dS between CORUM genes that overlap CNVs and those that do not. I therefore tested whether there is a causal relationship between complex membership, low dN/dS and CNV overlap.

Like CORUM genes, the automatically clustered complexes also exhibited low dN/dS (Median 0.08, $P = 1.9 \cdot 10^{-30}$). It has been argued that proteins with obligate interactions are under stronger selective pressure (Mintseris and Weng, 2005), which could explain the low dN/dS in both CORUM and the automatically clustered complexes. Interestingly, Cooper *et al.* showed that CNVs and segmental duplications (SDs) are of fundamentally similar nature and frequently overlap. I thus hypothesised that the reduction in negative selection within CNVs is related to the higher copy number of some genes which have been recently duplicated in a fixed SD. If I split the genes in CNV regions into those that overlap a SD and those that do not, it can be measured that dN/dS ratios are highly significantly elevated in the genes that overlap SDs (MWU: $P < 2.2 \cdot 10^{-16}$), but not in the group outside SDs ($P = 0.017$).

Subsequently, I analysed the distribution of numbers of paralogs for human genes. I found that genes in CNV regions have significantly more paralogs than would be expected by chance (MWU, $P = 1.45 \cdot 10^{-9}$),whereas genes from CORUM have significantly fewer ($P < 2.2 \cdot 10^{-16}$). As with the evolutionary rate, the increase in numbers of paralogs is largely driven by CNVs that overlap SDs. Removing all genes inside SDs reduced the number of paralogs substantially (P-value reduced from $1.45 \cdot 10^{-9}$ to 0.0033). Conversely, the genes that are in both CNVs and SDs have significantly more paralogs than genes only found in CNV regions ($P = 4.3 \cdot 10^{-11}$). I conclude that

the increase in dN/dS in CNV regions is driven by an increase in gene copy number and thus does not explain the underrepresentation of dosage sensitive genes in CNV regions.

If SDs are largely responsible for the increased dN/dS within CNVs and the increase in number of paralogs, can I still detect the underrepresentation of CORUM genes in CNVs that do not overlap a SD? After removing all genes that overlap a SD, CORUM genes were still significantly underrepresented ($P = 3.3 \cdot 10^{-4}$) in CNV regions, indicating that negative selective pressure not only affects regions of segmental duplication but also other types of CNVs.

4.4 Discussion

4.4.1 Protein complexes are sensitive to alterations in gene expression

Correlated gene expression of interacting proteins is a well known phenomenon, to the extent that correlation analysis is used to validate high-throughput protein interaction experiments (Hahn *et al.*, 2005). Usually, expression data is gathered under diverse physiological conditions, *e.g.* at different stages of the cell cycle. In this analysis, I have compared data from 79 different human tissue types. As expected, I observe strong correlation between the changes in gene expression for members of the same protein complex in different tissues. This observation hints at the importance of tightly regulated gene expression for the correct functioning of protein complexes.

However, it does not directly verify if the stoichiometry of complexes is under the same strong regulation. I therefore measured the variation in expression levels for interacting proteins in different HapMap individuals. Expressional noise of protein complexes has been analysed in *S. cerevisiae* and *D. melanogaster*(Lemos *et al.*, 2004), but the HapMap gene expression data allow the first systematic evaluation of protein complex expression in human. I find that genes in CORUM exhibit significantly

lower variation in expression than the rest of the genome. This is direct evidence that expression of complex genes is under tighter regulation than the rest of the genome. Furthermore, I find that genes in large complexes maintain particularly low expression variation. While I cannot rule out that this observation is due to functional constraints on the particular complexes, it does suggest that sensitivity to expressional noise is related to the number of subunits a complex maintains.

When I analysed the composition of genes in CNV regions, I made the curious observation that the small number of CORUM genes that overlap a CNV (165 genes in total) are biased towards deletions rather than duplications. If I assume that negative selection is acting on CNVs, the intuitive biological explanation for this phenomenon would be that CORUM genes are at least as sensitive to duplication as to deletion, which in turn supports the concept that members of protein complexes are sensitive not just to under- but also to overexpression.

I made another observation that supports this hypothesis. When comparing a manually curated set of dosage sensitive genes derived from the scientific literature, I found that a significantly larger than expected proportion of these genes were members of a protein complex as defined by the CORUM database. Taken together, these findings indicate that stoichiometric fluctuations negatively affect protein complexes.

4.4.2 CNVs affect expression levels of contained genes

A key proposition that underpins our understanding of dosage sensitivity is that duplication or deletion of the genomic region containing a gene will result in a significant up- or downregulation of expression of the gene. There have been previous reports of widespread expressional silencing of chromosomal amplifications (Platzer *et al.*, 2002). In contrast, I observed lower average gene expression in deletion CNVs compared to duplication CNVs (Figure 4.5). It has to be noted, though, that these differences in expression are small for the majority of genes within a CNV. Furthermore, there are

numerous cases where deletions seemingly result in increased expression and vice versa. Figures 4.7c and 4.7d exemplify how noisy the expression data for a gene can be, despite a visible expression difference between deletion and duplication genotypes. Sensitivity to detect expression differences at low concentration is not the main source of this variability in gene expression. Rather, I suspect there to be inherent fluctuations between the different cell lines used in the analysis (Blake *et al.*, 2003). Expressional noise alone does not explain that some CNVs seem not to affect gene expression at all. Rather, the inaccurate prediction of start and end coordinates of CNVs is likely to be largely responsible for the lack of correlation between CNVs and gene expression. Individuals with a CNV genotype falsely labelled as unaffected, or a gene erroneously placed inside a CNV, will skew the distribution of expression ratios.

I speculate, however, that there could also be a physiological explanation for the unexpectedly low change in gene expression upon copy-number variation. It is conceivable that the cell attempts to compensate changes in copy number on gene expression by *e.g.* increasing or decreasing transcription or modulating mRNA degradation. Such autosomal dosage compensation was first observed in *D. melanogaster* (Devlin *et al.*, 1982) and a general mechanism for dosage regulation has been proposed (Birchler *et al.*, 2005). According to this theory, dosage balance is achieved through a network of regulatory genes which themselves are therefore dosage sensitive. The enrichment of CORUM for regulatory and transcription related functions might thus explain its sensitivity to copy-number variation and the low effect of CNVs on gene expression at the same time. Interestingly, Kind *et al.* (2008) recently described the formation and binding properties of a dosage-regulatory complex in *D. melanogaster*. They note that the components of the complex are not only conserved in mammals, but there is also autosomal activity of the respective proteins which is not fully understood. With the arrival of new CNV datasets featuring improved breakpoint accuracy, it should become possible to better distinguish between false positive predictions and genes that are actually sub-

ject to dosage compensation. Subsequently, this will make it possible to determine the frequency of autosomal dosage compensation of copy-number varied genes.

4.4.3 CNVs as the source of recent duplications

It has been noted (Nguyen *et al.*, 2006) that genes within CNV regions exhibit higher than expected dN/dS ratios, suggesting a relaxation of selective pressure. On the contrary, complex genes, dosage sensitive genes and highly connected genes in general, show very low dN/dS ratios, irrespective of whether they overlap CNVs or not. Stronger selective constraints in highly connected proteins have previously been attributed to functional constraints on the protein surface in order to maintain multiple binding sites (Mintseris and Weng, 2005).

Interestingly, I also show that genes in CNV regions have significantly more paralogs than expected by chance, while genes in protein complexes possess, on average, fewer paralogs (Yang *et al.*, 2003). This suggests that CNV regions have been hot-spots of large scale variation for a prolonged period of time, as it has also been shown that gene-rich CNV regions correspond well with regions of segmental duplications (Cooper *et al.*, 2007). In fact, I found that those CNV regions that overlap segmental duplications are primarily (though not exclusively) responsible for the high number of paralogs.

Conversely, the reason for the increase in dN/dS in many genes within CNV regions could be attributed to their higher number of paralogous sequences: Even a partial relaxation of selection pressure due to an additional gene copy is likely to increase the observed dN/dS ratios. In fact, genes in CNVs overlapping segmental duplications are again primarily, but not exclusively, responsible for the elevated dN/dS ratios. These observations underline that CNV regions are a frequent source of gene duplicates which occasionally get fixed over the course of evolution and thus drive evolution of some gene families.

4.4.4 Dosage sensitivity and negative selection on CNVs

I observed that CNV regions are less likely to contain genes encoding protein complexes, as well as other dosage sensitive genes. Furthermore, CNVs which occur in multiple individuals and can thus be assumed to be older than unique CNVs are particularly depleted of CORUM genes. Hence, it appears that pressures on correct dosage limit the set of genes which can sustain variation in copy-number, even though the effect of CNVs on gene expression is not straightforward.

Dang *et al.* (2008) reported that haploinsufficient genes are seldom found between two regions of segmental duplication. These results shed new light on this finding: It seems that dosage sensitive genes in general are biased against regions in which they are prone to suffer from copy-number variation. Segmental duplications are the most common source of such rearrangements, however I show that other CNVs not related to segmental duplications are also depleted of dosage sensitive genes. This indicates that rearrangements due to CNVs are subject to negative selection.

These findings offer a partial but consistent explanation for the biased composition of CNV regions. In addition to that, the correlation between dosage sensitivity and protein complex membership provides a convenient way to predict which genes are likely to be important in diseases which involve genomic rearrangements. The enrichment of CORUM for genes upregulated in cancer clearly hints towards this possibility. Future investigations should focus on the involvement of CNVs of putative dosage sensitive genes in cancer and complex diseases.

Chapter 5

Concluding Remarks

In the first part of this thesis, I have attempted to evaluate the potential and the limitations of using structure information for the study of protein interactions. I have shown that protein domains known to be part of an interaction interface in a protein structure can be projected onto the protein interaction network. This reveals that while our current knowledge of interacting domain pairs is small, these domain pairs are significantly overrepresented in experimentally verified protein interactions in both eukaryotes as well as prokaryotes. There is also significant conservation of domain pairs between species, even though only approximately 5% of the protein interaction network is covered by the structural data. This presents a strong argument for solving the structures of more novel interacting domain pairs. A substantially higher coverage could for example provide enough information to identify the most likely binary pairs of interacting proteins in complexes identified using affinity-purification methods: those protein pairs with known interacting domain pairs can be assumed to be more likely to really interact.

In the following chapter, I demonstrated that the existing structural data can be employed successfully to investigate disease mutations on a molecular level. I described several genetic diseases which are the result of point mutations in a domain which is known to be involved in an interaction through a homologous structure. In the future,

binding kinetics experiments will hopefully confirm my predictions. My approach already exemplifies the power of structural homology based approaches applied to protein interactions. Within the possibilities of the incomplete datasets available, I estimated that 4% of all known disease mutations affect a protein interaction. Increased numbers of structural templates and more stringently defined domains, representing only a particular binding geometry or binding partner, could improve the sensitivity and specificity of my method further.

Interestingly, many of the mutations in interaction interfaces are inherited in a dominant fashion. In the last part of this thesis, I extended my analysis beyond structure-based domains to study the evolutionary pressures governing protein complexes in human. Specifically, I investigated the distribution of protein complexes with respect to large insertion and deletion polymorphisms often referred to as copy-number variations (CNVs). It is known that proteins vary regarding their duplicability and sensitivity to homozygous deletion. It has been argued that many dosage sensitive proteins are members of protein complexes. I observed in human that expression variation in members of protein complexes is significantly lower than in other selected proteins. Furthermore, I could show that members of protein complexes are rarely found inside CNVs. Combined, these two facts suggest that frequently, purifying selection acts against CNVs that contain genes encoding protein complexes, or genes in protein complexes have evolved to reside outside regions which are enriched for CNVs. It seems likely that such evolutionary pressures have been acting for some time, as the set of protein complex genes also has fewer paralogs on average than other genes. In congruence with the duplication/divergence theory of gene evolution, the studied genes of members of protein complexes are under stronger negative selection than the rest of the genome, as indicated by their low dN/dS rates.

An interesting alternative approach to the same question could be the analysis of known knock-out mice mutants. With the increasing availability of knock-out models

for various genes, it could be envisaged to differentiate between heterozygous as opposed to purely homozygous phenotypes, in a similar way as dominant and recessive mutations are defined in human disease. From my initial results presented in this thesis, I expect knock-outs of genes in protein complexes to be more often phenotypicaly active than other genes.

In summary, it can be said that the investigation of protein interactions has already brought about many exciting insights and fostered interconnections between previously unrelated fields. Combining structure information with protein interactions to explain genetic diseases is an example of such an integrative approach that will probably become more common in the coming years. Similarly, my analysis of large scale genomic variation in the context of protein interactions shows how network biology can provide insights into such fundamental questions as gene duplicability. However, as the field of protein interaction research is still in a comparatively early stage of development, many basic assertions still need to be made and many obstacles need to be overcome. Our understanding of the evolution of protein interactions is still incomplete. Being able to trace the processes that shaped the interaction networks of higher organisms would not only shed light on the origins of organismal complexity, but could also be of practical use: it is still unclear to what extent protein interactions are conserved between species. Moreover, it is also not yet fully understood what distinguishes a protein interaction interface from other surface regions. As a result of that, our ability to validate or even predict protein interactions is still limited. My findings point towards the possibility of reducing the complexity of protein interaction networks down to domain interaction networks as a more conserved unit of interaction evolution.

Bibliography

Albert R and Barabási A. Statistical mechanics of complex networks. *Rev Mod Physics*, 74:47–97, 2002. 1.1.4

Aloy P, Ceulemans H, Stark A and Russell RB. The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, 332:989–998, 2003. 1.1.1.4, 2.3.4, 3.3.2.1, 3.4.1

Aloy P and Russell RB. The third dimension for protein interactions and complexes. *Trends Biochem Sci*, 27:633–638, 2002. 3.4.1

Aloy P and Russell RB. Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, 22:1317–1321, 2004. 1.3.1, 2.1, 2.3.7, 2.3.7, 2.4.1, 2.4.2

Aury JM, Jaillon O, Duret L, Noel B, Jubin C *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, 444:171–178, 2006. 4.1

Axelsen JB, Lotem J, Sachs L and Domany E. Genes overexpressed in different human solid cancers exhibit different tissue-specific expression profiles. *Proc Natl Acad Sci USA*, 104:13122–13127, 2007. 4.2.4, 4.3.2

Barabasi A and Albert R. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. 1.1.4

Barabasi AL and Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5:101–113, 2004. 1.1.4

Barbujani G, Magagni A, Minch E and Cavalli-Sforza LL. An apportionment of human DNA diversity. *Proc Natl Acad Sci USA*, 94:4516–4519, 1997. 1.2.2

Bartel PL and Fields S. *The Yeast Two-Hybrid System (Advances in Molecular Biology)*. Oxford University Press, USA, 1st edition, 1997. 1.1.1.2

Basu MK, Carmel L, Rogozin IB and Koonin EV. Evolution of protein domain promiscuity in eukaryotes. *Genome Res*, 18:449–461, 2008. 2.4.2

Batzer MA and Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet*, 3:370–379, 2002. 1.2.1

Berggård T, Linse S and James P. Methods for the detection and analysis of protein–protein interactions. *Proteomics*, 7:2833–2842, 2007. 1.1.1.1, 1.1.1.2

Berman HM. The Protein Data Bank: a historical perspective. *Acta Crystallogr A*, 64:88–95, 2008. 1.5

Birchler J. The genetic basis of dosage compensation of alcohol dehydrogenase-1 in maize. *Genetics*, 97:625–637, 1981. 4.1

Birchler JA, Bhadra U, Bhadra MP and Auger DL. Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev Biol*, 234:275–288, 2001. 4.1

Birchler JA, Riddle NC, Auger DL and Veitia RA. Dosage balance in gene regulation: biological implications. *Trends Genet*, 21:219–226, 2005. 4.1, 4.4.2

Birchler JA, Yao H and Chudalayandi S. Biological consequences of dosage dependent gene regulatory systems. *Biochim Biophys Acta*, 1769:422–428, 2007. 4.1

Blake WJ, Kaern M, Cantor CR and Collins JJ. Noise in eukaryotic gene expression. *Nature*, 422:633–637, 2003. 4.4.2

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol*, 18:883–889, 2008. 1.2.3.1

Botstein D and Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics*, 33:228–237, 2003. 1.2.3.1

Boyd W. Genetics and the human race: Definition of race on the basis of gene frequencies supplements definition from morphological characters. *Science*, 140:1057–1064, 1963. 1.2.3

Branden C and Tooze J. *Introduction to Protein Structure*. Garland Publishing, 2nd edition, 1991. 1.1.1.4

Bravo J and Aloy P. Target selection for complex structural genomics. *Curr Opin Struct Biol*, 16:385–392, 2006. 2.4.1

Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A *et al.* The BioGRID interaction database: 2008 update. *Nucleic Acids Res*, 36:D637–640, 2008. 1.1.3, 2.2.1, 3.3.3

Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S and Birney E. The HGNC database in 2008: a resource for the human genome. *Nucleic Acids Res*, 36:D445–448, 2008. 4.2.1

Buratti E, Baralle M and Baralle FE. Defective splicing, disease and therapy: searching for master checkpoints in exon definition. *Nucleic Acids Res*, 34:3494–3510, 2006. 3.1

Burckstummer T, Bennett KL, Preradovic A, Schutze G, Hantschel O, Superti-Furga G and Bauch A. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat Methods*, 3:1013–1019, 2006. 1.1.1.1

Bushell KM, Sollner C, Schuster-Böckler B, Bateman A and Wright GJ. Large-scale screening for novel low-affinity extracellular protein interactions. *Genome Res*, 18:622–630, 2008. 1.1.1.5, 1.4

Camon E, Magrane M, Barrell D, Lee V, Dimmer E *et al.* The Gene Ontology Annotation (GOA) database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Res*, 32:D262–266, 2004. 4.2.9

Chakrabarti P and Janin J. Dissecting protein–protein recognition sites. *Proteins*, 47:334–343, 2002. 3.3.4.4, 3.9

Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L and Cesareni G. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 35:D572–574, 2007. 1.1.3, 2.2.1

Cheung SW, Shaw CA, Yu W, Li J, Ou Z *et al.* Development and validation of a CGH microarray for clinical cytogenetic diagnosis. *Genet Med*, 7:422–432, 2005. 4.2.4

Chiti F and Dobson CM. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem*, 75:333–366, 2006. 3.1

Chothia C. One thousand families for the molecular biologist. *Nature*, 357:543–544, 1992. 1.3

Chothia C and Lesk AM. The relation between the divergence of sequence and structure in proteins. *EMBO J*, 5:823–826, 1986. 1.1.1.4, 3.3.2.1

Collins FS, Guyer MS and Chakravarti A. Variations on a theme: Cataloging human DNA sequence variation. *Science*, 278:1580–1581, 1997. 3.1

Conrad DF, Andrews TD, Carter NP, Hurles ME and Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, 38:75–81, 2006. 1.2.3

Conrad DF and Hurles ME. The population genetics of structural variation. *Nat Genet*, 39:S30–36, 2007. 4.1

Cooper GM, Nickerson DA and Eichler EE. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet*, 29:S22–29, 2007. 4.1, 4.3.3, 4.3.4, 4.4.3

Cunningham BC and Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, 244:1081–1085, 1989. 1.1, 3.2.6

Cusick ME, Klitgord N, Vidal M and Hill DE. Interactome: gateway into systems biology. *Hum Mol Genet*, 14:R171–181, 2005. 2.1

Dang V, Kassahn K, Marcos A and Ragan M. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur J Hum Genet*, 2008. 4.4.4

Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, 1st edition, 1859. 1.2

Deane CM, Salwinski L, Xenarios I and Eisenberg D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1:349–356, 2002. 1.1.2

Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, Davis RW, Nislow C and Giaever G. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, 169:1915–1925, 2005. 4.1

Devlin RH, Holm DG and Grigliatti TA. Autosomal dosage compensation in *Drosophila melanogaster* strains trisomic for the left arm of chromosome 2. *Proc Natl Acad Sci USA*, 79:1200–1204, 1982. 4.1, 4.4.2

Disotell TR. Human genomic variation. *Genome Biol*, 1:5, 2000. 1.2.2

van Dongen S. *Graph clustering by flow simulation.* Ph.D. thesis, University of Utrecht, The Netherlands, 2000. 4.2.3

Durbin R, Eddy SR, Krogh A and Mitchison G. *Biological Sequence Analysis.* Cambridge University Press, 1998. 1.3

Eddy SR. *HMMER User's Guide: Biological sequence analysis using profile hidden Markov models, version 2.2.* Washington University School of Medicine, http://hmmer.wustl.edu, 2001. 3.2.4

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32:1792–1797, 2004. 3

Eisenberg E and Levanon E. Preferential attachment in the protein network evolution. *Phys Rev Lett*, 91:138701, 2003. 1.1.4

El Ghouzzi V, Legeai-Mallet L, Aresta S, Benoist C, Munnich A, De Gunzburg J and Bonaventure J. Saethre-Chotzen mutations cause TWIST protein degradation or impaired nuclear location. *Hum Mol Genet*, 9:813–819, 2000. 3.3.6.3

Enright AJ, Iliopoulos I, Kyrpides NC and Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402:86–90, 1999. 1.3.1, 2.2.4

Ewing RM, Chu P, Elisma F, Li H, Taylor P *et al.* Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol Syst Biol*, 3:89, 2007. 4.2.3

Fawcett T. An introduction to ROC analysis. *Pattern Recog Lett*, 27:861–874, 2006. 3.3.2.3

Feldman MW, Lewontin RC and King MC. Race: a genetic melting-pot. *Nature*, 424:374, 2003. 1.2.2

Ferrer-Costa C, Orozco M and de la Cruz X. Characterization of disease–associated single amino acid polymorphisms in terms of sequence and structure properties. *J Mol Biol*, 315:771–786, 2002. 3.1, 3.4.2

Fields S and Song O. A novel genetic system to detect protein–protein interactions. *Nature*, 340:245–246, 1989. 1.1.1.2

Finn RD, Marshall M and Bateman A. iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 21:410–412, 2005. 1.3.1, 2.1, 2.2.4, 3.2.2, 3.3.1

Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res*, 34:D247–251, 2006. 1.4

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ *et al.* The Pfam protein families database. *Nucleic Acids Res*, 36:D281–288, 2008. 1, 1.3, 2.1

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA *et al.* Copy number variation: new insights in genome diversity. *Genome Res*, 16:949–961, 2006. 1.2.3

Fu L. Alteration of protein–protein interactions of congenital cataract crystallin mutants. *Investig Ophthalmology & Vis Sci*, 44:1155–1159, 2003. 3.1

Gandhi TKB, Zhong J, Mathivanan S, Karthick L, Chandrika KN *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38:285–293, 2006. 1, 2.4.3

Gavin AC, Aloy P, Grandi P, Krause R, Boesche M *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, 2006. 1.1.2, 2.3.5

Gerhart J and Schachman H. Distinct subunits for the regulation and catalytic activity of aspartate transcarbamylase. *Biochemistry*, 4:1054–1062, 1965. 1.1

Gerhart JC and Pardee AB. The enzymology of control by feedback inhibition. *J Biol Chem*, 237:891–896, 1962. 1.1

Giorgini F and Muchowski PJ. Connecting the dots in Huntington's disease with protein interaction networks. *Genome Biol*, 6:210–211, 2005. 3.1

Glaser F, Rosenberg Y, Kessel A, Pupko T and Ben-Tal N. The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, 58:610–617, 2005. 3.2.3

Grace CRR, Perrin MH, Gulyas J, Digruccio MR, Cantle JP, Rivier JE, Vale WW and Riek R. Structure of the N-terminal domain of a type B1 G protein-coupled receptor in complex with a peptide ligand. *Proc Natl Acad Sci USA*, 104:4858–4863, 2007. 3.2.6

Grigoriev A. On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res*, 31:4157–4161, 2003. 2.1

Guindon S and Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52:696–704, 2003. 4

Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes H and Stumpflen V. MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*, 34:D436–441, 2006. 2.2.1, 3.3.3

Hahn A, Rahnenfuhrer J, Talwar P and Lengauer T. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6:1, 2005. 4.3.2, 4.4.1

Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33:D514–517, 2005. 1.2.3.1, 3.1, 3.2.1

Han JDJ, Dupuy D, Bertin N, Cusick ME and Vidal M. Effect of sampling on topology predictions of protein–protein interaction networks. *Nat Biotechnol*, 23:839–844, 2005. 1.1.4

Hart GT, Ramani AK and Marcotte EM. How complete are current yeast and human protein-interaction networks? *Genome Biol*, 7:120, 2006. 1, 1.1.2

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L *et al.* The HUPO PSI's molecular interaction format — a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22:177–183, 2004. 2.2.1

Hong B, Senisterra G, Rabeh W, Vedadi M, Leonardi R *et al.* Crystal structures of human pantothenate kinases: Insights into allosteric regulation and mutations linked to a neurodegeneration disorder. *J Biol Chem*, 282:27984–27993, 2007. 3.1

Huber LA. Is proteomics heading in the wrong direction? *Nat Rev Mol Cell Biol*, 4:74–80, 2003. 1.3

Hughes AL and Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335:167–170, 1988. 1.2.1

Humphrey W, Dalke A and Schulten K. VMD – Visual Molecular Dynamics. *J Mol Graph*, 14:33–38, 1996. 3.2.9

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW and Lee C. Detection of large-scale variation in the human genome. *Nat Genet*, 36:949–951, 2004. 1.2.3, 4.2.7

Ispolatov I, Yuryev A, Mazo I and Maslov S. Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res*, 33:3629–3635, 2005. 3.3.3

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 98:4569–4574, 2001. 1.1.2, 1.1.4

Itzhaki Z, Akiva E, Altuvia Y and Margalit H. Evolutionary conservation of domain–domain interactions. *Genome Biol*, 7:R125, 2006. 2.1

James LC, Keeble AH, Khan Z, Rhodes DA and Trowsdale J. Structural basis for PRYSPRY-mediated tripartite motif (TRIM) protein function. *Proc Natl Acad Sci USA*, 104:6200–6205, 2007. 3.2.6

Janin J, Rodier F, Chakrabarti P and Bahadur RP. Macromolecular recognition in the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, 63:1–8, 2007. 1.1

Jansen R, Greenbaum D and Gerstein M. Relating whole-genome expression data with protein–protein interactions. *Genome Res*, 12:37–46, 2002. 4.1, 4.3.2

Jeong H, Mason SP, Barabasi AL and Oltvai ZN. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. 1.1.4

Jhoti H. High-throughput structural proteomics using x-rays. *Trends Biotechnol*, 19:S67–71, 2001. 1.1.1.4

Jimenez-Sanchez G, Childs B and Valle D. Human disease genes. *Nature*, 409:853–855, 2001. 3.3.4.3, 1

Jones S and Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 93:13–20, 1996. 1.1

Jothi R, Cherukuri PF, Tasneem A and Przytycka TM. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J Mol Biol*, 362:861–875, 2006. 2.1, 2.4.1, 2.4.3

Junttila MR, Saarinen S, Schmidt T, Kast J and Westermarck J. Single-step Strep-tag purification for the isolation and identification of protein complexes from mammalian cells. *Proteomics*, 5:1199–1203, 2005. 1.1.1.1

Kafatos FC, Efstratiadis A, Forget BG and Weissman SM. Molecular evolution of human and rabbit beta-globin mRNAs. *Proc Natl Acad Sci USA*, 74:5618–5622, 1977. 1.2.1

Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS and Koonin EV. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol*, 2:18, 2002. 1.1.4

Kellis M, Birren BW and Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004. 4.1

Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H and Phillips DC.

A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958. 1.1.1.4

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D. The human genome browser at UCSC. *Genome Res*, 12:996–1006, 2002. 4.4

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, 35:D561–565, 2007. 1.1.3, 2.2.1, 3.3.3, 4.2.3

Kersey P, Bower L, Morris L, Horne A, Petryszak R *et al.* Integr8 and genome reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res*, 33:D297–302, 2005. 2.2.3

Kestler HA. ROC with confidence — a Perl program for receiver operator characteristic curves. *Comput Methods Programs Biomed*, 64:133–136, 2001. 3.7

Khanin R and Wit E. How scale–free are biological networks? *J Comput Biol*, 13:810–818, 2006. 1.1.4

Kind J, Vaquerizas JM, Gebhardt P, Gentzel M, Luscombe NM, Bertone P and Akhtar A. Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in *Drosophila*. *Cell*, 133:813–828, 2008. 4.4.2

Kissinger CR, Rejto PA, Pelletier LA, Thomson JA, Showalter RE *et al.* Crystal structure of human ABAD/HSD10 with a bound inhibitor: implications for design of Alzheimer's disease therapeutics. *J Mol Biol*, 342:943–952, 2004. 3.3.6

Klein C, Philippe N, Le Deist F, Fraitag S, Prost C, Durandy A, Fischer A and Griscelli C. Partial albinism with immunodeficiency (Griscelli syndrome). *J Pediatr*, 125:886–895, 1994. 3.3.6.1

Kleywegt GJ. Validation of protein crystal structures. *Acta Crystallogr D Biol Crystallogr*, 56:249–265, 2000. 1.1.1.4

Klotz IM, Langerman NR and Darnall DW. Quaternary structure of proteins. *Annual Review of Biochemistry*, 39:25–62, 1970. 1.1

Koegl M and Uetz P. Improving yeast two-hybrid screening systems. *Brief Funct Genomic Proteomic*, 6:302–312, 2007. 1.1.1.2

Kondrashov FA and Koonin EV. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet*, 20:287–290, 2004. 1.2.3.1, 4.1

Kondrashov FA, Rogozin I, Wolf Y and Koonin E. Selection in the evolution of gene duplications. *Genome Biol*, 3:2, 2002. 4.1

Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE and Berman HM. The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*, 34:D302–305, 2006. 1.1.1.4, 2.1, 3.3.1

Krissinel E and Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, 372:774–797, 2007. 1.1.1.4

Krogan NJ, Cagney G, Yu H, Zhong G, Guo X *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440:637–643, 2006. 4.2.3

Krogh A, Brown M, Mian IS, Sjolander K and Haussler D. Hidden Markov models in computational biology. applications to protein modeling. *J Mol Biol*, 235(5):1501–1531, 1994. 1.3

Kuhn K, Baker SC, Chudin E, Lieu MH, Oeser S *et al.* A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res*, 14:2347–2356, 2004. 4.2.5

Lamolet B, Pulichino AM, Lamonerie T, Gauthier Y, Brue T, Enjalbert A and Drouin J. A pituitary cell-restricted T box factor, Tpit, activates POMC transcription in cooperation with Pitx homeoproteins. *Cell*, 104:849–859, 2001. 3.3.6.2

Leandro J, Nascimento C, de Almeida IT and Leandro P. Co-expression of different subunits of human phenylalanine hydroxylase: Evidence of negative interallelic complementation. *Biochim Biophys Acta*, 1762:544–550, 2006. 3.4.3

Lee H, Deng M, Sun F and Chen T. An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, 7, 2006. 2.1, 2.4.1

Lee JA and Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*, 52:103–121, 2006. 4.1

Lehner B, Crombie C, Tischler J, Fortunato A and Fraser AG. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat Genet*, 38:896–903, 2006. 1.1.1.5

Lemos B, Meiklejohn CD and Hartl DL. Regulatory evolution across the protein interaction network. *Nat Genet*, 36:1059–1060, 2004. 4.4.1

Lewontin RC. The apportionment of human diversity. *Evolutionary Biology*, 6:391–398, 1972. 1.2.3

Littler SJ and Hubbard SJ. Conservation of orientation and sequence in protein domain–domain interactions. *J Mol Biol*, 345:1265–1279, 2005. 2.1

Livingstone F. Anthropological implications of sickle cell gene distribution in west africa. *Am Anthropol*, 60:533–562, 1958. 1.2.3

Logsdon NJ, Jones BC, Allman JC, Izotova L, Schwartz B, Pestka S and Walter MR. The IL-10R2 binding hot spot on IL-22 is located on the N-terminal

helix and is dependent on N-linked glycosylation. *J Mol Biol*, 342:503–514, 2004. 3.2.6

Lu X, Shaw CA, Patel A, Li J, Cooper ML *et al.* Clinical implementation of chromosomal microarray analysis: summary of 2513 postnatal cases. *PLoS ONE*, 2:e327, 2007. 3.1

Luscombe NM, Qian J, Zhang Z, Johnson T and Gerstein M. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol*, 3:R8, 2002. 1.1.4

Lynch M and Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155, 2000. 4.1

MacBeath G and Schreiber SL. Printing proteins as microarrays for high-throughput function determination. *Science*, 289:1760–1763, 2000. 1.1.1.5

Mani R, St Onge RP, Hartman JLt, Giaever G and Roth FP. Defining genetic interaction. *Proc Natl Acad Sci USA*, 105:3461–3466, 2008. 1.1.1.5

Markham K, Bai Y and Schmitt-Ulms G. Co-immunoprecipitations revisited: an update on experimental concepts and their implementation for sensitive interactome investigations of endogenous proteins. *Anal Bioanal Chem*, 389:461–473, 2007. 1.1.1.5

Maslov S and Sneppen K. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002. 2.2.6

Menasche G, Pastural E, Feldmann J, Certain S, Ersoy F *et al.* Mutations in RAB27A cause Griscelli syndrome associated with haemophagocytic syndrome. *Nat Genet*, 25:173–176, 2000. 3.3.6.1

Mendel J. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 4:3–47, 1865. 1.2

von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S and Bork P. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417:399–403, 2002. 1.1.2, 2.1

Mewes HW, Dietmann S, Frishman D, Gregory R, Mannhaupt G *et al.* MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res*, 36:D196–201, 2008. 1.1.1.3, 1.1.3

Mintseris J and Weng Z. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc Natl Acad Sci USA*, 102:10930–10935, 2005. 4.3.4, 4.4.3

Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S *et al.* Human protein reference database–2006 update. *Nucleic Acids Res*, 34:D411–414, 2006. 1.1.3, 2.2.1, 3.3.3

Müller CW and Herrmann BG. Crystallographic structure of the T domain-DNA complex of the brachyury transcription factor. *Nature*, 389:884–888, 1997. 3.3.6.2

Nair SK and Burley SK. X-ray structures of Myc-Max and Mad-Max recognizing DNA: Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, 112:193–205, 2003. 3.3.6.3

Neduva V and Russell RB. Linear motifs: evolutionary interaction switches. *FEBS Lett*, 579:3342–3345, 2005. 2.3.5

Ng PC and Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31:3812–3814, 2003. 3.3.2.2

Ng SK, Zhang Z, Tan SH and Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 31:251–4, 2003. 2.1

Nguyen DQ, Webber C and Ponting CP. Bias of selection on human copy-number variants. *PLoS Genet*, 2:e20, 2006. 4.1, 4.3.3, 4.3.4, 4.4.3

Nye TMW, Berzuini C, Gilks WR, Babu MM and Teichmann SA. Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21:993–1001, 2005. 2.1

Ofman R, Ruiter JPN, Feenstra M, Duran M, Poll-The BT *et al.* 2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency is caused by mutations in the HADH2 gene. *Am J Hum Genet*, 72:1300–1307, 2003. 3.3.6

Ofran Y and Rost B. Analysing six types of protein–protein interfaces. *J Mol Biol*, 325:377–387, 2003. 3.3.4.4

Ofran Y and Rost B. Protein–protein interaction hotspots carved into sequences. *PLoS Comput Biol*, 3:e119, 2007. 3.3.2.4

Orr H. A test of Fisher's theory of dominance. *Proc Natl Acad Sci USA*, 88:11413–11415, 1991. 4.1

Ostermeier C and Brunger AT. Structural basis of Rab effector specificity: crystal structure of the small G protein Rab3A complexed with the effector domain of rabphilin-3a. *Cell*, 96:363–374, 1999. 3.3.6.1

Pagel P, Wong P and Frishman D. A domain interaction map based on phylogenetic profiling. *J Mol Biol*, 344:1331–46, 2004. 2.1

Papp B, Pal C and Hurst LD. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424:194–197, 2003. 3.4.3, 4.1

Park J, Lappe M and Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J Mol Biol*, 307:929–38, 2001. 2.1

Peng K, Obradovic Z and Vucetic S. Exploring bias in the protein data bank using contrast classifiers. *Pac Symp Biocomput*, 435–446, 2004. 2.1

Perutz M, Rossmann M, Cullis A, Muirhead H, Will G and North A. Structure of haemoglobin: A three-dimensional fourier synthesis at 5.5Å resolution, obtained by X-ray analysis. *Nature*, 185:416–422, 1960. 1.1.1.4

Platzer P, Upender MB, Wilson K, Willis J, Lutterbaugh J *et al.* Silence of chromosomal amplifications in colon cancer. *Cancer Res*, 62:1134–1138, 2002. 4.4.2

Powell AJ, Read JA, Banfield MJ, Gunn-Moore F, Yan SD *et al.* Recognition of structurally diverse substrates by type II 3-hydroxyacyl-CoA dehydrogenase (HADH II)/amyloid-beta binding alcohol dehydrogenase (ABAD). *J Mol Biol*, 303:311–327, 2000. 3.3.6

Prince VE and Pickett FB. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet*, 3:827–837, 2002. 4.1

Pulichino AM, Vallette-Kasic S, Couture C, Gauthier Y, Brue T *et al.* Human and mouse TPIT gene mutations cause early onset pituitary ACTH deficiency. *Genes Dev*, 17:711–716, 2003. 3.3.6.2

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0. 3.2.8, 4.2.12

Randles LG, Lappalainen I, Fowler SB, Moore B, Hamill SJ and Clarke J. Using model proteins to quantify the effects of pathogenic mutations in Ig-like proteins. *J Biol Chem*, 281:24216–24226, 2006. 3.3.2.4

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH *et al.* Global variation in

copy number in the human genome. *Nature*, 444:444–454, 2006. 1.2.3, 3.1, 4.4, 4.3.1, 4.7, 4.3.3

Reguly T, Breitkreutz A, Boucher L, Breitkreutz B, Hon G *et al.* Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, 5:11, 2006. 1.1.1.3

Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17:1030–1032, 1999. 1.1.1.1

Riley R, Lee C, Sabatti C and Eisenberg D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 6:R89, 2005. 2.1, 2.4.1

Risch N and Merikangas K. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996. 1.2.3.1

Ross ED, Minton A and Wickner RB. Prion domains: sequences, structures and interactions. *Nat Cell Biol*, 7:1039–1044, 2005. 3.1

Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437:1173–1178, 2005. 1.1.2, 4.2.3

Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 36:D646–650, 2008. 4.2.2

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 32:D449–451, 2004. 1.1.3, 2.2.1, 3.3.3

Sammut SJ, Finn RD and Bateman A. Pfam 10 years on: 10,000 families and still growing. *Brief Bioinform*, 9:210–219, 2008. 1.3

Sanchez C, Lachaize C, Janody F, Bellon B, Roder L, Euzenat J, Rechenmann F and Jacq B. Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an internet database. *Nucleic Acids Res*, 27:89–94, 1999. 1.1.4

Sankar P and Cho MK. Genetics. toward a new vocabulary of human genetic variation. *Science*, 298:1337–1338, 2002. 1.2.2

Savas S, Tuzmen S and Ozcelik H. Human SNPs resulting in premature stop codons and protein truncation. *Hum Genomics*, 2:274–286, 2006. 3.1

Schuster-Böckler B and Bateman A. Visualizing profile–profile alignment: pairwise HMM logos. *Bioinformatics*, 21:2912–2913, 2005. 1.4

Schuster-Böckler B and Bateman A. An introduction to hidden Markov models. *Curr Protoc Bioinformatics*, A3, 2007a. 1.3

Schuster-Böckler B and Bateman A. Reuse of structural domain-domain interactions in protein networks. *BMC Bioinformatics*, 8:259, 2007b. 1.4

Schuster-Böckler B and Bateman A. Protein interactions in human genetic diseases. *Genome Biol*, 9:9, 2008. 1.4

Schuster-Böckler B, Schultz J and Rahmann S. HMM Logos for visualization of protein families. *BMC Bioinformatics*, 5, 2004. 3.2.5

Sebat J, Lakshmi B, Troge J, Alexander J, Young J *et al.* Large-scale copy number polymorphism in the human genome. *Science*, 305:525–528, 2004. 1.2.3

Seto ML, Lee SJ, Sze RW and Cunningham ML. Another TWIST on Baller-Gerold syndrome. *Am J Med Genet*, 104:323–330, 2001. 3.3.6.3

Shany E, Saada A, Landau D, Shaag A, Hershkovitz E and Elpeleg ON. Lipoamide dehydrogenase deficiency due to a novel mutation in the interface domain. *Biochem Biophys Res Commun*, 262:163–166, 1999. 3.3.4.2

Sharan R and Ideker T. Modeling cellular machinery through biological network comparison. *Nat Biotech*, 24:427–433, 2006. 1

She X, Jiang Z, Clark RA, Liu G, Cheng Z *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, 431:927–930, 2004. 4.2.8

Shinawi M and Cheung SW. The array CGH and its clinical applications. *Drug Discov Today*, 13:760–770, 2008. 1.2.3

Shy ME, Jani A, Krajewski K, Grandis M, Lewis RA *et al.* Phenotypic clustering in MPZ mutations. *Brain*, 127:371–384, 2004. 3.1

Sidhu SS, Fairbrother WJ and Deshayes K. Exploring protein-protein interactions with phage display. *Chembiochem*, 4:14–25, 2003. 1.1.1.5

Sing T, Sander O, Beerenwinkel N and Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21:3940–3941, 2005. 3.6

Siva N. 1000 Genomes project. *Nat Biotechnol*, 26:256, 2008. 1.2.2

Smith EA and Corn RM. Surface plasmon resonance imaging as a tool to monitor biomolecular interactions in an array based format. *Appl Spectrosc*, 57:320A–332A, 2003. 1.1.1.5

Sopko R, Huang D, Preston N, Chua G, Papp B *et al.* Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell*, 21:319–330, 2006. 4.1

Sprinzak E, Sattath S and Margalit H. How reliable are experimental protein–protein interaction data? *J Mol Biol*, 327:919–923, 2003. 1.1.2

Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS *et al.* Systematic screen for human disease genes in yeast. *Nat Genet*, 31:400–404, 2002. 4.1

Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH *et al.* A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–968, 2005. 4.2.3

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315:848–853, 2007. 4.2.5, 4.3.1

Strom M, Hume AN, Tarafder AK, Barkagianni E and Seabra MC. A family of Rab27-binding proteins. melanophilin links Rab27a and myosin Va function in melanosome transport. *J Biol Chem*, 277:25423–25430, 2002. 3.3.6.1

Stumpf MPH, Thorne T, de Silva E, Stewart R, An HJ, Lappe M and Wiuf C. Estimating the size of the human interactome. *Proc Natl Acad Sci USA*, 105:6959–6964, 2008. 1.1.2

Stumpf MPH, Wiuf C and May RM. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc Natl Acad Sci USA*, 102:4221–4224, 2005. 1.1.4

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*, 101:6062–6067, 2004. 4.2.5

Svedberg T. Mass and size of protein molecules. *Nature*, 123:871, 1929. 1.1

The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003. 1.2.3

Thorn KS and Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17:284–285, 2001. 1.1, 3.2.6, 3.3.2

Tong AHY, Lesage G, Bader GD, Ding H, Xu H *et al.* Global mapping of the yeast genetic interaction network. *Science*, 303:808–813, 2004. 1.1.1.5

Uetz P, Giot L, Cagney G, Mansfield T, Judson R *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000. 1.1.2, 1.1.4

Veitia RA. Exploring the etiology of haploinsufficiency. *BioEssays*, 24:175–184, 2002. 3.4.3

Veitia RA. Gene dosage balance: deletions, duplications and dominance. *Trends Genet*, 21:33–35, 2005. 4.1

Velankar S, McNeil P, Mittard-Runte V, Suarez A, Barrell D, Apweiler R and Henrick K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*, 33:D262–265, 2005. 1.3.1

Vitkup D, Sander C and Church GM. The amino-acid mutational spectrum of human genetic disease. *Genome Biol*, 4:R72, 2003. 3.3.4.4

Walsh STR and Kossiakoff AA. Crystal structure and site 1 binding energetics of human placental lactogen. *J Mol Biol*, 358:773–784, 2006. 3.2.6

Wang Z and Moult J. SNPs, protein structure, and disease. *Hum Mutat*, 17:263–270, 2001. 3.1

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007. 1.2.3.1

Williams AD, Shivaprasad S and Wetzel R. Alanine scanning mutagenesis of Aβ(1–40) amyloid fibril stability. *J Mol Biol*, 357:1283–1294, 2006. 3.2.6

Wolf YI, Rogozin IB and Koonin EV. Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res*, 14:29–36, 2004. 1.3

Wong JMS, Ionescu D and Ingles CJ. Interaction between BRCA2 and replication protein A is compromised by a cancer-predisposing mutation in BRCA2. *Oncogene*, 22:28–33, 2003. 3.3.4.2

Wright S. Physiological and evolutionary theories of dominance. *Am Nat*, 68:24–53, 1934. 1.2.3.1, 4.1

Wu C, Apweiler R, Bairoch A, Natale D, Barker W *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 34:D187–191, 2006. 1.3, 3.1, 3.2.1

Yang J, Lusk R and Li WH. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci USA*, 100:15661–15665, 2003. 4.1, 4.4.3

Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13:555–556, 1997. 4.2.11

Yue P, Li Z and Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol*, 353:459–473, 2005. 3.1

Zdobnov EM, Lopez R, Apweiler R and Etzold T. The EBI SRS server — Recent developments. *Bioinformatics*, 18:368–373, 2002. 3.2.1

Zhu H, Domingues FS, Sommer I and Lengauer T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:27, 2006. 2.2.5, 3.2.3

Zuckerkandl E and Pauling L. Molecular disease, evolution, and genetic heterogeneity. *Horiz Biochem*, 189–225, 1962. 1.2.1

Appendices

Appendix A

Table A.1: 20 most frequent *i*Pfam domain pairs in protein interactions of 5 species.

Accession A	**Name A**	**Accession B**	**Name B**	**Frequency**
		E. coli		
PF00005	ABC_tran	PF00005	ABC_tran	21
PF00072	Response_reg	PF00072	Response_reg	19
PF00126	HTH_1	PF00126	HTH_1	17
PF03466	LysR_substrate	PF00126	HTH_1	16
PF03466	LysR_substrate	PF03466	LysR_substrate	16
PF00271	Helicase_C	PF00271	Helicase_C	15
PF00313	CSD	PF00313	CSD	14
PF00106	adh_short	PF00106	adh_short	12
PF00532	Peripla_BP_1	PF00532	Peripla_BP_1	11
PF00293	NUDIX	PF00293	NUDIX	10
PF00271	Helicase_C	PF00270	DEAD	10
PF00216	Bac_DNA_binding	PF00216	Bac_DNA_binding	9
PF00392	GntR	PF00392	GntR	9
PF00575	S1	PF00575	S1	9
PF00009	GTP_EFTU	PF00009	GTP_EFTU	9
PF00158	Sigma54_activat	PF00158	Sigma54_activat	9
PF02518	HATPase_c	PF02518	HATPase_c	9
PF03144	GTP_EFTU_D2	PF03144	GTP_EFTU_D2	9
PF03144	GTP_EFTU_D2	PF00009	GTP_EFTU	9
PF00270	DEAD	PF00270	DEAD	9
		S. cerevisiae		
PF00069	Pkinase	PF00069	Pkinase	266
PF00400	WD40	PF00400	WD40	141
PF00227	Proteasome	PF00227	Proteasome	96
PF01423	LSM	PF01423	LSM	84
PF00076	RRM_1	PF00076	RRM_1	79
PF00271	Helicase_C	PF00271	Helicase_C	74
PF00134	Cyclin_N	PF00069	Pkinase	65

Accession A	Name A	Accession B	Name B	Frequency
PF00271	Helicase_C	PF00270	DEAD	51
PF00018	SH3_1	PF00018	SH3_1	49
PF00004	AAA	PF00004	AAA	46
PF00270	DEAD	PF00270	DEAD	41
PF02984	Cyclin_C	PF00069	Pkinase	35
PF00069	Pkinase	PF00023	Ank	32
PF00433	Pkinase_C	PF00069	Pkinase	30
PF00172	Zn_clus	PF00172	Zn_clus	27
PF05739	SNARE	PF00957	Synaptobrevin	26
PF02985	HEAT	PF02985	HEAT	25
PF00125	Histone	PF00125	Histone	24
PF00271	Helicase_C	PF00176	SNF2_N	21
PF00575	S1	PF00069	Pkinase	20
		C. elegans		
PF00105	zf-C4	PF00105	zf-C4	33
PF00104	Hormone_recep	PF00104	Hormone_recep	31
PF00105	zf-C4	PF00104	Hormone_recep	31
PF00595	PDZ	PF00595	PDZ	12
PF00076	RRM_1	PF00076	RRM_1	12
PF00227	Proteasome	PF00227	Proteasome	11
PF00069	Pkinase	PF00069	Pkinase	11
PF02932	Neur_chan_memb	PF02932	Neur_chan_memb	9
PF02931	Neur_chan_LBD	PF02931	Neur_chan_LBD	9
PF02932	Neur_chan_memb	PF02931	Neur_chan_LBD	9
PF00004	AAA	PF00004	AAA	6
PF00412	LIM	PF00018	SH3_1	5
PF01423	LSM	PF01423	LSM	5
PF02188	GoLoco	PF00503	G-alpha	5
PF00651	BTB	PF00651	BTB	4
PF01849	NAC	PF01849	NAC	4
PF00412	LIM	PF00412	LIM	4
PF00595	PDZ	PF00071	Ras	4
PF01849	NAC	PF00627	UBA	4
PF01466	Skp1	PF00646	F-box	4
		D. melanogaster		
PF00096	zf-C2H2	PF00096	zf-C2H2	117
PF00134	Cyclin_N	PF00069	Pkinase	63
PF00076	RRM_1	PF00076	RRM_1	54
PF00010	HLH	PF00010	HLH	54
PF01466	Skp1	PF00646	F-box	48
PF00069	Pkinase	PF00069	Pkinase	38
PF00595	PDZ	PF00071	Ras	22
PF00788	RA	PF00071	Ras	21
PF02984	Cyclin_C	PF00069	Pkinase	21
PF00612	IQ	PF00036	efhand	21
PF00179	UQ_con	PF00097	zf-C3HC4	21
PF00046	Homeobox	PF00046	Homeobox	20

Accession A	Name A	Accession B	Name B	Frequency
PF00069	Pkinase	PF00023	Ank	20
PF00651	BTB	PF00651	BTB	14
PF01423	LSM	PF01423	LSM	14
PF00063	Myosin_head	PF00036	efhand	13
PF00134	Cyclin_N	PF00134	Cyclin_N	11
PF00018	SH3_1	PF00017	SH2	11
PF03931	Skp1_POZ	PF00560	LRR_1	10
PF02179	BAG	PF00012	HSP70	10
		H. sapiens		
PF07714	Pkinase_Tyr	PF00017	SH2	464
PF00069	Pkinase	PF00069	Pkinase	386
PF07714	Pkinase_Tyr	PF00018	SH3_1	318
PF00018	SH3_1	PF00017	SH2	241
PF00017	SH2	PF00017	SH2	200
PF00018	SH3_1	PF00018	SH3_1	179
PF07714	Pkinase_Tyr	PF07714	Pkinase_Tyr	162
PF00076	RRM_1	PF00076	RRM_1	147
PF00433	Pkinase_C	PF00069	Pkinase	112
PF00010	HLH	PF00010	HLH	95
PF00069	Pkinase	PF00023	Ank	74
PF00105	zf-C4	PF00104	Hormone_recep	72
PF00104	Hormone_recep	PF00104	Hormone_recep	71
PF00096	zf-C2H2	PF00096	zf-C2H2	71
PF00089	Trypsin	PF00079	Serpin	66
PF00105	zf-C4	PF00105	zf-C4	60
PF07714	Pkinase_Tyr	PF00102	Y_phosphatase	58
PF00169	PH	PF00071	Ras	56
PF00046	Homeobox	PF00046	Homeobox	54
PF00619	CARD	PF00619	CARD	54

Table A.2: 20 most frequent *i*Pfam domain pairs in protein interactions of 5 species, excluding intrachain structures.

Accession A	Name A	Accession B	Name B	Frequency
		E. coli		
PF00005	ABC_tran	PF00005	ABC_tran	21
PF00072	Response_reg	PF00072	Response_reg	19
PF00126	HTH_1	PF00126	HTH_1	17
PF03466	LysR_substrate	PF00126	HTH_1	16
PF03466	LysR_substrate	PF03466	LysR_substrate	16
PF00271	Helicase_C	PF00271	Helicase_C	15
PF00313	CSD	PF00313	CSD	14
PF00106	adh_short	PF00106	adh_short	12
PF00532	Peripla_BP_1	PF00532	Peripla_BP_1	11

Accession A	Name A	Accession B	Name B	Frequency
PF00293	NUDIX	PF00293	NUDIX	10
PF00392	GntR	PF00392	GntR	9
PF02518	HATPase_c	PF02518	HATPase_c	9
PF00575	S1	PF00575	S1	9
PF00009	GTP_EFTU	PF00009	GTP_EFTU	9
PF00158	Sigma54_activat	PF00158	Sigma54_activat	9
PF00270	DEAD	PF00270	DEAD	9
PF03144	GTP_EFTU_D2	PF03144	GTP_EFTU_D2	9
PF00216	Bac_DNA_binding	PF00216	Bac_DNA_binding	9
PF03144	GTP_EFTU_D2	PF00009	GTP_EFTU	9
PF00004	AAA	PF00004	AAA	8
		S. cerevisiae		
PF00069	Pkinase	PF00069	Pkinase	266
PF00400	WD40	PF00400	WD40	141
PF00227	Proteasome	PF00227	Proteasome	96
PF01423	LSM	PF01423	LSM	84
PF00076	RRM_1	PF00076	RRM_1	79
PF00271	Helicase_C	PF00271	Helicase_C	74
PF00134	Cyclin_N	PF00069	Pkinase	65
PF00018	SH3_1	PF00018	SH3_1	49
PF00004	AAA	PF00004	AAA	46
PF00270	DEAD	PF00270	DEAD	41
PF02984	Cyclin_C	PF00069	Pkinase	35
PF00069	Pkinase	PF00023	Ank	32
PF00433	Pkinase_C	PF00069	Pkinase	30
PF00172	Zn_clus	PF00172	Zn_clus	27
PF05739	SNARE	PF00957	Synaptobrevin	26
PF02985	HEAT	PF02985	HEAT	25
PF00125	Histone	PF00125	Histone	24
PF00575	S1	PF00069	Pkinase	20
PF01138	RNase_PH	PF01138	RNase_PH	19
PF03725	RNase_PH_C	PF01138	RNase_PH	19
		C. elegans		
PF00105	zf-C4	PF00105	zf-C4	33
PF00105	zf-C4	PF00104	Hormone_recep	31
PF00104	Hormone_recep	PF00104	Hormone_recep	31
PF00076	RRM_1	PF00076	RRM_1	12
PF00595	PDZ	PF00595	PDZ	12
PF00227	Proteasome	PF00227	Proteasome	11
PF00069	Pkinase	PF00069	Pkinase	11
PF02932	Neur_chan_memb	PF02932	Neur_chan_memb	9
PF02931	Neur_chan_LBD	PF02931	Neur_chan_LBD	9
PF02932	Neur_chan_memb	PF02931	Neur_chan_LBD	9
PF00004	AAA	PF00004	AAA	6
PF01423	LSM	PF01423	LSM	5
PF00412	LIM	PF00018	SH3_1	5
PF02188	GoLoco	PF00503	G-alpha	5

Accession A	Name A	Accession B	Name B	Frequency
PF00595	PDZ	PF00071	Ras	4
PF01849	NAC	PF01849	NAC	4
PF01466	Skp1	PF00646	F-box	4
PF00651	BTB	PF00651	BTB	4
PF00210	Ferritin	PF00210	Ferritin	3
PF00017	SH2	PF00017	SH2	3
		D. melanogaster		
PF00096	zf-C2H2	PF00096	zf-C2H2	117
PF00134	Cyclin_N	PF00069	Pkinase	63
PF00076	RRM_1	PF00076	RRM_1	54
PF00010	HLH	PF00010	HLH	54
PF01466	Skp1	PF00646	F-box	48
PF00069	Pkinase	PF00069	Pkinase	38
PF00595	PDZ	PF00071	Ras	22
PF02984	Cyclin_C	PF00069	Pkinase	21
PF00179	UQ_con	PF00097	zf-C3HC4	21
PF00788	RA	PF00071	Ras	21
PF00612	IQ	PF00036	efhand	21
PF00046	Homeobox	PF00046	Homeobox	20
PF00069	Pkinase	PF00023	Ank	20
PF00651	BTB	PF00651	BTB	14
PF01423	LSM	PF01423	LSM	14
PF00063	Myosin_head	PF00036	efhand	13
PF00134	Cyclin_N	PF00134	Cyclin_N	11
PF00018	SH3_1	PF00017	SH2	11
PF02179	BAG	PF00012	HSP70	10
PF02196	RBD	PF00071	Ras	10
		H. sapiens		
PF07714	Pkinase_Tyr	PF00017	SH2	464
PF00069	Pkinase	PF00069	Pkinase	386
PF00018	SH3_1	PF00017	SH2	241
PF00017	SH2	PF00017	SH2	200
PF00018	SH3_1	PF00018	SH3_1	179
PF07714	Pkinase_Tyr	PF07714	Pkinase_Tyr	162
PF00076	RRM_1	PF00076	RRM_1	147
PF00433	Pkinase_C	PF00069	Pkinase	112
PF00010	HLH	PF00010	HLH	95
PF00069	Pkinase	PF00023	Ank	74
PF00105	zf-C4	PF00104	Hormone_recep	72
PF00104	Hormone_recep	PF00104	Hormone_recep	71
PF00096	zf-C2H2	PF00096	zf-C2H2	71
PF00089	Trypsin	PF00079	Serpin	66
PF00105	zf-C4	PF00105	zf-C4	60
PF07714	Pkinase_Tyr	PF00102	Y_phosphatase	58
PF00169	PH	PF00071	Ras	56
PF00046	Homeobox	PF00046	Homeobox	54
PF00619	CARD	PF00619	CARD	54

Accession A	Name A	Accession B	Name B	Frequency
PF00531	Death	PF00531	Death	53

Appendix B

Table B.1: All structures

E coli			Yeast			Worm			Fly			Human		
Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.
PF00005	ABC_tran	49	PF00069	Pkinase	113	PF00069	Pkinase	82	PF00096	zf-C2H2	234	PF00069	Pkinase	346
PF00072	Response_reg	36	PF00400	WD40	87	PF00105	zf-C4	63	PF00069	Pkinase	171	PF00096	zf-C2H2	205
PF02518	HATPase_c	31	PF00271	Helicase_C	70	PF00104	Hormone_recep	57	PF00076	RRM_1	124	PF00169	PH	174
PF00126	HTH_1	30	PF00172	Zn_clus	49	PF01391	Collagen	46	PF00400	WD40	97	PF00076	RRM_1	172
PF03466	LysR_substrate	28	PF00270	DEAD	48	PF00076	RRM_1	39	PF00046	Homeobox	79	PF00018	SH3_1	170
PF00672	HAMP	20	PF00076	RRM_1	47	PF00400	WD40	32	PF00089	Trypsin	72	PF00400	WD40	151
PF00512	HisKA	19	PF00096	zf-C2H2	36	PF00595	PDZ	28	PF00036	efhand	67	PF00001	7tm_1	145
PF00486	Trans_reg_C	16	PF00004	AAA	33	PF00097	zf-C3HC4	27	PF00097	zf-C3HC4	65	PF00595	PDZ	132
PF00532	Peripla_BP_1	16	PF00005	ABC_tran	30	PF00096	zf-C2H2	25	PF00595	PDZ	65	PF00097	zf-C3HC4	132
PF00271	Helicase_C	15	PF00153	Mito_carr	30	PF02798	GST_N	25	PF00023	Ank	64	PF00047	ig	131
PF00392	GntR	14	PF02985	HEAT	26	PF00043	GST_C	23	PF00018	SH3_1	60	PF00023	Ank	117
PF00106	adh_short	13	PF00071	Ras	24	PF00646	F-box	23	PF00560	LRR_1	60	PF00017	SH2	109
PF00158	Sigma54_activat	13	PF00169	PH	24	PF00651	BTB	22	PF00271	Helicase_C	57	PF00041	fn3	109
PF00196	GerE	13	PF00018	SH3_1	22	PF00018	SH3_1	20	PF00047	ig	56	PF00036	efhand	107
PF00037	Fer4	12	PF00702	Hydrolase	22	PF00169	PH	20	PF00169	PH	54	PF07686	V-set	104
PF00165	HTH_AraC	12	PF00665	rve	21	PF00023	Ank	19	PF07679	I-set	52	PF07714	Pkinase_Tyr	99
PF04055	Radical_SAM	11	PF00097	zf-C3HC4	21	PF00004	AAA	19	PF00651	BTB	51	PF00008	EGF	92
PF00293	NUDIX	11	PF00515	TPR_1	21	PF00271	Helicase_C	19	PF00010	HLH	45	PF00046	Homeobox	92
PF00455	DeoR	10	PF00149	Metallophos	21	PF00046	Homeobox	19	PF00041	fn3	43	PF00560	LRR_1	91
PF02954	HTH_8	10	PF00226	DnaJ	21	PF00149	Metallophos	19	PF00515	TPR_1	43	PF00168	C2	87

Table B.2: Interchain only

E coli			Yeast			Worm			Fly			Human		
Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.
PF00005	ABC_tran	49	PF00069	Pkinase	113	PF00069	Pkinase	82	PF00096	zf-C2H2	234	PF00069	Pkinase	346
PF00072	Response_reg	36	PF00400	WD40	87	PF00105	zf-C4	63	PF00069	Pkinase	171	PF00096	zf-C2H2	205
PF02518	HATPase_c	31	PF00271	Helicase_C	70	PF00104	Hormone_recep	57	PF00076	RRM_1	124	PF00169	PH	174
PF00126	HTH_1	30	PF00172	Zn_clus	49	PF01391	Collagen	46	PF00400	WD40	97	PF00076	RRM_1	172
PF03466	LysR_substrate	28	PF00270	DEAD	48	PF00076	RRM_1	39	PF00046	Homeobox	79	PF00018	SH3_1	170
PF00672	HAMP	20	PF00076	RRM_1	47	PF00400	WD40	32	PF00089	Trypsin	72	PF00400	WD40	151
PF00486	Trans_reg_C	16	PF00096	zf-C2H2	36	PF00595	PDZ	28	PF00036	efhand	67	PF00001	7tm_1	145
PF00532	Peripla_BP_1	16	PF00004	AAA	33	PF00097	zf-C3HC4	27	PF00097	zf-C3HC4	65	PF00595	PDZ	132
PF00271	Helicase_C	15	PF00005	ABC_tran	30	PF02798	GST_N	25	PF00595	PDZ	65	PF00097	zf-C3HC4	132
PF00392	GntR	14	PF02985	HEAT	26	PF00096	zf-C2H2	25	PF00023	Ank	64	PF00047	ig	131
PF00106	adh_short	13	PF00071	Ras	24	PF00043	GST_C	23	PF00018	SH3_1	60	PF00023	Ank	117
PF00158	Sigma54_activat	13	PF00169	PH	24	PF00646	F-box	23	PF00560	LRR_1	60	PF00017	SH2	109
PF00196	GerE	13	PF00018	SH3_1	22	PF00651	BTB	22	PF00271	Helicase_C	57	PF00041	fn3	109
PF00165	HTH_AraC	12	PF00702	Hydrolase	22	PF00018	SH3_1	20	PF00047	ig	56	PF00036	efhand	107
PF00037	Fer4	12	PF00226	DnaJ	21	PF00169	PH	20	PF00169	PH	54	PF07686	V-set	104
PF00293	NUDIX	11	PF00097	zf-C3HC4	21	PF00023	Ank	19	PF07679	I-set	52	PF07714	Pkinase_Tyr	99
PF00270	DEAD	10	PF00665	rve	21	PF00271	Helicase_C	19	PF00651	BTB	51	PF00008	EGF	92
PF00455	DeoR	10	PF00149	Metallophos	21	PF00004	AAA	19	PF00010	HLH	45	PF00046	Homeobox	92
PF02954	HTH_8	10	PF00515	TPR_1	21	PF00149	Metallophos	19	PF00041	fn3	43	PF00560	LRR_1	91
PF00155	Aminotran_1_2	10	PF08240	ADH_N	20	PF00046	Homeobox	19	PF00515	TPR_1	43	PF00168	C2	87

Table B.3: No crystal contacts

E coli			Yeast			Worm			Fly			Human		
Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.	Accession	Name	Freq.
PF00005	ABC_tran	49	PF00069	Pkinase	113	PF00069	Pkinase	82	PF00096	zf-C2H2	234	PF00069	Pkinase	346
PF00072	Response_reg	36	PF00400	WD40	87	PF00105	zf-C4	63	PF00069	Pkinase	171	PF00096	zf-C2H2	205
PF02518	HATPase_c	31	PF00271	Helicase_C	70	PF00104	Hormone_recep	57	PF00076	RRM_1	124	PF00169	PH	174
PF00126	HTH_1	30	PF00172	Zn_clus	49	PF01391	Collagen	46	PF00400	WD40	97	PF00076	RRM_1	172
PF03466	LysR_substrate	28	PF00270	DEAD	48	PF00076	RRM_1	39	PF00046	Homeobox	79	PF00018	SH3_1	170
PF00672	HAMP	20	PF00076	RRM_1	47	PF00400	WD40	32	PF00089	Trypsin	72	PF00400	WD40	151
PF00532	Peripla_BP_1	16	PF00096	zf-C2H2	36	PF00595	PDZ	28	PF00036	efhand	67	PF00001	7tm_1	145
PF00271	Helicase_C	15	PF00004	AAA	33	PF00097	zf-C3HC4	27	PF00595	PDZ	65	PF00097	zf-C3HC4	132
PF00392	GntR	14	PF00005	ABC_tran	30	PF02798	GST_N	25	PF00097	zf-C3HC4	65	PF00595	PDZ	132
PF00106	adh_short	13	PF02985	HEAT	26	PF00096	zf-C2H2	25	PF00023	Ank	64	PF00047	ig	131
PF00158	Sigma54_activat	13	PF00071	Ras	24	PF00043	GST_C	23	PF00018	SH3_1	60	PF00023	Ank	117
PF00196	GerE	13	PF00169	PH	24	PF00646	F-box	23	PF00560	LRR_1	60	PF00017	SH2	109
PF00165	HTH_AraC	12	PF00018	SH3_1	22	PF00651	BTB	22	PF00271	Helicase_C	57	PF00041	fn3	109
PF00037	Fer4	12	PF00702	Hydrolase	22	PF00018	SH3_1	20	PF00047	ig	56	PF00036	efhand	107
PF00293	NUDIX	11	PF00665	rve	21	PF00169	PH	20	PF00169	PH	54	PF07686	V-set	104
PF00155	Aminotran_1_2	10	PF00097	zf-C3HC4	21	PF00023	Ank	19	PF07679	I-set	52	PF07714	Pkinase_Tyr	99
PF00171	Aldedh	10	PF00149	Metallophos	21	PF00004	AAA	19	PF00651	BTB	51	PF00046	Homeobox	92
PF00270	DEAD	10	PF00515	TPR_1	21	PF00046	Homeobox	19	PF00010	HLH	45	PF00008	EGF	92
PF02954	HTH_8	10	PF08240	ADH_N	20	PF00149	Metallophos	19	PF00041	fn3	43	PF00560	LRR_1	91
PF07992	Pyr_redox_2	10	PF00107	ADH_zinc_N	17	PF00271	Helicase_C	19	PF00515	TPR_1	43	PF00168	C2	87

Appendix C

Table C.1: The 30 most frequent *i*Pfam domain architectures per species. The left column lists the sequence of *i*Pfam domains that comprises a distinct domain architecture, separated by a "—". Non-*i*Pfam domains are omitted to underline the effect of domain architecture on *i*Pfam domain pair frequency. The right column contains the frequency of the architecture, defined as the number of sequences which share the same architecture.

Architecture	**Frequency**
E. coli	
ABC_tran	32
HTH_1 — LysR_substrate	28
Peripla_BP_1	15
adh_short	13
Response_reg — Trans_reg_C	13
ABC_tran — ABC_tran	11
DeoR	10
NUDIX	10
HAMP — HisKA — HATPase_c	9
Aldedh	9
Aminotran_1_2	8
DEAD — Helicase_C	8
Response_reg — GerE	7
CSD	7
GntR	7
S1	6
Response_reg	6
TPP_enzyme_N — TPP_enzyme_M — TPP_enzyme_C	6
ADH_N — ADH_zinc_N	6
Aminotran_3	6
Fe-ADH	6
GerE	6
Acetyltransf_1	6
Glycos_transf_2	6
Hydrolase	6
Radical_SAM	6

Architecture	Frequency
4HBT	5
NTP_transferase	5
Hydrolase_3	5
Pribosyltran	5
S. cerevisiae	
Pkinase	93
Zn_clus	47
DEAD — Helicase_C	42
RRM_1	29
Mito_carr — Mito_carr — Mito_carr	26
Ras	24
zf-C2H2 — zf-C2H2	22
rve	20
Metallophos	20
WD40 — WD40	18
ADH_N — ADH_zinc_N	16
LSM	16
WD40	14
Aldo_ket_red	14
PH	14
Proteasome	14
HSP70	14
DnaJ	13
AAA	13
UQ_con	13
zf-C3HC4	13
Abhydrolase_1	12
WD40 — WD40 — WD40 — WD40 — WD40	11
ABC_tran — ABC_tran	11
SH3_1	11
adh_short	11
WD40 — WD40 — WD40 — WD40	11
Aminotran_1_2	11
Acetyltransf_1	11
Hydrolase	11
C. elegans	
Pkinase	60
zf-C4 — Hormone_recep	56
Collagen — Collagen — Collagen	30
zf-C3HC4	23
RRM_1	22
GST_N — GST_C	22
F-box	20
Metallophos	18
Collagen — Collagen	15
Homeobox	15
Kinesin	14

Architecture	**Frequency**
AAA	14
p450	13
K_tetra	12
Proteasome	12
Neur_chan_LBD — Neur_chan_memb	11
BTB	11
Motile_Sperm	10
Filament	10
Ras	10
Arrestin_N — Arrestin_C	10
zf-CCCH — zf-CCCH	10
zf-C2H2	9
ubiquitin	9
MATH — BTB	9
COesterase	9
7tm_1	9
Aminotran_1_2	8
RRM_1 — RRM_1 — RRM_1	8
DEAD — Helicase_C	8
D. melanogaster	
Pkinase	109
Trypsin	66
RRM_1	55
Homeobox	55
zf-C3HC4	45
RRM_1 — RRM_1	44
HLH	42
Ras	38
zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2	37
zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2	36
p450	34
UQ_con	29
DEAD — Helicase_C	28
GST_N — GST_C	27
BTB	26
adh_short	25
zf-C2H2 — zf-C2H2 — zf-C2H2	24
Proteasome	23
7tm_1	22
Metallophos	22
zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2	21
Kinesin	20
zf-C2H2	20
zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2 — zf-C2H2	19
zf-C4 — Hormone_recep	19

Architecture	**Frequency**
COesterase	17
AMP-binding	17
zf-C2H2 — zf-C2H2	16
Tetraspannin	16
efhand — efhand — efhand	16
H. sapiens	
Pkinase	200
7tm_1	141
Ras	84
zf-C3HC4	76
RRM_1	68
Homeobox	66
HLH	57
zf-C4 — Hormone_recep	52
RRM_1 — RRM_1	51
IL8	41
Filament	41
DEAD — Helicase_C	38
SH3_1	37
zf-C2H2 — zf-C2H2 — zf-C2H2	32
UQ_con	31
K_tetra	30
PH	28
PDZ	28
MHC_I — C1-set	27
SH2	26
V-set	26
Trypsin	25
UCH	24
Lectin_C	23
C2 — C2	23
TGF_beta	23
WD40 — WD40 — WD40 — WD40 — WD40 — WD40	22
bZIP_1	22
Kinesin	22
RRM_1 — RRM_1 — RRM_1	22

Appendix D

Table D.1: All *i*Pfam domain pairs that are shared between *E. coli*, *S. cerevisiae* and *H. sapiens*.

Accession A	**Name A**	**Accession B**	**Name B**
PF00004	AAA	PF00004	AAA
PF00005	ABC_tran	PF00005	ABC_tran
PF00009	GTP_EFTU	PF00009	GTP_EFTU
PF00011	HSP20	PF00011	HSP20
PF00013	KH_1	PF00013	KH_1
PF00027	cNMP_binding	PF00027	cNMP_binding
PF00035	dsrm	PF00035	dsrm
PF00043	GST_C	PF00043	GST_C
PF00044	Gp_dh_N	PF00044	Gp_dh_N
PF00056	Ldh_1_N	PF00056	Ldh_1_N
PF00085	Thioredoxin	PF00085	Thioredoxin
PF00091	Tubulin	PF00091	Tubulin
PF00106	adh_short	PF00106	adh_short
PF00107	ADH_zinc_N	PF00107	ADH_zinc_N
PF00117	GATase	PF00117	GATase
PF00118	Cpn60_TCP1	PF00118	Cpn60_TCP1
PF00132	Hexapep	PF00132	Hexapep
PF00149	Metallophos	PF00149	Metallophos
PF00155	Aminotran_1_2	PF00155	Aminotran_1_2
PF00156	Pribosyltran	PF00156	Pribosyltran
PF00160	Pro_isomerase	PF00160	Pro_isomerase
PF00166	Cpn10	PF00118	Cpn60_TCP1
PF00171	Aldedh	PF00171	Aldedh
PF00180	Iso_dh	PF00180	Iso_dh
PF00183	HSP90	PF00183	HSP90
PF00185	OTCace	PF00185	OTCace
PF00199	Catalase	PF00199	Catalase
PF00202	Aminotran_3	PF00202	Aminotran_3
PF00204	DNA_gyraseB	PF00204	DNA_gyraseB
PF00205	TPP_enzyme_M	PF00205	TPP_enzyme_M
PF00206	Lyase_1	PF00206	Lyase_1

Accession A	Name A	Accession B	Name B
PF00208	ELFV_dehydrog	PF00208	ELFV_dehydrog
PF00224	PK	PF00224	PK
PF00227	Proteasome	PF00227	Proteasome
PF00254	FKBP_C	PF00254	FKBP_C
PF00258	Flavodoxin_1	PF00258	Flavodoxin_1
PF00270	DEAD	PF00270	DEAD
PF00271	Helicase_C	PF00176	SNF2_N
PF00271	Helicase_C	PF00270	DEAD
PF00271	Helicase_C	PF00271	Helicase_C
PF00288	GHMP_kinases_N	PF00288	GHMP_kinases_N
PF00289	CPSase_L_chain	PF00289	CPSase_L_chain
PF00291	PALP	PF00291	PALP
PF00293	NUDIX	PF00293	NUDIX
PF00300	PGAM	PF00300	PGAM
PF00317	Ribonuc_red_lgN	PF00317	Ribonuc_red_lgN
PF00328	Acid_phosphat_A	PF00328	Acid_phosphat_A
PF00334	NDK	PF00334	NDK
PF00365	PFK	PF00365	PFK
PF00378	ECH	PF00378	ECH
PF00383	dCMP_cyt_deam_1	PF00383	dCMP_cyt_deam_1
PF00389	2-Hacid_dh	PF00389	2-Hacid_dh
PF00438	S-AdoMet_synt_N	PF00438	S-AdoMet_synt_N
PF00448	SRP54	PF00448	SRP54
PF00456	Transketolase_N	PF00456	Transketolase_N
PF00462	Glutaredoxin	PF00462	Glutaredoxin
PF00479	G6PD_N	PF00479	G6PD_N
PF00483	NTP_transferase	PF00483	NTP_transferase
PF00488	MutS_V	PF00488	MutS_V
PF00491	Arginase	PF00491	Arginase
PF00515	TPR_1	PF00515	TPR_1
PF00533	BRCT	PF00533	BRCT
PF00534	Glycos_transf_1	PF00534	Glycos_transf_1
PF00542	Ribosomal_L12	PF00542	Ribosomal_L12
PF00570	HRDC	PF00570	HRDC
PF00571	CBS	PF00571	CBS
PF00578	AhpC-TSA	PF00578	AhpC-TSA
PF00583	Acetyltransf_1	PF00583	Acetyltransf_1
PF00586	AIRS	PF00586	AIRS
PF00587	tRNA-synt_2b	PF00587	tRNA-synt_2b
PF00596	Aldolase_II	PF00596	Aldolase_II
PF00625	Guanylate_kin	PF00625	Guanylate_kin
PF00627	UBA	PF00009	GTP_EFTU
PF00627	UBA	PF00627	UBA
PF00636	Ribonuclease_3	PF00035	dsrm
PF00664	ABC_membrane	PF00005	ABC_tran
PF00664	ABC_membrane	PF00664	ABC_membrane
PF00676	E1_dh	PF00676	E1_dh
PF00679	EFG_C	PF00009	GTP_EFTU

Accession A	Name A	Accession B	Name B
PF00702	Hydrolase	PF00702	Hydrolase
PF00731	AIRC	PF00731	AIRC
PF00899	ThiF	PF00899	ThiF
PF00923	Transaldolase	PF00923	Transaldolase
PF00929	Exonuc_X-T	PF00929	Exonuc_X-T
PF01000	RNA_pol_A_bac	PF00562	RNA_pol_Rpb2_6
PF01039	Carboxyl_trans	PF01039	Carboxyl_trans
PF01053	Cys_Met_Meta_PP	PF01053	Cys_Met_Meta_PP
PF01063	Aminotran_4	PF01063	Aminotran_4
PF01138	RNase_PH	PF01138	RNase_PH
PF01182	Glucosamine_iso	PF01182	Glucosamine_iso
PF01192	RNA_pol_Rpb6	PF00623	RNA_pol_Rpb1_2
PF01193	RNA_pol_L	PF00562	RNA_pol_Rpb2_6
PF01193	RNA_pol_L	PF01000	RNA_pol_A_bac
PF01193	RNA_pol_L	PF01193	RNA_pol_L
PF01227	GTP_cyclohydroI	PF01227	GTP_cyclohydroI
PF01230	HIT	PF01230	HIT
PF01259	SAICAR_synt	PF01259	SAICAR_synt
PF01423	LSM	PF01423	LSM
PF01467	CTP_transf_2	PF01467	CTP_transf_2
PF01546	Peptidase_M20	PF01546	Peptidase_M20
PF01612	3_5_exonuc	PF00570	HRDC
PF01624	MutS_I	PF01624	MutS_I
PF01751	Toprim	PF00270	DEAD
PF01842	ACT	PF01842	ACT
PF01926	MMR_HSR1	PF01926	MMR_HSR1
PF01965	DJ-1_PfpI	PF01965	DJ-1_PfpI
PF02142	MGS	PF02142	MGS
PF02463	SMC_N	PF02463	SMC_N
PF02518	HATPase_c	PF00183	HSP90
PF02518	HATPase_c	PF00204	DNA_gyraseB
PF02518	HATPase_c	PF01119	DNA_mis_repair
PF02518	HATPase_c	PF02518	HATPase_c
PF02729	OTCace_N	PF00185	OTCace
PF02729	OTCace_N	PF02729	OTCace_N
PF02769	AIRS_C	PF00586	AIRS
PF02769	AIRS_C	PF02769	AIRS_C
PF02772	S-AdoMet_synt_M	PF00438	S-AdoMet_synt_N
PF02772	S-AdoMet_synt_M	PF02772	S-AdoMet_synt_M
PF02773	S-AdoMet_synt_C	PF00438	S-AdoMet_synt_N
PF02773	S-AdoMet_synt_C	PF02772	S-AdoMet_synt_M
PF02773	S-AdoMet_synt_C	PF02773	S-AdoMet_synt_C
PF02775	TPP_enzyme_C	PF00205	TPP_enzyme_M
PF02775	TPP_enzyme_C	PF02775	TPP_enzyme_C
PF02776	TPP_enzyme_N	PF00205	TPP_enzyme_M
PF02776	TPP_enzyme_N	PF02775	TPP_enzyme_C
PF02776	TPP_enzyme_N	PF02776	TPP_enzyme_N
PF02779	Transket_pyr	PF00456	Transketolase_N

Accession A	Name A	Accession B	Name B
PF02779	Transket_pyr	PF00676	E1_dh
PF02779	Transket_pyr	PF02779	Transket_pyr
PF02780	Transketolase_C	PF00456	Transketolase_N
PF02780	Transketolase_C	PF02779	Transket_pyr
PF02780	Transketolase_C	PF02780	Transketolase_C
PF02781	G6PD_C	PF00479	G6PD_N
PF02781	G6PD_C	PF02781	G6PD_C
PF02786	CPSase_L_D2	PF00117	GATase
PF02786	CPSase_L_D2	PF00289	CPSase_L_chain
PF02786	CPSase_L_D2	PF00988	CPSase_sm_chain
PF02786	CPSase_L_D2	PF02142	MGS
PF02786	CPSase_L_D2	PF02786	CPSase_L_D2
PF02787	CPSase_L_D3	PF00117	GATase
PF02787	CPSase_L_D3	PF00289	CPSase_L_chain
PF02787	CPSase_L_D3	PF00988	CPSase_sm_chain
PF02787	CPSase_L_D3	PF02786	CPSase_L_D2
PF02787	CPSase_L_D3	PF02787	CPSase_L_D3
PF02798	GST_N	PF00043	GST_C
PF02798	GST_N	PF02798	GST_N
PF02800	Gp_dh_C	PF00044	Gp_dh_N
PF02800	Gp_dh_C	PF02800	Gp_dh_C
PF02812	ELFV_dehydrog_N	PF00208	ELFV_dehydrog
PF02812	ELFV_dehydrog_N	PF02812	ELFV_dehydrog_N
PF02826	2-Hacid_dh_C	PF00389	2-Hacid_dh
PF02826	2-Hacid_dh_C	PF02826	2-Hacid_dh_C
PF02852	Pyr_redox_dim	PF02817	E3_binding
PF02852	Pyr_redox_dim	PF02852	Pyr_redox_dim
PF02866	Ldh_1_C	PF00056	Ldh_1_N
PF02866	Ldh_1_C	PF02866	Ldh_1_C
PF02867	Ribonuc_red_lgC	PF00317	Ribonuc_red_lgN
PF02867	Ribonuc_red_lgC	PF02867	Ribonuc_red_lgC
PF02881	SRP54_N	PF00448	SRP54
PF02881	SRP54_N	PF02881	SRP54_N
PF02887	PK_C	PF00224	PK
PF02887	PK_C	PF02887	PK_C
PF02978	SRP_SPB	PF00448	SRP54
PF02978	SRP_SPB	PF02881	SRP54_N
PF02978	SRP_SPB	PF02978	SRP_SPB
PF03129	HGTP_anticodon	PF00587	tRNA-synt_2b
PF03129	HGTP_anticodon	PF03129	HGTP_anticodon
PF03144	GTP_EFTU_D2	PF00009	GTP_EFTU
PF03144	GTP_EFTU_D2	PF03144	GTP_EFTU_D2
PF03372	Exo_endo_phos	PF03372	Exo_endo_phos
PF03477	ATP-cone	PF00317	Ribonuc_red_lgN
PF03477	ATP-cone	PF02867	Ribonuc_red_lgC
PF03725	RNase_PH_C	PF01138	RNase_PH
PF03725	RNase_PH_C	PF03725	RNase_PH_C
PF03764	EFG_IV	PF00009	GTP_EFTU

Accession A	Name A	Accession B	Name B
PF03764	EFG_IV	PF00679	EFG_C
PF03764	EFG_IV	PF03144	GTP_EFTU_D2
PF03807	F420_oxidored	PF03807	F420_oxidored
PF03953	Tubulin_C	PF00091	Tubulin
PF03953	Tubulin_C	PF03953	Tubulin_C
PF04983	RNA_pol_Rpb1_3	PF01192	RNA_pol_Rpb6
PF04997	RNA_pol_Rpb1_1	PF01192	RNA_pol_Rpb6
PF04998	RNA_pol_Rpb1_5	PF01192	RNA_pol_Rpb6
PF05188	MutS_II	PF00488	MutS_V
PF05188	MutS_II	PF01624	MutS_I
PF05190	MutS_IV	PF05190	MutS_IV
PF05192	MutS_III	PF00488	MutS_V
PF05192	MutS_III	PF01624	MutS_I
PF05192	MutS_III	PF05188	MutS_II
PF05192	MutS_III	PF05190	MutS_IV
PF06026	Rib_5-P_isom_A	PF06026	Rib_5-P_isom_A
PF06418	CTP_synth_N	PF00117	GATase
PF06418	CTP_synth_N	PF06418	CTP_synth_N
PF07687	M20_dimer	PF01546	Peptidase_M20
PF07687	M20_dimer	PF07687	M20_dimer
PF07973	tRNA_SAD	PF00587	tRNA-synt_2b
PF07992	Pyr_redox_2	PF02852	Pyr_redox_dim
PF07992	Pyr_redox_2	PF07992	Pyr_redox_2
PF08240	ADH_N	PF00107	ADH_zinc_N
PF08240	ADH_N	PF08240	ADH_N
PF08544	GHMP_kinases_C	PF00288	GHMP_kinases_N
PF08544	GHMP_kinases_C	PF08544	GHMP_kinases_C

Appendix E

Table E.1: Most frequent Gene Ontology annotations on all *i*Pfam families shared between *E. coli*, *S. cerevisiae* and *H. sapiens*.

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF00291	catalytic activity	9	metabolic process	13		
PF00702	catalytic activity	9	metabolic process	13		
PF01063	catalytic activity	9	metabolic process	13		
PF00171	oxidoreductase activity	9	metabolic process	13		
PF00378	catalytic activity	9	metabolic process	13		
PF00106	oxidoreductase activity	9	metabolic process	13		
PF00180	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	3	metabolic process	13		
PF00389	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	3	metabolic process	13		
PF00289	ligase activity	2	metabolic process	13		
PF02817	acyltransferase activity	2	metabolic process	13		
PF01842	amino acid binding	2	metabolic process	13		
PF00676	oxidoreductase activity, acting on the aldehyde or oxo group of donors, disulfide as acceptor	1	metabolic process	13		
PF00583	N-acetyltransferase activity	1	metabolic process	13		

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF00623	DNA-directed RNA polymerase activity	8	transcription	7	nucleus	1
PF04998	DNA-directed RNA polymerase activity	8	transcription	7		
PF01193	DNA-directed RNA polymerase activity	8	transcription	7		
PF01000	DNA-directed RNA polymerase activity	8	transcription	7		
PF04997	DNA-directed RNA polymerase activity	8	transcription	7		
PF04983	DNA-directed RNA polymerase activity	8	transcription	7		
PF00562	DNA-directed RNA polymerase activity	8	transcription	7		
PF01624	ATP binding	26	mismatch repair	6		
PF05190	ATP binding	26	mismatch repair	6		
PF00488	ATP binding	26	mismatch repair	6		
PF05188	ATP binding	26	mismatch repair	6		
PF01119	ATP binding	26	mismatch repair	6		
PF05192	ATP binding	26	mismatch repair	6		
PF00208	oxidoreductase activity	9	amino acid metabolic process	5		
PF02812	oxidoreductase activity	9	amino acid metabolic process	5		
PF01053	pyridoxal phosphate binding	4	amino acid metabolic process	5		
PF02887	magnesium ion binding	3	glycolysis	5		
PF00044	NAD binding	3	glycolysis	5		
PF02800	NAD binding	3	glycolysis	5		
PF00224	magnesium ion binding	3	glycolysis	5		
PF00185	carboxyl- and carbamoyltransferase activity	2	amino acid metabolic process	5		
PF02729	carboxyl- and carbamoyltransferase activity	2	amino acid metabolic process	5		

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF00365	6-phosphofructokinase activity	1	glycolysis	5	6-phosphofructokinase complex	1
PF00166	ATP binding	26	protein folding	4		
PF00155	pyridoxal phosphate binding	4	biosynthetic process	4		
PF01467	nucleotidyltransferase activity	2	biosynthetic process	4		
PF00483	nucleotidyltransferase activity	2	biosynthetic process	4		
PF00183	unfolded protein binding	1	protein folding	4		
PF00534			biosynthetic process	4		
PF00254			protein folding	4		
PF00160			protein folding	4		
PF00438	ATP binding	26	one-carbon compound metabolic process	3		
PF02772	ATP binding	26	one-carbon compound metabolic process	3		
PF02773	ATP binding	26	one-carbon compound metabolic process	3		
PF03725	RNA binding	7	RNA processing	3		
PF00636	RNA binding	7	RNA processing	3		
PF01138	RNA binding	7	RNA processing	3		
PF03129	ATP binding	26	translation	2		
PF02852	oxidoreductase activity	9	cell redox homeostasis	2	cytoplasm	3
PF00056	oxidoreductase activity	9	tricarboxylic acid cycle intermediate metabolic process	2		
PF02866	oxidoreductase activity	9	tricarboxylic acid cycle intermediate metabolic process	2		
PF02881	GTP binding	8	SRP-dependent cotranslational protein targeting to membrane	2	signal recognition particle, endoplasmic reticulum targeting	2
PF00448	GTP binding	8	SRP-dependent cotranslational protein targeting to membrane	2	membrane	1

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF02781	glucose-6-phosphate 1-dehydrogenase activity	2	glucose metabolic process	2		
PF00479	glucose-6-phosphate 1-dehydrogenase activity	2	glucose metabolic process	2		
PF00542	structural constituent of ribosome	1	translation	2	intracellular	6
PF00199	catalase activity	1	electron transport	2		
PF00462	protein disulfide oxidoreductase activity	1	cell redox homeostasis	2		
PF03807			electron transport	2		
PF01182			carbohydrate metabolic process	2		
PF00923			carbohydrate metabolic process	2		
PF02463	ATP binding	26	DNA metabolic process	1	chromosome	2
PF00204	ATP binding	26	DNA topological change	1	chromosome	2
PF00664	ATP binding	26	transport	1	integral to membrane	1
PF00334	ATP binding	26	UTP biosynthetic process	1		
PF00118	ATP binding	26	cellular protein metabolic process	1		
PF00587	ATP binding	26	tRNA aminoacylation for protein translation	1		
PF00288	ATP binding	26	phosphorylation	1		
PF00988	ATP binding	26	nitrogen compound metabolic process	1		
PF03953	GTP binding	8	protein polymerization	1	protein complex	1
PF01192	DNA-directed RNA polymerase activity	8	transcription, DNA-dependent	1		
PF02978	RNA binding	7	protein targeting	1	signal recognition particle, endoplasmic reticulum targeting	2
PF01751	nucleic acid binding	5	DNA modification	1		

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF02787	carbamoyl-phosphate synthase activity	2	arginine biosynthetic process	1	cytoplasm	3
PF01227	GTP cyclohydrolase I activity	1	aromatic compound biosynthetic process	1	cytoplasm	3
PF00731	phosphoribosylaminoimidazole carboxylase activity	1	de novo' IMP biosynthetic process	1	phosphoribosylaminoimidazole carboxylase complex	1
PF02867	ribonucleoside-diphosphate reductase activity	1	DNA replication	1	ribonucleoside-diphosphate reductase complex	1
PF00227	threonine endopeptidase activity	1	ubiquitin-dependent protein catabolic process	1	proteasome core complex (sensu Eukaryota)	1
PF01259	phosphoribosylaminoimidazolesuccinocarboxamide synthase activity	1	purine nucleotide biosynthetic process	1		
PF01546	metallopeptidase activity	1	proteolysis	1		
PF06026	ribose-5-phosphate isomerase activity	1	pentose-phosphate shunt, non-oxidative branch	1		
PF06418	CTP synthase activity	1	pyrimidine nucleotide biosynthetic process	1		
PF01423			mRNA metabolic process	1	ribonucleoprotein complex	1
PF00156			nucleoside metabolic process	1		
PF00004	ATP binding	26				
PF02786	ATP binding	26				
PF02518	ATP binding	26				
PF00270	ATP binding	26				
PF00271	ATP binding	26				
PF00176	ATP binding	26				
PF00005	ATP binding	26				
PF02775	catalytic activity	9				
PF00586	catalytic activity	9				
PF00206	catalytic activity	9				
PF00258	oxidoreductase activity	9				
PF00899	catalytic activity	9				
PF00117	catalytic activity	9				

Accession	Function	Freq	Process	Freq	Compartment	Freq
PF00578	oxidoreductase activity	9				
PF01926	GTP binding	8			intracellular	6
PF00679	GTP binding	8				
PF00009	GTP binding	8				
PF03144	GTP binding	8				
PF03764	GTP binding	8				
PF00013	RNA binding	7				
PF00570	nucleic acid binding	5			intracellular	6
PF01612	nucleic acid binding	5			intracellular	6
PF00383	hydrolase activity	4				
PF00202	pyridoxal phosphate binding	4				
PF00149	hydrolase activity	4				
PF07687	hydrolase activity	4				
PF00293	hydrolase activity	4				
PF02776	thiamin pyrophosphate binding	3				
PF02826	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	3				
PF00205	magnesium ion binding	3				
PF00596	metal ion binding	2				
PF01039	ligase activity	2				
PF00132	acyltransferase activity	2				
PF00491	metal ion binding	2				
PF00035	double-stranded RNA binding	1			intracellular	6
PF00328	acid phosphatase activity	1				
PF00533					intracellular	6

Appendix F

Table F.1: List of disease mutations linked to protein interaction defects, derived from the scientific literature.

Mutation	Variant	Seq ID	Description	Inh.	Mech.
604312	.0001	P01034	In patients with Icelandic-type cerebroarterial amyloidosis (105150), Abrahamson et al. (1987) identified a 358T-A transversion in the CST3 gene, resulting in a leu68-to-gln (L68Q) substitution.The dimerization was highly temperature-dependent, with a rise in incubation temperature from 37 to 40 degrees centigrade resulting in a 150% increase in dimerization rate.		GF
107300	.0021	P01008	Antithrombin III defficiency	AD	LF
121011	.0020	P29033	gap-junction protein (no direct functional link)	AD	LF
123580	.0001	P02489	Crystallin change of preference in polymers	AD	CF
123590	.0001	P02511	Crystallin change of preference in polymers	AD	CF
123680	.0001	Q53R50	Crystallin change of preference in polymers in Coppock cataract	AD	LF
125647	.0002	Q4LE79	Desmoplakin; This region of the desmoplakin protein interacts with intermediate filaments to anchor them to the desmosome	AR	LF
134850	.0010	P02679	Fibrinogen G, impaired polymerisation	AR	LF
134850	.0017	P02679	Fibrinogen G, impaired polymerisation	AR	LF
138040	.0009	P04150	GLUCOCORTICOID receptor, reduced cofactor binding	AD	LF
139250	.0020	P01241	Growth Hormone, in a prepubertal Spanish child with familial short stature (604271), Lewis et al. (2004) found an ile179-to-met (I179M) amino acid substitution. Molecular modeling studies suggested that the I179M substitution might perturb interactions between GH and the GH receptor loop containing residue trp169, thereby affecting signal transduction.	AD	LF
139320	.0032	Q5JWD2	GNAS	IM	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
139350	.0004	P04264	Keratin	AD	LF
139350	.0015	P04264	Keratin	AD	LF
141800	.0179	Q5R9M5	HBA1; Hb Yuda has a very low oxygen affinity and slightly decreased cooperative subunit interaction.		LF
141900	.0038	P68871	HBB; HEMOGLOBIN C [HBB, GLU6LYS]	IM	GF
147545	.0002	P35568	IRS1	AR	LF
147557	.0014	P16144	Koster et al. (2001) reported that this mutation renders integrin beta-4 unable to interact with plectin (601282) and prevents the localization of plectin in hemidesmosomes.	AR	LF
147557	.0015	P16144	Koster et al. (2001) reported that this mutation renders integrin beta-4 unable to interact with plectin (601282) and prevents the localization of plectin in hemidesmosomes.	AR	LF
600576	.0001	P43694	Garg et al. (2003) demonstrated that GATA4 (600576) interacts with TBX5 and showed that a missense mutation in GATA4, G296S (600576.0001), abrogated this interaction.	AD	LF
235200	.0011	NP_620575	By performing immunoprecipitation studies in HeLa cells, Ka et al. (2005) found that the Q283P mutation prevented the normal interaction between HFE protein and beta-2-microglobulin (B2M; 109700) and between HFE protein and transferrin receptor (TFRC; 190010).	CH	LF
300300	.0025	Q32ML5	de Weers et al. (1994) identified a C-to-T transition at position 993, resulting in a substitution of tryptophan for arginine-288. This mutation was found in the SH2-like domain where arg288 is highly conserved and crucial for the interaction with the aromatic ring of phosphotyrosine. Therefore, the replacement of arg288 by a nonpolar tryptophan would entirely abrogate the formation of the high-affinity comosine binding pocket. The change to a neutral glycine residue is highly likely to disrupt the binding potential of this region. This patient has less than 1% B cells and undetectable immunoglobulin levels, indicating that the replacement of this highly conserved arginine residue completely abolishes the functioning of Btk.	XL	LF
300490	.0013	O60880	SH2 Domain Protein 1A; Based on the molecular structure of the SH2D1A-SLAM (603492) interaction, this mutation was predicted to disrupt binding between the SH2 domain of SH2D1A and the cytoplasmic domain of SLAM. The mutation was also predicted to interfere with SH2D1A-2B4 (605554) binding because of the strong amino acid homology shared by SLAM and 2B4.	XL	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
305371	.0002	P15976	Freson et al. (2001) described a family with isolated X-linked macrothrombocytopenia without anemia but with some dyserythropoietic features (see 300367) in 13 males in 9 sibships of 3 generations connected through carrier females. A novel mutation in the GATA1 gene, asp218 to gly (D218G), resulted in a weaker interaction with FOG1	XL	LF
305371	.0005	P15976	Freson et al. (2002) described a 2-generation family with deep macrothrombocytopenia (see 300367), marked anemia, and early mortality. The mutation is predicted to result in substitution of tyrosine for aspartate-218 (D218Y). Zinc finger interaction studies revealed a stronger loss of affinity of D218Y-GATA1 than of D218G-GATA1 (305371.0002) for the essential transcription factor FOG1 (601950) and a disturbed GATA1 self-association.	XL	LF
600160	.0016	P42771	CDK inhib 2a; A val59-to-gly mutation in the CDKN2A gene was found in 4 families segregating cutaneous malignant melanoma; The mutation, which occurs in a hydrophobic region with the second ankyrin repeat, impairs p16-INK4a function, as shown by studies of protein-protein interactions and cell proliferation assays.	AD	LF
601130	.0002	P11712	Cytochrome P450; the CYP2C9*3 variant is less than 5% as efficient as the wildtype enzyme, while CYP2C9*2 shows about 12% of wildtype activity, apparently as a result of the amino acid substitution altering the interaction of the enzyme with cytochrome P450 oxidoreductase. Aithal et al. (1999) studied the frequency of the 2 variant alleles in individuals with a low warfarin dose requirement; see 122700. Patients in the low-dose group were more likely to have difficulties at the time of induction of warfarin therapy and had an increased risk of major bleeding complications.	PM	LF
601769	.0010	P11473	Whitfield et al. (1996) identified a mutation in the VDR gene, resulting in an ile314-to-ser (I314S) substitution in the hormone-binding domain of the protein. The mutation caused decreased 1,25-(OH)2D3-dependent transactivation of the VDR and impaired heterodimeric interaction with the retinoid X receptor	AR	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
601769	.0011	P11473	In a patient with vitamin D-dependent rickets type II (277440), Whitfield et al. (1996) identified a mutation in the VDR gene, resulting in an arg391-to-cys (R391C) substitution in the hormone-binding domain of the protein. The mutation caused decreased 1,25-(OH)2D3-dependent transactivation of the VDR and impaired heterodimeric interaction with the retinoid X receptor (RXR; 180245).	AR	LF
603273	.0009	Q9H3D4	In a 6-year-old patient with Hay-Wells syndrome (106260) who lacked any limb defects, McGrath et al. (2001) identified an A-to-T transversion at nucleotide 1542 of the TP63 gene, resulting in a leu518-to-phe substitution in the sterile alpha motif (SAM) domain. Molecular modeling suggested that the substitution would alter protein-protein interactions.	AD	LF
603273	.0010	Q9H3D4	In a 10-month-old infant with typical features of Hay-Wells syndrome (106260), McGrath et al. (2001) identified a T-to-G transversion at nucleotide 1564 of the TP63 gene, resulting in a cys526-to-gly substitution in the sterile alpha motif (SAM) domain. Molecular modeling suggested that the substitution would alter protein-protein interactions.	AD	LF
603714	.0002	O95343	Laflamme et al. (2004) demonstrated that the SIX3 protein carrying this mutation did not interact with NOR1 (600542) in vivo.	AD	LF
606860	.0002	P05155	Complement Component 1 Inhib; Davis et al. (1992) showed that the dysfunction demonstrated by this mutation results from a block in the interaction with target protease.		LF
608014	.0001	Q9UJY1	HS 22kd Prot, Increased binding!	AD	GF
608014	.0002	Q9UJY1	HS 22kd Prot, Increased binding!	AD	GF
608537	.0019	P40337	Ang et al. (2002) concluded that the R200W substitution impairs the interaction of VHL with HIF1-alpha	AR	LF
103850	.0002	P04075	Mutation of Glu to Arg in subunit interface. However, this is not proved to disrupt protein-protein interaction but it seems likely and the authors argue this is the case	AR	LF
256540	.0014	P10619	A structural model of the mutant PPCA was constructed by amino acid substitution of 453glutamic acid for lysine in the crystal structure of the wild type PPCA precursor reported. The results show that the K453E mutation is located at the dimer interface of the PPCA and reduces the hydrogen bond formation in the dimer. This structural change may cause instability of the PPCA dimer.		

Mutation	Variant	Seq ID	Description	Inh.	Mech.
305900	.0051	NP_000393	In a study of the causative mutation in 12 cases of G6PD deficiency associated with chronic nonspherocytic hemolytic anemia, Vulliamy et al. (1998) found 1 patient to have a novel mutation, which they called G6PD Serres: a 1082C-T change, causing an ala361-to-val substitution in the dimer interface where most other severe G6PD mutations are found. Blood. 2000 Feb 15;95(4):1499-501.	XL	LF
193400	.0013	NP_000543	J Biol Chem. 1991 Jul 25;266(21):13499-502. In previous studies, we have mapped the epitope for an anti-vWF monoclonal antibody which inhibits the interaction between FVIII and vWF to a region spanning Thr78 to Thr96 of the mature protein. We now report the identification of a mutation within this region of vWF that results in decreased FVIII binding.	CH	LF
606869	.0009	P06865	Paw et al. (1990) identified a G-to-A transition at nucleotide 1511 resulting in substitution of histidine for arginine at position 504 in the HEXA molecule. Cultured fibroblasts from the patient synthesized an alpha subunit that could acquire mannose 6-phosphate and be secreted, but which failed to associate with the beta-subunit to form the enzymatically active heterodimer.	AR	LF
193400	.0024	NP_000543	Schneppenheim et al. (1996) demonstrated a heterozygous cys2010-to-arg mutation in the mature vWF subunit causing the type IID von Willebrand disease phenotype in 2 unrelated patients. Recombinant expression of mutant vWF fragments demonstrated that the mutation was responsible for defective disulfide bonding of the C-terminal domains, thus impairing dimer formation.	AD	LF
141850	.0005	P69905	Goossens et al. (1982) described another nondeletion mechanism: mutation in the 125th codon of the alpha-2 gene resulted in substitution of proline for leucine in a region of the H helix of the alpha-globin chain, which is critical for alpha-beta contact, resulting in impediment to alpha-beta dimer formation, the initial step in hemoglobin tetramer assembly.	AR	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
125660	.0006	Q53SB5	The leu345-to-pro mutation (L345P) in this kindred was located in an evolutionarily highly conserved position of the desmin coiled-coil rod domain important for dimer formation. L345P desmin was incapable of forming filamentous networks in transfected HeLa and SW13 cells. Sjoberg et al. (1999) concluded that the L345P mutation causes myopathy by interfering in a dominant-negative manner with the dimerization-polymerization process of intermediate filament assembly.	AD	LF
190160	NA	P10828	Proc Natl Acad Sci U S A. 1997 Jan 7;94(1):248-53 Here we describe a novel leucine to valine mutation in codon 454 (L454V) of the thyroid hormone beta receptor. [...] indicating that the interaction of this residue with accessory proteins is critical for transcriptional activation. (Not in OMIM nor UniProt)	AD	LF
235200	.0001	NP_620575	J Biol Chem. 1997 May 30;272(22):14025-8: Co-immunoprecipitation studies demonstrate that wild-type HLA-H binds beta2-microglobulin and that the C282Y mutation completely abrogates this interaction.	AR	LF
607008	.0001	P11310	MCAD DEFICIENCY; the amount of K329E tetramer formed was distinctly less than wildtype at any point up to 60 minutes after import, indicating that the assembly of K329E was defective. After further incubation, K304E decayed more rapidly than did wildtype, indicating a reduced stability. In similar studies K329R behaved like the wildtype, while K329D closely resembled K329E, indicating that a basic residue at 304 is essential for tetramer formation and intramitochondrial stability of mature MCAD.	AR	LF
176300	.0039	P02766	Jenne et al. (1996) identified a 'new' amyloidogenic val20ile mutation of the TTR gene. tetramer stability was significantly reduced in agreement with the expected change in the interactions between 2 opposing dimers via the side chain of ile20.		CF
174763	.0002	P54098	J Biol Chem. 2005 Sep 9;280(36):31341-6: the A467T mutant enzyme failed to interact with and was not stimulated by the accessory subunit.	AR	LF
157140	.0003	P10636	J Neurochem. 2000 Jun;74(6):2583-9: Mutated tau is less phosphorylated than its normal counterpart at serines 396 and 404. Furthermore, the phosphorylated mutant protein is unable to bind to microtubules.	AD	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
191044	.0001	P19429	Biochemistry. 2002 Jun 11;41(23):7267-74: the affinity is reduced by approximately 14-fold by the T142 phosphorylation and approximately 4-fold by the mutation R145G.	AD	LF
P51587	VAR_020705	P51587	Oncogene. 2003 Jan 9;22(1):28-33: the cancer-predisposing mutation Y42C in BRCA2 significantly compromised the interaction between RPA and BRCA2		LF
276000	.0006	Q5NV57	Hum Mutat. 2004 Jan;23(1):22-31: E79K markedly inhibited autoactivation of cationic trypsinogen. Remarkably, however, E79K trypsin activated anionic trypsinogen PRSS2 (601564) 2-fold.	AD	CF
P00156	VAR_013653	P00156	European Journal of Biochemistry, Volume 271, Issue 7, April 2004, Pages 1292-1298: The mitochondrial cytochrome b missense mutation, G167E, has been reported in a patient with cardiomyopathy. The residue G167 is located in an extramembranous helix close to the hinge region of the iron-sulfur protein. Analysis of the enzyme activity indicated that the mutation affected its stability, which could be the result of an altered binding of the iron-sulfur protein on the complex. [...]This suggested that the mutation G167E could hinder the movement of the iron-sulfur protein, probably by distorting the structure of the hinge region.		LF
238331	NA	P09622	Biochem Biophys Res Commun. 1999 Aug 19;262(1):163-6: Asp for Val at position 479 of the precursor form - the mutation resides within the interface domain and likely perturbs stable dimerization		LF
P04275	VAR_005800	P04275	PubMed=1409710: von Willebrand disease type B: a missense mutation selectively abolishes ristocetin-induced von Willebrand factor binding to platelet glycoprotein Ib; J Thromb Haemost. 2006 Feb;4(2):417-25: The interaction of von Willebrand factor-A1 domain with collagen: mutation G1324S (type 2M von Willebrand disease) impairs the conformational change in A1 domain induced by collagen	CH	LF
606672	.0003	P07359	J Thromb Haemost. 2003 Oct;1(10):2198-205: The 125I-labeled VWF binding to mutant compared with the wild type displayed three patterns, gain-of-function (G233S, G233V, and M239V), equivalent function (G233A), and loss-of-function (G233K and G233D)	AD	GF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
193400	.0018	NP_000543	Proc Natl Acad Sci U S A. 1992 Oct 15;89(20):9846-9: the type B variant VWF displayed an aberrant interaction with the gpIb platelet receptor that seemed to be independent of multimeric structure.	AD	LF
193400	.0008	NP_000543	J Biol Chem. 1992 Oct 15;267(29):21187-92: the Arg578–¿Gln mutation increases the affinity of vWF for GPIb but does not directly impair vWF interaction with collagen or heparin. Arg578 may therefore be necessary to prevent normal vWF from interacting with GPIb.	AD	GF
193400	.00012	NP_000543	Blood. 1992 Feb 1;79(3):563-7: These results illustrate the importance of Arg 53 of the mature vWF subunit for the binding of FVIII to vWF	AR	LF
O75695	VAR_008499	O75695	Structure 2006 Feb;14:367-378: The abilities of RP2 to bind Arl3 and cause retinitis pigmentosa seem to be correlated, since both the R118H and E138G mutants show drastically reduced affinity to Arl3		LF
O75695	VAR_018074	O75695	Structure 2006 Feb;14:367-378: The abilities of RP2 to bind Arl3 and cause retinitis pigmentosa seem to be correlated, since both the R118H and E138G mutants show drastically reduced affinity to Arl3		LF
603693	.0001	Q8WW38	Although the mutant protein retained the ability to bind the partner protein GATA4 (600576) and repress GATA4-mediated gene activation, it was subtly impaired in this function.		LF
607759	NA	P08514	J Thromb Haemost. 2004 Jul;2(7):1167-75: A novel Phe171Cys mutation in the alpha(IIb) gene of patients with GT is associated with abrogation of alpha(IIb)beta(3) complex formation	AR	LF
300384	.0008	P50402	Hum Genet. 1999 Mar;104(3):262-8: Biochemical analysis has demonstrated that the mobility and expression levels of the mutant forms of emerin are indistinguishable from that of wild-type emerin, but that they have weakened interactions with nuclear lamina components		LF
300384	.0009	P50402	Hum Genet. 1999 Mar;104(3):262-8: Biochemical analysis has demonstrated that the mobility and expression levels of the mutant forms of emerin are indistinguishable from that of wild-type emerin, but that they have weakened interactions with nuclear lamina components		LF
605906	.0009	O75112	J Biol Chem. 2004 Feb 20;279(8):6746-52: o reveal the biochemical changes due to the mutation, we performed a yeast two-hybrid assay and a pull-down assay. It was demonstrated by both assays that the D626N mutation of Cypher/ZASP increased the affinity of the LIM domain for protein kinase C	AD	GF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
603959	.0015	Q9Y5I7	Am J Hum Genet. 2003 Dec;73(6):1293-301: The T233R mutation was found to abolish binding of CLDN16 to ZO1	AR	LF
193400	NA	NP_000543	J Thromb Haemost. 2004 Jul;2(7):1135-42: The mutation L1503Q does not significantly disrupt the conformation of the protein; thus the subtle loss of multimers in this patient may be due to altered interactions with the ADAMTS13 protease.		LF
P04275	VAR_010242	P04275	Blood. 2000 Jul 15;96(2):560-8: Multimer analysis showed that rVWFR273W failed to form high-molecular-weight multimers present in wild-type rVWF	AR	LF
P98172	VAR_023135	P98172	Twigg S.R.F., Kan R., Babbs C., Bochukova E.G., Robertson S.P., Wall S.A., Morriss-Kay G.M., Wilkie A.O.M. "Mutations of ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome." Proc. Natl. Acad. Sci. U.S.A. 101:8652-8657(2004)	AR	LF
P98172	VAR_023130	P98172	Twigg S.R.F., Kan R., Babbs C., Bochukova E.G., Robertson S.P., Wall S.A., Morriss-Kay G.M., Wilkie A.O.M. "Mutations of ephrin-B1 (EFNB1), a marker of tissue boundary formation, cause craniofrontonasal syndrome." Proc. Natl. Acad. Sci. U.S.A. 101:8652-8657(2004)	AR	LF
P98172	VAR_023131	P98172	Am J Hum Genet. June 2004; 74(6): 12091215	AR	LF
P98172	VAR_023128	P98172	Am J Hum Genet. June 2004; 74(6): 12091215	AR	LF
Q99574	VAR_008520	Q99574	Nature. 1999 Sep 23;401(6751):376-9; polymerization disease: Familial dementia caused by polymerization of mutant neuroserpin.	AD	GF
P98172	VAR_023132	P98172	Am J Hum Genet. June 2004; 74(6): 12091215	AR	LF
P98172	VAR_023133	P98172	Am J Hum Genet. June 2004; 74(6): 12091215	AR	LF
P98172	VAR_023134	P98172	Am J Hum Genet. June 2004; 74(6): 12091215	AR	LF
610550	.0007	Q00266	J. Biol. Chem., Vol. 276, Issue 17, 13803-13809, April 27, 2001; Chamberlin et al. (1997) identified a heterozygous 791G-A transition in the MAT1A gene, resulting in an arg264-to-his (R264H) substitution. In vitro studies suggested that residue 264 is involved in salt bridge formation essential for subunit dimerization and that the dominant effect of the R264H mutation is exerted by the formation of enzymatically inactive R264/R264H dimers.	AD	LF
610550	.0009	Q00266	Unlike the R264H (610550.0007) mutation, which behaves as an autosomal dominant, the authors found that the R264C mutation behaves as an autosomal recessive.	AR	LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
134850	.0004	P02679	Ebert and Bell (1988) identified Baltimore-3 as a congenital abnormal fibrinogen with defective fibrin monomer polymerization. Bantia et al. (1990) demonstrated an asn308-to-ile mutation. Polymerization is also affected by asn308-to-lys (Kyoto-1).	AD	LF
134850	.0005	P02679	Ebert and Bell (1988) identified Baltimore-3 as a congenital abnormal fibrinogen with defective fibrin monomer polymerization. Bantia et al. (1990) demonstrated an asn308-to-ile mutation. Polymerization is also affected by asn308-to-lys (Kyoto-1).	AD	LF
P05166	VAR_000280	P05166	Molecular Genetics and Metabolism, Volume 74, Number 4, December 2001, pp. 476-483(8): To clarify the molecular effect associated with gene alterations causing propionic acidemia, 12 different mutations affecting the PCCB gene were analyzed for their involvement in alpha-beta heteromeric and beta-beta homomeric assembly.	AR	LF
P05166	VAR_000281	P05166	Molecular Genetics and Metabolism, Volume 74, Number 4, December 2001, pp. 476-483(8): To clarify the molecular effect associated with gene alterations causing propionic acidemia, 12 different mutations affecting the PCCB gene were analyzed for their involvement in alpha-beta heteromeric and beta-beta homomeric assembly.	AR	LF
P05166	VAR_009086	P05166	Molecular Genetics and Metabolism, Volume 74, Number 4, December 2001, pp. 476-483(8): To clarify the molecular effect associated with gene alterations causing propionic acidemia, 12 different mutations affecting the PCCB gene were analyzed for their involvement in alpha-beta heteromeric and beta-beta homomeric assembly.	AR	LF
600160	.0007	P42771	Oncogene. 1999 Sep 23;18(39):5423-34; Harland et al. (1997) identified a met53-to-ile mutation in the CDKN2A gene in affected members of a family with melanoma. They showed that the protein expressed from this previously described mutation did not bind to CDK4/CDK6, confirming its role as a causal mutation in melanoma.	AD	LF
P42771	VAR_001409	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001410	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001411	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001420	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001424	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001449	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001447	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF
P42771	VAR_001448	P42771	Oncogene. 1999 Sep 23;18(39):5423-34		LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
603868	.0010	P51159	In a study of the spectrum of mutations in children with primary emophagocytic lymphohistiocytosis (267700), Zur Stadt et al. 2006) identified 2 mutations in RAB27A in 3 patients with Griscelli syndrome type 2 (607624), which can present with hemophagocytic lymphohistiocytosis. One of these was the missense mutation ala87 to pro (A87P). In functional studies using a mammalian 2-hybrid system, they found that the A87P mutation in RAB27A and leu403 to pro in UNC13D (608897.0007) each prevented the formation of a stable UNC13D/RAB27A complex in vitro.	AR	LF
608897	0007	Q70J99	In a study of the spectrum of mutations in children with primary emophagocytic lymphohistiocytosis (267700), Zur Stadt et al. 2006) identified 2 mutations in RAB27A in 3 patients with Griscelli syndrome type 2 (607624), which can present with hemophagocytic lymphohistiocytosis. One of these was the missense mutation ala87 to pro (A87P). In functional studies using a mammalian 2-hybrid system, they found that the A87P mutation in RAB27A and leu403 to pro in UNC13D (608897.0007) each prevented the formation of a stable UNC13D/RAB27A complex in vitro.	AR	LF
Q99574	VAR_008521	Q99574	Nature. 1999 Sep 23;401(6751):376-9; polymerization disease: Familial dementia caused by polymerization of mutant neuroserpin.	AD	GF
100710	.0003	P11230	Quiram et al. (1999) demonstrated that the mutation impairs AChR assembly by disrupting a specific interaction between the beta and delta (100720) subunits.	CH	LF
O15273	VAR_029446		J Am Coll Cardiol. 2004 Dec 7;44(11):2192-201: Two TCAP mutations, T137I and R153H, were found in patients with HCM, and another TCAP mutation, E132Q, was identified in a patient with DCM. It was demonstrated by the qualitative assays that the HCM-associated mutations augment the ability of Tcap to interact with titin and calsarcin-1, whereas the DCM-associated mutations impair the interaction of Tcap with MLP, titin, and calsarcin-1		
104760	.0001	P05067	In Levy et al. (1990) identified a 1852G-C transversion in the APP gene, resulting in a glu693-togln (E693Q) substitution. Miravalle et al. (2000) demonstrated in vitro that the E693Q mutation resulted in a high content of beta-sheet amyloid conformation and fast aggregation/fibrillization properties.	AD	GF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
104760	.0013	P05067	In a patient with early-onset familial Alzheimer disease (104300), Kamino et al. (1992) identified an A-to-G transition in the APP gene, resulting in a glu693-to-gly (E693G) substitution. n vitro, the Arctic mutant form of A-beta forms protofibrils and fibrils at higher rates and in larger quantities than wildtype A-beta. In transgenic mice that expressed the Arctic mutant in neurons, Cheng et al. (2004) found that amyloid plaques formed faster and were more extensive compared to control mice. Cheng et al. (2004) concluded that the Arctic mutation is highly amyloidogenic in vivo.	AD	GF
176640	.0001	Q53YK7	The PRNP gene has an unstable region of 5 variant tandem octapeptide coding repeats between codons 51 and 91. Extension of this repeat causes rapid formation of amyloid plaques and neurodegeneration	AD	GF
141900	.0243	P68871	HEMOGLOBIN S [HBB, GLU6VAL] The classic sickle cell anaemia	IM	GF
P00439	VAR_000900	P00439	Molecular Genetics and Metabolism 73, 230 238 (2001) : The R157N mutation, associated here with the most marked decrease in two-hybrid interaction, also showed in all other expression systems the most severe effects, including rapid and very extensive aggregation.	AD	GF
P00441	VAR_007131	P00441	Proc Natl Acad Sci U S A. 2004 April 20; 101(16): 59765981: The crystal structures of the A4V and I113T mutants of SOD1 reveal a significant reorientation of the two subunits at the monomermonomer interface. This destabilization of the dimeric interface may result in an increased tendency to unfold or lose metals in vivo.	AD	LF
P00441	VAR_007155	P00441	Proc Natl Acad Sci U S A. 2004 April 20; 101(16): 59765981: The crystal structures of the A4V and I113T mutants of SOD1 reveal a significant reorientation of the two subunits at the monomermonomer interface. This destabilization of the dimeric interface may result in an increased tendency to unfold or lose metals in vivo.	AD	LF
602533	.0002	Q99497	Mutations in DJ-1, a human gene with homologues in organisms from all kingdoms of life, have been shown to be associated with autosomal recessive, early onset Parkinson's disease. The structure suggests that the loss of function caused by the Parkinson's-associated mutation L166P in DJ-1 is due to destabilization of the dimer interface.	AR	LF
P00492	VAR_006756	P00492	Human Mutation 23 (6), pp. 599-611: Destroys the helix thus the dimerization		LF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
P00492	VAR_006802	P00492	Human Mutation 23 (6), pp. 599-611: Interrupts hydrogen bond in A-B interface		LF
P00492	VAR_006803	P00492	Human Mutation 23 (6), pp. 599-611: Interrupts hydrogen bond in A-B interface		LF
P00492	VAR_006765	P00492	Human Mutation 23 (6), pp. 599-611: Removes one hydrogen bond, but not severe effect		LF
P01241	VAR_015805	P01241	Hum Mutat. 2003 Apr;21(4):424-40: two of the amino acids involved (K41 and T175) are among eight key residues identified as being necessary for tight binding affinity between site 1 of GH and the GHR.	AR	LF
P01241	VAR_015814	P01241	Hum Mutat. 2003 Apr;21(4):424-40: two of the amino acids involved (K41 and T175) are among eight key residues identified as being necessary for tight binding affinity between site 1 of GH and the GHR.	AR	LF
107680	.0016	P02647	In an English family with autosomal dominant nonneuropathic systemic amyloidosis, Soutar et al. (1992) identified a CTG (leu)-to-CGG (arg) transversion at codon 60. The affected individuals were heterozygotes.	AD	GF
107680	.0010	P02647	n a family of English-Scottish-Irish extraction, Van Allen et al. (1968) studied a form of amyloidosis in which neuropathy dominated the clinical picture early in the course and nephropathy late in the course.	AD	GF
107680	.0024	P02647	Hamidi Asl et al. (1999) found that autosomal dominant hereditary amyloidosis with a unique cutaneous and cardiac presentation and death from heart failure by the sixth or seventh decade was associated with a 1389T-C transition in exon 4 of the APOA1 gene. The predicted substitution of leu90-to-pro (L90P) substitution was confirmed by structural analysis of amyloid protein isolated from cardiac deposits of amyloid. The subunit protein was composed exclusively of NH2-terminal fragments of the variant APOA1 with the longest ending at residue 94 in the wildtype sequence. Amyloid fibrils derived from 4 previously described APOA1 variants were composed of similar fragments with carboxy-terminal heterogeneity, but contrary to those variants, which all carry one extra positive charge, the leu90-to-pro substitution did not result in any charge modification.	AD	GF

Mutation	Variant	Seq ID	Description	Inh.	Mech.
601145	.0004	P04080	Lalioti et al. (1997) identified a homozygous G-to-C transversion at nucleotide 426 in exon 1 of the cystatin B gene in non-Finnish EPM1 (254800) families from northern Africa and Europe. The mutation resulted in a gly4-to-arg substitution and was the first missense mutation described in association with EPM1. Molecular modeling predicted that this substitution severely affected the contact of cystatin B with papain. Alakurtti et al. (2005) transiently expressed the G4R mutation in BHK-21 cells. The mutant protein failed to associate with lysosomes.	AR	LF
152780	.0004	P01229	In a 30-year-old man who presented with delayed puberty and infertility and was found to have hypogonadism associated with absence of circulating luteinizing hormone, Valdes-Socin et al. (2004) identified a homozygous gly36-to-asp (G36D) substitution in the LHB gene; the mutation disrupted a vital cysteine knot motif and abrogated the heterodimerization and secretion of luteinizing hormone.	AR	LF
O15273	VAR_029447		J Am Coll Cardiol. 2004 Dec 7;44(11):2192-201: Two TCAP mutations, T137I and R153H, were found in patients with HCM, and another TCAP mutation, E132Q, was identified in a patient with DCM. It was demonstrated by the qualitative assays that the HCM-associated mutations augment the ability of Tcap to interact with titin and calsarcin-1, whereas the DCM-associated mutations impair the interaction of Tcap with MLP, titin, and calsarcin-1		

Appendix G

Table G.1: All predicted interacting mutations with the respective structural template sequences, percentage identity between query and target sequence as well as the predicted crystal contact status of the template interaction.

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
100650	.0001	P05091	504	2.27	P05091	504	100.00	No
100690	.0003	NP_000070	274	3.75	P02711	278	81.77	No
100690	.0006	NP_000070	269	2.58	P02711	273	81.77	No
100690	.0009	NP_000070	276	3.79	P02711	280	81.77	No
100710	.0001	P11230	289	2.93	Q6S3I0	285	60.53	No
100720	.0002	Q07001	271	3.54	P02711	260	34.31	No
102540	.0002	P68032	363	3.71	P68135	363	98.93	Yes
102560	.0003	NP_001605	332	2.59	P07830	332	94.91	No
102600	.0004	NP_000476	65	2.02	P49435	67	47.41	No
102610	.0002	P68133	117	2.62	P68135	117	100.00	No
102610	.0006	P68133	359	2.63	P68135	359	100.00	Yes
102610	.0010	P68133	336	3.26	P68135	336	100.00	No
102610	.0010	P68133	336	3.26	P68139	336	100.00	No
102610	.0013	P68133	334	2.59	P68135	334	100.00	No
102610	.0013	P68133	334	2.59	P68139	334	100.00	No
103600	.0007	P02768	143	2.26	P02768	143	100.00	Yes
103600	.0011	P02768	345	2.44	P02768	345	100.00	No
103850	.0001	P04075	128	3.95	P00883	128	99.14	No
103850	.0002	P04075	206	3.20	P00883	206	99.14	No
107280	.0001	NP_001076	414	3.77	P01011	414	100.00	No
107300	.0007	P01008	416	2.84	P05619	335	40.27	No
107300	.0010	P01008	425	2.09	P01008	425	100.00	No
107300	.0011	P01008	426	2.65	P01008	426	100.00	No
107300	.0019	P01008	439	4.17	P05619	357	40.27	No
107300	.0020	P01008	425	2.09	P01008	425	100.00	No
107300	.0021	P01008	425	2.09	P01008	425	100.00	No
107300	.0022	P01008	414	2.94	P05619	333	40.27	No
107300	.0027	P01008	416	2.84	P05619	335	40.27	No
107300	.0041	P01008	402	3.09	P05120	357	35.39	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
107400	.0004	P01009	400	2.37	P01009	400	100.00	No
107400	.0012	P01009	76	3.59	P01009	76	100.00	No
107400	.0013	P01009	288	3.24	P01009	288	100.00	No
107400	.0014	P01009	393	4.17	P01009	393	100.00	No
107400	.0017	P01009	76	3.59	P01009	76	100.00	No
107400	.0019	P01009	280	2.40	P01009	280	100.00	No
107400	.0026	P01009	382	2.73	P01009	382	100.00	No
107400	.0029	P01009	360	2.98	P01011	361	45.55	No
107400	.0037	P01009	280	2.40	P01009	280	100.00	No
107400	.0039	P01009	77	3.58	P01009	77	100.00	No
107680	.0005	P02647	131	2.04	P02647	131	100.00	No
107680	.0016	P02647	84	2.73	P02647	84	100.00	No
107680	.0021	P02647	74	3.82	P02647	74	100.00	No
107680	.0022	P02647	180	3.02	P02647	180	100.00	No
107680	.0024	P02647	114	2.44	P02647	114	100.00	No
107680	.0026	P02647	198	2.75	P02647	198	100.00	No
107930	.0003	P20711	309	2.41	P80041	309	91.82	No
109270	.0003	P02730	327	2.02	P02730	327	100.00	No
114240	.0010	P20807	490	3.92	Q07009	416	56.86	No
114800	.0002	P00915	246	4.22	O43570	275	36.80	No
118504	.0004	P43681	280	2.61	P02711	272	50.00	No
120130	.0001	P02462	1408	3.86	P02452	148	38.60	Yes
120130	.0002	P02462	921	3.86	P02452	151	38.60	No
120140	.0044	NP_001835	717	3.86	P02452	145	38.60	No
120150	.0021	P02452	1178	3.86	P02452	145	43.86	Yes
120160	.0008	NP_000080	907	3.86	P02452	142	40.35	No
120160	.0010	NP_000080	547	3.86	P02452	145	42.11	No
120160	.0015	NP_000080	976	3.86	P02452	151	38.60	No
120160	.0030	NP_000080	661	3.86	P02452	139	42.11	No
120190	.0003	P05997	960	3.86	P02452	136	42.11	No
120290	.0004	NP_542412	977	3.86	P02452	148	43.86	No
120550	.0001	P02745	208	2.06	P02746	213	37.82	No
120580	.0002	NP_958850	534	2.12	P00734	467	36.20	No
121050	.0008	NP_001990	1169	5.87	Q9JJS8	152	30.77	Yes
122500	.0002	P08185	389	2.51	P01011	405	47.98	No
123101	.0003	P35548	172	3.65	P06601	243	46.43	No
123610	.0002	P05813	91	3.67	P53674	118	52.44	No
123620	.0001	NP_000487	155	2.02	P02522	154	95.12	No
123690	.0001	P07320	14	2.66	P08209	14	87.34	No
123690	.0004	P07320	23	3.23	P62697	129	39.24	No
123690	.0006	P07320	23	3.23	P62697	129	39.24	No
123940	.0003	P19013	449	3.97	P08670	395	38.76	No
124020	.0003	P33261	212	3.14	P10632	212	77.97	No
125240	.0001	P08174	87	6.30	P20023	75	30.19	No
125270	.0004	P13716	240	2.28	P13716	240	100.00	No
125660	.0003	NP_001918	393	3.13	P08670	387	73.38	No
125660	.0006	NP_001918	345	2.79	P08670	339	73.38	No
125660	.0007	NP_001918	406	4.16	P08670	400	73.38	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
125660	.0010	NP_001918	385	2.43	P08670	379	73.38	No
125660	.0011	NP_001918	389	3.89	P08670	383	73.38	Yes
130130	.0002	P08246	206	2.49	P00761	189	36.27	No
130130	.0006	P08246	139	3.20	P00761	114	36.27	No
130130	.0007	P08246	101	2.42	P08246	101	100.00	No
130130	.0009	P08246	71	5.41	P08246	71	100.00	No
130410	.0001	NP_001976	164	3.92	P38117	164	100.00	No
130410	.0003	NP_001976	128	3.97	P38117	127	100.00	No
130410	.0003	NP_001976	128	3.97	P38117	128	100.00	No
131399	.0001	NP_000493	286	4.08	P05164	314	72.21	No
131550	.0004	NP_005219	719	3.81	Q06187	408	36.48	Yes
131550	.0005	NP_005219	719	3.81	Q06187	408	36.48	Yes
134370	.0007	P08603	1207	2.30	P68638	240	31.37	Yes
134797	.0005	NP_000129	1249	5.87	Q9JJS8	152	37.50	Yes
134797	.0011	NP_000129	723	4.22	P07204	441	42.42	No
134850	.0001	P02679	301	2.72	P02679	301	100.00	No
134850	.0002	P02679	301	2.72	P02679	301	100.00	No
134850	.0004	P02679	334	3.47	P02679	334	100.00	No
134850	.0005	P02679	334	3.47	P02679	334	100.00	No
134850	.0006	P02679	336	4.27	P02679	336	100.00	No
134850	.0018	P02679	191	2.95	Q02020	227	43.46	No
134850	.0019	P02679	335	3.07	P02679	335	100.00	No
136351	.0003	NP_004110	835	2.30	P06213	1183	38.35	Yes
136351	.0004	NP_004110	835	2.30	P06213	1183	38.35	Yes
136351	.0005	NP_004110	835	2.30	P06213	1183	38.35	Yes
136351	.0006	NP_004110	835	2.30	P06213	1183	38.35	Yes
136351	.0007	NP_004110	835	2.30	P06213	1183	38.35	Yes
136352	.0003	NP_891555	1041	4.05	Q07912	256	35.04	Yes
136352	.0005	NP_891555	1114	4.23	Q06187	596	36.69	Yes
136530	.0002	P01225	69	5.86	P01225	69	100.00	No
136850	.0006	NP_000134	343	3.15	P05042	296	60.79	No
136850	.0007	NP_000134	233	3.61	P05042	186	60.79	No
136850	.0008	NP_000134	233	3.61	P05042	186	60.79	No
137780	.0010	P14136	362	3.97	P08670	395	63.96	No
137780	.0012	P14136	352	3.39	P08670	385	63.96	No
138079	.0001	P35557	279	2.31	P05708	283	51.88	No
139250	.0020	P01241	205	2.01	P01241	205	100.00	No
139320	.0003	P63092	272	3.01	P04896	272	99.74	No
139320	.0008	P63092	201	4.27	P63096	177	41.62	No
139320	.0009	P63092	201	4.27	P63096	177	41.62	No
139320	.0010	P63092	227	4.62	P10824	203	41.91	No
139320	.0012	P63092	227	4.62	P10824	203	41.91	No
139320	.0013	P63092	201	4.27	P63096	177	41.62	No
139320	.0018	P63092	170	3.46	P63096	146	41.62	No
139320	.0020	P63092	231	4.27	P04896	231	99.74	No
139320	.0021	P63092	201	4.27	P63096	177	41.62	No
139330	.0001	NP_653082	38	3.58	P63096	41	67.44	No
139340	.0001	NP_005263	79	2.81	P63096	78	69.74	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
139350	.0003	P04264	481	4.94	P08670	399	38.76	No
139360	.0001	NP_002061	179	4.27	P63096	177	87.90	No
139360	.0002	NP_002061	179	4.27	P63096	177	87.90	No
139360	.0003	NP_002061	179	4.27	P63096	177	87.90	No
139360	.0004	NP_002061	200	3.56	P10824	198	87.90	No
140100	.0005	P00738	265	2.36	P00742	341	32.38	No
141800	.0028	NP_000549	75	2.25	P02089	79	42.19	No
141800	.0095	NP_000549	75	2.25	P02089	79	42.19	No
141800	.0100	NP_000549	75	2.25	P02089	79	42.19	No
141800	.0122	NP_000549	75	2.25	P02089	79	42.19	No
141800	.0157	NP_000549	75	2.25	P02089	79	42.19	No
141850	.0006	P69905	62	3.12	P02089	67	42.19	No
141850	.0007	P69905	109	2.56	P02089	114	42.19	No
141850	.0008	P69905	61	2.40	P02089	66	42.19	No
141850	.0009	P69905	27	2.10	P01958	27	87.69	No
141850	.0011	P69905	16	2.30	P01990	16	67.69	Yes
141850	.0012	P69905	47	2.60	P02208	57	30.77	Yes
141850	.0025	P69905	47	2.60	P02208	57	30.77	Yes
141850	.0031	P69905	104	3.59	P69905	104	100.00	No
141850	.0034	P69905	74	2.25	P02089	79	42.19	No
141850	.0035	P69905	80	2.20	P02089	85	42.19	No
141850	.0037	P69905	126	2.03	P69905	126	100.00	No
141850	.0042	P69905	20	2.28	P02118	19	41.41	Yes
141850	.0045	P69905	66	2.37	P02089	71	42.19	No
141850	.0049	P69905	72	3.17	P02089	77	42.19	No
141850	.0052	P69905	95	3.08	P69905	95	100.00	No
141850	.0053	P69905	37	3.98	P01965	37	83.85	No
141850	.0055	P69905	31	2.86	P69905	31	100.00	No
141850	.0060	P69905	65	2.18	P02089	70	42.19	No
141850	.0065	P69905	59	2.52	P02089	64	42.19	No
141900	.0005	P68871	19	2.34	P02118	19	69.40	Yes
141900	.0019	P68871	15	5.78	P02118	15	69.40	Yes
141900	.0021	P68871	102	2.60	P68871	102	100.00	No
141900	.0025	P68871	88	2.86	P68871	88	100.00	No
141900	.0026	P68871	119	2.23	P68871	119	100.00	No
141900	.0027	P68871	127	2.16	P68871	127	100.00	No
141900	.0028	P68871	100	3.08	P68871	100	100.00	No
141900	.0030	P68871	67	3.12	P02089	67	79.85	No
141900	.0046	P68871	99	2.45	P68871	99	100.00	No
141900	.0048	P68871	66	2.40	P02089	66	79.85	No
141900	.0064	P68871	19	2.34	P02118	19	69.40	Yes
141900	.0068	P68871	98	2.88	P68871	98	100.00	No
141900	.0071	P68871	26	2.10	P68871	26	100.00	No
141900	.0078	P68871	77	3.17	P02089	77	79.85	No
141900	.0079	P68871	79	2.25	P02089	79	79.85	No
141900	.0084	P68871	79	2.25	P02089	79	79.85	No
141900	.0096	P68871	127	2.16	P68871	127	100.00	No
141900	.0104	P68871	26	2.10	P68871	26	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
141900	.0109	P68871	37	3.09	P68871	37	100.00	No
141900	.0114	P68871	99	2.45	P68871	99	100.00	No
141900	.0116	P68871	66	2.40	P02089	66	79.85	No
141900	.0118	P68871	92	5.44	P02089	92	79.85	No
141900	.0119	P68871	92	5.44	P02089	92	79.85	No
141900	.0120	P68871	17	2.30	P68871	17	100.00	No
141900	.0126	P68871	64	2.52	P02089	64	79.85	No
141900	.0130	P68871	128	2.12	P68871	128	100.00	No
141900	.0131	P68871	77	3.17	P02089	77	79.85	No
141900	.0143	P68871	132	2.60	P68871	132	100.00	No
141900	.0144	P68871	30	2.86	P68871	30	100.00	No
141900	.0145	P68871	102	2.60	P68871	102	100.00	No
141900	.0146	P68871	99	2.45	P68871	99	100.00	No
141900	.0148	P68871	124	2.47	P68871	124	100.00	No
141900	.0151	P68871	98	2.88	P68871	98	100.00	No
141900	.0158	P68871	36	3.98	P68871	36	100.00	No
141900	.0162	P68871	17	2.30	P68871	17	100.00	No
141900	.0163	P68871	67	3.12	P02089	67	79.85	No
141900	.0164	P68871	92	5.44	P02089	92	79.85	No
141900	.0168	P68871	19	2.34	P02118	19	69.40	Yes
141900	.0169	P68871	97	3.47	P68871	97	100.00	No
141900	.0172	P68871	114	2.56	P02089	114	79.85	No
141900	.0184	P68871	97	3.47	P68871	97	100.00	No
141900	.0186	P68871	92	5.44	P02089	92	79.85	No
141900	.0192	P68871	17	2.30	P68871	17	100.00	No
141900	.0193	P68871	97	3.47	P68871	97	100.00	No
141900	.0195	P68871	100	3.08	P68871	100	100.00	No
141900	.0197	P68871	92	5.44	P02089	92	79.85	No
141900	.0199	P68871	36	3.98	P68871	36	100.00	No
141900	.0201	P68871	98	2.88	P68871	98	100.00	No
141900	.0203	P68871	52	2.02	P02118	52	69.40	No
141900	.0212	P68871	52	2.02	P02118	52	69.40	No
141900	.0213	P68871	117	3.30	P68871	117	100.00	No
141900	.0220	P68871	35	2.99	P68871	35	100.00	No
141900	.0229	P68871	78	2.52	P02089	78	79.85	No
141900	.0230	P68871	99	2.45	P68871	99	100.00	No
141900	.0234	P68871	15	5.78	P02118	15	69.40	Yes
141900	.0236	P68871	102	2.60	P68871	102	100.00	No
141900	.0241	P68871	37	3.09	P68871	37	100.00	No
141900	.0250	P68871	117	3.30	P68871	117	100.00	No
141900	.0253	P68871	88	2.86	P68871	88	100.00	No
141900	.0256	P68871	70	2.18	P02089	70	79.85	No
141900	.0269	P68871	102	2.60	P68871	102	100.00	No
141900	.0272	P68871	52	2.02	P02118	52	69.40	No
141900	.0273	P68871	36	3.98	P68871	36	100.00	No
141900	.0274	P68871	67	3.12	P02089	67	79.85	No
141900	.0276	P68871	26	2.10	P68871	26	100.00	No
141900	.0278	P68871	30	2.86	P68871	30	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
141900	.0281	P68871	79	2.25	P02089	79	79.85	No
141900	.0288	P68871	124	2.47	P68871	124	100.00	No
141900	.0289	P68871	124	2.47	P68871	124	100.00	No
141900	.0294	P68871	123	2.16	P68871	123	100.00	No
141900	.0300	P68871	97	3.47	P68871	97	100.00	No
141900	.0301	P68871	99	2.45	P68871	99	100.00	No
141900	.0302	P68871	132	2.60	P68871	132	100.00	No
141900	.0307	P68871	99	2.45	P68871	99	100.00	No
141900	.0311	P68871	17	2.30	P68871	17	100.00	No
141900	.0313	P68871	15	5.78	P02118	15	69.40	Yes
141900	.0315	P68871	37	3.09	P68871	37	100.00	No
141900	.0318	P68871	35	2.99	P68871	35	100.00	No
141900	.0319	P68871	127	2.16	P68871	127	100.00	No
141900	.0320	P68871	127	2.16	P68871	127	100.00	No
141900	.0394	P68871	119	2.23	P68871	119	100.00	No
141900	.0397	P68871	114	2.56	P02089	114	79.85	No
141900	.0404	P68871	92	5.44	P02089	92	79.85	No
141900	.0405	P68871	99	2.45	P68871	99	100.00	No
141900	.0411	P68871	17	2.30	P68871	17	100.00	No
141900	.0424	P68871	114	2.56	P02089	114	79.85	No
141900	.0427	P68871	92	5.44	P02089	92	79.85	No
141900	.0428	P68871	18	2.32	P02118	18	69.40	Yes
141900	.0433	P68871	79	2.25	P02089	79	79.85	No
141900	.0438	P68871	67	3.12	P02089	67	79.85	No
141900	.0440	P68871	37	3.09	P68871	37	100.00	No
141900	.0447	P68871	67	3.12	P02089	67	79.85	No
141900	.0448	P68871	127	2.16	P68871	127	100.00	No
141900	.0452	P68871	98	2.88	P68871	98	100.00	No
141900	.0453	P68871	79	2.25	P02089	79	79.85	No
141900	.0466	P68871	26	2.10	P68871	26	100.00	No
141900	.0469	P68871	77	3.17	P02089	77	79.85	No
141900	.0481	P68871	124	2.47	P68871	124	100.00	No
141900	.0487	P68871	36	3.98	P68871	36	100.00	No
141900	.0490	P68871	36	3.98	P68871	36	100.00	No
141900	.0492	P68871	122	3.09	P68871	122	100.00	No
141900	.0494	P68871	117	3.30	P68871	117	100.00	No
141900	.0495	P68871	123	2.16	P68871	123	100.00	No
141900	.0499	P68871	128	2.12	P68871	128	100.00	No
141900	.0500	P68871	128	2.12	P68871	128	100.00	No
141900	.0512	P68871	64	2.52	P02089	64	79.85	No
141900	.0518	P68871	97	3.47	P68871	97	100.00	No
141900	.0525	P68871	26	2.10	P68871	26	100.00	No
141900	.0531	P68871	52	2.02	P02118	52	69.40	No
142000	.0004	P02042	99	2.45	P68871	99	92.54	No
142000	.0016	P02042	98	2.88	P68871	98	92.54	No
142000	.0029	P02042	30	2.86	P68871	30	92.54	No
142000	.0034	P02042	26	2.10	P68871	26	92.54	No
142000	.0035	P02042	37	3.09	P68871	37	92.54	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
142000	.0039	NP_000510	37	3.98	P68871	36	92.54	No
142000	.0041	NP_000510	89	2.86	P68871	88	92.54	No
142200	.0001	P69891	75	2.38	P02089	75	70.15	No
142200	.0002	P69891	128	2.12	P69891	128	100.00	No
142200	.0006	P69891	37	3.09	P02070	36	74.63	No
142200	.0006	P69891	37	3.09	P68871	37	74.63	No
142200	.0007	P69891	79	2.25	P02089	79	70.15	No
142200	.0008	P69891	97	3.47	P02070	96	74.63	No
142200	.0008	P69891	97	3.47	P68871	97	74.63	No
142200	.0016	P69891	36	3.98	P02070	35	74.63	No
142200	.0016	P69891	36	3.98	P68871	36	74.63	No
142200	.0018	P69891	75	2.38	P02089	75	70.15	No
142200	.0032	P69891	75	2.38	P02089	75	70.15	No
142250	.0009	P69892	77	3.17	P02089	77	70.90	No
142250	.0014	P69892	117	3.30	P68871	117	75.37	No
142250	.0019	P69892	26	2.10	P68871	26	75.37	No
142250	.0021	P69892	125	2.06	P69891	125	99.25	No
142250	.0022	P69892	66	2.40	P02089	66	70.90	No
142250	.0031	P69892	66	2.40	P02089	66	70.90	No
142250	.0034	P69892	92	5.44	P02089	92	70.90	No
142250	.0036	P69892	15	5.78	P02118	15	73.13	Yes
142250	.0039	P69892	75	2.38	P02089	75	70.90	No
142250	.0045	P69892	75	2.38	P02089	75	70.90	No
142250	.0048	P69892	17	2.30	P68871	17	75.37	No
142250	.0049	P69892	19	2.34	P02118	19	73.13	Yes
142360	.0004	P05546	462	2.49	P05546	462	100.00	No
142410	.0005	P20823	272	4.16	P40424	288	34.62	No
142984	.0001	P28358	319	2.82	Q6B2C0	185	37.04	Yes
142989	.0004	P35453	314	2.38	P02836	500	33.93	No
142989	.0007	P35453	298	3.65	P06601	243	30.36	No
142993	.0001	P58304	200	4.16	P40424	288	32.14	No
142993	.0002	P58304	200	4.16	P40424	288	32.14	No
142994	.0008	P50219	248	3.39	P02836	459	50.00	Yes
147450	.0001	P00441	37	3.51	P00441	37	100.00	No
147450	.0002	P00441	38	2.87	P00441	38	100.00	No
147450	.0003	P00441	41	3.40	P00441	41	100.00	Yes
147450	.0004	P00441	41	3.40	P00441	41	100.00	Yes
147450	.0006	P00441	85	3.55	P53636	117	30.71	No
147450	.0007	P00441	93	3.69	P00441	93	100.00	Yes
147450	.0008	P00441	93	3.69	P00441	93	100.00	Yes
147450	.0011	P00441	113	2.73	P00441	113	100.00	No
147450	.0016	P00441	104	3.07	P00441	104	100.00	Yes
147450	.0017	P00441	144	2.22	P00446	167	31.16	Yes
147450	.0020	P00441	6	3.81	P00442	6	83.33	No
147450	.0026	P00441	126	3.05	P00441	126	100.00	No
147450	.0033	P00441	93	3.69	P00441	93	100.00	Yes
148040	.0016	P13647	472	3.36	P08670	401	37.13	No
148041	.0004	NP_005545	469	3.39	P08670	403	37.13	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
148066	.0001	NP_000517	384	2.70	P08670	368	35.83	No
148066	.0011	NP_000517	415	4.94	P08670	399	35.83	No
148066	.0012	NP_000517	419	3.39	P08670	403	35.83	No
148067	.0008	NP_005548	354	2.41	P08670	336	34.53	No
148080	.0012	P13645	439	3.93	P08670	389	36.16	No
150330	.0017	P02545	377	4.16	P08670	400	30.39	No
151385	.0006	Q01196	107	2.92	Q01196	107	100.00	No
151385	.0008	Q01196	58	2.90	Q01196	58	100.00	No
152780	.0001	P01229	74	4.14	P01233	74	86.54	No
152780	.0004	P01229	56	3.85	P01233	56	86.54	No
153450	.0002	P61626	85	2.78	P61626	85	100.00	No
153450	.0003	P61626	82	6.08	P61626	82	100.00	No
153450	.0005	P61626	82	6.08	P61626	82	100.00	No
156845	.0003	NP_000239	217	3.51	Q12772	343	40.00	No
160710	.0002	P13533	795	4.13	P13538	795	60.00	No
160760	.0001	P12883	403	2.33	P10587	405	51.93	No
160760	.0014	P12883	403	2.33	P10587	405	51.93	No
160760	.0015	P12883	403	2.33	P10587	405	51.93	No
160760	.0022	P12883	532	2.27	P13538	534	81.80	No
160760	.0024	P12883	743	2.52	P10587	753	51.93	Yes
162280	.0001	NP_006149	333	2.92	P08670	342	53.25	No
164761	.0013	NP_066124	918	4.05	Q06187	563	35.74	Yes
164790	.0001	P01111	13	3.01	P01112	13	91.88	No
164790	.0002	P01111	61	4.56	P01112	61	91.88	No
164840	.0003	P04198	394	3.51	Q12772	343	34.00	No
164860	.0007	P08581	1136	5.25	P00520	295	40.40	Yes
171060	.0007	P21439	1161	3.24	Q9CHL8	473	46.15	Yes
171760	.0002	P05186	71	4.02	Q9BHT8	45	47.20	No
171760	.0004	P05186	71	4.02	Q9BHT8	45	47.20	No
172400	.0001	NP_000166	158	3.79	P06744	157	100.00	No
172400	.0002	P06744	346	2.58	P06744	346	100.00	No
172400	.0003	P06744	524	2.99	P06744	524	100.00	No
172400	.0004	NP_000166	539	3.10	P06744	538	100.00	No
172471	.0002	P15735	189	3.47	P05132	200	33.06	No
172471	.0002	P15735	189	3.47	P00517	200	33.06	No
173110	.0001	P28069	172	4.06	P10037	172	98.65	No
173110	.0005	P28069	143	4.06	P14859	299	58.11	Yes
173350	.0005	P00747	616	5.08	P00761	42	45.59	No
173350	.0007	P00747	751	2.42	P00747	751	100.00	Yes
173515	.0005	P14770	24	5.79	P07359	24	34.62	No
175100	.0008	P25054	713	2.29	P35222	646	32.43	Yes
176300	.0007	P02766	131	2.94	O93330	133	57.27	No
176300	.0008	P02766	136	4.70	P02766	136	100.00	No
176300	.0011	P02766	134	3.84	P02766	134	100.00	No
176300	.0033	P02766	134	3.84	P02766	134	100.00	No
176300	.0034	P02766	127	2.12	P02766	127	100.00	No
176300	.0046	P02766	73	3.86	P02766	73	100.00	No
176300	.0047	P02766	38	4.03	P02766	38	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
176300	.0050	P02766	89	4.94	O93330	91	57.27	No
176730	.0001	NP_000198	49	3.54	P01308	49	100.00	No
176730	.0002	NP_000198	48	3.24	P01308	48	100.00	No
176730	.0003	NP_000198	34	3.49	P01308	34	100.00	No
176860	.0002	P04070	444	5.62	P00735	615	40.61	No
176860	.0005	P04070	301	3.00	P00735	467	40.61	No
176860	.0008	P04070	343	3.37	P00735	508	40.61	No
176860	.0011	P04070	334	2.89	P00735	499	40.61	No
176860	.0012	P04070	289	2.88	P00735	454	40.61	No
176860	.0019	P04070	226	3.30	P00735	381	40.61	No
176860	.0022	P04070	339	2.75	P00735	504	40.61	No
176860	.0024	P04070	149	3.20	P00743	138	54.84	No
176930	.0004	P00734	425	2.37	P00734	425	100.00	No
176930	.0005	P00734	601	2.93	P00734	601	100.00	No
176947	.0005	P43403	465	4.05	Q07912	256	38.98	Yes
180200	.0003	P06400	445	3.83	P06400	445	100.00	Yes
180200	.0004	P06400	567	3.66	P06400	567	100.00	No
180200	.0019	P06400	661	4.22	P06400	661	100.00	No
188540	.0001	P01222	49	3.85	P01225	48	42.31	No
188540	.0002	P01222	32	3.57	P01225	31	42.31	No
188540	.0004	P01222	69	4.14	P01225	66	42.31	No
188826	.0002	P35625	191	2.22	P16035	200	45.88	No
189980	.0006	P00519	351	5.05	P54763	742	42.80	Yes
190020	.0001	P01112	12	2.54	P01112	12	100.00	No
190020	.0002	P01112	61	4.56	P01112	61	100.00	No
190020	.0003	P01112	12	2.54	P01112	12	100.00	No
190020	.0004	P01112	12	2.54	P01112	12	100.00	No
190020	.0005	P01112	13	3.01	P01112	13	100.00	No
190070	.0001	NP_004976	12	2.54	P01112	12	94.38	No
190070	.0002	NP_004976	12	2.54	P01112	12	94.38	No
190070	.0003	NP_004976	13	3.01	P01112	13	94.38	No
190070	.0004	NP_004976	59	3.71	P01112	59	94.38	No
190070	.0005	NP_004976	12	2.54	P01112	12	94.38	No
190070	.0006	NP_004976	12	2.54	P01112	12	94.38	No
190070	.0007	NP_004976	12	2.54	P01112	12	94.38	No
190070	.0009	NP_004976	60	3.86	P01112	60	94.38	No
190070	.0011	NP_004976	58	4.11	P01112	58	94.38	No
190070	.0013	NP_004976	34	3.06	P01112	34	94.38	No
190450	.0005	P60174	170	3.99	P04789	172	53.59	No
190990	.0001	P07951	91	3.43	P42639	91	85.59	No
190990	.0002	P07951	147	4.37	P42639	147	85.59	No
190990	.0003	P07951	117	2.08	P42639	117	85.59	Yes
191010	.0002	P09493	175	2.93	P42639	175	98.73	No
191010	.0003	P09493	95	2.48	P42639	95	98.73	No
191030	.0001	NP_689476	9	3.30	P04268	50	78.39	No
191044	.0003	NP_000354	82	2.49	P19429	81	99.24	No
191170	.0001	P04637	248	3.89	P04637	248	100.00	No
191170	.0003	P04637	245	3.65	P02340	242	88.66	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
191170	.0006	P04637	249	3.89	P04637	249	100.00	Yes
191170	.0008	P04637	242	5.61	P04637	242	100.00	No
191170	.0009	P04637	245	3.65	P02340	242	88.66	Yes
191170	.0010	P04637	248	3.89	P04637	248	100.00	No
191170	.0013	P04637	241	3.27	P04637	241	100.00	No
191170	.0019	P04637	245	3.65	P02340	242	88.66	Yes
191170	.0024	P04637	280	3.89	P04637	280	100.00	No
191170	.0030	P04637	175	3.89	P04637	175	100.00	Yes
191170	.0032	P04637	138	3.26	P04637	138	100.00	Yes
191170	.0038	P04637	189	2.04	P04637	189	100.00	Yes
191306	.0001	NP_002244	1147	3.80	P08631	497	42.57	Yes
191315	.0008	NP_001007793	604	4.59	P32577	304	37.19	Yes
217030	.0001	P05156	418	4.66	P03951	469	37.73	No
218030	.0007	P80365	227	2.70	P19992	147	30.52	No
227500	.0003	P08709	238	5.53	P00763	48	42.99	No
227500	.0004	P08709	307	2.12	P00760	109	42.06	Yes
227500	.0006	P08709	304	3.00	P00761	94	41.12	No
227500	.0007	P08709	117	2.99	P00740	104	61.29	No
227500	.0018	P08709	121	5.88	P00740	108	61.29	No
227500	.0023	P08709	414	2.45	Q9Y5Y6	816	37.67	Yes
229700	.0004	NP_000498	30	2.48	P09467	29	99.69	No
232050	.0005	P05166	168	2.88	Q8GBW6	146	52.71	No
232050	.0008	P05166	435	4.05	Q9X4K7	417	57.77	No
232800	.0003	NP_000280	39	3.77	P00512	25	48.00	No
232800	.0004	NP_000280	543	3.07	P00512	140	35.96	No
232800	.0006	NP_000280	39	3.77	P00512	25	48.00	No
234000	.0001	P00748	590	5.02	P00747	784	43.93	Yes
238331	.0002	P09622	488	4.27	P09624	479	77.98	No
250850	.0001	Q00266	322	3.04	P13444	323	97.08	No
250850	.0002	Q00266	55	2.81	P13444	56	97.98	No
250850	.0007	Q00266	264	3.43	P13444	265	97.08	No
250850	.0009	Q00266	264	3.43	P13444	265	97.08	No
256540	.0009	P10619	132	3.11	Q8W4X3	166	30.87	No
259730	.0007	NP_000058	40	3.94	P23589	70	52.36	No
264900	.0010	P03951	430	2.81	P00761	47	40.38	No
264900	.0011	P03951	594	3.22	P03951	594	100.00	No
264900	.0014	P03951	418	3.53	P00761	35	40.38	No
264900	.0015	P03951	587	4.76	P00766	207	40.18	No
300039	.0003	P49335	202	3.78	P14859	296	78.38	Yes
300075	.0017	P51812	268	3.79	P49137	263	30.96	No
300104	.0002	P31150	70	3.80	P39958	78	55.76	No
300206	.0002	NP_055086	487	5.60	Q9NZN1	487	100.00	No
300300	.0001	NP_000052	525	4.05	Q07912	256	41.30	Yes
300300	.0005	NP_000052	28	3.31	Q06187	27	100.00	No
300300	.0021	NP_000052	252	5.44	P08631	114	54.72	Yes
300300	.0022	NP_000052	255	2.28	O89100	304	33.33	No
300300	.0025	NP_000052	288	3.80	O60880	13	30.14	No
300300	.0026	NP_000052	307	4.27	P35235	32	32.43	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
300300	.0027	NP_000052	334	3.75	P27986	670	33.80	No
300300	.0032	NP_000052	408	2.92	Q06187	407	100.00	Yes
300300	.0036	NP_000052	520	4.10	P08069	1134	34.94	Yes
300300	.0037	NP_000052	520	4.10	P08069	1134	34.94	Yes
300300	.0047	NP_000052	613	3.26	P00520	455	48.19	No
300300	.0047	NP_000052	613	3.26	P00519	455	48.19	Yes
300382	.0008	Q96QS3	373	2.62	P06601	258	69.64	No
300382	.0015	Q96QS3	333	3.39	P02836	459	42.86	Yes
300382	.0016	Q96QS3	369	2.89	P06601	254	69.64	No
300386	.0003	P29965	227	2.84	P29965	227	100.00	No
300461	.0004	NP_000522	111	2.51	P04391	76	43.26	No
300461	.0025	NP_000522	129	3.02	P04391	94	43.26	No
300490	.0001	O60880	55	2.35	O60880	55	100.00	No
300490	.0004	O60880	32	4.27	P35235	32	30.14	No
300490	.0013	O60880	55	2.35	O60880	55	100.00	No
303900	.0001	P04000	247	2.14	P02699	231	45.97	No
305900	.0011	NP_000393	216	4.53	P11413	215	100.00	No
305900	.0015	NP_000393	410	3.54	P11413	409	100.00	No
305900	.0024	NP_000393	213	2.72	P11413	212	100.00	No
305900	.0027	NP_000393	227	3.21	P11413	226	100.00	Yes
305900	.0029	NP_000393	463	2.34	P11413	462	100.00	Yes
305900	.0035	NP_000393	227	3.21	P11413	226	100.00	Yes
305900	.0039	NP_000393	410	3.54	P11413	409	100.00	No
305900	.0040	NP_000393	439	3.55	P11413	438	100.00	No
305900	.0050	NP_000393	467	3.56	P11413	466	100.00	Yes
306900	.0015	P00740	75	2.42	P00741	29	92.68	No
306900	.0016	P00740	75	2.42	P00741	29	92.68	No
306900	.0022	P00740	106	3.30	P00740	106	100.00	No
306900	.0024	P00740	160	3.29	P09871	161	31.43	No
306900	.0062	NP_000124	363	2.78	P00743	370	47.00	No
308000	.0016	NP_000185	70	2.67	P00492	69	100.00	No
308000	.0017	NP_000185	71	3.25	Q26997	81	38.10	No
312865	.0007	O15266	173	2.86	P02836	510	48.21	No
313700	.0024	P10275	608	4.26	P03372	234	53.33	No
314200	.0003	P05543	303	2.59	P01011	311	44.39	No
516020	.0007	P00156	166	3.28	P00157	166	81.54	No
516030	.0008	P00395	196	2.01	P00396	196	93.64	No
516050	.0006	P00414	58	5.16	P06030	66	50.00	No
600046	.0014	O95477	935	3.62	Q9YGA6	38	31.25	No
600046	.0015	O95477	935	3.62	Q9YGA6	38	31.25	No
600194	.0004	P35908	485	3.45	P08670	398	35.18	No
600194	.0006	P35908	482	3.97	P08670	395	35.18	No
600211	.0010	Q13950	200	3.21	Q01196	149	91.04	No
600211	.0012	Q13950	169	3.35	Q01196	118	91.04	No
600225	.0002	P30793	134	2.33	P30793	134	100.00	No
600225	.0008	P30793	144	3.94	P30793	144	100.00	No
600225	.0015	P30793	135	2.72	P30793	135	100.00	No
600225	.0017	P30793	211	4.32	P22288	202	97.12	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
600509	.0003	NP_000343	716	3.84	P68187	39	33.72	No
600509	.0011	NP_000343	1506	3.85	Q9KQW9	506	33.33	Yes
600529	.0001	NP_001689	197	2.66	Q13825	197	100.00	No
600584	.0001	P52952	178	2.30	P02836	494	48.21	Yes
600584	.0013	P52952	190	4.16	P40424	288	30.36	No
600644	.0001	NP_976030	185	6.16	P15151	179	31.33	No
600871	.0002	Q99684	403	2.74	P03001	166	31.82	No
600983	.0010	P08235	645	5.86	P06536	482	89.19	No
600993	.0001	Q13485	358	2.11	Q13485	358	100.00	No
600993	.0003	Q13485	493	2.48	Q13485	493	100.00	No
600993	.0011	Q13485	352	3.51	Q13485	352	100.00	No
601107	.0001	Q92887	768	3.65	Q58206	153	31.65	Yes
601107	.0005	Q92887	1382	4.20	Q9CHL8	430	37.99	Yes
601145	.0004	P04080	4	3.43	P04080	4	100.00	No
601538	.0006	O75360	88	4.46	P40424	252	33.93	No
601538	.0011	O75360	99	3.65	P06601	243	67.86	No
601538	.0012	O75360	99	3.65	P06601	243	67.86	No
601542	.0005	NP_700476	91	4.16	P40424	288	33.93	No
601545	.0001	NP_000421	149	4.99	P62871	53	33.33	No
601545	.0006	NP_000421	31	3.95	P63005	30	100.00	No
601615	.0005	Q99758	568	3.62	P68187	38	33.33	No
601622	.0010	Q15672	156	3.54	P01106	403	44.90	No
601687	.0005	Q99456	429	4.94	P08670	399	33.11	No
601769	.0002	P11473	73	4.26	P03372	234	46.67	No
601769	.0011	P11473	391	2.95	Q13133	415	39.44	No
601789	.0002	Q92968	326	2.44	P08631	127	32.08	Yes
601802	.0001	Q9UBX0	160	4.16	P40424	288	33.93	No
601928	.0003	O43790	402	3.97	P08670	395	37.66	No
601928	.0005	O43790	402	3.97	P08670	395	37.66	No
602018	.0001	Q99748	191	2.04	Q07731	205	45.26	No
602049	.0001	P15153	57	4.23	P15153	57	100.00	No
602153	.0002	Q14533	402	3.97	P08670	395	37.66	No
602225	.0001	O43186	80	2.89	P06601	254	64.29	No
602225	.0005	O43186	41	3.12	P06601	215	64.29	No
602225	.0006	O43186	41	3.12	P06601	215	64.29	No
602298	.0001	P51149	129	2.78	P62825	126	32.91	No
602298	.0002	P51149	162	3.47	P62826	156	32.91	No
602298	.0002	P51149	162	3.47	P62826	157	32.91	No
602298	.0003	P51149	161	3.86	P11233	163	36.25	No
602421	.0010	P13569	549	3.70	P13569	549	100.00	No
602421	.0011	P13569	549	3.70	P13569	549	100.00	No
602421	.0012	P13569	549	3.70	P13569	549	100.00	No
602421	.0022	P13569	1282	2.52	Q9CHL8	421	31.67	Yes
602421	.0032	P13569	1303	3.39	Q9CHL8	442	31.67	Yes
602421	.0048	P13569	1291	4.20	Q9CHL8	430	31.67	Yes
602421	.0063	P13569	1283	2.74	Q9CHL8	422	31.67	Yes
602421	.0114	P13569	1303	3.39	Q9CHL8	442	31.67	Yes
602438	.0003	Q9ULV5	20	2.65	P22121	196	36.98	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
602445	.0001	NP_005016	49	3.58	O35684	49	86.93	No
602445	.0002	NP_005016	52	3.19	O35684	52	86.93	No
602533	.0004	Q99497	149	3.28	Q99497	149	100.00	Yes
602575	.0001	O60663	246	4.42	P02836	504	37.50	No
602575	.0002	O60663	198	3.12	P06601	215	35.71	No
602575	.0011	O60663	226	3.65	P06601	243	35.71	No
602765	.0001	P78385	407	3.97	P08670	395	37.66	No
602821	.0002	Q12840	280	4.20	P33173	307	44.04	Yes
603234	.0017	O95255	1339	2.74	Q9CHL8	422	36.11	Yes
603470	.0008	P00966	363	3.87	Q9X2A1	361	60.00	No
603470	.0009	P00966	390	3.83	Q9X2A1	388	60.00	No
603470	.0010	P00966	304	3.22	Q9X2A1	302	60.00	No
603470	.0012	P00966	86	2.52	Q9X2A1	84	60.00	No
603470	.0013	P00966	279	4.22	Q9X2A1	277	60.00	No
603470	.0016	P00966	362	3.21	Q9X2A1	360	60.00	No
603470	.0019	P00966	310	3.26	Q9X2A1	308	60.00	No
603851	.0005	Q99453	100	3.12	P06601	215	69.64	No
603868	.0001	P51159	73	5.75	P63012	76	46.88	No
603868	.0006	P51159	152	3.49	P01112	134	33.96	Yes
604277	.0004	Q9UBP0	499	3.83	Q01853	637	40.66	No
604720	.0005	Q9UP52	690	2.50	P02786	658	53.38	No
605020	.0001	Q9NZR4	166	3.12	P06601	215	62.50	No
605271	.0001	Q9UK55	324	5.69	P01011	299	32.34	No
605481	.0005	Q8IZT6	3060	4.58	Q02440	775	36.84	Yes
605481	.0006	Q8IZT6	1326	4.17	Q02440	778	35.00	Yes
605481	.0008	Q8IZT6	2063	3.41	Q02440	787	35.00	Yes
605511	.0003	P57727	251	5.08	P00761	42	40.38	No
605511	.0004	P57727	404	3.88	P07338	216	41.28	No
606765	.0005	NP_783651	453	3.61	P05164	462	47.52	No
606873	.0012	P07686	183	2.03	P07686	183	100.00	Yes
606885	.0005	P16219	383	3.62	P15651	383	94.63	No
606989	.0002	NP_000241	173	4.41	P05164	173	100.00	No
606989	.0003	NP_000241	251	3.39	P05164	251	100.00	No
606999	.0008	P07902	171	3.78	P09148	151	52.30	No
606999	.0011	P07902	183	2.41	P09148	163	52.30	No
606999	.0016	P07902	194	3.47	P09148	174	52.30	No
607379	.0005	P35240	535	2.05	P26038	517	38.76	No
607379	.0006	P35240	538	3.26	P26038	520	38.76	No
607809	.0002	P24752	183	3.07	P07097	146	43.14	No
608053	.0002	P13804	266	3.56	P13804	266	100.00	No
608310	.0002	P04424	286	2.79	P04424	286	100.00	No
608348	.0003	P12694	290	3.12	P84129	227	37.04	No
608537	.0025	P40337	155	2.93	P40337	155	100.00	No
608801	.0007	Q92947	337	3.08	Q06319	290	31.03	No
608845	.0003	Q9H0F7	31	3.88	P84080	30	42.20	No
608845	.0003	Q9H0F7	31	3.88	P84079	30	42.20	No
608845	.0003	Q9H0F7	31	3.88	P84077	30	42.20	No
608845	.0005	Q9H0F7	31	3.88	P84080	30	42.20	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
608845	.0005	Q9H0F7	31	3.88	P84079	30	42.20	No
608845	.0005	Q9H0F7	31	3.88	P84077	30	42.20	No
609014	.0002	Q9HCC0	99	3.09	Q8GBW6	61	35.11	No
609014	.0003	Q9HCC0	155	2.11	Q9X4K7	123	32.40	No
609712	.0003	NP_870986	353	4.05	P30613	384	100.00	No
609712	.0004	P30613	384	4.05	P30613	384	100.00	No
609712	.0006	P30613	479	2.41	P11974	435	59.17	Yes
O15266	VAR_012346	O15266	173	2.86	P02836	510	48.21	No
O43186	VAR_003750	O43186	41	3.12	P06601	215	64.29	No
O43186	VAR_003751	O43186	80	2.89	P06601	254	64.29	No
O43186	VAR_007946	O43186	41	3.12	P06601	215	64.29	No
O43790	VAR_018126	O43790	402	3.97	P08670	395	37.66	No
O43790	VAR_018127	O43790	402	3.97	P08670	395	37.66	No
O60663	VAR_004203	O60663	226	3.65	P06601	243	35.71	No
O60663	VAR_004205	O60663	246	4.42	P02836	504	37.50	No
O60806	VAR_018387	O60806	128	3.32	P24781	127	81.36	No
O60880	VAR_005612	O60880	32	4.27	P35235	32	30.14	No
O60880	VAR_018307	O60880	55	2.35	O60880	55	100.00	No
O95255	VAR_013390	O95255	1339	2.74	Q9CHL8	422	36.11	Yes
O95255	VAR_013391	O95255	1347	4.20	Q9CHL8	430	36.11	Yes
O95342	VAR_013334	O95342	461	4.04	Q9YGA6	42	30.82	No
O95477	VAR_009150	O95477	935	3.62	Q9YGA6	38	31.25	No
P00156	VAR_013653	P00156	166	3.28	P00157	166	81.54	No
P00414	VAR_002167	P00414	78	3.54	P00415	78	87.84	No
P00441	VAR_007132	P00441	7	3.01	P00441	7	100.00	No
P00441	VAR_007136	P00441	37	3.51	P00441	37	100.00	No
P00441	VAR_007137	P00441	38	2.87	P00441	38	100.00	No
P00441	VAR_007138	P00441	41	3.40	P00441	41	100.00	Yes
P00441	VAR_007139	P00441	41	3.40	P00441	41	100.00	Yes
P00441	VAR_007144	P00441	85	3.55	P53636	117	30.71	No
P00441	VAR_007146	P00441	93	3.69	P00441	93	100.00	Yes
P00441	VAR_007147	P00441	93	3.69	P00441	93	100.00	Yes
P00441	VAR_007148	P00441	93	3.69	P00441	93	100.00	Yes
P00441	VAR_007149	P00441	93	3.69	P00441	93	100.00	Yes
P00441	VAR_007155	P00441	113	2.73	P00441	113	100.00	No
P00441	VAR_007156	P00441	115	3.75	P00441	115	100.00	No
P00441	VAR_007157	P00441	125	3.92	P00441	125	100.00	Yes
P00441	VAR_007159	P00441	139	4.10	P00441	139	100.00	Yes
P00441	VAR_007160	P00441	144	2.22	P00446	167	31.16	Yes
P00441	VAR_007161	P00441	148	3.38	P00441	148	100.00	No
P00441	VAR_007162	P00441	148	3.38	P00441	148	100.00	No
P00441	VAR_007163	P00441	149	3.53	P00441	149	100.00	No
P00441	VAR_007164	P00441	151	2.66	P00441	151	100.00	No
P00441	VAR_008717	P00441	6	3.81	P00442	6	83.33	No
P00441	VAR_008719	P00441	93	3.69	P00441	93	100.00	Yes
P00441	VAR_008720	P00441	104	3.07	P00441	104	100.00	Yes
P00441	VAR_008722	P00441	124	4.20	P00441	124	100.00	Yes
P00441	VAR_008724	P00441	144	2.22	P00446	167	31.16	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P00441	VAR_013524	P00441	38	2.87	P00441	38	100.00	No
P00441	VAR_013526	P00441	49	2.28	P00441	49	100.00	No
P00441	VAR_013529	P00441	76	2.97	P00441	76	100.00	Yes
P00441	VAR_013531	P00441	86	4.15	P00441	86	100.00	No
P00441	VAR_013532	P00441	89	2.76	P00441	89	100.00	Yes
P00441	VAR_013535	P00441	105	2.16	P00441	105	100.00	Yes
P00441	VAR_013536	P00441	108	3.52	P00441	108	100.00	No
P00441	VAR_013538	P00441	114	3.85	P00441	114	100.00	No
P00441	VAR_013539	P00441	126	3.05	P00441	126	100.00	No
P00451	VAR_015134	P00451	2307	2.35	P12259	2183	42.54	Yes
P00480	VAR_004864	P00480	90	3.76	P04391	55	43.26	No
P00480	VAR_004875	P00480	126	4.22	P04391	91	43.26	No
P00480	VAR_004876	P00480	129	3.02	P04391	94	43.26	No
P00480	VAR_004922	P00480	264	2.20	P04391	232	38.06	No
P00480	VAR_004923	P00480	264	2.20	P04391	232	38.06	No
P00480	VAR_004924	P00480	267	2.66	P00480	267	100.00	Yes
P00480	VAR_004925	P00480	268	4.55	P00480	268	100.00	Yes
P00480	VAR_004926	P00480	269	3.04	P00480	269	100.00	Yes
P00492	VAR_006773	P00492	69	2.67	P00492	69	100.00	No
P00492	VAR_006774	P00492	70	3.25	Q26997	81	38.10	No
P00533	VAR_019297	P00533	719	3.81	Q06187	408	36.48	Yes
P00734	VAR_006715	P00734	425	2.37	P00734	425	100.00	No
P00734	VAR_006719	P00734	601	2.93	P00734	601	100.00	No
P00740	VAR_006543	P00740	91	4.57	P00741	45	92.68	No
P00740	VAR_006548	P00740	102	5.88	P09871	143	37.93	No
P00740	VAR_006549	P00740	106	3.30	P00740	106	100.00	No
P00740	VAR_006550	P00740	108	5.88	P00740	108	100.00	No
P00740	VAR_006564	P00740	160	3.29	P09871	161	31.43	No
P00740	VAR_006575	P00740	241	3.30	P00761	23	45.71	No
P00740	VAR_006576	P00740	253	3.36	P00761	34	45.71	No
P00740	VAR_006577	P00740	253	3.36	P00761	34	45.71	No
P00740	VAR_006578	P00740	265	3.24	P00761	46	45.71	No
P00740	VAR_006580	P00740	283	2.03	P00761	62	45.71	No
P00740	VAR_006584	P00740	302	4.66	P00761	81	45.71	No
P00740	VAR_006585	P00740	316	2.96	P00761	93	45.71	No
P00740	VAR_006586	P00740	321	2.66	P00761	98	45.71	No
P00740	VAR_006587	P00740	333	3.26	P00761	110	45.71	No
P00740	VAR_006591	P00740	356	5.33	P00761	129	45.71	No
P00740	VAR_006592	P00740	357	3.37	P00761	130	45.71	No
P00740	VAR_006594	P00740	363	2.78	P00743	370	47.00	No
P00740	VAR_006600	P00740	390	2.38	P08709	383	42.79	No
P00740	VAR_006601	P00740	394	3.78	P00761	168	45.71	No
P00740	VAR_006604	P00740	407	5.21	P00761	181	45.71	No
P00740	VAR_006605	P00740	413	3.64	P00761	187	45.71	No
P00740	VAR_006609	P00740	430	3.22	P00761	200	45.71	No
P00740	VAR_006610	P00740	431	4.69	P00761	201	45.71	No
P00740	VAR_006611	P00740	431	4.69	P00761	201	45.71	No
P00740	VAR_006612	P00740	432	3.14	P00761	202	45.71	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P00740	VAR_006613	P00740	432	3.14	P00761	202	45.71	No
P00740	VAR_006620	P00740	450	3.32	P00761	220	45.71	No
P00740	VAR_017308	P00740	75	2.42	P00741	29	92.68	No
P00740	VAR_017312	P00740	252	5.53	P00761	33	45.71	No
P00740	VAR_017315	P00740	306	2.61	P00761	85	45.71	No
P00740	VAR_017316	P00740	357	3.37	P00761	130	45.71	No
P00740	VAR_017317	P00740	397	2.61	P00761	171	45.71	No
P00740	VAR_017318	P00740	410	3.74	P00761	184	45.71	No
P00740	VAR_017319	P00740	411	3.65	P00761	185	45.71	No
P00740	VAR_017320	P00740	411	3.65	P00761	185	45.71	No
P00740	VAR_017321	P00740	414	3.88	P00761	188	45.71	No
P00740	VAR_017322	P00740	442	2.93	P00761	212	45.71	No
P00740	VAR_017324	P00740	453	5.62	P00761	223	45.71	No
P00740	VAR_017344	P00740	52	2.02	P00741	6	92.68	Yes
P00740	VAR_017346	P00740	106	3.30	P00740	106	100.00	No
P00740	VAR_017352	P00740	241	3.30	P00761	23	45.71	No
P00740	VAR_017353	P00740	252	5.53	P00761	33	45.71	No
P00740	VAR_017354	P00740	318	2.64	P00761	95	45.71	No
P00740	VAR_017355	P00740	333	3.26	P00761	110	45.71	No
P00740	VAR_017362	P00740	407	5.21	P00761	181	45.71	No
P00740	VAR_017363	P00740	412	3.87	P00761	186	45.71	No
P00740	VAR_017364	P00740	435	5.02	P00761	205	45.71	No
P00740	VAR_017365	P00740	442	2.93	P00761	212	45.71	No
P00747	VAR_006629	P00747	620	3.24	P00761	46	45.59	No
P00747	VAR_006630	P00747	751	2.42	P00747	751	100.00	Yes
P00748	VAR_006624	P00748	590	5.02	P00747	784	43.93	Yes
P00790	VAR_006483	P00790	92	2.09	P07339	95	50.00	No
P00813	VAR_002222	P00813	141	2.39	P56658	141	89.05	No
P00966	VAR_000683	P00966	86	2.52	Q9X2A1	84	60.00	No
P00966	VAR_000688	P00966	272	4.22	Q9X2A1	270	60.00	No
P00966	VAR_000690	P00966	304	3.22	Q9X2A1	302	60.00	No
P00966	VAR_000692	P00966	363	3.87	Q9X2A1	361	60.00	No
P00966	VAR_000693	P00966	363	3.87	Q9X2A1	361	60.00	No
P00966	VAR_000694	P00966	390	3.83	Q9X2A1	388	60.00	No
P00966	VAR_015892	P00966	86	2.52	Q9X2A1	84	60.00	No
P00966	VAR_015900	P00966	265	3.86	Q9X2A1	263	60.00	No
P00966	VAR_015901	P00966	269	2.91	Q9X2A1	267	60.00	No
P00966	VAR_015903	P00966	310	3.26	Q9X2A1	308	60.00	No
P00966	VAR_015904	P00966	362	3.21	Q9X2A1	360	60.00	No
P00966	VAR_016008	P00966	279	4.22	Q9X2A1	277	60.00	No
P00966	VAR_016009	P00966	310	3.26	Q9X2A1	308	60.00	No
P00966	VAR_016010	P00966	363	3.87	Q9X2A1	361	60.00	No
P00966	VAR_016011	P00966	363	3.87	Q9X2A1	361	60.00	No
P00966	VAR_016015	P00966	119	4.03	P59846	116	54.26	No
P01008	VAR_007042	P01008	90	4.16	O35684	29	32.34	No
P01008	VAR_007044	P01008	112	4.27	P05619	32	40.27	No
P01008	VAR_007047	P01008	133	3.75	P07385	41	30.91	No
P01008	VAR_007053	P01008	158	2.95	P05619	73	40.27	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P01008	VAR_007056	P01008	198	4.46	O35684	132	32.34	No
P01008	VAR_007062	P01008	283	5.14	P01008	283	100.00	No
P01008	VAR_007063	P01008	302	2.12	P01008	302	100.00	No
P01008	VAR_007065	P01008	334	2.34	P01012	262	31.81	No
P01008	VAR_007069	P01008	414	2.94	P05619	333	40.27	No
P01008	VAR_007070	P01008	416	2.84	P05619	335	40.27	No
P01008	VAR_007071	P01008	416	2.84	P05619	335	40.27	No
P01008	VAR_007074	P01008	425	2.09	P01008	425	100.00	No
P01008	VAR_007075	P01008	425	2.09	P01008	425	100.00	No
P01008	VAR_007076	P01008	425	2.09	P01008	425	100.00	No
P01008	VAR_007077	P01008	426	2.65	P01008	426	100.00	No
P01008	VAR_007078	P01008	434	3.53	P05619	352	40.27	No
P01008	VAR_007079	P01008	434	3.53	P05619	352	40.27	No
P01008	VAR_007080	P01008	434	3.53	P05619	352	40.27	No
P01008	VAR_007082	P01008	437	3.08	P01008	437	100.00	No
P01008	VAR_007083	P01008	438	3.31	P01008	438	100.00	No
P01008	VAR_007084	P01008	439	4.17	P05619	357	40.27	No
P01008	VAR_007085	P01008	439	4.17	P05619	357	40.27	No
P01008	VAR_007087	P01008	456	3.57	P05619	374	40.27	No
P01008	VAR_007088	P01008	457	2.58	P01008	457	100.00	Yes
P01008	VAR_009258	P01008	438	3.31	P01008	438	100.00	No
P01008	VAR_012316	P01008	95	4.09	P05619	16	40.27	No
P01009	VAR_006980	P01009	58	2.83	P01009	58	100.00	No
P01009	VAR_006985	P01009	77	3.58	P01009	77	100.00	No
P01009	VAR_006986	P01009	84	3.01	P05120	38	31.15	No
P01009	VAR_006999	P01009	280	2.40	P01009	280	100.00	No
P01009	VAR_007001	P01009	354	3.32	P05120	352	31.15	No
P01009	VAR_007005	P01009	382	2.73	P01009	382	100.00	No
P01009	VAR_007006	P01009	386	2.33	P01009	386	100.00	No
P01009	VAR_007007	P01009	386	2.33	P01009	386	100.00	No
P01009	VAR_007009	P01009	393	4.17	P01009	393	100.00	No
P01111	VAR_006845	P01111	13	3.01	P01112	13	91.88	No
P01111	VAR_006846	P01111	61	4.56	P01112	61	91.88	No
P01111	VAR_021194	P01111	12	2.54	P01112	12	91.88	No
P01112	VAR_006836	P01112	12	2.54	P01112	12	100.00	No
P01112	VAR_006837	P01112	12	2.54	P01112	12	100.00	No
P01112	VAR_006838	P01112	61	4.56	P01112	61	100.00	No
P01116	VAR_006839	P01116	12	2.54	P01112	12	94.38	No
P01116	VAR_006840	P01116	12	2.54	P01112	12	94.38	No
P01116	VAR_006841	P01116	61	4.56	P01112	61	94.38	No
P01116	VAR_016026	P01116	12	2.54	P01112	12	94.38	No
P01116	VAR_016027	P01116	12	2.54	P01112	12	94.38	No
P01116	VAR_016028	P01116	12	2.54	P01112	12	94.38	No
P01116	VAR_016029	P01116	13	3.01	P01112	13	94.38	No
P01116	VAR_016030	P01116	59	3.71	P01112	59	94.38	No
P01130	VAR_005361	P01130	327	3.20	P09871	145	38.24	No
P01130	VAR_005362	P01130	329	5.88	P09871	147	38.24	No
P01130	VAR_005367	P01130	343	3.29	P09871	161	38.24	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P01130	VAR_005373	P01130	364	5.87	Q9JJS8	152	40.63	No
P01241	VAR_015814	P01241	201	2.34	P01241	201	100.00	No
P01241	VAR_015815	P01241	209	3.80	P01241	209	100.00	No
P01308	VAR_003971	P01308	34	3.49	P01308	34	100.00	No
P01308	VAR_003972	P01308	48	3.24	P01308	48	100.00	No
P01308	VAR_003973	P01308	49	3.54	P01308	49	100.00	No
P01308	VAR_003976	P01308	92	3.28	P01308	92	100.00	No
P01857	VAR_003888	P01857	241	2.07	P01865	240	61.63	No
P02042	VAR_003104	P02042	26	2.10	P68871	26	92.54	No
P02042	VAR_003113	P02042	98	2.88	P68871	98	92.54	No
P02042	VAR_003114	P02042	99	2.45	P68871	99	92.54	No
P02452	VAR_001644	P02452	221	3.86	P02452	151	47.37	No
P02452	VAR_001646	P02452	263	3.76	P02452	133	35.09	No
P02452	VAR_001647	P02452	263	3.76	P02452	133	35.09	No
P02452	VAR_001648	P02452	272	3.86	P02452	142	35.09	No
P02452	VAR_001649	P02452	275	3.86	P02452	145	35.09	No
P02452	VAR_001650	P02452	332	3.86	P02452	142	45.61	No
P02452	VAR_001654	P02452	383	3.76	P02452	133	40.35	No
P02452	VAR_001655	P02452	389	3.86	P02452	139	40.35	No
P02452	VAR_001656	P02452	389	3.86	P02452	139	40.35	No
P02452	VAR_001657	P02452	398	3.86	P02452	148	40.35	No
P02452	VAR_001658	P02452	398	3.86	P02452	148	40.35	No
P02452	VAR_001659	P02452	401	3.86	P02452	151	40.35	No
P02452	VAR_001672	P02452	569	3.86	P02452	139	43.86	No
P02452	VAR_001675	P02452	638	3.86	P02452	148	43.86	No
P02452	VAR_001677	P02452	701	3.86	P02452	151	40.35	No
P02452	VAR_001683	P02452	743	3.76	P02452	133	40.35	No
P02452	VAR_001684	P02452	743	3.76	P02452	133	40.35	No
P02452	VAR_001688	P02452	809	3.86	P02452	136	42.11	No
P02452	VAR_001689	P02452	815	3.86	P02452	142	42.11	No
P02452	VAR_001690	P02452	821	3.86	P02452	148	42.11	No
P02452	VAR_001696	P02452	869	3.86	P02452	136	43.86	No
P02452	VAR_001697	P02452	884	3.86	P02452	151	43.86	No
P02452	VAR_001699	P02452	926	3.76	P02452	133	42.11	No
P02452	VAR_001708	P02452	1049	3.86	P02452	136	43.86	Yes
P02452	VAR_001709	P02452	1058	3.86	P02452	145	43.86	Yes
P02452	VAR_001710	P02452	1061	3.86	P02452	148	43.86	Yes
P02452	VAR_001711	P02452	1061	3.86	P02452	148	43.86	Yes
P02452	VAR_001719	P02452	1106	3.76	P02452	133	45.61	Yes
P02452	VAR_001720	P02452	1124	3.86	P02452	151	45.61	Yes
P02452	VAR_001725	P02452	1166	3.76	P02452	133	43.86	Yes
P02452	VAR_001726	P02452	1172	3.86	P02452	139	43.86	Yes
P02452	VAR_001727	P02452	1181	3.86	P02452	148	43.86	Yes
P02452	VAR_001728	P02452	1184	3.86	P02452	151	43.86	Yes
P02452	VAR_008118	P02452	866	3.76	P02452	133	43.86	No
P02458	VAR_001742	P02458	285	3.86	P02452	142	49.12	No
P02458	VAR_001749	P02458	705	3.86	P02452	142	42.11	No
P02458	VAR_001752	P02458	822	3.86	P02452	136	47.37	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P02458	VAR_001757	P02458	948	3.86	P02452	142	45.61	No
P02458	VAR_001761	P02458	1074	3.86	P02452	148	45.61	Yes
P02458	VAR_001764	P02458	1119	3.76	P02452	133	43.86	Yes
P02458	VAR_001765	P02458	1128	3.86	P02452	142	43.86	Yes
P02458	VAR_017641	P02458	702	3.86	P02452	139	42.11	No
P02458	VAR_017642	P02458	711	3.86	P02452	148	42.11	No
P02458	VAR_017644	P02458	825	3.86	P02452	139	47.37	No
P02458	VAR_017646	P02458	879	3.76	P02452	133	43.86	No
P02458	VAR_023929	P02458	648	3.86	P02452	145	38.60	No
P02458	VAR_023931	P02458	828	3.86	P02452	142	47.37	No
P02458	VAR_024820	P02458	648	3.86	P02452	145	38.60	No
P02458	VAR_024821	P02458	702	3.86	P02452	139	42.11	No
P02461	VAR_001769	P02461	201	3.86	P02452	139	49.12	No
P02461	VAR_001773	P02461	567	3.86	P02452	139	42.11	No
P02461	VAR_001780	P02461	756	3.86	P02452	148	42.11	No
P02461	VAR_001783	P02461	804	3.76	P02452	133	40.35	No
P02461	VAR_001786	P02461	936	3.86	P02452	145	42.11	No
P02461	VAR_001787	P02461	936	3.86	P02452	145	42.11	No
P02461	VAR_001788	P02461	939	3.86	P02452	148	42.11	No
P02461	VAR_001791	P02461	996	3.86	P02452	145	38.60	No
P02461	VAR_001793	P02461	1050	3.86	P02452	139	43.86	Yes
P02461	VAR_001797	P02461	1104	3.76	P02452	133	43.86	Yes
P02461	VAR_001798	P02461	1164	3.76	P02452	133	38.60	Yes
P02461	VAR_001799	P02461	1167	3.86	P02452	136	38.60	Yes
P02461	VAR_001800	P02461	1170	3.86	P02452	139	38.60	Yes
P02461	VAR_001801	P02461	1173	3.86	P02452	142	38.60	Yes
P02461	VAR_001802	P02461	1176	3.86	P02452	145	38.60	Yes
P02461	VAR_001803	P02461	1182	3.86	P02452	151	38.60	Yes
P02461	VAR_011098	P02461	204	3.86	P02452	142	49.12	No
P02461	VAR_011099	P02461	204	3.86	P02452	142	49.12	No
P02461	VAR_011100	P02461	210	3.86	P02452	148	49.12	No
P02461	VAR_011111	P02461	264	3.86	P02452	136	40.35	No
P02461	VAR_011112	P02461	267	3.86	P02452	139	40.35	No
P02461	VAR_011113	P02461	321	3.76	P02452	133	42.11	No
P02461	VAR_011114	P02461	327	3.86	P02452	139	42.11	No
P02461	VAR_011117	P02461	444	3.86	P02452	136	36.84	No
P02461	VAR_011119	P02461	501	3.76	P02452	133	42.11	No
P02461	VAR_011120	P02461	519	3.86	P02452	151	42.11	No
P02461	VAR_011124	P02461	636	3.86	P02452	148	42.11	No
P02461	VAR_011128	P02461	699	3.86	P02452	151	43.86	No
P02461	VAR_011131	P02461	744	3.86	P02452	136	42.11	No
P02461	VAR_011134	P02461	879	3.86	P02452	148	52.63	No
P02461	VAR_011135	P02461	882	3.86	P02452	151	52.63	No
P02461	VAR_011140	P02461	924	3.76	P02452	133	42.11	No
P02461	VAR_011141	P02461	942	3.86	P02452	151	42.11	No
P02461	VAR_011144	P02461	984	3.76	P02452	133	38.60	No
P02461	VAR_011145	P02461	999	3.86	P02452	148	38.60	No
P02461	VAR_011149	P02461	1044	3.76	P02452	133	43.86	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P02461	VAR_011150	P02461	1050	3.86	P02452	139	43.86	Yes
P02461	VAR_011155	P02461	1164	3.76	P02452	133	38.60	Yes
P02461	VAR_011156	P02461	1164	3.76	P02452	133	38.60	Yes
P02461	VAR_011157	P02461	1170	3.86	P02452	139	38.60	Yes
P02461	VAR_011158	P02461	1173	3.86	P02452	142	38.60	Yes
P02461	VAR_011159	P02461	1179	3.86	P02452	148	38.60	Yes
P02533	VAR_003843	P02533	383	2.70	P08670	368	35.83	No
P02533	VAR_003844	P02533	414	4.94	P08670	399	35.83	No
P02533	VAR_003845	P02533	418	3.39	P08670	403	35.83	No
P02533	VAR_010450	P02533	376	2.96	P08670	361	35.83	No
P02533	VAR_010451	P02533	387	3.13	P08670	372	35.83	No
P02533	VAR_023724	P02533	407	3.26	P08670	392	35.83	No
P02533	VAR_023725	P02533	412	3.03	P08670	397	35.83	No
P02538	VAR_017076	P02538	468	3.39	P08670	403	37.13	No
P02545	VAR_009985	P02545	358	3.66	P08670	381	30.39	No
P02545	VAR_009986	P02545	371	2.86	P08670	394	30.39	No
P02545	VAR_016205	P02545	377	4.16	P08670	400	30.39	No
P02647	VAR_000610	P02647	84	2.73	P02647	84	100.00	No
P02647	VAR_000616	P02647	132	4.16	P02647	132	100.00	No
P02647	VAR_021362	P02647	180	3.02	P02647	180	100.00	No
P02679	VAR_002409	P02679	301	2.72	P02679	301	100.00	No
P02679	VAR_002410	P02679	301	2.72	P02679	301	100.00	No
P02679	VAR_002412	P02679	334	3.47	P02679	334	100.00	No
P02679	VAR_002413	P02679	334	3.47	P02679	334	100.00	No
P02679	VAR_002414	P02679	336	4.27	P02679	336	100.00	No
P02679	VAR_015853	P02679	335	3.07	P02679	335	100.00	No
P02708	VAR_000285	P02708	299	3.75	P02711	278	81.77	No
P02708	VAR_021207	P02708	294	2.58	P02711	273	81.77	No
P02708	VAR_021208	P02708	301	3.79	P02711	280	81.77	No
P02730	VAR_000800	P02730	327	2.02	P02730	327	100.00	No
P02730	VAR_013786	P02730	147	2.32	P02730	147	100.00	Yes
P02766	VAR_007548	P02766	38	4.03	P02766	38	100.00	No
P02766	VAR_007549	P02766	38	4.03	P02766	38	100.00	No
P02766	VAR_007551	P02766	44	4.22	P02766	44	100.00	No
P02766	VAR_007577	P02766	89	4.94	O93330	91	57.27	No
P02766	VAR_007592	P02766	127	2.12	P02766	127	100.00	No
P02766	VAR_007594	P02766	131	2.94	O93330	133	57.27	No
P02766	VAR_007595	P02766	134	3.84	P02766	134	100.00	No
P02766	VAR_007596	P02766	136	4.70	P02766	136	100.00	No
P02766	VAR_007597	P02766	136	4.70	P02766	136	100.00	No
P02766	VAR_007598	P02766	134	3.84	P02766	134	100.00	No
P02768	VAR_000511	P02768	143	2.26	P02768	143	100.00	Yes
P02768	VAR_000523	P02768	345	2.44	P02768	345	100.00	No
P03951	VAR_012093	P03951	430	2.81	P00761	47	40.38	No
P03951	VAR_012096	P03951	594	3.22	P03951	594	100.00	No
P04070	VAR_006648	P04070	108	2.19	P00740	105	43.33	No
P04070	VAR_006649	P04070	109	3.30	P00740	106	43.33	No
P04070	VAR_006657	P04070	147	5.88	P00742	136	48.39	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P04070	VAR_006658	P04070	149	3.20	P00743	138	54.84	No
P04070	VAR_006670	P04070	226	3.30	P00735	381	40.61	No
P04070	VAR_006671	P04070	243	2.86	P00735	399	40.61	No
P04070	VAR_006673	P04070	253	5.33	P00735	409	40.61	No
P04070	VAR_006679	P04070	289	2.88	P00735	454	40.61	No
P04070	VAR_006681	P04070	298	3.22	P00735	464	40.61	No
P04070	VAR_006682	P04070	301	3.00	P00735	467	40.61	No
P04070	VAR_006683	P04070	301	3.00	P00735	467	40.61	No
P04070	VAR_006687	P04070	321	3.20	P00735	487	40.61	No
P04070	VAR_006691	P04070	334	2.89	P00735	499	40.61	No
P04070	VAR_006693	P04070	343	3.37	P00735	508	40.61	No
P04070	VAR_006695	P04070	367	2.03	P00735	533	40.61	No
P04070	VAR_006697	P04070	385	3.78	P00735	551	40.61	No
P04070	VAR_006698	P04070	388	2.61	P00761	171	37.67	No
P04070	VAR_006699	P04070	388	2.61	P00761	171	37.67	No
P04070	VAR_006702	P04070	401	3.74	P00735	570	40.61	No
P04070	VAR_006704	P04070	423	3.14	P00735	594	40.61	No
P04070	VAR_006705	P04070	426	5.02	P00735	597	40.61	No
P04070	VAR_006706	P04070	433	2.93	P00735	604	40.61	No
P04070	VAR_006707	P04070	436	2.77	P00735	607	40.61	No
P04070	VAR_006708	P04070	441	3.32	P00735	612	40.61	No
P04070	VAR_006709	P04070	444	5.62	P00735	615	40.61	No
P04075	VAR_000550	P04075	128	3.95	P00883	128	99.14	No
P04080	VAR_002206	P04080	4	3.43	P04080	4	100.00	No
P04181	VAR_000568	P04181	93	3.59	P04181	93	100.00	No
P04181	VAR_000569	P04181	154	3.07	P04181	154	100.00	No
P04181	VAR_000570	P04181	180	3.11	P04181	180	100.00	No
P04181	VAR_000579	P04181	319	4.95	P04181	319	100.00	No
P04264	VAR_017825	P04264	478	4.09	P08670	396	38.76	No
P04264	VAR_017826	P04264	478	4.09	P08670	396	38.76	No
P04264	VAR_017827	P04264	481	4.94	P08670	399	38.76	No
P04264	VAR_017828	P04264	485	3.39	P08670	403	38.76	No
P04275	VAR_005802	P04275	1374	3.59	P04275	1374	100.00	Yes
P04275	VAR_005803	P04275	1374	3.59	P04275	1374	100.00	Yes
P04424	VAR_000677	P04424	111	2.39	P11447	107	47.62	No
P04424	VAR_000678	P04424	193	2.22	P24058	195	71.77	No
P04424	VAR_000679	P04424	286	2.79	P04424	286	100.00	No
P04629	VAR_009630	P04629	649	4.10	P08069	1134	43.61	Yes
P04629	VAR_009631	P04629	654	4.05	Q07912	256	39.92	Yes
P04629	VAR_009632	P04629	674	2.30	P06213	1183	44.91	Yes
P04637	VAR_005880	P04637	137	2.41	P04637	137	100.00	Yes
P04637	VAR_005881	P04637	138	3.26	P04637	138	100.00	Yes
P04637	VAR_005923	P04637	172	3.34	P04637	172	100.00	Yes
P04637	VAR_005927	P04637	174	2.74	P04637	174	100.00	Yes
P04637	VAR_005928	P04637	175	3.89	P04637	175	100.00	Yes
P04637	VAR_005929	P04637	175	3.89	P04637	175	100.00	Yes
P04637	VAR_005930	P04637	175	3.89	P04637	175	100.00	Yes
P04637	VAR_005931	P04637	175	3.89	P04637	175	100.00	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P04637	VAR_005932	P04637	175	3.89	P04637	175	100.00	Yes
P04637	VAR_005933	P04637	176	5.61	P04637	176	100.00	Yes
P04637	VAR_005934	P04637	176	5.61	P04637	176	100.00	Yes
P04637	VAR_005935	P04637	177	4.12	P04637	177	100.00	No
P04637	VAR_005939	P04637	184	2.31	P04637	184	100.00	No
P04637	VAR_005943	P04637	189	2.04	P04637	189	100.00	Yes
P04637	VAR_005944	P04637	190	2.53	P02340	187	88.66	Yes
P04637	VAR_005952	P04637	198	3.54	P04637	198	100.00	Yes
P04637	VAR_005955	P04637	213	3.89	P04637	213	100.00	Yes
P04637	VAR_005965	P04637	237	4.78	P04637	237	100.00	Yes
P04637	VAR_005969	P04637	241	3.27	P04637	241	100.00	No
P04637	VAR_005970	P04637	242	5.61	P04637	242	100.00	No
P04637	VAR_005971	P04637	245	3.65	P02340	242	88.66	Yes
P04637	VAR_005972	P04637	245	3.65	P02340	242	88.66	Yes
P04637	VAR_005973	P04637	245	3.65	P02340	242	88.66	Yes
P04637	VAR_005974	P04637	245	3.65	P02340	242	88.66	Yes
P04637	VAR_005975	P04637	245	3.65	P02340	242	88.66	Yes
P04637	VAR_005980	P04637	247	4.00	P04637	247	100.00	No
P04637	VAR_005981	P04637	248	3.89	P04637	248	100.00	No
P04637	VAR_005982	P04637	248	3.89	P04637	248	100.00	No
P04637	VAR_005983	P04637	248	3.89	P04637	248	100.00	No
P04637	VAR_005984	P04637	248	3.89	P04637	248	100.00	No
P04637	VAR_005985	P04637	249	3.89	P04637	249	100.00	Yes
P04637	VAR_005986	P04637	249	3.89	P04637	249	100.00	Yes
P04637	VAR_006000	P04637	277	5.61	P04637	277	100.00	No
P04637	VAR_006007	P04637	280	3.89	P04637	280	100.00	No
P04637	VAR_006008	P04637	280	3.89	P04637	280	100.00	No
P04637	VAR_006009	P04637	280	3.89	P04637	280	100.00	No
P05164	VAR_015377	P05164	173	4.41	P05164	173	100.00	No
P05164	VAR_015378	P05164	251	3.39	P05164	251	100.00	No
P05165	VAR_009088	P05165	52	2.72	P24182	16	55.36	Yes
P05166	VAR_000274	P05166	165	3.66	Q8GBW6	143	52.71	No
P05166	VAR_000275	P05166	168	2.88	Q8GBW6	146	52.71	No
P05166	VAR_000278	P05166	410	2.90	Q9X4K7	392	57.77	No
P05166	VAR_000279	P05166	497	2.64	Q9X4K7	488	57.77	No
P05166	VAR_000281	P05166	519	2.12	Q8GBW6	503	52.71	No
P05166	VAR_009082	P05166	205	2.47	Q9X4K7	185	57.77	No
P05166	VAR_009086	P05166	536	2.68	Q9X4K7	527	57.77	No
P05166	VAR_023849	P05166	112	3.86	Q9X4K7	92	57.77	No
P05166	VAR_023851	P05166	165	3.66	Q8GBW6	143	52.71	No
P05166	VAR_023852	P05166	188	2.45	Q9X4K7	168	57.77	No
P05166	VAR_023856	P05166	435	4.05	Q9X4K7	417	57.77	No
P05166	VAR_023857	P05166	439	3.66	Q9X4K7	421	57.77	No
P05166	VAR_023858	P05166	468	2.95	Q9X4K7	450	57.77	No
P05186	VAR_006149	P05186	71	4.02	Q9BHT8	45	47.20	No
P05186	VAR_006150	P05186	71	4.02	Q9BHT8	45	47.20	No
P05186	VAR_011087	P05186	381	3.47	P00634	394	32.61	No
P05186	VAR_013975	P05186	71	4.02	Q9BHT8	45	47.20	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P05997	VAR_013588	P05997	960	3.86	P02452	136	42.11	No
P06213	VAR_015927	P06213	1158	4.10	P08069	1134	82.71	Yes
P06213	VAR_015928	P06213	1158	4.10	P08069	1134	82.71	Yes
P06400	VAR_005579	P06400	567	3.66	P06400	567	100.00	No
P06400	VAR_005581	P06400	654	2.94	P06400	654	100.00	No
P06400	VAR_005582	P06400	661	4.22	P06400	661	100.00	No
P06400	VAR_010049	P06400	530	3.97	P06400	530	100.00	No
P06400	VAR_010050	P06400	657	2.96	P06400	657	100.00	No
P06744	VAR_002528	P06744	342	3.27	P06744	342	100.00	No
P06744	VAR_002529	P06744	346	2.58	P06744	346	100.00	No
P06744	VAR_002530	P06744	346	2.58	P06744	346	100.00	No
P06744	VAR_002531	P06744	374	2.60	P06744	374	100.00	No
P06744	VAR_002532	P06744	388	5.37	P06744	388	100.00	No
P06744	VAR_002536	P06744	516	2.65	P06744	516	100.00	No
P06744	VAR_002537	P06744	524	2.99	P06744	524	100.00	No
P06744	VAR_002538	P06744	538	3.10	P06744	538	100.00	No
P06865	VAR_003203	P06865	39	2.01	P07686	72	38.46	Yes
P07195	VAR_004177	P07195	171	2.01	P07195	171	100.00	No
P07195	VAR_011634	P07195	68	2.07	P07195	68	100.00	No
P07195	VAR_011636	P07195	171	2.01	P07195	171	100.00	No
P07196	VAR_009703	P07196	331	2.92	P08670	342	53.25	No
P07202	VAR_006060	P07202	453	3.61	P05164	462	47.52	No
P07202	VAR_021623	P07202	240	3.95	P05164	262	47.52	No
P07202	VAR_021625	P07202	326	2.35	P05164	339	47.52	No
P07202	VAR_021629	P07202	493	3.30	P05164	501	47.52	No
P07202	VAR_021632	P07202	660	4.30	P05164	668	47.52	No
P07320	VAR_010733	P07320	14	2.66	P08209	14	87.34	No
P07320	VAR_021145	P07320	23	3.23	P62697	129	39.24	No
P07477	VAR_011656	P07477	139	4.20	P00761	124	78.97	No
P07741	VAR_006747	P07741	64	2.02	P49435	67	47.41	No
P07902	VAR_002553	P07902	51	3.44	P09148	31	52.30	No
P07902	VAR_002554	P07902	55	2.49	P09148	35	52.30	No
P07902	VAR_002559	P07902	97	3.15	P09148	77	52.30	No
P07902	VAR_002560	P07902	98	3.15	P09148	78	52.30	No
P07902	VAR_002563	P07902	117	3.46	P09148	97	52.30	No
P07902	VAR_002564	P07902	118	2.31	P09148	98	52.30	No
P07902	VAR_002583	P07902	171	3.78	P09148	151	52.30	No
P07902	VAR_002584	P07902	179	3.23	P09148	159	52.30	No
P07902	VAR_002585	P07902	183	2.41	P09148	163	52.30	No
P07902	VAR_002589	P07902	194	3.47	P09148	174	52.30	No
P07902	VAR_002594	P07902	201	2.00	P09148	181	60.36	No
P07902	VAR_002596	P07902	209	4.60	P09148	189	60.36	No
P07902	VAR_002597	P07902	209	4.60	P09148	189	60.36	No
P07902	VAR_002599	P07902	217	2.50	P09148	197	60.36	No
P07902	VAR_002601	P07902	231	3.48	P09148	211	60.36	No
P07902	VAR_002602	P07902	249	6.28	P09148	229	60.36	No
P07902	VAR_002618	P07902	323	3.43	P09148	300	60.36	No
P07902	VAR_002619	P07902	323	3.43	P09148	300	60.36	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P07902	VAR_008042	P07902	45	2.27	P09148	25	52.30	No
P07902	VAR_023328	P07902	51	3.44	P09148	31	52.30	No
P07949	VAR_006338	P07949	873	4.10	P08069	1134	41.57	Yes
P07949	VAR_006342	P07949	918	4.05	Q06187	563	35.74	Yes
P07949	VAR_006345	P07949	946	3.01	Q07912	325	37.01	Yes
P07949	VAR_006347	P07949	973	4.13	P11362	722	55.31	No
P07951	VAR_013468	P07951	117	2.08	P42639	117	85.59	Yes
P07951	VAR_013469	P07951	147	4.37	P42639	147	85.59	No
P07951	VAR_016086	P07951	91	3.43	P42639	91	85.59	No
P07954	VAR_002447	P07954	312	2.47	Q9LCC6	265	46.34	No
P07954	VAR_013501	P07954	233	3.61	P05042	186	60.79	No
P08123	VAR_001862	P08123	433	3.86	P02452	151	42.11	No
P08123	VAR_001866	P08123	547	3.86	P02452	145	40.35	No
P08123	VAR_001874	P08123	670	3.86	P02452	148	42.11	No
P08123	VAR_001878	P08123	730	3.86	P02452	145	40.35	No
P08123	VAR_001879	P08123	736	3.86	P02452	151	40.35	No
P08123	VAR_001884	P08123	778	3.76	P02452	133	47.37	No
P08123	VAR_001885	P08123	784	3.86	P02452	139	47.37	No
P08123	VAR_001886	P08123	787	3.86	P02452	142	47.37	No
P08123	VAR_001887	P08123	790	3.86	P02452	145	47.37	No
P08123	VAR_001888	P08123	796	3.86	P02452	151	47.37	No
P08123	VAR_001900	P08123	1078	3.76	P02452	133	45.61	Yes
P08123	VAR_001901	P08123	1096	3.86	P02452	151	45.61	Yes
P08123	VAR_008120	P08123	973	3.86	P02452	148	38.60	No
P08185	VAR_016223	P08185	389	2.51	P01011	405	47.98	No
P08237	VAR_006063	P08237	38	3.77	P00512	25	48.00	No
P08237	VAR_006064	P08237	38	3.77	P00512	25	48.00	No
P08237	VAR_006067	P08237	542	3.07	P00512	140	35.96	No
P08246	VAR_009538	P08246	32	3.08	P00747	583	38.31	No
P08246	VAR_009539	P08246	177	2.19	P00747	724	38.31	Yes
P08519	VAR_006633	P08519	4193	4.61	P00747	173	58.11	Yes
P08559	VAR_004952	P08559	167	2.36	P08559	167	100.00	No
P08559	VAR_004954	P08559	205	3.02	P08559	205	100.00	Yes
P08559	VAR_004957	P08559	231	3.70	P08559	231	100.00	No
P08581	VAR_006290	P08581	1228	2.30	P06213	1183	41.18	Yes
P08581	VAR_006291	P08581	1228	2.30	P06213	1183	41.18	Yes
P08581	VAR_006292	P08581	1230	3.29	P06213	1185	41.18	Yes
P08581	VAR_006293	P08581	1230	3.29	P06213	1185	41.18	Yes
P08581	VAR_006294	P08581	1250	4.05	Q06187	563	35.34	Yes
P08603	VAR_019406	P08603	959	5.88	P08174	253	37.74	No
P08603	VAR_025865	P08603	630	5.88	P08174	225	37.74	Yes
P08603	VAR_025868	P08603	951	2.61	P08174	245	37.74	No
P08603	VAR_025876	P08603	1142	3.98	P68638	180	38.46	Yes
P08603	VAR_025877	P08603	1157	6.30	P20023	75	36.54	Yes
P08709	VAR_006506	P08709	238	5.53	P00763	48	42.99	No
P08709	VAR_006508	P08709	304	3.00	P00761	94	41.12	No
P08709	VAR_006509	P08709	307	2.12	P00760	109	42.06	Yes
P08709	VAR_006516	P08709	402	3.76	P00763	198	42.99	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P08709	VAR_006517	P08709	402	3.76	P00763	198	42.99	No
P08709	VAR_014407	P08709	121	5.88	P00740	108	61.29	No
P08709	VAR_014416	P08709	304	3.00	P00761	94	41.12	No
P08709	VAR_014417	P08709	307	2.12	P00760	109	42.06	Yes
P08709	VAR_014419	P08709	391	2.73	P00747	747	36.57	Yes
P08709	VAR_014420	P08709	435	2.93	P00763	227	42.99	No
P08709	VAR_015141	P08709	312	2.32	P00761	102	41.12	No
P08709	VAR_015143	P08709	363	2.71	P00761	149	41.12	No
P08709	VAR_015144	P08709	403	3.74	P00763	199	42.99	No
P08779	VAR_017067	P08779	353	2.41	P08670	336	34.53	No
P09417	VAR_006965	P09417	145	2.70	P11348	142	96.84	No
P09493	VAR_007601	P09493	175	2.93	P42639	175	98.73	No
P09622	VAR_006908	P09622	488	4.27	P09624	479	77.98	No
P10153	VAR_013150	P10153	156	5.45	P61823	145	39.47	No
P10275	VAR_004685	P10275	608	4.26	P03372	234	53.33	No
P10275	VAR_009746	P10275	601	5.86	P15207	584	100.00	No
P10275	VAR_009747	P10275	604	3.09	P34021	309	42.67	Yes
P10275	VAR_009749	P10275	611	5.86	P06536	492	76.00	No
P10275	VAR_009783	P10275	720	3.13	P10275	720	100.00	No
P10275	VAR_009788	P10275	725	4.05	P10275	725	100.00	No
P10275	VAR_009792	P10275	733	3.70	P10275	733	100.00	No
P10619	VAR_001386	P10619	65	5.35	Q8W4X3	97	30.87	No
P10619	VAR_001389	P10619	395	4.19	Q8W4X3	409	30.87	No
P10721	VAR_004107	P10721	791	4.10	P08069	1134	35.71	Yes
P10721	VAR_004109	P10721	816	2.30	P06213	1183	39.10	Yes
P10721	VAR_023828	P10721	816	2.30	P06213	1183	39.10	Yes
P11177	VAR_021057	P11177	132	2.52	P11177	132	100.00	No
P11217	VAR_014004	P11217	291	3.32	P00490	267	47.36	No
P11230	VAR_000287	P11230	285	3.28	Q6S3I0	281	60.53	No
P11230	VAR_000288	P11230	289	2.93	Q6S3I0	285	60.53	No
P11362	VAR_017890	P11362	666	6.26	Q06187	562	36.55	Yes
P11362	VAR_017891	P11362	719	3.75	P00519	458	41.83	Yes
P11413	VAR_002470	P11413	175	3.95	P11413	175	100.00	Yes
P11413	VAR_002476	P11413	211	2.66	P11413	211	100.00	No
P11413	VAR_002477	P11413	212	2.72	P11413	212	100.00	No
P11413	VAR_002478	P11413	215	4.53	P11413	215	100.00	No
P11413	VAR_002479	P11413	226	3.21	P11413	226	100.00	Yes
P11413	VAR_002480	P11413	226	3.21	P11413	226	100.00	Yes
P11413	VAR_002482	P11413	256	3.89	P11413	256	100.00	Yes
P11413	VAR_002483	P11413	273	3.59	P11413	273	100.00	Yes
P11413	VAR_002484	P11413	277	2.74	P11413	277	100.00	No
P11413	VAR_002503	P11413	409	3.54	P11413	409	100.00	No
P11413	VAR_002504	P11413	409	3.54	P11413	409	100.00	No
P11413	VAR_002506	P11413	438	3.55	P11413	438	100.00	No
P11413	VAR_002514	P11413	462	2.34	P11413	462	100.00	Yes
P11473	VAR_004662	P11473	73	4.26	P03372	234	46.67	No
P11473	VAR_004667	P11473	391	2.95	Q13133	415	39.44	No
P11488	VAR_009279	P11488	37	3.58	P63096	41	67.44	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P11498	VAR_015200	P11498	451	2.07	P24182	410	31.43	No
P12107	VAR_013583	P12107	625	3.86	P02452	148	42.11	No
P12107	VAR_013584	P12107	676	3.86	P02452	139	52.63	No
P12107	VAR_013587	P12107	1516	3.86	P02452	136	42.11	Yes
P12694	VAR_004969	P12694	190	2.44	P12694	190	100.00	No
P12694	VAR_015101	P12694	290	3.12	P84129	227	37.04	No
P12883	VAR_004573	P12883	403	2.33	P10587	405	51.93	No
P12883	VAR_004574	P12883	403	2.33	P10587	405	51.93	No
P12883	VAR_004586	P12883	731	2.19	P13538	733	81.80	No
P12883	VAR_014199	P12883	743	2.52	P10587	753	51.93	Yes
P12883	VAR_017747	P12883	532	2.27	P13538	534	81.80	No
P12883	VAR_020803	P12883	320	2.88	P10587	322	51.93	Yes
P13569	VAR_000167	P13569	504	2.89	P26361	504	85.96	No
P13569	VAR_000176	P13569	549	3.70	P13569	549	100.00	No
P13569	VAR_000177	P13569	549	3.70	P13569	549	100.00	No
P13569	VAR_000178	P13569	549	3.70	P13569	549	100.00	No
P13569	VAR_000197	P13569	579	3.99	P13569	579	100.00	No
P13569	VAR_000200	P13569	613	2.15	P13569	613	100.00	No
P13569	VAR_000201	P13569	614	3.21	P13569	614	100.00	No
P13569	VAR_000261	P13569	1282	2.52	Q9CHL8	421	31.67	Yes
P13569	VAR_000262	P13569	1283	2.74	Q9CHL8	422	31.67	Yes
P13569	VAR_000264	P13569	1291	4.20	Q9CHL8	430	31.67	Yes
P13569	VAR_000265	P13569	1291	4.20	Q9CHL8	430	31.67	Yes
P13569	VAR_000266	P13569	1303	3.39	Q9CHL8	442	31.67	Yes
P13569	VAR_000267	P13569	1303	3.39	Q9CHL8	442	31.67	Yes
P13645	VAR_003833	P13645	442	3.26	P08670	392	36.16	No
P13645	VAR_010510	P13645	439	3.93	P08670	389	36.16	No
P13645	VAR_010511	P13645	446	4.09	P08670	396	36.16	No
P13647	VAR_003876	P13647	463	3.26	P08670	392	37.13	No
P13647	VAR_010466	P13647	467	4.09	P08670	396	37.13	No
P13647	VAR_023726	P13647	404	2.34	P08670	333	37.13	No
P13716	VAR_003635	P13716	240	2.28	P13716	240	100.00	No
P13804	VAR_002368	P13804	266	3.56	P13804	266	100.00	No
P13942	VAR_010655	P13942	808	3.76	P02452	133	40.35	No
P14136	VAR_017475	P14136	362	3.97	P08670	395	63.96	No
P14770	VAR_024997	P14770	24	5.79	P07359	24	34.62	No
P15153	VAR_017452	P15153	57	4.23	P15153	57	100.00	No
P15735	VAR_009518	P15735	189	3.47	P05132	200	33.06	No
P15735	VAR_009518	P15735	189	3.47	P00517	200	33.06	No
P15735	VAR_020854	P15735	157	2.62	P49137	190	35.77	No
P16144	VAR_011297	P16144	336	2.73	P05106	347	35.37	No
P16219	VAR_000316	P16219	383	3.62	P15651	383	94.63	No
P17661	VAR_007902	P17661	392	3.13	P08670	387	73.38	No
P17661	VAR_009189	P17661	344	2.79	P08670	339	73.38	No
P17661	VAR_018771	P17661	384	2.43	P08670	379	73.38	No
P17661	VAR_018772	P17661	388	3.89	P08670	383	73.38	Yes
P19013	VAR_016038	P19013	449	3.97	P08670	395	38.76	No
P19429	VAR_016078	P19429	81	2.49	P19429	81	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P19438	VAR_013410	P19438	59	5.87	P19438	102	33.33	No
P19438	VAR_013411	P19438	62	5.87	P19438	62	100.00	No
P19438	VAR_019302	P19438	59	5.87	P19438	102	33.33	No
P19438	VAR_019303	P19438	62	5.87	P19438	62	100.00	No
P19438	VAR_019304	P19438	99	5.87	P19438	99	100.00	Yes
P19438	VAR_019329	P19438	51	2.26	P19438	91	33.33	No
P20594	VAR_022584	P20594	115	3.94	P18910	128	41.89	No
P20807	VAR_001367	P20807	490	3.92	Q07009	416	56.86	No
P20807	VAR_009560	P20807	214	3.81	Q07009	189	55.44	Yes
P20807	VAR_009561	P20807	215	2.37	Q07009	190	55.44	Yes
P20807	VAR_009574	P20807	440	3.45	Q07009	366	56.86	No
P20807	VAR_009584	P20807	490	3.92	Q07009	416	56.86	No
P20807	VAR_009589	P20807	567	2.66	Q07009	494	56.86	No
P20807	VAR_009595	P20807	705	4.21	Q64537	154	57.14	No
P20807	VAR_009596	P20807	705	4.21	Q64537	154	57.14	No
P20823	VAR_003759	P20823	272	4.16	P40424	288	34.62	No
P20823	VAR_010537	P20823	12	2.61	P22361	12	94.86	No
P20823	VAR_010553	P20823	200	3.27	P06601	214	36.54	No
P20823	VAR_010556	P20823	229	3.65	P06601	243	36.54	No
P20823	VAR_010563	P20823	272	4.16	P40424	288	34.62	No
P20823	VAR_012483	P20823	20	3.33	P22361	20	94.86	No
P20933	VAR_005069	P20933	60	2.94	P20933	60	100.00	No
P20933	VAR_005071	P20933	101	2.95	P20933	101	100.00	No
P20933	VAR_005075	P20933	306	2.91	Q47898	304	37.41	No
P20933	VAR_015429	P20933	135	3.73	P20933	135	100.00	No
P20933	VAR_015432	P20933	257	4.01	P20933	257	100.00	No
P21439	VAR_023504	P21439	1161	3.24	Q9CHL8	473	46.15	Yes
P21953	VAR_004974	P21953	206	2.83	P84130	139	53.11	No
P22033	VAR_004416	P22033	368	2.01	P11653	345	63.48	No
P22033	VAR_004417	P22033	369	4.22	P11653	346	63.48	No
P22830	VAR_002385	P22830	267	2.22	P22830	267	100.00	No
P23760	VAR_003804	P23760	238	4.46	P40424	252	35.71	No
P23760	VAR_003805	P23760	265	2.61	P02836	500	37.50	No
P23760	VAR_003806	P23760	271	4.16	P40424	288	35.71	No
P23760	VAR_017537	P23760	271	4.16	P40424	288	35.71	No
P23760	VAR_017538	P23760	271	4.16	P40424	288	35.71	No
P24752	VAR_007500	P24752	158	2.24	P07097	120	43.14	No
P24752	VAR_007501	P24752	183	3.07	P07097	146	43.14	No
P25054	VAR_005040	P25054	1027	4.37	P25054	1027	100.00	Yes
P25054	VAR_005044	P25054	1176	4.29	P25054	1024	53.33	Yes
P26367	VAR_003812	P26367	44	3.43	Q02548	56	77.42	Yes
P26440	VAR_015966	P26440	411	2.65	P26440	411	100.00	No
P28069	VAR_003778	P28069	143	4.06	P14859	299	58.11	Yes
P28358	VAR_022582	P28358	319	2.82	Q6B2C0	185	37.04	Yes
P28360	VAR_003754	P28360	196	3.65	P06601	243	42.86	No
P29400	VAR_001915	P29400	129	3.76	P02452	133	47.37	No
P29400	VAR_001916	P29400	129	3.76	P02452	133	47.37	No
P29400	VAR_001923	P29400	325	3.86	P02452	145	42.11	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P29400	VAR_001924	P29400	325	3.86	P02452	145	42.11	No
P29400	VAR_001929	P29400	383	3.76	P02452	133	42.86	No
P29400	VAR_001930	P29400	400	3.86	P02452	151	42.86	No
P29400	VAR_001939	P29400	521	3.86	P02452	136	38.60	No
P29400	VAR_001940	P29400	521	3.86	P02452	136	38.60	No
P29400	VAR_001942	P29400	609	3.86	P02452	151	43.86	No
P29400	VAR_001947	P29400	684	3.86	P02452	151	42.11	No
P29400	VAR_001950	P29400	796	3.86	P02452	142	47.37	No
P29400	VAR_001956	P29400	1104	3.86	P02452	142	40.35	Yes
P29400	VAR_001964	P29400	1421	3.76	P02452	133	43.86	Yes
P29400	VAR_001968	P29400	1517	4.30	Q7SIB2	61	58.23	Yes
P29400	VAR_001973	P29400	1649	2.97	Q7SIB2	193	58.02	Yes
P29400	VAR_001973	P29400	1649	2.97	P02462	1633	58.02	Yes
P29400	VAR_001974	P29400	1677	4.20	Q7SIB2	221	58.02	Yes
P29400	VAR_007992	P29400	331	3.86	P02452	151	42.11	No
P29400	VAR_008000	P29400	669	3.86	P02452	136	42.11	No
P29400	VAR_008008	P29400	1107	3.86	P02452	145	40.35	Yes
P29400	VAR_008009	P29400	1161	3.86	P02452	139	42.11	Yes
P29400	VAR_008011	P29400	1220	3.86	P02452	139	40.35	Yes
P29400	VAR_008012	P29400	1333	3.76	P02452	133	36.84	Yes
P29400	VAR_008013	P29400	1427	3.86	P02452	139	43.86	Yes
P29400	VAR_011221	P29400	192	3.86	P02452	136	54.39	No
P29400	VAR_011222	P29400	204	3.86	P02452	148	54.39	No
P29400	VAR_011229	P29400	319	3.86	P02452	139	42.11	No
P29400	VAR_011237	P29400	524	3.86	P02452	139	38.60	No
P29400	VAR_011241	P29400	603	3.86	P02452	145	43.86	No
P29400	VAR_011242	P29400	609	3.86	P02452	151	43.86	No
P29400	VAR_011249	P29400	681	3.86	P02452	148	42.11	No
P29400	VAR_011253	P29400	802	3.86	P02452	148	47.37	No
P29400	VAR_011269	P29400	1036	3.76	P02452	133	42.11	Yes
P29400	VAR_011270	P29400	1039	3.86	P02452	136	42.11	Yes
P29400	VAR_011271	P29400	1045	3.86	P02452	142	42.11	Yes
P29400	VAR_011275	P29400	1158	3.86	P02452	136	42.11	Yes
P29400	VAR_011276	P29400	1167	3.86	P02452	145	42.11	Yes
P29400	VAR_011277	P29400	1170	3.86	P02452	148	42.11	Yes
P29400	VAR_011281	P29400	1229	3.86	P02452	148	40.35	Yes
P29400	VAR_011290	P29400	1677	4.20	Q7SIB2	221	58.02	Yes
P29965	VAR_007524	P29965	227	2.84	P29965	227	100.00	No
P29965	VAR_017923	P29965	170	3.56	P29965	170	100.00	No
P29965	VAR_017927	P29965	174	4.04	P29965	174	100.00	No
P29965	VAR_017938	P29965	226	2.87	P29965	226	100.00	No
P30613	VAR_004042	P30613	337	4.23	P30613	337	100.00	No
P30613	VAR_004043	P30613	337	4.23	P30613	337	100.00	No
P30613	VAR_004044	P30613	339	4.19	P30613	339	100.00	No
P30613	VAR_004045	P30613	341	3.84	P30613	341	100.00	No
P30613	VAR_004052	P30613	384	4.05	P30613	384	100.00	No
P30613	VAR_004053	P30613	392	3.08	P30613	392	100.00	No
P30613	VAR_004054	P30613	393	4.41	P30613	393	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P30613	VAR_004055	P30613	393	4.41	P30613	393	100.00	No
P30613	VAR_004061	P30613	431	2.16	P30613	431	100.00	No
P30613	VAR_004075	P30613	559	2.22	P11974	515	59.17	Yes
P30613	VAR_004076	P30613	566	4.20	P30613	566	100.00	No
P30613	VAR_011445	P30613	222	3.84	P30613	222	100.00	No
P30613	VAR_011454	P30613	341	3.84	P30613	341	100.00	No
P30613	VAR_011455	P30613	342	2.93	P30613	342	100.00	No
P30613	VAR_011456	P30613	348	2.24	P30613	348	100.00	No
P30613	VAR_011459	P30613	376	3.70	P30613	376	100.00	No
P30613	VAR_011460	P30613	387	3.93	P30613	387	100.00	No
P30613	VAR_011461	P30613	390	4.19	P30613	390	100.00	No
P30613	VAR_011478	P30613	385	4.23	P30613	385	100.00	No
P30613	VAR_011480	P30613	479	2.41	P11974	435	59.17	Yes
P30613	VAR_011482	P30613	569	3.48	P30613	569	100.00	No
P30793	VAR_002638	P30793	134	2.33	P30793	134	100.00	No
P30793	VAR_002640	P30793	144	3.94	P30793	144	100.00	No
P30793	VAR_002644	P30793	186	3.10	P22288	177	97.12	No
P30793	VAR_002647	P30793	211	4.32	P22288	202	97.12	No
P30793	VAR_016896	P30793	135	2.72	P30793	135	100.00	No
P30793	VAR_016902	P30793	199	3.70	P22288	190	97.12	No
P30793	VAR_016903	P30793	211	4.32	P22288	202	97.12	No
P30793	VAR_016904	P30793	213	4.22	P30793	213	100.00	No
P31271	VAR_017775	P31271	371	4.16	P02836	503	37.50	No
P31271	VAR_017776	P31271	372	4.42	P02836	504	37.50	No
P33527	VAR_011489	P33527	671	3.57	P13569	451	42.69	No
P35240	VAR_000814	P35240	117	2.42	P11171	304	31.02	No
P35240	VAR_000825	P35240	535	2.05	P26038	517	38.76	No
P35240	VAR_000826	P35240	538	3.26	P26038	520	38.76	No
P35453	VAR_015953	P35453	314	2.38	P02836	500	33.93	No
P35520	VAR_002172	P35520	87	4.06	P35520	87	100.00	No
P35520	VAR_002176	P35520	130	2.28	P35520	130	100.00	Yes
P35520	VAR_008051	P35520	84	2.63	Q9WZD3	7	46.15	No
P35520	VAR_008064	P35520	151	2.48	Q9WZD3	66	46.15	No
P35520	VAR_021792	P35520	108	2.07	P35520	108	100.00	No
P35548	VAR_010201	P35548	172	3.65	P06601	243	46.43	No
P35555	VAR_002278	P35555	129	5.88	P00740	108	50.00	No
P35555	VAR_002291	P35555	723	4.22	P07204	441	42.42	No
P35555	VAR_002323	P35555	1249	5.87	Q9JJS8	152	37.50	Yes
P35555	VAR_002331	P35555	1893	2.77	P07204	443	44.83	Yes
P35555	VAR_002339	P35555	2258	5.87	Q9JJS8	152	40.63	Yes
P35555	VAR_017971	P35555	154	5.88	P05106	562	34.78	No
P35555	VAR_017974	P35555	560	3.48	P07204	469	45.16	Yes
P35555	VAR_017985	P35555	723	4.22	P07204	441	42.42	No
P35555	VAR_017986	P35555	734	5.87	Q9JJS8	152	38.24	No
P35555	VAR_017988	P35555	776	5.87	Q9JJS8	152	34.38	No
P35555	VAR_017989	P35555	776	5.87	Q9JJS8	152	34.38	No
P35555	VAR_017991	P35555	816	5.88	P09871	143	37.50	No
P35555	VAR_017995	P35555	921	5.87	Q9JJS8	152	30.77	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P35555	VAR_018007	P35555	1374	5.87	Q9JJS8	152	40.63	Yes
P35555	VAR_018019	P35555	1796	3.48	P07204	469	36.84	Yes
P35555	VAR_023865	P35555	541	5.87	Q9JJS8	152	34.38	No
P35555	VAR_023871	P35555	832	5.88	P08709	132	52.00	No
P35555	VAR_023873	P35555	1058	3.48	P07204	469	40.63	Yes
P35555	VAR_023881	P35555	1333	5.87	Q9JJS8	152	35.14	Yes
P35555	VAR_023884	P35555	1475	3.48	P07204	469	51.61	Yes
P35555	VAR_023885	P35555	1475	3.48	P07204	469	51.61	Yes
P35555	VAR_023895	P35555	1900	5.87	Q9JJS8	152	40.00	Yes
P35556	VAR_002350	P35556	1252	5.87	Q9JJS8	152	41.03	Yes
P35556	VAR_010741	P35556	1252	5.87	Q9JJS8	152	41.03	Yes
P35557	VAR_003698	P35557	175	2.67	P19367	179	48.04	No
P35557	VAR_003709	P35557	279	2.31	P05708	283	51.88	No
P35557	VAR_003711	P35557	300	3.58	P05708	304	51.88	No
P35557	VAR_003712	P35557	300	3.58	P05708	304	51.88	No
P35557	VAR_010586	P35557	108	4.34	P19367	560	55.61	No
P35557	VAR_010587	P35557	137	2.30	P19367	589	55.61	No
P35625	VAR_007509	P35625	191	2.22	P16035	200	45.88	No
P35908	VAR_009186	P35908	482	3.97	P08670	395	35.18	No
P35908	VAR_009187	P35908	485	3.45	P08670	398	35.18	No
P35908	VAR_010516	P35908	490	3.39	P08670	403	35.18	No
P35916	VAR_018413	P35916	1041	4.05	Q07912	256	35.04	Yes
P35916	VAR_018415	P35916	1114	4.23	Q06187	596	36.69	Yes
P35916	VAR_018416	P35916	1137	4.13	P11362	722	52.38	Yes
P36897	VAR_022344	P36897	200	3.76	P36897	200	100.00	No
P37173	VAR_022352	P37173	336	2.75	P36897	291	41.05	Yes
P37231	VAR_010728	P37231	495	2.39	Q07869	458	69.23	Yes
P38117	VAR_002369	P38117	163	3.92	P38117	164	100.00	No
P38117	VAR_025804	P38117	127	3.97	P38117	127	100.00	No
P38117	VAR_025804	P38117	127	3.97	P38117	128	100.00	No
P40337	VAR_005742	P40337	155	2.93	P40337	155	100.00	No
P40337	VAR_005743	P40337	156	4.15	P40337	156	100.00	No
P40337	VAR_005744	P40337	156	4.15	P40337	156	100.00	No
P40337	VAR_005746	P40337	157	3.07	P40337	157	100.00	No
P40337	VAR_005748	P40337	158	2.61	P40337	158	100.00	No
P40337	VAR_005749	P40337	158	2.61	P40337	158	100.00	No
P40337	VAR_008101	P40337	155	2.93	P40337	155	100.00	No
P40692	VAR_004438	P40692	64	2.54	P54278	71	37.11	No
P42771	VAR_001412	P42771	23	2.65	P42771	23	100.00	No
P42771	VAR_001440	P42771	74	2.61	Q60773	71	51.61	No
P42771	VAR_001441	P42771	74	2.61	Q60773	71	51.61	No
P42771	VAR_001453	P42771	89	2.65	P42771	89	100.00	No
P42771	VAR_001454	P42771	89	2.65	P42771	89	100.00	No
P43034	VAR_007724	P43034	148	4.99	P62871	53	33.33	No
P43034	VAR_015398	P43034	30	3.95	P63005	30	100.00	No
P43246	VAR_004488	P43246	834	3.36	P23909	779	48.91	No
P43403	VAR_015538	P43403	465	4.05	Q07912	256	38.98	Yes
P43681	VAR_000295	P43681	280	2.61	P02711	272	50.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P43681	VAR_017531	P43681	280	2.61	P02711	272	50.00	No
P43699	VAR_015189	P43699	213	4.16	P40424	288	32.14	No
P49748	VAR_000349	P49748	366	3.75	P15651	297	38.10	No
P49748	VAR_000350	P49748	366	3.75	P15651	297	38.10	No
P49748	VAR_000356	P49748	453	3.62	P15651	383	38.10	No
P49748	VAR_000357	P49748	454	3.23	P15651	384	38.10	No
P49748	VAR_000358	P49748	456	2.95	P15651	386	38.10	No
P49748	VAR_000359	P49748	459	2.02	P15651	389	38.10	No
P49748	VAR_000361	P49748	469	2.65	Q06319	374	36.49	No
P49748	VAR_000362	P49748	469	2.65	Q06319	374	36.49	No
P50219	VAR_017876	P50219	248	3.39	P02836	459	50.00	Yes
P50219	VAR_017879	P50219	292	4.16	P02836	503	50.00	No
P50219	VAR_017881	P50219	295	4.16	P40424	288	35.71	No
P50219	VAR_017882	P50219	295	4.16	P40424	288	35.71	No
P51149	VAR_018722	P51149	129	2.78	P62825	126	32.91	No
P51149	VAR_018723	P51149	162	3.47	P62826	156	32.91	No
P51149	VAR_018723	P51149	162	3.47	P62826	157	32.91	No
P51159	VAR_010654	P51159	73	5.75	P63012	76	46.88	No
P51159	VAR_011335	P51159	152	3.49	P01112	134	33.96	Yes
P51587	VAR_020718	P51587	1524	4.41	P51587	1524	100.00	Yes
P51587	VAR_020725	P51587	2072	2.87	P51587	1538	44.12	Yes
P51812	VAR_006196	P51812	431	3.87	P05132	52	33.75	No
P52333	VAR_010498	P52333	910	3.26	Q06187	481	32.26	Yes
P52952	VAR_003752	P52952	178	2.30	P02836	494	48.21	Yes
P52952	VAR_010117	P52952	188	4.42	P02836	504	48.21	No
P53634	VAR_009541	P53634	249	3.15	P53634	249	100.00	No
P53634	VAR_009542	P53634	252	4.57	P53634	252	100.00	No
P53634	VAR_009544	P53634	301	3.78	P07711	181	38.54	No
P53634	VAR_016936	P53634	429	6.32	P53634	429	100.00	No
P53634	VAR_019038	P53634	236	4.19	P53634	236	100.00	No
P53634	VAR_019041	P53634	300	3.79	O46427	183	42.79	No
P53634	VAR_019042	P53634	300	3.79	O46427	183	42.79	No
P53634	VAR_019043	P53634	301	3.78	P07711	181	38.54	No
P53634	VAR_019046	P53634	319	3.20	P07711	199	38.54	No
P53634	VAR_019047	P53634	412	4.46	P53634	412	100.00	No
P55084	VAR_021130	P55084	118	4.26	P28790	73	37.50	No
P55084	VAR_021131	P55084	122	2.70	P28790	77	37.50	No
P55084	VAR_021132	P55084	134	3.58	P28790	89	37.50	No
P57727	VAR_011678	P57727	251	5.08	P00761	42	40.38	No
P57727	VAR_011679	P57727	404	3.88	P07338	216	41.28	No
P57727	VAR_013495	P57727	407	4.23	P07338	219	41.28	No
P58304	VAR_011618	P58304	200	4.16	P40424	288	32.14	No
P58304	VAR_011619	P58304	200	4.16	P40424	288	32.14	No
P60174	VAR_007535	P60174	72	3.80	P00939	72	98.32	No
P60174	VAR_007539	P60174	170	3.99	P04789	172	53.59	No
P61457	VAR_005530	P61457	96	2.39	P61459	96	100.00	No
P61626	VAR_004281	P61626	85	2.78	P61626	85	100.00	No
P62070	VAR_006848	P62070	72	4.56	P01112	61	61.88	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P63092	VAR_003441	P63092	201	4.27	P63096	177	41.62	No
P63092	VAR_003442	P63092	201	4.27	P63096	177	41.62	No
P63092	VAR_003443	P63092	227	4.62	P10824	203	41.91	No
P63092	VAR_017844	P63092	201	4.27	P63096	177	41.62	No
P63092	VAR_017845	P63092	201	4.27	P63096	177	41.62	No
P63092	VAR_017846	P63092	201	4.27	P63096	177	41.62	No
P63092	VAR_017847	P63092	227	4.62	P10824	203	41.91	No
P63092	VAR_017848	P63092	231	4.27	P04896	231	99.74	No
P68032	VAR_012857	P68032	101	2.27	P68135	101	98.93	No
P68032	VAR_012861	P68032	333	2.90	P68139	333	98.93	No
P68032	VAR_012862	P68032	363	3.71	P68135	363	98.93	Yes
P68133	VAR_011682	P68133	117	2.62	P68135	117	100.00	No
P68133	VAR_015579	P68133	42	2.20	P68135	42	100.00	No
P68133	VAR_015583	P68133	258	2.20	P68135	258	100.00	No
P68133	VAR_015586	P68133	288	3.61	P68135	288	100.00	No
P68133	VAR_015587	P68133	359	2.63	P68135	359	100.00	Yes
P68871	VAR_002878	P68871	15	5.78	P02118	15	69.40	Yes
P68871	VAR_002879	P68871	15	5.78	P02118	15	69.40	Yes
P68871	VAR_002882	P68871	17	2.30	P68871	17	100.00	No
P68871	VAR_002883	P68871	17	2.30	P68871	17	100.00	No
P68871	VAR_002884	P68871	17	2.30	P68871	17	100.00	No
P68871	VAR_002885	P68871	18	2.32	P02118	18	69.40	Yes
P68871	VAR_002886	P68871	19	2.34	P02118	19	69.40	Yes
P68871	VAR_002887	P68871	19	2.34	P02118	19	69.40	Yes
P68871	VAR_002888	P68871	19	2.34	P02118	19	69.40	Yes
P68871	VAR_002907	P68871	26	2.10	P68871	26	100.00	No
P68871	VAR_002908	P68871	26	2.10	P68871	26	100.00	No
P68871	VAR_002914	P68871	30	2.86	P68871	30	100.00	No
P68871	VAR_002919	P68871	35	2.99	P68871	35	100.00	No
P68871	VAR_002920	P68871	36	3.98	P68871	36	100.00	No
P68871	VAR_002921	P68871	36	3.98	P68871	36	100.00	No
P68871	VAR_002922	P68871	36	3.98	P68871	36	100.00	No
P68871	VAR_002923	P68871	37	3.09	P68871	37	100.00	No
P68871	VAR_002924	P68871	37	3.09	P68871	37	100.00	No
P68871	VAR_002925	P68871	37	3.09	P68871	37	100.00	No
P68871	VAR_002943	P68871	52	2.02	P02118	52	69.40	No
P68871	VAR_002944	P68871	52	2.02	P02118	52	69.40	No
P68871	VAR_002961	P68871	66	2.40	P02089	66	79.85	No
P68871	VAR_002962	P68871	67	3.12	P02089	67	79.85	No
P68871	VAR_002963	P68871	67	3.12	P02089	67	79.85	No
P68871	VAR_002969	P68871	70	2.18	P02089	70	79.85	No
P68871	VAR_002979	P68871	77	3.17	P02089	77	79.85	No
P68871	VAR_002980	P68871	77	3.17	P02089	77	79.85	No
P68871	VAR_002981	P68871	77	3.17	P02089	77	79.85	No
P68871	VAR_002982	P68871	78	2.52	P02089	78	79.85	No
P68871	VAR_002983	P68871	79	2.25	P02089	79	79.85	No
P68871	VAR_002993	P68871	88	2.86	P68871	88	100.00	No
P68871	VAR_002994	P68871	88	2.86	P68871	88	100.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P68871	VAR_003001	P68871	92	5.44	P02089	92	79.85	No
P68871	VAR_003002	P68871	92	5.44	P02089	92	79.85	No
P68871	VAR_003003	P68871	92	5.44	P02089	92	79.85	No
P68871	VAR_003004	P68871	92	5.44	P02089	92	79.85	No
P68871	VAR_003013	P68871	97	3.47	P68871	97	100.00	No
P68871	VAR_003014	P68871	97	3.47	P68871	97	100.00	No
P68871	VAR_003015	P68871	97	3.47	P68871	97	100.00	No
P68871	VAR_003016	P68871	97	3.47	P68871	97	100.00	No
P68871	VAR_003017	P68871	98	2.88	P68871	98	100.00	No
P68871	VAR_003018	P68871	99	2.45	P68871	99	100.00	No
P68871	VAR_003019	P68871	100	3.08	P68871	100	100.00	No
P68871	VAR_003020	P68871	100	3.08	P68871	100	100.00	No
P68871	VAR_003025	P68871	102	2.60	P68871	102	100.00	No
P68871	VAR_003026	P68871	102	2.60	P68871	102	100.00	No
P68871	VAR_003040	P68871	117	3.30	P68871	117	100.00	No
P68871	VAR_003041	P68871	117	3.30	P68871	117	100.00	No
P68871	VAR_003042	P68871	119	2.23	P68871	119	100.00	No
P68871	VAR_003051	P68871	123	2.16	P68871	123	100.00	No
P68871	VAR_003052	P68871	124	2.47	P68871	124	100.00	No
P68871	VAR_003053	P68871	124	2.47	P68871	124	100.00	No
P68871	VAR_003054	P68871	124	2.47	P68871	124	100.00	No
P68871	VAR_003058	P68871	127	2.16	P68871	127	100.00	No
P68871	VAR_003059	P68871	127	2.16	P68871	127	100.00	No
P68871	VAR_003060	P68871	128	2.12	P68871	128	100.00	No
P68871	VAR_003069	P68871	132	2.60	P68871	132	100.00	No
P68871	VAR_003070	P68871	132	2.60	P68871	132	100.00	No
P68871	VAR_010144	P68871	114	2.56	P02089	114	79.85	No
P68871	VAR_010145	P68871	114	2.56	P02089	114	79.85	No
P68871	VAR_025399	P68871	117	3.30	P68871	117	100.00	No
P69891	VAR_003141	P69891	36	3.98	P02070	35	74.63	No
P69891	VAR_003141	P69891	36	3.98	P68871	36	74.63	No
P69891	VAR_003142	P69891	37	3.09	P02070	36	74.63	No
P69891	VAR_003142	P69891	37	3.09	P68871	37	74.63	No
P69891	VAR_003163	P69891	79	2.25	P02089	79	70.15	No
P69891	VAR_003168	P69891	97	3.47	P02070	96	74.63	No
P69891	VAR_003168	P69891	97	3.47	P68871	97	74.63	No
P69891	VAR_003175	P69891	128	2.12	P69891	128	100.00	No
P69892	VAR_003131	P69892	15	5.78	P02118	15	73.13	Yes
P69892	VAR_003139	P69892	26	2.10	P68871	26	75.37	No
P69892	VAR_003156	P69892	66	2.40	P02089	66	70.90	No
P69892	VAR_003157	P69892	66	2.40	P02089	66	70.90	No
P69892	VAR_003162	P69892	77	3.17	P02089	77	70.90	No
P69892	VAR_003166	P69892	92	5.44	P02089	92	70.90	No
P69892	VAR_003171	P69892	117	3.30	P68871	117	75.37	No
P69892	VAR_003174	P69892	125	2.06	P69891	125	99.25	No
P69892	VAR_020646	P69892	17	2.30	P68871	17	75.37	No
P69892	VAR_020647	P69892	19	2.34	P02118	19	73.13	Yes
P69892	VAR_020651	P69892	75	2.38	P02089	75	70.90	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P69905	VAR_002729	P69905	11	2.28	P02118	12	41.41	Yes
P69905	VAR_002731	P69905	14	5.78	P02118	15	41.41	Yes
P69905	VAR_002733	P69905	16	2.30	P01990	16	67.69	Yes
P69905	VAR_002734	P69905	16	2.30	P01990	16	67.69	Yes
P69905	VAR_002739	P69905	20	2.28	P02118	19	41.41	Yes
P69905	VAR_002740	P69905	20	2.28	P02118	19	41.41	Yes
P69905	VAR_002748	P69905	27	2.10	P01958	27	87.69	No
P69905	VAR_002749	P69905	27	2.10	P01958	27	87.69	No
P69905	VAR_002750	P69905	27	2.10	P01958	27	87.69	No
P69905	VAR_002752	P69905	31	2.86	P69905	31	100.00	No
P69905	VAR_002754	P69905	37	3.98	P01965	37	83.85	No
P69905	VAR_002756	P69905	40	2.44	P02074	38	46.09	No
P69905	VAR_002756	P69905	40	2.44	P02070	38	46.09	No
P69905	VAR_002759	P69905	44	2.17	P69905	44	100.00	No
P69905	VAR_002760	P69905	44	2.17	P69905	44	100.00	No
P69905	VAR_002761	P69905	45	3.01	P02208	54	30.77	Yes
P69905	VAR_002762	P69905	45	3.01	P02208	54	30.77	Yes
P69905	VAR_002763	P69905	47	2.60	P02208	57	30.77	Yes
P69905	VAR_002764	P69905	47	2.60	P02208	57	30.77	Yes
P69905	VAR_002765	P69905	47	2.60	P02208	57	30.77	Yes
P69905	VAR_002766	P69905	47	2.60	P02208	57	30.77	Yes
P69905	VAR_002774	P69905	56	2.63	P02208	71	30.77	No
P69905	VAR_002775	P69905	56	2.63	P02208	71	30.77	No
P69905	VAR_002779	P69905	59	2.52	P02089	64	42.19	No
P69905	VAR_002782	P69905	61	2.40	P02089	66	42.19	No
P69905	VAR_002783	P69905	61	2.40	P02089	66	42.19	No
P69905	VAR_002784	P69905	62	3.12	P02089	67	42.19	No
P69905	VAR_002790	P69905	72	3.17	P02089	77	42.19	No
P69905	VAR_002791	P69905	74	2.25	P02089	79	42.19	No
P69905	VAR_002792	P69905	74	2.25	P02089	79	42.19	No
P69905	VAR_002793	P69905	74	2.25	P02089	79	42.19	No
P69905	VAR_002801	P69905	80	2.20	P02089	85	42.19	No
P69905	VAR_002808	P69905	87	5.44	P02089	92	42.19	No
P69905	VAR_002809	P69905	87	5.44	P02089	92	42.19	No
P69905	VAR_002814	P69905	94	2.45	P69905	94	100.00	No
P69905	VAR_002815	P69905	95	3.08	P69905	95	100.00	No
P69905	VAR_002816	P69905	95	3.08	P69905	95	100.00	No
P69905	VAR_002817	P69905	97	2.60	P69905	97	100.00	No
P69905	VAR_002821	P69905	109	2.56	P02089	114	42.19	No
P69905	VAR_002823	P69905	112	3.30	P01958	112	87.69	No
P69905	VAR_002823	P69905	112	3.30	P01966	112	87.69	No
P69905	VAR_002825	P69905	114	2.07	P69905	114	100.00	No
P69905	VAR_002826	P69905	114	2.07	P69905	114	100.00	No
P69905	VAR_002827	P69905	114	2.07	P69905	114	100.00	No
P69905	VAR_002835	P69905	122	2.46	P69905	122	100.00	No
P69905	VAR_002837	P69905	126	2.03	P69905	126	100.00	No
P69905	VAR_002838	P69905	126	2.03	P69905	126	100.00	No
P69905	VAR_002839	P69905	127	2.60	P01958	127	87.69	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
P69905	VAR_025002	P69905	31	2.86	P69905	31	100.00	No
P69905	VAR_025389	P69905	94	2.45	P69905	94	100.00	No
P69905	VAR_025392	P69905	126	2.03	P69905	126	100.00	No
P78363	VAR_008430	P78363	965	3.62	Q9YGA6	38	32.72	No
P78363	VAR_008431	P78363	978	3.11	Q9YGA6	51	32.72	No
P78363	VAR_008436	P78363	1087	3.85	Q9YGA6	165	32.72	Yes
P78363	VAR_012547	P78363	971	3.73	Q9YGA6	44	32.72	No
P78363	VAR_012558	P78363	1063	3.70	Q9YGA6	141	32.72	Yes
P78363	VAR_012559	P78363	1087	3.85	Q9YGA6	165	32.72	Yes
P78385	VAR_023052	P78385	407	3.97	P08670	395	37.66	No
P78504	VAR_013203	P78504	386	2.12	P00740	105	41.94	No
P80365	VAR_015639	P80365	227	2.70	P19992	147	30.52	No
P80365	VAR_015640	P80365	237	2.81	P19992	157	30.52	No
P80365	VAR_015642	P80365	250	2.52	P19992	170	30.52	No
P80404	VAR_008883	P80404	220	3.11	P80147	220	95.95	No
P82279	VAR_011642	P82279	250	5.88	P08709	132	48.39	No
P82279	VAR_022943	P82279	195	5.88	P09871	143	41.38	No
P82279	VAR_022946	P82279	383	5.88	P08709	130	41.94	No
P82279	VAR_022954	P82279	681	5.88	P09871	143	31.03	No
P82279	VAR_022966	P82279	894	2.65	P00740	100	46.67	No
P82279	VAR_022977	P82279	1205	3.29	P09871	161	34.48	Yes
P82279	VAR_022980	P82279	1321	5.88	P08709	130	51.61	Yes
P98172	VAR_023131	P98172	111	2.41	P52800	114	61.15	No
P98172	VAR_023132	P98172	115	4.26	P52800	118	61.15	No
P98172	VAR_023133	P98172	119	4.10	P52800	122	61.15	No
P98172	VAR_023134	P98172	119	4.10	P52800	122	61.15	No
P98172	VAR_023135	P98172	119	4.10	P52800	122	61.15	No
Q00266	VAR_006935	Q00266	55	2.81	P13444	56	97.98	No
Q00266	VAR_006937	Q00266	264	3.43	P13444	265	97.08	No
Q00266	VAR_006939	Q00266	322	3.04	P13444	323	97.08	No
Q01955	VAR_011212	Q01955	1207	3.86	P02452	139	36.84	Yes
Q01955	VAR_011217	Q01955	1334	3.86	P02452	148	43.86	Yes
Q01955	VAR_011219	Q01955	1661	4.20	Q7SIB2	221	49.38	Yes
Q01974	VAR_010771	Q01974	620	4.44	Q06187	525	37.75	Yes
Q02388	VAR_001825	Q02388	2073	3.86	P02452	145	40.35	Yes
Q02388	VAR_001826	Q02388	2076	3.86	P02452	148	40.35	Yes
Q02388	VAR_001827	Q02388	2079	3.86	P02452	151	40.35	Yes
Q02388	VAR_001830	Q02388	2569	3.86	P02452	148	43.86	Yes
Q02388	VAR_001832	Q02388	2623	3.86	P02452	142	33.33	Yes
Q02388	VAR_001836	Q02388	2749	3.86	P02452	139	42.86	Yes
Q02388	VAR_011169	Q02388	1812	3.86	P02452	139	47.37	Yes
Q02388	VAR_011184	Q02388	2064	3.86	P02452	136	40.35	Yes
Q02388	VAR_011185	Q02388	2079	3.86	P02452	151	40.35	Yes
Q02388	VAR_011188	Q02388	2207	3.86	P02452	139	42.86	Yes
Q02388	VAR_011190	Q02388	2263	3.86	P02452	136	42.11	Yes
Q02388	VAR_011194	Q02388	2366	3.86	P02452	142	42.86	Yes
Q02388	VAR_011195	Q02388	2369	3.86	P02452	145	42.86	Yes
Q02388	VAR_015520	Q02388	1815	3.86	P02452	142	47.37	Yes

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
Q03692	VAR_001844	Q03692	598	3.71	Q00780	661	61.29	No
Q06124	VAR_015601	Q06124	42	2.34	P35235	42	100.00	No
Q06124	VAR_015613	Q06124	139	2.38	O89100	84	42.47	No
Q06187	VAR_006220	Q06187	27	3.31	Q06187	27	100.00	No
Q06187	VAR_006221	Q06187	27	3.31	Q06187	27	100.00	No
Q06187	VAR_006227	Q06187	287	3.80	O60880	13	30.14	No
Q06187	VAR_006231	Q06187	306	4.27	P35235	32	32.43	No
Q06187	VAR_006232	Q06187	333	3.75	P27986	670	33.80	No
Q06187	VAR_006239	Q06187	407	2.92	Q06187	407	100.00	Yes
Q06187	VAR_006249	Q06187	508	5.05	P54763	742	40.16	Yes
Q06187	VAR_006251	Q06187	519	4.10	P08069	1134	34.94	Yes
Q06187	VAR_006254	Q06187	524	4.05	Q07912	256	41.30	Yes
Q06187	VAR_006255	Q06187	524	4.05	Q07912	256	41.30	Yes
Q06187	VAR_006256	Q06187	525	4.44	Q06187	525	100.00	Yes
Q06187	VAR_006267	Q06187	591	2.29	Q07912	325	41.30	Yes
Q06187	VAR_006268	Q06187	593	3.57	Q07912	327	41.30	Yes
Q06187	VAR_006269	Q06187	593	3.57	Q07912	327	41.30	Yes
Q06187	VAR_006270	Q06187	597	4.47	Q07912	331	41.30	Yes
Q06187	VAR_006272	Q06187	612	3.26	P00520	455	48.19	No
Q06187	VAR_006272	Q06187	612	3.26	P00519	455	48.19	Yes
Q06187	VAR_006273	Q06187	618	4.13	P08631	478	41.53	Yes
Q06187	VAR_006276	Q06187	632	5.76	P08631	492	41.53	Yes
Q06187	VAR_006277	Q06187	640	4.27	P08631	500	41.53	Yes
Q06187	VAR_006278	Q06187	640	4.27	P08631	500	41.53	Yes
Q06187	VAR_006280	Q06187	646	2.31	P08631	506	41.53	Yes
Q06187	VAR_008293	Q06187	27	3.31	Q06187	27	100.00	No
Q06187	VAR_008305	Q06187	287	3.80	O60880	13	30.14	No
Q06187	VAR_008307	Q06187	306	4.27	P35235	32	32.43	No
Q06187	VAR_008319	Q06187	508	5.05	P54763	742	40.16	Yes
Q06187	VAR_008323	Q06187	524	4.05	Q07912	256	41.30	Yes
Q06187	VAR_008326	Q06187	562	6.26	Q06187	562	100.00	Yes
Q06187	VAR_008330	Q06187	618	4.13	P08631	478	41.53	Yes
Q06187	VAR_008331	Q06187	618	4.13	P08631	478	41.53	Yes
Q07001	VAR_021211	Q07001	271	3.54	P02711	260	34.31	No
Q09428	VAR_000100	Q09428	715	3.84	P68187	39	33.72	No
Q09428	VAR_008540	Q09428	1492	3.35	Q9CHL8	499	35.75	Yes
Q09428	VAR_015009	Q09428	1505	3.85	Q9KQW9	506	33.52	Yes
Q13253	VAR_011361	Q13253	35	4.16	Q13253	35	100.00	No
Q13253	VAR_018324	Q13253	35	4.16	Q13253	35	100.00	No
Q13402	VAR_009328	Q13402	503	3.77	P13538	529	41.12	No
Q13402	VAR_024047	Q13402	519	2.04	P13538	545	41.12	No
Q13402	VAR_024048	Q13402	756	4.13	P10587	803	40.00	No
Q13485	VAR_011380	Q13485	493	2.48	Q13485	493	100.00	No
Q13485	VAR_019571	Q13485	352	3.51	Q13485	352	100.00	No
Q13608	VAR_007918	Q13608	812	3.51	Q01853	585	52.46	No
Q13608	VAR_007919	Q13608	812	3.51	Q01853	585	52.46	No
Q13950	VAR_012132	Q13950	113	2.98	Q01196	62	91.04	No
Q13950	VAR_012133	Q13950	118	2.83	Q01196	67	91.04	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
Q13950	VAR_012137	Q13950	169	3.35	Q01196	118	91.04	No
Q13950	VAR_012142	Q13950	193	3.89	Q01196	142	91.04	Yes
Q13950	VAR_012145	Q13950	200	3.21	Q01196	149	91.04	No
Q13950	VAR_012146	Q13950	205	3.07	Q01196	154	91.04	No
Q13950	VAR_012147	Q13950	209	2.48	Q01196	158	91.04	No
Q14533	VAR_018116	Q14533	402	3.97	P08670	395	37.66	No
Q15672	VAR_004496	Q15672	131	3.34	P01106	377	44.90	No
Q15672	VAR_015219	Q15672	156	3.54	P01106	403	44.90	No
Q16667	VAR_013849	Q16667	187	3.46	Q16667	187	100.00	No
Q16836	VAR_024081	Q16836	258	3.56	Q16836	258	100.00	No
Q5IJ48	VAR_022986	Q5IJ48	116	2.77	P01135	54	43.33	No
Q6XZB0	VAR_023760	Q6XZB0	55	3.82	P29183	62	35.37	No
Q8NBP7	VAR_025453	Q8NBP7	253	3.05	P04072	104	33.04	No
Q92838	VAR_011080	Q92838	332	3.13	Q92838	332	100.00	No
Q92838	VAR_013487	Q92838	302	3.03	Q92838	302	100.00	No
Q92887	VAR_000099	Q92887	768	3.65	Q58206	153	31.65	Yes
Q92887	VAR_010756	Q92887	1382	4.20	Q9CHL8	430	37.99	Yes
Q92947	VAR_000394	Q92947	309	2.30	Q06319	262	31.03	No
Q92947	VAR_000396	Q92947	333	2.79	Q06319	286	31.03	No
Q92947	VAR_000408	Q92947	390	3.77	P15651	368	31.51	No
Q92947	VAR_000409	Q92947	390	3.77	P15651	368	31.51	No
Q92968	VAR_009306	Q92968	326	2.44	P08631	127	32.08	Yes
Q99456	VAR_008528	Q99456	429	4.94	P08670	399	33.11	No
Q99497	VAR_020496	Q99497	149	3.28	Q99497	149	100.00	Yes
Q99574	VAR_008520	Q99574	49	3.58	O35684	49	86.93	No
Q99574	VAR_008521	Q99574	52	3.19	O35684	52	86.93	No
Q99684	VAR_016213	Q99684	403	2.74	P03001	166	31.82	No
Q99697	VAR_003765	Q99697	115	3.65	P06601	243	62.50	No
Q99697	VAR_003766	Q99697	137	4.16	P40424	288	33.93	No
Q99758	VAR_023498	Q99758	568	3.62	P68187	38	33.33	No
Q9GZU5	VAR_013876	Q9GZU5	264	3.75	P41391	132	38.10	No
Q9H3D4	VAR_020870	Q9H3D4	243	3.89	P02340	172	57.22	Yes
Q9H3D4	VAR_020871	Q9H3D4	243	3.89	P02340	172	57.22	Yes
Q9H3D4	VAR_020873	Q9H3D4	318	3.89	P04637	248	56.19	No
Q9H3D4	VAR_020874	Q9H3D4	319	3.89	P04637	249	56.19	Yes
Q9HCC0	VAR_012792	Q9HCC0	99	3.09	Q8GBW6	61	35.11	No
Q9HCC0	VAR_012793	Q9HCC0	155	2.11	Q9X4K7	123	32.40	No
Q9NZR4	VAR_014246	Q9NZR4	166	3.12	P06601	215	62.50	No
Q9UBP0	VAR_010198	Q9UBP0	499	3.83	Q01853	637	40.66	No
Q9UBX0	VAR_010225	Q9UBX0	160	4.16	P40424	288	33.93	No
Q9UBX5	VAR_017153	Q9UBX5	227	2.97	Q9JJS8	162	41.94	No
Q9ULV5	VAR_017558	Q9ULV5	20	2.65	P22121	196	36.98	Yes
Q9UM47	VAR_012878	Q9UM47	146	5.88	P08709	132	38.71	No
Q9UM47	VAR_012886	Q9UM47	222	5.88	P08709	130	58.06	No
Q9UM47	VAR_012887	Q9UM47	224	5.88	P08709	132	58.06	No
Q9UM47	VAR_012900	Q9UM47	1261	5.88	P00740	108	38.71	Yes
Q9Y458	VAR_021832	Q9Y458	183	3.80	O15119	191	50.55	Yes
Q9Y5X4	VAR_010025	Q9Y5X4	97	4.26	P03372	234	44.00	No

Mut acc	Variant	Prot Acc	Resid	Cons	Templ acc	Templ resid	% id	Cryst Cont?
Q9Y6D9	VAR_019714	Q9Y6D9	516	2.08	Q9Y6D9	516	100.00	No

Appendix H

Table H.1: List of diseases and dosage sensitive genes compiled by the Baylor College of Medicine Medical Genetics Laboratory.

Disease description	**Gene**
1q41q42 deletion	DISP1
van der Woude syndrome	IRF6
Short stature, pituitary and cerebellar defects, and small sella turcica	LHX4
Pituitary anomalies with holoprosencephaly-like features	GLI2
Synpolydactyly/Syndactyly II//Split hand foot malformation 5 (SHFM 5)	HOXD13
Feingold	MYCN
nephronophthisis	NPHP1
SATB2, cleft palate	SATB2
Severe myoclonic epilepsy of infancy (SMEI) or Dravet syndrome	SCN1A
Holoprosencephaly 2, SIX3	SIX3
ASHG 2006	SUMO1
Mowat-Wilson	ZEB2
Noonan	SOS1
Heterotaxy 2	CFC1
Hypertension with CHD	BMPR2
Blepharophimosis	FOXL2
Waardenburg syndrome type II (WS2A)	MITF
3q29 microdeletion	PAK2
microphthalmia	SOX2
forebrain defects, left-right laterality defects	TDGF1
	TGFBR2
TP73L, split food/split hand 4	TP63
Dandy-Walker syndrome	ZIC1, ZIC4
Noonan	RAF1
Rieger	PITX2
alfa synuclein	SCNA
Cornelia de Lange	NIPBL
microcephaly, CHD	NKX2-5
microcephaly, CHD	NPM1
Sotos	NSD1

Disease description	Gene
Treacher Collins syndrome	TCOF1
ADLD adult onset aut. dom. leukodystrophy	LMNB1
	EGR2
Chronic pancreatitis	SPINK1
Congenital 21-alpha hydroxylase deficiency	CYP21A2
Cleidocranial dysplasia	RUNX2
Prader-Willi-like phenotype	SIM1
VEGF	VEGF
Transient neonatal diabetes loci on 6q24 (OMIM 601410)	ZAC
Iridogoniodysgenesis anomaly, Axenfeld-Rieger syndrome	FKHL7 (FOXC1)
COL1A2	COL1A2
Williams	ELN
speech delay	FOXP2
Greig	GLI3
Williams	LIMK1
Split hand/foot	SHFM1
Holoprosencephaly 3, SHH	SHH
Saethre Chotzen	TWIST1
Hereditary pancreatitis	PRSS1
Schizophrenia & epilepsy	CNTNAP2
CHARGE	CHD7
Langer-Giedion	EXT1
Branchiootorenal (BOR)/Melnick-Fraser/Oto-facio-cervical)OFC)	EYA1
Congenital heart disease	GATA4
Bipolar disorder	IMPA1
Langer Giedion	TRPS1
Tetralogy of Fallot	ZFPM2/FOG2
9q34 microdeletion	EHMT1
GPR51, overgrowth	GABBR2
Nail-Patella	LMX1B
9q34 microdeletion	NOTCH1
Gorlin syndrome/Holoprosencephaly 7	PTCH1
Robinow/brachydactyly 1 Olivieri et al	ROR2
Sex reversal - Steroidogenic factor SF-1	SF-1
Loeys–Dietz syndrome	TGFBR1
Tuberous sclerosis	TSC1
Split food split hand 3	FBXW4
hypoparathyroidism, sensorineural deafness, and renal disease, HDR	GATA3
GRID1, 10q22q23 deletion	GRID1
Nebulette	NEBL
NRG3, 10q22q23 deletion	NRG3
PTEN, Cowden syndrome, Bannayan-Zonana syndrome	PTEN
Hirschsprung	RET
Potocki-Shaffer	ALX4
behavioral problems and autistic spectrum disorder. (OMIM 114130)	CALC1
behavioral problems and autistic spectrum disorder. (OMIM 114130)	CALC2
Potocki-Shaffer	EXT2
Beckwith-Wiedeman	H19

Disease description	Gene
Beckwith-Wiedeman	IGF2
Beckwith-Wiedeman	KCNQ1
Mitochondrial complex 1 deficiency	NDUFV1
Beckwith-Wiedeman	p57 (CDKN1C)
WAGR, Aniridia, PAX6	PAX6
Craniosynostosis	SOX6
WAGR, Wilms tumor, WT1	WT1
Stickler syndrome	COL2A1
Osteopoikilosis, short stature and MR	HMGA2
Osteopoikilosis, short stature and MR	LEMD3
Microduplication, Ruiter et al 2007	NOS1
Noonan	PTPN11
Microduplication, Ruiter et al 2008	RFC5
Microduplication, Ruiter et al 2006	THRAP2
Timothy	CACNA1C
Holt-Oram	TBX5
ulnar-mammary syndrome	TBX3
GPC5, brachydactyly and other skeletal anomalies	GPC5
GPC6, brachydactyly and other skeletal anomalies	GPC6
Retinoblastoma	RB
Holoprosencephaly 5, ZIC2	ZIC2
Hirschsprung	EDNRB
Anophthalmia, pituitary hypoplasia, and ear anomalies	BMP4
14q11.2 deletion syndrome	CHD8
FOXG1B	FOXG1B
14q11.2 deletion syndrome	SUPT16H
Branchiootic syndrome-3	SIX1
Oculoauriculovertebral spectrum (?)	SIX6
15q13.3 microdeletion	CHRNA7
Marfan	FBN1
Severe IUGR, developmental delay, postnatal growth retardation	IGF1R
NR2F2, Diaphragmatic hearnia	NR2F2
PML	PML
PWS/AS	SNRPN
PWS/AS	UBE3A
Rubinstein-Taybi	CREBBP
Rubinstein-Taybi	DNASE1
alpha thalasemia-MR syndrome	HBA1
alpha thalasemia-MR syndrome	HBA2
Tuberous sclerosis	PKD1
Polycystic kidney disease	TSC2
Townes-Brocks	SALL1
Osteogenesis imperfecta type IV	COL1A1
17q21.31 microdeletion	CRHR1
Cystinosis	CTNS
Miller-Dieker	LIS1
17q21.31 microdeletion	MAPT
NF1	NF1

Disease description	Gene
CMT1A	PMP22
SMS	RAI1
Campomelic dysplasia	SOX9
TCF2, renal cysts and diabetes	TCF2
Miller-Dieker	YWHAE
Dyggve Melchior Clausen	DYM
Holoprosencephaly 4	TGIF1
Pitt-Hopkins	TCF4
BMP2	BMP2
Brachydactyly C	GDF5
Alagille	JAG1
Coloboma	SNAP25
Alzheimer - early onset	APP
SIM2	SIM2
Holoprosencephaly 1	TMEM1
Metachromatic leukodystrophy	ARSA
NF2	NF2
22q13.3 deletion	SHANK3
DGS	TBX1