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.

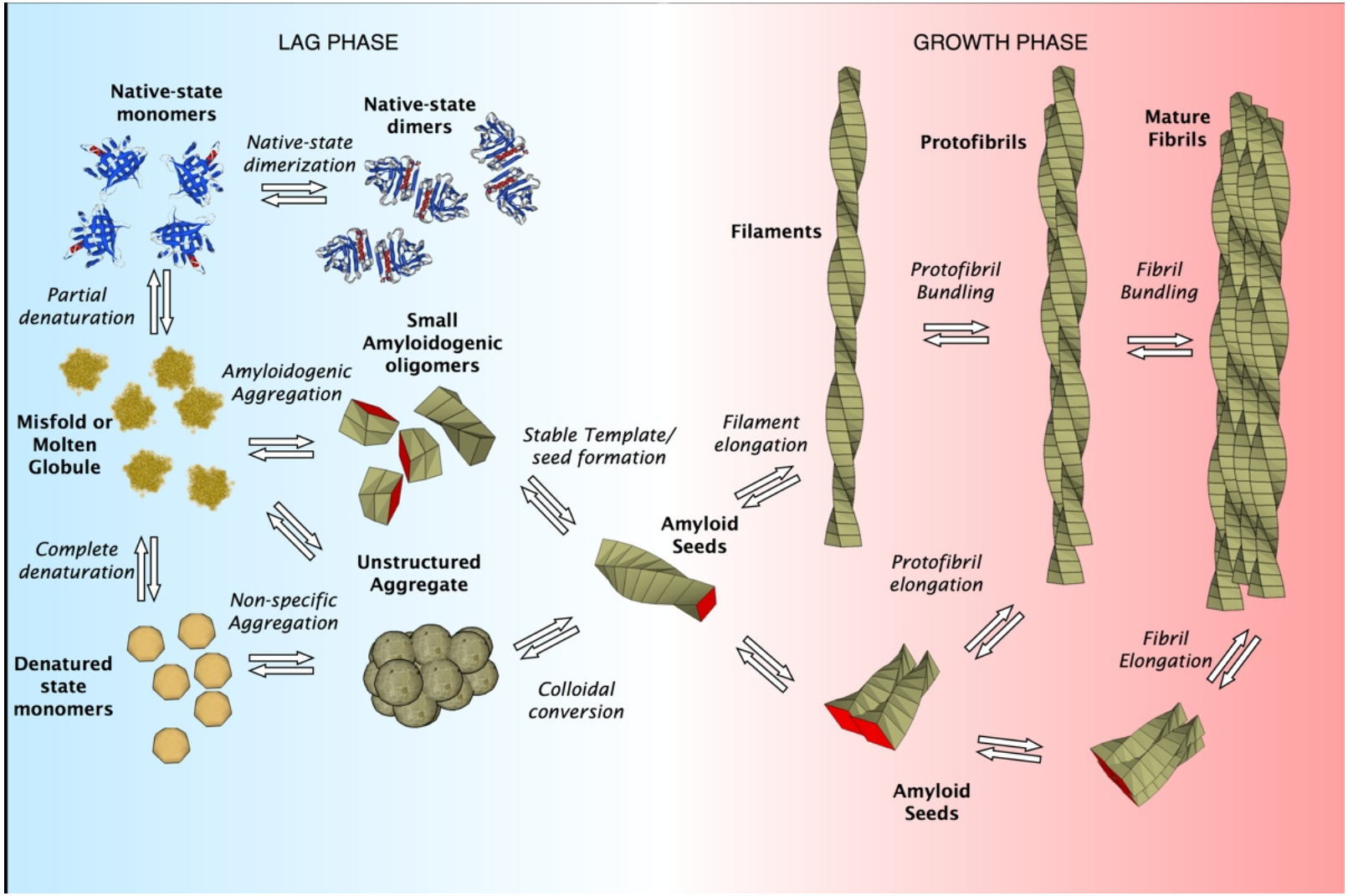

Figure 3.1: Disease pathways of amyloid disorders. In the native state, amyloid precursor proteins maintain an equlibrium between monomeric and dimeric state. Misfolding of a fraction of the protein leads to aggregation of denaturated monomers which subsequently clump into increasingly larger structures 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.

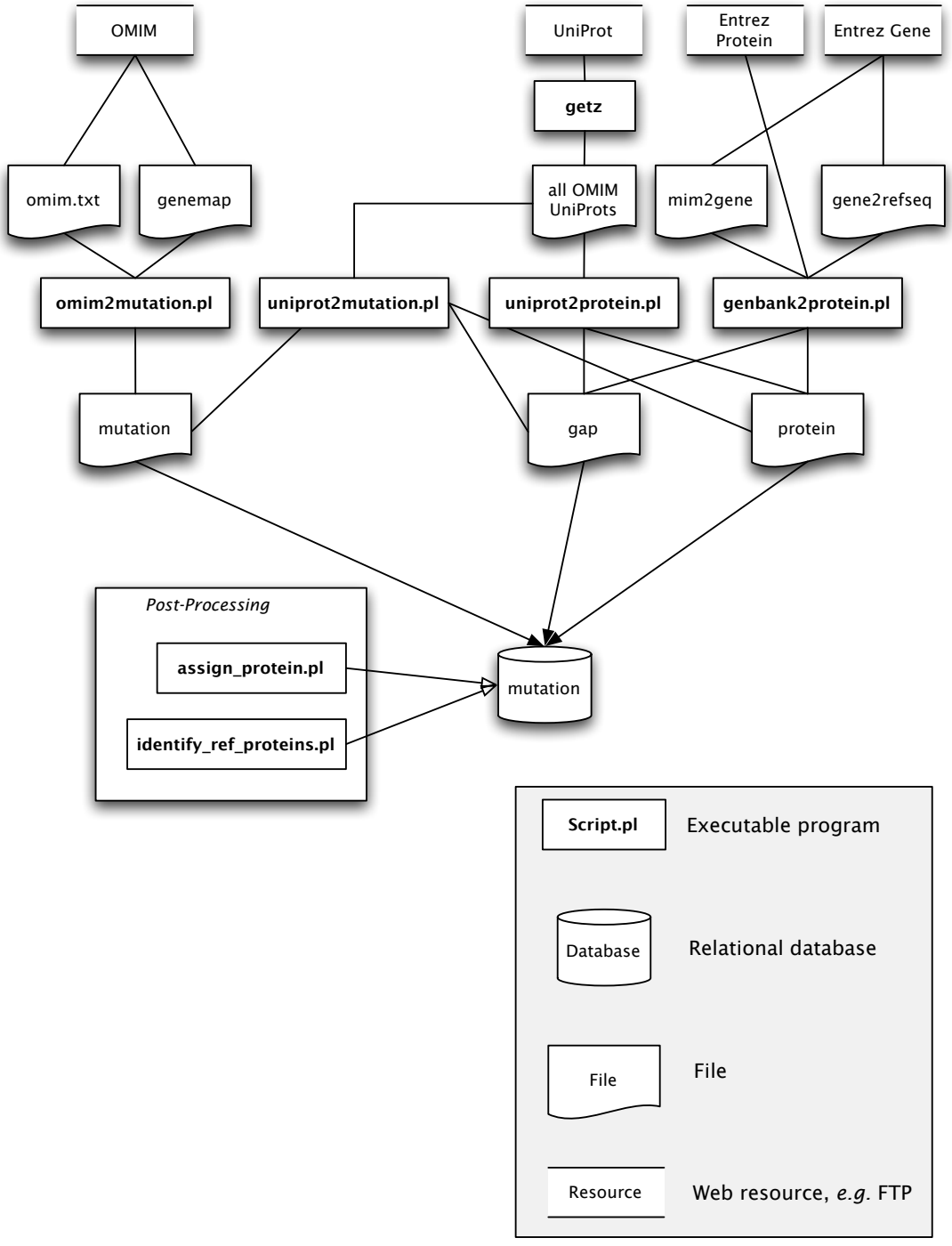

Figure 3.2: Workflow for generation the mutation database from OMIM and UniProt. Several Perl scripts merge and format the data to be imported into a relational database. The post-processing scripts then identify the sequence/post-translational modification combination that best matches the observed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

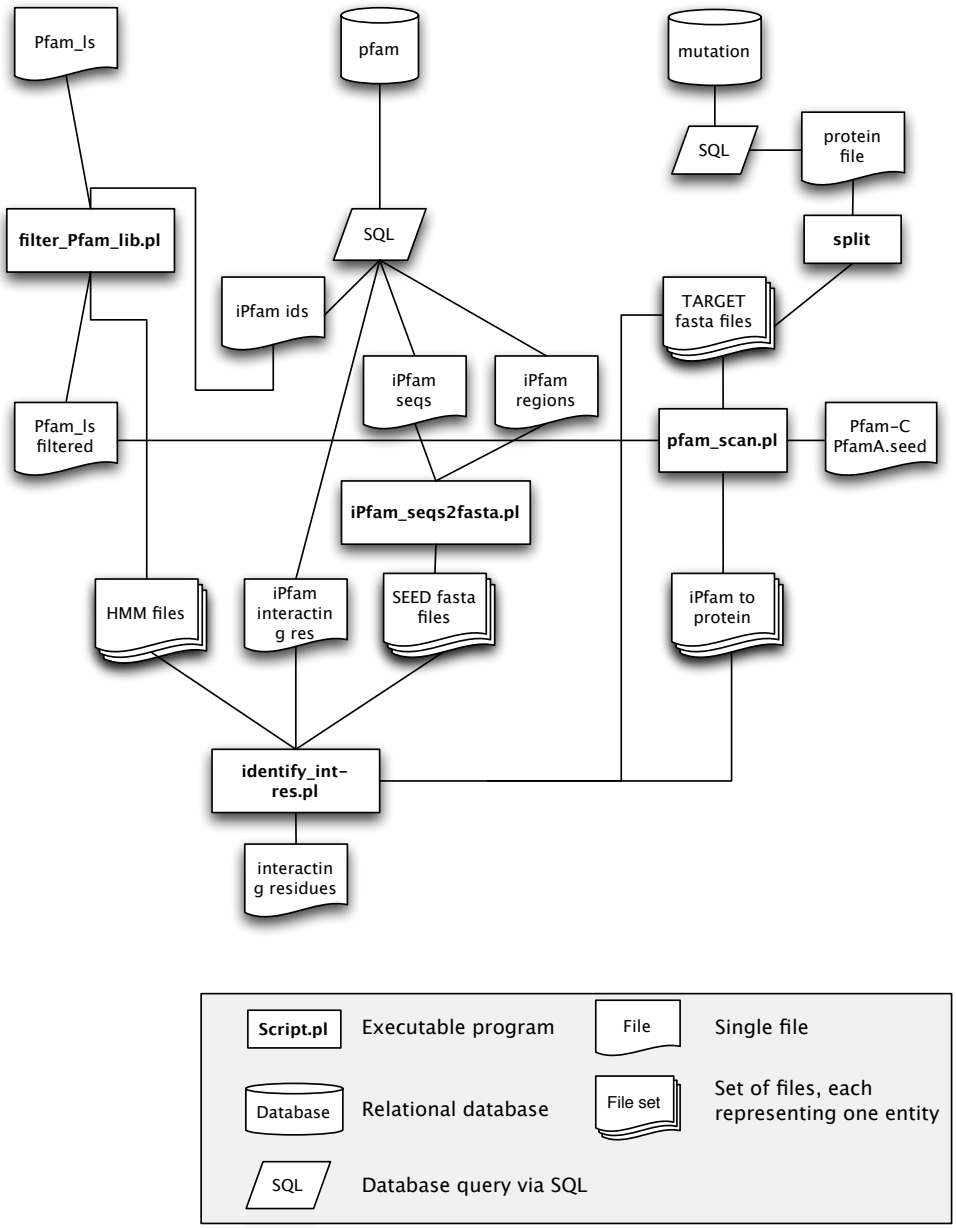

Figure 3.3: Outline of the computational steps leading to the mapping of interacting residues to known disease mutations. The central script is called `identify_int-res.pl` and takes an HMM library file and two sets of fasta files corresponding to domain regions, one containing the structural seeds and 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.

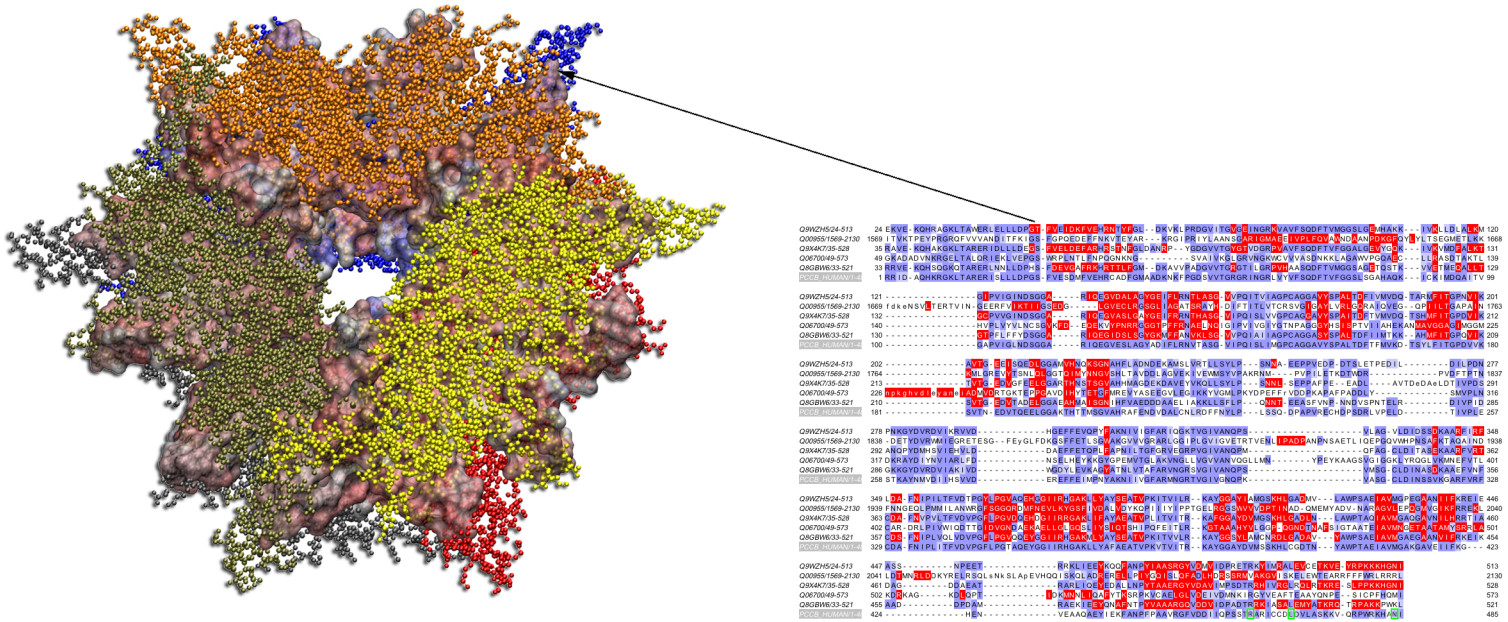

Figure 3.4: Predicting potentially interacting residues from structure. To the left: Structure of Propionyl-CoA carboxylase beta chain with the interaction interface shown as a surface (PDB 1vrg). To the right: Alignment of all sequences matching Pfam domain *Carboxyl_trans* (PF01039) for which a structure 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).

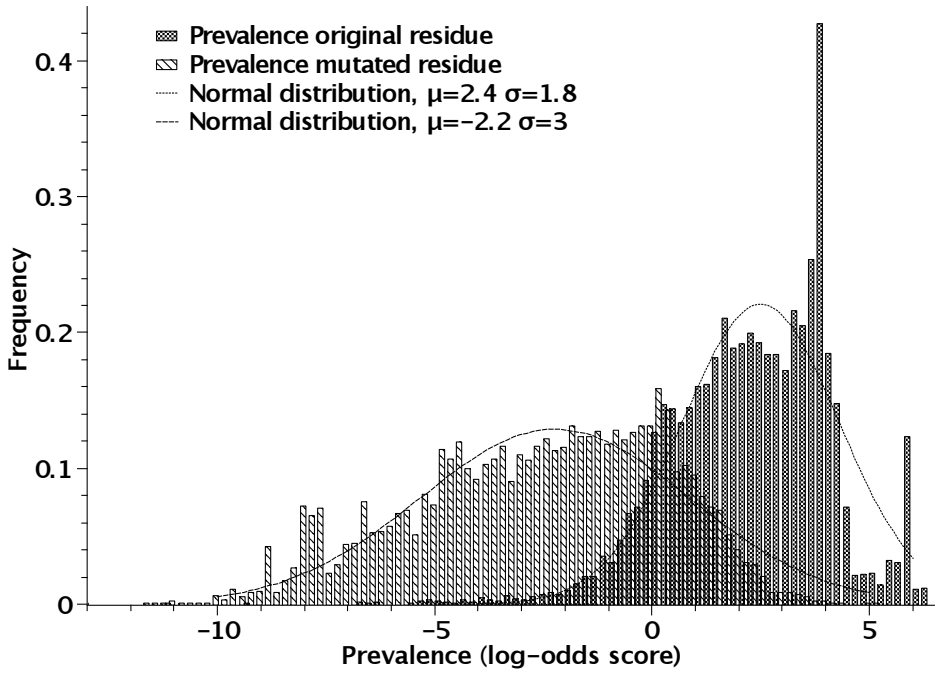

Figure 3.5: Histogram of prevalence of wild-type and mutated residues. The prevalence score distributions of mutated and wild-type residues are clearly separated. They intersect around 0, suggesting that residues whose frequency is similar to the background distribution are as common in mutations as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

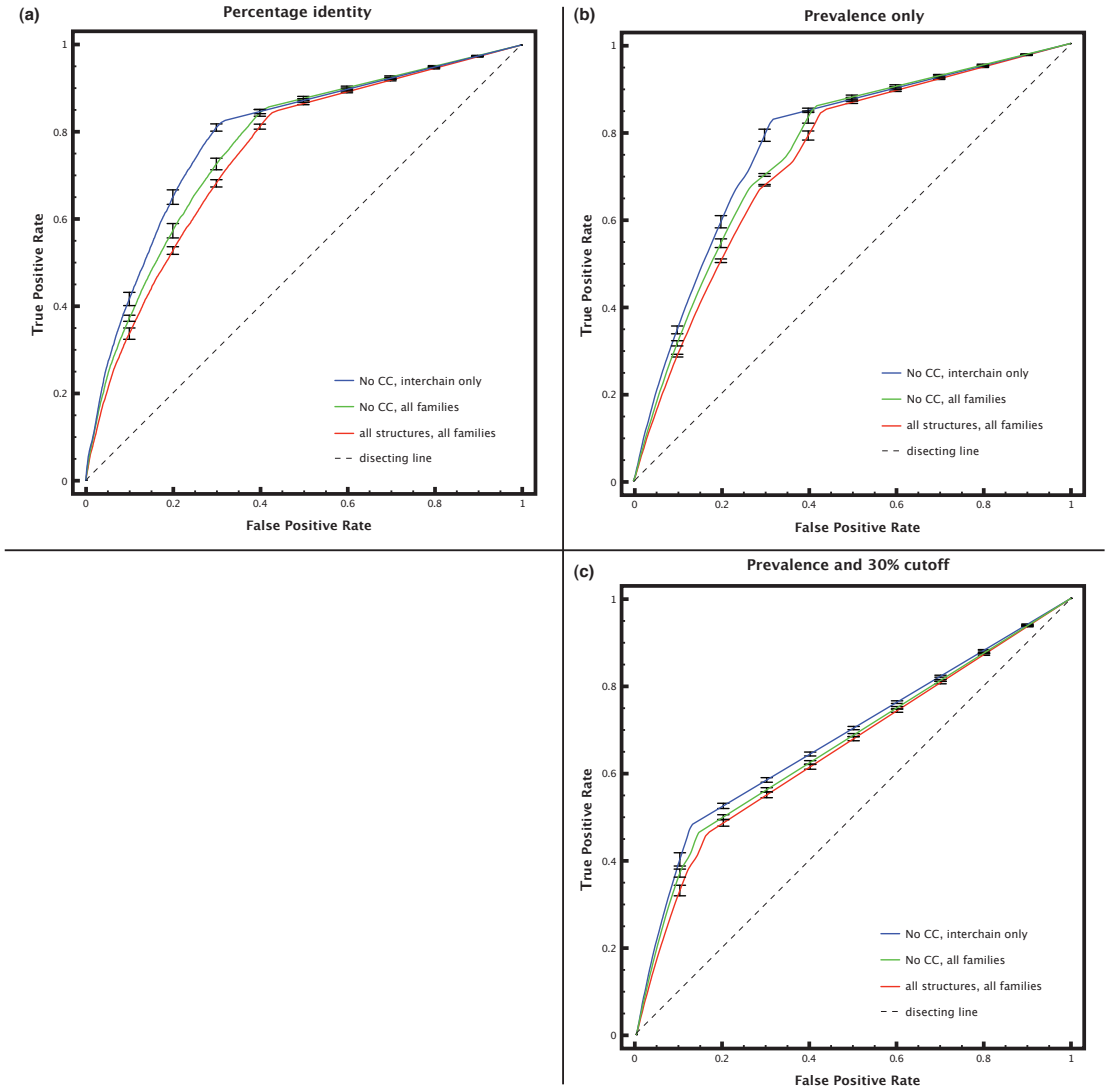

Figure 3.6: Receiver Operator Characteristic (ROC) curves calculated on cross-validation results. Each curve is the combined classification result of all predictions made on the sum of all the individual *i*Pfam families. Bars reflecting standard deviation between repetitions with different training/target 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

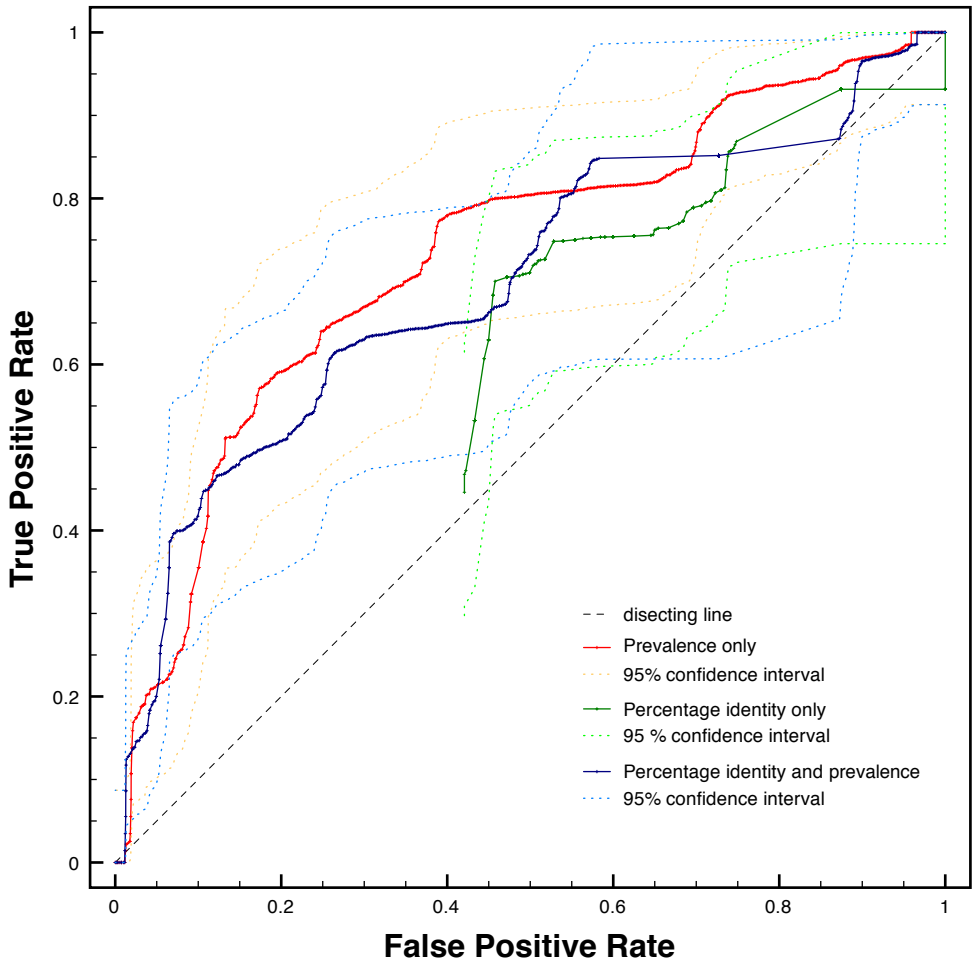

Figure 3.7: Receiver Operator Characteristic (ROC) curves calculated on a set of alanine scanning experiments. The red line represents the performance of my algorithm when changing only the residue prevalence threshold, applying no percentage identity cutoff. The green line shows the performance using 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.

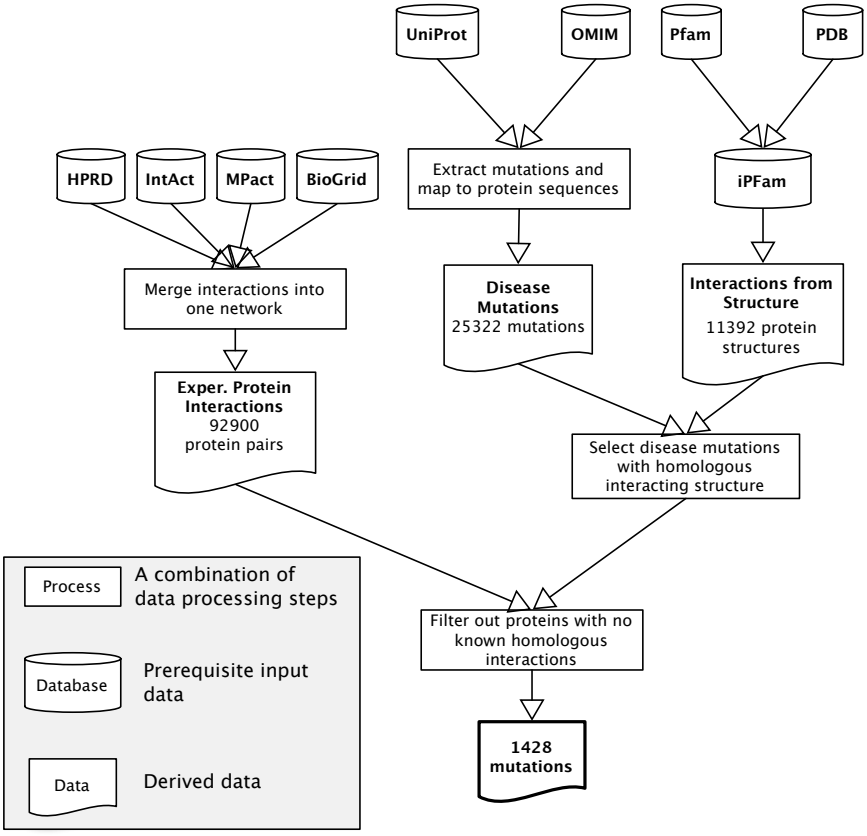

Figure 3.8: Schematic outline of data integration for the prediction of interacting residues. Mutations from OMIM and UniProt for which a residue in a homologous structure is involved in an interaction are selected. This set is restricted further by searching for homologous proteins with known interactions, 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

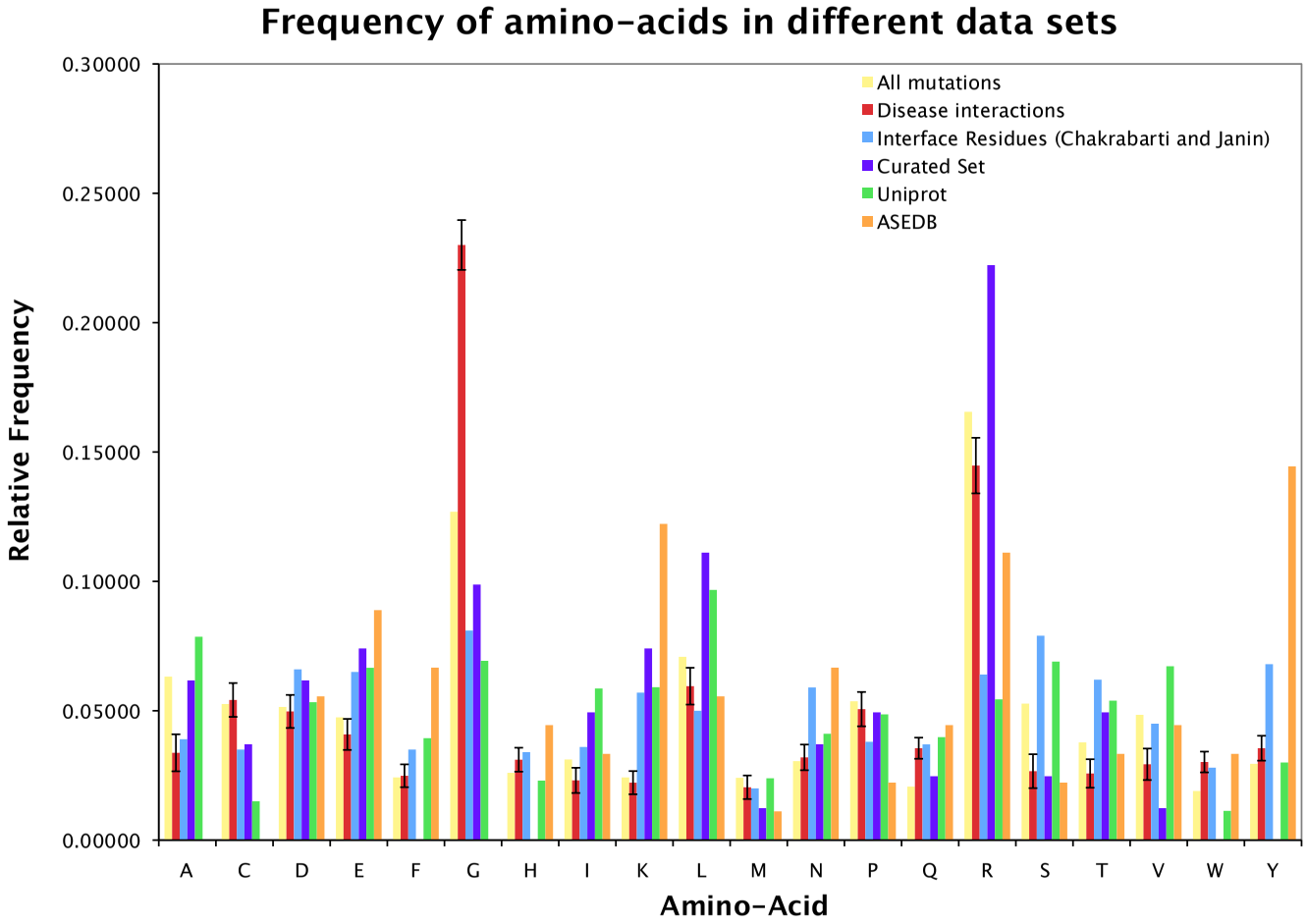

Figure 3.9: Distributions of residue frequencies for all mutations in OMIM and Uniprot (wild type), the predicted set (wild type), the curated set, for interface residues as described by Chakrabarti and Janin (2002), the whole of UniProt and for residues from ASEdb with $\Delta\Delta G > 2$kcal/mol. 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.

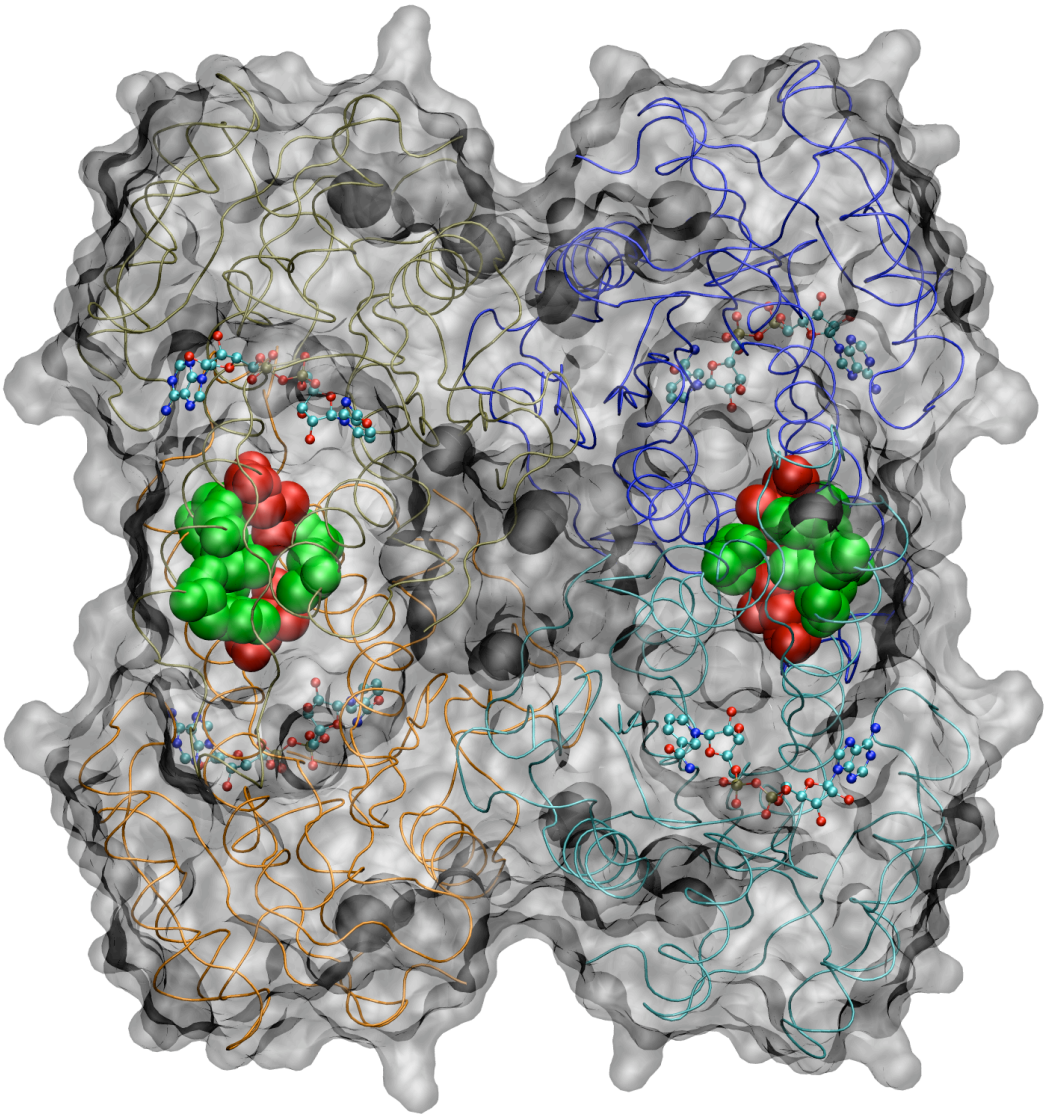

Figure 3.10: Structure of Rat brain 3-hydroxyacyl-CoA dehydrogenase with bound NADH [PDB: 1e3s]. The molecule is composed of 4 monomers, shown as different coloured ribbons. The Leu122 residue is highlighted in red with its binding partners shown in green. As Leu122 also interacts with the Leu122 of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

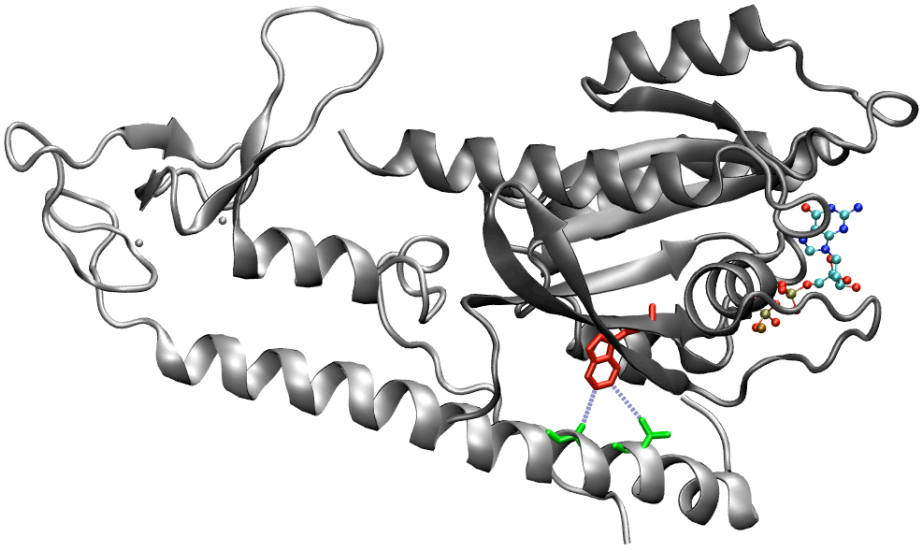

Figure 3.11: The small G protein Rab3A with bound GTP interacting with the effector domain of rabphilin-3A. The residue corresponding to the mutated Trp73 from human RAB27A, is highlighted in red, while the two residues in contact with it are coloured green.

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

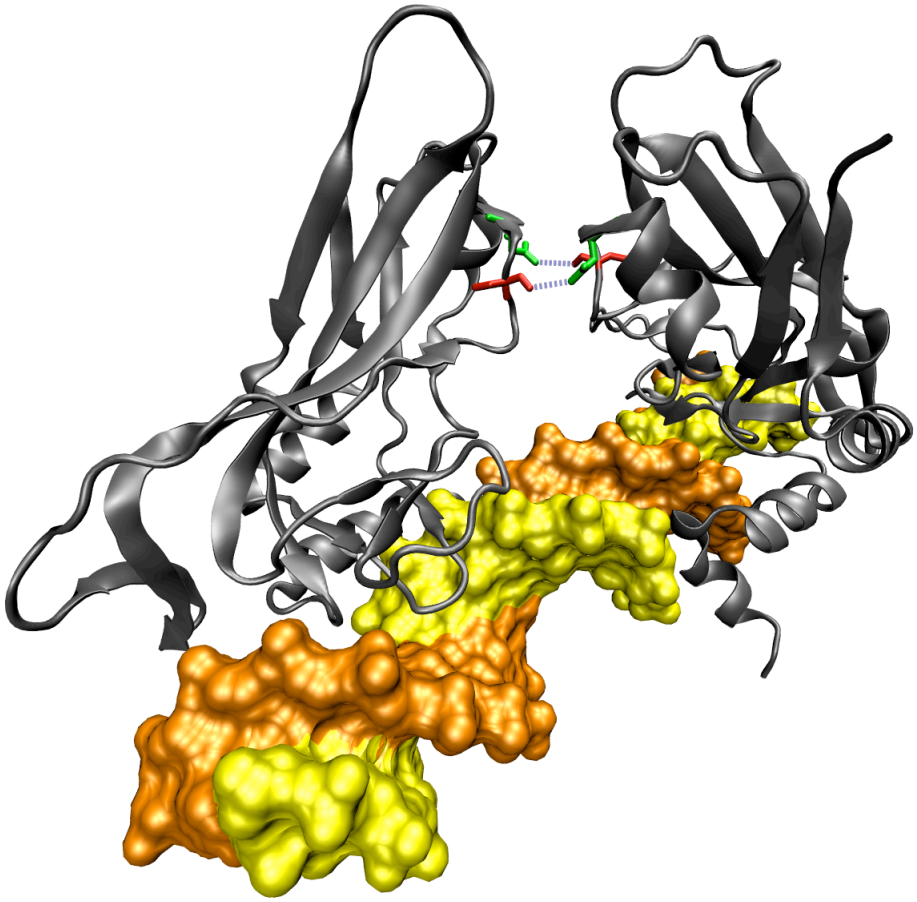

Figure 3.12: The crystal structure of a T-domain from *Xenopus laevis* bound to DNA. The residues highlighted in red are the mutated Ser128, with green residues representing the contact residues in the partner protein. Blue dashed lines show residue contacts.

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.

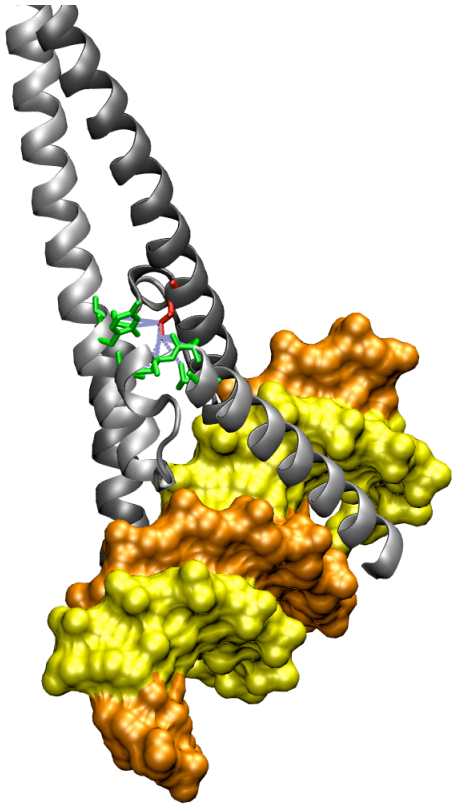

Figure 3.13: Both Myc-c and Max form a basic helix-loop-helix motif. They dimerize mainly through their extended helix II regions. The residue that corresponds to Ile156 in H-Twist is Ile550, shown in red. The residue sits at a key position of the interface, forming bonds with seven residues in Max,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