

Yeast variation effect on proteins

Hoe verandert gist zijn eiwitten?

Promoters:

Prof. Vera van Noort

Prof. Kevin Verstrepen

Department of Microbial and Molecular Systems

Centre of Microbial and Plant Genetics

Dissertation presented in
Fulfillment of the requirements
for the degree of Master of Science in
Bioinformatics

Lynn Verbeke

June 2015

"This dissertation is part of the examination and has not been corrected after defence for eventual errors. Use as a reference is permitted subject to written approval of the promotor stated on the front page."

Yeast variation effect on proteins

Hoe verandert gist zijn eiwitten?

Promoters:

Prof. Vera van Noort

Prof. Kevin Verstrepen

Department of Microbial and Molecular Systems

Centre of Microbial and Plant Genetics

Dissertation presented in
Fulfillment of the requirements
for the degree of Master of Science in
Bioinformatics

Lynn Verbeke

June 2015

Scientific summary

When a nsSNP is located at an interaction site it can stabilize or destabilize PPIs and can lead to a loss, a gain or a change of function [23]. The importance of nsSNPs in PPIs has recently been shown: lots of mutations that cause diseases in humans affect interactions [133]. For this research, the data set used was obtained in the collaborative project NATAR and contains SNPs of yeast in an environment with increasing ethanol levels.

Firstly, all amino acids part of an interaction site are selected by downloading all known complexes of *S. cerevisiae* proteins from the PDB [29]. Using the structural data, all amino acids that are closer than 5Å and belong to a different chain are considered to be interacting [20]. Secondly, gene information for all SNPs is obtained from BioMart [40]. Finally the amino acid number of the SNPs is calculated and compared to the interacting amino acids. For all SNPs that are located at a known interaction site, the mutated amino acid can be determined. Comparison between the original and the mutated residue divides the 33 SNPs into 4 nonsense, 6 silent and 23 non-synonymous mutations.

Apart from some SNPs occurring in vacuole, mitochondrion and cytoskeleton, most of the nsSNPs can be linked either to Central Dogma processes or to protein breakdown and refolding. This can be explained by the fact that ethanol affects both transcription and translation and that high levels of misfolded proteins are observed in stress conditions [72] [131]. This indicates a reaction from the cell to the effects of ethanol.

The nsSNPs are the most interesting as they can affect the PPI. Different algorithms have been developed to predict this effect. Three predictions are compared: two from the prediction tools SNP-IN and PoPMuSiC as they belong to the best performing tools that are publicly available, and a literature-based opinion, for which different sources are considered [24] [25] [29] [34] [49] [50] [51] [53]. Although all methods are relatively successful in the prediction of deleterious mutations, a lot of improvement still has to be made in successfully predicting beneficial mutations [24] [25]. Prediction accuracies cannot be measured as there is no interaction data to compare the prediction results with, but predictions can be compared. Because all structural information of the observed mutations in Mdv1 and Rpn8 gives positive results in both tools, it is highly likely that those interactions will be preserved. Also there is a nsSNP in Qcr7 that has a likelihood of 1 in the SNP-IN tool to be deleterious to the interaction, so this interaction will probably be broken.

This work can be extended in two directions, either in validating the predictions, or in increasing the prediction accuracies. Validation is preferentially done with Y2H to test the direct interaction with the mutated gene and its interacting partner. Increasing the prediction accuracy can be achieved by integrating the results of different methods, different PDB files from the same interaction and evolutionary conservation of the mutated residue into a new classifier.

Glossary

Abbreviation	Full word
δG	free energy
AD	activator domain
alt	alternative
AP-MS	affinity-purification mass spectrometry
ASA	accessible surface area
chr	chromosome
DBD	DNA binding domain
DNA	deoxyribonucleic acid
IDR	intrinsically disordered region
MoRE	molecular recognition element
mRNA	messenger RNA
mRNP	messenger ribonucleoprotein
N/A	not available
NAD+	nicotinamide adenine dinucleotide
NADH	reduced nicotinamide adenine dinucleotide
NATAR	Natural and Artificial Genetic Variation in Microbes
NP	not preserved
nsSNP	non-synonymous single nucleotide polymorphism
P	preserved
PDB	protein data bank
PE	phosphatidylethanolamine
pos	position
PPI	protein-protein interaction
PTM	post-translational modification
ref	reference
RNA	ribonucleic acid

Abbreviation	Full word
ROS	reactive oxygen species
<i>S. cerevisiae</i>	Saccharomyces cerevisiae
SGD	Saccharomyces Genome Databank
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
tRNA	transfer ribonucleic acid
UAS	Upstream Activator Sequence
UV	ultraviolet
Y2H	Yeast 2-hybrid
ZAFFI	Zlab AFFinity enhancement

Contents

Scientific summary	vii
Glossary	ix
1 Introduction	1
1.1 The effect of ethanol on <i>S. cerevisiae</i>	1
1.2 Goal of this project	3
I Literature review	5
2 Protein-protein interactions	7
2.1 From DNA to protein	7
2.2 Interactions between proteins	8
2.3 Structural characteristics of the PPI interface	10
3 The effect of genetic variation on PPIs	15
3.1 Conservation of the PPI interface	15
3.2 nsSNPs in PPIs	16
3.3 Assessment of the effect of nsSNPs on PPIs	17
II Materials and methods	21
4 Overview of methodology	23
5 Extraction of interacting amino acids	25
5.1 Collection of potential complexes	25
5.2 Selection of chains	26
5.3 Extraction of amino acids	27
6 Processing of SNP data	29
6.1 Preparation of SNP data	29
6.2 Obtaining gene information	31
6.3 Connecting the SNP data to the gene information	32

7	Connecting SNPs to structure	35
7.1	Sorting and filtering the data per chromosome	35
7.2	Intron-containing genes	36
7.3	Calculation of the amino acid number	37
7.4	Matching the SNPs to the structural data	38
7.5	Determination of the new amino acid	39
8	Automatization of the process	43
8.1	Obtaining gene information	43
8.2	Calculation of the amino acid number	44
8.3	Connecting SNPs to structure	45
8.4	Updating of the available information	46
III	Results	47
9	Overview of the results	49
10	Nonsense and silent mutations	51
10.1	Nonsense mutations	51
10.1.1	Fermentor 4	51
10.1.2	Fermentor 6	51
10.2	Silent mutations	53
10.2.1	Synonymous mutations in the population of fermentor 6	54
10.2.2	Synonymous mutations in the clones of fermentor 2	54
10.2.3	Synonymous mutations in the clones of fermentor 6	54
11	Non-synonymous mutations	57
11.1	In the population	59
11.2	In the clones at 100 generations	63
11.3	In the clones at 200 generations	70
11.3.1	Fermentor 3	70
11.3.2	Fermentor 6	71
12	Discussion and future work	79
12.1	Affected processes	79
12.1.1	Central Dogma processes	80
12.1.2	Defolded proteins	83
12.2	Affected processes in the different fermentors	83
12.3	The predictions	84
12.4	Future work	85
12.4.1	Validation of the obtained predictions	86
12.4.2	Increasing the prediction accuracy	88
A	Detailed prediction results	99
	Summary in layman's terms	103

Chapter 1

Introduction

Saccharomyces cerevisiae, a type of yeast, is the most studied eukaryotic organism in the world. As an unicellular organism, *S. cerevisiae* possesses certain properties that make it advantageous for research: it is a small organism with a short generation time and it can easily be cultured, thereby decreasing cost and space requirements compared to higher eukaryotes [1]. Next to that, it has the same internal structure as every other eukaryotic organism but a smaller genome: a human genome contains about 3.2 gigabases [2], the *S. cerevisiae* genome on the contrary is only 12 megabases long [3]. Understanding cellular processes in *S. cerevisiae* increases the level of understanding in higher eukaryotes, because knowledge can be transferred due to homology and conservation [4]. Additionally, *S. cerevisiae* is able to reproduce both in a sexual as well as in a non-sexual way, permitting it to study sexual genetics. Finally, *S. cerevisiae* is frequently used in industry, raising the interest in research also on an economic level [1].

S. cerevisiae is able to produce high concentrations of ethanol, and therefore it is often used in both the food industry as well as for production of bio-ethanol, a source of bio-energy [5]. The capacity to produce high ethanol levels is rare because there are only a few other organisms that show some ethanol tolerance [6]. However, too high levels of ethanol have a negative impact on growth and viability, thereby limiting ethanol gain, which makes understanding of the mechanisms behind ethanol tolerance important for improvement of production in a cost-effective manner [5] [7]. A distinction can be made between the reaction on high ethanol levels and ethanol tolerance. The ethanol stress response describes the reaction of a cell on high ethanol levels (*what does the cell do to survive?*), ethanol tolerance can be defined as the survival performance during chronic ethanol exposure (*does the cell manage to survive?*) [5]. Although all read literature emphasizes this distinction, it remains unclear how to properly distinguish between both: what is called an ethanol stress response in one paper, is called part of the tolerance process in another paper. Furthermore, the stress response on ethanol can be divided into a general stress response and a specific ethanol stress response [7]. This distinction is displayed in figure 1.1.

1.1 The effect of ethanol on *S. cerevisiae*

Ethanol seems to target mainly the membrane, both in function as well as in structure. The membrane becomes more fluid and therefore transport along the membrane is facilitated.

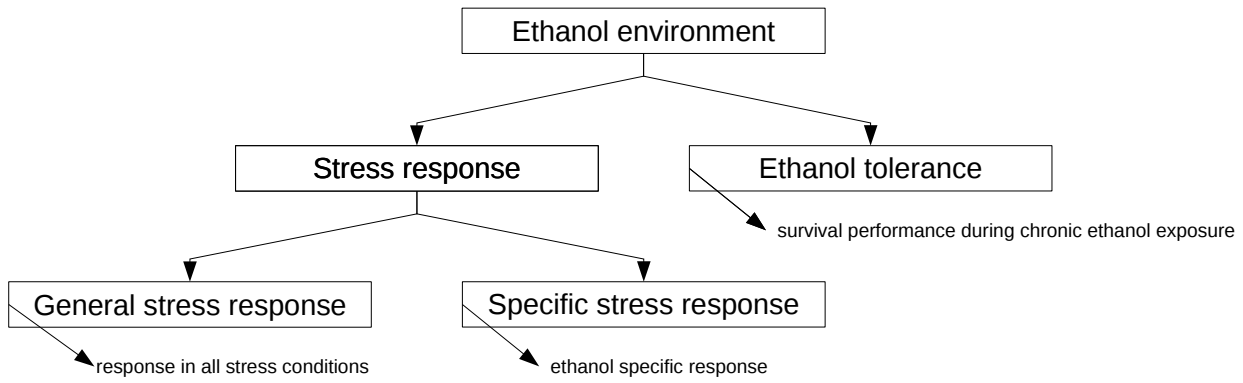


Figure 1.1: Reaction of *S. cerevisiae* in an ethanol environment

Besides the membrane, the vacuole becomes one single, large organelle instead of showing a segregated structure [7]. The vacuole is a single-membrane fluid-filled organelle responsible for storage of small molecules like water and ions and for degradation of waste products and salts [8]. At low ethanol concentrations, inhibition of growth, viability and a decreased cell volume can be observed; while at high concentrations, cell death can occur [7]. However, yeast has learned to handle this kind of stress by introducing signal transduction pathways that adapt cell cycle, gene expression profiles and metabolic activities [7].

General stress response in *S. cerevisiae* Cells act in order to survive in non-perfect conditions like fluctuations in nutrients, pH, temperature, and exposure to UV irradiation or potentially toxic compounds [9]. When investigating the gene expression profile of yeast under different stress conditions, there are around 900 genes (14 % of the total number of genes) that have an altered expression similar for all stress environments. Around 600 genes have a decreased transcription [7]. Those genes are generally involved in energy consuming and growth-related processes, mRNA metabolism and protein synthesis [7] [9]. The other 300 genes are upregulated in stress conditions and are related to carbohydrate metabolism, detoxification of reactive oxygen species, protein folding, DNA damage repair, vacuolar and mitochondrial functions and production of protective proteins like heat shock proteins [7] [9]. Heat shock proteins possess a variety of functions, e.g. remodeling of unfolded proteins or preserving membrane integrity [7]. Summarized, when *S. cerevisiae* is exposed to a stressful environment, it will act by saving energy by decreasing the transcription of processes that are unnecessary for immediate survival. This energy can then be used to increase the expression of proteins that aid in cell protection.

Specific ethanol stress response in *S. cerevisiae* Next to the mechanism that acts in general stress conditions, there is also a response specific to the kind of stress. Three reactions have been found to take place in ethanol stress and not (yet) in other stress environments, so they are assigned as being specific to ethanol stress [7]. First of all, the cell enters a pseudo-starvation state [7]. The reason for this is not yet investigated, however a hypothesis is formulated. In *S. cerevisiae*, pyruvate is converted to acetaldehyde which is in turn converted into ethanol. This last chemical reaction requires the oxidation of NADH to NAD⁺ [10]. However, during ethanol stress, the level of acetaldehyde had appeared to

be much lower in the cell, decreasing the speed of the last chemical reaction. This not only decreases the production of ethanol, it also decreases the amount of NAD⁺ in the cell, which happens to be a cofactor in the glycolysis process [7]. Secondly, ethanol is known to reduce the pH inside the cell. This is counteracted by V-ATPases, the most abundant proteins in ethanol stress conditions [7] [11]. These are proton pumps that try to keep a constant pH both within the cell as well as within cell organelles [11]. Thirdly, there is an increased synthesis of tryptophan [7]. Here too, it is not exactly known how tryptophan affects the cell under ethanol stress, but it is believed that the loss in membrane function could result in a shortage of amino acids [7]. A summary of the total stress response is shown in 1.2.

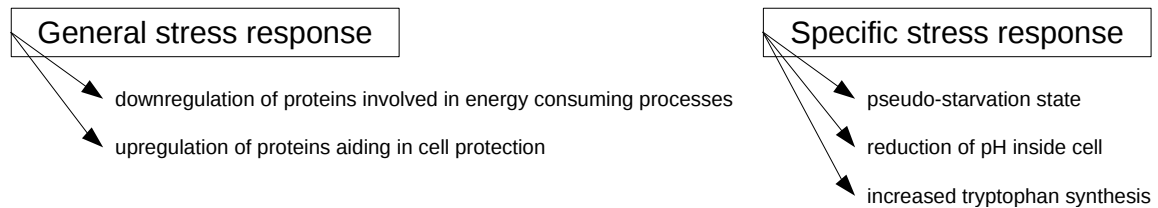


Figure 1.2: The ethanol stress response in *S. cerevisiae*

Ethanol tolerance in *S. cerevisiae* Ethanol tolerance is a process which is very hard to understand. It has been shown that a greater tolerance and a more rapid response can be achieved by pre-exposing the yeast to a low amount of ethanol. This effect can also be accomplished by exposing yeast to other stressful environments, which enforces the idea of a general response mechanism that becomes more efficient upon stimulation. Secondly, gene expression profiles change in later stages of exposure. This shows the difficulty of understanding this complicated biological process. Finally, there is some evidence that sensitivity to ethanol is dependent on its concentration [7].

1.2 Goal of this project

Up to date, most of the effort of studying the mechanisms of ethanol stress response and tolerance has been put in understanding what happens at the transcriptome level or by doing gene knock out studies [5] [7] [12]. However, a complete understanding can only be achieved by acquiring full knowledge of all levels in the cell. During this project, it is aimed to gain some understanding on the effect of ethanol on protein-protein interactions. The data set used for this research was obtained in a collaborative project called NATAR, that includes several groups at KU Leuven and contains SNPs of *S. cerevisiae* in an environment with increasing ethanol levels. Division of these nsSNPs based on their position in the protein (at the interface of an interaction, at the surface of the protein or in the core of the protein) is necessary to further investigate the effect of the observed nsSNPs on protein-protein interactions. This could reveal new mechanisms that are either part of the (general) stress response or that lead to ethanol tolerance.

Part I

Literature review

Chapter 2

Protein-protein interactions

2.1 From DNA to protein

Cells are usually classified as prokaryotic or eukaryotic. Prokaryotes are unicellular organisms that can survive in almost every environment. They have two membranes but no internal compartments, not even a nucleus. The genetic material, which is very often a single circular DNA molecule, is located in a nucleoid region. Eukaryotes are organisms like plants, animals, fungi and protists. Their cells are much bigger and contain different internal compartments. One of those compartments is the nucleus where the genetic material, consisting of different linear chromosomes, is located [8].

DNA is transcribed into RNA which is then translated into a protein. This whole process is called the “Central Dogma” of molecular biology. While this process proceeds for the prokaryotes completely at one place, it is a little bit more complicated for eukaryotes, which is explained below [8].

Parts of the DNA are transcribed by a RNA polymerase into RNA. Because RNA is a very unstable molecule, a cap is added at the 5' end and a poly(A) tail is added at the 3' end. Also, some internal fragments, called introns, are removed from the transcript. After that, the mature mRNA is transported via the nuclear pore complex outside the nucleus to a ribosome. During translation, the amino acids matching the codons of the RNA are transported to the ribosome by tRNAs. While the new amino acids are attached to the end of the chain, the beginning already starts to fold into the native protein structure. When the mature mRNA is completely translated, some post-translational modifications can occur before the protein is transported to its location where it can perform its function [8]. A summary of this process is shown in figure 2.1.

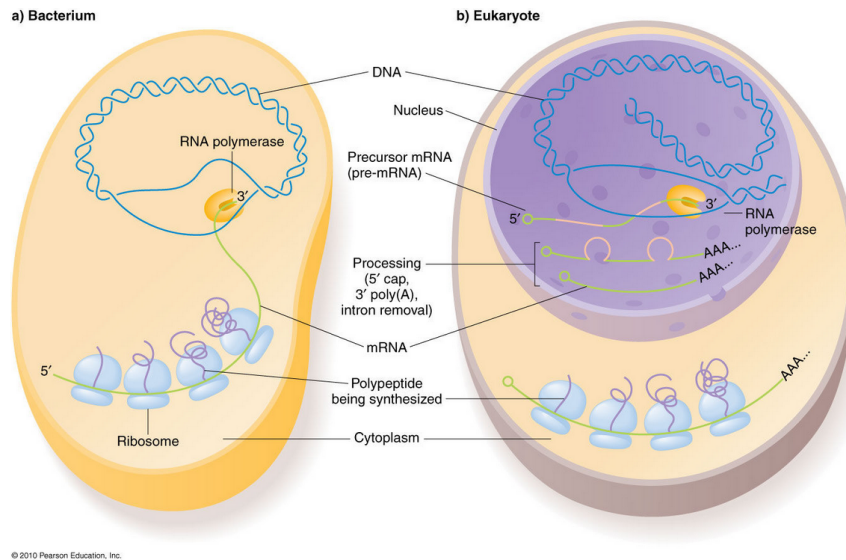


Figure 2.1: The Central Dogma of biology in bacteria and eukaryotes [13]

2.2 Interactions between proteins

Protein-protein interactions (PPIs) are formed when two or more proteins interact. Interactions are very important in cellular processes. Often, all proteins in the complex are needed for one particular function, but complexes might also evolve without a functional reason. This interaction is controlled by three factors: encounter, local concentration and local physicochemical environment. First of all, two proteins can only interact when they are co-localized in time and space. This can happen when the two proteins are co-expressed or localized in the same compartment, but also via diffusion or transport. Secondly, the local concentration should be high enough. This can be controlled at the gene level through regulated expression but also at the protein level through protein degradation, diffusion, and temporary storage. Finally, the affinity can be increased in the presence of an effector molecule or by a change in physiological conditions like pH, temperature or concentrations of ions. Most PPIs are very specific, even though there are some proteins that have multiple binding partners at overlapping interfaces [14]. The most important forces that stabilize protein-protein interactions are hydrophobic interactions, disulfide bonds and salt bridges [15]. Usually, there is only a small fraction of interface residues that contribute to most of the binding energy. These residues are called hot spots [15] [16].

Protein-protein interactions can be categorized based on different properties. First of all, all proteins involved can be the same or they can differ, called homo-complexes and hetero-complexes. Homocomplexes can be further divided into isologous or heterologous associations: while the two chains have exactly the same interface in an isologous association, this is not necessary during heterologous association. Next to that, the difference can be made between non-obligate complexes, when the proteins also exist independently, and obligate complexes. Finally, the PPIs can be categorized based on the lifetime of the complex: there are permanent interactions which are very stable and there are transient complexes. These transient complexes can be further distinguished in weak and strong. Weak transient complexes associate and dissociate constantly, while strong transient interactions require

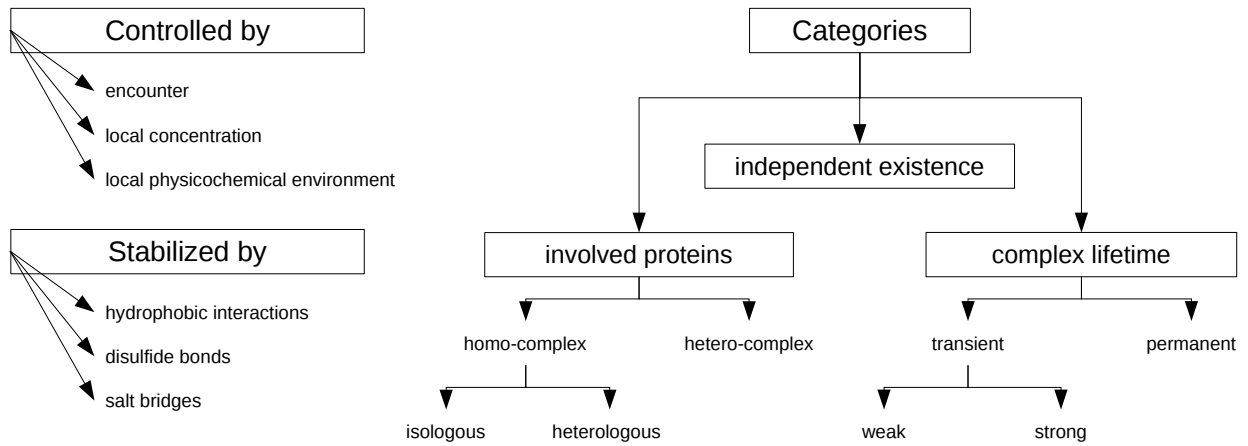


Figure 2.2: Some characteristics of protein-protein interactions

a molecular trigger for association or dissociation which stabilizes the complex, e.g. an effector [14] [17]. Obligate complexes are usually permanent, while non-obligate complexes can be either transient or permanent [14]. These properties are schematically displayed in figure 2.2.

All PPIs of one species can be integrated in a big and complex network. Although there is only limited interaction data available, of which experimental data is not always very reliable, it is generally assumed that biological networks are scale-free networks. Scale-free networks are characterized by the presence of hubs. A hub is a node which is highly connected. This is advantageous because it makes the network more robust to random defects and any two points are closer to each other. Due to the limited amount of hubs compared to the other nodes, the probability that one of these hubs will fail is a lot smaller. However, the network is a lot more sensitive to failure of a hub than to any other node. Both a random network and a scale-free network are shown in figure 2.3. In the random network, almost all nodes have approximately the same number of connections. In the scale-free network on the other hand, most nodes only have a couple of connections while there are a few highly connected nodes. In both figures 2.3 A and B, the five most highly connected nodes are coloured red and their first neighbours green. Almost twice as much nodes are reached by the red nodes in the scale-free network (60%) compared to the random network (27%). In the *S. cerevisiae* network that was known in 2005, 93% of the nodes had 5 or less connections. Of those nodes, only 21% appeared to be essential to survival. On the other hand, 62% of the 0.7% of the nodes that had more than 15 connections are essential to survival. The *S. cerevisiae* protein interaction network as known in 2005 is shown in figure 2.3 C. The color of the node indicates whether removal of the corresponding protein is lethal (red), nonlethal (green), slowing down growth (orange) or unknown (yellow).

It can be expected that such highly connected hubs possess some structural features that allow them to interact with a lot of different proteins. It is assumed that these hubs are intrinsically disordered (ID) proteins or at least contain some ID regions. An ID region is a protein region that lacks a fixed 3D structure. These regions are defined by some specific features like amino acid composition, charge, and exhibit properties like high specificity,

low affinity and binding diversity. When binding to a certain protein, part of the ID region becomes structured in such a way that the interaction is optimal. This is called a molecular recognition element (MoRE), and folds into an α -helix, a β -strand or an irregular structure. Four types of hubs can be distinguished: mostly disordered hubs that lack a complete structure without an interacting partner, partially disordered hubs where the disordered regions contribute mostly to the interactions, ordered hubs which seem to interact mostly with disordered partners and hubs that contain a disordered linker that can connect different functional domains [18].

2.3 Structural characteristics of the PPI interface

Jones and Thornton [19] have been trying to discriminate the interfaces of a PPI from the rest of the surface of a protein when the complex is not known. In a later study by Changhui Yan e.a. [15], the characteristics of the interface are compared to those of the rest of the surface and of the protein core using 6545 interfaces. In extension to this, they tried to get rid of the impact of the difference in solvent accessibility, so that the differences in characteristics can really be accounted to the difference in functionality. This is done by comparing a set of residues at the surface with the same distribution of solvent accessibility compared to the set residues at the interface [15]. Following properties can be compared:

size and shape

Δ ASA (accessible surface area) is the change in ASA when going from unbounded state to a complex. The size of the interface can be calculated as the mean Δ ASA, which is half the sum of the total Δ ASA for both molecules. A residue is defined to be at the interface when its ASA has decreased more than $1\text{\AA}^2$ [19] [16], and is said to be interacting to another residue when the distance between them is less than 5 [20] or $6\text{\AA}$ [15]. The average size of all studied interfaces is $1227\text{\AA}^2$, although all sizes span a broad range, which is shown in figure 2.4 [15]. The shape of the interface is explained both in terms of planarity and circularity. Planarity is measured by the root mean square deviation from the least squares plane of all interface atoms, while the circularity is calculated as the ratio of the principal axes of this plane [19].

complementarity between surfaces

Complementarity is defined by a gap index, which is the gap volume between molecules divided by the interface ASA [19]. There are four highly preferred types of complementarity in interacting interfaces: two cysteine residues due to the creation of disulfide bonds, two hydrophobic residues, two residues with opposite charges due to the formation of salt bridges and two aromatic residues [15].

residue interface propensities

When a residue has the tendency to appear more on an interface than on the regular surface, the interface residue propensity should be bigger than 1. This propensity is calculated by taking the ratio of the fraction of ASA that a particular amino acid contributes to the interface to the fraction of ASA that this amino acid contributes to the surface [19]. This definition can be extended to comparing the presence of an amino acid in one environment to the presence of that amino acid in the whole dataset

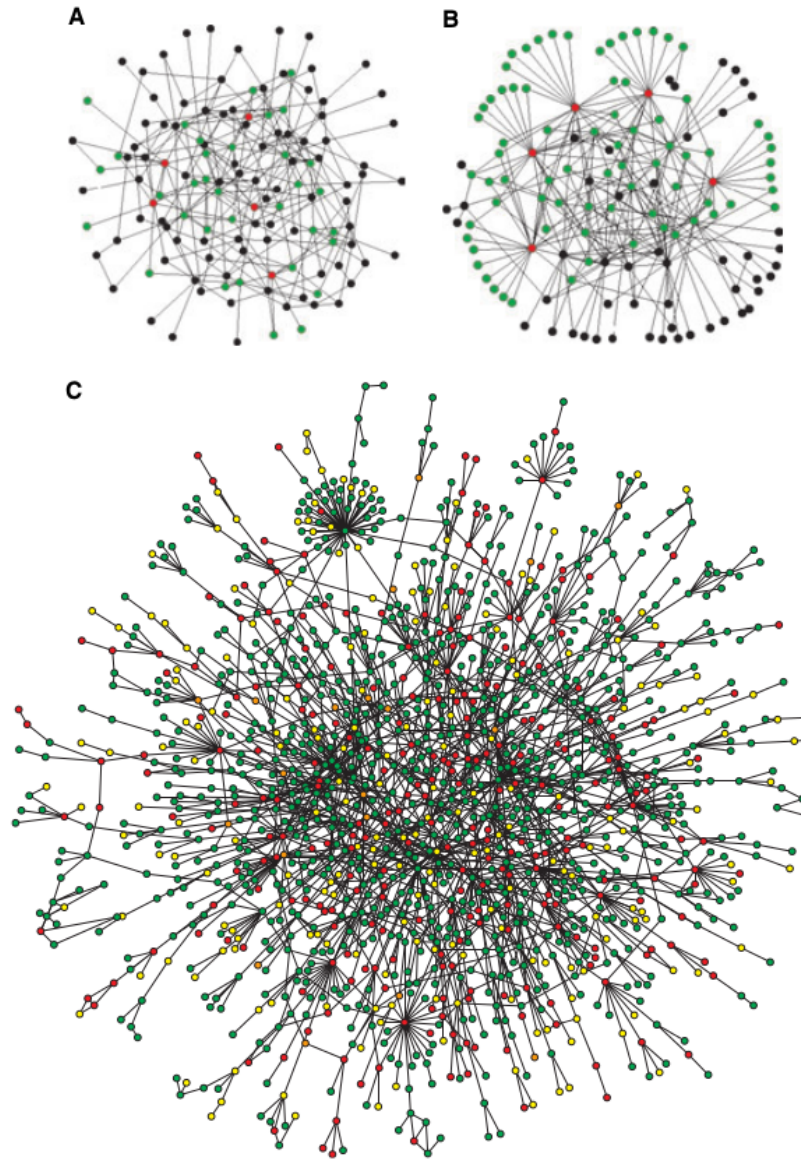


Figure 2.3: A) simulated random network; B) scale-free network; C) scale-free network of *S. cerevisiae* [18]

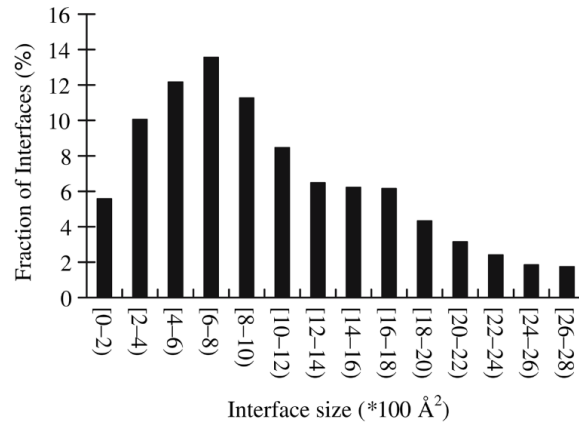
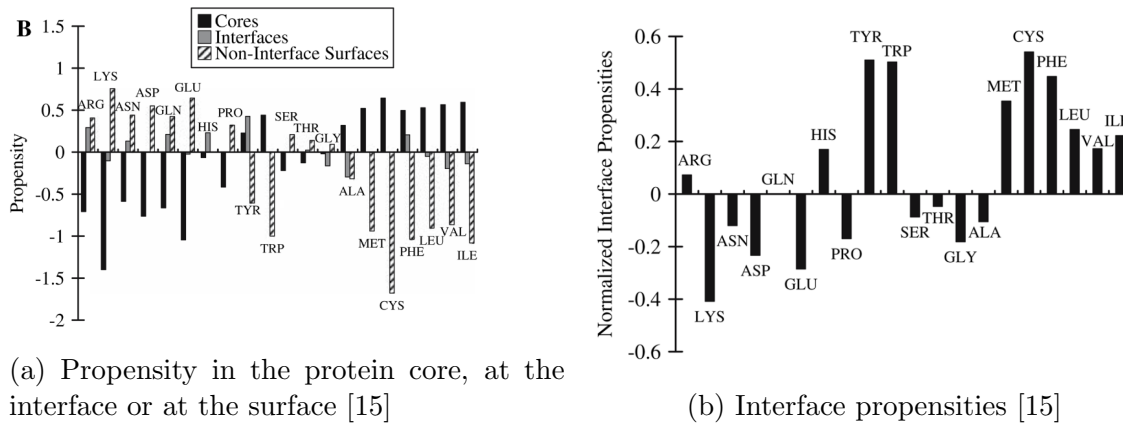


Figure 2.4: Interface size distribution for 6545 interfaces [15]



(a) Propensity in the protein core, at the interface or at the surface [15]

(b) Interface propensities [15]

Figure 2.5: Propensities for all amino acids sorted by increasing hydrophobicity

[15]. When comparing propensities of the residues in the different environments, it can be seen in figure 2.5a that there is a high propensity for hydrophobic amino acids in the protein core, while there is a high propensity for hydrophilic amino acids at the surface. The interface propensities are mostly in between those of the core and the surface, except for histidine, tyrosine and glycine [15]. When looking at figure 2.5b, which compares the propensities of amino acids at the interface in a data set that cancels out the effect of solvent accessibility, it is clear that the hydrophobic residues are highly preferred at interfaces, just like the aromatic residues and cysteine. This can be explained by the preferred complementarity between the interfaces as clarified above [15].

hydrophobicity including hydrogen bonding

In general, the hydrophobicity of an interface is in between the hydrophobicity of the core and the rest of the surface, which is hydrophilic. This is illustrated in figure 2.5a. Also, the mean number of hydrogen bonds per 100 Å² of Δ ASA is calculated to compare hydrogen bonding [19].

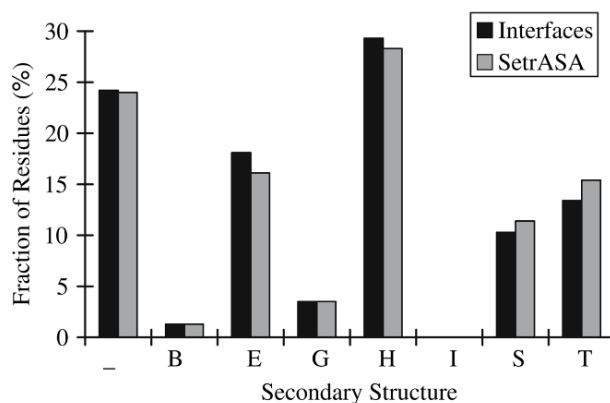


Figure 2.6: The composition of interfaces in terms of secondary structure: β -bridge (B), extended β -sheet (E), 3_10 -helix (G), α -helix (H), π -helix (I), bend (S), helix-turn (T) and other (-) [15]

segmentation and secondary structure

The segmentation of an interface can be defined as the number of segments in an interface of a protein. Two amino acids at the interface separated by more than 5 residues are defined to be in different segments. This segmentation is important, because a lowly segmented interface can easily be mimicked by a small peptide which decreases the probability of being an interface. For the secondary structure, an equal proportion of helices, β -strands and coils is observed in all complexes [19]. This result is confirmed by Changhui Yan e.a. As can be derived from figure 2.6, the difference in secondary structures between interfaces and surfaces with the same solvent accessibility distribution is very small, and is not statistically significant [15].

conformational changes on complex formation

There are different conformational changes possible during complex formation, from no change at all to changes of side chains to even changes of the main chain or movements of the domains. However, up to date it is not clear how formation of a complex induces conformational changes, although conformational changes are very likely when the contact area is $1000\text{\AA}^2$ or bigger [14] [17] [19].

There seems to be no single property that can truly predict the interface location, but all measured properties together could possibly distinguish an interface from the rest of the surface. Therefore, it could be possible for a protein whose interaction interface is not known to calculate all the properties for the surface and rank parts of the surface based on the probability that it will be part of the interface [19].

Jones and Thornton [19] also compared the interfaces of two types of PPIs, homodimers and heterodimers, based on these properties. Two remarks should be noted: first of all, this research is limited to complexes of only two proteins, but it is generally assumed that conclusions can be extended to complexes with more than two proteins. Also, in the paper, the distinction is made between homodimers and heterodimers. However, it seems more appropriate to make a distinction based on a different characteristic of the interaction: obligate or non-obligate. This is what is done as well in later work [14] [17]. The conclusions

of those two interaction types drawn by Jones and Thornton [19] are briefly summarized here.

Obligate complexes In obligate complexes which are mostly permanent, there is a relation between the molecular weight and the size of the interface: larger molecules tend to have larger interfaces. Also, the interfaces are less planar but relatively circular. The two proteins are more complementary compared to transient interactions. Next to that, the interfaces have the greatest preference for hydrophobic residues. This makes the interface more hydrophobic but this can be expected since the interface is permanently located inside the complex. Finally, there is a high segmentation for all interfaces [14] [19].

Non-obligate complexes Non-obligate complexes on the other hand, have smaller interfaces compared to obligate interactions. As already mentioned, the interfaces are more planar and less complementary to each other. Here too, the interfaces are relatively circular and highly segmented, except for the inhibitors in enzyme-inhibitor complexes. They have only 2-5 interacting segments and the interfaces are less circular. Because the proteins also exist independently, their interfaces are less hydrophobic, but this is balanced by a greater preference for polar residues, leading to a higher number of salt bridges and hydrogen bonds per ΔASA . When making the difference between weak and strong transient interactions, the interfaces of weak interactions are even smaller and more planar [14] [17] [19].

Chapter 3

The effect of genetic variation on PPIs

3.1 Conservation of the PPI interface

Some positions in proteins are conserved. This means that similar protein sequences exist either in the same organism or in different organisms that show the same or a similar amino acid at analogous positions. The more conserved a position is, the more important the residue is believed to be [21]. This conservation can be expressed in terms of entropy, which is a measure of chaos or unpredictability. The lower this value, the more conserved a residue is. Due to the functional importance of the residues needed for binding, it can be expected that the interface of the protein is more conserved than the rest of the surface [16]. Normalized entropies for residues in the three types of environment in proteins (core, interface, rest of the surface) are displayed in figure 3.1a. All entropy values range between 0 and 100%. The core appears to be the most conserved environment, which can be explained by the fact that changes inside the core can disrupt the whole protein structure. The surface on the other hand shows the smallest fraction of highly conserved residues [15]. This difference is still present when the effect of solvent accessibility is removed, see figure 3.1b [15].

When comparing interface residues with other surface residues, the conservation is in most cases higher at the interface [15] [16]. In protein structures where this is not the case, it could be argued that there are other, yet unknown interaction sites. Also, a catalytic site in an enzyme has a conservation even higher than any other interface. These two observations, together with the type of alignment used, can influence the conservation of the interface relative to the surface [16]. Daniel R. Caffrey e.a. also compared the interface with parts of the surface that contain more or less the same number of residues, called surface patches, to investigate the possibility of predicting interfaces based on conservation scores [16]. However, the difference in conservation between interfaces and surface patches is not statistically significant, indicating that is practically impossible to predict interaction sites using conservation alone [16].

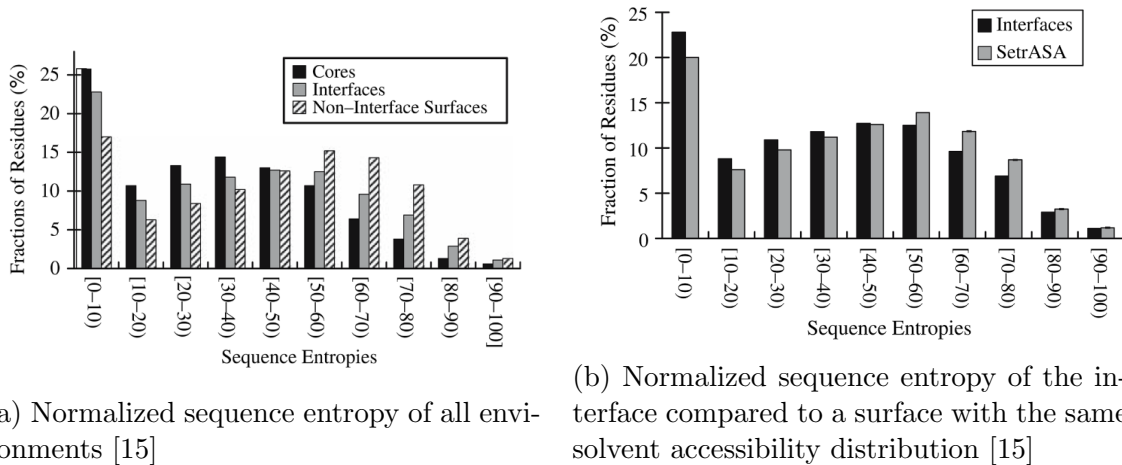


Figure 3.1: Normalized sequence entropies of different environments

3.2 Non-synonymous single nucleotide polymorphisms in PPIs

Different organisms from the same species do not have completely the same genetic material. This difference can be good, as it allows flexibility in phenotype or prepares for changing environments. There are different types of genetic variation, but in this work only single nucleotide polymorphisms (SNPs) are discussed. A SNP is a variation at one position in the DNA [22]. When the SNP is present in a gene-encoding DNA sequence, it can affect the amino acid sequence of the encoded protein. This is then called a non-synonymous SNP (nsSNP) and is often associated with disease [23].

nsSNPs can occur in the core of the protein, at a protein-protein interface or at the rest of the protein surface [20]. nsSNPs associated with a disease are most often found at the core of the protein, because the amino acids in the core are usually hydrophobic and have constraint space. By mutating one of these residues into a polar or hydrophilic one, or by changing a smaller amino acid into a bigger one (or vice versa), destabilization or misfolding of the protein is very likely to occur. On the surface on the contrary, more variation is possible, unless this variation occurs at an interaction site. PPI interfaces are more conserved than the rest of the surface, increasing the probability for a mutation to have some effect compared to the rest of the surface [23].

Generally, mutations inside an interface can stabilize or destabilize PPIs. When a protein has multiple interfaces to connect with different proteins, it is obvious that changes in one interface do not have the same effects as changes in another interface. Also, disease-causing nsSNPs often occur in proteins that are located centrally in the PPI network. This increases the influence the nsSNP has on the cell [23]. In this section, the effect of nsSNPs on intermolecular contacts, post-translational modifications and intrinsically disordered regions in PPIs is described.

Intermolecular contacts A mutation in the interface of a PPI can lead to a loss, a gain or a change of function. First of all, a loss of function is observed when an interaction is disrupted. As explained before, the residues at an interface are more hydrophobic compared

to the rest of the surface. When one of the hydrophobic residues mutates to a charged or polar residue, the interaction might be harmed. Interaction can also be impaired by mutating residues that are involved in noncovalent interactions like salt bridges or hydrogen bonds or in covalent interactions like a disulfide bridge. The secondary structures inside an interface can also be affected when a mutation results in a proline residue. Proline is a cyclic amino acid that introduces abrupt changes in the main chain of the protein. The distorted backbone in the interface can make it hard to interact. Secondly, a gain of function can be obtained when new interactions are introduced. This is the case when a mutation in an interface causes the interface to be recognized by another protein than its usual partner. This could also cause a switch of function. Incorrect interactions might also lead to aggregation of the complex, which influences the behavior of the cell [23].

Post-translational modifications (PTMs) Not only do nsNSPs affect intermolecular contacts, post-translational modifications (PTMs) can also be modified. Transient interactions are often regulated by PTMs. A very common PTM in signaling cascades is phosphorylation, which occurs at a serine, threonine or tyrosine residue. When one of these residues is mutated to another residue that cannot be phosphorylated, phosphorylation will not take place, leading to a loss of the regular interactions. Glycosylation can also be affected when mutations modify or introduce recognition motifs for glycosylase enzymes. A gain of glycosylation could lead to new binding partners, while a loss of glycosylation could lead to interaction losses [23].

Intrinsically disordered regions (IDRs) Some proteins contain regions that lack an ordered 3D structure and obtain this during interaction. These are intrinsically disordered regions and mutations in such an IDR could create more order, thereby leading to an easier interaction and aggregation. Very often, IDRs contain a lot of PTMs, which can also affect PPIs as explained before [23].

3.3 Assessment of the effect of nsSNPs on PPIs

When trying to investigate the effect of nsSNPs on PPIs, the process can be divided into two parts: first all the nsSNPs located on a PPI interface should be extracted, and then it can be tried to determine what type of effect the nsSNP has on the PPI. This is explained in this section.

Determination of nsSNPs located at an interface Alessia David e.a. identified the preference for nsSNPs to be in the core, on the interface or on the surface of a human protein. To do so, they had to identify the location of all SNPs. First, they collected all structures from the protein data bank. To extend their analysis, they also included homologous complexes when the human structure is not available. After that, the residues are categorized as interface residue, surface residue or core residue based on the following properties: when 2 atoms of 2 residues are closer than $5 \text{ \AA}$, the 2 residues are defined to be on the interface. However, when all interface residues are found, the interfaces with less than 5 residues are excluded because this is assumed to be too small for an interface. All residues with a solvent accessible surface area of less than $5 \text{ \AA}^2$ are defined to be in the core

of the protein. All other residues are considered to be at the surface. When this is done, the nsSNPs can be mapped on the structures and frequencies can be calculated to analyze the preference location of nsSNPs [20].

Prediction of the effect of an nsSNP on the PPI Different algorithms have been developed for prediction of the effect of mutations on PPIs, dividing the nsSNPs into three categories: deleterious, neutral or beneficial. The names of these categories do not refer to their biological impact, but rather to the effect the nsSNP has on the PPI: a detrimental nsSNP disrupts the PPI, a neutral nsSNP does not change the binding complex and a beneficial nsSNP increases the binding affinity of the complex [24] [25]. In a community-wide blind prediction experiment, twenty-two algorithms were compared against each other. Three methods outperformed the others, and the following aspects distinguish these methods from the rest: explicit consideration of stability and modeling of packing, electrostatics and solvation. Different approaches are used: machine learning by Fernandez-Recio, atom-level energy functions by Weng and coarse-grained models by Dehouck and Rooman [24]. Another algorithm published slightly after this broad evaluation is called the SNP-IN tool [25]. These methods are briefly described here. Although all 4 methods mentioned are relatively successful in the prediction of deleterious mutations, a lot of improvement still has to be made in successfully predicting beneficial mutations [24] [25].

Fernandez-Recio

Machine learning is used by the lab of Fernandez-Recio e.a. At first, the structure of the mutated protein is predicted using FoldX [26]. Out of this mutant structure, 85 descriptors are derived. The probability for the mutation to be stabilizing or non-stabilizing is calculated as the mean confidence value of 5 classifiers. Those 5 classifiers are selected as the best performing of different regression learners that aim at predicting ΔG and different binomial classifiers that classify a mutation as stabilizing or non-stabilizing [24].

ZAFFI, Weng

The lab of Zhiping Weng e.a. developed the ZAFFI protocol (Zlab AFFInity enhancement). This protocol is based on the concept of free energy (ΔG): a reaction occurs spontaneously when its ΔG is smaller than 0. This ΔG consists of an enthalpic term and an entropic term. In practice, it is very difficult to quantify the entropy and therefore other approaches that approximate ΔG must be used. In this case, the ZAFFI function is used to calculate the energy of the wildtype complex and the mutant complex. When the difference of both is smaller than 0, the ZAFFI filter is used to separate true positives from false positives. The cut-off value of the filter is set to 0.05: when the ZAFFI filter value is smaller than 0.05, a mutation passes the filter. The ZAFFI function consists out of different terms on the atomic level, which are weighted using multi-linear regression. This process is summarized in figure 3.2 [27] [24].

Dehouck and Rooman

Here too, the concept of ΔG is used, but instead of using terms at the atomic level, the structure is rather described by descriptors at the residue level. First, the effect of the

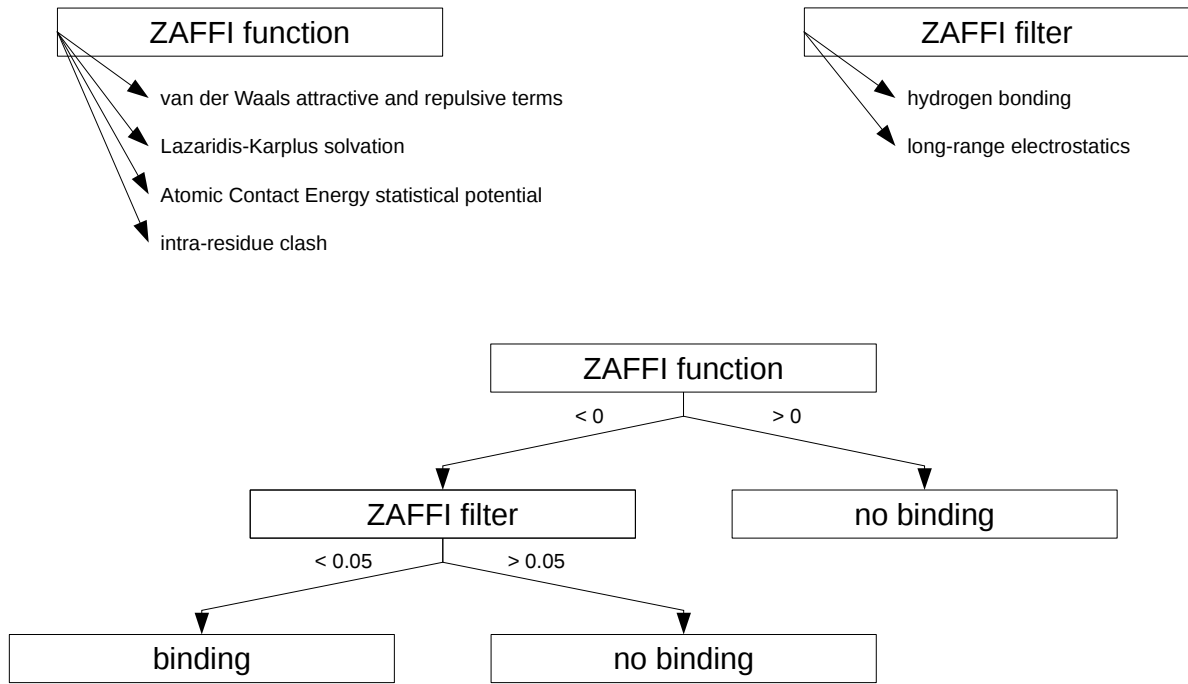


Figure 3.2: The ZAFFI protocol

mutation on the stability of the complex and of the single molecule is examined. This effect is calculated via the tool PoPMuSiC, which incorporates 13 potentials, volume changes to account for packing defects and an independent term. The potentials are functions of four descriptors: three for each residue - amino acid type, torsion angles, solvent accessibility - and one for each pair of residues, the distance between the average geometric centres of the side chains. These 16 terms are weighted using sigmoid functions, allowing a smooth transition from the protein core to the protein surface. Secondly, the effect on the binding affinity can be calculated by using a weighted combination of the effect when the mutation is located at the interface and the effect when the mutation is located anywhere. This last value is an approximation of the ΔG of the reaction and whether or not binding happens depends on interpretation of this value [28] [24].

SNP-IN tool, Zhao

The SNP-IN tool is a tool that predicts the effect of a nsSNP based on three classification problems. In the first problem, it is assumed that the nsSNP affects the function of the PPI and it is determined whether the nsSNP strengthens or weakens the binding. In the second problem, the nsSNP is classified as either disrupting or preserving the PPI. The third classification problem classifies the nsSNP in 1 of the following 3 categories: detrimental, neutral or beneficial. This can be seen in figure 3.3. Different supervised and semi-supervised learning methods are compared to each other. In supervised learning, all the nsSNPs are previously labeled, while in semi-supervised learning, unlabeled nsSNPs are also taken into account to increase the performance. This is represented in figure 3.4. Every nsSNP, labeled or unlabeled,

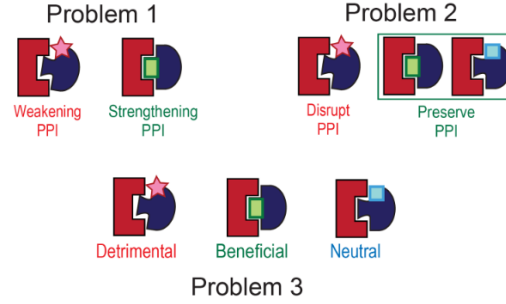


Figure 3.3: Three classification problems of SNP-IN tool [25]

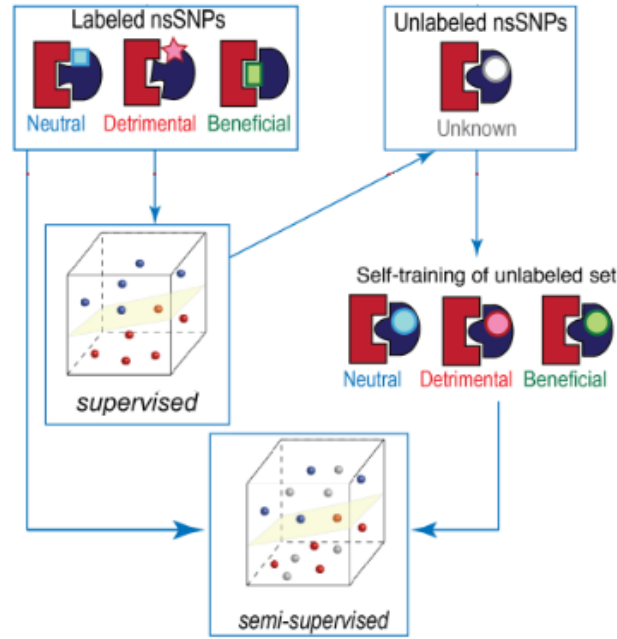


Figure 3.4: Semi-supervised learning in SNP-IN tool [25]

is represented as a set of 33 features taking into account structural information of the complex. Only for the first problem, supervised learning seems to perform better than semi-supervised learning. All problems use random forest classifiers, because their performance is better than support vector machines. For the third problem, an increase in performance is obtained by adding the classification of the first two problems as additional features. The tool itself requires as input a complex and a mutation, and it outputs the likelihood of a disruption of the PPI and the likelihood of the type of effect. Also, 3D structures of both the regular interaction as well as the interaction with the mutation are shown. When comparing this tool to other existing methods, SNP-IN seems to outperform all other methods. However, this tool only focuses on the intermolecular contacts that have changed due to the mutation. Incorporation of the effects on PTMs or IDRs is still a challenge [25].

Part II

Materials and methods

Chapter 4

Overview of methodology

The methodology to get all nsSNPs at interacting sites is briefly explained here and a more detailed explanation can be found in the remaining chapters of this part:

1. **Extraction of all interacting amino acids**

In a first phase, all amino acids that are part of an interaction site must be selected. To do so, all known complexes of *S. cerevisiae* proteins are downloaded from the protein databank. Using the structural data, all amino acids that are close together and belong to a different chain are considered to be interacting. This is described in detail in chapter 5.

2. **Processing of nsSNP data**

Secondly, the SNP data of the NATAR project is processed. All SNPs are linked to their Ensembl identifier and gene information is obtained from BioMart. This is explained in chapter 6.

3. **nsSNPs affecting an interacting amino acid**

Finally it is possible to calculate the amino acid number of the SNPs and to link this list to the list of interacting amino acids. For all SNPs that are located at a known interaction site, the mutated amino acid can be calculated. Comparison between the original and the mutated amino acid divides the SNPs into three categories: nonsense, silent and non-synonymous mutations. More about this can be found in chapter 7.

Once all calculations are performed, the SNPs located at interacting sites are investigated in depth and the results are summarized in Part III. Also, the methodology can be automated so that it becomes very easy to see whether a SNP is located at the interface of an interaction and whether the SNP is a nonsense, silent or non-synonymous mutations. This is shown in chapter 8.

Chapter 5

Extraction of interacting amino acids

In this chapter, it is explained how amino acids at interacting sites are extracted from protein complexes. A schematic overview can be seen in figure 5.1.

5.1 Collection of potential complexes

In order to list all amino acids known to be at a protein-protein interaction site, all studied *S. cerevisiae* protein complexes must be used. These can be found in the RCSB protein data bank via the advanced search field [29]. Two requirements are explicitly queried: the source organism and the number of entities. The source organism is set to *S. cerevisiae* *S288C* because it is the most frequent strain of *S. cerevisiae*. The number of entities of the type *protein* is set to be at least two to ensure that all selected structures are complexes. On the 26th of November 2014, this results in a list of 4035 unique chains mapping to 359 PDB files. All those files can be downloaded together and collected in one folder. Every PDB file has a unique identifier consisting of four letters and/or numbers. The name of every file is this ID together with the extension *.pdb*.

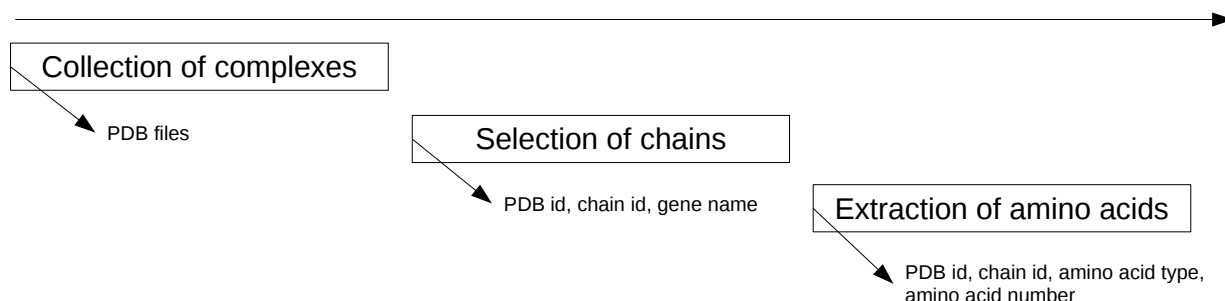


Figure 5.1: Schematic overview of amino acid extraction

5.2 Selection of chains

After collecting all potential complexes, two problems are revealed:

- A complex is selected as soon as one of the chains originates from *S. cerevisiae*. However, it might be possible that a complex exists out of one *S. cerevisiae* chain and a synthetic construct or a chain originating from another organism. These interaction sites should not be taken into account as they might not reflect a *S. cerevisiae* complex.
- A PDB file does not contain the name of the genes that encode for the proteins present in that file. It is important to know the name of the gene in order to connect the nsSNPs to the interaction sites.

To solve these issues, connection with the webpage of every PDB file is necessary. This is done with a Python script, *linkchaintogene.py* that is explained here:

1. All PDB identifiers are obtained by appending the names of all files to a list.
2. A loop going over all PDB files is created, doing the following:
 - (a) The *urllib* module of Python is used to retrieve the html page from an url that is created by adding the unique PDB id behind
`http://www.rcsb.org/pdb/explore/explore.do?structureId=` [30].
 - (b) The Python library *Beautiful Soup* is used to extract the necessary data out of the html page [31]. Out of this data, three lists are created: one that contains the chain identifiers, one that contains the gene names and one that contains the source organisms.
 - (c) Now it is checked whether at least one gene of *S. cerevisiae* is present. If not, the PDB identifier is placed on a list with files that have to be removed. This is the case for 47 files.
 - (d) For the other files, the length of the three lists are compared. In most cases, the lists are equally long and can be pasted next to each other. Together with the PDB identifier, a table is thus created where every row contains a PDB identifier, a source organism, chain identifier(s) and a gene name. In case multiple chains belong to the same gene, the list of chain identifiers is split up, so that every row contains one chain identifier.
 - (e) In 123 of the cases, the three lists have different lengths, indicating that one of the fields on the html page was not filled in. Very often this is because a synthetic construct is also present in the complex, which misses a gene name and/or a source organism. However, because there is no consistency in what is present and what is missing, it is very hard to program this. Therefore, those files have to be checked manually, so their PDB identifiers are also put on a list.
3. After looping over all files, the long table existing of 4 columns is printed on to a text file. The files that have to be checked manually, are checked now and the text file is modified according to the information present on the web pages of the PDB files. The first lines of the resulting text file *chain_gene.txt* are shown in figure 5.2.

1zuk	Saccharomyces cerevisiae	A	BBC1 MTI1 YJL020C J1286 J1305 YJL021C
1zuk	Saccharomyces cerevisiae	B	BBC1 MTI1 YJL020C J1286 J1305 YJL021C
1zuk	Saccharomyces cerevisiae S288c	C	LAS17 YOR181W
2ZPN	Saccharomyces cerevisiae	A	ATG8 APG8 AUT7 CVT5 YBL078C YBL0732
2ZPN	Saccharomyces cerevisiae	B	ATG8 APG8 AUT7 CVT5 YBL078C YBL0732
2ZPN	Saccharomyces cerevisiae	C	ATG8 APG8 AUT7 CVT5 YBL078C YBL0732
2ZPN	Saccharomyces cerevisiae	D	ATG8 APG8 AUT7 CVT5 YBL078C YBL0732
2ZPN	Saccharomyces cerevisiae S288c	E	ATG19 CVT19 YOL082W O0980 YOL01

Figure 5.2: Head of *chain_gene.txt*

This file is further modified by deleting all lines that do not contain *Saccharomyces* using another script, *deletenonschains.py*. Since it is sure now that all chains of all PDB files originate from *S. cerevisiae*, the column of source organism can also be removed.

5.3 Extraction of amino acids

In order to get all interacting amino acids belonging to one of the chains of the list, a new script, *findaa.py* is executed. This script follows these steps:

1. All the identifiers of the residual PDB files are put into one list.
2. A loop goes over every file and performs the following actions:
 - (a) For all lines starting with 'ATOM', it is checked whether or not the chain that atom belongs to is listed into the text file. If it is, the following information of that line will be stored: the chain identifier, the type of amino acid, the number of the amino acid and its coordinates.
 - (b) When all the lines of the PDB file are processed and all information of the amino acids is stored, a new loop is started. This loop calculates the euclidean distance of every two amino acids that do not belong to the same chain. If this distance is smaller than 5 Å, both amino acids are stored in a list [20]. This list contains 4 columns: PDB identifier, chain identifier, type of amino acid and number of amino acid.
3. When the loop is finished, all double lines are deleted from the resulting list. This final list is then written to a text file and contains 205328 amino acids.

Finally, the interacting amino acids should be connected to their genes. It is decided to work with the Ensembl gene identifier of the gene, instead of the associated gene name. This is mainly because the Ensembl gene identifier is unique, while there might be different names of one gene. This is done in the following way:

1. The last column of *chain_gene.txt* is extracted and split based on the spaces between all possible gene names. The first four columns of all possible names are concatenated in one big list.
2. This list is introduced in the BioMart ID conversion tool [32] in two directions: first, the list is used to translate all gene names in their Ensembl gene identifier and then

1zuk	A	PHE	14	YJL020C
1zuk	C	PRO	5	YOR181W
1zuk	C	ALA	4	YOR181W
1zuk	C	PRO	6	YOR181W
1zuk	C	PRO	3	YOR181W
1zuk	A	TYR	16	YJL020C
1zuk	C	PRO	8	YOR181W
1zuk	C	ARG	11	YOR181W
1zuk	A	LYS	17	YJL020C
1zuk	B	PRO	15	YJL020C

Figure 5.3: Head of *interactingaagene.txt*

the list is used to translate all Ensembl gene identifiers in their associated gene name. These two results are concatenated together into 1 table - *idnamelist.txt* - consisting of 2 columns, one with all gene names and one with all Ensembl gene identifiers.

3. The script *attachidtochain* compares for every line of *chain_gene_result.txt* the first gene name of the last column with the columns of a line in *idnamelist.txt*. If there is a match, the Ensembl gene identifier is placed before this line of *chain_gene_result.txt*. If there is no match and there is a second gene name, then this is compared to the columns of the same line of *idnamelist.txt*. This is also checked for the third and the fourth column. If there is still no match, then the first field is left empty. The table is then written to *chainidlist.txt*.
4. Only for 8 lines the first field is left empty, and this is summarized in table 5.1. The first gene, CENPK1137D_4440 [33], belongs to baker's yeast and cannot be linked to a gene name with SGD, UniProt, UCSC Genome Browser or BioMart [34] [35] [36] [32]. HSP90B, although listed in the PDB as belonging to *S. cerevisiae*, turns out to be a human protein [37]. Therefore, these 4 lines still have an empty Ensembl gene identifier field, and they will be left out later. The other 4 lines belong to gene NAB1A, which is an alias for RPS0A with Ensembl gene identifier YGR214W which is filled in in the table [38].

Gene name	Line number
CENPK1137D_4440	547, 548
HSP90B	849, 850
NAB1A	969, 1144, 1363, 2604

Table 5.1: Unrecognized gene names and their line numbers

Finally, the Ensembl gene identifier can be linked to every amino acid based on the combination of PDB id and chain number: the script *connectaatogene.py* loops through every line of *interactingAA.txt* and searches for the combination of PDB id and chain number in *chainidlist*. The resulting Ensembl gene identifier is pasted at the end of the line, and the new lines are written to *interactingaagene.txt*. The result is shown in figure 5.3.

Chapter 6

Processing of SNP data

As mentioned in chapter 1.2, the data set used for this research is from the NATAR project. The data set is obtained during an experiment in an environment with increasing levels of ethanol (from 6% to 12%) and aims at investigating the adaptive evolution of *S. cerevisiae*. The increasing ethanol concentration ensures that a constant pressure is present, and that the cells keep on adapting to the changing environment. The experiment is set up in 6 fermentors, of which two are containing haploid yeast cells, two are containing diploid yeast cells and two are containing tetraploid yeast cells. However, there is a rapid convergence to diploidy. At different time points, the whole population in every fermentor is sequenced to see how common a mutation is. This results in a list of mutations with their frequency in the population (ranging from 0% to 100%) at different moments in time. These time points can be seen in figure 6.1. At the same time points, three ethanol-tolerant clones from every fermentor are sequenced to see which mutations occur in the same lineage. The resulting frequencies then indicate whether the mutation is absent (0%), heterozygous (50%) or homozygous (100%). The data set consists of two folders called clonal and population that contain the mutation frequencies of respectively the clones and the whole population. Each folder contains another two subfolders, one with SNP data and one with indels (insertions and deletions of bases in the genome [39]). Every subfolder contains six files, one for every fermentor. The whole processing of the SNP data is summarized in figure 6.2 and is explained in this chapter.

6.1 Preparation of SNP data

First of all, the most convenient way to use the SNP data later on is to have all data in one file. To do so, a script, *SNPprep.py*, has been written, which goes through all twelve files in the SNP subfolders, and stores all non-intergenic SNPs in one file, listing for every SNP its origin (clonal or population), fermentor (1 to 6), the chromosome, gene and its position, and the reference and alternative base pair. This results in 5653 unique SNPs in the file *preppedsnps.txt*. At the same time, a list is drafted that keeps track of all genes in which a SNP has been observed. This list is called *genelist.txt* and contains 3010 unique genes. Heads of both lists are shown in figure 6.3 and figure 6.4.

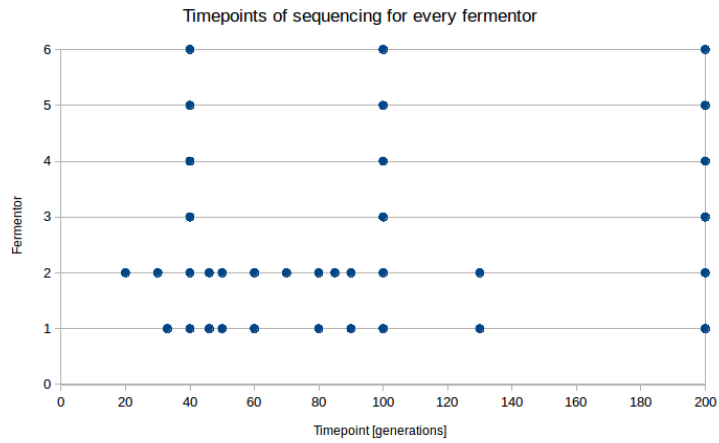


Figure 6.1: Timepoints of sequencing for every fermentor

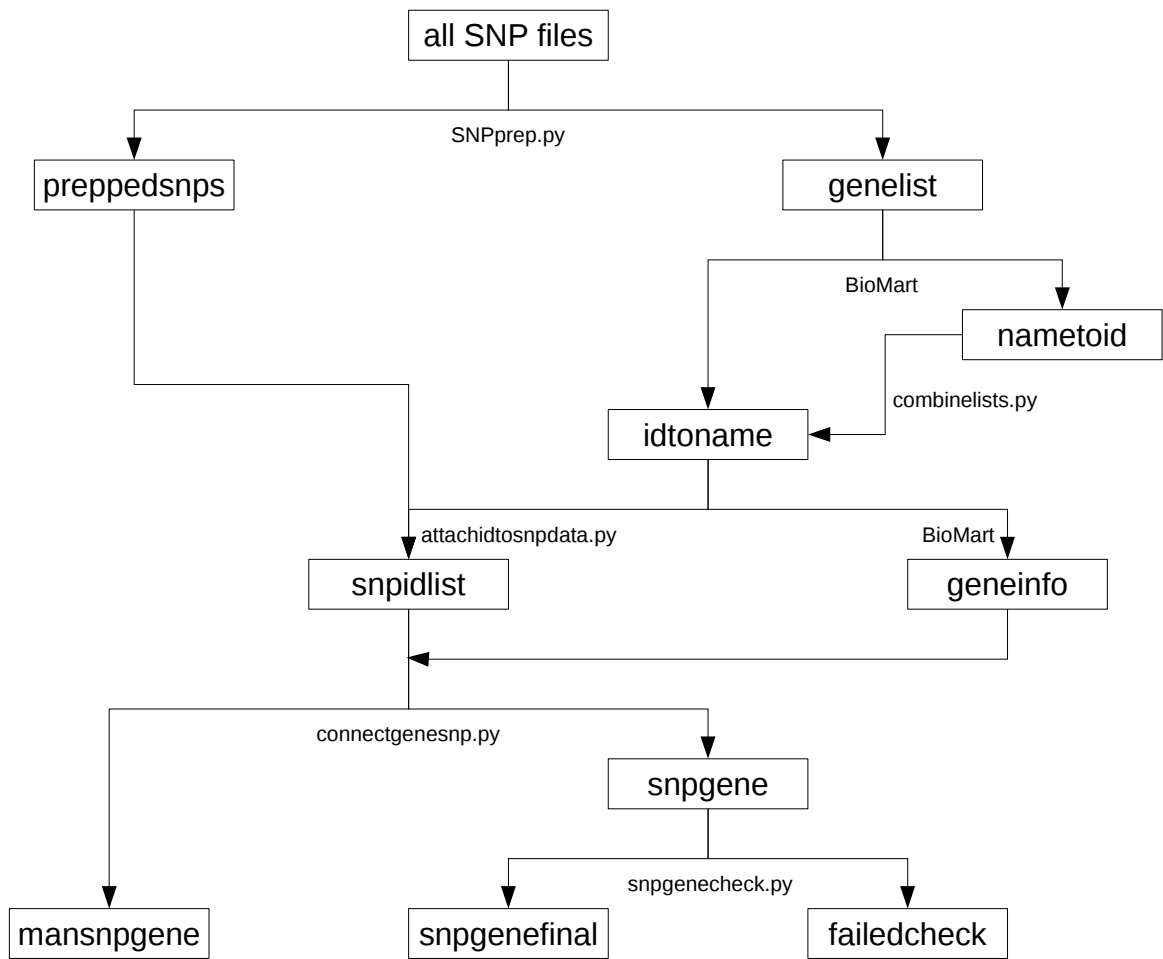


Figure 6.2: Summary of SNP processing

Clonal/Population			Fermentor		Chromosome	Gene name	Position	Reference	Alternative
C	F3	chrIX	ASG1	103484	G	T			
C	F3	chrVII	tK	122270	C	T			
C	F3	chrXVI	GRE1	129020	C	A			
C	F3	chrXVI	CBP3	148364	C	A			

Figure 6.3: Head of *preppedsnps.txt*

```

Gene
10/01/14
AAC1
AAD14
AAD4

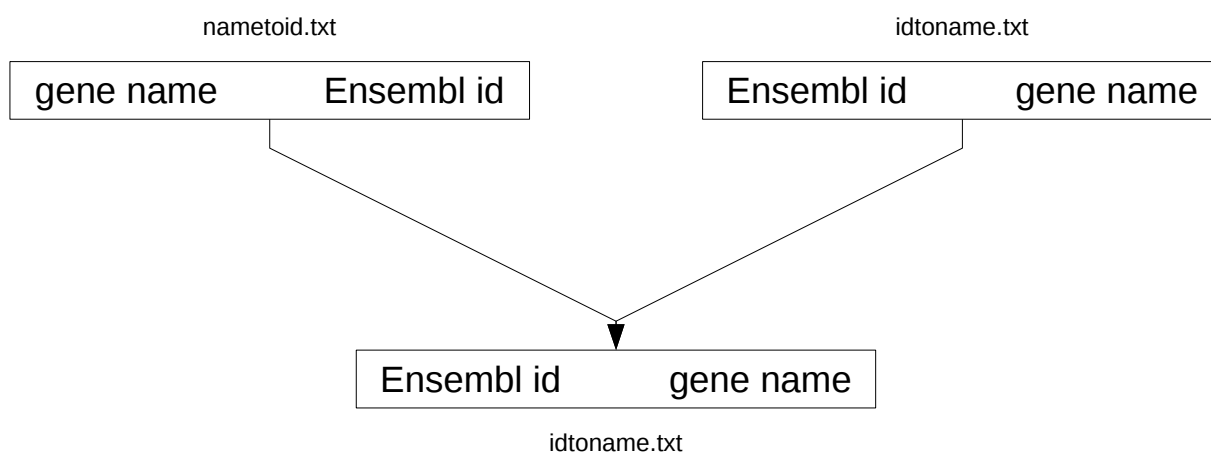
```

Figure 6.4: Head of *genelist.txt*

6.2 Obtaining gene information

To obtain the necessary gene information, the list with all genes in which a SNP is observed, *genelist.txt*, is used. By deeper investigation it has become clear that the list consists both of gene names and Ensembl identifiers. To further use this list in an automatic way, it is necessary that the list consists either only of gene names or of Ensembl identifiers. Therefore, the ID conversion tool of BioMart is used [32]. The whole list is converted from associated gene name to Ensembl gene identifier, resulting in a new list - *idtoname.txt* - of 2537 gene names with their associated Ensembl gene identifier. Next to that, the original list is converted from Ensembl gene identifier to associated gene name, producing a new list - *nametoid.txt* - of 456 combinations of gene name and Ensembl identifier. Those two results are combined together in *idtoname.txt*, as shown in figure 6.5. This gives a list of 2993 gene names with the corresponding id, and a total of 17 genes that were not recognized by BioMart.

The 2993 Ensembl gene identifiers are used to obtain all the necessary gene information. For this, the BioMart tool to browse data is used [40]. The identifier list is uploaded and as output the following information is requested: Ensembl gene identifier, associated gene name, chromosome, gene start, gene end and strand. The resulting list is called *geneinfo.txt* and the head of it is shown in figure 6.6.

Figure 6.5: Appending *nametoid.txt* to *idtoname.txt*

Ensembl	Gene ID	Gene	Start (bp)	Strand	Gene End (bp)	Associated Gene Name	Chromosome Name
YPR161C	864449	-1	866422	SGV1	XVI		
YOL138C	61325	-1	65350	RTC1	XV		
YDR395W	1263324	1	1266158	SXM1	IV		
YPR165W	875368	1	875997	RH01	XVI		
YPL015C	525810	-1	526883	HST2	XVI		
YCL050C	37836	-1	38801	APA1	III		
YMR193W	650036	1	650812	MRPL24	XIII		
YER087C-B		332582	-1	332830	SBH1	V	
YGR097W	678695	1	682135	ASK10	VII		
YKL025C	390240	-1	392279	PAN3	XI		
YKR011C	461635	-1	462696		XI		
YEL004W	146951	1	147979	YEA4	V		

Figure 6.6: Head of *geneinfo.txt*

6.3 Connecting the SNP data to the gene information

Now that all gene information is obtained, this information can be connected to the SNP data available. A first script, *attachidtosnpdata.py* uses *idtoname.txt* to attach the Ensembl gene identifier to the SNP data. The script loops through all lines in the SNP file, searches for the gene in the list of identifiers and gene names, and once a match is obtained, the Ensembl gene identifier is attached to the line. If there is no match, the id field is left empty. It is now straightforward to connect the list with gene information to the list of the SNP data based on the Ensembl gene id, as can be seen in figure 6.7. This is done with the script *connectgenesnp.py*. For 42 SNPs there is no matching identifier so these are put on a separate list for manual control later on. The other 5611 SNPs are all connected to the gene information. As the two lists are just connected together, every line contains two times the chromosome number and the gene name, which can be seen in figure 6.7. As an extra control, it is checked with *snpgenecheck.py* whether the two chromosomes are the same and whether the field gene corresponds to either the Ensembl gene identifier or the standard gene name. It appears that for 1 SNP two different chromosomes are listed, so this is saved for later investigation in *failedcheck.txt*.

Summarizing, the SNPs are divided into three lists: one that contains 5610 SNPs that showed no problem processing, one containing 1 SNP that listed two different chromosomes and one consisting of 42 SNPs of which the gene was not recognized by BioMart. These last two lists are processed manually in order not to lose any information:

failedcheck.txt

The SNP file listed a SNP with IMP2 as gene name - a gene located on chromosome XIII - at position 54334 on chromosome IX. By navigating to this position in the UCSC Genome Browser [36], the gene at this location is identified as IMP2' [41]. This is adjusted and manually added to *snpgenefinal.txt*.

mansnpgene.txt

For the 42 SNPs in genes for which no Ensembl gene identifier was found, the name of the gene could also be retrieved on the UCSC Genome Browser [36] based on the chromosome and its position. This gene name can then be converted to its Ensembl gene identifier by the BioMart ID conversion tool [32] or by searching the gene on the

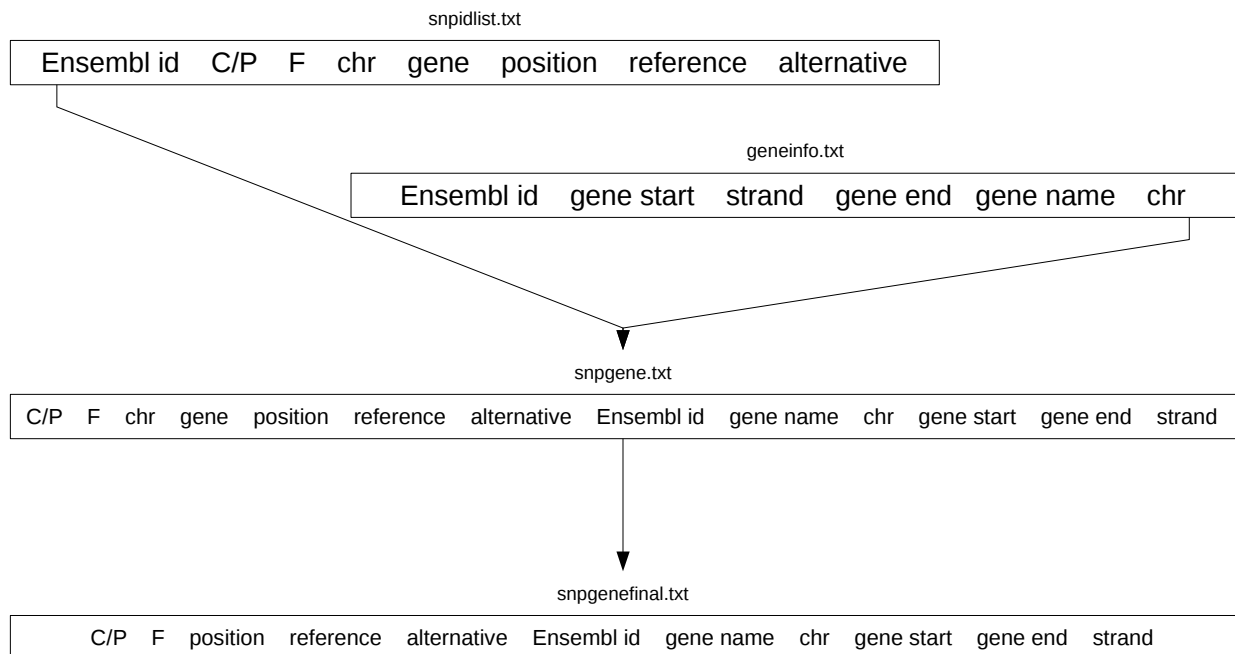


Figure 6.7: How the gene information is attached to the SNP data

CP	FERM	POS	REF	ALT	ENSEMBL	GENE	CHR	START	END	STRAND
C	F3	103484	G	T	YIL130W	ASG1	chrIX	102782	105676	1
C	F3	129020	C	A	YPL223C	GRE1	chrXVI	128632	129138	-1
C	F3	148364	C	A	YPL215W	CBP3	chrXVI	147416	148423	1
C	F3	196710	T	G	YPL184C	MRN1	chrXVI	195950	197788	-1
C	F3	209604	G	A	YDL140C	RP021	chrIV	205360	210561	-1
C	F3	240185	C	A	YML016C	PPZ1	chrXIII	239458	241536	-1
C	F3	243501	C	T	YNL216W	RAP1	chrXIV	241689	244172	1
C	F3	26801	G	C	YEL066W	HPA3	chrV	26667	27206	1
C	F3	269744	G	C	YER056C-A	RPL34A	chrV	269423	270185	-1

Figure 6.8: Head of the *snpgenefinal.txt*

Saccharomyces Genome Database, SGD [34]. Three reasons why the gene could not be recognized by BioMart are recognized:

1. Not the whole gene name is given.
2. *10/01/14* was used instead of *OCT1*, which can be explained by the fact that Excel recognized *OCT1* as a date. This default conversion to the date format is a common problem in bioinformatics, and one should pay attention because this conversion is irreversible [42].
3. An alias is used instead of the standard name.

The concrete changes that are applied are listed in table 6.1 and all SNPs could be added to the final list. The head of the list is shown in figure 6.8.

SNP line number	Gene (SNP data)	Standard gene name
1	tK	tK(CUU)G1
2,4,5	KEM1	XRN1
3	tS	tS(GCU)O
6,7,8,9,38	10/01/14	OCT1
10	tD	tD(GUC)J1
11	tI	tI(AAU)I2
12	tQ	tQ(UUG)E1
13,14,15	ARG5	ARG5,6
16	tV	tV(AAC)E1
17	tV	tV(UAC)D
18	tD	tD(GUC)G1
19,39	tI	tI(AAU)E2
20,21,40	ADE5	ADE5,7
22	tF	tF(GAA)P2
23	tH	tH(GUG)H
24	tN	tN(GUU)N2
25,26,27,28,29,30,31,41	DUR1	DUR1,2
32	tA	tA(AGC)L
33	tT	tT(UGU)G1
34	tI	tI(AAU)D
35	tS	tS(GCU)L
36	tL	TL(UAG)J
37	tL	TL(GAG)G

Table 6.1: Correct gene name for SNPs of which the gene was not recognized by BioMart

Chapter 7

Connecting the SNPs to the interacting amino acids

Now that all information is gathered in a useful way, the SNPs can be connected to the interacting sites of the proteins. To do so, all the data is first sorted per chromosome. Although this is not really necessary, it helps to oversee the big amount of data. After this sorting and some additional filtering, the genes containing introns are separated from the rest of the data. Then, the amino acid number can be calculated, automatically for the genes that do not contain introns, and manually for intron-containing genes. This combination of gene Ensembl id and amino acid number can be matched to the list of interacting amino acid. Finally, for all matches the new - mutated - amino acid is determined. This is all described in the remainder of this chapter.

7.1 Sorting and filtering the data per chromosome

To get a better hold on the SNP data and the amino acid data, first the SNP data is separated based on chromosome number: the script *sortsn.py* loops through every line of *snpgenefinal.txt* and reads on the seventh position of this line the chromosome number. The line is then written to a file *SNPchrK.txt* where K represents any of the possible chromosomes (I-XVI). In case something goes wrong and the chromosome number is not recognized, the line is written to *SNPwrong.txt*, but this turns out to be an empty file. With the help of Excel, a list of unique Ensembl identifiers per chromosome is obtained. These lists are concatenated together with the chromosome the gene is located in, in *idchr.txt*. A summary of this is shown in table 7.1.

The amino acid data is sorted in a similar way, except that some extra sorting is applied: before adding the line to the file of the correct chromosome, it is checked whether the Ensembl identifier is also in the list of Ensembl ids of the SNPs. After all, it is of no use to keep interacting data about genes without an observed SNP. This results in 16 files, *IAAchrK.txt*, where K again represents any of the possible chromosome numbers. This result is summarized in table 7.2. For now, there are 172 genes that could be affected by mutations in the interacting site of the protein.

Chromosome	Number of SNPs	Number of genes
I	112	58
II	421	218
III	162	77
IV	835	432
V	263	160
VI	161	72
VII	511	271
VIII	230	127
IX	187	108
X	347	185
XI	282	163
XII	505	260
XIII	379	214
XIV	367	185
XV	504	269
XVI	387	217
TOTAL	5653	3016

Table 7.1: Number of SNPs and unique genes per chromosome

7.2 Separating the SNPs located in intron-containing genes

The calculation of the amino acid number is straightforward when there are no introns in the gene. Although the number of intron-containing genes is small in *S. cerevisiae*, it is best to separate the intron-containing genes from the other genes in order not to lose any information. According to Parenteau e.a. there are 283 intron-containing genes [43]. Jason E Stajich e.a. describe the intron density as 0.05/gene [44]. When assuming that *S. cerevisiae* contains about 6000 genes [43], there should be 300 introns. It is also possible to request YeastMine for a list containing all Ensembl identifiers of genes with introns [45]. This list consists of 281 identifiers. Although all those number are slightly different, it is decided to go on with the list of SGD and YeastMine for two reasons: the list is very easily obtained and SGD is assumed the general standard of genome data of *S. cerevisiae*. This list is called *introngenes.txt*.

A script called *usefulsnpselection.py* now goes through all the SNP files (1 per chromosome, as explained in 7.1) and does the following:

1. Per line, it is checked whether there is structural data available for this gene. This is done by matching the Ensembl identifier with the list of 172 genes.
2. If there is no structural data, the line is written to *othersnps.txt*. If there is structural data, it is checked whether the gene contains an intron. This is done by matching the Ensembl identifier with the list of 281 intron-containing genes.
3. If the gene contains an intron, the line is written to a file called *manualsnps.txt*. If the

Chromosome	Number of iAAs	Number of genes
I	135	2
II	3865	11
III	674	6
IV	10774	26
V	3504	8
VI	7666	8
VII	14879	18
VIII	1859	9
IX	1389	9
X	4591	13
XI	487	5
XII	1710	12
XIII	9224	10
XIV	660	5
XV	21252	20
XVI	8165	10
No matching id	114494	
TOTAL	205328	172

Table 7.2: Number of interacting amino acids and unique genes per chromosome that are also included in the SNP data

gene does not contain an intron, it is written to *USNPchrK.txt* where K represents any of the possible chromosome numbers.

This results in 293 SNPs that are located in a gene of which there is interaction data, 26 SNPs that are located in an intron-containing gene of which there is interaction data and 5334 SNPs of which there is no interaction data available in the PDB.

7.3 Calculation of the amino acid number

For the 293 SNPs that are located in a gene that does not contain introns and for which there is structural data available, the amino acid number can now be calculated. This is done using the following algorithm:

1. Check whether the gene is located on the forward or the backward strand.
2. If the gene is located on the forward strand, calculate x and r as follows:

$$x = \left\lfloor \frac{SNP\ position - start\ of\ gene + 1}{3} \right\rfloor \quad (7.1)$$

$$r = (SNP\ position - start\ of\ gene + 1) \mod 3 \quad (7.2)$$

If the gene is located on the backward strand, then calculate x and r in this way:

$$x = \left\lfloor \frac{\text{end of gene} - \text{SNP position} + 1}{3} \right\rfloor \quad (7.3)$$

$$r = (\text{end of gene} - \text{SNP position} + 1) \bmod 3 \quad (7.4)$$

The basic idea is to get the amino acid number by calculating how far away the SNP position is from the beginning of the gene and to divide this by three. The remainder of this division can then be used to determine the exact position in the amino acid.

3. When $r = 0$:

$$\text{amino acid number} = x \quad (7.5)$$

else:

$$\text{amino acid number} = x + 1 \quad (7.6)$$

This is exactly what *snpaanumber.py* does: it goes over all 293 SNPs, calculates the amino acid number and the remainder, and concatenates those numbers to every line.

7.4 Matching the SNPs to the structural data

It can be tried now to link all 293 SNPs with interacting amino acids based on Ensembl identifier and amino acid number. The script *matchaasnp.py* checks for every chromosome whether there are lines in the SNP data and lines in the interacting amino acid data that contain the same Ensembl id and amino acid number. If that is the case, all information in both lines is attached next to each other and written to a new file, *matches.txt*. This file shows 192 matches, of which there are 32 unique SNPs: some genes are found in multiple PDB files, so their interacting residues appear more than once in the interacting amino acid data.

However, there are still 26 SNPs left that need manual checking because they are located in an intron-containing gene. The calculations of the amino acid number are similar as above, except that the length of the intron should also be extracted from the nominator of equations 7.1 to 7.4 in case the SNP is located after the intron. Information about the position and lengths of the introns is found on the SGD-page of every gene [34]. Given the chromosome, Ensembl identifier and amino acid number, a match with the interacting amino acids is searched using Excel. Only 5 more SNPs could be added to the 32 unique SNPs, leading to 37 unique SNPs in total. The other 21 SNPs are either located in the intron, or not at a (known) interacting site, see results in table 7.3.

SNP position	Number of SNPs
in intron	9
at interacting site	5
not at interacting site	12
TOTAL	26

Table 7.3: Position of SNPs in intron-containing genes

7.5 Determination of the new amino acid

For the 37 SNPs left, the amino acid to which the residue is mutated can be determined. This is done in the following manner:

1. Look at the remainder that was previously determined in equations 7.2 and 7.4. There are three possible cases:
 - $r = 0$ The SNP can be located as the last base pair of the amino acid number. This means that the amino acid number should just be multiplied by 3 to obtain the base pair position of the SNP.
 - $r = 1$ The first base pair of the amino acid number is the one that is mutated. The exact position can be obtained by subtracting one from the amino acid number, multiplying this by three and take the next position, which is the first base pair of the residue of interest.
 - $r = 2$ When the remainder equals two, the base pair position is located in the middle of the residue. By subtracting one from the amino acid number and multiplying this by three and add two, the exact position of the SNP is obtained.
2. Take the three correct base pairs that together make up for the residue in which the SNP is located by going to the SGD page of the gene and looking at the coding DNA strand.
3. Check, to be sure, whether the translated codon corresponds to the amino acid in the PDB file.
4. Apply the mutation by switching the correct position to the mutated base pair in the forward strand, or by switching the correct position to the inverted mutated base pair in the backward strand.
5. Determine the new amino acid using the DNA codon table.

The SNPs can now be categorized in four groups:

- A group of 4 mutations where the PDB file does not apply the same numbering as expected (meaning that the first residue of the coding sequence is not numbered as being one). In this case, the calculated numbering of the SNP must be rescaled to the numbering of the PDB file and might not obtain overlap anymore with the data set of interacting amino acids.

- A group of 4 nonsense mutations, where the residue is replaced by a STOP codon, which finishes the translation too early and most likely results in a short, non-functional peptide [46].
- A group of 6 synonymous, or silent, mutations where the mutated residue is the same as the original one and the protein is not changed [47].
- 23 SNPs that do result in a change of amino acid, which changes the protein and could potentially change the interaction with another protein.

The results of this process are now summarized in table 7.4. For every SNP, the Ensembl identifier as well as the amino acid number in which the mutation takes place and the remainder are shown in bold. Beneath, the three base pair positions of the amino acid of interest are shown, with in italic the position that is mutated. The third line then shows the original base pairs and the mutated combination. The original and mutated residue are displayed in the last line, calculated by using the codon table.

YHR013C	AA=204	r=2	YDL140C	AA=320	r=1	YPR189W	AA=957	r=2
610	<i>611</i>	612	<i>958</i>	959	960	2869	<i>2870</i>	2871
TTC	->	TAC	CGT	->	TGT	ACT	->	AAT
PHE	->	TYR	ARG	->	CYS	THR	->	ASN
YCR035C	AA=218	r=1	YDR529C	AA=43	r=2	YPR191W	AA=22	r=0
<i>652</i>	653	654	127	<i>128</i>	129	64	65	<i>66</i>
GTA	->	ATA	CTC	->	CAC	AGA	->	AGG
VAL	->	ILE	LEU	->	HIS	ARG	->	ARG
YER018C	AA=180	r=2	YDL140C	AA=817	r=1	YNR052C	AA=238	r=0
538	<i>539</i>	540	<i>2449</i>	2450	2451	712	713	<i>714</i>
GTG	->	GGG	GCA	->	ATA	GAC	->	GAT
ALA	->	GLY	ALA	->	ILE	ASP	->	ASP
YER018C	AA=135	r=2	YDR372C	AA=176	r=1	YOR157C	AA=213	r=1
403	<i>404</i>	405	<i>526</i>	527	528	<i>637</i>	638	639
GCC	->	GTC	CAA	->	TAA	GCA	->	ALA
ALA	->	VAL	GLN	->	STOP	ALA	->	THR
YLL039C	AA=28	r=0	YJL112W	AA=260	r=2	YOR261C	AA=227	r=2
82	83	<i>84</i>	778	<i>779</i>	780	679	<i>680</i>	681
TCG	->	TCA	AAT	->	ATT	GGC	->	GAC
SER	->	SER	ASN	->	ILE	GLY	->	ASP
YLR181C	AA=150	r=1	YJR123W	AA=187	r=0	YOR290C	AA=619	r=0
<i>448</i>	449	450	559	560	<i>561</i>	1855	1856	<i>1857</i>
CGA	->	TGA	ATC	->	ATT	AGA	->	AGG
ARG	->	STOP	ILE	->	ILE	ARG	->	ARG
YLL042C	AA=41	r=1	YJL008C	AA=279	r=2	YOR332W	AA=73	r=1
<i>121</i>	122	123	835	<i>836</i>	837	<i>217</i>	218	219
CGT	->	AGT	GCC	->	GAC	CAA	->	TAA
ARG	->	SER	GLY	->	ASP	GLN	->	STOP
YBR130C	AA=106	r=2	YKL210W	AA=910	r=0	YGR183C	AA=25	r=1
316	<i>317</i>	318	2728	2729	<i>2730</i>	<i>73</i>	74	75
AAC	->	ACC	GAA	->	GAT	GCC	->	ACC
ASN	->	THR	GLU	->	ASP	ALA	->	THR
YBR154C	AA=212	r=2	YKL210W	AA=910	r=0	YIL018W	AA=184	r=2
634	<i>635</i>	636	2728	2729	<i>2730</i>	550	<i>551</i>	552
AGA	->	AAA	GAA	->	GAT	CGT	->	CAT
ARG	->	LYS	GLU	->	ASP	ARG	->	HIS
YBR154C	AA=212	r=2	YKL206C	AA=46	r=1	YLR199C	AA=199	r=1
634	<i>635</i>	636	<i>136</i>	137	138	<i>595</i>	596	597
AGA	->	AAA	GGA	->	TGA	AGT	->	TGT
ARG	->	LYS	GLY	->	STOP	SER	->	CYS
YMR293C	AA=240	r=1	YPL015C	AA=193	r=0	YOR293W	AA=58	r=2
<i>718</i>	719	720	577	578	<i>579</i>	172	<i>173</i>	174
GTA	->	AGA	TCG	->	TCA	CAA	->	CTA
VAL	->	ARG	SER	->	SER	GLN	->	LEU

Table 7.4: The mutated positions: First line consists always of the Ensembl identifier, the amino acid number and the remainder; second line shows the base pair positions of which the mutated one is in italic; third line indicates the original and the mutated basepairs; fourth line then shows the original and the mutated residue.

Chapter 8

Automatization of the process

It could come in handy to have some kind of tool that would perform the actions explained in the previous chapters automatically, i.e. a tool that requires as input a SNP and returns the mutated amino acid together with the residues of all genes it interacts with. This is now possible by combining a few scripts together with some new code and is explained in this chapter.

Four input arguments are passed on to the script: the first one is the gene name or Ensembl identifier of the gene in which the SNP is located, the second argument is the base pair position of the SNP, the third argument is the reference base pair and the last argument is the alternative base pair. First, all gene information needed is gathered. Secondly, the amino acid number is calculated. This is then used to get the original base pairs and the mutated base pairs. Finally, the amino acid number and Ensembl identifier are also used to find all PDB files containing this gene and to calculate all residues the mutated residue is interacting with.

8.1 Obtaining gene information

As explained before in chapter 6, the BioMart browse data tool can be used to obtain gene information [40]. However, another tool of BioMart, called sequence retrieval, also allows to obtain the coding sequence, which will be useful for later [48]. Therefore, from the Ensemble database of *S. cerevisiae* genes, all coding sequences are obtained with the following header information: Ensembl gene id, associated gene name, chromosome, gene start, gene end, strand and gene description. All the information is downloaded from BioMart and stored in the file *geneinfoseq.txt*. This file is stored in the same folder as the script.

As just explained, the script requires four input arguments. When a different number of arguments is passed on, the script stops running and gives the message that exactly four arguments are needed. The first argument is either the gene name or the Ensembl identifier and is used to obtain the necessary gene information: the script loops through *geneinfoseq.txt* and when the first argument matches either the gene name or the Ensembl identifier, all gene information is stored into variables and the file is closed. Matching the first argument happens in a case-insensitive way. When the first argument is not recognized after looping through the whole file, the script stops running with the message that the gene name was not recognized.

8.2 Calculation of the amino acid number

Before proceeding, it is now checked whether the second argument - the position of the SNP - is an integer and whether the position of the SNP is really between the first and the last base pair of the gene. If this is not the case, the program exits giving a message related to the occurred error.

The next thing that is checked is whether the gene contains introns. In chapter 7, a list of Ensembl identifiers of all genes containing an intron obtained from YeastMine is used [45]. However, it is also possible to obtain for every intron the Ensembl identifier, its starting position and its ending position. This is very useful for calculation of the amino acid number. Therefore, this list is downloaded and saved as *introns.txt*. This list contains 311 introns. As there are only 281 intron-containing genes, some genes must contain more than 1 intron.

The script loops through the intron list, and every time it finds an intron that is located in the gene of interest, it is checked where the SNP is located compared to the intron. If the SNP is located in the intron, the script stops running while telling the user that the SNP is located inside an intron. When the gene is located on the forward (reverse) strand, the intron length should only be taken into account when the SNP is located after (before) the intron. This is checked and when the length of the intron should be taken into account for calculation of the amino acid number, the length is added to the variable *intronlength*. This means that when there are multiple introns inside the gene of interest, the variable *intronlength* is the sum of the lengths of all introns that are located before the SNP (seen in the direction of transcription).

After this, part of the code of *snpaanumber.py* of chapter 7 is used, with one slight change: when calculating the amino acid number, *intronlength* is also subtracted from the nominator of equations 7.1 to 7.4. Again, both the amino acid number and the remainder are stored.

Because the complete coding sequence is available, the original base pairs can be obtained from the sequence by selecting their positions based on calculations explained in chapter 7.5. By using the file *dnacodontable.txt* the original amino acid can also be determined: the file consists of 2 columns, the first one listing all codons and the second one all corresponding amino acids. The script loops through the file until it finds the correct codon and the amino acid on the same row is then outputted as the original amino acid.

Based on the third and fourth input argument converted to their upper case character - the reference and the alternative - the mutated codon and amino acid can be calculated. Based on the remainder that was calculated together with the amino acid number, it is possible to know which of the three base pairs is the mutated one. The procedure is dependent on whether the gene is located on the forward or the backward strand. When the gene is located on the forward strand, it is first checked whether or not both the reference and the alternative are one of the four base pairs. If not, the script stops running giving an indication that reference or alternative is not correct. If they both are, it is checked if the base pair that is obtained from the available sequence information corresponds to the input reference. If this is not the case, the program will stop running indicating to the user that the reference does not correspond to the available information. If it does correspond, the reference base pair in the original codon is replaced by the alternative and using this mutated codon and *dnacodontable.txt*, the mutated amino acid can be retrieved.

The procedure for the backward strand is actually the same, but before starting the reference and the alternative are reverted ($A \leftrightarrow T$, $C \leftrightarrow G$).

A result file is created and some basic information is already written to it. The result file is named as *[ensembl id]_[SNP position]_[reference]_[alternative]*. The first, second and third line contain some basic information about the gene and the SNP respectively. The fourth line indicates whether or not the gene contains introns. The following 5 lines give respectively the affected amino acid number, the original amino acid, the original base pairs, the mutated base pairs and the mutated amino acid.

8.3 Connecting the SNP to the available structural information

After this, it can be investigated if the SNP is located at an interaction site. First, all PDB files that have the amino acid of interest at an interacting site are determined. This is done by going through the list *interactingaagene.txt*. As explained before in chapter 5.3, this list consists of 5 columns: PDB identifier, chain identifier, amino acid, amino acid number and Ensembl gene identifier. A loop goes through every line and for every combination of Ensembl id and amino acid number that corresponds to the Ensembl id and amino acid number of the SNP, both the PDB file and the chain identifier are stored into a new list. If this list is empty at the end of the loop, the last line of the result file says that the SNP is not located at an interacting site and the script finishes. Otherwise, a new loop is started that runs *findiaaofaa.py* (explained in chapter 9) for every combination of PDB id and chain id. This script requires the amino acid number next to the PDB id and the chain id. Of course this number is just the amino acid number that was previously calculated. The first loop of this script goes through *chaindilist.txt* to find all chain identifiers in the PDB file of interest that do belong to *S. cerevisiae*. After that, a new loop goes through all atoms of the current PDB file to make a list of all atoms that do not belong to the same chain as the requested one. Also, a list containing the coordinates of all atoms of the amino acid of interest are stored. Next, all distances are calculated and when the distance is small enough, the residue, its number, the chain and the euclidean distance can be stored. The only difference made to the code of *findiaaofaa.py* is that instead of the chain identifier the Ensembl identifier of every gene is listed. Also, a second list of all Ensembl identifiers together with the gene name and description of all those interacting genes is kept. The results are then written to the result file in the following way: first the result file indicates from what PDB file the structural information is obtained and what chain is considered as the chain of interest. Also, a list of all residues that are interacting with the amino acid of interest is printed, giving the following information: Ensembl id of the gene, the amino acid, the residue number and the minimal distance to the amino acid of interest. Finally, a list of all the interacting genes is printed with the ensembl identifier and a gene description as extra information.

A snapshot of a result file is shown in figure 8.1.

```

GENE INFO - Ensembl: YLL042C, gene name: ATG10, chrXII, gene start: 52087, gene end: 52590, strand: -1
GENE DESCRIPTION: Conserved E2-like conjugating enzyme; mediates formation of the Atg12p-Atg5p
conjugate, which is a critical step in autophagy [Source:SGD;Acc:S000003965]
SNP INFO - position: 52470, reference: G, alternative: T
Intron containing gene: 0
Affected amino acid number: 41
Original amino acid: ARG
Original base pairs: CGT
Mutated base pairs: AGT
Mutated amino acid: SER

INTERACTING AMINO ACIDS
PDB file 1 - 4gsk, chain Z:
Ensembl id      amino acid      residue number  distance
YHR171W PHE      16
YHR171W LEU      17
YHR171W ASP      18
YHR171W PHE      61

PDB file 2 - 4gsk, chain Y:
Ensembl id      amino acid      residue number  distance
YHR171W PHE      16
YHR171W LEU      17
YHR171W ASP      18

INTERACTING GENE INFO:
YHR171W ATG7      Autophagy-related protein and dual specificity member of the E1 family; mediates the
attachment of Atg12p to Atg5p and Atg8p to phosphatidylethanolamine which are required steps in
autophagosome formation; E1 enzymes are also known as ubiquitin-activating enzymes; involved in
methionine restriction extension of chronological lifespan in an autophagy-dependent manner
[Source:SGD;Acc:S000001214]

```

Figure 8.1: Resulting text file *YLL042C_52470_G_T.txt* when running *python auto.py atg10 52470 G T*

8.4 Updating of the available information

To end this chapter, one remark needs to be made. This script works very fast, but it has one big drawback: the information used for all calculations is only the information that is available at the moment of writing. In some time, when there are more structures available on the PDB, or new genes have been discovered, this information will not be complete anymore. There are two solutions for this. The first one would be to update every once in a while all files that need an update (*chainidlist*, *geneinfoseq.txt*, *interactingaagene.txt* and *introns.txt*). A second solution would be to create these files in real time when the script is run. This however is not really interesting as it increases the calculation time tremendously (about a week is needed to obtain *interactingaagene.txt*).

Part III

Results

Chapter 9

Overview of the results

After determining which SNPs are located at the interface of a PPI, 33 SNPs are left: 4 nonsense mutations, 6 synonymous mutations and 23 non-synonymous mutations. The effect of those mutations on the interactions is described in the two following chapters.

Chapter 10 describes the effect of the nonsense mutations and the synonymous mutations. As mentioned before, a nonsense mutation means that a residue is replaced by a STOP codon, resulting in a too short and most likely non-functional peptide [46]. Therefore, it is useful to not only investigate the function of the complex or interaction, but also of the individual protein. After that, it is explained how the effect of the silent mutations is a lot harder to predict as it is not yet completely known how such a mutation exactly affects protein folding and expression levels. Therefore, the proteins and complexes in which the 6 synonymous mutations occur will just be described, without making any implications on the effect.

Chapter 11 deals with the most interesting SNPs and their possible effects. Each of the 23 non-synonymous mutations is described in depth, as well as their possible effect on the complex or pathway. Also, predictions made by both PoPMuSiC and the SNP-IN tool are calculated to compare their results to an opinion based on literature [25] [28].

A couple of comments should be made about processing the results. First, different sources are used to get the necessary information, of which PDB [29], PubMed [49], SGD [34], STRING [50] and KEGG pathways [51] are the most consulted. Also, when visualization of the interaction is necessary, PyMOL is used [52]. This is an open source program that allows molecular visualization [52] and is chosen over other programs because it is free and previously used in other courses. In case of a non-synonymous mutation, the BLOSUM value is mentioned. The BLOSUM matrix is a substitution matrix that is used for measuring the similarity between two protein sequences. Every possible mutation is given a score based on the logodds ratio of the observed and expected probabilities of occurrence of each pair. The logodds ratio is then multiplied by some scaling factor and rounded to the nearest integer. Sequences are clustered by blocks based on a % of similarity and the observed substitutions between blocks are used to calculate a BLOSUM matrix that is named after the similarity in the blocks [53]. A whole family of BLOSUM matrices is obtained, but the standard one, which is also the one used in BLAST [54], is BLOSUM62, which is based on blocks with sequences of at least 62% identity [53]. When the observed frequency is as expected for an amino acid pair, its corresponding value is 0. When the

observed frequency is less (more) than expected, its corresponding value is lower (higher) than 0 [53]. Although here no sequences are compared, it could be useful to look at the BLOSUM value corresponding to a non-synonymous mutation: a very frequent mutation is less likely to affect the interaction compared to a rare mutation. Finally, in order to have a correct idea of the interactions taking place, some more information is useful. Therefore, two more scripts have been written:

findiaaofaa.py

This script will return a list of all amino acids that a particular residue is interacting with. When running the script, three arguments are required: identifier of the PDB file, chain identifier and number of the residue of interest. Two lists are created out of the PDB file: one that stores all the atoms of the residue of interest, and one that stores all other atoms. Then a loop is created that calculates the distance between any two atoms of the two different lists. When the euclidean distance is smaller than 5 Å, the residue information should be written to a new list. This is done by first checking whether this residue is already present in this list. If it is not, the following information is stored: chain identifier, residue number, residue and euclidean distance. If it is, the euclidean distance just calculated is compared to the euclidean distance present in the list and the smallest one is stored in the list. When all calculations are performed, the resulting list is written to the following file: *[PDB identifier]_[chain identifier]_[residue number].txt*.

findiaaofch.py

When two mutations occur in the interaction between two proteins, it is interesting to know whether they do belong to the same interacting surface or not. If they do, the two mutations have a combined effect and must be investigated together. If they both belong to a different surface and are a bit more separated, the two mutations have no influence on each other and can be investigated separately. Therefore, a script is written that returns a list of all interacting residues of two chains. This can be visualized using PyMOL [52] to see whether they belong to the same interface or not. The script requires three arguments: the PDB identifier and the two chain identifiers. In a first step, two lists are built: one containing all the atoms of the first chain and one containing all the atoms of the second chain. A loop then calculates euclidean distances between any two atoms of the two different lists, and when this distance is smaller than 5 Å, the chain identifier, residue number and residue type of both atoms are saved. Before writing the list to a file, all duplicate lines are removed so every amino acid occurs only once. The list is then written to a file called *[PDB identifier]_[first chain identifier]_[second chain identifier].txt*.

Chapter 10

The effect of the nonsense and silent mutations

10.1 Nonsense mutations

All nonsense mutations are observed as heterozygous mutations in the clones at 200 generations. One mutation is found in the clones of fermentor 4 and the other three in the clones of fermentor 6.

10.1.1 Fermentor 4

YDR372C (VPS74); chrIV 1221624; Q176STOP

Vps74 is a vacuolar protein sorting protein and interacts with Sac1 [55]. Structural information of this interaction can be retrieved from the PDB with identifier 4TU3. Although a decreased resistance to 8% ethanol is observed, the cell remains viable after deletion of this gene [55] [56]. Another mutant phenotype is a small defect in vacuolar fragmentation [55] [57]. Vps74 regulates Sac1, a phosphatase that hydrolyzes PtdIns4P in the early and medial Golgi apparatus. In the absence of this complex, there will be a higher amount of PtdIns4P in the Golgi apparatus, increasing the probability that this lipid travels to the cell body [58]. This lipid is currently known to act as a signaling molecule, but suggestions have been made that it also acts as inducer and stabilizer of membrane deformations [59].

10.1.2 Fermentor 6

YLR181C (VTA1); chrXII 517223; R150STOP

This nonsense mutation in Vta1 is located at the interaction site with Vps60, a protein involved in vacuolar transport [60], and structural information about this interaction can be found with PDB identifier 2LUH. Vta1 is a multivesicular body protein and a positive regulator of Vps4, which disassembles the ESCRT complexes. These complexes form a machinery that sorts proteins and deforms membranes during processes like cell division, virus budding and multivesicular body formation [57]. The interaction between Vps60 and Vta1 does not result in conformational changes of Vta1, but enhances the function of Vps4 in a way that is not understood yet [61]. Without Vta1, the cell is still viable [62]. It has

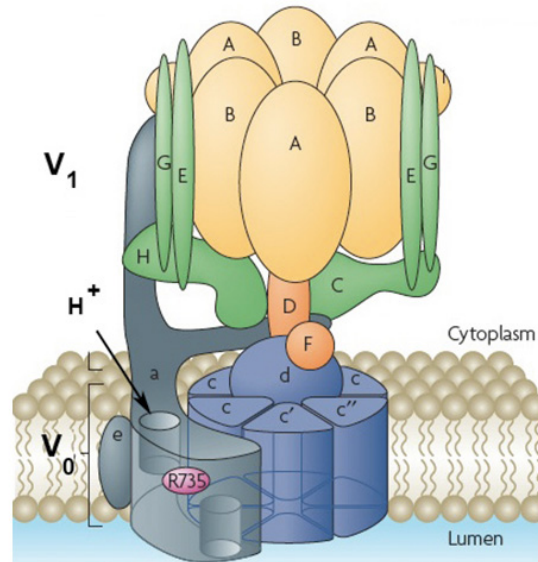


Figure 10.1: Structure of V-ATPase. Vma4 encodes for subunit E of the V1 domain, Vma10 is represented by subunit G of the V1 domain. [66]

been shown that Vta1 mutants show a strong defect in vacuolar fragmentation [57]. This is also a known effect of ethanol, although this is only been investigated during a fluorescence microscopy experiment [63].

YKL206C (ADD66); chrXI 49674; G46STOP

Add66 forms together with Pba1 a heterodimer which is involved in the assembly of the 20S proteasome [64]. Structural information of this heterodimer can be found in PDB file 4G4S. The mutation acts on the 46th residue which interacts with four residues of Pba1. Knock-out of this gene has shown an increased stress resistance and although the knock-out is sensitive to 8% ethanol, it is not sensitive to 5% ethanol [56] [64].

YOR332W (VMA4); chrXV 943872; Q73STOP

In PDB files 4EFA and 4DL0, structural information about Vma4 can be found. Vma4 interacts with Vma10 where the nonsense mutation takes place. The Vma4 gene encodes subunit E of the V1 domain of V-ATPase [65]. The structure of V-ATPase is shown in figure 10.1. As mentioned in chapter 1, this protein pump is very important in ethanol stress because it maintains the intracellular pH [11]. Although a knock-out of this gene results in a cell that is still viable, the resistance to ethanol has been observed to have decreased because of a lack of V-ATPase activity [65].

10.2 Silent mutations

For the longest time, synonymous mutations were ignored in biology because they do not change the amino acid sequence and therefore were expected not to have any effect. However, it turns out that the impact of synonymous mutations on proteins and their structure is a lot bigger than previously assumed. First of all, it is the combination of all base pairs that determines the mRNA structure. Mutating even one base pair can have an impact on this structure. Secondly, there is a bias to use certain codons over other codons, called codon usage bias: every codon coding for the same amino acid appears with a certain frequency, and some codons are a lot rarer than others. There is less tRNA available for a rare codon compared to a common one, and therefore translation occurs slower. Not only the codon itself, also its context is important to translational speed. If there are a lot of positively charged amino acids near a codon, translation is likely to go slower due to interaction with the exit tunnel of the ribosome. Also, the rate of the peptide bond formation is dependent on the previous and the current amino acid. Even more, some parts of proteins are recognized as stalling sequences, where translation pauses [67] [68].

At some places, translation is required to go slow so the peptide can already start folding or some interactions can already start taking place. This is achieved by using a rare codon or by a particular mRNA structure that induces translational pauses. Also, the use of codons with different frequencies is used to regulate protein expression levels: the introduction of rare codon slows down the translation and decreases the expression. Next to that, stability of the mRNA structure also regulates protein expression levels: unstable mRNA is degraded a lot easier than stable mRNA which lowers protein expression [67] [68].

This brief introduction should make it clear that silent mutations can have an effect on protein folding and on expression levels. However, due to the only recent discovery of the effects of synonymous mutations, there is still a lot of research going on in this area and so far, no clear conclusions can be drawn from these findings. Therefore, this section will list the synonymous mutations that have been found in the interaction sites without trying to make any biological implications. Also, in table 10.1, the codon frequencies of the original codon and the mutated codon are listed. These frequencies are obtained from the codon usage table on SGD. There are two tables available, one that only uses known open reading frames and one that also uses predicted open reading frames. It is chosen to work with the most recent one (1999) that only uses known open reading frames to ensure that the data used to build the table is correct, although maybe not complete [69]. On one hand, it can be thought that an increasing frequency results in a faster translation, and thus higher expression, of the protein. On the other hand, this faster translation can result in misfolding of the protein, leading to a non-functional protein. This is the reason that no conclusion can be made.

Gene	Frequency original codon	Frequency mutated codon
Rps5	27%	47%
Qcr2	49%	21%
Snf2	49%	21%
Pop2	35%	65%
Hst2	9%	21%
Ubi4	9%	21%

Table 10.1: Frequencies of the original and the mutated codons

10.2.1 Synonymous mutations in the population of fermentor 6

YNR052C (POP2); chrXIV 719934; D238D

In fermentor 6 of the population data, a mutation in Pop2 has been observed. This mutation has a frequency of 22% and has appeared at 200 generations. On the Protein Data Bank with identifier 4B8C it can be seen that this mutation takes place in the interaction site with another chain of Pop2 on the following residues: alanine at position 173, valine at position 174 and glutamine at position 177. Pop2 is a ribonuclease responsible for RNA degradation and part of the Ccr4-Not complex that performs mRNA deadenylation. The cell is viable without Pop2 but shows a decreased resistance to 6% ethanol [70].

10.2.2 Synonymous mutations in the clones of fermentor 2

YLL039C (UBI4); chrXII 65124; S28S

A heterozygous mutation in Ubi4 at the interaction site with Vps27 is seen at 80 generations, but this mutation is not present anymore at 100 generations. Interacting information is available in PDB file 1Q0W. The residues of Vps27 interacting with Ubi4 at position 28 are: glutamate at positions 261, 268 and 273, isoleucine at positions 263 and 267, arginine at positions 264 and 275, lysine at positions 265 and 272, alanine at positions 266 and 278, leucine at positions 269 and 271, serine at positions 270, 274 and 277 and asparagine at position 276. Ubi4 is also known as ubiquitin and is attached to proteins to indicate that those proteins need to be degraded [71]. This is essential in stress conditions as a lot of proteins will defold and need to be degraded or refolded correctly [72]. Therefore, it is not surprising that an increased expression is observed in stress conditions [71]. The interacting partner of Ubi4, Vps27, is a protein involved in sorting the ubiquitinated proteins that have to be degraded [73].

10.2.3 Synonymous mutations in the clones of fermentor 6

YPR191W (QCR2); chrXVI 919446; R22R

In Qcr2, a heterozygous synonymous mutations is present at 100 generations, but 100 generations later it is not there anymore. Qcr2 is a subunit of the bc1 complex (also called complex III or ubiquinol cytochrome-c reductase). This complex is located in the inner mitochondrial membrane as part of the electron transport chain [74]. The interaction site

in which the mutation takes place, is engaged in interaction with subunit 1 of this complex, Cor1. Structural data is stored in PDB file 4PD4. The specific interacting residues in Cor1 are serine at position 45 and histidine at position 47. Mutants of Qcr2 are still viable but show a decreased resistance to 8% of ethanol [74].

YOR290C (SNF2); chrXV 858402; R619R

Another heterozygous mutation that is present at 100 generations but not anymore at 200 generations is a mutation in Snf2. Snf2 is a subunit of the SWI/SNF complex. This complex is capable of remodeling chromatin and possesses DNA-stimulated ATPase activity. This remodeling allows binding of the transcription factor to the DNA, and therefore it is also involved in regulation of transcription [75] [76]. At position 619, Snf2 interacts with the following residues of Arp7: valine at position 24, glutamate at position 25, leucine at position 26, aspartate at position 431, tryptophan at position 437 and leucine at positions 438 and 441. The interaction can be visualized with PDB file 4I6M. Without this gene, a decreased resistance to 6% of ethanol has been observed, but the cell is still viable [76].

YPL015C (HST2); chrXVI 526305; S193S

The serine residue at position 193 of Hst2 interacts with residues lysine at position 12 and glycine at position 13 of Hhf1. Structural data of this interaction can be found in PDB files 1Q1A, 1SZC, 1SZD, 2QQG and 2QQF. The synonymous mutation at this position has been observed at 100 generations and not at 200 generations. Hst2 is a protein deacetylase and regulates nucleolar and telomeric silencing. Mutant studies have shown that it is a non-essential gene [77].

YJR123W (RPS5); chrX 652461; I187I

A heterozygous mutation is also present in Rps5 at 200 generations. The Rps5 gene encodes for a protein that is part of the small ribosomal unit and is essential for viability [78]. The small ribosomal subunit finds the mRNA that needs to be translated, and after binding with a large subunit, it will make sure that the right tRNA is found for every codon [79]. At this position, there is interaction with certain residues from Rps25a. There is a lot of structural information available for this interaction, because it is part of the ribosome which has been studied in depth. The following PDB identifiers contain information about this interaction: 3U5C, 3U5G, 4UJF, 4UJH, 4UJK, 4UJM, 4UJP, 4UJR, 4UJU, 4UJW, 4UJZ, 4UK1, 4UK4, 4UK6, 4UK9, 4UKB, 4UKF, 4UKH, 4UKK, 4UKM, 4UKP, 4UKR, 4UKU, 4UKW, 4UKZ, 4UL1, 4UL4, 4UL6, 4UL9, 4ULB, 4ULE, 4ULG, 4ULJ, 4ULM, 4ULP and 4ULR. The residues of Rps25a that residue 187 of Rps5 is interacting with, depend slightly on the PDB file: in all files, there is interaction with serine at position 63. All but 4UJW also show interaction with valine at position 66. Then there are four PDB files - 4ULP, 4UK4, 4UL4 and 4UJF - that also show interaction with valine at position 62. The last PDB file, 4UJF, and 4UKP also have aspartate at position 67 as interacting residue. Finally, 4UKM and 4ULM indicate histidine at position 38 to be interacting with Rps5.

Chapter 11

The effect of the non-synonymous mutations

In this chapter, the non-synonymous mutations and their possible effect on PPIs will be described. First of all, the non-synonymous mutations that occur in the whole population are sorted in the order of decreasing frequency: it is expected that the impact of the mutation on ethanol tolerance will be higher when it occurs more frequently in the population. The non-synonymous mutations that occur only in the clones are sorted both on time point and on fermentor. In this way, it can easily be seen which mutations occur together in a cell: if two mutations in the same complex or pathway occur at the same point in time in the same cell, they probably have a combined effect. Both mutations can then be investigated together.

In table 11.1 three predictions are compared: an opinion based on literature (importance of the complex in stress conditions, similarity between the original and the mutated amino acid, ...), a prediction by the SNP-IN tool and a prediction by PoPMuSiC [25] [28]. Only these two algorithms of the four explained in chapter 3.3 are used, because they have an easy-to-use graphical interface that is publicly available, which is not the case for the other two tools. In this table, the first column indicates the gene where the SNP is located in, the second column indicates how many times the chain can be found in the structural information, the third column is a personal opinion based on literature and similarity of the original and the mutated amino acid which can be either preserved (P), not preserved (NP) or not known (P/NP). The two last columns indicate how many times the interaction is preserved (P), not preserved (NP) or how many times results are not available (N/A). In the case of SNP-IN, the lack of results can often be explained by the fact that the PDB file no longer exists because it is merged with other files. However, this does not really make any sense because one of the input arguments is the PDB file itself and not just the PDB id. This problem is absent in PoPMuSiC, but PoPMuSiC always shows problems when there are two chains that have the same id, one lowercase and one uppercase. This indicates that PoPMuSiC is probably a case sensitive tool. An extended table containing all information retrieved from these tools can be found in Appendix A.

A first comment can be made about the opinion which is only based on the information from literature. All these opinions are explained in the remainder of this chapter, but they are still somewhat subjective. Therefore, the column opinion does not state the true effect of the nsSNP on the interaction, but rather the expected effect based on the existing

knowledge. Secondly, when looking at the results, it can already be seen that for both tools, the number of positive predictions (preserved) is much smaller than the number of negative predictions (not-preserved). This is a major drawback of existing prediction tools, as is already explained in chapter 3.3. Also, there is no consistency in the results when there are different PDB files containing the same interaction, which makes it even harder to come at a solid conclusion.

GENE	CHAINS	OPINION	SNP-IN			PoPMuSiC		
			P	NP	N/A	P	NP	N/A
UBA1	2	P	2			2		
RPB5	7	P	7			7		
CCT8	4	P	4			4		
ARD1	3	P	2	1		3		
RRP43	2	P	2			2		
ATG10	2	NP	2			2		
SHE3	2	P/NP	2			2		
MDV1	2	P/NP	2			2		
RPS10A	34	P/NP	34			18	16	
PUP1	75	P	2	55	18	69	6	
RPO21 POS 320	8	P	8			4	4	
SPC25 POS 180	2	P	2			2		
PBA1	1	P	1			1		
SPC25 POS 135	2	P	2			2		
HER2	2	NP	2			2		
QCR7	1	P/NP	1			1		
RPO21 POS 817	7	P	1	6		7		
SKI3	2	P/NP	2			2		
RPN8	2	P	2			2		
QCR9	1	P	1			1		
RPL2B	1	P/NP	1			1		

Table 11.1: Results of prediction tools of the effect of nsSNPs on the interactions (P for preserving and NP for not preserving)

11.1 Non-synonymous mutations in the population of fermentor 6

YKL210W (UBA1); chrXI 41892; E910D

This mutation is present in 47% of the organisms at 100 generations and even in 73% of the organisms at 200 generations. It occurs in the interaction between Uba1 and Ubi4, which is ubiquitin. Structural information of this interaction can be found with PDB identifier 4NNJ, where both chains A and C represent Uba1. Uba1 is the ubiquitin activating enzyme 1 and plays an important role in protein degradation [80]. By activating ubiquitin, it performs the first step in the attachment of ubiquitin to proteins that need to be degraded, which is shown in figure 11.1 [81]. Even more, this protein is essential for viability and has an increased transcription in stress conditions [80]. The reason why the ubiquitination process is so important in stress conditions can be explained as follows: some proteins inside the cell will defold due to the different environment. This causes both loss of the function of that protein, as well as protein aggregation due to the exposed hydrophobic residues that will stick together in an aqueous environment. In order to avoid this aggregation, the denatured proteins have to be degraded as soon as possible or have to be properly folded again [72]. This is why it can be expected that a protein degradation process is highly important in stress conditions.

The amino acid at position 910, glutamate, is mutated to aspartate. This is the most common mutation because the chemical properties of both amino acids are very similar: they are both negatively charged and have the same structure except that glutamate contains an extra methylene group compared to aspartate [82] [83]. However, due to this difference, glutamate is more hydrophobic than aspartate [83]. Also, aspartate is slightly more rigid within a protein structure, and therefore it has a slightly higher propensity to be localized at an interaction site [84]. The residue is supposed to interact with three residues of Ubi4: glycine at position 47, lysine at position 48 and glutamine at position 49. Lysine is a positively charged amino acid and because of the negatively charged glutamate and aspartate, there will be high complementarity between both surfaces, even after mutation. Because of the importance of the ubiquitination process in stress conditions, it is very unlikely that the interaction will be lost after this mutation.

YBR154C (RPB5); chrII 548375; R212K

The second most occurring mutation has a frequency of 46% at 100 generations, a frequency of 49% at 200 generations and is located in Rpb5. Seven PDB files - 3RZD, 3S14, 3S2H, 4A3K, 3RZO, 3S2D and 4A3F - contain structural information of this protein. Position 212 is part of the interaction site with Rpo21, which is the largest subunit of RNA polymerase II [85]. Rpb5 itself is a subunit of all RNA polymerases (I, II and III) and acts as a jaw to contact DNA and activate transcription [86]. An image of RNA polymerase II is shown in figure 11.2 where Rpb5 is subunit ABC1, interacting with Rpo21, also called B1. Although Rpb5 is shared amongst the three RNA polymerases, only the interaction in RNA polymerase II is possibly influenced by this mutation. Half of the transcription done by this polymerase is for ribosomal proteins. Also lots of small nucleolar RNA (snoRNA) and most of the mRNAs are transcribed by polymerase II [86]. Without Rpb5 the cell is inviable.

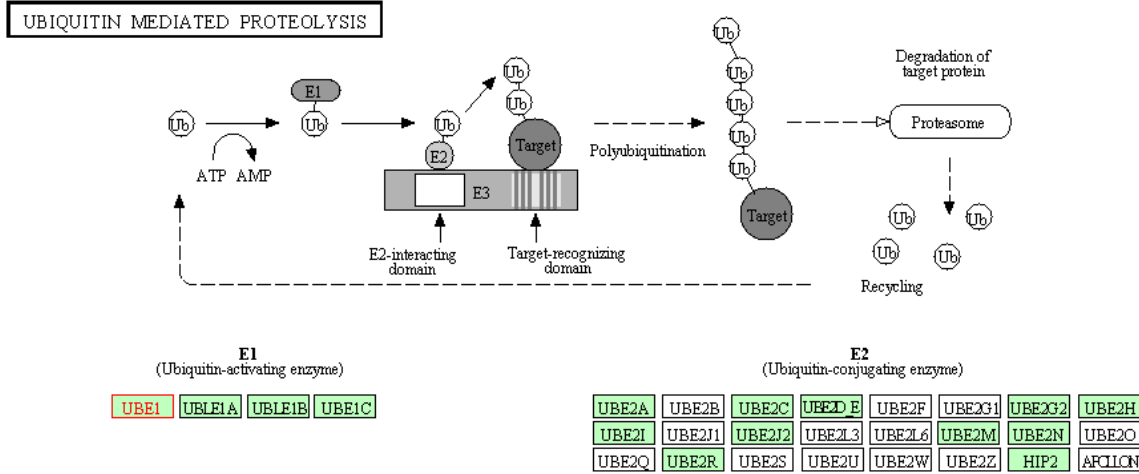


Figure 11.1: Ubiquitination process, UBA1 is indicated in red [51]

The mutation results in a lysine residue instead of an arginine residue. This amino acid interacts with isoleucine at position 1341, glutamate at position 1342, threonine at positions 1376 and 1377, glutamine at position 1378 and glycine at position 1379. The residue the closest to the arginine residue is glutamate, which is negatively charged, while both arginine and lysine are positively charged. A difference between arginine and lysine is the fact that arginine has two positions to interact with, while lysine has only one. This is shown in figure 11.3. The green sticks are the residues of Rpo21 that are interacting with arginine. In figure 11.3a, it can be seen that some of the interacting residues are closer to one end branch of arginine while other interacting residues are closer to the other end branch. Also, it can be seen here that for residue 1378 only the distance from the backbone atom N to an end branch of arginine is smaller than $5\text{\AA}$. In figure 11.3b, the arginine residue is replaced by lysine. Although the atoms of threonine at position 1377, glutamine at position 1378 and glycine at position 1379 are now too far to be interacting with the lysine residue (distances 6.9, 8.1 and $6.7\text{\AA}$ respectively), the interaction with the other residues, especially the negatively charged glutamate, is still kept. Finally, the BLOSUM substitution matrix assigns a positive value to this mutation, indicating that this mutation is more likely to occur. This could indicate that the complex is not affected by the mutation. On the other hand, it is known that there is a decreased transcription of energy-consuming processes in stress conditions, which includes the biosynthesis of ribosomes [9]. This could explain why the interaction with Rpo21 is broken: transcription of unnecessary genes in stress conditions can be avoided by disrupting part of the polymerase that performs transcription. Next to that, RNA polymerase II is responsible for transcription of ribosomal proteins, which are repressed in stress conditions and snoRNAs that all help in ribosome maturation [86]. However, it should be noted that RNA polymerase II is also responsible for transcription of the necessary genes, so a complete disruption of the whole RNA polymerase II complex is impossible [85].

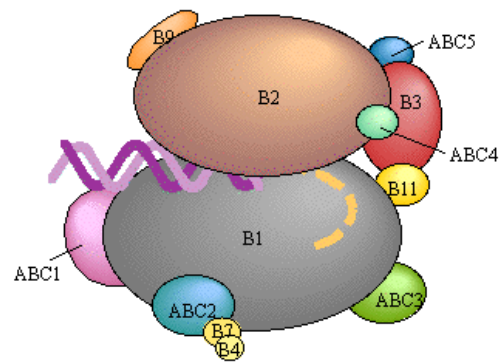
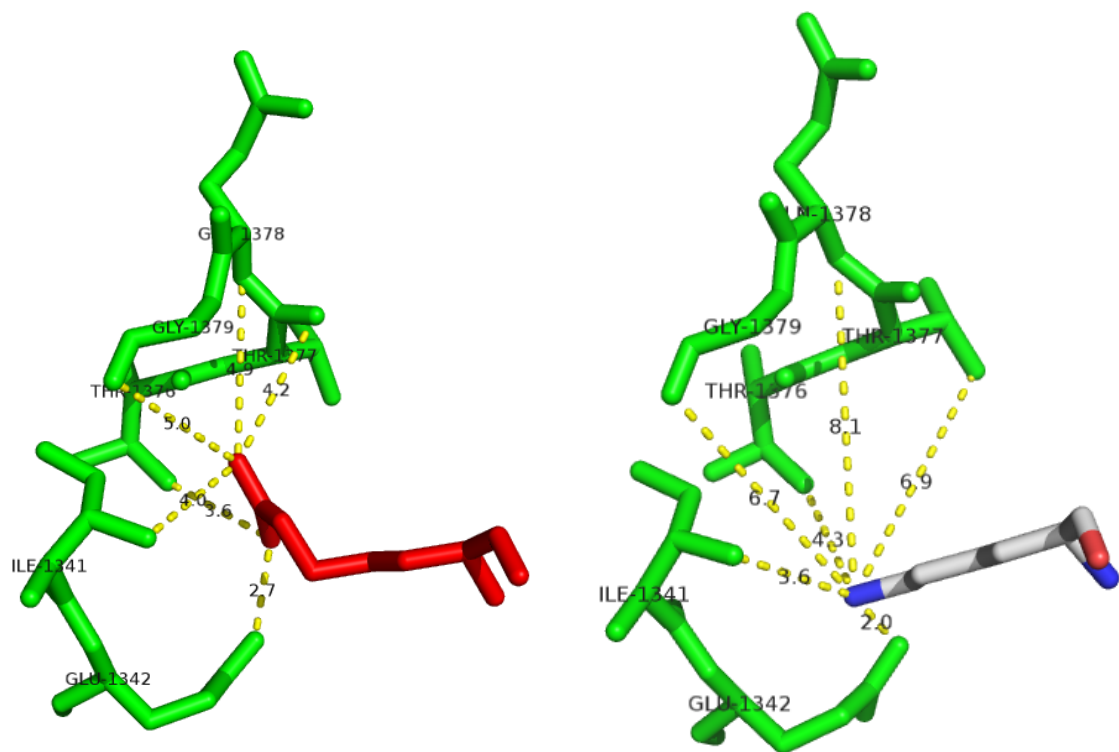


Figure 11.2: RNA polymerase II, with genes Rpb5 (ABC1), Rpo21 (B1) and Rpb2 (B2) [51]



(a) The original arginine (red) and its interacting partners of Rpo21 (green)

(b) The mutated lysine and its interacting partners of Rpo21 (green)

Figure 11.3: Visualization of the mutation in Rpb5 using PDB file 3RZD

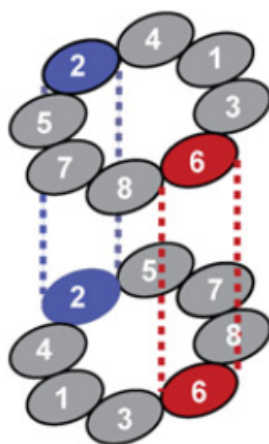


Figure 11.4: Schematic picture of the CCT complex [88]

YJL008C (CCT8); chrX 420828; G279D

The gene *Cct8* contains a mutation with a population frequency of 35% at 200 generations. Two PDB files, 4D8R and 4D8Q, have interaction information about this protein, but these two files are merged on 10/12/2014 into the file 4V94. In this file, there are 4 chains of this protein, all interacting with an alanine residue at position 256 of *Cct6*. *Cct8* is part of the CCT chaperonin, a protein complex consisting of 8 subunits that aids in correct folding of about 10% of the proteome and in assembly of actin and tubulin [87] [88]. The complex consists of two rings, each containing one chain of every subunit. The way that they are linked together, is schematically shown in figure 11.4. This complex is thus essential, especially in stress conditions where there will be an increased amount of denatured proteins [72]. Also other chaperones, like heat shock proteins, know an increased transcription in ethanol stress, as mentioned in chapter 1 [12].

The mutation taking place at position 279, is from glycine to aspartate. While glycine is a non-polar amino acid, aspartate is a positively charged amino acid. Interaction takes place with alanine, also a non-polar amino acid. The BLOSUM matrix assigns a value of -1 to this mutation, indicating that it is less likely to occur. This might indicate that the mutation will cause changes in the interaction of *Cct8* and *Cct6*. However it is expected that, due to the importance of the complex, the mutation will not be deleterious for the interaction.

11.2 Non-synonymous mutations in the clones of fermentor 6 at 100 generations

YKL210W (UBA1); chrXI 41892; E910D

This mutation is already completely explained in section 11.1.

YBR154C (RPB5); chrII 548375; R212K

This mutation is already completely explained in section 11.1.

YHR013C (ARD1); chrVIII 130836; F204Y

The NatA acetyltransferase is an N-terminal acetyltransferase consisting of a catalytic subunit, Ard1, and an auxiliary subunit, Nat1, and can be studied using PDB files 4HNX and 4HNY. Next to NatA, there are two other complexes, NatB and NatC, that also catalyze the reaction of adding an acetyl group to N-terminal residues cotranslationally. An increased expression of these proteins is observed when there is a faster protein synthesis in the cell. This N-terminal acetylation happens for 50% of the yeast proteins and neutralizes positive charges, stability, interactions or other protein modifications that occur later on. Cells lacking either Ard1 or Nat1 show, among other, a slower growth and an increased sensitivity to temperature [89]. Besides that, increased expression levels of Ard1 have been observed in stress conditions [90]. Residue 204, phenylalanine, of Ard1 is interacting with residue 307, arginine, of Nat1. A heterozygous mutation in residue 204 is found in the clone cells of fermentor 6 at 100 generations, but it has disappeared by the time 200 generations have been reached.

Here, a phenylalanine residue is converted into a tyrosine residue. Both are aromatic residues differing only in the fact that tyrosine has an additional OH attached to the fourth carbon of the benzene ring. However, this makes tyrosine less hydrophobic, although it has a higher propensity to be located at the interface. This mutation is rather likely to occur as it is assigned a positive value of 3 in the BLOSUM matrix. Also the increased expression levels of Ard1 in stress conditions indicate the importance of this complex. Therefore it is assumed that the interaction will not be destroyed due to the mutation.

YCR035C (RRP43); chrIII 192367; V218I

At 100 generations, a heterozygous non-synonymous mutation can be found in the Rrp43 gene. The affected residue is interacting with Mtr3. There are two PDB files that give slightly different information about this interaction: while 4IFD indicates interaction of the specific residue with leucine at position 101, 4OO1 rather indicates interaction with residue 100, glycine. Rrp43 is one of the core parts of the exosome that breaks down RNA in the 3'-to-5' direction. A schematic representation of the exosome is shown in figure 11.5. The cell is inviable in case of null mutants of Rrp43 and shows increased expression levels in response to DNA replication stress [91]. Inviability is also obtained in case of null mutants of Mtr3 [92].

The mutation occurring converts a valine residue into an isoleucine residue. Both amino acids are non-polar with isoleucine being slightly more hydrophobic compared to valine.

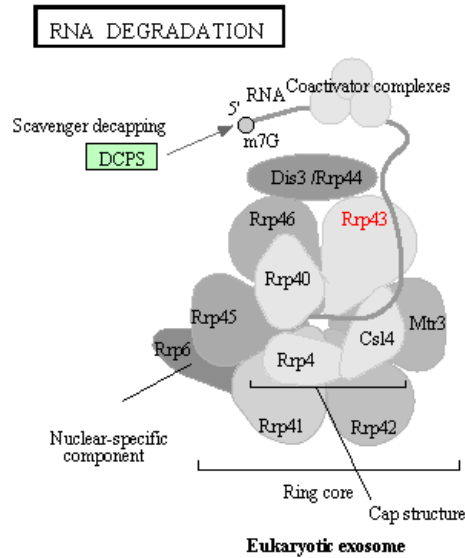


Figure 11.5: Schematic view of the exosome with Rrp43 indicated in red [51]

Also, a value of 3 is assigned to this mutation by BLOSUM. It is expected that the binding will not be broken because of the resemblance between the two amino acids, but also because of the function of the complex. A higher expression level of the Rrp43 gene in stress conditions shows that this complex is important in stress conditions, probably because it is a way to control the expression levels of both necessary and unnecessary genes.

YLL042C (ATG10); chrXII 52470; R41S

Residue 41 of Atg10 interacts with Atg7. PDB file 4GSK contains structural information about two Atg10 chains and two Atg7 chains. In the first chain of Atg10, the arginine residue at position 41 interacts with residues 16, 17 and 18 of one chain of Atg7, which are respectively phenylalanine, leucine and aspartate. In the second chain of Atg10 at position 41, there is interaction observed with all the same residues and with residue 61, phenylalanine, of the other Atg7 chain. This complex is part of the pathway that regulates autophagy [51]. Autophagy is a highly conserved pathway that brings part of the cytoplasm, like unnecessary proteins or organelles, to the lysosome for degradation [93]. The part of interest in this pathway ensures creation of an autophagosome and is an ubiquitin-like cascade of events. In this step, Atg7 is an E1-like enzyme, which means that it is similar to an ubiquitin-activating enzyme, Atg10 is an E2-like enzyme, similar to an ubiquitin-conjugating enzyme, and Atg12 is the ubiquitin-like protein. In a first step, Atg12 is transferred from Atg7, where it is activated, to Atg10. The similar reaction with Atg8 acting as ubiquitin-like protein and Atg3 as an E2-like enzyme is happening simultaneously. In the case of ubiquitination, an E3-enzyme, also called ubiquitin ligase, will connect the ubiquitin to the protein that needs to be degraded. However, in the case of autophagy, this enzyme is not required and Atg12 will automatically conjugate to Atg5. After conjugation with two copies of Atg16, another Atg5-Atg12 complex will be attracted to the final complex which acts as an E3-like enzyme and facilitates transfer of Atg8 from Atg3 to phosphatidylethanolamine (PE) which is a lipid in the membrane. Also, the Atg5-Atg12-Atg16 complex will target the membrane

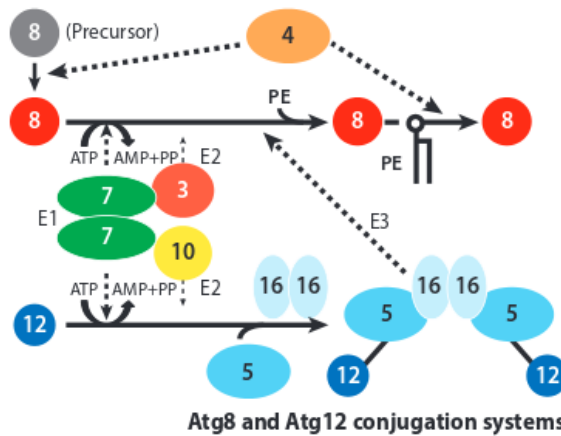


Figure 11.6: Part of interest in the autophagosome formation process [95]

used for autophagosomal formation and expansion [94] [95]. A schematic overview of this explanation can be seen in figure 11.6. Autophagy in case of starvation is very common because all unnecessary proteins can be degraded allowing to reuse the amino acids they are composed of [95]. As mentioned before in chapter 1, one of the specific ethanol stress responses in *S. cerevisiae* is that the cell enters a pseudo-starvation state [7]. Therefore, it is not surprising that null mutants of *Atp10* show a decreased resistance to ethanol [93]. The mutation is found in only one of the two copies of the gene in the clones of fermentor 6 at 100 generations, but has disappeared at 200 generations.

Arginine is a positively charged amino acid and is located close enough to aspartate to form salt bridges. By mutating arginine to serine, a polar and uncharged residue, this salt bridge will disappear. The BLOSUM value for this mutation is -1. It is very likely that this interaction will be lost although the importance of the complex in stress conditions can be clearly seen.

YBR130C (SHE3); chrII 496553; N106T

At 100 generations, a heterozygous mutation in *She3* in the clones of fermentor 6 is observed. This mutation has disappeared at 200 generations. *She3* is important for mRNA localization, which means that the mRNA is first transported to a certain destination before it is translated. This is done for a diversity of reasons. Firstly, it is more efficient to immediately translate the mRNA at the right location than having to transport the protein after translation. Secondly, during the mRNA transport, translation is often repressed which avoids accumulation of the protein on locations in the cell where it could be harmful. Also, response on a stimulus is much quicker if only translation of the localized mRNA has to be activated instead of initiating transcription. Finally, the localized mRNA often serves as a scaffold for protein complex assembly [96]. For a long time it was believed that this mechanism only took place in large and specialized cells like neurons or embryos, because it is a way to locally control expression levels. However, now it is known that mRNA localization is not a specialized process but happens in every cell with most mRNAs [96] [97]. The process of mRNA localization happens in a few steps. First, the mRNA

requires a zipcode that indicates the exact location. This zipcode must then be recognized by some RNA-binding proteins. This complex of mRNA and RNA-binding proteins, also called mRNP (messenger ribonucleoprotein), must then be properly transported through diffusion or active transport. At the final destination, the RNA is anchored. Finally, when translation is necessary, a certain mechanism has to ensure that the mRNA is released from the proteins that repress its translation [96]. One of the mechanisms to transport the mRNA is actin-myosin based, which is used in *S. cerevisiae* to transport mRNA from mother cell to daughter cell [96] [97]. In this case, the proteins that form the motor complex are Myo4 - the motor -, She2 - the protein that binds to the zipcode of the mRNA -, and She3 - the adaptor that connects She2-mRNA to Myo4 - [97]. To briefly summarize the function of She3, it can be said that it is an adaptor that links Myo4 and She2-mRNA in order to transport mRNA. Without She3, a cell is still viable [98]. In PDB file 4LL8 structural information about the stable complex of one Myo4 and two She3 proteins is available. This structural information indicates that residue 106 of one chain of She3 is interacting with residue 105, lysine, of the other chain of She3. It must be noted that the numbering of the residues in the PDB file is slightly different from what would be expected: residue 61 of the PDB numbering corresponds to residue 82 of the expected numbering. This means that, according to the PDB numbering, it is residue 85 of one She3 chain that is interacting with residue 84 of the other She3 chain and the mutation is located at residue 85.

The occurred mutation converts asparagine into threonine. Both residues are polar and uncharged, but threonine is more hydrophobic compared to asparagine. Although both residues have a negative propensity to be located at the interface, the presence of threonine is slightly more likely compared to asparagine. The BLOSUM value for this mutation is 0. The mutated residue interacts with lysine, a positively charged residue. Before mutation there were no salt bridges, so these cannot be destroyed by the mutation. Taking all this into account, it is expected that the mutation will probably have no effect on the interaction. It must be noted that it is also quite possible that the interaction will be lost after the mutation, because this provides a way of selective mRNA export out of the nucleus and can be used to regulate expression levels of proteins, which is - as indicated before - very important in stress conditions as generally only the proteins essential for survival will be generated [99].

Because this mutation happens twice, on every chain of She3, and both mutations are located close to each other, it could be interesting to investigate the effect of both mutations on the interaction. Unfortunately this is not possible in either of the two tools, but the mutations can be visualized. Figure 11.7 shows interacting residues of both chains and the interaction distance. Figure 11.8 on the other hand shows the mutated threonine residues in blue. It can be seen that, depending on the conformation of the side chain of threonine, the interaction is still present or is lost.

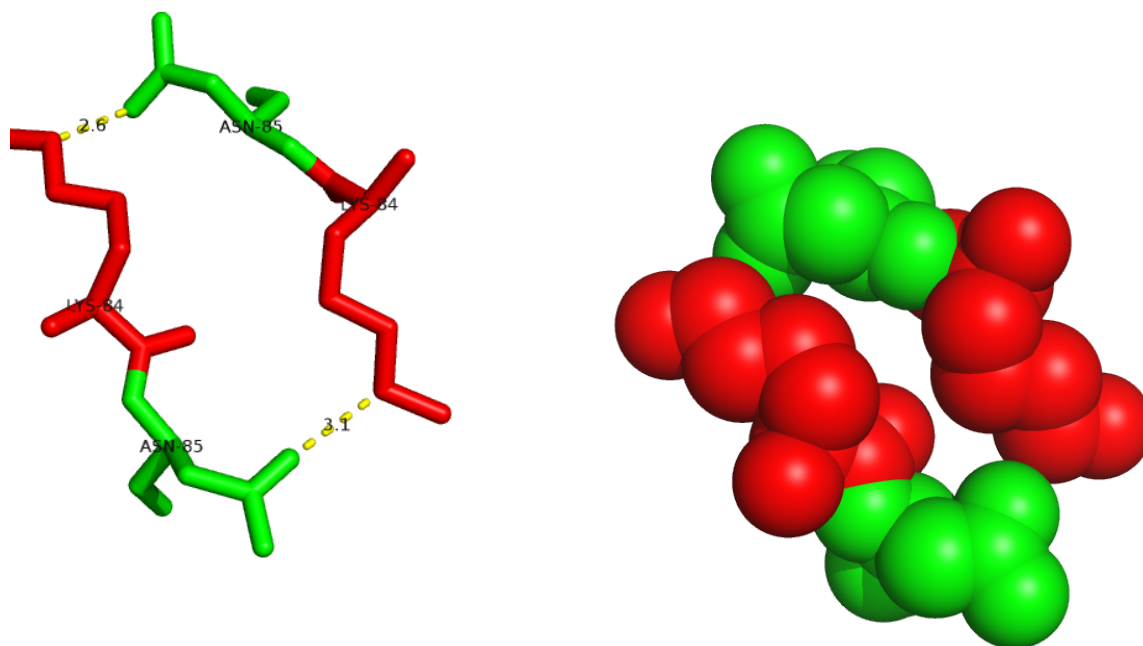


Figure 11.7: The original asparagine residue (green) with lysine (red) of two She3 chains, both in stick and in sphere visualization

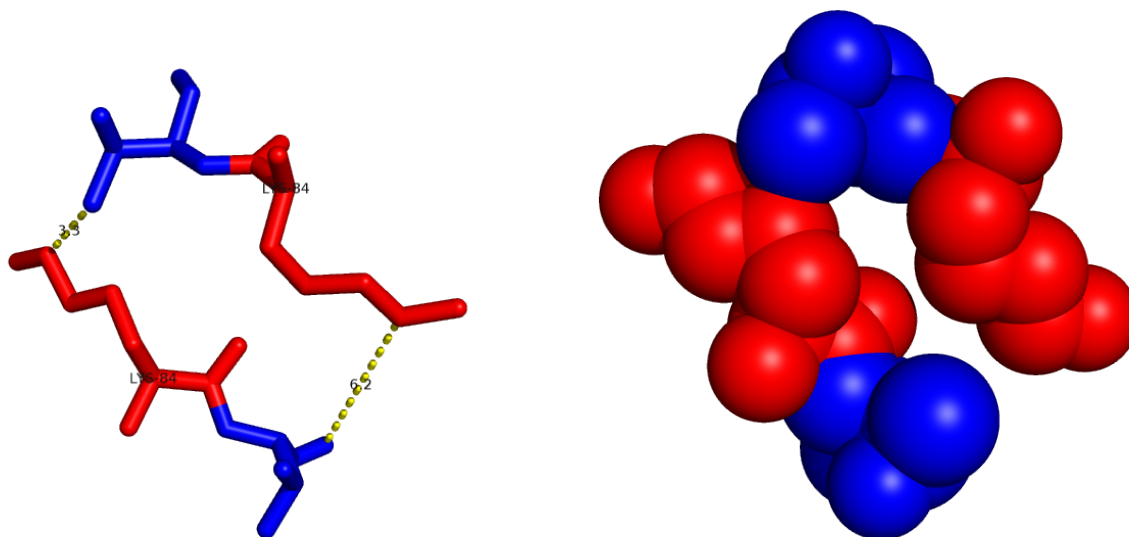


Figure 11.8: The mutated threonine residue in two different side chain conformations (blue) with lysine (red) of two She3 chains, both in stick and in sphere visualization

YJL112W (MDV1); chrX 206083; N260I

Another heterozygous mutation in the clones of fermentor 6 at 100 generations is located in the Mdv1 gene. Mdv1 acts as an adaptor that connects Fis1 with Dnm1 during mitochondrial fission. Fis1 is a receptor located at the outer membrane of the mitochondrion [100]. Dnm1 on the other hand is a dynamin-related GTPase that mediates constriction and fission [101]. Although originally Mdv1 was believed to be only the adaptor linking Fis1 and Dnm1, it also acts as a scaffold for other protein interactions important for mitochondrial fission [101]. Null mutants of Mdv1 are viable, but null mutants of Fis1 show a decreased viability [100] [102]. Even more, Fis1 inhibits cell death in ethanol stress by decreasing the amount of ROS (reactive oxygen species) that is generated in ethanol conditions. Next to that, Fis1 is responsible for mitochondrial fragmentation in ethanol stress [103]. Fis1 has thus an important function in ethanol stress which is less the case for Mdv1 and Dnm1 as null mutants of those genes did not show an altered sensitivity to ethanol [103].

The mutated residue in Mdv1 is interacting with glutamate at position 73 of Fis1 which is calculated from PDB file 3UUX. The original residue, asparagine, is a polar and uncharged residue. After mutation, asparagine is replaced by isoleucine. Isoleucine on the other hand is a non-polar residue. It is a lot more hydrophobic and has a higher propensity to be located at the interface compared to asparagine. The value present in the BLOSUM matrix for this mutation is -3. All of this indicates that the mutation will have a big impact on the interaction between both residues, but it is not clear if the the interaction will be strengthened or broken as null mutants of Mdv1 do not show altered sensitivity to ethanol.

YOR293W (RPS10A); chrXV 867707; Q58L

In the clones of fermentor 6, there is a heterozygous mutation present in Rps10a at 100 generations that has disappeared 100 generations later. The mutation takes place at residue 58 of Rps10a where there is interaction with Rps3. Information about this interaction is available in the following 34 PDB files: 3U5C, 3U5G, 4UJF, 4UJH, 4UJK, 4UJM, 4UJP, 4UJR, 4UJU, 4UJW, 4UJZ, 4UK1, 4UK4, 4UK6, 4UK9, 4UKB, 4UKF, 4UKH, 4UKK, 4UKM, 4UKP, 4UKR, 4UKU, 4UKW, 4UKZ, 4UL1, 4UL4, 4UL6, 4UL9, 4ULB, 4ULE, 4ULG, 4ULJ, 4ULM, 4ULP and 4ULR. Of all those files, 11 files - 4UJF, 4UJU, 4UK4, 4UKF, 4UKP, 4UKU, 4UL4, 4ULE, 4ULJ, 4ULP and 4ULR - indicate an interaction of the 58th residue of Rps10a with arginine at position 27 and glutamate at position 28 of Rps3. The other 23 files show an additional interaction with residue 24, phenylalanine, of Rps3. Both Rps10a and Rps3 are parts of the small ribosomal unit [104] [105]. As previously explained in chapter 10, the small ribosomal subunit is responsible for detecting the mRNA and for finding the right tRNA for every codon after binding with the large subunit [79]. Although Rps10a is not an essential protein, Rps3 is. This is probably because Rps3 is a multifunctional protein, not only involved in translation, but also in DNA repair, metastasis and apoptosis [106].

The occurred mutation transforms a glutamine residue into a leucine residue. Glutamine is a polar residue and leucine is a non-polar residue, both are uncharged. There is no possibility of a broken salt bridge or disulfide bond. However, leucine is a lot more hydrophobic compared to glutamine which could be beneficial for the interaction. The value in the BLOSUM matrix for this mutation is -2, indicating that this is a less common mutation and

will therefore have some effect. It could be argued that this interaction cannot be broken because Rps3 is essential to the cell so probably its interactions will also be important, but the fact that Rps3 is essential is already explained by its multifunctionality.

YOR157C (PUP1); chrXV 631115; A214T

Another heterozygous mutation that is observed in the clones of fermentor 6 at 100 generations is located in the Pup1 gene. Pup1 is synthesized as a propeptide that is necessary for efficient assembly of the 20S proteasome [107]. Cleavage between residue 29 and 30 is needed to activate the trypsin-like activity [107] [108]. This is the reason that the amino acid numbering in PDB files containing this protein starts at 1 in the active form, which is actually residue 30 in the propeptide. This implies that the mutation in residue 214 of the propeptide is actually residue 184 in the active form of the protein. There are 38 PDB files containing structural information of this protein: 1G65, 3OEU, 3OEV, 3SDI, 3SDK, 3SHJ, 3TDD, 3UN4, 3UN8, 3WXR, 4EU2, 4FZC, 4FZG, 4G4S, 4GK7, 4HNP, 4HRC, 4HRD, 4JSQ, 4JSU, 4JT0, 4LTC, 4NNN, 4NNW, 4NO1, 4NO6, 4NO8, 4NO9, 4Q1S, 4QBY, 4QLQ, 4QLS, 4QLT, 4QLU, 4QLV, 4R02, 4R17 and 4R18. All but 4G4S contain two chains of Pup1, resulting in 75 residues of which interaction partners can be calculated. It appears that, although the numbering is slightly different in some PDB files, residue 214 of Pup1 is always interacting with residues 223, 224 and 225, respectively glycine, tyrosine and threonine of Pre8. Both Pup1 and Pre8 are components of the proteasome, represented by β 2 and α 2 in figure 11.9. They are both essential for viability and needed for a correct structure of the proteasome complex [107] [109]. The proteasome is a complex that degrades the proteins that are ubiquitinated [110]. Activity of the proteasome is generated by three peptidases showing chymotrypsin-like, trypsin-like or postglutamyl peptide hydrolyzing activity [108]. Pup1 is responsible for the trypsin-like activity and cleaves the peptides that have to be degraded after basic residues [107].

The mutation that occurs in Pup1 mutates an alanine residue into a threonine residue. Alanine is slightly bigger than threonine, and is also a bit more hydrophobic. While alanine is non-polar, threonine is a polar residue. However, the BLOSUM matrix assigns a value of 0 to this mutation. It is expected that this mutation will be neutral or beneficial to the interaction due to the importance of both proteins and their necessity for a correct assembly of the proteasome.

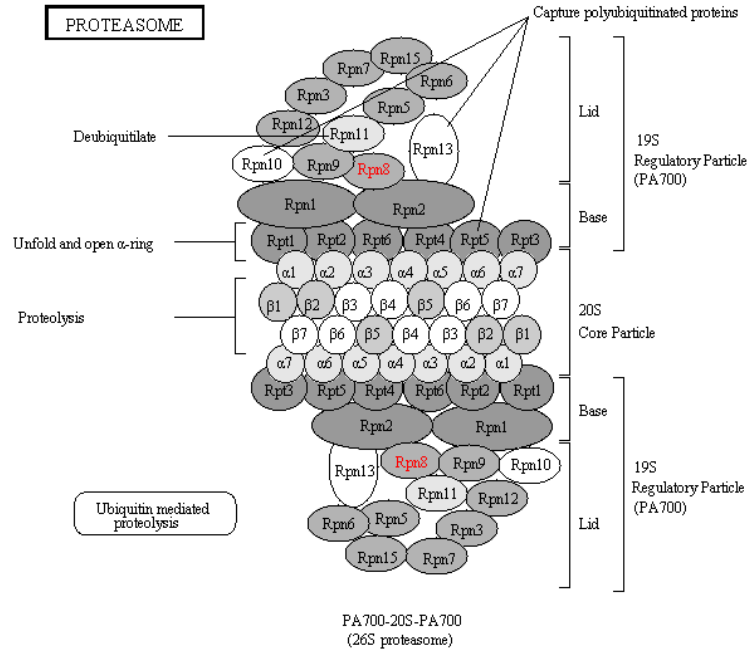


Figure 11.9: 26S Proteasome [51]

11.3 Non-synonymous mutations in the clones at 200 generations

11.3.1 Fermentor 3

YDL140C (RPO21); chrIV 209604; R320C

In the clones of fermentor 3, a mutation at position 209604 of chromosome IV is observed. This corresponds with residue 320 of Rpo21, where there is interaction with two proteins, Rpb2 and Rpb5, supported by PDB files 3S14, 3S2H, 3RZO, 3RZD and 3S2D. According to all PDB files, there is interaction with arginine at residue 14 of Rpb5. In PDB file 3S2H there is an additional interaction with lysine at position 470 of Rpb2, while in PDB file 3RZO there is interaction with lysine at position 471. PDB files 3RZD and 3S2D show interaction with both of those lysine residues of Rpb2 and also with a glutamine residue at position 469. Finally, PDB file 3S2D shows an additional interaction with Rpb5 at residue 7 which is arginine. In figure 11.2, these interactions are schematically visualized: Rpb5 is subunit ABC1, Rpo21 is subunit B1 and Rpb2 is represented as B2. As mentioned before in this chapter, Rpo21 is the largest subunit of RNA polymerase II. This subunit, together with Rpb2, Rpb3 and Rpb11 (respectively B3 and B11 in figure 11.2), are the core subunits of RNA polymerase II, and mutations in one of those subunits can affect both the assembly of the complex and proper functioning [85]. Knock out of this gene results in an inviable cell [85].

The arginine residue at position 320 mutates into a cysteine residue. There is some difference between both residues: arginine is a positively charged residue, while cysteine is

a polar residue containing sulfur. Also, the BLOSUM matrix assigns a value of -3 to this mutation, indicating that it is less likely to occur. However, when looking at the residues this amino acid interacts with, it can be seen that all these residues are positively charged (lysine, arginine) or uncharged (glutamine). Two positively charged residues repel each other, and therefore the mutation from a positively charged to a neutral residue could be beneficial for binding. In section 11.1, the interaction between Rpb5 and Rpo21 is already discussed. Next to that, the interaction between Rpo21 and Rpb2 cannot be disrupted as both are core subunits of RNA polymerase II and needed to ensure proper assembly and transcription of necessary genes [85].

11.3.2 Fermentor 6

YKL210W (UBA1); chrXI 41892; E910D

This mutation is already completely explained in section 11.1.

YBR154C (RPB5); chrII 548375; R212K

This mutation is already completely explained in section 11.1.

YER018C (SPC25); chrV 192086; V180G

A heterozygous mutation in Spc25 is observed in the clones of fermentor 6 at 200 generations. At residue 180, there is interaction with arginine at position 67 of Cnn1. Structural data can be retrieved from PDB file 4GEQ. Spc25 is part of the highly conserved Ndc80 complex [111] [112]. This complex is responsible for the binding of microtubules to the kinetochore during cell division. Kinetochores are large complexes that assemble to the centromeres of the chromosome to separate sister chromatids. The interaction between Spc25 and Cnn1 is part of the connection between the Ndc80 complex and the centromeres and is schematically pictured in figure 11.10 [112]. The microtubule is shown in green and the centromere as a black line. Knock out of Spc25 results in inviability of the cell [111].

A valine residue is mutated into a glycine residue. These two amino acids are both non-polar amino acids, with valine being a lot more hydrophobic than glycine. The BLOSUM matrix shows that this is a less likely mutation to occur. Also, the whole Ndc80 complex is conserved, indicating the importance of all proteins involved in this complex. Therefore, it can be expected that the mutation will not break the interaction between the proteins.

YLR199C (PBA1); chrXII 548089; S199C

Another heterozygous mutation observed in the clones of fermentor 6 is located in the gene Pba1. Residue 199 of Pba1 is interacting with glutamine at position 43 of Add66 based on the structural information in PDB file 4G4S. In chapter 10, a STOP mutation in Add66 at position 46 is already described. This STOP mutation is also observed in the clones of fermentor 6 at 200 generations. In figure 11.11a, the interacting surfaces of Pba1 and Add66 before any mutation are colored in red and blue. It can be seen that there is only one big interaction surface. This means that the interaction is definitely influenced by the STOP mutation. In figure 11.11b, the same surface of Pba1 is shown in red, but for Add66

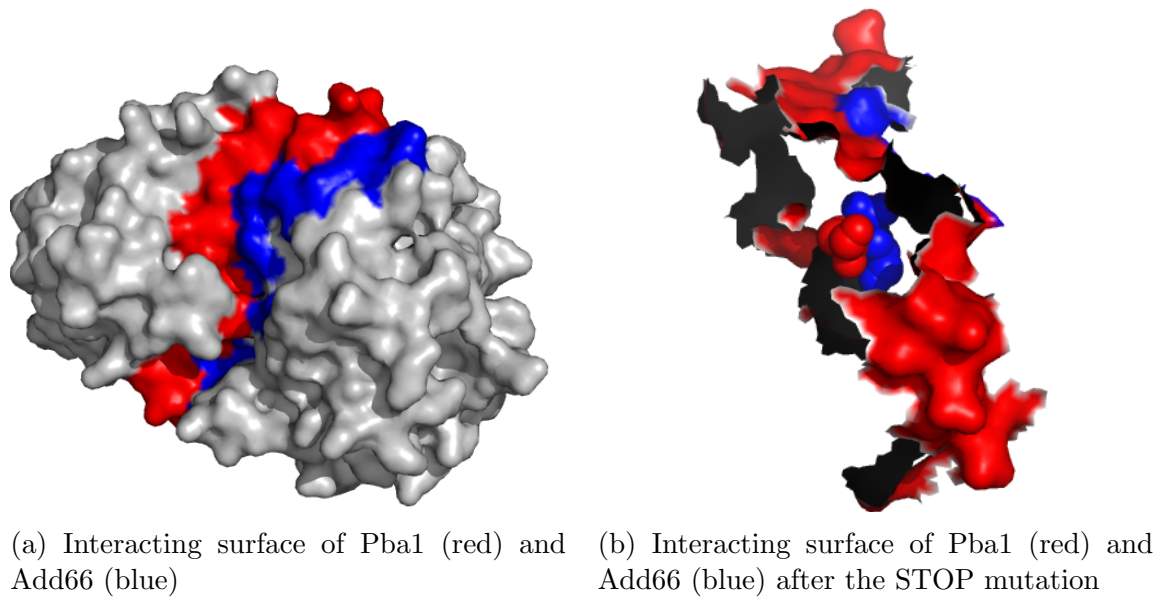


Figure 11.11: Interaction between Pba1 and Add66

chapter, the Ndc80 complex binds to the microtubules to connect them to the kinetochores to ensure a correct separation of the sister chromatids [112]. Both Spc25 and Spc24 are essential in *S. cerevisiae*.

Residue 135 of Spc25 interacts with an asparagine residue at position 158 of Spc24. Both the original amino acid, alanine, and the mutated amino acid, valine, are non-polar and uncharged residues. Valine is slightly bigger and less hydrophobic than alanine, and the BLOSUM matrix assigns a value of 0 to this mutation. Due to the fact that both proteins are essential and that the complex is highly conserved, it is expected that the interaction will not be lost.

YMR293C (HER2); chrXIII 856076; V240R

A heterozygous mutation has also occurred at residue 240 of Her2 and is located at the interface with Gtf1. Her2 and Gtf1 form together with Pet112 a tRNA-dependent amidotransferase of which the structure is present in PDB files 4N0I and 4N0H. This enzyme is needed in the tRNA pathway: in every species, there is the need for at least one paired aminoacyl-tRNA per amino acid. Connection of each tRNA with its corresponding amino acid can be done either directly or indirectly. Most of the aminoacyl-tRNA is made in the direct way by an aminoacyl-tRNA synthetase that catalyzes acylation between tRNA and amino acid. However, there is also an indirect pathway during which conversion of a precursor mischarged aminoacyl-tRNA takes place. In the case of glutamyl-tRNA, both pathways have been observed: while the direct pathway is present in eukaryotes, the indirect pathway is present in prokaryotes. However, in the mitochondrion of *S. cerevisiae*, it seems that the indirect pathway is used instead of the expected direct pathway. This indirect pathway is regulated by two enzymes. First, there is a non-discriminating glutamyl-tRNA

synthetase which creates a Glu-tRNA^{Gln}. In a second step, a tRNA-dependent amidotransferase transamidates this Glu-tRNA^{Gln} into Gln-tRNA^{Gln} [115]. This second enzyme is called the GatFAB tRNA-dependent amidotransferase and is the complex where the mutation takes place. Although not mentioned on SGD, this complex is also used for the conversion from Asp-tRNA^{Asn} to Asn-tRNA^{Asn}, see the red boxes in figure 11.12. The GatFAB complex consists of three proteins: GatA (Her2), GatB (Pet112) and GatF (Gtf1). Although GatA and GatB are very conserved accross eukaryotes and bacteria, GatF seems to be only present in fungi. There it has the same function as GatC in eukaryotes, although there is only little sequence homology between both proteins. GatF in yeast has two functions: the first one is to modulate the folding and stability of GatA as this protein is more unstable in yeast than in other organisms. Therefore it needs extensive interactions with GatF. The second function of GatF is to maintain the assembly of GatA and GatB [116]. It can be understood that GatF has an important function for correct assembly of the complex, but null mutants are also viable. Even more, an increased stress resistance is observed in GatF null mutants [117].

S. cerevisiae is viable without Her2, but a decreased resistance to 5% of ethanol has been observed [56] [118]. The mutation replaces a valine residue by an arginine residue. They differ a lot from each other: valine is a lot smaller and non-polar, while arginine is a positively charged residue. Interaction at this position takes place with a tyrosine residue at position 45 of Gtf1. Also, the BLOSUM matrix indicates a value of -3, indicating that the mutation is not so likely to occur. Therefore it is assumed that the mutation will impact the interaction. An argument for a strengthened interaction would be that lots of interactions between Her2 and Gtf1 are needed due to the unstable nature of Her2 in yeast, but the increased stress resistance in cells lacking Gtf1 shows that these interactions might not be that important in stress conditions. Therefore, it is very likely that the mutation breaks the interaction.

YDR529C (QCR7); chrIV 1496421; L43H

In chapter 10, it is explained that a silent mutation in Qcr2 at 100 generations is observed in the clones of fermentor 6. At 200 generations, this mutation has disappeared but a heterozygous non-synonymous mutation in this same complex has appeared in subunit 7, called Qcr7. As mentioned before, this complex is part of the electron transport chain and is located in the inner mitochondrial membrane. Subunit 7 would play a role in the assembly of this complex and is essential for activity [119]. Next to that, Qcr7 is conserved accross eukaryotes [119]. Structural data is given in PDB file 4PD4. This mutation is found at position 43 where the leucine residue interacts with the following residues of Cob: leucine at position 210, isoleucine at position 212 and phenylalanine at position 318. Although a decreased resistance to 5% is observed [56], the cell would still be viable without Qcr7 [119]. The observed mutation changes leucine into histidine. Leucine is a non-polar residue while histidine is a positively charged one. The BLOSUM matrix assigns a value of -3 to this mutation. Due to this big change it is very likely that the interaction is altered. It could be argued that the presence of Qcr7 is essential for activity of the complex and thus the interaction will not be broken, but because *S. cerevisiae* is still viable without Qcr7, the interaction might not be as important as expected and could thus be disrupted.

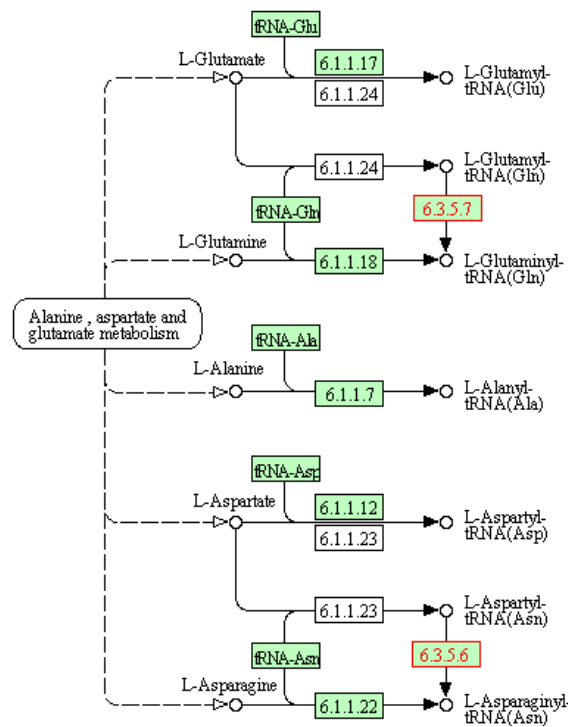


Figure 11.12: aminoacyl-tRNA biosynthesis for glutamine and asparagine [51]

YDL140C (RPO21); chrIV 208113; A817I

Next to a non-synonymous mutation in Rpo21 in the clones of fermentor 3, a different non-synonymous and heterozygous mutation in this gene is observed in the clones of fermentor 6. At position 817, the alanine residue interacts with leucine at position 514, proline at position 524 and serine at position 764 of Rpb2. Structural information is available in PDB files 3RZD, 3S14, 3S2H, 4A3K, 3RZO, 3S2D and 4A3F. As explained previously in this chapter, both genes belong to the core subunits of the RNA polymerase II complex, which is illustrated in figure 11.2. Rpo21 is essential for viability and mutations in any of the core subunits could affect assembly or function of the complex [85].

Since this complex is necessary for transcription of most of the genes, it has been argued already that this interaction cannot be disrupted. The mutation taking place changes the alanine residue into an isoleucine residue. Although the BLOSUM matrix assigns a value of -1 to this mutation, both are non-polar residues with alanine slightly smaller in size. Also, isoleucine appears to be more hydrophobic and has a higher propensity to be localized at the interface compared to alanine. Also, no salt bridges or disulfide bonds are lost due to this mutation. Therefore, it is possible that this mutation enforces the interaction between Rpo21 and Rpb2.

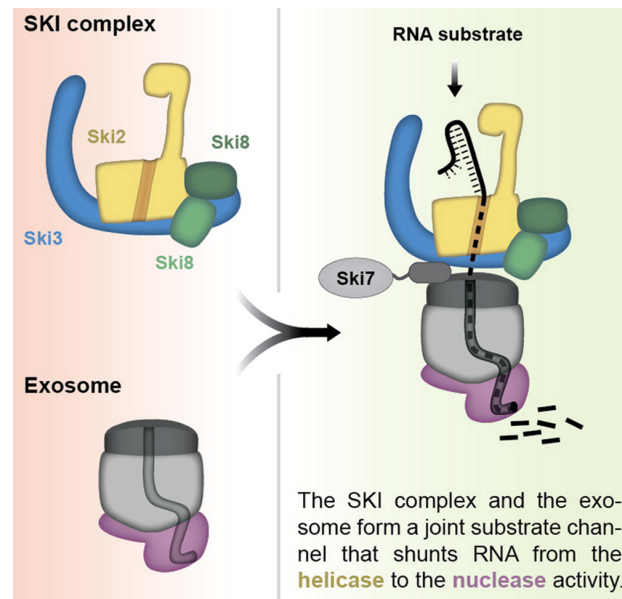


Figure 11.13: The Ski complex with the exosome [120]

YPR189W (SKI3); chrXVI 915533; T957N

One Ski3 protein forms with one Ski2 protein and two Ski8 proteins the Ski complex. This Ski complex is evolutionary conserved and is linked to the exosome via Ski7. The complex guides RNA to the exosome for 3'-to-5' degradation. This interaction is visualized in figure 11.13. Mutants of one of the proteins involved in the Ski complex are lethal in *S. cerevisiae* when the 5'-to-3' exoribonuclease is deleted [120]. However, null mutants of just Ski3 show an increased stress resistance [121]. In PDB file 4BUJ structural information about this complex is available.

On the interaction site between Ski3 and Ski2, a non-synonymous mutation is found in Ski3 in the clones of fermentor 6. This mutation is heterozygous and present at 200 generations and changes a threonine residue into an asparagine residue at position 957. This position interacts with a valine residue at position 94 of Ski2. Both threonine and asparagine are polar uncharged amino acids and the mutation is assigned a value 0 in BLOSUM. Although this could indicate that the impact of the mutation is rather small, it is still possible that the interaction is broken. A disrupted interaction of this complex will result in RNA accumulation because it is not degraded as fast anymore. This could be a way of the cell to regulate expression levels of some proteins that are very necessary in stress conditions.

YOR261C (RPN8); chrXV 816252; G227D

PDB file 3J47 contains structural information of the lid of the 26S proteasome, of which the composition is shown in figure 11.9. In a subunit of this lid a heterozygous mutation in the clones of fermentor 6 at 200 generations is observed at residue 227 of Rpn8, colored red in figure 11.9. This residue interacts with both a methionine residue at position 366 of Rpn9 and a histidine residue at position 425 of Rpn5. As explained before, the function of a proteasome is to degrade all ubiquitinated proteins. Rpn8 is an essential subunit of

this proteasome and the cell is inviable without it [122]. Next to that, *S. cerevisiae* is also inviable in case of null mutants of Rpn5 and Rpn9 [123] [124].

The occurred mutation switches glycine into aspartate. Glycine is the smallest amino acid and is non-polar. Aspartate on the other hand is negatively charged. However, since histidine is positively charged, this mutation could induce salt bridges that enforce the interaction. Other reasons that lead to this belief, are the function of this complex - which is highly important in stress conditions - and the fact that all these proteins are essential for survival.

YGR183C (QCR9); chrVII 859191; A25T

Together with the mutation in Qcr7 explained earlier in this section, a heterozygous mutation in Qcr9 or the ninth subunit of the ubiquinol cytochrome-c reductase is observed in the clones of fermentor 6 at 200 generations. At position 25, Qcr9 is interacting with residues 65, 68 and 69 of Rip1, which are respectively leucine, serine and alanine. This interaction can be found in PDB file 4PD4. Qcr9 is, just like Qcr7, conserved across eukaryotes and essential for both structure and function of the complex [125]. Also, Rip1 is one of the catalytic subunits of the complex and is known to be unstable with reduced function in absence of Qcr9 [125]. Another phenotypic characteristic of a null mutant of Qcr9 is a decreased resistance to 8% of ethanol [56].

Although alanine is a nonpolar amino acid while threonine is a polar one, both are uncharged and BLOSUM assigns a value of 0 to this mutation. Next to that, to ensure proper function of Rip1, interaction with Qcr9 is necessary. Therefore, it is assumed that this interaction will not be disrupted.

YIL018W (RPL2B); chrIX 317718; R184H

Rpl2b is a protein of the large subunit of the ribosome. This subunit contains the active site of the ribosome where the peptide bonds are formed [79]. A heterozygous mutation in the interaction site with Rpl34A is found in the clones of fermentor 6. Next to this heterozygous mutation at 200 generations, a silent mutation in Rps5 of the small ribosomal subunit has also been found in these clones, see chapter 10. Rpl2b null mutants are viable and show an increased stress resistance [126]. At position 184, Rpl2b interacts with residues 14, tyrosine, and 30, glutamate, of Rpl34a, which is another protein of the large ribosomal subunit. 3D data of the complex can be retrieved from the PDB with identifier 4B6A.

The arginine-to-histidine mutation is assigned a value of 0 in the BLOSUM matrix. Both residues are positively charged, which could indicate that the interaction is not disrupted, especially because the salt bridge with glutamate is not lost. However, since null mutants of Rpl2b have an increased stress resistance, it means that this protein is not really essential for correct assembly nor function of the ribosome. Therefore, it is also possible that the interaction is broken.

Chapter 12

Discussion and future work

First of all, it is interesting to look at all affected pathways link this to known effects of ethanol. Then, it can be studied whether the affected processes occur in all fermentors, or if different processes are affected in the different fermentors. Also, something must be said about the prediction results. Finally, it is explained how to proceed this research.

12.1 Affected processes

When looking at all mutations (nonsense, silent and non-synonymous) that occur at interaction sites, regardless of fermentor or time point, one thing is very clear: most of the SNPs occur either in central dogma processes (transcription, translation) or have to do with breakdown or refolding of proteins. All genes participating in one of those processes can be found in figure 12.1 and a broader explained can be found in the following 2 subsections. The other genes that have SNPs at interaction sites are divided based on their localization, displayed in figure 12.2. Genes related to protein synthesis, vacuole and mitochondrion have an altered expression level during ethanol stress as mentioned in chapter 1, so it is already known that ethanol influences processes there.

First of all, there is the vacuole. It has been mentioned in chapter 1 that the vacuole is responsible for storage of small molecules and for degradation of waste products. Also, one of the effects of ethanol on *S. cerevisiae* is the fact that the vacuole loses its segregated structure [7] [8]. Furthermore, it has been seen that a reduced pH caused by ethanol is counteracted by V-ATPases [7]. Next to that, an increased transcription of proteins related to vacuolar functions is seen [7]. Therefore, it is surprising that all mutations found in vacuolar proteins are nonsense mutations. However, the viability of the cell without any of these proteins show that these genes are not essential. This is the foundation of the belief that this is a way of the cell to save energy by not synthesizing unnecessary proteins while keeping important processes working.

Secondly, the mitochondrion is known to be the organelle where ATP, the energy source of the cell, is generated. However, mitochondria are also involved in many other processes like programmed cell death [127]. The discovered mutations can be again divided into two categories: three mutations are found in genes of the electron transport chain and one mutation is found in a gene involved in mitochondrial fission. The electron transport chain is the main pathway of ATP synthesis, and a loss of one of its components can inhibit the

whole pathway and lower ATP levels [128]. Therefore, it is expected that the observed mutations ensure a perfectly working electron transport chain in ethanol conditions, so enough energy is present to survive. In chapter 1 it was explained that energy is saved by adjusting expression levels of unnecessary genes. Another way of making sure that there is enough energy is of course to produce more ATP. Mitochondrial fission has an important function as it separates non-working mitochondria: mitochondria that are unable to fuse again are dysfunctional and can be broken down through an autophagy-related process [127]. Increased fission is important to ensure only fully functional mitochondria is present, while the material of the dysfunctional mitochondria can be reused [127].

Finally, two non-synonymous mutations in Spc25 have been observed. This gene is present in the cytoskeleton and is involved in correct separation of sister chromatids for the daughter cells [112]. Different genome-wide studies have shown the importance of genes related to cytoskeleton biogenesis [129] [130]. This can be related to the function of the cytoskeleton which is to support cell shape.

12.1.1 Central Dogma processes

Most of the proteins that have mutations in an interaction site are related to transcription and translation. This is well explained by a paper that appeared very recently. Haft e.a. have investigated the effects of ethanol on transcription and translation in *E. coli* using RNAseq and ribosome profiling [131]. A summary of their findings can be found in figure 12.3. *E. coli* is a prokaryote and has thus no nucleus. This means that transcription and translation happen at the same place in the cell. The upper part of figure 12.3 shows how transcription and translation happen in the absence of ethanol, while the lower part shows the effect of ethanol on transcription and translation. Three main effects have been found: increased misreading, increased stalling and increased pausing [131]. First of all, the RNA polymerase is affected by ethanol in such a way that it moves slower across the DNA, thereby decreasing the transcription speed. This makes it possible for Rho to terminate transcription early [131]. Rho is a prokaryotic factor that binds to RNA that is being described, unwinds it from the DNA template and terminates transcription [132]. Next to that, the ribosome conformation is altered in ways that increase misreading and stalling during translation. Misreading is the process where a wrong amino acid is added to the growing peptide chain. Stalling of the ribosome happens at certain codons (during this experiment stalling happens mostly at non-starting methionine codons), indicating that their translation is inhibited by ethanol. This stalling can lead to early termination of translation or affect the folding of the protein (as explained in chapter 10.2) [131]. All three effects have their impact on protein synthesis and they increase the levels of misfolded and incorrect proteins [131].

This experiment was performed on *E. coli*, so some clear differences with eukaryotes are present: first of all, in eukaryotes transcription happens in the nucleus while translation happens in the cytoplasm. Also, transcription is not terminated by a rho factor. However, it is highly likely that the direct effects of ethanol on transcription and translation in *E. coli* are also present in *S. cerevisiae*. Therefore, it makes sense that a lot of mutations take place in the Central Dogma processes: if ethanol affects transcription and translation, the cell must find a way to correct for this in order to survive. The introduction of mutations

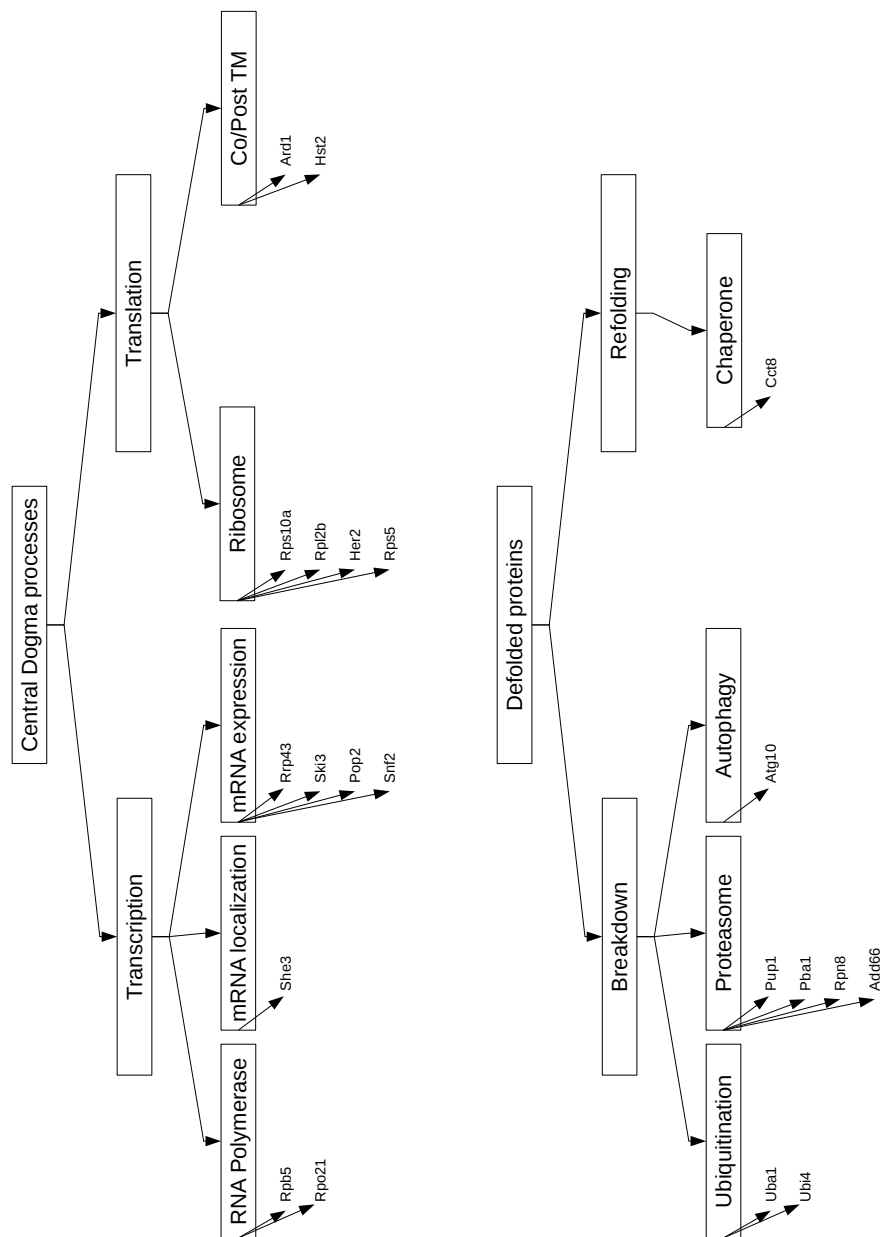


Figure 12.1: All mutated genes that are involved in central dogma processes or breakdown and refolding of proteins

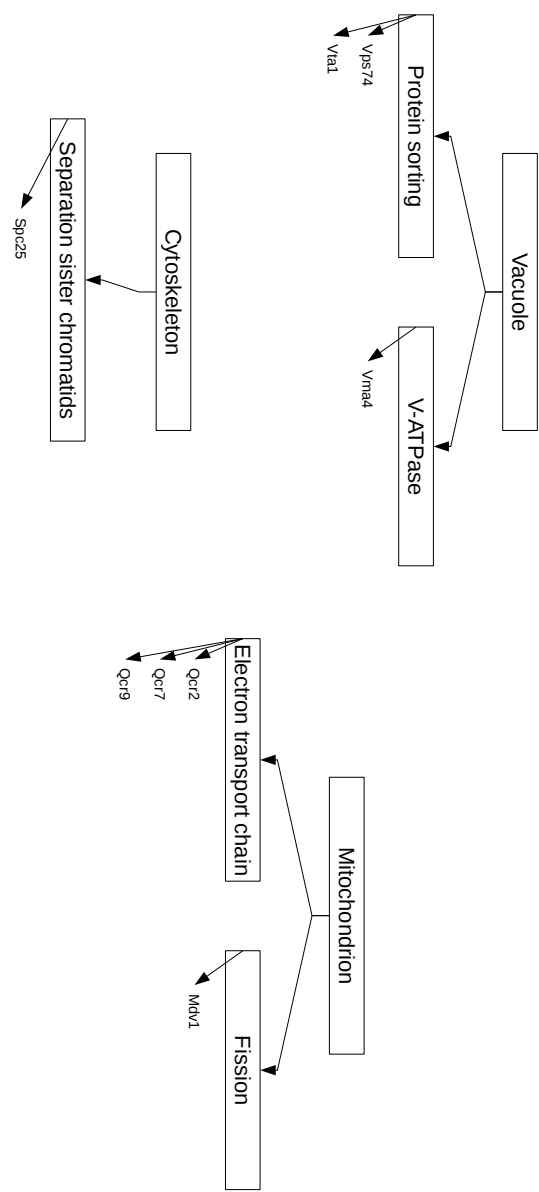


Figure 12.2: All mutated genes that are not involved in central dogma processes or break-down and refolding of proteins

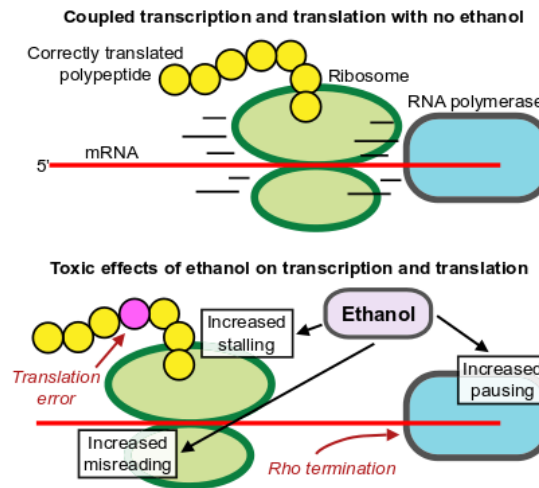


Figure 12.3: How ethanol affects transcription and translation in *E. coli* [131]

in these complexes could be one of the solutions.

12.1.2 Defolded proteins

As previously mentioned, an increased amount of defolded proteins is observed in stress conditions [72]. This is dangerous as it causes both loss of protein function as well as aggregation of misfolded proteins [72]. There are two solutions to this: either energy is spent in trying to refold the protein properly or the protein is broken down. Refolding is performed by chaperones. Next to that, chaperones also aid in protein folding during translation. Breakdown of proteins is performed by labeling the protein with ubiquitin. The labeled protein is then a target for the proteasome that breaks down the protein [72].

During the NATAR project RNA sequencing of wild type yeast cells (haploid and diploid) in the presence of ethanol was performed, and this showed an upregulation of pathways that indicate protein misfolding (personal communication with Karin Voordeckers). This already indicates that a lot of proteins will be defolded in ethanol stress conditions. To counteract this, both refolding and breakdown of proteins will be highly important and it can be expected that the mutations occurring in the involved complexes try to make these processes more efficient.

12.2 Affected processes in the different fermentors

Another interesting point of view is investigating the separate fermentors to see whether all discussed pathways are affected in all fermentors, or whether every fermentor affects a different pathway. However, when dividing the SNPs based on the fermentor, it can be seen in figure 12.4 that almost all SNPs are located in fermentor 6. This can be explained by the fact that fermentor 2 and 6 have the mutator phenotype: an increased number of mutations in those fermentors is observed due to more mistakes during DNA replication or due to decreased DNA repair. While in fermentor 2 there are more indels, in fermentor 6 there is a higher number of SNPs.

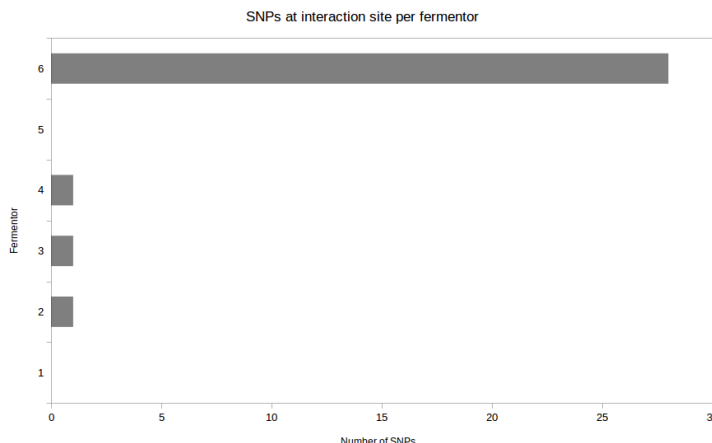


Figure 12.4: The number of SNPs at interaction sites of protein-protein interactions per fermentor

However, because almost all SNPs are located in the same fermentor, it can be stated that it is a combination of all discussed pathways that is affected by ethanol. If the pathways could be clearly divided over the different fermentors, it could be stated that mutations in only a limited amount of pathways are needed to acquire ethanol tolerance. However, now it becomes clear that all discussed pathways are affected by ethanol and changes in all of them are necessary to survive.

12.3 The predictions

Although the prediction accuracy of both SNP-IN and PoPMuSiC cannot be measured as there is no interaction data to compare the prediction results with, it is still possible to study the prediction results in detail. Comparisons between the results of SNP-IN, PoPMuSiC and literature-based opinion are listed in tables 12.1, 12.2 and 12.3.

When looking at table 12.1, it is clear that both prediction tools predict the mutation to be deleterious in most cases. However, there are 4 cases where both tools predict the interaction to be preserved: two chains of Mdv1 and two chains of Rpn8. Because all structural information of the observed mutations in both genes gives positive results in both tools, it is highly likely that those interactions will not be lost. In the case of Rpn8, the literature-based opinion also indicated that the interaction is probably preserved. In the case of Mdv1 however, it was not clear what conclusion to draw from the literature.

The two other tables 12.2 and 12.3 also show some expected trends: the number of cases where the opinion predicts a broken interaction while the tool predicts a preserved interaction, is 0 for both tools. The opposite situation shows high numbers for both tools (82 and 100). This can be easily explained: the tools are bad in predicting neutral or beneficial mutations and will predict more often that the interaction is lost. The literature-based opinion on the other hand is a lot more careful in predicting interactions to be lost.

When comparing the three results, there are 5 cases where the 3 predictions are the same: the mutation in the two chains of Atg10 are all predicted to be deleterious, just as the mutation in the chain of Pba1. Both chains of Rpn8 on the other hand will still be

SNP-IN	PoPMuSiC	
	NP	P
	NP	88
	P	10
		4

Table 12.1: SNP-IN versus PoPMuSiC (56 without both)

OPINION	SNP-IN	
	NP	P
	NP	3
	P	82
		10

Table 12.2: Opinion versus SNP-IN (67 without both)

OPINION	PoPMuSiC	
	NP	P
	NP	3
	P	100
		6

Table 12.3: Opinion versus PoPMuSiC (53 without both)

interacting after the mutation.

Finally, there is 1 mutation that has a likelihood of 1 in the SNP-IN tool to be deleterious: L43H in Qcr7. Therefore, it can be expected that this interaction will definitely be broken.

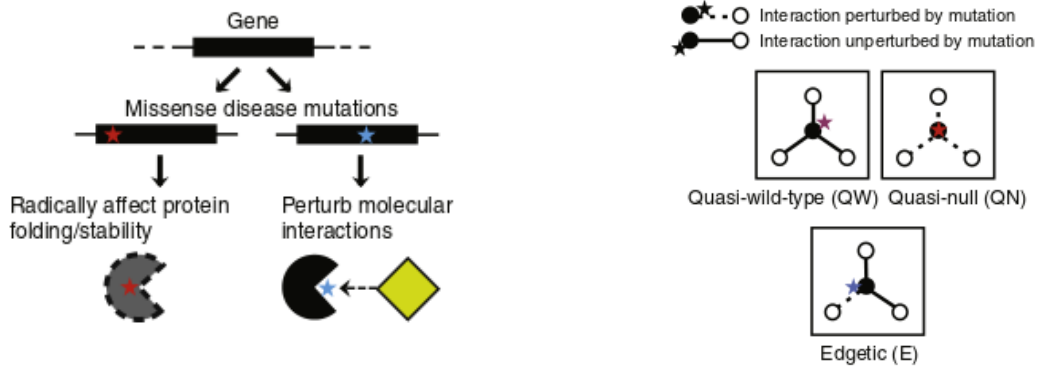
12.4 Future work

The importance of non-synonymous mutations at the interaction sites of protein-protein interactions is very recently investigated by Sahni e.a. [133]. They investigated a whole data set of disease-associated mutations in humans. Previously, these mutations were thought to result in a complete loss of function of the protein. This however, is a simplification of the reality as different mutations in the same gene can lead to different diseases. Therefore, all non-synonymous mutations were thought to either affect the protein stability or to perturb molecular interactions, as shown in figure 12.5a [133].

First, they characterized protein-chaperone interactions with the hypothesis that an increased binding to chaperones is an indicator for misfolding or instability, leading to a decreased or even a complete loss of function. It turns out that for all mutated genes that show an increased binding to chaperones, significantly more mutations are located in the core of the protein [133]. This was already explained in chapter 3.2: due to the constraint space and the high hydrophobicity in the protein core, mutations in the core are more likely to result in destabilization or misfolding of the protein [20]. However, the majority of mis-sense disease mutations does not belong to this category [133].

Therefore, a second step was to investigate whether these mutations disturbed PPIs. The effect on interactions can be categorized in three groups: quasi-wildtype, when there is almost no effect on interactions, edgetic, when some interactions are lost, or quasi-null, when all interactions are lost. This is visualized in figure 12.5b. It appeared that in the case of a quasi-null PPI profile, the protein showed a decreased expression and an increased binding to chaperones. In the other two PPI profiles on the other hand, no significant change in expression levels or chaperone binding was noted. This indicates that when all interactions are lost, the mutation most likely affects protein structure or stability, while otherwise the mutation affects some or all interactions. Diseases caused by a different mutation in the same gene can then be explained by the fact that different interactions will be perturbed in both diseases [133].

When comparing disease-causing mutations to non-disease mutations, they discovered



(a) Two hypotheses for the missense disease mutations: either they affect protein folding or they disturb interactions [133]

(b) The three types of PPI profiles for mutations [133]

Figure 12.5: Mutations in protein-protein interactions

that disease-causing mutations are more likely to disturb PPIs. The missense disease mutations can be divided into two big categories: one group of mutations that result in a complete loss of interaction due to misfolding or instability and another group of mutations that results in partial loss of interactions. The fact that different non-synonymous mutations in the same gene lead to different diseases due to different interaction losses clearly indicates the importance of PPIs for the phenotype [133].

The study by Sahni e.a. has clearly demonstrated that different phenotypes can be obtained by changing the interactome network and gives meaning to the study performed. This work can be extended in two directions, either in validating the predictions, or in increasing prediction accuracies. These two are discussed now.

12.4.1 Validation of the obtained predictions

The predictions obtained can be validated in order to increase the understanding of the effect of ethanol on PPIs. There are a lot of methods that can be used to detect PPIs, but in a relatively recent paper from 2009 it is stated that yeast two-hybrid (Y2H) and affinity-purification mass spectrometry (AP-MS) are the two most frequently used methods and therefore those two will be discussed here [134].

Y2H

The Y2H system is based on the transcriptional activator Gal4, which consists out of a DNA binding domain (DBD) and an activation domain (AD). Both domains are independent and work individually but they are both needed for a functional Gal4. In the Y2H system, the protein of interest, X, is attached to the DBD. In this case, X would be the mutated gene. The potential interacting partner is attached to the AD. If there is interaction between X and Y, then the transcriptional activator binds to the upstream activator sequence (UAS) and attracts RNA polymerase II which then starts transcription of a reporter gene, illustrated in figure 12.6. This reporter gene must encode for a protein of which the function can easily be tested, for example an auxotrophic marker: when the reporter gene is not transcribed, the cell will not be

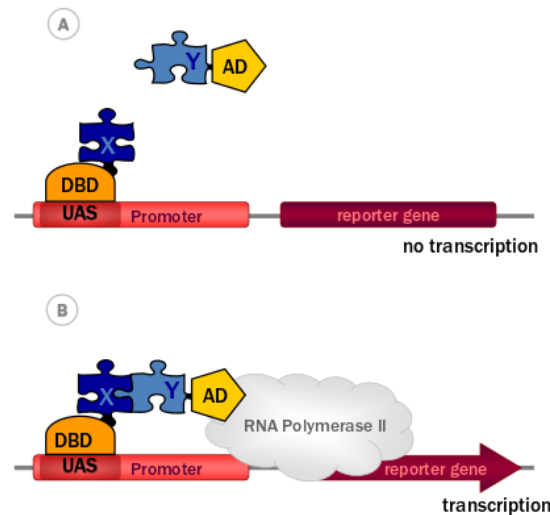


Figure 12.6: The yeast two-hybrid system, (A) when there is no interaction, (B) when there is interaction [134]

able to grow on minimal media [134]. There are some drawbacks like the fact that only binary interactions can be measured and that it must be possible for both the protein of interest and its potential interacting partner to enter the nucleus. Also, a confirmed interaction in the Y2H system does not necessarily mean that there is *in vivo* co-localization in time and space. Next to that, only limited PTMs are possible inside the nucleus. But, as this system is relatively cheap and easily accessible, this is a popular method and lots of variations of this system have been created that deal with these problems [134].

AP-MS

AP-MS is first affinity purification followed by mass spectrometry. Affinity-purification is the capture of biological material in soluble phase via a high-affinity ligand that is coupled to a solid support [135]. First, cells that express the protein of interest are grown, the cells are broken and the content is poured into a solid support. A ligand, for example an antibody, is attached to this solid support, separating the protein of interest and its interacting partners from the rest of the cells material. With this complex, mass spectrometry is performed. During mass spectrometry, the proteins are first cut into smaller peptides. Of these peptides, the *mass-over-charge ratio* (m/z) is measured. This can be done a second time (MS-MS), and the resulting spectra can be compared to a database to identify the proteins in the complex. The advantages of this system are that the interactions happen *in vivo*, that a whole complex is purified and that condition-specific interactions can be measured. However, it is not trivial to identify multiple proteins from a peptide list. Also, an antibody (or other ligand) of the protein of interest has to be known [135]. This can be solved by fusing the open reading frame of the protein of interest to the DNA sequence of a tag. This tag is ideally a protein of which purification can be done very efficiently. Finally, the result of purifying a whole complex lists all participating proteins, but there is no way to distinguish the directly from the indirectly interacting proteins [135].

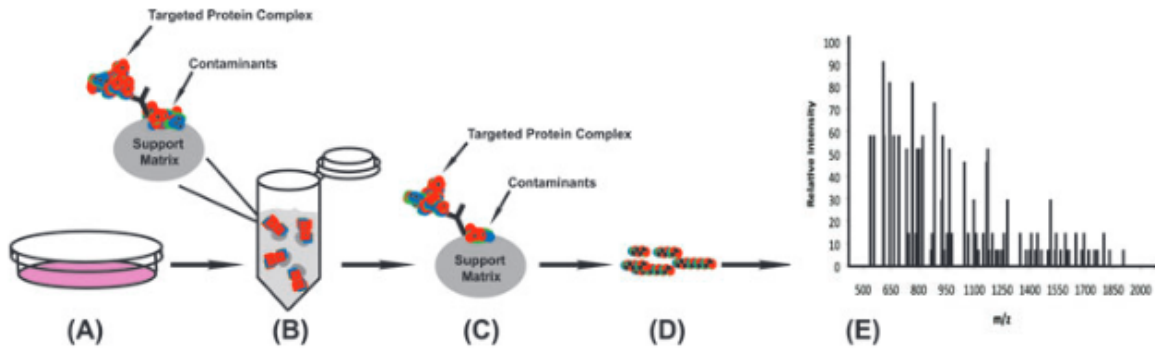


Figure 12.7: AP-MS [135]

When comparing both methods, it seems that for this research Y2H is preferred: the drawback of the binary interactions is not a limitation in this case. It is actually the binary interaction that needs to be demonstrated. Even more, it could be possible that a mutation disturbs a particular interaction but that both proteins are still part of the complex due to their interactions with other proteins of the complex. This cannot be captured by AP-MS. However, when it is known that a mutation perturbs an interaction, AP-MS can be used to investigate whether both proteins are still part of the complex.

12.4.2 Increasing the prediction accuracy

As indicated already in chapter 3.3, the prediction tools used are not always optimal in predicting the effect of nsSNPs on PPIs. Therefore, the results of different methods could be used together to increase the performance: combining the results of all algorithms into a new classifier could yield predictions that are more accurate than the single predictions. To do so, a deeper understanding of the algorithms that exist currently, and of their results, is needed. The focus should mainly be on the four algorithms that seem to perform best at the moment: methods by Fernandez-Recio, Weng, Dehouck and Rooman, and the SNP-IN tool by Zhao [24] [25]. Because all methods still have trouble predicting mutations that strengthen binding, extra care should be taken here [25].

In addition to combining the results of different prediction tools, the prediction results of different PDB files can also be combined. For now, a different result is obtained for every specific interaction of which structural information is known. However, lots of complexes are investigated more than once, and have thus slightly different structural information. Also, lots of interactions appear more than once in one structure. However, it can be expected that the effect of a mutation in interaction with a certain gene will be the same. This makes it possible to combine all results of the different specific interactions into one output that has an accuracy higher than a prediction based on one structure.

Finally, there is one thing that none of the tools take into account: evolutionary conservation. As explained in chapter 3.1, the more conserved residues are thought to be more important. Also, it is expected that interface residues are more conserved than other surface residues [15]. Sahni e.a. also discovered that interaction-perturbing mutations are more conserved across species [133]. Therefore, it makes sense to also investigate the conservation of the mutated residues and to take this into account when making a prediction.

Bibliography

- [1] Wikipedia. Saccharomyces cerevisiae. http://en.wikipedia.org/wiki/Saccharomyces_cerevisiae. Accessed: 2014-12-30.
- [2] John D. McPherson e.a. A physical map of the human genome. *Nature*, 2001.
- [3] Wu Wei e.a. Genome sequencing and comparative analysis of saccharomyces cerevisiae strain yjm789. *Proceedings of the National Academy of Sciences of the USA*, 2007.
- [4] David Botstein and Gerald R. Fink. Yeast: An experimental organism for 21st century biology. *Genetics*, 2011.
- [5] Ceyda Kasavi, Serpil Eraslan, Kazim Yalcin Arga, Ebru Toksoy Oner, and Betul Kirdar. A system based network approach to ethanol tolerance in saccaromyces cerevisiae. *BMC Systems Biology*, 2014.
- [6] Thiago M. Pais, Maria R. Foulquie-Moreno, Georg Hubmann, Jorge Duitama, Steve Swinnen, Annelies Goovaerts, Yudi Yang, Francoise Dumortier, and Johan M. Thevelein. Comparative polygenic analysis of maximal ethanol accumulation capacity and tolerance to high ethanol levels of cell proliferation in yeast. *PLOS Genetics*, 2013.
- [7] D. Stanley, A. Bandara, S. Fraser, P.J. Chambers, and G.A. Stanley. The ethanol stress response and ethanol tolerance of saccharomyces cerevisiae. *Journal of Applied Microbiology*, 2010.
- [8] Laurence A. Moran, Robert A. Horton, Gray Scrimgeour, and Marc Perry. *Principles of Biochemistry*.
- [9] Dongrong Chen, W. Mark Toone, Juan Mata, Rachel Lyne, Gavin Burns, Katja Kivinen, Alvis Brazma, Nic Jones, and Jurg Bahler. Global transcriptional responses of fission yeast to environmental stress. *Molecular Biology of the Cell*, 2003.
- [10] Wikipedia. Acetaldehyde. <http://en.wikipedia.org/wiki/Acetaldehyde>. Accessed: 2015-01-03.
- [11] Katsuhide Fujita, Akinobu Matsuyama, Yoshinori Kobayashi, and Hitoshi Iwahashi. The genome-wide screening of yeast deletion mutants to identify the genes required for tolerance to ethanol and other alcohols. *FEMS Yeast Res*, 2006.

- [12] Menggen Ma and Z. Lewis Lius. Mechanisms of ethanol tolerance in *saccharomyces cerevisiae*. *Applied Microbiology and Biotechnology*, 2010.
- [13] Ncbi molecular biology review the central dogma. <http://www.edupdfinfo.com/blog/ncbi-molecular-biology-review-the-central-dogma>. Accessed: 2015-02-12.
- [14] Irene M.A. Nooren and Janet M. Thornton. Diversity of protein-protein interactions. *The EMBO Journal*, 2003.
- [15] Changhui Yan, Feihong Wu, Robert L. Jernigan, Drena Dobbs, and Vasant Honavar. Characterization of protein-protein interfaces. *The Protein Journal*, 2008.
- [16] Daniel R. Caffrey, Shyamal Somaroo, Jason D. Hughes, Julian Mintseris, and Enoch S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 2004.
- [17] Irene M.A. Nooren and Janet M. Thornton. Structural characterisation and functional significance of transient protein-protein interactions. *Journal of molecular biology*, 2003.
- [18] A. Keith Dunker, Marc S. Cortese, Pedro Romero, Lilia M. Iakoucheva, and Vladimir N. Uvesky. Flexible nets. the roles of intrinsic disorder in protein interaction networks. 2005.
- [19] Susan Jones and Janet M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the USA*, 1996.
- [20] Alessia David, Rozami Razali, Mark N. Wass, and Michael J.E. Sternberg. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous snps. *Human Mutation*, 2011.
- [21] Conserved sequence. http://en.wikipedia.org/wiki/Conserved_sequence. Accessed: 2015-03-10.
- [22] Nature. Snp. <http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295>. Accessed: 2014-12-29.
- [23] Christopher M. Yates and Michael J. E. Sternberg. The effects of non-synonymous single nucleotide polymorphisms (nssnps) on protein-protein interactions. *Journal of molecular biology*, 2013.
- [24] Rocco Moretti, Sarel J. Fleishman, et al. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. *Proteins: Structure, Function and Bioinformatics*, 2013.
- [25] Nan Zhao, Jing Ginger Han, Chi-Ren Shyu, and Dmitry Korkin. Determining effects of non-synonymous snps on protein-protein interactions using supervised and semi-supervised learning. *PLOS Computational Biology*, 2014.

- [26] Foldx. <http://foldx.crg.es/>. Accessed: 2015-02-17.
- [27] Thom Vreven, Howook Hwang, Brian G. Pierce, and Zhiping Weng. Prediction of protein-protein binding free energies. *Protein Science*, 2012.
- [28] Yves Dehouck, Jean Marc Kwasigroch, Dimitri Gilis, and Marianne Rooman. Pop-music 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 2011.
- [29] RCSB. Protein data bank. <http://www.rcsb.org/pdb/home/home.do>. Accessed: 2014-11-26.
- [30] Mark Pilgrim. Extracting data from html documents. http://www.diveintopython.net/html_processing/extracting_data.html. Accessed: 2014-11-12.
- [31] Leonard Richardson. Beautiful soup documentation - beautiful soup 4.2.0 documentation. <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>. Accessed: 2014-11-12.
- [32] Biomart central portal - id conversion. http://central.biomart.org/converter/#!/ID_converter/gene_ensembl_config_2. Accessed: 2015-02-18.
- [33] Cenpk1137d_4440 - cdc39p - *saccharomyces cerevisiae* (strain cen.pk113-7d) (baker's yeast). <http://www.uniprot.org/uniprot/N1P976>. Accessed: 2015-02-25.
- [34] *Saccharomyces* genome database. <http://www.yeastgenome.org/>. Accessed: 2015-02-19.
- [35] Uniprot. <http://www.uniprot.org/>. Accessed: 2015-02-25.
- [36] *S. cerevisiae* (*saccharomyces cerevisiae*) genome browser gateway. <http://genome-euro.ucsc.edu/cgi-bin/hgGateway?redirect=auto&source=genome.ucsc.edu>. Accessed: 2015-02-19.
- [37] Hsp90ab1 - heat shock protein hsp 90-beta - *homo sapiens* (human). <http://www.uniprot.org/uniprot/P08238>. Accessed: 2015-02-25.
- [38] Rps0a — sgd. <http://www.yeastgenome.org/locus/S000003446/overview>. Accessed: 2015-02-25.
- [39] Indel. <http://en.wikipedia.org/wiki/Indel>. Accessed: 2015-02-23.
- [40] Biomart central portal - search by database name. http://central.biomart.org/martwizard/#!/Search_by_database_name?mart=Ensembl+78+Genes+%28WT%2C+UK%29&datasets=scerevisiae_gene_ensembl. Accessed: 2015-02-18.
- [41] *S. cerevisiae* gene imp2' (yil154c) description and page index. http://genome-euro.ucsc.edu/cgi-bin/hgGene?hgg_gene=YIL154C&hgg_chrom=chrIX&hgg_start=53980&hgg_end=55021&hgg_type=sgdGene&db=sacCer3&hgside=202736754_YW8FVYSEZfTVdt1zw0awJXixoj3G. Accessed: 2015-02-19.

- [42] Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barret, and John N Weinstein. Mistaken identifiers: Gene name errors can be introduced inadvertently when using excel in bioinformatics. *BMC Bioinformatics*, 2004.
- [43] Julie Parenteau, Mathieu Durand, Steeve Véronneau, Andrée-Anne Lacombe, Geneviève Morin, Valérie Guérin, Bojana Cecez, Julien Gervais-Bird, Chu-Shin Koh, David Brunelle, Raymund J. Wellinger, Benoit Chabot, and Sherif Abou Elela. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Molecular Biology of the Cell*, 2008.
- [44] Jason E. Stajich, Fred S. Dietrich, and Scott W. Roy. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biology*, 2007.
- [45] Yeastmine: Template query: All genes of a selected feature type—genes with introns. http://yeastmine.yeastgenome.org/yeastmine/template.do?name=Gene_FeatureType_Introns&scope=all. Accessed: 2015-02-27.
- [46] nonsense mutation — learn science at scitable. <http://www.nature.com/scitable/definition/nonsense-mutation-228>. Accessed: 2015-03-18.
- [47] silent mutation — learn science at scitable. <http://www.nature.com/scitable/definition/silent-mutation-10>. Accessed: 2015-03-18.
- [48] Biomart portal. http://central.biomart.org/sequence/?gui=Sequence_retrieval&mart=metaseq_mart_63_config. Accessed: 2015-04-21.
- [49] Home - pubmed - ncbi. <http://www.ncbi.nlm.nih.gov/pubmed>. Accessed: 2015-04-02.
- [50] String: functional protein association networks. <http://string-db.org/>. Accessed: 2015-04-02.
- [51] Kegg pathway database. <http://www.genome.jp/kegg/pathway.html>. Accessed: 2015-03-19.
- [52] Pymol — www.pymol.org. <https://www.pymol.org/>. Accessed: 2015-04-02.
- [53] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 1992.
- [54] Blast: Basic local alignment search tool. <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. Accessed: 2015-04-22.
- [55] Vps74 — sgd. <http://www.yeastgenome.org/locus/S000002780/overview>. Accessed: 2015-03-25.
- [56] Katsunori Yoshikawa, Tadamasa Tanaka, Chikara Furusawa, Keisuke Nagahisa, Takashi Hirasawa, and Hiroshi Shimizu. Comprehensive phenotypic analysis for identification of genes affecting growth under ethanol stress in *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 2009.

- [57] Lydie Michailat and Andreas Mayer. Identification of genes affecting vacuole membrane fragmentation in *saccharomyces cerevisiae*. *PLOS ONE*, 2013.
- [58] Yiying Cai, Yongqiang Deng, Florian Horenkamp, Karin M. Reinisch, and Christopher G. Burd. Sac1-vps74 structure reveals a mechanism to terminate phosphoinositide signaling in the golgi apparatus. *The Journal of Cell Biology*, 2014.
- [59] Samuel Furse, Nicholas J. Brooks, Annela M. Seddon, Rüdiger Woscholski, Richard H. Templer, Edward W. Tate, Piers R.J. Gaffney, and Oscar Ces. Lipid membrane curvature induced by distearoyl phosphatidylinositol 4-phosphate. *The Royal Society of Chemistry*, 2012.
- [60] Vps60 — sgd. <http://www.yeastgenome.org/locus/S000002894/overview>. Accessed: 2015-03-24.
- [61] Zhongzheng Yang, Cody Vild, Jiaying Ju, Xu Zhang, Jianping Liu, Jie Shen, Bin Zhao, Wenxian Lan, Fuchun Gong, Maili Liu, Chunyang Cao, and Zhaohui Xu. Structural basis of molecular recognition between escrt-iii-like protein vps60 and aaa-atpase regulator vta1 in the multivesicular body pathway. *The Journal of Biological Chemistry*, 2012.
- [62] Vta1 — sgd. <http://www.yeastgenome.org/locus/S000004171/overview>. Accessed: 2015-03-24.
- [63] Philip G. Meaden, Nils Arneborg, Lars U. Guldfieldt, Hendrik Siegumfeldt, and Mogens Jakobsen. Endocytosis and vacuolar morphology in *saccharomyces cerevisiae* are altered in response to ethanol stress or heat shock. *Yeast*, 1999.
- [64] Add66 — sgd. <http://www.yeastgenome.org/locus/S000001689/overview>. Accessed: 2015-03-24.
- [65] Vma4 — sgd. <http://www.yeastgenome.org/locus/S000005859/overview>. Accessed: 2015-03-24.
- [66] The michael forgac lab — sackler. <http://sackler.tufts.edu/Faculty-and-Research/Faculty-Research-Pages/Michael-Forgac>. Accessed: 2015-03-24.
- [67] Ryan C. Hunt, Vijaya L. Simhadri, Matthew Iandoli, Zuben E. Sauna, and Chava Kimchi-Sarfaty. Exposing synonymous mutations. *Trends in Genetics*, 2014.
- [68] Julie L. Chaney and Patricia L. Clarck. Roles for synonymous codon usage in protein biogenesis. *Annual Review of Biophysics*, 2015.
- [69] *S. cerevisiae* codong usage tables. http://www.yeastgenome.org/community/codon_usage.shtml. Accessed: 2015-03-27.
- [70] Pop2 — sgd. <http://www.yeastgenome.org/locus/S000005335/overview>. Accessed: 2015-03-27.

- [71] Ubi4 — sgd. <http://www.yeastgenome.org/locus/S000003962/overview>. Accessed: 2015-03-31.
- [72] *Stress Proteins*, chapter Ubiquitin and the Stress Response. Springer, 1999.
- [73] Vps27 — sgd. <http://www.yeastgenome.org/locus/S000005289/overview>. Accessed: 2015-03-31.
- [74] Qcr2 — sgd. <http://www.yeastgenome.org/locus/S000006395/overview>. Accessed: 2015-03-27.
- [75] Swi/snf - wikipedia, the free encyclopedia. <http://en.wikipedia.org/wiki/SWI/SNF>. Accessed: 2015-03-27.
- [76] Snf2 — sgd. <http://www.yeastgenome.org/locus/S000005816/overview>. Accessed: 2015-03-27.
- [77] Hst2 — sgd. <http://www.yeastgenome.org/locus/S000005936/overview>. Accessed: 2015-03-27.
- [78] Rps5 — sgd. <http://www.yeastgenome.org/locus/S000003884/overview>. Accessed: 2015-03-31.
- [79] Rcsb pdb-101. <http://www.rcsb.org/pdb/101/motm.do?momID=10>. Accessed: 2015-03-31.
- [80] Uba1 — sgd. <http://www.yeastgenome.org/locus/S000001693/overview>. Accessed: 2015-03-19.
- [81] Antje Schäfer, Monika Kuhn, and Hermann Schindelin. Structure of the ubiquitin-activating enzyme loaded with two ubiquitin molecules. *Acta Crystallographica*, 2014.
- [82] *Atlas of Protein Sequence and Structure*, chapter A Model of Evolutionary Change in Proteins. 1978.
- [83] Vernon R. Young and Alfred M. Ajami. Glutamate: an amino acid of particular distinction. In *The Journal of Nutrition*.
- [84] *Bioinformatics for Geneticists*, chapter Amino Acid Properties and Consequences of Substitutions. John Wiley and Sons, 2003.
- [85] Rpo21 — sgd. <http://www.yeastgenome.org/locus/S000002299/overview>. Accessed: 2015-03-23.
- [86] Rpb5 — sgd. <http://www.yeastgenome.org/locus/S000000358/overview>. Accessed: 2015-03-23.
- [87] Volker Stoldt, Felicitas Rademacher, Verena Kehren, Joachim F. Ernst, David A. Pearce, and Fred Sherman. Review: The cct eukaryotic chaperonin subunits of *saccharomyces cerevisiae* and other yeasts. *Yeast*, 1998.

- [88] Alexander Leitner, Lukasz A. Joachimiak, Andreas Bracher, Leonie Mönkemeyer, Thomas Walzthoeni, Bryan Chen, Sebastian Pechmann, Susan Holmes, Yao Cong, Boxue Ma, Steve Ludtke, Wah Chiu, F. Ulrich Hartl, Ruedi Aebersold, and Judith Frydman. The molecular architecture of the eukaryotic chaperonin tric/cct. *Structure*, 2012.
- [89] Bogdan Polevoda and Fred Sherman. Composition and function of the eukaryotic n-terminal acetyltransferase subunits. *Biochemical and biophysical research communications*, 2003.
- [90] Ard1 — sgd. <http://www.yeastgenome.org/locus/S0000001055/overview>. Accessed: 2015-04-13.
- [91] Rrp43 — sgd. <http://www.yeastgenome.org/locus/S0000000631/overview>. Accessed: 2015-04-14.
- [92] Mtr3 — sgd. <http://www.yeastgenome.org/locus/S0000003390/overview>. Accessed: 2015-04-14.
- [93] Atg10 — sgd. <http://www.yeastgenome.org/locus/S0000003965/overview>. Accessed: 2015-04-14.
- [94] Stephen E Kaiser, Kai Mao, Asad M Taherbhoy, Shanshan Yu, Jennifer L Olszewski, David M Duda, Igor Kurinov, Alan Deng, Timothy D Fenn, Daniel J Klionsky, and Brend A Schulman. Noncanonical e2 recruitment by the autophagy e1 revealed by atg7-atg3 and atg7-atg10 structures. *Nature Structural and molecular biology*, 2012.
- [95] Nobuo N. Noda and Fuyuhiko Inagaki. Mechanisms of autophagy. *The Annual Review of Biophysics*, 2015.
- [96] Michael D. Blower. Molecular insights into intracellular rna localization. *The International Review of Cell and Molecular Biology*, 2013.
- [97] Hang Shi, Nimisha Singh, Filipp Esselborn, and Günter Blobel. Structure of a myosin adaptor complex and pairing by cargo. *PNAS*, 2014.
- [98] She3 — sgd. <http://www.yeastgenome.org/locus/S0000000334/overview>. Accessed: 2015-04-16.
- [99] Reiko Takemura, Yoshiharu, and Shingo Izawa. Stress response in yeast mrna export factor: reversible changes in rat8p localization are caused by ethanol stress but not heat shock. *Journal of Cell Science*, 2004.
- [100] Fis1 — sgd. <http://www.yeastgenome.org/locus/S0000001327/overview>. Accessed: 2015-04-16.
- [101] Yan Zhang, Nickie C. Chan, Huu B. Ngo, Harry Gristick, and David C. Chan. Crystal structure of mitochondrial fission complex reveals scaffolding function for mitochondrial division 1 (mdv1) coiled coil. *Journal of Biological Chemistry*, 2012.

- [102] Mdv1 — sgd. <http://www.yeastgenome.org/locus/S000003648/overview>. Accessed: 2015-04-16.
- [103] Hiroshi Kitagaki, Yoshio Araki, Kouichi Funato, and Hitoshi Shimoi. Ethanol-induced death in yeast exhibits features of apoptosis mediated by mitochondrial fission pathway. *FEBS Letters*, 2007.
- [104] Rps10a — sgd. <http://www.yeastgenome.org/locus/S000005819/overview>. Accessed: 2015-04-20.
- [105] Rps3 — sgd. <http://www.yeastgenome.org/locus/S000005122/overview>. Accessed: 2015-04-20.
- [106] Ki Moon Seong, Sang-Oun Jung, Hag Dong Kim, Hee Ju Kim, You-Jin Jung, Soo-Young Choi, and Joon Kim. Yeast ribosomal protein s3 possesses a β -lyase activity on damaged dna. *FEBS Letters*, 2012.
- [107] Pup1 — sgd. <http://www.yeastgenome.org/locus/S000005683/overview>. Accessed: 2015-04-20.
- [108] Cassandra S. Arendt and Marck Hochstrasser. Eukaryotic 20s proteasome catalytic subunit propeptides prevent active site inactivation by n-terminal acetylation and promote particle assembly. *The EMBO Journal*, 1999.
- [109] Pre8 — sgd. <http://www.yeastgenome.org/locus/S000004557/overview>. Accessed: 2015-04-20.
- [110] Pba1 — sgd. <http://www.yeastgenome.org/locus/S000004189/overview>. Accessed: 2015-04-01.
- [111] Spc25 — sgd. <http://www.yeastgenome.org/locus/S000000820/overview>. Accessed: 2015-03-31.
- [112] Francesca Malvezzi, Gabriele Litos, Alexander Schleiffer, Alexander Heuck, Karl Mechtler, Tim Clausen, and Stefan Westermann. A structural basis for kinetochore recruitment of the ndc80 complex via two distinct centromere receptors. *The EMBO Journal*, 2013.
- [113] Karen E. Gascoigne and Iain M. Cheeseman. T time for point centromeres. *Nature Cell Biology*, 2012.
- [114] Beth M. Stadtmueller, Erik Kish-Trier, Katherine Ferrell, Charisse N. Petersen, Howard Robinson, David G. Myszka, Debra M. Eckert, Tim Formosa, and Christopher P. Hill. Structure of a proteasome pba1-pba2 complex. *The Journal of Biological Chemistry*, 2012.
- [115] Mathieu Frechin, Bruno Senger, Mélanie Brayé, Daniel Kern, Robert Pierre Martin, and Hubert Dominique Becker. Yeast mitochondrial gln-trna^{Gln} is generated by a gatfab-mediated transamidation pathway involving arc1p-controlled subcellular sorting of cytosolic glurs. *Genes and development*, 2009.

- [116] Yuhei Arais, Jonathan L Huot, Takuya Sekiguchi, Mathieu Frechin, Frédéric Fischer, Ludovic Enkler, Bruno Senger, Ryuichiro Ishitani, Hubert D. Becker, and Osamu Nureki. Crystal structure of *Saccharomyces cerevisiae* mitochondrial gatfab reveals a novel subunit assembly in trna-dependent amidotransferases. *Nucleic Acids Research*, 2014.
- [117] Gtf1 — sgd. <http://www.yeastgenome.org/locus/S000003334/overview>. Accessed: 2015-04-08.
- [118] Her2 — sgd. <http://www.yeastgenome.org/locus/S000004907/overview>. Accessed: 2015-04-08.
- [119] Qcr7 — sgd. <http://www.yeastgenome.org/locus/S000002937/overview>. Accessed: 2015-04-08.
- [120] Felix Halbach, Peter Reichelt, Michaela Rode, and Elena Conti. The yeast ski complex: Crystal structure and rna channeling to the exosome complex. *Cell*, 2013.
- [121] Ski3 — sgd. <http://www.yeastgenome.org/locus/S000006393/overview>. Accessed: 2015-04-09.
- [122] Rpn8 — sgd. <http://www.yeastgenome.org/locus/S000005787/overview>. Accessed: 2015-04-09.
- [123] Rpn5 — sgd. <http://www.yeastgenome.org/locus/S000002306/overview>. Accessed: 2015-04-09.
- [124] Rpn9 — sgd. <http://www.yeastgenome.org/locus/S000002835/overview>. Accessed: 2015-04-09.
- [125] Qcr9 — sgd. <http://www.yeastgenome.org/locus/S000003415/overview>. Accessed: 2015-04-09.
- [126] Rpl2b — sgd. <http://www.yeastgenome.org/locus/S000001280/overview>. Accessed: 2015-04-13.
- [127] Matthias Müller, Kaihui Lu, and Andreas S Reichert. Mitophagy and mitochondrial dynamics in *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta*, 2015.
- [128] Donald D. Newmeyer and Shelagh Ferguson-Miller. Mitochondria: Releasing power for life and unleashing the machineries of death. *Cell*, 2003.
- [129] C. Auesukaree, A. Damnernsawad, M. Kruatrachue, P. Pokethitiyook, C. Boonchird, Y. Kaneko, and S. Harashima. Genome-wide identification of genes involved in tolerance to various environmental stress in *Saccharomyces cerevisiae*. *Journal of Applied Genetics*, 2009.
- [130] Miguel C. Teixeira, Luís R. Raposo, Nuno P. Mira, artur B. Lourenco, and Isabel Sá-Correia. Genome-wide identification of *Saccharomyces cerevisiae* genes required for maximal tolerance to ethanol. *Applied and environmental microbiology*, 2009.

- [131] Rembrandt J. F. Haft, David H. Keating, Tyler Schwaegler, Michael S. Schwalbach, Jeffrey Vinokur, Mary Tremaine, Jason M. Peters, Matthew V. Kotlajich, Edward L. Pohlmann, Irene M. Ong, Jeffrey A. Grass, Patricia J. Kiley, and Robert Landick. Correcting direct effects of ethanol on translation and transcription machinery confers ethanol tolerance in bacteria. *PNAS*, 2014.
- [132] Rho factor - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Rho_factor. Accessed: 2015-05-05.
- [133] Nidhi Sahni, Song Yi, Mikko Taipale, Juan I. Fuxman Bass, Jasmin Coulombe-Huntington, Fan Yang, Jian Peng, Jochen Weile, Georgios I. Karras, Yang Wang, István A. Kovács, Atanas Kamburov, Irina Krykbaeva, Mandy H. Lam, George Tucker, Vikram Khurana, Amitabh Sharma, Yang-Yu Liu, Nozomu Yachie, Quan Zhong, Yun Shen, Alexandre Palagi, Adriana San-Miguel, Changyu Fan, Dawit Balcha, Amelie Dricot, Daniel M. Jordan, Jennifer M. Walsh, Akash A. Shah, Xinpeng Yang, Ani K. Stoyanova, Alex Leighton, Michael A. Calderwood, Yves Jacob, Michael E. Cusick, Kouros Salehi-Ashtiani, Luke J. Whitesell, Shamil Sunyaev, Bonnie Berger, Albert-László Barabási, Benoit Charlotiaux, David E. Hill, Tong Hao, Frederick P. Roth, Yu Xia, Albertha J.M. Walhout, Susan Lindquist, and Marc Vidal. Widespread macromolecular interaction perturbation in human genetic disorders. *Cell*, 2015.
- [134] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 2009.
- [135] Wade H. Dunham, Michael Mullin, and Anne-Claude Gingras. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics*, 2012.

Appendix A

Detailed prediction results

This appendix lists all prediction results obtained from the SNP-IN tool and PoPMuSiC [25] [28]. The following information can be found in the table:

- The gene in which the mutation is observed.
- The mutation that has occurred.
- The gene that interacts with the mutated gene at the mutated residue.
- The PDB identifier of the structural information that is used to come to these results.
- The chain identifier of the mutated gene.
- The chain identifier of the interacting gene.
- A personal opinion based on literature and similarity of the original and the mutated residue. This opinion can be that the interaction is preserved (P), that it is lost (NP) or no clear conclusions can be drawn from the obtained knowledge (P/NP).
- Results of the SNP-IN tool. The first two columns indicate the classification of the 2-class problem and its likelihood (the mutation disrupts (D) or preserves (P) the interaction); and the other two columns show the classification of the 3-class problem and its likelihood (the mutation is detrimental (D), neutral (N) or beneficial (B) to the interaction).
- Result of PoPMuSiC. This tool outputs a difference in free energy. When this difference is smaller than 0, the mutation is assigned to preserve the interaction (P), otherwise the interaction is predicted to be lost (NP).

GENE	MUTATION	INTERACTING GENE	PDB ID	CHAIN ID OF GENE	CHAIN ID OF INTERACTING GENE	OPINION	SNP-IN				PoPMuSiC	
							2 CLASS	LIKELIHOOD	3-CLASS	LIKELIHOOD	RESULT	ddG
UBA1	E910D	UBI4	4NNJ	A C	B D	P	P P	0.56 0.61	D D	0.72 0.66	NP	1.5
RPB5	R212K	RPO21	3RZD	E	A	P	D	0.94	D	0.97	NP	2.24
			3S14				D	0.55	D	0.77	NP	2.18
			3S2H				D	0.98	D	0.91	NP	2.04
			4A3K				D	0.95	D	0.97	NP	2.22
			3RZO				D	0.75	D	0.89	NP	2.45
			3S2D				D	0.94	D	0.87	NP	2.4
			4A3F				D	0.96	D	0.97	NP	2.07
CCT8	G279D	CCT6	4D8R	h	f	P					NP	0.65
			D	n								
			H	F						NP	0.65	
ARD1	F204Y	NAT1	4HNX	B	A	P	P	0.79	N	0.57	NP	0.45
			4HNY	B	A		P	0.85	D	0.4	NP	0.48
			D	D	C		P	0.86	N	0.56	NP	0.48
RRP43	V218I	MTR3	4IFD	C	F	P	P	0.92	N	0.74	NP	0.62
			4OO1	C	F		P	0.92	N	0.65	NP	0.44
ATG10	R41S	ATG7	4GSK	Z Y	A B	NP	D	0.92	D	0.92	NP	1.8
D	D	0.88	D	0.93	NP		1.8					
SHE3	N106T	SHE3	4LL8	B E	E B	P/NP	P D	0.65 0.77	D D	0.52 0.78	NP NP	0.38 0.38
MDV1	N260I	FIS1	3UUX	B D	A C	P/NP	D P	0.56 0.97	N N	0.52 0.82	P P	-0.16 -0.16
RPS10A	Q58L	RPS3	4ULJ	L	E	P/NP					NP	0.14
			4UJK	L	E						NP	0.14
			4ULB	o	h							
			4UKP	L	E						NP	0.22
			4UKR	o	h							
			4UKK	L	E						NP	0.24
			4UJP	L	E						NP	0.3
			4UJR	o	h							
			4UJH	o	h							
			4UKB	o	h							
			4ULP	L	E						NP	0.24
			4UL6	o	h							
			4UJM	o	h							
			4UKZ	L	E						NP	0.31
			4ULG	o	h							
			4UL4	L	E						NP	0.18
			4ULE	L	E						NP	0.05
			3U5G	K	D						NP	0.28
			4ULR	o	h							
			4UL9	L	E						NP	0.21
			4UKW	o	h							
			4UK9	L	E						NP	0.19
			4UJW	o	h							
			4UJF	L	E						NP	0.33
			4UK6	o	h							
			4UKF	L	E						NP	0.22
			4UKM	o	h							
			4ULM	o	h							
			3U5C	K	D						NP	0.35
			4UL1	o	h							
			4UKH	o	h							
			4UJU	L	E						NP	0.49
4UKU	L	E					NP	0.31				
4UK4	L	E					NP	0.3				

GENE	MUTATION	INTERACTING GENE	PDB ID	CHAIN ID OF GENE	CHAIN ID OF INTERACTING GENE	OPINION	SNP-IN				PoPMuSiC	
							2 CLASS	LIKELIHOOD	3-CLASS	LIKELIHOOD	RESULT	ddG
PUP1	A214T	PRE8	1G65	H	A	P	P	0.7	D	0.79	NP	0.66
				V	O		P	0.95	D	0.69	NP	0.66
			3OEU	H	A		P	0.85	D	0.68	NP	0.77
				V	O		D	0.81	D	0.89	NP	0.77
			3OEV	H	A		D	0.95	D	0.9	NP	0.82
				V	O		D	0.86	D	0.88	NP	0.82
			3SDI	H	A		P	0.93	D	0.77	NP	0.96
				V	O		P	0.52	D	0.79	NP	0.96
			3SDK	H	A		D	0.97	D	0.9	NP	0.84
				V	O		D	0.95	D	0.89	NP	0.84
			3SHJ	H	A		D	0.95	D	0.98	NP	1
				V	O		D	0.92	D	0.94	NP	1
			3TDD	H	A		D	0.91	D	0.98	NP	0.76
				V	O		D	0.72	D	0.84	NP	0.76
			3UN4	H	A						NP	0.86
				V	O						NP	0.86
			3UN8	H	A						NP	0.7
				V	O						NP	0.7
			3WXR	I	B		P	0.63	D	0.85	NP	0.7
				W	P		D	0.99	D	0.95	NP	0.7
			4EU2	I	B		D	0.91	D	0.93	NP	0.82
				W	P		P	0.86	D	0.82	NP	0.82
			4FZC	H	A		P	0.93	N	0.48	NP	0.86
				V	O		P	0.83	D	0.44	NP	0.86
			4FZG	H	A		D	0.91	D	0.89	NP	0.93
				V	O		D	0.67	D	0.83	NP	0.93
			4G4S	I	B		D	0.91	D	0.91	NP	0.59
				I	B		D	0.86	D	0.91	NP	0.8
			4GK7	H	A		D	0.89	D	0.86	NP	0.8
				V	O		D	0.91	D	0.89	NP	0.65
			4HNP	H	A		P	0.67	D	0.78	NP	0.65
				V	O		P	0.69	D	0.7	NP	0.68
			4HRC	H	A		D	0.89	D	0.93	NP	0.68
				V	O						NP	0.78
			4HRD	H	A						NP	0.78
				V	O						NP	0.78
			4JSQ	H	A		D	0.91	D	0.88	NP	0.51
				V	O		D	0.57	D	0.76	NP	0.51
			4JSU	H	A		D	0.93	D	0.96	NP	0.72
				V	O		D	0.69	D	0.78	NP	0.72
			4JT0	H	A		P	0.88	D	0.5	NP	0.64
				V	O		P	0.85	D	0.75	NP	0.64
			4LTC	H	A						NP	0.76
				V	O						NP	0.76
			4NNN	H	A						NP	0.92
				V	O						NP	0.92
			4NNW	H	A		P	0.84	D	0.36	NP	0.76
				V	O		P	0.89	D	0.4	NP	0.76
			4NO1	H	A		P	0.82	D	0.69	NP	0.82
				V	O		P	0.85	D	0.42	NP	0.82
			4NO6	H	A		P	0.92	D	0.46	NP	0.74
				V	O		P	0.92	D	0.48	NP	0.74
			4NO8	H	A		P	0.88	D	0.55	NP	0.71
				V	O		P	0.84	D	0.65	NP	0.71
			4NO9	H	A		P	0.85	D	0.48	NP	0.68
				V	O		P	0.89	B	0.52	NP	0.68
			4Q1S	H	A		D	0.91	D	0.92	NP	0.87
				V	O		D	0.96	D	0.96	NP	0.87
			4QBY	H	A		P	0.9	D	0.45	NP	0.75
				V	O		P	0.91	D	0.53	NP	0.75
			4QLQ	H	A						NP	0.73
				V	O						NP	0.73
			4QLS	H	A		P	0.9	D	0.59	NP	0.69
				V	O		P	0.82	D	0.57	NP	0.69
			4QLT	H	A		D	0.88	D	0.92	NP	0.75
				V	O		D	0.91	D	0.95	NP	0.75
			4QLU	H	A		D	0.96	D	0.95	NP	0.44
				V	O		D	0.94	D	0.94	NP	0.44
			4QLV	H	A		P	0.87	D	0.61	NP	0.86
				V	O		P	0.67	D	0.82	NP	0.86
			4R02	H	A							
				V	O							
			4R17	H	A							
				V	O							
			4R18	H	A							
				V	O							

GENE	MUTATION	INTERACTING GENE	PDB ID	CHAIN ID OF GENE	CHAIN ID OF INTERACTING GENE	OPINION	SNP-IN				PoPMuSiC	
							2 CLASS	LIKELIHOOD	3-CLASS	LIKELIHOOD	RESULT	ddG
RPO21 POS 320	R320C	RPB2	3RZD	A	B	P	P	0.58	D	0.73	NP	1.05
			3S14	A	B		D	0.56	D	0.75	P	-0.09
			3S2H	A	B		P	0.57	D	0.75	P	-0.17
		RPB5	3RZD	A	E	P	P	0.6	D	0.63	NP	1.05
			3S14	A	E		P	0.58	D	0.58	P	-0.09
			3S2H	A	E		P	0.57	D	0.56	P	-0.17
			3RZO	A	E		P	0.59	D	0.75	NP	0.84
			3S2D	A	E		P	0.6	D	0.6	NP	0.38
SPC25 POS 180	V180G	CNN1	4GEQ	A	F	P	D	0.73	D	0.9	NP	2.9
				C	E		D	0.84	D	0.82	NP	2.9
PBA1	S199C	ADD66	4G4S	O	P	NP	D	0.54	D	0.87	NP	0.38
SPC25 POS 135	A135V	SPC24	4GEQ	A	B	P	P	0.88	N	0.5	NP	0.48
				C	D	P	0.89	N	0.78	NP	0.48	
HER2	V240R	GTF1	4N0I	A	F	NP	D	0.92	D	0.95	NP	2.1
			4N0H	A	F		D	0.93	D	0.97	NP	1.9
QCR7	L43H	COB	4PD4	G	C	P/NP	D	0.99	D	1	NP	2.96
RPO21 POS 817	A817I	RPB2	3RZD	A	B	P	P	0.67	D	0.64	NP	0.68
			3S14	A	B		P	0.82	N	0.37	NP	0.74
			3S2H	A	B		P	0.75	D	0.43	NP	0.57
			4A3K	A	B		P	0.87	D	0.46	NP	0.5
			3RZO	A	B		P	0.69	D	0.64	NP	0.55
			3S2D	A	B		D	0.96	D	0.94	NP	0.57
			4A3F	A	B		D	0.7	D	0.82	NP	0.47
SKI3	T957N	SKI2	4BUJ	B	A	P/NP	P	0.55	D	0.7	NP	1.12
				F	E		P	0.62	D	0.65	NP	1.12
RPN8	G227D	RPN9	3J47	U	O	P	P	0.95	N	0.94	P	-0.09
		RPN5			P		P	0.96	N	0.86	P	-0.09
QCR9	A25T	RIP1	4PD4	I	E	P	P	0.56	D	0.82	NP	0.73
RPL2B	R184H	RPL34A	4B6A	A	p	P/NP	P	0.96	N	0.75	NP	0.89

Summary in layman's terms

A single nucleotide polymorphism, SNP, is the most common form of genetic variation. A SNP is a change in base pair at one position in the DNA and when it is present in a gene-encoding sequence, it can change the amino acid sequence of the encoded protein. This is called a non-synonymous SNP (nsSNP) [22]. When a nsSNP is located at an interaction site it can stabilize or destabilize protein-protein interactions (PPIs) and can lead to a loss, a gain or a change of function [23]. The importance of nsSNPs in PPIs has recently been shown: lots of mutations that cause diseases in humans affect interactions [133].

First the nsSNPs located at an interaction site are extracted out of a SNP data set and then their effect on the interaction is researched. The data set used was obtained in the collaborative project NATAR and contains SNPs of yeast in an environment with increasing ethanol levels. Yeast is able to produce high concentrations of ethanol, and therefore it is used in both the food industry as well as for bio-ethanol production [5]. Not only is the understanding of ethanol tolerance important for economic reasons, biological knowledge in this one-cellular organism can also be transferred to higher-order organisms [20].

To investigate which SNPs are located at an interaction site, all known structures of yeast complexes are used to extract protein parts that belong to an interaction site. This information is linked to the list of SNPs: if the SNP location belongs to the extracted information, the interaction might be changed, so it should be investigated.

When looking at all complexes that are possibly disturbed by SNPs, it is noticed that most of the SNPs occur either in processes related to the production of proteins, either to the breakdown of proteins. The production of a protein is a process called the Central Dogma, which consists of transcribing the DNA into RNA and translating this RNA into an amino acid sequence. It has been seen that ethanol affects both transcription and translation, so it can be expected that the SNPs are a way to correct for what ethanol is doing [131]. Another effect of ethanol is that lots of proteins lose their structure. This can be solved either by trying to fix this or by breaking down the proteins [72]. Therefore it makes sense that the cell is trying to optimize those processes by introducing SNPs in interactions.

A few algorithms have been developed for prediction of the effect of nsSNPs on PPIs, dividing them into three categories based on their effect on the PPI: deleterious, neutral or beneficial. Although those methods are relatively successful in the prediction of deleterious nsSNPs, a lot of improvement still has to be made in successfully predicting beneficial nsSNPs [24] [25]. Further work could be focused on improving the accuracy of predictions by combining results of different prediction methods and of all available structures from the same interaction. Also, the obtained results can be validated by biological experiments, so the predictions can be replaced by knowledge concerning ethanol tolerance in yeast.