RICE UNIVERSITY

Molecular Basis of Gene Dosage Sensitivity

by

Jianping Chen

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Ariel Fernández, Chair
Karl F. Hasselmann Professor
Department of Bioengineering
Rice University

Michael W. Deem

Michael W. Deem, John W. Cox Professor
Department of Bioengineering
Department of Physics and Astronomy
Rice University

Jianpeng Ma, Professor
Bioengineering at Rice University
Biochemistry at Baylor College of Medicine

Laura Segatori

Laura Segatori, T.N. Law Assistant Professor
Chemical and Biomolecular Engineering
Rice University

HOUSTON, TEXAS

JANUARY 2009

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Molecular Basis of Gene Dosage Sensitivity

by

Jianping Chen

Deviation of gene expression from normal levels has been associated with diseases. Both under- and overexpression of genes could lead to deleterious biological consequences. Dosage balance has been proposed to be a key issue of determining gene expression phenotype. Gene deletion or overexpression of any component in a protein complex produces abnormal phenotypes. As a result, interacting partners should be co-expressed to avoid dosage imbalance effects. The strength of transcriptional co-regulation of interacting partners is supposed to reflect gene dosage sensitivity. Although many cases of dosage imbalance effects have been reported, the molecular attributes determining dosage sensitivity remain unknown. This thesis uses a protein structure analysis protocol to explore the molecular basis of gene dosage sensitivity, and studies the post-transcriptional regulation of dosage sensitive genes.

Solvent-exposed backbone hydrogen bond (SEBH or called as dehydron) provides a structure marker for protein interaction. Protein structure vulnerability, defined as the ratio of SEBHs to the overall number of backbone hydrogen bonds, quantifies the extent to which protein relies on its binding partners to maintain structure integrity. Genes encoding vulnerable proteins need to be highly co-expressed with their interacting partners. Pro-

tein structure vulnerability may hence serves as a structure marker for dosage sensitivity. This hypothesis is examined through the integration of gene expression, protein structure and interaction data sets. Both gene co-expression and protein structure vulnerability are calculated for each interacting subunits from human and yeast complexes. It turns out that structure vulnerability quantifies dosage sensitivity for both temporal phases (yeast) and tissue-specific (human) patterns of mRNA expression, determining the extent of co-expression similarity of binding partners.

Highly dosage sensitive genes encode proteins which are vulnerable to water attack. They are subject to tight post-transcriptional regulation. In human, this extra regulation is achieved through extensive microRNA targeting of genes coding for extremely vulnerable proteins. In yeast, on the other hand, our results imply that such a regulation is likely achieved through sequestration of the extremely vulnerable proteins into aggregated states. The 85 genes encoding extremely vulnerable proteins contain the five confirmed yeast prions. It has been proposed that yeast prion protein aggregation could produce multiple phenotypes important for cell survival in some particular circumstances. These results suggest that extremely vulnerable proteins resorting to aggregation to buffer the deleterious consequences of dosage imbalance. However, a rigorous proof will require a structure-based integration of information drawn from the interactome, transcriptome and post-transcriptional regulome.

Acknowledgments

I would like to thank my advisor, Dr. Ariel Fernández, for his encouragement, guidance and support throughout this work. This thesis would not have been possible without his help. I would also like to thank to my other thesis committee members Dr. Michael Deem, Dr. Jianpeng Ma and Dr. Laura Segatori for consenting to be on my committee.

I would like to thank Dr. Kristina Rogale Plazonic for providing research insights and data support. Thanks to Dr. Alejandro Crespo, Xi Zhang and Natalia Pietrosemoli for invaluable discussions and continuous support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Gene expression phenotype has attracted significant interest, due to the advent of high-throughput techniques such as DNA microarray. These new techniques allow quantitative expression measurement of thousands of genes simultaneously. Expression profiling has been widely used to detect disease-related genes [1, 2]. Many studies have established a relationship between human diseases and specific changes in gene expression [3, 4]. Both under- and overexpression of genes could lead to deleterious biological consequences. Dosage balance has been proposed to be a key issue of determining gene expression phenotype [5]. Spatially or chemically isolated functional modules such as protein complexes are responsible for discrete functions. Therefore, gene deletion or overexpression of any component in a protein complex results in a dosage imbalance, which could lead to disease. According to the dosage balance theory, interacting partners should be co-expressed to avoid dosage imbalance effects, and the strength of transcriptional co-regulation of interacting partners is supposed to reflect dosage sensitivity.

Although numerous dosage imbalance effects have been documented in a variety of species, the molecular attributes determining the magnitude of these deleterious effects (i. e., the dosage sensitivity) remain unknown. We tackle this problem exploiting a structure marker of protein interactions. This structural marker, can be identified by solvent-exposed backbone hydrogen bonds (SEBHs or dehydrons). SEBHs are backbone hydrogen bonds poorly protected by surrounding nonpolar side chains [6]. SEBHs are exposed to water attack, and hence are weakly bonded. However, they can be stabilized upon approach of nonpolar groups. SEBHs are enriched in protein binding interface, and become well wrapped upon protein association [7]. The number of SEBHs quantifies the extent of protein connectivity. The more SEBHs a protein possesses, the more interactive it becomes. In this sense, the number of SEBHs also marks the level of protein structure vulnerability. Proteins rich in SEBHs depend on their binding partners for structural integrity, and are hence structurally vulnerable.

Vulnerable proteins rely on their binding partners to maintain structure integrity. Changes in relative expression levels of vulnerable proteins and their interacting partners are likely to induce dosage imbalance effects. According the dosage balance theory, genes encoding vulnerable proteins are sensitive to dosage changes. This thesis examines this prediction through the integration of gene expression, protein structure and interaction data sets.

In Chapter 2, we give a brief description of gene dosage effects and dosage balance hypothesis. We first provide several examples of gene expression phenotypes. Then we describe dosage balance theory that has been proposed to explain gene dosage effects.

In Chapter 3, we define protein structure vulnerability on the basis of solvent exposed backbone hydrogen bonds. To do this, we first give the statistical definition of this type of hydrogen bonds. Then we discuss their important role of marking protein interactions. At last, we give the definition of protein structure vulnerability.

In Chapter 4, we study gene dosage sensitivity from protein structure perspective. This chapter presents the main results of this thesis. structural marker introduced in Chapter 3 is related to gene co-expression. We show that vulnerability quantifies dosage sensitivity for both temporal phases (yeast) and tissue-specific (human) patterns of mRNA expression, determining the extent of co-expression similarity of binding partners.

In Chapter 5, we discuss the post-transcriptional regulation of expression of genes encoding extremely vulnerable proteins. Gene expression is subject to regulation at the post-transcriptional stage. Genes encoding extremely vulnerable proteins need to be under tight control to avoid gene dosage imbalance effects. This chapter discusses the differences of regulatory mechanisms in human and yeast.

In Chapter 6, we verify the relationship between protein structure vulnerability and gene dosage sensitivity by examining the effect of protein structure vulnerability on gene duplication. According to our prediction, duplicates of genes encoding highly vulnerable proteins should be more likely to cause dosage imbalance and hence be less frequently to be retained in evolution. Therefore, genes encoding vulnerable proteins should have less paralogs than genes encoding proteins with good packing quality. This chapter presents some important results on this issue.

In Chapter 7, we give conclusions and discuss future work. Structure vulnerability provides a molecular basis for gene dosage imbalance effects. Genes encoding extremely vulnerable proteins are subject to strong post-transcriptional regulation. In human, this extra regulation is achieved through extensive microRNA targeting of genes coding for extremely vulnerable proteins. In yeast, on the other hand, our results imply that such a regulation is likely achieved through sequestration of the extremely vulnerable proteins into aggregated states. These results imply that protein aggregation can buffer the deleterious effects of dosage imbalances. This chapter will give a brief discussion about this topic and suggest future work needed to elucidate this puzzle.

Chapter 2

Gene Dosage Effect and Dosage Balance Hypothesis

Gene expression is the process by which the genetic information encoded on DNA is transferred to protein or RNA. The genetic information is not always accurately transferred in the gene expression process. Deviation of gene expression from normal levels, i.e. under- or overexpression, can arise from genetic, environmental, developmental or random biological effects. Gene expression variation usually leads to new phenotypes. Reduced gene dosage produces haploinsufficient effect, while increased gene quantity also results in abnormal phenotype. This chapter gives a brief description of gene expression and gene dosage effects, and then discusses dosage balance theory that is proposed to explain gene dosage effects.

2.1 Gene expression

Gene expression involves two major stages: transcription and translation. At first, DNA sequence is transcribed into a complementary nucleotide RNA strand called messenger RNA (mRNA). Then, the mRNA codon sequence is translated into a chain of amino acids that form a protein.

Gene expression levels can be evaluated by measuring mRNA levels. There are several ways to detect mRNA expression levels. One traditional technique is northern blotting [8], a process in which a sample of RNA separated on an agarose gel is hybridized to a radio-labeled RNA probe that is complementary to the target sequence. Northern blotting quantifies mRNA levels by measuring band strength in an image of a gel, which may result in lower quality data. Despite its shortages, northern blotting is still often used due to some particular benefits it offers, such as the ability to discriminate alternately spliced transcripts.

In contrast to traditional methods which measure mRNA levels individually, modern techniques perform expression profiling in which transcript levels for many genes are measured at once. DNA microarray technology is one widely used expression profiling technique [9]. This high-throughput technology consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides. Each spot contains picomoles of a specific DNA sequence used as probes to hybridize a target (a cDNA or cRNA sample) under high-stringency conditions. Targets are usually fluorophore-labeled and quantified by fluorescence-based detection. The probes are attached to a solid surface by a covalent bond to a chemical matrix in standard microarrays (Figure 2.1). The solid surface is usually a chip made of glass or silicon, commonly referred to as gene chip.

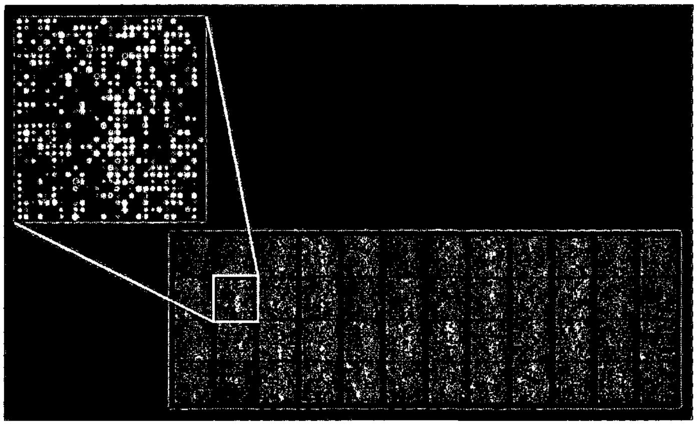


Figure 2.1: A microarray chip with approximately 40,000 probes. The upper left corner shows one enlarged part of the chip. (from WIKI)

2.2 Gene dosage effect

Gene dosage effect refers to the relationship between genotype and phenotype. Deviation of gene dosage from the normal level can produce new phenotypes. Reduced gene dosage produces haploinsufficiency effect, and gene over-expression may also lead to diseases.

2.2.1 Haploinsufficiency

Haploinsufficiency is one phenomenon arising from the total, or partial, lack of activity of one copy of gene at a diploid locus. The Single functional gene copy only produces half of the normal amount of the gene product, leads to an abnormal phenotype.

Several cases of haploinsufficiency have been documented in man. Collagens IIA1 and VA1 participates in the formation of connective tissue. The assembly of collagen fibril involves an initiation of micro-fibrils through a helical cooperative mechanism. It has been shown that type I collagen self-associates efficiently only if the concentration of the protein exceeds a threshold value [10]. Haploinsufficiency effect has also been seen in human elastin, another polymeric component of the connective tissue [11]. Its polymerization also involves a highly cooperative monomer->oligomer self-association process [12].

In additional to human, other species also exhibit haploinsufficiency effects. It is known that haploinsufficiency of protamins 1 or 2 results in fertility in mice. Protamine is a major DNA-binding protein in the nucleus of sperm in most vertebrates. They help to constrain the DNA into a small space less than 5% of a somatic cell nucleus. The protamine-DNA interaction is a highly cooperative process [13]. A decrease in the amount of protamine leads to disruption of sperm nuclear formation and abnormal sperm function [14].

The above cases involve genes encoding proteins that are synthesized and required in large amount. There are also many examples of transcription factors which regulate the expression of target genes and normally work close to a threshold. Haploinsufficiency of the Wilms' tumor gene-1 (WT1) contributes to male-to-female reversal [15], while haploinsufficiency of Steroidogenic factor 1 (SF1) is associated with adrenal failure [16].

2.2.2 Increased gene dosage effect

In additional to reduced gene dosage, increased gene expression also causes phenotypic consequences. PLP and PMP22 are two genes associated with myelin formation. The former encodes the proteolipid protein PLP of the central nervous system, and the later produces the peripheral myelin protein 22 (PMP22). Duplication of these two genes results in Pelizaeus-Merzbacher disease/spastic paraplegia type II and the type 1A Charcot-Marie Tooth syndrome respectively [17].

There are more examples of increased gene dosage leading to a gain-of-function that produces new phenotypes. SRY-related HMG box 9 (SOX9) is an essential transcription factor in chondrogenesis [18]. HI of SOX9 causes anomalies and gonadal dysgenesis in a 46, XY background [19]. However, duplication of a genomic region containing SOX9 is responsible for female to male sex reversal [20]. Another example is manifested by the constitutive over-expression of ID1, an inhibitor of the DNA binding capacity of bHLH proteins. The resulted phenotype resembles that of the null mutation for E2A, a factor involved in B cell development [21].

2.3 Dosage balance hypothesis

Section 2.2 discusses dosage effects caused by reduced or increased gene dosage. In those cases, we consider the changes of absolute expression levels. However, the dosage balance has been proposed to be a key issue. Let us take a look at an example in the budding yeast. Mlc1p is a light chain for the myosin Myo2p. Mlc1p contributes to the structural stability of Myo2p, and displays haploinsufficiency. However, reduced amounts of Myo2p can suppress the haploinsufficiency exhibited by Mlc1p. It is the relative excess of Myo2p that is more likely to be responsible for the "toxic effect". It is clear that the stoichiometric balance also plays an important role in determining phenotypic effects.

2.3.1 Gene dosage balance in macromolecular complexes

Veitia proposed that the subunits of a complex should be balanced to avoid dominant fitness defects [5]. Accordingly, both under- and overexpression of individual components within a complex tend to lower fitness. Consider a complex consist of proteins A and B. An excess of A induces dosage imbalance and hence may be deleterious ([22]): A could form homodimers which may disrupt pathways, it could interfere with interaction between B and other proteins, it leads to irreversible AB binding with an abnormal function, or it could produce toxic precipitates. Another more complicated example is a trimer A-B-C, where B is a bridge. An increase in the amount of B may lead to the irreversible formation of subcomplexes AB and BC. Stoichiometric imbalances in macromolecular complexes, therefore, can be a source of dominant phenotype.

The dosage balance theory proposes several predictions ([22]): "adaptations should tend to minimize the degree of imbalance, heterozygous deletions or over-expression of one subunit should be deleterious, the strength of transcriptional coregulation of subunits is expected to reflect dosage sensitivity, and an imbalance caused by halving (or increasing) gene dosage in one gene will be rescued, at least in part, by reducing (or increasing) expression of the interacting partner."

Rapid degradation of unassembled ribosomal subunits may be one case of the first prediction, while the fourth prediction is evidenced by the example of Mlc1p and Myo2p mentioned above.

The other twos have been supported by a phenotype and gene expression analysis of protein complex in yeast (Saccharomyces cerevisiae) [23]. Protein complex annotation was extracted from MIPS comprehensive Yeast Genome Database. Fitness effect was measured by the growth rates of heterozygous and homozygous diploid strains for single-gene deletions in the yeast genome. Only essential genes were considered to minimize measurement biases. It was found that genes with low heterozygous fitness are more enriched for components in complexes. Dosage sensitive genes are at least twice more likely to encode proteins involved in complexes than dosage insensitive genes (Figure 2.2) [23]. On the other hand, components of protein complexes counts 47% of the genes whose overexpression in wild-type cell is lethal [23]. This is a highly significant excess as compared to genes whose overexpression has no detrimental effect on fitness. These two facts strongly support the second prediction stating that under- and overexpression of subunits of a complex can be deleterious.

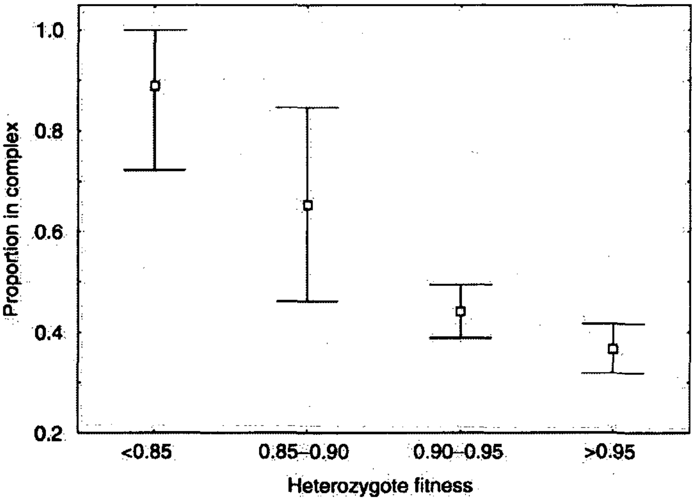


Figure 2.2: Correlation between dosage sensitivity and proportion of genes in protein complex. Dosage sensitive genes (low heterozygote fitness) have a higher percentage of genes encoding components in protein complex than dosage insensitive genes [23].

Papp *et al.* test the third prediction by examining co-expression of interacting gene pairs [23]. They found that the interacting pairs with high fitness deficiency are much more co-expressed than the others (Figure 2.3 [23]). Only 20% of the interacting pairs with low fitness deficiency (less than 5%) are co-expressed, while more than 80% of the subunit pairs with high fitness deficiency (more than 15%) show co-expression evidence [23]. Dosage sensitivity is shown to affect the strength of transcriptional co-regulation of subunits.

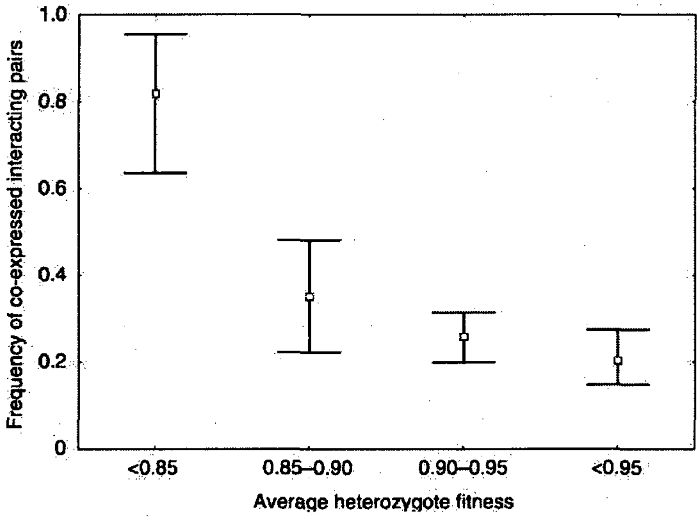


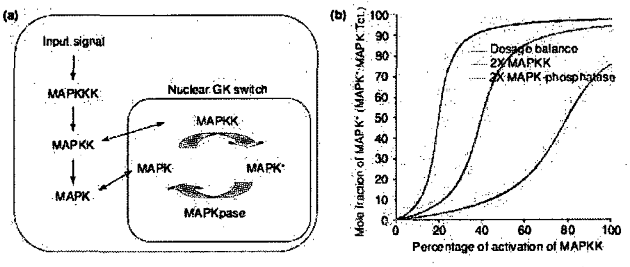
Figure 2.3: Correlation between dosage sensitivity and frequency of co-expressed interacting pairs. Dosage sensitive genes (low heterozygote fitness) are more co-expressed than dosage insensitive genes [23].

Another important work that Papp *et al.* has done is to study the co-evolution of protein subunits. According to the balance theory, single gene duplications of subcomponents induce dosage imbalance and hence can be harmful. Therefore, interacting pairs should either remain sole copies or undergo gene duplication in the same time. Consistently Papp *et al.* found a large excess of solo copy pairs and interacting pairs with the same number of paralogues [23]. They also noticed that genes involved in complexes rarely have many paralogues.

There is more evidence showing the impact of dosage balance on gene duplication. Yang and colleagues demonstrated that the proportion of duplicated genes decreases with the size of protein complexes [24]. Duplication of subunits in large protein complex is more likely to cause dosage imbalance, since there is less chance of synchronous duplication of all components. Another evidence is the clustering of genes encoding subunits of stable complexes on the yeast chromosomes [25].

2.3.2 Gene dosage balance in informational pathways

Dosage imbalance problems are not restricted to protein complexes, but also exist in signal transduction and genetic pathways [26]. One example is the mitogen-activated protein kinase (MAPK) signaling module. This signal pathway includes a phosphorylation cascade involving three main levels: MAPKKK, MAPKK and MAPK, along with their corresponding phosphatases (Figure 2.4). It is a Goldbeter-Koshland (GK) switch system [27], which responds to extracellular stimuli in a switch-like manner. When MAPKK and nuclear MAPK-phosphatase are saturated, the response of the system, represented by the quantity of MAPK's phosphorylated form (MAPK*), is dependent on the ratio of active MAPKK to MAPK-phosphatase. A sharp transition occurs in the signal-response curve, as the ratio exceed a critical value. A change in the ratio of MAPKK and MAPK-phosphatase leads to a shift in the response curve, and probably a fitness defect. On the contrary, the parallel change of both proteins, which keeps the ratio constant, does not alter the position of the threshold or the shape of the curve, as long as they are saturated. Another example is established in genetic circuits in which a dosage balance between repressors is required for bi-stability [26].



TRENDS in Genetics

Figure 2.4: Illustration of gene dosage balance in signaling pathways. (a) Schematic representation of a mitogen-activated protein kinase (MAPK) pathway. The smaller rectangle represents a Goldbeter-Koshland (GK) switch [27]. (b) The response of the GK switch system as a function of the ratio of active MAPKK to MAPK-phosphatase. The blue line represents the normal response, while the pink and red lines represent the response for double the gene dosage of either MAPKK or MAPK-phosphatase respectively [26].

2.4 Summary

Gene expression is a process of transferring information from DNA to proteins. Sometimes, genes are expressed in reduced or increased quantity. Both under- and overexpression of genes could produce abnormal phenotypes. Dosage balance has been proposed to be a key factor in determining gene expression phenotype. Stoichiometric imbalances in macromolecular complexes and informational pathways are a source of dominant phenotypes. Protein subunits from the same complex should be co-expressed to avoid dosage imbalance effects. The strength of transcriptional co-regulation of interacting pairs is expected to reflect dosage sensitivity. Several experimental results have provided strong support to gene dosage theory.

Chapter 3

Protein structure vulnerability

This chapter describes a structure marker of protein interactivity referred to Solvent Exposed Backbone Hydrogen bond (SEBH) or dehydron, and quantifies the extent of protein structure vulnerability based on this new structure feature.

3.1 Solvent exposed backbone hydrogen bond

Hydrogen bonding is one major component that determines the protein structure stability. While most backbone hydrogen bonds are well protected, a few of them are exposed to solvent desolvation. These solvent exposed hydrogen bonds are under-wrapped, and vulnerable to water attack.

3.1.1 An introduction to protein structure

Proteins are building blocks of a living cell. They catalyze biochemical reactions. They are structural components of muscle, membranes and membrane channels. They also play an important role in immune responses, cell signaling and the cell cycle. The proper function of proteins demands an adoption of fairly rigid spatial structures. A minor shift in three-dimensional (3D) structures can lead a loss of or dramatic changes in protein activities.

Proteins are polymers of amino acids. An amino acid is a molecule that contains both amine and carboxyl groups. The general formula for the amino acid is $H_2NCHRCOOH$, where R stands for organic substitute. There are 20 common amino acids: alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalaine, proline, serine, threonine, tryptophan, tyrosine, and valine. The carboxyl group (-COOH) of one amino acid reacts with the amino group ($-NH_2$) of another amino acid, which produces a molecule of water (H_2O) and a peptide bond (CO-NH) (Figure 3.1).

Amino acids can be classified as being hydrophilic or hydrophobic, according to the polarity of the side chain (Table 3.1). The physical properties of the side chains are important in protein structure and protein-protein interactions. Nonpolar groups tend to cluster together to minimize the solvent exposure area and the entropy cost of forming hydrogen-bond network. The clustering of nonpolar groups is referred as hydrophobic effects. On the contrary, polar groups interact with each other through hydrogen bonding or other interactions.

Polypeptides and proteins are chains of amino acids connected by peptide bonds. The

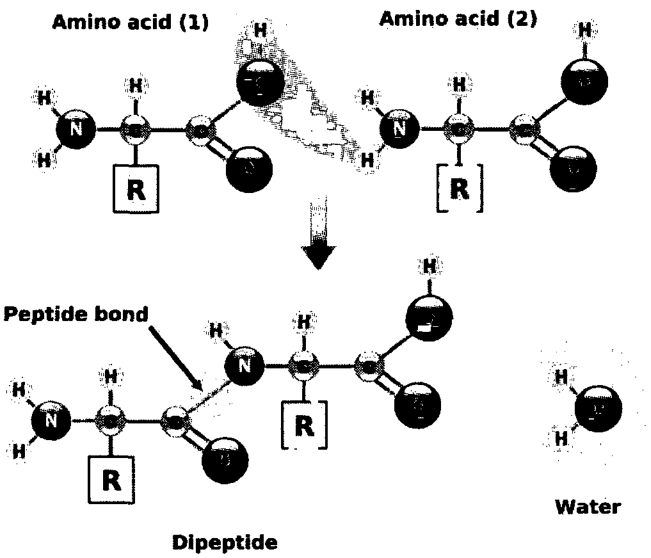


Figure 3.1: Illustration of peptide bond formation (from WIKI).

sequences of amino acids, which are encoded on DNA, determine the 3D structure of proteins. A protein sequence is called its primary structure. The primary structure is then folded to 3D structure, which is often composed of regular secondary structures, namely, α-helix and β-sheet (Figure 3.2 (a), α-helix (red), β-sheet (yellow)). The 3D structure of a single protein molecule is referred to tertiary structure. Complexation of several protein molecules produce a large protein complex regarded as the quaternary structure of a protein (Figure 3.2 (b)). The tertiary structure of the protein is determined by the distribution of hydrophilic and hydrophobic amino acids, and the quaternary structure is influenced by amino acids on protein surface. For example, soluble proteins' surfaces are often rich in polar amino acids like serine and threonine, while integral membrane proteins tend to place hydrophobic amino acids on their surface that helps them enter the lipid bilayer. Similarly, proteins binding to positively-charged molecules have many negatively charged

Table 3.1: 20 amino acids and their side chain properties (from WIKI).

Amino Acid	Side chain polarity	Side chain acidity or basicity of neutral species
Glycine	nonpolar	neutral
Alanine	nonpolar	neutral
Valine	nonpolar	neutral
Leucine	nonpolar	neutral
IsoLeucine	nonpolar	neutral
Methionine	nonpolar	neutral
Proline	nonpolar	neutral
Phenylalanine	nonpolar	neutral
Tyrosine	nonpolar	neutral
Tryptophan	nonpolar	neutral
Serine	polar	neutral
Threonine	polar	neutral
Cysteine	polar	neutral
Asparagine	polar	neutral
Glutamine	polar	neutral
Aspartate	polar	acidic
Glutamate	polar	acidic
Lysine	polar	basic
Arginine	polar	basic(strongly)
Histidine	polar	basic(weakly)

amino acids on their surface, while proteins interacting with negatively-charged molecules have surfaces rich with positively charged chains.

The folding of ploymers into 3D structures is driven by a number of noncovalent interactions including hydrogen bonding, ionic interactions, Van der Waals forces, and hydrophobic packing. A hydrogen bond forms between an electronegative atom and a hydrogen atom bonded to nitrogen, oxygen or fluorine. Ionic interaction is the electrostatic interaction between metal and non-metal ions. Van der Waals force is the force between molecules which includes dipole-dipole force, instantaneous dipole-induced dipole force.

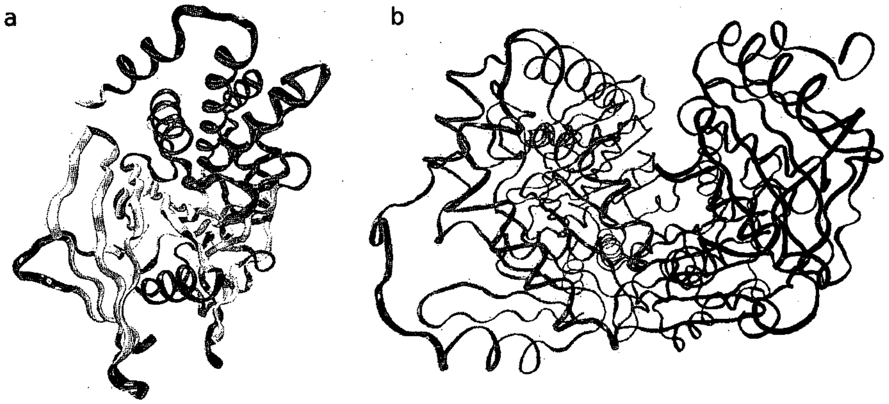


Figure 3.2: (a) The 3D structure of a single protein (PDB.2VEU). α-helix is colored as red, and β-sheet is colored as yellow. (b)The 3D structure of a protein complex (PDB.1KB9). The cyan chain represents cytochrome b-c1 complex subunit 1, and the blue chain represents Cytochrome c oxidase subunit 2.

Hydrophobic interaction is the tendency of nonpolar components to form aggregates to minimize their contacts with water. In addition to protein folding, these interactions also contribute to the stability of protein structure. Among these interactions, hydrogen bonding is a major component in stabilizing protein structure. A hydrogen bond is typically 5 to 30 kJ/mol [28], comparable to that of weak covalent bonds (155 kJ/mol, [29]). Hydrogen bonds are often found in α-helices and β-sheets as shown in Figure 3.3. They were attributed to a critical role in formation of α-helix and β-sheet, when Pauling proposed the secondary structure model of protein [30, 31].

3.1.2 The wrapping of hydrogen bonds

Protein folding involves burial of hydrophobic residues, which provide protection for hydrogen bonds from water attack. In proteins, most of the backbone hydrogen bonds

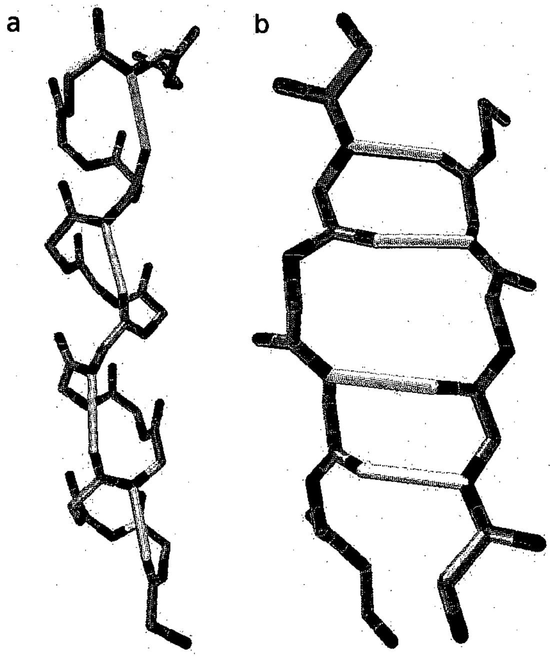


Figure 3.3: Backbone hydrogen bonds (shown as yellow bonds) in α-helix (a) and β-sheet (b) (PDB.1JAT). The cyan, blue and red balls represent covalent bonds contributed by carbon, nitrogen and oxygen atoms respectively. Graph was prepared using VMD.

are well protected by the hydrophobic groups of the side chains. These kinds of hydrogen bonds have a high bonding energy and hence are essential for stabilizing the protein conformation. Both statistical and theoretical approaches have been employed to quantify the extent to which the hydrogen bonds are protected by proteins [6, 7, 32]. The results showed that about 92% of backbone hydrogen bonds were well protected and about 8% of hydrogen bonds are vulnerable to water attack [32]. Those under-wrapped hydrogen bonds are proposed to be central to protein-protein interaction [6, 7, 32]. They are termed as Solvent Exposed Backbone Hydrogen bond (SEBH) or dehydron. When a hydrophobic group approaches to a dehydron, the water molecules around the dehydron are excluded and the dehydron turn into a well wrapped hydrogen bond which has a high bonding energy. The net energy gain in this process has been experimentally determined to be close to 4kJ/mol [33], which is significant and comparable to the strength of hydrophobic interaction.

The majority of backbone hydrogen bonds are well protected by the surrounding non-polar groups. The level of wrapping could be quantified by counting the number of hydrophobic groups in the dehydration domain of the hydrogen bond. The dehydration domain is defined as two spheres of radius R centered at the alpha-carbons of the residues paired by the hydrogen bond (see Figure 3.4). This value of R is related to the characteristic length Λ of the solvent-structuring effect due to the presence of a vicinal hydrophobe [34]. By fixing Λ at 1.8 Å (the effective thickness of a single-layer water cavity) and assuming structuring influence decays exponentially, R is set to be 6.0Å (∼ 3Λ) to reduce the structuring influence to 1% of its maximum value. The level of wrapping of the hydrogen bond ρ is measured by the number of hydrophobic groups in the domain.

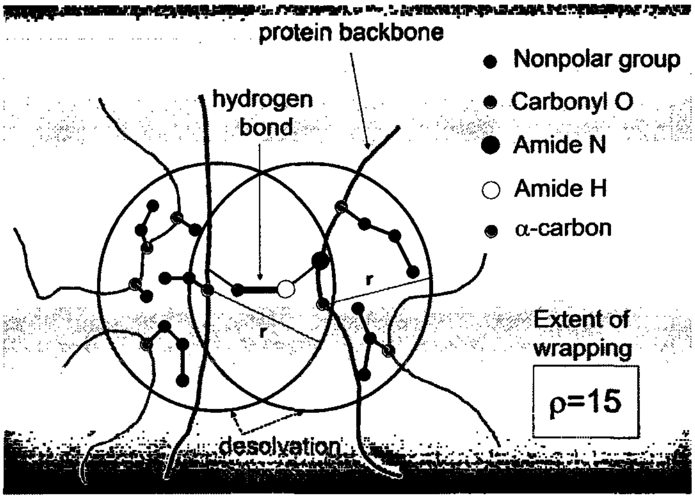


Figure 3.4: Illustration of the dehydration domain of a hydrogen bond. As shown in this figure, the wrapping level of the hydrogen bond ρ = 15.

In structures of PDB-reported soluble proteins, at least two thirds of the backbone hydrogen bonds are protected on average by $\rho = 26.6 \pm 7.5$ side-chain nonpolar groups for a desolvation ball radius 6 Å. SEBHs are those hydrogen bonds whose ρ lies in the tails of the distribution, i.e. their microenvironment contains 19 or fewer nonpolar groups, so their ρ-value is below the mean (ρ = 26.6) minus one standard deviation (σ = 7.5).

The nearly constant ρ value reflects the generic composition of the protein chains. On the other hand, the dispersion σ is largely due to the variation of the size of the side chains. The large hydrophobic residues provide better protection for the hydrogen bond than small residues, and the number of large hydrophobic residues required to protect the hydrogen bond is less than that of small residues. No backbone hydrogen bond has less than two hydrophobic residues or more than eight in its desolvation domain [34].

3.1.3 Wrapping and disorder score

SEBHs or dehydrons are hydrogen bonds whose wrapping levels are significantly lower than average hydrogen bonds. They are vulnerable to water attack. As a result, regions containing dehydrons are supposed to be disordered to some extent. The relationship between wrapping and disorder is studied using the highly accurate disorder prediction program PONDR [35]. PONDR predicts structural disorder from sequence by quantifying sequence attributes over windows of 9 to 21 amino acids [36, 37, 38]. Those attributes including hydropathy and sequence complexity are averaged over windows and the values are used to guide the prediction [37]. The predictor assigns a disorder score λ_D, ranging from 0 to 1, to each residues along the sequence. The disorder score λ_D represents the propensity of the residue to be in a disordered region: λ_D = 0, the residue absolutely belongs to an ordered region; $\lambda_D = 1$, the residue is absolutely resided in a disordered region. There is 6% of false positive predictions of disorder in sequence windows of >= 40 amino acids for more than 1, 100 nonhomologous PDB proteins. However, this value is overestimated due to the fact that many disordered regions in monomeric chains become ordered upon ligand binding or in crystal contacts.

Disorder analysis has been performed over 2,806 nonredundant nonhomologous PDB domains, and the disorder score for each residue has been obtained. Residues were grouped to 45 bins ($8 <= \rho <= 52$), according to the level of wrapping of hydrogen bonds in which these residue are engaged. Disorder scores were averaged over each group of residues. Figure 3.5 shows the correlation between disorder score at a particular residue site and the extent of wrapping of the hydrogen bond engaging that residue (if any). The strong correlation implies that regions rich in dehydrons tend to adopt a natively disordered state. The backbone hydrogen bonds need enough protection from other parts of protein and contribute to the stability of protein structures.

The strong correlation between disorder score and the wrapping level of hydrogen bonds also enable us to predict the existence of dehydrons on the sequence basis. For regions with a disorder score $\lambda_D>0.35$, the accuracy for dehydron prediction is 94%. Dehydrons are a structural feature, and structures can not be obtained from sequence. It might be surprising that dehydrons can be inferred from sequence. However, PONDR employs a learning strategy that incorporates sequence windows in its training set together with the structural context in which such windows occur. The inclusion of structure information in disorder predictions provides the basis of high accuracy of predicting dehydrons.

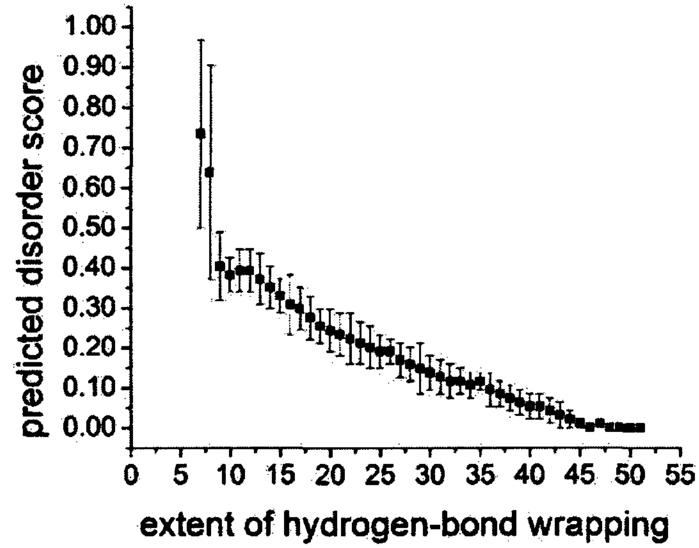


Figure 3.5: Correlation between disorder score at a residue and the wrapping of the hydrogen bond in which the residue is engaged.

3.2 Role of Dehydrons in protein interactions

Dehydrons are identified as packing defects in protein structure. They are highly sensitive to the water removal, and are important in protein association. Upon protein association, dehydrons in the binding interface become dehydrated and stable, which then contribute to stability of protein structures. The number of dehydrons therefore serves as a quantifier of protein interactions.

3.2.1 Dehydrons as a determinant of protein association

As defined in the second method, dehydrons are those bonds whose Coulomb energy could change dramatically with the presence of a new hydrophobe from a binding partner. Therefore, dehydrons could serve as good indicator for binding sites. By examining the protein-protein interface of 212 complexes from an exhaustive database, it is found that the density of dehydrons in the interface δ_{int} is higher than the overall density of dehydrons δ: 77 have $\delta_{int}/\delta > 1.5$, some even has 7 times higher density of dehydrons in the interface. 92.9% of the PDB complexes have higher density of dehydrons at the protein-protein interface than the average density for individual monomeric partners [7]. The dehydrons in the interface of monomeric proteins become well wrapped in the complexes. The high density of dehydrons in the interface indicates that the exclusion of water from the structurally defective region play an important role in protein-protein association.

3.2.2 Dehydrons as an indicator of protein interactivity

As a strong indicator of protein interactivity, dehydrons could further provide insight into the pattern of proteomic connectivity. A systematic investigation of the dehydron and interaction patterns of all monomeric PDB domains from the yeast proteome found that domain connectivity is proportional to the average number of dehydrons in the family [39]. Figure 3.6 shows a correlation of the number of dehydrons in all monomeric PDB domains from the yeast proteome and the number of their interacting partners obtained from the Database of Interacting Proteins (DIP).

The almost linear correlation between the number of dehydrons and the number of interaction also holds for structural families (or Structural Classification of Proteins (SCOP) superfamilies) (see Figure 3.7 A) [39]. The numbers in the figure are the SCOP IDs for some protein domains. These two results suggest that the domain connectivity is measured by the average number of dehydrons $< r >$ in a given family. Figure 3.7 B displays the distribution of protein families according to their $< r >$. The fraction $f = f(< r >)$ of

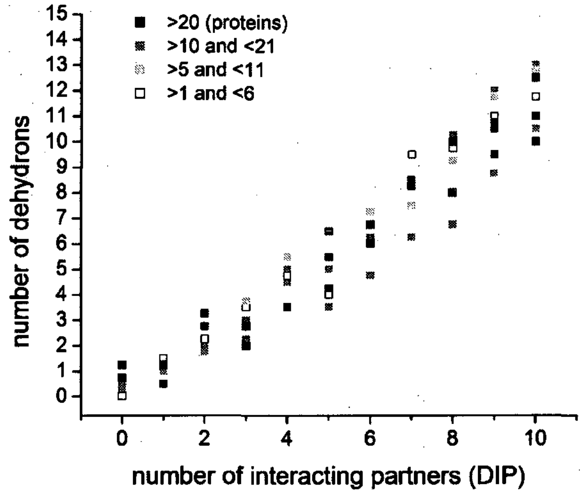


Figure 3.6: The number of dehydrons of a protein domain as a function of protein interactivity. The number of dehydrons of a given domain fold is obtained by averaging over all proteins in the domain, and protein interaction data are from Database of Interacting Proteins (DIP). The correlation is quite strong, with a correlation coefficient of 0.88 [39].

protein families follows the same power law $f(<r>) \propto <r>^{-\gamma}$ as the distribution of protein connectivity, which further confirm the relationship between $<r>$ and proteomic interactivity. The index γ for H. sapiens, M.musculus and E. coli are respectively 1.44, 1.49 and 2.1 [39].

3.3 Definition of protein structure vulnerability

Proteins with large number of dehydrons are more likely to be involved in protein association. Protection from their binding partners stabilizes dehydrons at binding interfaces by excluding water molecules in their microenvironment. The number of dehydrons therefore could measure the extent of protein structure vulnerability. The higher percentage of dehydrons a protein has, the more vulnerable this protein becomes.

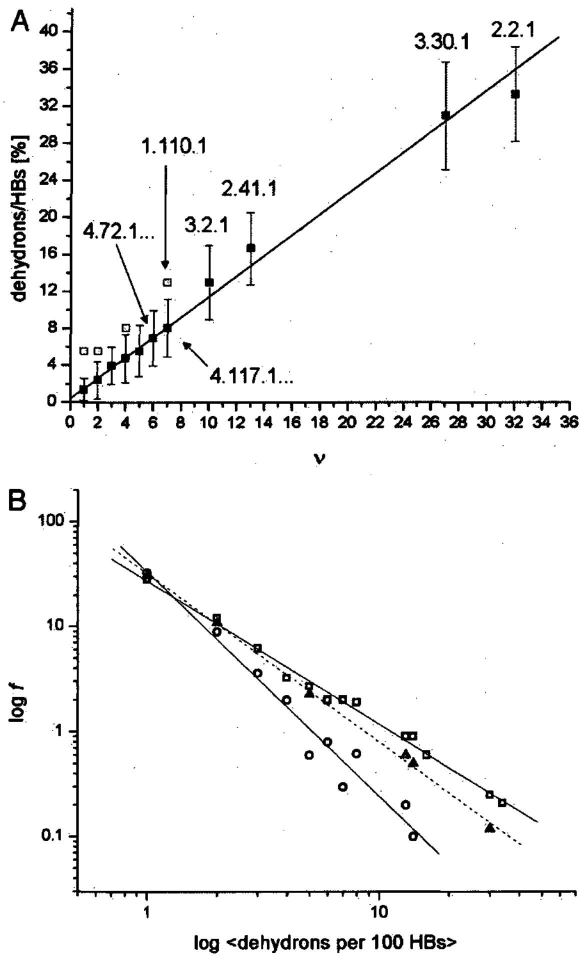


Figure 3.7: (A) Correlation between the average percentage of dehydrons and the connectivity ν of SCOP families represented in the PDB. The numbers in the graph are the SCOP IDs. (B) Distribution of SCOP families according to their average number $< r >$ of dehydrons per 100 hydrogen bonds. $f = f(< r >)$ is the fraction of families with $< r >$ dehydrons. □ - H.sapiens; ▲ - M. musculus; ○ - E. coli [39].

Protein structure vulnerability is defined as the ratio ν of dehydrons or SEBHs to the total number of backbone hydrogen bonds. According to this definition, the most vulnerable protein is the potassium channel scorpion toxin HSTX1 ($\nu = 100\%$, see Figure 3.8). It is highly active on voltage-gated Kv1.3 potassium channels. This protein belongs to the scorpion short toxin family, which essentially contains potassium channel blockers of 29 to 39 amino acids and three disulfide bridges. These proteins are characterized by their high affinities and different specificities for several types of potassium channels. Furthermore, it has the particularity to possess a fourth disulfide bridge. HsTX1 binds with a picomolar affinity to the Kv1.3 channels [40].

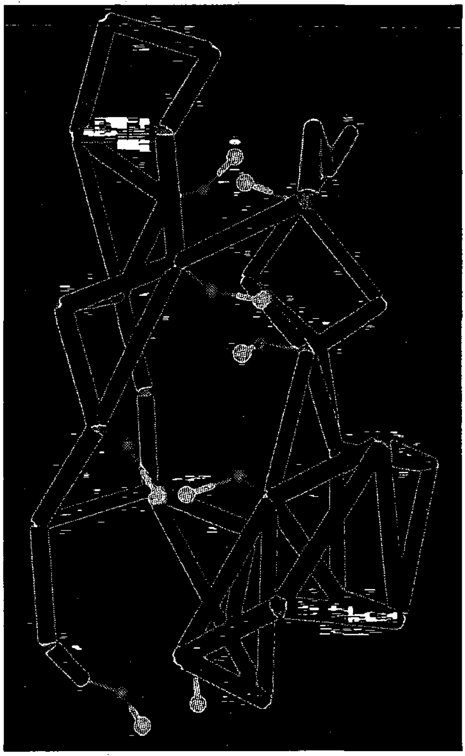


Figure 3.8: The wrapping pattern of the toxin protein (PDB.1QUZ). The backbone is represented as virtual bonds (shown as blue segments) joining consecutive α carbon atoms (grey spheres), and the green segments represent solvent exposed backbone hydrogen bonds. The side chains with yellow spheres represent cystine residues which form four disulfide bridges [41].

3.4 Summary

Hydrogen bonding makes a major contribution to protein structure stability. Protein folds in such a way that most backbone hydrogen bonds are well wrapped. Those well-wrapped hydrogen bonds are stable and essential for stabilizing protein conformation. On the contrary, some backbone hydrogen bonds are exposed to solvent desolvation, and are vulnerable to water attack. Those solvent exposed hydrogen backbone hydrogen bonds (SEBHs) or dehydrons can be stabilized upon protection by nonpolar residues from interacting proteins. SEBHs serve as a structural marker of protein interactivity. Protein structural vulnerability is quantified by the ratio of SEBHs over the total number of backbone hydrogen bonds. Proteins rich in SEBHs are vulnerable to water attack, and rely on their interacting partners to maintain structural integrity.

Chapter 4

Protein structure vulnerability as dosage sensitivity quantifier

Biology system consists of separate functional modules, which results from spatial or specificity isolation. Protein complex is a typical spatial defined module, whose components interact in the same time to promote a specific function. Dosage imbalance between subunits of a protein complex has become a key issue of genetic dominance. Both under- and overexpression of subcomponents can be deleterious. The strength of transcriptional co-regulation of interacting pairs is assumed to reflect the dosage sensitivity. On the other hand, we found that structure vulnerability quantifies the extent to which proteins rely on their binding partners to maintain their structure integrity. Highly under-wrapped protein is vulnerable to water attack, and is highly needy for protection from other proteins. Therefore, vulnerability should provide a molecular basis for co-expression of binding partners, and hence for dosage sensitivity. This chapter examines this prediction, by integrating transcriptomics, metabolomics, and proteomics data sets of human and yeast (Saccharomyces cerevisiae).

4.1 Materials and Methods

4.1.1 Expression data sources

Gene expression levels are assessed by DNA microarray analysis. DNA microarray is a high-throughput technology which measures the mRNA levels of thousands of genes simultaneously. Human expression profiles were obtained from Novartis gene expression atlas [42]. This expression dataset contains an extensive collection of human characterized and uncharacterized genes. For each gene, there are expression data from 79 tissue samples. We discarded six cancer tissues: ColorectalAdenocarcinoma, leukemialymphoblastic(molt4), lymphomaburkittsRaji, leukemiapromyelocytic, lymphomaburkittsDaudi, leukemiachronicmyelogenous (k562). Yeast expression data was obtained from the Saccharomyces Genome Database [43]. This data set contains mRNA expression levels during a transition from glucose fermentative to glycerol-based respiratory growth.

4.1.2 Protein interaction data sets

We compiled structure curated protein interaction datasets following steps of Gerstein et al. [44]. All proteins in the interaction data set are mapped to Pfam domains [45]. The Pfam domain interactions are annotated in iPfam [46], which employ structure information to define domain interactions. Two proteins were then considered to interact with each other, when their respective domains or homologs of their respective domains were found in a complex with PDB-reported structure. We obtained curated yeast protein domain interactions from the Structural Interaction Network (SIN) [44], and filtered them using recently published yeast protein complex data [47]. The human interaction data set was extracted from the protein complex list in MIPS/Mammalian Protein Complex Database (MPCDB) [48]. This database does not include data from high-throughput experiments, but only manually annotated mammalian protein complexes extracted from individual experiments described in the scientific literature. All interactions among components in complexes were then curated using Pfam and iPfam.

4.1.3 Calculation of expression correlation η

We use the Pearson correlation coefficients of expression vectors to determine similarity between expression profiles. For two expression vectors X and Y, the Pearson correlation coefficient Corr(X, Y) is given by

$$Corr(X,Y) = \frac{<(x-<x>)(y-<y>)>}{\sqrt{<x^2> - <x>^2}\sqrt{<y^2> - <y>^2}} \tag{4.1}$$

where x, y are generic coordinates in the vectors X and Y respectively, and < > indicates mean over the 73 normal tissues (human) [42] or over the 5 metabolic adaptation phases (yeast) [43]. The expression correlation for a protein-protein interaction is then normalized by the mean correlation over all gene pairs encoding for interacting domains. The normalization is necessary for comparative analysis across species because different species have different mean expression correlations and hence the significance of a correlation is necessarily a relative attribute. Given its statistical nature, the denominator is nonzero for any species since in a statistical sense, protein pairs that interact are expected to be positively correlated in their expression.

4.1.4 Calculation of vulnerability ν and identification of SEBHs for soluble proteins

The structural vulnerability ν of a protein is measured as the ratio of number of SEHBs to the total number of backbone hydrogen bonds. Identification of SEBHs was performed as described in Chapter 3. In this work, we adopted 6Å as the length of the desolvation ball radius. Analysis over all structures of PDB-reported soluble proteins showed that at least two thirds of the backbone hydrogen bonds are protected on average by $\rho = 26.6 \pm 7.5$ side-chain nonpolar groups. Thus, SEBHs lie in the tails of the distribution, i.e. their microenvironment contains 19 or fewer nonpolar groups, so their ρ-value is below the mean (ρ = 26.6) minus one standard deviation (σ = 7.5).

In cases where the protein structures were unavailable from the PDB, we generated atomic coordinates through homology threading using the program Modeller [49, 50, 51]. Modeller is a computer program that models 3D structures of proteins subject to spatial constraints [51], and was adopted for homology and comparative protein structure modeling. The homology threading was performed by adopting known homolog structures as templates. Yeast PDB homologs were obtained from the Saccharomyces Genome Database, and human PDB homologs were from Pfam. We generated the alignment of the target sequence to be modelled with the Pfam-homolog structure reported in PDB and the program computes a model containing all non-hydrogen atoms. The input for the computation consists of the set of constraints applied to the spatial structure of the amino acid sequence to be modeled and the output is the 3D structure that best satisfies these constraints. The 3D model is obtained by optimization of a molecular probability density function with a variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing.

4.2 Results

4.2.1 The protein complex containing the most correlated interacting pair

We quantitatively examined the relation between structural vulnerability of a protein and the extent of co-expression of genes encoding for its binding partners. First, we took a look at the mitochondrial respiratory chain complex III (Figure 4.1) who contains the most highly correlated interacting subunits ($\eta = 3.61$) among all interactions we examined in this work. This protein complex, which is located in the mitochondrial inner membrane, consists of four redox centers: cytochrome b/b6, cytochrome c1 and a 2Fe-2S cluster. The two highly co-expressed partners are subunits 1 and 2 from cytochrome b-c1 complex (subunit 1: Gene/ORF=COR1/YBL045C, shown in red; subunit 2: Gene/ORF=QCR2/YPR191W, blue). As we can see from Figure 4.1 (b), subunit 1 (red) is rich in SEBHs ($\nu = 57\%$) and hence structurally vulnerable. The high structure vulnerability of subunit 1 (red, cf. Figure b) renders it highly needy to interact with other subunits of the complex to maintain its structural integrity. The co-expression of subunits 1 and 2 is to ensure the protection of vulnerable structure. Figure 4.1 (c) shows the mutual protections of preformed SEBHs in the two subunits along part of their association interface (red: COR1 residues 42-119; blue: QCR2 residues 250-331). This intermolecular mutual "wrapping" of local weaknesses illustrates the fact that the association contributes to maintain structural integrity (Figure 4.1 (c)).

4.2.2 Correlation between protein structure vulnerability and gene expression in yeast

It is the more vulnerable protein that relies more on its partner to maintain structural integrity. Therefore, its structural vulnerability should be the driving force for co-expression of two genes. As it turns out, a tight correlation ($R^2 = 0.891$) between the maximum ν-value and the expression correlation η is obtained and shown to hold for all interacting pairs within the illustrative yeast complexes (Figure 4.2 a). This correlation is then found to hold across all 1,354 pairs of interacting proteins in the yeast interactome with Pfam representation (Figure 4.2 b, c).

4.2.3 Correlation between protein structure vulnerability and gene expression in human

Structure vulnerability is not only an organizing factor for the metabolic-adaptation transcriptome but also steers the organization of tissue-based transcriptomes. This is revealed by a similar comparative analysis of comprehensive gene expression and structure-filtered interaction data for human [42, 48]. Thus, a clear $(\eta - \nu)$-correlation is apparent between the co-expression of 607 gene pairs and the maximum structure vulnerability for each pair of interacting domains encoded in the ORFs of the respective genes (Figure 4.3).

The strong $(\eta - \nu)$-correlation implies that the protection of a functionally competent protein structure drives co-expression of its binding partners to an extent that is determined by the structure vulnerability. According to the gene dosage balance hypothesis, the extent of co-expression of interacting pairs reflects gene dosage sensitivity [26]. Therefore, the establishment of protein structure vulnerability as an organizing factor in yeast and human transcriptomes provides strong support for our hypothesis that protein structure vulnerability is the molecular basis of gene dosage sensitivity.

4.2.4 Protein intrinsic disorder and transcriptome organization

When an isolated protein fold is unable to protect solvent exposed hydrogen bonds from water attack, the protein structure becomes vulnerable and some regions are not in ordered state. This view of structural vulnerability is supported by a strong correlation between the degree of solvent exposure of intramolecular hydrogen bonds and the local propensity for structural disorder discussed in Chapter 3: In the absence of binding partners, the inability of a protein domain to exclude water intramolecularly from pre-formed hydrogen bonds may be causative of a loss of structural integrity, and this tendency is marked by the disorder propensity of the domain [35]. These findings lead us to regard the predicted extent of disorder in a protein domain as a likely surrogate for its vulnerability and contrast it with the extent of expression correlation of its interactive partners.

As described in Chapter 3, the disorder propensity is determined by the disorder score f_d (f_d = 1, certainty of disorder; f_d = 0, certainty of order) generated by PONDR-VSL2 [36, 37, 38]. The extent of intrinsic disorder of a domain was defined as the percentage of residues predicted to be disordered relative to a predetermined f_d - threshold ($f_d = 0.5$).

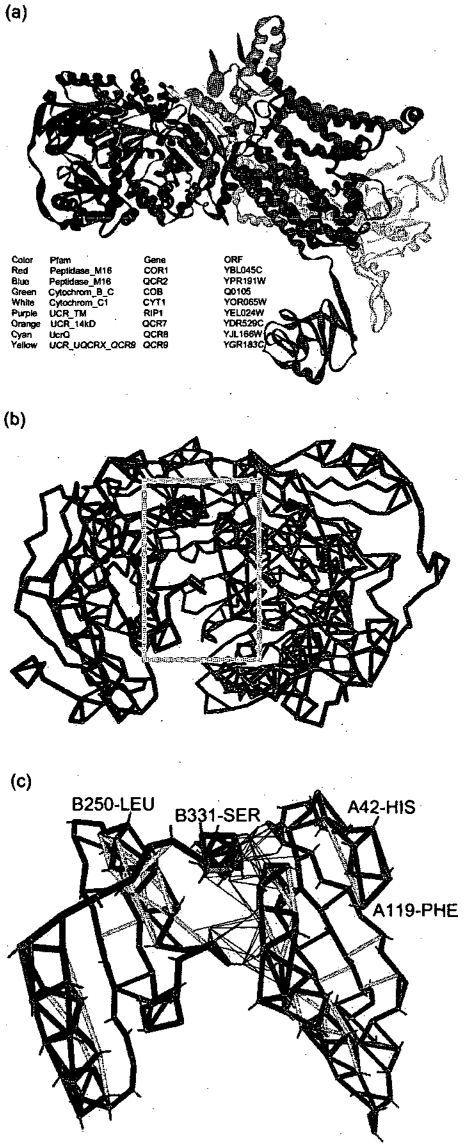
Reexamination of the expression correlations in the yeast and human transcriptomes was carried out taking into account a proteome-wide sequence-based attribution of the extent of disorder (% residues predicted to be disordered, or "disorder content") in interacting protein domains. The correlation results are shown in Figure 4.4. Although not strong, η-disorder correlations are still significant. The η-disorder correlation coefficient is high for yeast (R^2=0.752) (Figure 4.4 a), implying that disorder content determines degree of coexpression of binding partners to a significant extent. We should also notice that the large dispersion in disorder extent at high levels of coexpression (45% dispersion versus 15% for proteins with low disorder/low expression correlation). This fact indicates that highly disordered regions may adopt structures with very different levels of vulnerability depending on the complex in which they are involved. Therefore, the high dispersion to the η-disorder correlation reflects the nonlinear relationship between disorder extent and structure vulnerability.

The η-disorder correlation in human is considerably weaker (R^2=0.304, Figure 4.4 b) than in yeast. This is partly due to the fact that human proteins have a higher degree of disorder propensity than their yeast orthologs [35] and hence they are capable of significantly diversifying their structural adaptation (induced folding) in different complexes. In this context, the extent of disorder becomes a poor surrogate of structural vulnerability, as different ν-values may correspond to a single disorder prediction result.

The weaker η-disorder correlations is due to the fact that disorder score prediction is sequence-based. The disorder predictions did not include any structural information on induced fits arising upon complexation, and hence, unlike structure vulnerability, the predicted disorder score is independent of the complex under consideration. This fact introduces deviations in the estimation of vulnerability through disorder content for proteins with extensive disorder content since their conformational plasticity may enable diverse induced-fit conformations with different vulnerabilities.

4.3 Summary

Protein structure vulnerability quantifies the extent to which protein relies on its binding partner to maintain structural integrity. Interacting pairs containing vulnerable proteins are more co-expressed than the other interacting pairs. The strong correlation between co-expression and maximum structure vulnerability of interacting pairs supports our prediction that protein structure vulnerability quantifies the extent of gene dosage sensitivity.



Color	Pfam	Gene	ORF
Red	Peptidase_M16	COR1	YBL045C
Blue	Peptidase_M16	QCR2	YPR191W
Green	Cytochrom_B_C	COB	Q0105
White	Cytochrom_C1	CYT1	YOR065W
Purple	UCR_TM	RIP1	YEL024W
Orange	UCR_14kD	QCR7	YDR529C
Cyan	UcrQ	QCR8	YJL166W
Yellow	UCR_UQCRX_QCR9	QCR9	YGR183C

Figure 4.1: Mutual protections of SEBHs in the two subunits of mitochondrial respiratory chain complex III. (a) Ribbon representation of mitochondrial respiratory chain complex III (PDB. 1KB9, [52]). (b) SEBH pattern for subunit 1 (red) and subunit 2 (blue). The interacting pair is characterized by a very high expression correlation η = 3.61. The yellow square highlights the part of the interface shown in detail in (c). (c) Illustration of mutual protections of SEBHs in the two subunits along part of their interface. One side-chain bond (between α and β carbon) is displayed. The thin blue lines, which connect β-carbons in one protein with centers of hydrogen bonds in the other protein, represent mutual protections of hydrogen bonds across the protein-association interface. Thus, a thin line is shown whenever the side chain of one protein is contributing with nonpolar groups to the microenvironment of a preformed hydrogen bond in its binding partner.

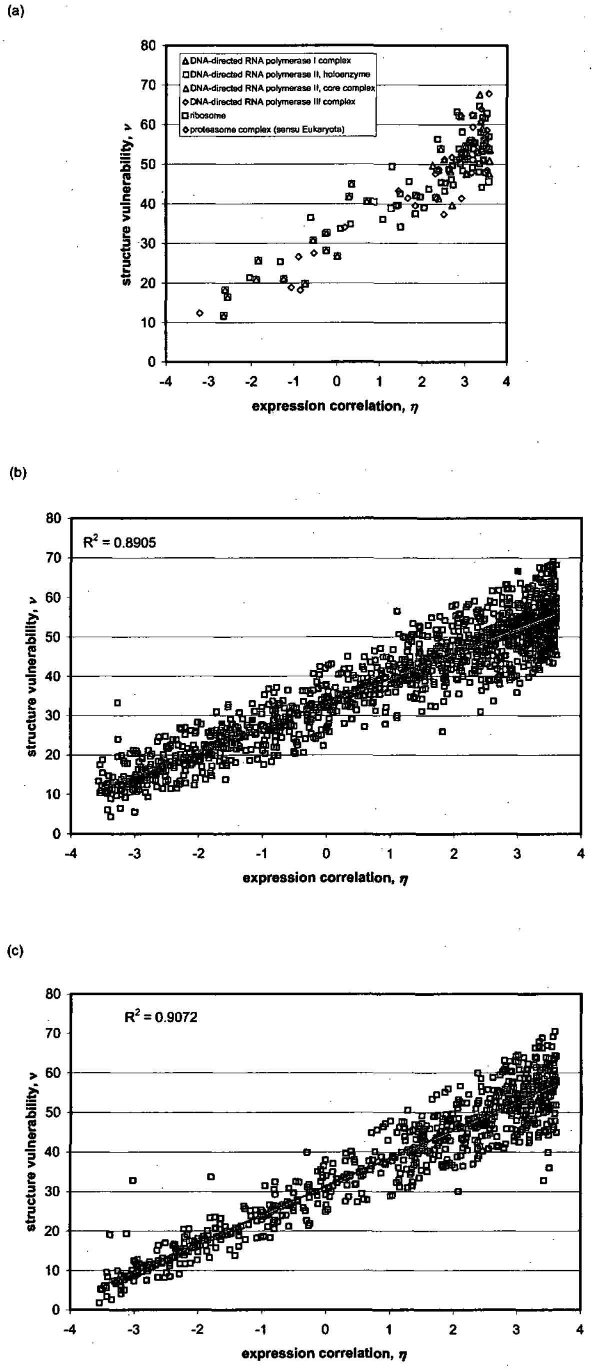


Figure 4.2: (a) Correlation between maximum structure vulnerability ν and co-expression similarity η for interactions within specific yeast complexes. (b) (η – ν)-correlation for all Pfam-filtered yeast protein interactions. Red points represent interactions involving extremely vulnerable proteins. (c) (η – ν)-correlation of Pfam-filtered yeast protein interactions involving only PDB-reported proteins. The red data point represents an interaction involving an extremely vulnerable protein, and the green point represents an interaction involving a prion protein (ERF2, [53]).

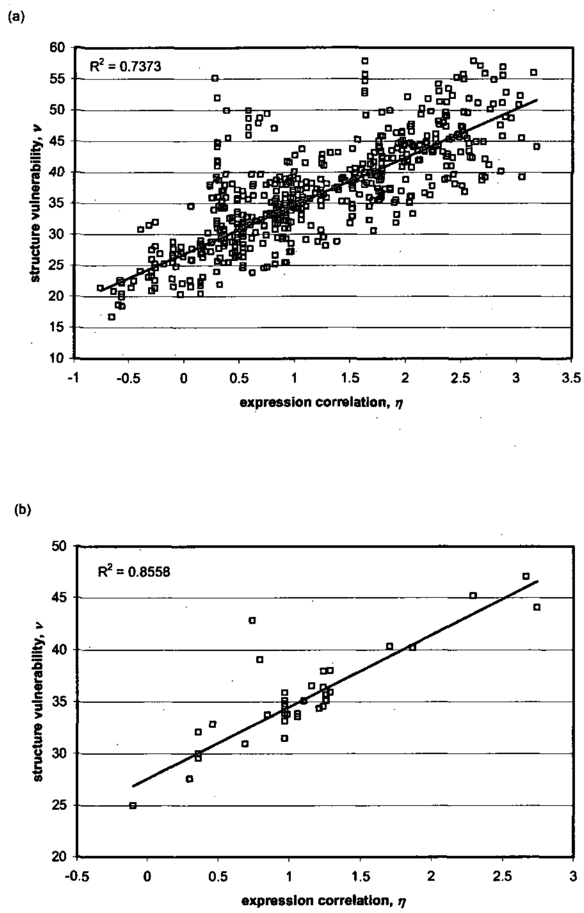


Figure 4.3: $(\eta - \nu)$-correlation for human protein interactions. (a) The $(\eta - \nu)$-correlation for all Pfam-filtered human protein interactions. Red points represent interactions involving extremely vulnerable proteins that will be discussed in Chapter 5. (b) The correlation over Pfam-filtered human protein interactions that involve only PDB-reported proteins. The red point represents interaction containing an extremely vulnerable protein.

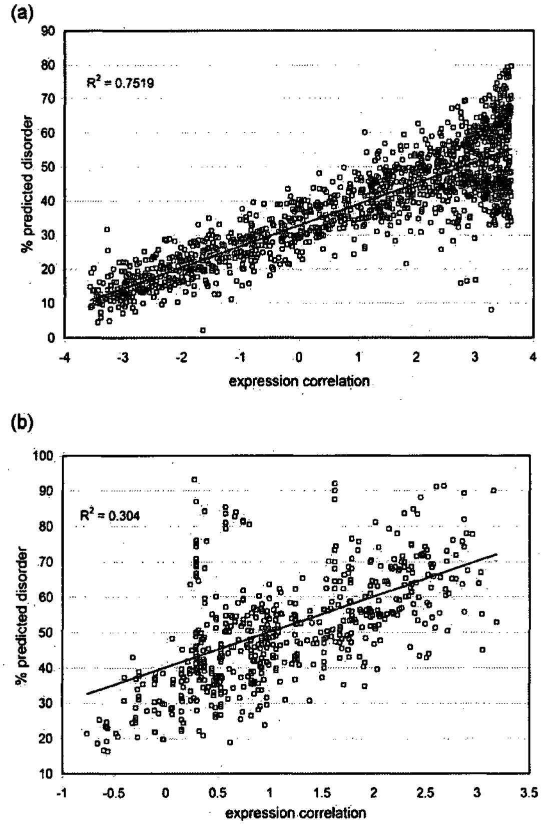


Figure 4.4: η-disorder correlation for yeast (a) and human (b) protein interactions. The disorder content is quantified by the percentage of predicted disordered residues.

Chapter 5

Post-transcriptional regulation of expression of genes encoding extremely vulnerable proteins

Messenger RNAs (mRNAs) are subjected to post-transcriptional regulation, after they are made from DNA. Gene expression may be repressed or silenced during post-transcriptional regulation. microRNAs (miRNAs) are a important class of post-transcriptional regulators. They regulate mRNA expression through two mechanisms: mRNA cleavage or translational repression. Genes encoding extremely vulnerable proteins are sensitive to dosage imbalance effects. Their expression should be tightly controlled by post-transcriptional regulators. This chapter discusses post-transcriptional regulation of genes encoding extremely vulnerable proteins in human and yeast.

5.1 Extremely vulnerable proteins

The ($\eta - \nu$)-correlations for human interactions are weaker than correlations for yeast interacting pairs. There are a few but significant outlier pairs (Figure 4.3, red data points) beyond the confidence band defined by a width of two Gaussian dispersions from the linear ($\eta - \nu$)-fit. An examination of proteins sequences revealed that those outliers involve proteins containing regions rich with poor protectors of backbone hydrogen bonds including gluatamine (Q) and asparagine (N). Regions rich with poor protectors could not adopt 3D structures, because they do not provide enough protection for backbone hydrogen bonds. Proteins with those regions are extremely vulnerable. They are prone to aggregate and form fibrils. Protein aggregation could lead to a loss-of-function and cause diseases. This is manifested by the Q-rich Huntington protein whose aggregation results in neurodegenerative disorder referred to Huntington disease.

A census of regions rich with poor protectors was performed, and 115 proteins were found to contain those regions (Table 5.1). In addition to Q and N, we also took into account other residues: G, A, S, Y, and P. These poor protectors possess side chains with insufficient nonpolar groups, with polar groups too close to the backbone (thus precluding hydrogen-bond protection through clustering of nonpolar groups) [39] or with amphiphilic aggregation-nucleating character (Y) [53, 54, 55]. Charged backbone de-protecting side chains (D, E) are excluded since they would entail negative design relative to protein self-aggregation. The (poor protector)-rich region spans 30 amino acids. This value was chosen to be consistent with the threshold used in a census of Q/N-rich regions [56]. In principle, a sizable window of residues unable to protect backbone hydrogen bonds produces a poor folder, yielding a highly vulnerable structure [39, 57]. Thus, these sequences are either probably unable to sustain a stable soluble structure, or prone to relinquish the folding information encoded in the amino acid sequence in favor of self-aggregation [57].

Table 5.1: List of extremely vulnerable proteins in human

Gene Symbol	Entrez Gene ID	SwissProt ID	Protein Name
HRNR	388697	Q86YZ3	Hornerin
ARID1B	57492	Q8NFD5	AT-rich interactive domain-containing protein 1B
ARID1A	8289	O14497	AT-rich interactive domain-containing protein 1A
RBM14	10432	Q96PK6	RNA-binding protein 14
FUS	2521	P35637	RNA-binding protein FUS
ILF3	3609	Q12906	Interleukin enhancer-binding factor 3
EP400	57634	Q96L91	E1A-binding protein p400
COL3A1	1281	P02461	Collagen alpha-1(III) chain precursor
HNRPUL1	11100	Q9BUJ2	Heterogeneous nuclear ribonucleoprotein U-like protein 1
KHSRP	8570	Q92945	Far upstream element-binding protein 2
MN1	4330	Q10571	Probable tumor suppressor protein MN1
KRT9	3857	P35527	Keratin, type I cytoskeletal 9
KRT10	3858	P13645	Keratin, type I cytoskeletal 10
TAF4	6874	O00268	Transcription initiation factor TFIID subunit 4
PEF1	553115	Q9UBV8	Peflin
RANBP9	10048	Q96S59	Ran-binding protein 9
MED15	51586	Q96RN5	Positive cofactor 2 glutamine/Q-rich-associated protein
MAML3	55534	Q96JK9	Mastermind-like protein 3
MED12	9968	Q93074	Mediator of RNA polymerase II transcription subunit 12
TAF15	8148	Q92804	TATA-binding protein-associated factor 2N
TFG	10342	Q92734	Protein TFG
MAML2	84441	Q8IZL2	Mastermind-like protein 2
SAMD1	90378	Q6SPF0	Atherin
ZFHX3	463	Q15911	Alpha-fetoprotein enhancer-binding protein
SS18	6760	Q15532	SSXT protein
EWSR1	2130	Q01844	RNA-binding protein EWS

Continued on next page

Table 5.1: List of extremely vulnerable proteins in human

Gene Symbol	Entrez Gene ID	SwissProt ID	Protein Name
ANXA11	311	P50995	Annexin A11
LOR	4014	P23490	Loricrin
HNRPA2B1	3181	P22626	Heterogeneous nuclear ribonucleoproteins A2/B1
POU3F3	5455	P20264	POU domain, class 3, transcription factor 3
ANXA7	310	P20073	Annexin A7
LGALS3	3958	P17931	Galectin-3
KRT1	3848	P04264	Keratin, type II cytoskeletal 1
ZIC2	7546	O95409	Zinc finger protein ZIC 2
FOXD2	2306	O60548	Forkhead box protein D2
SHANK1	50944	Q9Y566	SH3 and multiple ankyrin repeat domains protein 1
SRRM2	23524	Q9UQ35	Serine/arginine repetitive matrix protein 2
NOVA2	4858	Q9UNW9	RNA-binding protein Nova-2
COL17A1	1308	Q9UMD9	Collagen alpha-1(XVII) chain
PRR12	57479	Q9ULL5	Proline-rich protein 12
ZMIZ1	57178	Q9ULJ6	Zinc finger MIZ domain-containing protein 1
HCN2	610	Q9UL51	Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2
DACH1	1602	Q9UI36	Dachshund homolog 1
KCNN3	3782	Q9UGI6	Small conductance calcium-activated potassium channel protein 3
HECA	51696	Q9UBI9	Headcase protein homolog
ZSWIM5	57643	Q9P217	Zinc finger SWIM domain-containing protein 5
BMP2K	55589	Q9NSY1	BMP-2-inducible protein kinase
FZD8	8325	Q9H461	Frizzled-8 precursor
WDR33	55339	Q9C0J8	WD repeat protein 33
HNRPAB	3182	Q99729	Heterogeneous nuclear ribonucleoprotein A/B
ATXN2	6311	Q99700	Ataxin-2
PHOX2B	8929	Q99453	Paired mesoderm homeobox protein 2B
FUBP1	8880	Q96AE4	Far upstream element-binding protein 1
MYST3	7994	Q92794	Histone acetyltransferase MYST3
PHLDA1	22822	Q8WV24	Pleckstrin homology-like domain family A member 1
PDCD6IP	10015	Q8WUM4	Programmed cell death 6-interacting protein
ZNF384	171017	Q8TF68	Zinc finger protein 384

Continued on next page

Table 5.1: List of extremely vulnerable proteins in human

Gene Symbol	Entrez Gene ID	SwissProt ID	Protein Name
NKX2-3	159296	Q8TAU0	Homeobox protein Nkx-2.3
C14orf32	93487	Q8NDC0	Uncharacterized protein C14orf32
ENAH	55740	Q8N8S7	Protein enabled homolog
SP8	221833	Q8IXZ3	Transcription factor Sp8
MED13L	23389	Q71F56	Thyroid hormone receptor-associated protein 2
AMOT	154796	Q4VCS5	Angiomotin
TRIM71	131405	Q2Q1W2	Tripartite motif-containing protein 71
FOXD1	2297	Q16676	Forkhead box protein D1
ATXN8	724066	Q156A1	Ataxin-8
SF1	7536	Q15637	Splicing factor 1
CDSN	1041	Q15517	Corneodesmosin precursor
SUZ12	23512	Q15022	Polycomb protein SUZ12
NCOA6	23054	Q14686	Nuclear receptor coactivator 6
HNRPD	3184	Q14103	Heterogeneous nuclear ribonucleoprotein D0
HNRPA0	10949	Q13151	Heterogeneous nuclear ribonucleoprotein A0
SOX4	6659	Q06945	Transcription factor SOX-4
EVX2	344191	Q03828	Homeobox even-skipped homolog protein 2
MLL	4297	Q03164	Zinc finger protein HRX
POU4F1	5457	Q01851	POU domain, class 4, transcription factor 1
NKX6-1	4825	P78426	Homeobox protein Nkx-6.1
POU6F2	11281	P78424	POU domain, class 6, transcription factor 2
FOXL2	668	P58012	Forkhead box protein L2
ATN1	1822	P54259	Atrophin-1
HNRPA3	220988	P51991	Heterogeneous nuclear ribonucleoprotein A3
SMARCA2	6595	P51531	Probable global transcription activator SNF2L2
VASP	7408	P50552	Vasodilator-stimulated phosphoprotein
GSK3A	2931	P49840	Glycogen synthase kinase-3 alpha
YLPM1	56252	P49750	YLP motif-containing protein 1
HD	3064	P42858	Huntingtin
COL18A1	80781	P39060	Collagen alpha-1(XVIII) chain precursor
KRT2	3849	P35908	Keratin, type II cytoskeletal 2 epidermal

Continued on next page

Table 5.1: List of extremely vulnerable proteins in human

Gene Symbol	Entrez Gene ID	SwissProt ID	Protein Name
OTX1	5013	P32242	Homeobox protein OTX1
HOXA13	3209	P31271	Homeobox protein Hox-A13
POLR2A	5430	P24928	DNA-directed RNA polymerase II largest subunit
SFPQ	6421	P23246	Splicing factor, proline- and glutamine-rich
RFX1	5989	P22670	MHC class II regulatory factor RFX1
COL5A1	1289	P20908	Collagen alpha-1(V) chain precursor
POU3F2	5454	P20265	POU domain, class 3, transcription factor 2
EGR1	1958	P18146	Early growth response protein 1
AR	367	P10275	Androgen receptor
HNRPA1	3178	P09651	Heterogeneous nuclear ribonucleoprotein A1
SYP	6855	P08247	Synaptophysin
COL2A1	1280	P02458	Collagen alpha-1(II) chain precursor
COL1A1	1277	P02452	Collagen alpha-1(I) chain precursor
FMNL1	752	O95466	Formin-like protein 1
SS18L1	26039	O75177	SS18-like protein 1
LDB3	11155	O75112	LIM domain-binding protein 3
BRD4	23476	O60885	Bromodomain-containing protein 4
PHLPP	23239	O60346	PH domain leucine-rich repeat-containing protein phosphatase
WIPF1	7456	O43516	WAS/WASL interacting protein family member 1
HOXA3	3200	O43365	Homeobox protein Hox-A3
SETD1A	9739	O15047	Histone-lysine N-methyltransferase, H3 lysine-4 specific SET1
SYN3	8224	O14994	Synapsin-3
HGS	9146	O14964	Hepatocyte growth factor-regulated tyrosine kinase substrate
TCERG1	10915	O14776	Transcription elongation regulator 1
SOX1	6656	O00570	SOX-1 protein
WASL	8976	O00401	Neural Wiskott-Aldrich syndrome protein
FOXE1	2304	O00358	Forkhead box protein E1

All outlier interactions in the human $(\eta - \nu)$-correlation involve genes with extreme vulnerability (Figure 4.3 and Table 5.1). The $(\eta - \nu)$-correlation reported in Figure 4.3 for human is weaker than the yeast counterpart likely because, in contrast with yeast, mRNA levels are not a reliable surrogate for protein expression levels in human. Expression of human genes is subject to post-transcriptional regulation, and those genes encoding extremely vulnerable proteins should be under tight regulation. Recently, microRNA-mediated gene regulation has emerged as an important mechanism of post-transcriptional regulation. Here we study the post-transcriptional regulation of human genes encoding extremely vulnerable proteins by examining their microRNA targeting.

5.2 microRNA

microRNAs (miRNAs) are a novel class of non-protein coding RNAs that serve as post-transcriptional gene regulators in a wide variety of organisms [58, 59]. These endogenous 22 nucleotide RNAs negatively regulate gene expression by base-pairing with the 3' untranslated regions (3' UTRs) of target mRNAs [58]. miRNAs are ubiquitously expressed, and regulate in a number of cellular processes in worms, flies, fish, frogs, plants and mammals [60]. Over 6000 miRNAs have been discovered across all species, using molecular cloning and bioinformatics prediction [61]. Although the role of majority of miRNAs remain unclear, experimental data on a few of them show that they are involved in embryonal stem cell development, fat metabolism, neuronal differentiation, and cancer development [62, 63, 64, 65].

5.2.1 Discovery of miRNA

The first miRNA was discovered by the Ambros and Ruvkun laboratories in 1993, when researchers studied heterochronic gene *lin-4*-mediated temporal regulation of another heterochronic gene *lin-14* in C. elegans [66, 67]. It has been shown that a 22 nucleotide small RNA encoded in lin-4 has multiple imperfect complimentary sites in the 3' UTR region of *lin-14* mRNA [66, 67]. This finding suggests that *lin-4* regulates the protein levels of *lin-14* by binding to the 3' UTR region of the corresponding mRNA. As lin-4's homologs are not found in other species, this unique RNA based regulatory mechanism was thought to be present only in C. elegans. Things did not change until researchers discovered another heterochronic gene *let-7* which has homologs in other species including human and drosophila. Similar to *lin-4*, *let-7* encodes a 22 nucleotide regulatory small RNA [68]. This small RNA interacts with the 3' UTR of *lin-41*, and regulates gene expression of *lin-41* [68]. The identification of heterochronic gene *let-7* in *C. elegans* and other species revealed a new class of RNA molecules performing regulation of gene expression. This new class of small RNAs are then referred as microRNAs, abbreviated miRNAs, as several novel small RNAs with similar regulatory roles were identified [69, 70, 71].

5.2.2 Biogenesis of miRNA

miRNA is made from precursor miRNA (pre-miRNA), which in turn is the product of a miRNA primary transcript (pri-miRNA) (Figure 5.1). pri-miRNAs are transcribed from the genome, and then processed to the 60-70 nucleotide pre-miRNA in the nucleus. The former process is brought about by RNA polymerase II [72], and the latter process is promoted by the microprocessor complex consist of the nuclease Drosha and the double-stranded RNA binding protein Pasha [73]. The pre-miRNA is processed to the mature 22 nucleotide miRNA:miRNA* duplex by another Rnase III enzyme, Dicer. The mature miRNA is then released and incorporated into RNA-induced silencing complex (RISC), a ribonucleoprotein complex, whereas the miRNA* strand is typically degraded [74, 75]. The preference of the mature miRNA over the passenger strand by RISC may be partly due to the differences in the thermodynamic stability between the two strands [74]. The less thermodynamically stable 5' end of the mature miRNA renders it more unstable, and hence more favorable by the RISC.

5.2.3 Regulatory mechanism

MicroRNAs guide the RISC to regulate gene expression by either of two mechanisms: mRNA cleavage or translational repression. The choice of regulatory mechanisms is widely believed to be determined by the degree of complementarity of the miRNA with a certain region of its mRNA target. If the miRNA 22 nucleotides perfectly or near-perfectly match the mRNA sequence, the target mRNA can be cleaved and degraded; otherwise, its translation is repressed.

An endonuclease called Argonaute 2 is required for site-specific cleavage of the target. This protein contains a PAZ and a PIWI domain, which are characteristic of the proteins of the Argonaute family and the Dicer family. The human Argonaute family contains four Argonaute proteins, Argonaute 1-4 (Ago 1-4). Though all of four proteins bind to miRNAs with similar affinities, it is only AGO2 that displays endonuclease activity [76]. Structural data of AGO2 revealed that its PIWI domain has a strikingly similarity to Rnase H type

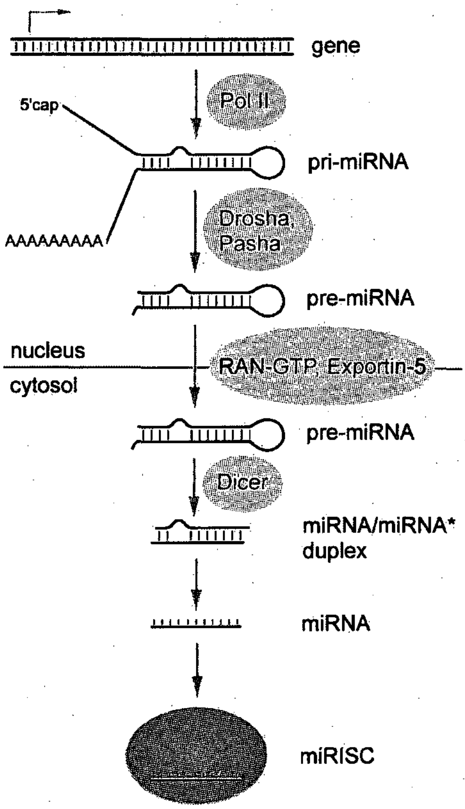


Figure 5.1: miRNA biogenesis process (from WIKI).

enzymes. This unique Rnase H-like PIWI domain is a key factor that is responsible for the endonuclease and site specific cleavage activity of AGO2. One mRNA cleavage case has been reported in animal [77].

Most animal miRNAs performs translational repression rather than cleavage on their mRNA targets due to the imperfect base-pairings between them and their targets. In the translational repression mode, mRNAs are not degraded of target mRNAs but can be destabilized as a result of deadenylation and subsequent decapping. The mechanism of translational repression by miRNA remains elusive. Controversy rises over the step at which miRNAs block translation. There is evidence for translational initiation block by miRNA, whereas other studies suggest that miRNA blocks the elongation of transcripts [78, 79]. Another issue is the role of processing bodies (P-bodies). P-bodies are cytoplasmic foci where ribosomal components do not exist and mRNAs can stay without being translated. Some researchers proposed that translational repression is mediated by the interaction between proteins in P-bodies and Argonaute proteins bound to miRNAs and their target mRNAs [80]. However, other argued that P-bodies may serve as temporary storage sites of translationally repressed mRNAs [81].

5.2.4 miRNA and disease

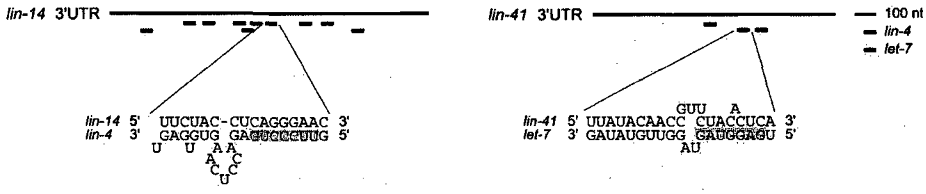
miRNAs play an important role in regulation of gene expression. Dysregulation of miRNA function, therefore, can lead to deleterious effects. Absence of mature miRNAs is lethal in animals [82, 83]. In *C. elegans*, mutation of miRNA-producing *dicer-1* leads to defects in germ-line development [84]. In Drosophila, depletion of Loquacious, the partner of Dicer-1, is responsible for female sterility [82]. In mammals, misexpression of miRNAs leads to deleterious biological consequences. Overexpression of the pancreatic islet-specific miR-375 suppressed glucose-induced insulin secretion. The deletion of its target, myotrophin, produce the same effect [85]. Conversely, inhibition of endogenous miR-375 function increases myotrophin levels and enhances insulin secretion, which indicates that miR-375 is an inhibitor of glucose-stimulated insulin secretion.

miRNAs are also associated with diseases in human. One example is the neuropsychiatric disorder Tourette's syndrome (TS) caused by the mutation in the 3' UTR of SLITRK1 [86]. A GU base pair is replaced by AU pairing, which results in stronger regulation by the miRNA. miRNAs may serve as tumor suppressors, which is implied by the loss of miRNA in cancer tissue. Chromosome region 13q14, location of miR-15a and miR-16-1 genes, is deleted in most of chronic lymphocytic leukemia cases [87]. Those two miRNAs target the antiapoptotic gene Bcl2, which indicates that depletion of miR-15a and miR-16-1 may lead to the inhibition of apoptosis and produce malignancies [88]. miRNAs can also be potential oncogenes. The miR17-92 locus 12q31 is overexpressed in some tumors [65]. Amplification of this cluster in a mouse model of human B cell lymphoma accelerated the formation of c-Myc-induced tumor [65].

5.3 microRNA targeting of genes encoding extremely vulnerable proteins

5.3.1 Target Identification

The first miRNA targets were identified from genetic interaction data in Caenorbabditis elegans. The mutation of heterochronic mRNA *lin-14* suppresses the phenotype caused by the mutation of miNRA *lin-4*. This fact led to the identification of sequence complementarity between the 3' UTR of *lin-14* and the 5' portion of *lin-4* (Figure 5.2). Despite its power, the genetic approach can identify only those targets whose overexpression results in the miRNA mutant phenotypes. There are few examples of this type. It remains unclear whether this sort of relationship is a general rule. Subsequently, the miRNA-target interactions were elucidated by the miRNA-target sites mutation and miRNA misexpression experiments [89, 90, 91]. These studies focused on the significance of pairing to the seed region located on the 5‘ end of the miRNA. The target sites on mRNAs can be grouped into two broad classes [92] : (a) 5’ dominant sites base pairing perfectly with the miRNA seed region, and (b) 3’ compensatory sites, with insufficient support from 5’ pairing to miRNAs’ 3’ region.


lin-14 3'UTR
lin-41 3'UTR
100 nt
lin-4
let-7
lin-14 5' UUCUAC-CUCAGGGAAC 3'
lin-4 3' GAGGUG GAGUCCCUUG 5'
U U AA
A C
C C
U
GUU A
lin-41 5' UUAUACAACC CUACCUCA 3'
let-7 3' GAUAUGUUGG GAUGGAGU 5'
AU

Bushati N, Cohen SM. 2007.
Annu. Rev. Cell Dev. Biol. 23:175–205

Figure 5.2: The miRNAs *lin-4* and *let-7* repress gene expression of their targets through imperfect base-pairing with the target 3UTRs [92].

5.3.2 Computational target prediction

Computational approach to the miRNA targets identification has been ongoing ever since the discovery of the first miRNA. Identification of hundreds of miRNAs in a variety of species and relatively small sets of targets pointed at the urgent need for accurate and efficient target prediction.

The characteristics of miRNAs give arise to some specific problems and difficulties that hinder accurate target prediction. First, miRNAs have only 22 nucleotides, and do not exhibit perfect complementarity to the 3’ UTRs of their target transcripts. This characteristic makes it inapproriate to implement standard sequence analysis techniques that were designed for searching long sequence match. Second, the location, extent or splice-variation of 3’ UTRs are not known for mammals. In human, there are roughly 30% of genes whose exact extent of the 3' UTR can not be delineated. Third, many target prediction approaches employ conservation of UTRs across species as a key filter for target detection, which leads to failure of identification of unconserved targets. For most mammalian miRNA, their targets' potential binding sites are conserved in orthologous URTs in multiple species. However, there exist classes of relatively recently evolved miRNAs (e.g. miR-430 in Zebrafish), whose targets do not share significant sequence similarity [93].

The computational target prediction methods developed so far fall into several categories. Their basic idea is to search sequence complementarity or favorable miRNA:target duplex thermodynamics. Most methods improve results by applying filters such as conservation of binding site and the presence of multiple sites. Many methods also require precise complementarity between the seed region of miRNAs and their target to further reduce false positives. After these filtering steps, a significance score is typically calculated for each potential target.

One of the most cited algorithms is TargetScan [91]. Firstly, it detects targets by examining their complementarity to the seed region of a miRNA. Only those perfectly match are considered for further analysis. The method then analyzes the extent of complementarity outside the seed region. Unlike many other algorithms, which tend to find all potential targets and then iteratively filter them, TargetScan seeks to eliminate false-positives as many as possible in the early stage. Groups of orthologous sequences are also used as input to filter out unconserved sites early on. The initial analysis using TargetScan predicted miRNA targets in Humans and performed conservation analysis using the Mouse, Rat and Fish genomes [91]. Shuffled sequences were then used to estimate a false-positive rate of between 22-31%. A very large scale and detailed validation of predicted targets found that

TargetScan not only predicted known miRNA binding sites but also novel sites [91]. Compared to other algorithms, TargetScan tends to reduce more false-positives, and is hence a good candidate for large-scale prediction. However, it probably misses those targets that do not pair perfectly in the seed regions or that are conserved poorly. A simpler version of TargetScan, TargetScanS, was developed later and exhibited higher target prediction fidelity [94].

5.3.3 Target prediction results

To obtain statistics on miRNA targeting, we identified putative target sites in the 3' UTRs (untranslated regions) of 17444 genes for 162 conserved miRNA families by using TargetScanS (version 4.0) [94]. 7,927 genes (45.4%) are predicted to contain at least one miRNA target site (Additional file 6), while 87 out of 105 (82.9%) extremely vulnerable genes are predicted to be targeted genes. Thus, human genes containing extremely vulnerable regions are more frequently targeted by miRNA ($P << 1.31x10^{-5}$, binomial test). In regards to miRNA regulation complexity, the mean number of miRNA target sites for human genes is 2.66 and the median is 0, while the mean number for extremely vulnerable genes is 6.01 and the median is 5. This significant difference ($P < 10^{-16}$, Wilcox rank test) strongly suggests that the deviation of extremely vulnerable genes from the $(\eta - \nu)$-correlation, with expression correlation evaluated at the level of mRNA expression, can be explained by a post-transcriptional miRNA regulation. This type of regulation influences the final protein expression level. In a broad sense, this analysis highlights the connection between protein structure and gene regulation: extremely vulnerable genes require tight control at the post-transcriptional level.

5.4 Primitive post-transcriptional regulation of genes encoding extremely vulnerable proteins in yeast

All outlier interactions in the human $(\eta - \nu)$-correlation involve genes with extreme vulnerability. When the same criterion for extreme vulnerability is applied to scan the yeast genome, 85 genes (Table 5.2) are identified whose ORFs contain the five confirmed prion proteins for this organism [54, 53, 55, 95]: PSI+ (SUP35), NU+ (NEW1), PIN+ (RNQ1), URE3 (URE2) and SWI+ (SWI1). Prions are originally found to be involved in mammalian neurodegenerative diseases where the aggregation of misfolded prion proteins causes neurodegenerative disorders. The prion concept was expanded to yeast to explain the unusual non-Mendelian behavior of some yeast genetic elements. The fact that five yeast proteins are identified to be extremely vulnerable indicates a relation between structural vulnerability of the soluble fold and aggregation propensity.

5.4.1 Prion diseases

Prions are infectious agents that are responsible for a variety of mammalian neurodegenerative diseases generally referred to as transmissible spongiform encephalopathies (TSE) or prion-diseases [96, 97, 98]. These diseases include: scrapie in sheep; bovine spongiform encephalopathy (also called "mad cow" disease) in cattle; chronic wasting disease (CWD) in deer and elk; Creutzfeldt-Jakob disease (CJD), Gerstmann-Straussler-Scheinker syndrome and kuru in human. Although the clinical, epidemiological, and neuropathological features of these diseases are very different, they all involve modification of the prion protein (PrP), a host encoded protein predominantly expressed in the central nervous system of the mammals [99]. Prion diseases cause neurodegenerative disorders that are sporadic, inherited and transmissible degenerative [97, 98].

5.4.2 Prion protein structure

The prion protein is a 253-residue protein encoded by the gene Prnp, which consists of a signal peptide for secretion, five octapeptide repeats near the ends of sequence, two glycosylation sites, and one disulfide bridge [100]. It is expressed in most tissues, but predominantly in neuronal tissues. The prion protein has two isoforms: the normal cellular prion protein PrP^C, rich in a-helical conformation, is soluble and protease-sensitive; the disease-associated misfolded prion protein PrP^{Sc}, rich in β-sheet conformation, is insoluble and mostly protease-resistant.

PrP^C consists of an unordered N-terminal fragment and a globular C-terminal domain. The N-terminal domain, the segment for residue 1-128, is characterized by the octapeptide repeats, while the C-terminal is made of three a-helices and two small b-strand regions. Despite a number of amino acid differences, the 3D structure of PrP^C is highly conserved across several species of mammals such as human, mouse, cattle, sheep, and so on [101].

Although the structure of PrP^{Sc} has not been fully understood, experimental data obtained by using X-ray and other biophysical techniques and computational modelling of small peptide fragments have provided insights into the structural rearrangement during PrP^{Sc} formation [102, 103]. The structure transformation mostly involves the conversion of α-helices into β-sheets in the globular C-terminal domain of the protein. The current models for PrP^{Sc} structure represents the antiparallel β-sheets conformation, which become stabilized upon oligomerization with other PrP^{Sc} proteins (Figure 5.3).

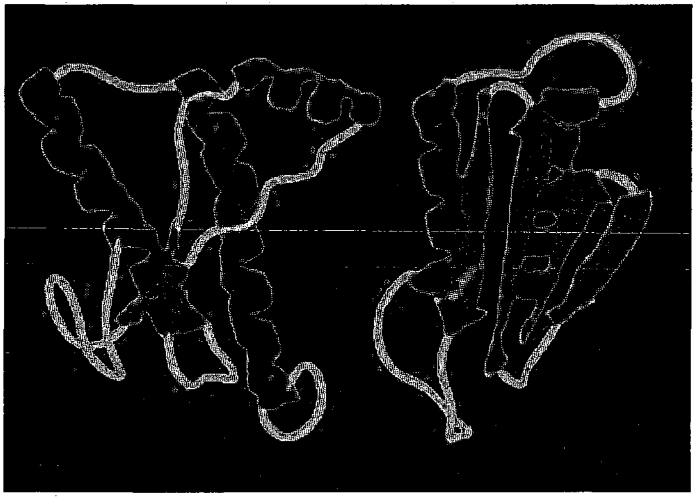


Figure 5.3: Two conformations of prion domain. Prion in normal condition PrP^C is displayed on the left, and its model of disease-related structure PrP^{Sc} is on the right (from Fred Cohen Laboratory, UCSF). [92].

5.4.3 The prion hypothesis

The nature of the transmissible agent of TSE has been extensively studied [104]. Initially, the agent was thought to be a slow virus, because the incubation period between the time of exposure to the pathogen and the onset of symptoms is unusually long compared to other virus diseases. However, further research has shown that the agent is not likely to a virus. The minimum molecular weight of the agent to maintain infectivity was much smaller than a virus or any other known type of infectious agent [105]. The normal treatments, which destroy nucleic acids, could not kill the infectious agents. Furthermore, the attempt to find a virus associated with the disease have been unsuccessful over the past 30 years [97]. These and other results led Griffith to propose the "protein-only" hypothesis. This hypothesis stated that the disease agent was a protein that was able to replicate itself in the body [106]. It gains great support from the successful isolation of a protease-resistant and misfolded protein from the infectious material by Stanley Prusiner's group in 1982 [107]. This protein was named as "prion", derived from proteinaceous and infectious [107].

There is a lot of evidence which supports the prion hypothesis. Many cases of the evidence were contributed by Prusiner's group. They have shown that the concentration of the protein was proportional to the infectivity titer [108]. Infectivity was reduced by agents, whose structures had been destroyed, as well as anti-PrP antibodies. In addition, infectivity was shown to be retained in highly purified PrP^{Sc} environment without other components. Another important evidence from Prusiner's group is the finding that mutation of PrP gene is linked to other cases of TSE. This result indicates that the genetic disease can be propagated in an infectious way. There are also supports from other groups. One particular strong support came from Charles Weissmann's group, who showed that the PrP-deleted mice were resistant to scrapie infection. There were no signs of scrapie nor propagation of the infectious agent in those mice. These and other results provide compelling evidence for the prion hypothesis and almost settle the debate over the nature of the infectious agent. The prion protein is the only component necessary to carry the infectivity.

5.4.4 Yeast prions

The prion concept was expanded by Reed Wickner in 1994, to explain the unusual non-Mendelian transmission of two yeast genetic elements termed [*URE3*] and [*PSI*$^+$] [109]. Those two traits were discovered 40 years ago, and could not be attributed to known non-Mendelian elements, like viruses, episomes or mitochondrial genes. Wickner proposed that [*URE3*] and [*PSI*$^+$] are the prion forms of the Ure2 and Sup35 proteins respectively.

Ure2p plays an important role in the cellular response to the nitrogen source, and Sup35p is involved in translation termination [110, 111]. Aggregation of these two proteins leads to the loss of function and produces prion phenotypes. Take for example Sup35p, which is a component of the translation termination complex. The protein aggregation occurs spontaneously at low frequency, and then recruits all normal Sup35p molecules into the prion state [112]. Figre 5.4 shows the mechanism of loss-of-function as a result of Sup35p aggregation. The prion state is passed on to the daughter cell when yeast divides. Since Wickner's proposal in 1994, extensive studies have provided strong support for the prion hypothesis in yeast [113, 114]. In addition, there are several other proteins in yeast and other fungi were found to exhibit the prion phenomenon [95, 113].

Like prions of mammals, Yeast prions transmit protein structural information in the absence of nucleic acid. They are all based upon the ability to self-replicate on their own. However, there are several important differences between two kinds of prions. In mammals, prions spread from cell to cell, whereas, in yeast, prions are transferred from mother cells to their daughters. Yeast prions do not kill the host cells like mammalian prions. They produce new metabolic phenotypes. Thus, yeast prions act as heritable determinants of phenotype. Although the research studies on yeast prions began much later, remarkable progress has been achieved and made a great contribution to understand the underlying biology of prions. One important progress comes from studies by Weissmann's group. The in vitro converted purified Sup35 prion domain was introduced to the cytoplasm of living yeast using a liposome transformation protocol. [PSI^+] prion appeared in 1 to 2% of transformed cells [116]. Another similar studies showed that the introduction of fibrils made in vitro from renatured recombinant HET-s to the mycelia of P. anserine induced efficient

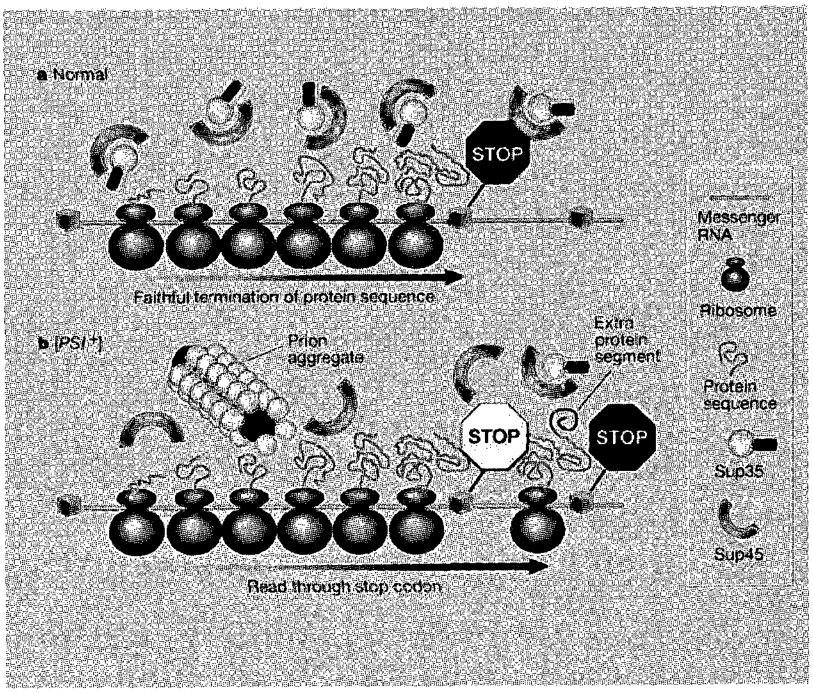


Figure 5.4: Sup35 protein aggregation results in a loss-of-function. a, The normal Sup35 protein functions as a translation terminator. It interacts with Sup45, and causes 'read-through' of stop codons. b, In [PSI+] cells, the misfolded Sup35 protein forms aggregates and fails to interact properly with the termination complex. As a result, stop codons are sometimes missed, producing increased amounts of proteins [115].

formation of the [Het-s] prions [117]. De novo generation of infectivity was demonstrated by introducing amyloid fibrils incubated with yeast-derived infectious aggregates into uninfected yeast hosts [118]. The fact that the amyloid fibres nucleated in vitro propagate the prion phenotype implies that the heritable information of distinct prion strains is based on the folding patterns of the same prion protein.

A common characteristic between mammalian and yeast prions is that the formation of β-sheet-rich aggregates that resemble amyloid fibrils. Protein aggregation is not only a typical characteristic of prions, but also a key step for prion propagation. For the five know yeast prion proteins, Sup35, Ure2, Rnq1, New1 and Swi1, aggregation is driven by their Gln/Asn-rich domains. However, only Gln/Asn-richness is not sufficient to induce protein aggregation. Gln/Asn-rich domains of other proteins have been appended to MC (the middle and C-terminal prion domain) of Sup35. Those domains from confirmed prions such as Rnq1 and New1 could replace the N terminal prion domain of Sup35 and form a [*PSI*$^+$]-like prion, whereas the domain from Pan1 could not [56, 119, 120]. Further research on prion domains of Sup35, Rnq1 and New1 showed that those prion domains have a positive bias for tyrosine, glycine and serine, and a negative bias for glutamate, aspartate, arginine [121].

Another important feature of yeast prions is their oligopeptide repeat sequences. There are five imperfect repeats (R1-R5) and one partial repeat (R6) in residues 41-97 of Sup35 that also compose the only sequence section similar to mammalian prion protein PrP. Deletion of two or more oligopeptide repeats in Sup35 destroy [*PSI*$^+$], whereas two additional copies of R2 increase dramatically the spontaneous appearance of prion state [122]. It has been suggested that the repeats might facilitate the correct alignment of intermolecular contacts between molecules [54]. Consistently, appending of a polyglutamine tract to the MC of Sup35, does not support [*PSI*$^+$], but addition of Sup35 repeats induces the prion formation , although proteins aggregates in both cases [123]. Their result indicates amyloid fibres are not necessarily prions.

Prion aggregation results in a loss-of-function of native proteins. For instance, Sup35 plays an important role in translation termination, and formation of [*PSI*$^+$] produces termination-defective phenotypes [124, 125]. [*PSI*$^+$] causes ribosomes to read through some nature occurring stop codons. It seems that [*PSI*$^+$]-mediated disruption to translation-termination could not lead to beneficial consequences. However, the conservation of Sup35 prion domain and its ability to switch to [*PSI*$^+$] state over several hundred million years implies that [*PSI*$^+$] might confer some advantages over the normal Sup35 protein (Figure 5.5 [126]).

An assessment of the fitness of [*PSI*$^+$] cells and [psi-] cells in 150 diverse growth conditions has provided important support to the hypothesis that [*PSI*$^+$] state could be beneficial [127]. The fitness of [*PSI*$^+$] cells increases in 25% of conditions in at least one genetic background, and decreases in another 25% of conditions [127]. Furthermore, [*PSI*$^+$] could induce profound alterations in colony morphology or stress tolerance [127, 128]. A variety of beneficial and heritable phenotypes arisen from subtle [*PSI*$^+$] alterations in translation-termination fidelity [127].

Studies on beneficial roles of prions have also been extended to other prions [129, 130]. The function of [*RNQ*$^+$], [*URE3*] and [*PIN*$^+$] is not well characterized. However, they can all induce the [*PSI*$^+$] formation [129, 130].

Prions are epigenetic, because their phenotypes can be inherited without modification of the genome. This characteristics provide a survive advantage in the fluctuation environments. Prions also serve as possible evolutionary capacitors and have essential roles in long-term memory formation [126].

5.4.5 Prion aggregation as a means of regulating gene expression of yeast extremely vulnerable proteins?

Extremely vulnerable proteins are subject to significant levels of post-transcriptional regulation. In human, this extra regulation is achieved through extensive miRNA targeting

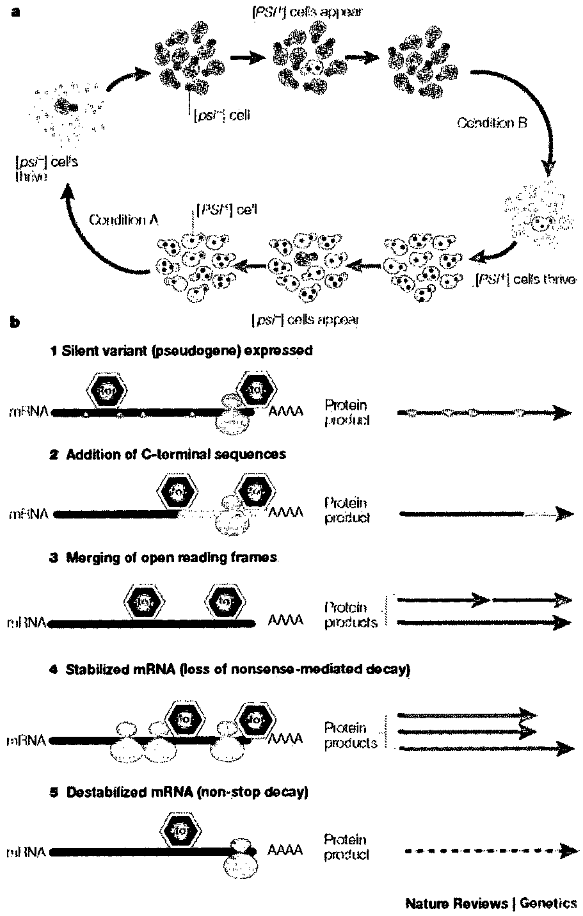


Figure 5.5: [PSI^+] may confer some advantages over the normal Sup35 protein [126]. a, Transition between [PSI^+] and [psi^-] states. [PSI^+] individuals appear spontaneously in a population of [psi^-] cells, and become dominant in some particular environment (condition B). When the situation changes (condition A), [psi^-] individuals thrive and [psi^-] cells die gradually. The transition between two states enables yeast cells to survive in some extreme circumstances. b, Expression of the usually silent genetic information as a result of readthrough of stop codons. (1) The expression of pseudogenes that are mutated in silent state may produce new functions. (2) C-terminal extensions on polypeptides perhaps alter protein function. (3) Two open reading frames are merged to yield new hybrid proteins. (4) Nonsense-mediated decay pathways are repressed to stabilize mRNAs. (5) Non-stop decay destabilizes mRNAs and alters the expression levels of proteisn.

of genes encoding extremely vulnerable proteins. In budding yeast, on the other hand, our results indicate that such a regulation is likely achieved through sequestration of the extremely vulnerable proteins into aggregated states. All five experimentally verified prions in budding yeast are found to be in the 85 extremely vulnerable protein list (Table 5.2). Unlike mammalian prions, yeast prions are potentially beneficial to the survival of cells in some specific circumstances [127]. Prion aggregation in yeast may provide some selective advantages [126]. The inclusion of five prion proteins in the extremely vulnerable proteins implies that if the extremely vulnerable proteins are themselves translational regulators, the sequestration to aggregated states may directly lead to epigenetic consequences and phenotypic polymorphism [127, 128]. Whether prion aggregation serves as a potential mechanism of gene expression regulation in yeast is an issue worthy of investigation. The experimental verification of other 80 extremely vulnerable proteins as prions can provide strong support to this hypothesis.

Table 5.2: List of extremely vulnerable proteins in yeast

SwissProt ID	Protein Name
P05453	Eukaryotic peptide chain release factor GTP-binding subunit (ERF2)
P23202	Protein URE2
P25367	[PIN+] prion protein RNQ1
Q08972	[NU+] prion formation protein 1
P09547	Transcription regulatory protein SWI1
P10591	Heat shock protein SSA1
P25339	Protein PUF4
P40467	Probable transcriptional regulatory protein YIL130W
P38216	Uncharacterized protein YBR016W
P19158	Inhibitory regulator protein IRA2
P40485	Phosphatidylinositol 4,5-bisphosphate-binding protein SLM1

Continued on next page

Table 5.2: List of extremely vulnerable proteins in yeast

SwissProt ID	Protein Name
P32505	Nuclear polyadenylated RNA-binding protein NAB2
P34761	Protein WHI3
P25299	mRNA 3'-end-processing protein RNA15
P25294	Protein SIS1
P38042	Anaphase-promoting complex subunit CDC27
P32334	Protein MSB2
P22470	Protein SAN1
P29295	Casein kinase I homolog HRR25
P32521	Protein PAN1
Q12489	Uncharacterized protein YDL012C
P50109	Protein PSP2
Q06449	[PSI+] inducibility protein 3
P11938	DNA-binding protein RAP1
P39105	Lysophospholipase 1 precursor
P19659	RNA polymerase II mediator complex subunit 15
P38741	Probable RNA-binding protein YHL024W
P47033	Protein PRY3
P40956	Protein GTS1
Q05672	RNA-binding suppressor of PAS kinase protein 1
P53281	LAS17-interacting protein 1
P32629	Mannan polymerase II complex ANP1 subunit
P22082	Transcription regulatory protein SNF2
Q08601	Metacaspase-1 precursor
P33417	Intrastrand cross-link recognition protein
P43582	WW domain-containing protein YFL010C
Q08954	Uncharacterized protein YPL199C
Q12224	Transcription factor RLM1
Q12221	Protein PUF2
P18899	Stress protein DDR48
P25644	Topoisomerase II-associated protein PAT1
Q03761	Transcription initiation factor TFIID subunit 12

Continued on next page

Table 5.2: List of extremely vulnerable proteins in yeast

SwissProt ID	Protein Name
P11746	Pheromone receptor transcription factor
P14680	Dual specificity protein kinase YAK1
Q03482	Uncharacterized protein YDR210W
Q02792	5'-3' exoribonuclease 2
Q02799	Zinc finger protein LEE1
P38429	Transcriptional regulatory protein SAP30
Q05854	Probable transcriptional regulatory protein YLR278C
Q04951	Probable family 17 glucosidase SCW10 precursor
P18480	Transcription regulatory protein SNF5
Q02630	Nucleoporin NUP116/NSP116
P80667	Peroxisomal membrane protein PAS20
Q12361	G protein-coupled receptor GPR1
P27654	Temperature shock-inducible protein 1 precursor
P38856	Hypothetical 71.7 kDa protein in REC104-SOL3 intergenic region
Q03825	Hypothetical 85.0 kDa protein in HLJ1-SMP2 intergenic region
P53894	Serine/threonine-protein kinase CBK1
P34756	1-phosphatidylinositol-3-phosphate 5-kinase FAB1
Q08732	Serine/threonine-protein kinase HRK1
P38266	Uncharacterized protein YBR108W
P53214	Hypothetical 57.5 kDa protein in VMA7-RPS25A intergenic region
P34217	RNA-binding protein PIN4
Q07800	Phosphatase PSR1
P38248	Extracellular matrix protein 33 precursor
P40552	Cell wall protein TIR3 precursor
P10863	Cold shock-induced protein TIR1 precursor
Q45U13	Yil130wp
P31384	Glucose-repressible alcohol dehydrogenase transcriptional effector
P38129	Transcription initiation factor TFIID subunit 5
P14922	Glucose repression mediator protein CYC8
P35732	Uncharacterized protein YKL054C
P19097	Fatty acid synthase subunit alpha

Continued on next page

Table 5.2: List of extremely vulnerable proteins in yeast

SwissProt ID	Protein Name
P54785	Transcriptional activator/repressor MOT3
P53438	Protein SOK2
P53334	Probable family 17 glucosidase SCW4 precursor
P40002	Uncharacterized protein YEL007W
Q08906	Facilitator of iron transport 2 precursor
P04050	DNA-directed RNA polymerase II largest subunit
Q05785	Epsin-2
P39743	Reduced viability upon starvation protein 167
P38180	Uncharacterized protein YBL081W
P32583	Suppressor protein SRP40
P41910	Repressor of RNA polymerase III transcription MAF1
P38996	Nuclear polyadenylated RNA-binding protein 3

5.5 Summary

The correlation between gene co-expression and protein structure vulnerability is weaker in human than the correlation in yeast. The outliers in human correlation involve extremely vulnerable proteins, which are more needy for protection from their binding partners. Genes encoding extremely vulnerable proteins are dosage sensitive, and hence require tighter post-transcriptional control. As shown in microRNA target prediction results, genes encoding extremely vulnerable proteins are more frequently targeted by microRNAs in human. In budding yeast, there are no signs of RNA interference. However, yeast extremely vulnerable protein list contains five conformed prions. Prion protein aggregation is potentially beneficial for cell survival, as it can produce diverse phenotypes in different environments. Post-transcriptional regulation of genes encoding extremely vulnerable proteins is proposed to be achieved through sequestration of the extremely vulnerable proteins into aggregated states. Experimental verification of those extremely vulnerable proteins as prions could further provide support to this theory.

Chapter 6

Protein structure vulnerability decreases gene duplicability

Chapter 4 tests the prediction that protein structure vulnerability provides molecular basis for gene dosage sensitivity, by examining the gene co-expression and structure vulnerability of interacting proteins. This prediction can also be verified by examining the effect of protein structure vulnerability on gene duplication. Protein structure vulnerability quantifies the extent to which protein structure relies on binding partners to maintain its integrity. Highly vulnerable proteins are needy for association with binding partners. According to our prediction, duplicates of genes encoding highly vulnerable proteins should be more likely to cause dosage imbalance and hence be less frequently to be retained in evolution. This chapter examines this deduction by analyzing the relationship between protein structure vulnerability and gene duplicability.

6.1 A negative effect of protein structure vulnerability on gene duplication

Gene duplication is one of the key factors producing new genetic variants [131, 132]. Recently, the evolutionary forces influencing gene duplicability have received intense interest. In particular, much effort has been devoted to explain gene duplication patterns at the genomic level using the gene dosage balance theory described in Chapter 2 [23]. According to the dosage balance hypothesis, dosage sensitive genes are less likely to be retained in evolution, and have fewer paralogs than dosage insensitive genes. In Chapter 4, we show that protein structure vulnerability quantifies the level of gene dosage sensitivity. As a result, we predict that the probability of retention of gene duplicates in evolution (i. e., gene duplicability) depends on the structure vulnerability of the protein encoded by the gene.

We collaborated with Li group to test this prediction [133]. We compiled a non-redundant set of proteins with PDB-reported structure in six organisms: *E. Coli*, yeast, worm, fly, human and thale cress. Then we determined both the structure vulnerability (or called as under-wrapping) and duplicability for each protein. Structure vulnerability was calculated as shown in chapter 3. Gene duplicability is quantified by the number of members in a gene family (i.e., the gene family size). *E. Coli* gene family annotation was obtained from Genome and Proteome Database, while family annotations for other five species were extracted from Ensembl Database [134]. We found that the under-wrapping decreases with increasing gene duplicability in all six organisms (Figure 6.1 and 6.2).

Genes with particular biological functions have been shown to duplicate more fre-

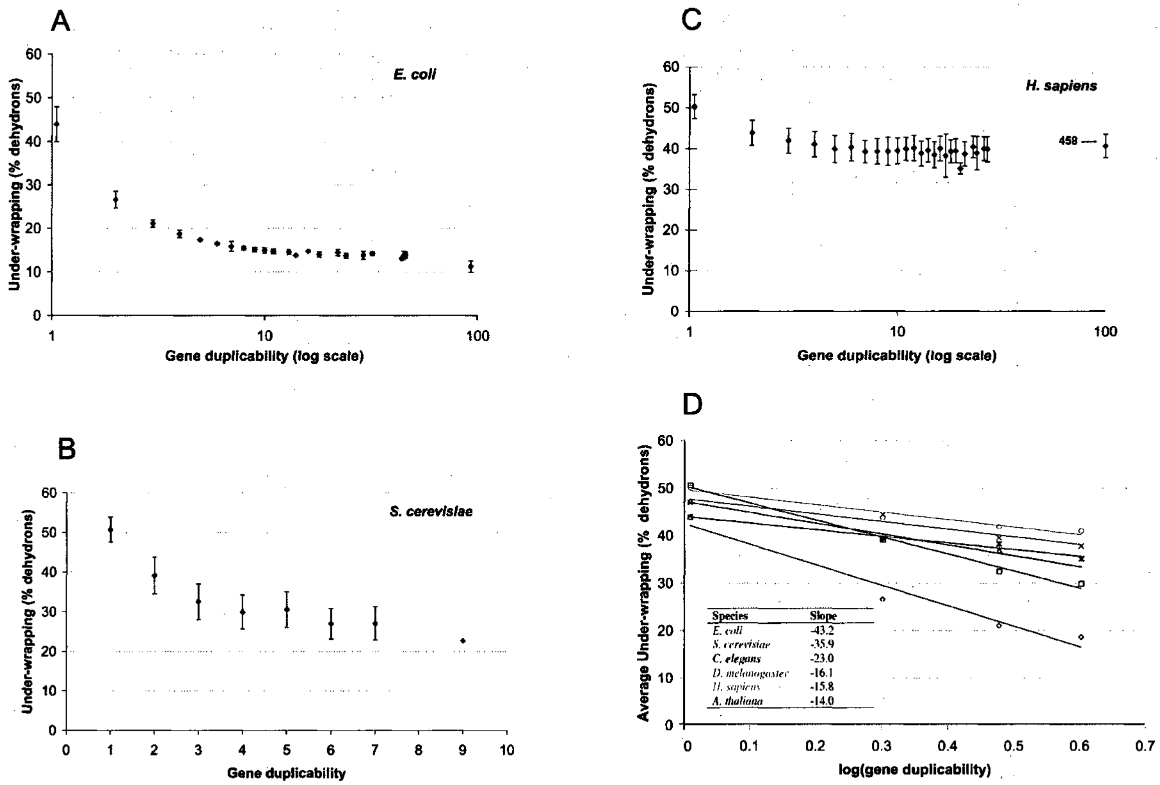


Figure 6.1: Anti-correlations between under-wrapping and gene duplicability in E. coli (A), in yeast (B) and, in human (C), and in slopes in six organisms (D). Gene duplicability is defined as the gene family size. The mean level of wrapping is calculated by averaging over all genes with the same duplicability.

quently in evolutionary history [135, 136]. We then investigated the potential influences of functional bias on our results, by comparing the under-wrapping levels between yeast singletons and duplicates for different functional categories. It turned out that singletons are consistently more under-wrapped than duplicates for each functional category. This result shows that the effect of the protein under-wrapping on gene duplicability does not depend on the gene function (Figure 6.3).

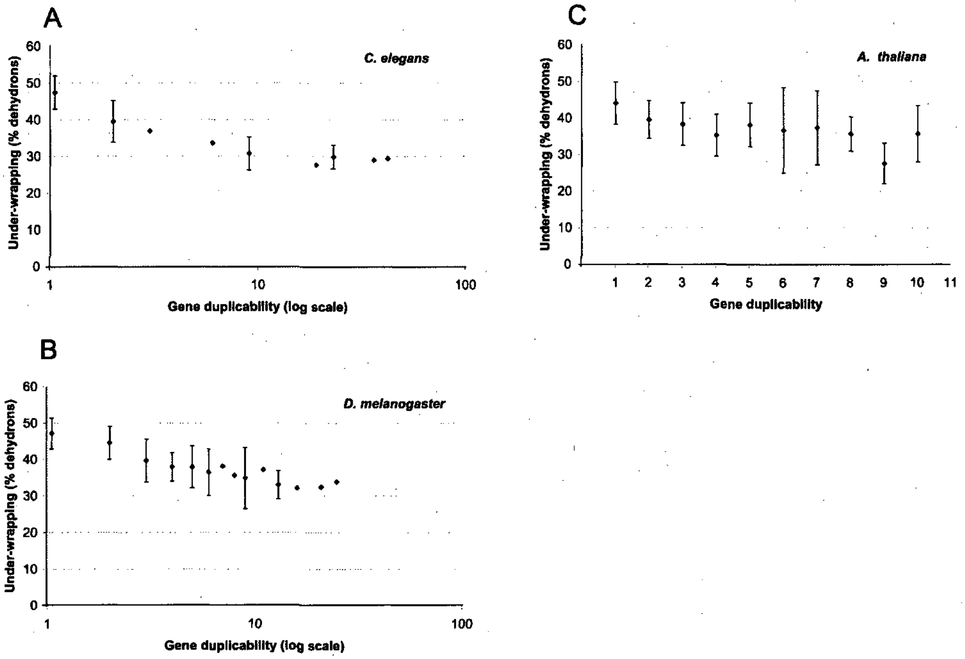


Figure 6.2: Anti-correlations between under-wrapping and gene duplicability in worm (A), fly (B), and thale cress (C).

6.2 Protein structure vulnerability dependency of gene duplication varies across species

Anti-correlations between protein under-wrapping and gene duplicability in six organisms strongly support our prediction that gene duplication is dependent on the level of protein under-wrapping. For all species, the decreasing trend is evident for genes with family size less than 5 and become less obvious at higher duplicability. However, the extent of correlation between two quantities varies across six organisms. The correlation differences

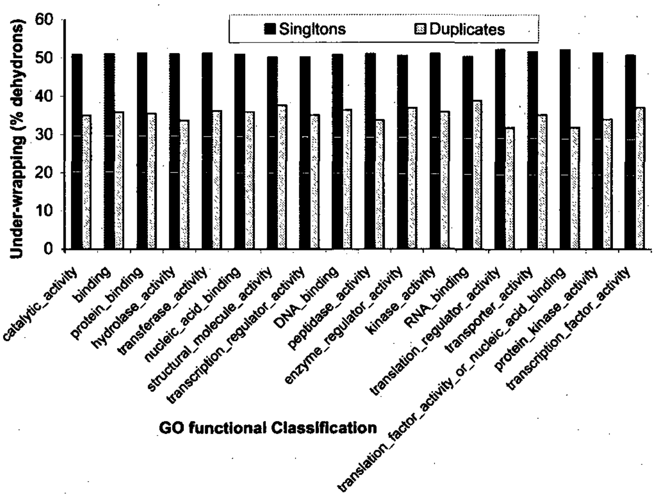


Figure 6.3: Yeast singletons are more under-wrapped than duplicates in all the functional categories.

can be seen more clearly using a linear regression between protein under-wrapping and gene duplicability in six organisms. As shown in Figure 6.1D, the effect of protein under-wrapping on gene duplicability decreases with increasing organismal complexity, that is, *E. Coli* > yeast > worm > fly $\sim$ human $\sim$ thale cress. It suggests that dosage imbalance may play a less important role in complex organisms. To investigate the correlation differences between organisms, we studied the under-wrapping distributions in *E. coli*, yeast and human. The results are shown in Figure 6.4. The under-wrapping of human proteins is mainly distributed between 35% and 55%, while the distributions of under-wrapping of proteins are wider in *E. Coli* and yeast. It implies that more human proteins are reliant on binding partners for structure integrity. The contrasting under-wrapping distributions provide some clue to understand the correlation difference between organisms. However, it still needs more investigation.

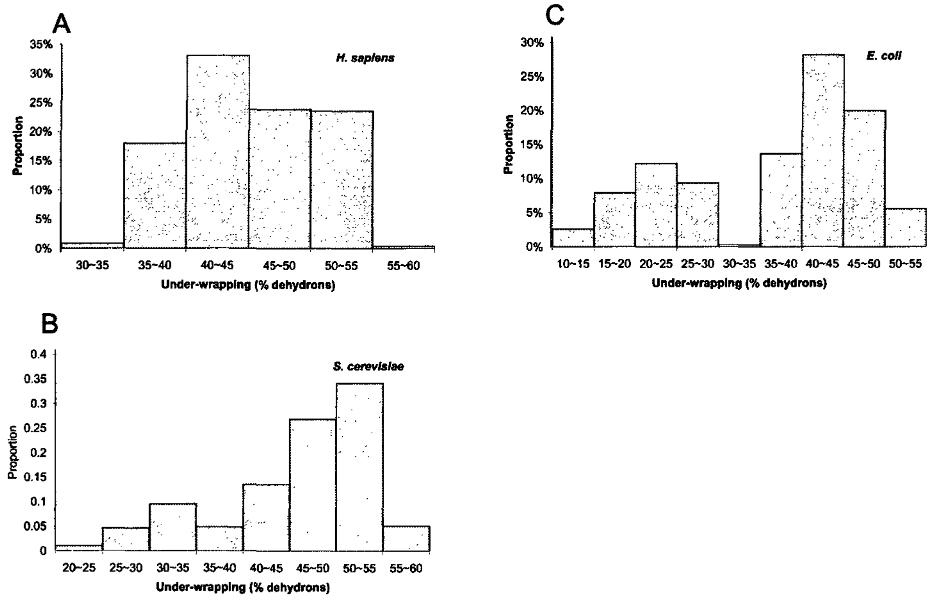


Figure 6.4: Distributions of percentage of families based on wrapping levels in Human (A), in Yeast (B), and in *E. Coli* (C).

In higher eukaryotes, genes encoding highly under-wrapped proteins have more paralogs, suggesting that complex organisms are less sensitive to the dosage imbalance effects. There are several possible reasons. First, there are more efficient expression regulatory systems in complex organisms. As we discussed in Chapter 5, microRNAs down-regulate expression of dosage sensitive genes. Second, sequence divergence in higher eukaryotes may be more significant, which help them to avoid dosage imbalance. Third, paralogs can interact with each other [137]. This phenomenon may be more prevalent in higher eukaryotes. Fourth, complex organisms generally have a smaller effective population size [138], which leads to less chance of dosage imbalance. The last possible factor is the positive selection due to functional diversification in complex organisms [139].

6.3 Vulnerability-duplicability correlation differences between whole-genome duplication (WGD) and non-WGD duplicates

The effect of protein under-wrapping on gene duplicability depends on the scale of duplication. In a whole-genome duplication (WGD) every gene in the genome is duplicated at the same time, while only part of the genome duplicates in a non-WGD (including individual or segmental duplication). Therefore, duplicates of highly under-wrapped proteins in WGD should be less likely to result in dosage imbalance, and have more chances of surviving from duplications. Focusing on the yeast proteins with only one paralog, we found a statistical significant difference between the under-wrapping levels in two kinds of duplication (Figure 6.5 A). Proteins surviving from WGDs have a higher under-wrapping level that those from non-WGDs, which implies that the dosage imbalance effect was relaxed for gene duplications in WGD. The gene ontology (GO) analysis found that this trend is present for genes from different function categories (Figure 6.5 B).

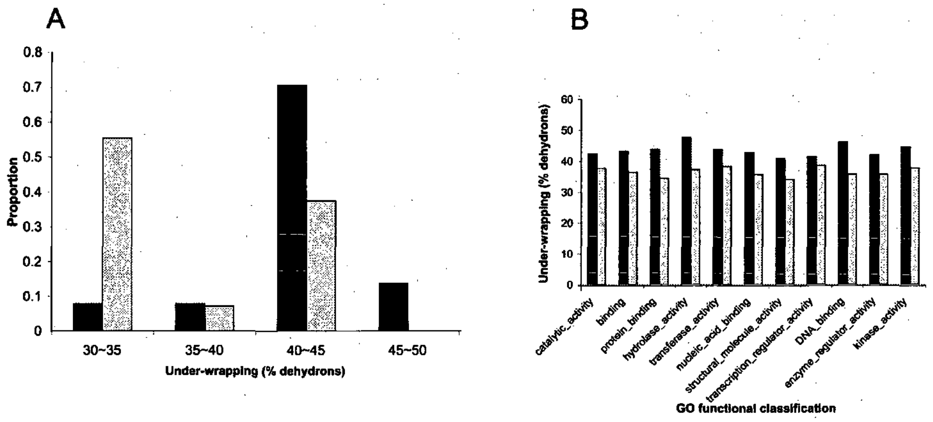


Figure 6.5: (A) Contrasting wrapping level distributions between WGD (Black bar) and Non-WGD (grey bar) duplicates. (B) Wrapping level distributions over gene functions for two groups. GO mapping for yeast genes is provides by a GO term analysis tool-GO term finder [140].

6.4 Summary

Gene duplication can produce new gene copies, and hence become one of main forces driving genetic innovations. However, segmental duplication may lead to a dosage imbalance between interacting partners. Highly vulnerable proteins, which are highly reliant on binding partners to maintain their structure integrity, are less likely to be duplicated during gene duplications. Results in this chapter provide additional support to our prediction that protein structure vulnerability serves as a molecular basis for the dosage imbalance effect.

Chapter 7

Conclusions and Ramifications

Genes are not always expressed at normal levels. Gene expression variation produces different phenotypes. Both under- and overexpression of genes could lead to abnormal phenotypes. Dosage balance theory has been proposed to explain gene dosage effects. Stoichiometric imbalances in macromolecular complexes can be a source of dominant phenotypes. Both gene deletion and overexpresion of a single subunit in a protein complex could be deleterious. The strength of transcriptional co-regulation of interacting pairs is expected to reflect dosage sensitivity. Gene dosage balance theory has been supported by several experimental results from Papp's group: dosage sensitive genes are at least twice more likely to encode proteins involved in complexes than dosage insensitive genes; components of protein complexes counts a significant portion of the genes whose overexpression in wild-type cell is lethal; interacting pairs with high dosage sensitivity are much more co-expressed than the others.

This work studies gene dosage sensitivity by examining the extent of protein structure vulnerability. Soluble protein structure may be more or less vulnerable to water attack depending on their packing quality. We quantify the structure vulnerability by determining the extent of solvent exposure of backbone hydrogen bonds. Within this scheme, local weaknesses in the protein structure may become protected upon complexation, as exposed backbone hydrogen bonds become exogenously dehydrated. Vulnerable structures are thus quantitatively reliant on binding partnerships to maintain their integrity, suggesting that vulnerability may be regarded as a structure-based indicator of gene dosage sensitivity. This observation is validated by establishing the significance of protein vulnerability or structure protection as an organizing factor in temporal phases (yeast) and tissue-based (human) transcriptomes. Specifically, this role was established by examining the degree of co-expressions of a protein with its binding partners in structure-represented interactions. Thus, for each Pfam-filtered binding partnership, the extent of co-expression across metabolic adaptation phases (yeast) or tissue types (human) was found to depend quantitatively on the structure vulnerability of the proteins involved. Hence, vulnerability may be regarded as an organizing factor encoded in the structure of gene products.

Furthermore, as shown in this work, the tight coordination between translation regulation and gene function dictates that extremely vulnerable, and hence "highly needy", proteins are subject to significant levels of post-transcriptional regulation. In human, this extra regulation is achieved through extensive miRNA targeting of genes coding for extremely vulnerable proteins. In yeast, on the other hand, our results imply that such a regulation is likely achieved through sequestration of the extremely vulnerable proteins into aggregated states. Intriguingly, the 85 yeast genes containing extremely vulnerable proteins included the five confirmed yeast prions. Unlike mammalian prions, as suggested by Lindquist, yeast prion protein aggregation provides phenotype plasticity and can confer some selective advantages. These results suggest that extremely vulnerable proteins resorting to aggregation to buffer the deleterious consequences of dosage imbalance. However, a rigorous proof will require a structure-based integration of information drawn from the interactome, transcriptome and post-transcriptional regulome.

If validated, the hypothesis that aggregation circumvents the deleterious effects of dosage imbalance would imply a selective advantage for yeast but this advantage may be significantly reduced in human, where self-templating aggregation traits are well known to be pathogenic. An evolutionary piece of evidence appears to support the hypothesis that aggregation may import a fitness advantage by mitigating dosage imbalance effects. In chapter 6, we established that genes coding for poorly wrapped yet structured proteins tend to be singletons, because the duplicates would be under severe selection pressure as they compete for obligatory binding partners. However, the average number of surviving paralogs for extremely vulnerable proteins is 3.2 in yeast and 5.1 in human. This suggests that the dosage imbalance brought about by gene duplication may be mitigated for extremely vulnerable proteins.

Aggregation may indeed constitute a primitive post-transcriptional regulatory element in unicellular eukaryotes like yeast, it is kept suppressed in human by an additional layer of post-transcriptional regulation. This would lead us to the far-reaching hypothesis that self-templating aggregation is a precursor of miRNA regulation. Proving this hypothesis is not an easy task. Yet, this picture needs to be explored and further validated by assessing the biological forces associated with dosage imbalances in yeast and human.

Bibliography

[1] R. Bals and B. Jany. Identification of disease genes by expression profiling. *European Respiratory Journal*, 18:882–889, 2001.

[2] J. W. Kim and X. W. Wang. Gene expression profiling of preneoplastic liver disease and liver cancer: a new era for improved early detection and treatment of these deadly diseases? *Carcinogenesis*, 24:363–369, 2003.

[3] E. T. Dermitzakis. From gene expression to disease risk. *Nature Genetics*, 40:492–493, 2008.

[4] V. Emilsson, G. Thorleifsson, B. Zhang, and et al. Genetics of gene expression and its effect on disease. *Nature*, 452:423–428, 2008.

[5] R. A. Veitia. Exploring the etiology of haploinsufficiency. *Bioessays*, 24:175–184, 2002.

[6] A. Fernandez and R. S. Berry. Extent of hydrogen-bond protection in folded proteins: a constraint on packing architectures. *Biophysical Journal*, 83:2475–2481, 2002.

[7] A. Fernandez and H. A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl. Acad. Sci. USA*, 100:113–118, 2003.

[8] J. C. Alwine, D. J. Kemp, and G. R. Stark. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proc. Natl. Acad. Sci. USA*, 74:5350–5354, 1977.

[9] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[10] G. C. Na, L. J. Butz, D. G. Bailey, and R. J. Carroll. In vitro collagen fibril assembly in glycerol solution: Evidence for a helical cooperative mechanism involving microfibrils. *Biochemistry*, 25:958–966, 1986.

[11] M. Tassabehji, K. Metcalfe, J. Hurst, G. S. Ashcroft, C. Kielty, C. Wilmot, D. Donnai, A. P. Read, and C. J. P. Jones. An elastin gene mutation producing abnormal tropoelastin and abnormal elastic fibres in a patient with autosomal dominant cutis laxa. *Human Molecular Genetics*, 7:1021–1028, 1998.

[12] P. Toonkool, D. G. Regan, P. W. Kuchel, M. B. Morris, and A. S. Weiss. Thermodynamic and hydrodynamic properties of human tropoelastin analytical ultracentrifuge and pulsed field-gradient spin-echo nmr studies. *Journal of Biological Chemistry*, 276:28042–28050, 2001.

[13] l. Willmitzer and K. G. Wagner. The binding of protamines to dna: role of protamine phosphorylation. *Biophysics of Structure and Mechanism*, 6:95–110, 1980.

[14] C. Cho, W. D. Willis, E. H. Goulding, H. Jung-Ha, Y.-C. Choi, N. B. Hecht, and E. M. Eddy. Haploinsufficiency of protamine-1 or -2 causes infertility in mice. *Nature Genetics*, 28:82–86, 2001.

[15] S. Barbaux, P. Niaudet, M.-C. Gubler, J.-P. Grunfeld, F. Jaubert, F. Kuttenn, C. N. Fekete, N. Souleyreau-Therville, E. Thibaud, M. Fellous, and K. McElreavey. Donor splice-site mutations in wt1 are responsible for frasier syndrome. *Nature Genetics*, 17:467–470, 1997.

[16] J. C. Achermann, M. Ito, P. C. Hindmarsh, and J. L. Jameson. A mutation in the gene encoding steroidogenic factor-1 causes xy sex reversal. *Nature Genetics*, 22:125–126, 1999.

[17] K. Woodward and S. Malcolm. Cns myelination and plp gene dosage. *Pharmacogenomics*, 2:263–272, 2001.

[18] W. Bi, J. M. Deng, Z. Zhang, R. R. Behringer, and B. de Crombrugghe. Sox9 is required for cartilage formation. *Nature Genetics*, 22:85–89, 1999.

[19] J. W. Foster, M. A. Dominguez-Steglich, S. Guioli, C. Kwok, P. A. Weller, M. Stevanovic, J. Weissenbach, S. Mansour, I. D. Young, P. N. Goodfellowparallel, J. D. Brook, and A. J. Schafer. Campomelic dysplasia and autosomal sex reversal caused by mutations in an sry-related gene. *Nature*, 372:525–529, 1994.

[20] B. Huang, S. Wang, Y. Ning, A. N. Lamb, and J. Bartley. Autosomal xx sex reversal caused by duplication of sox9. *American Journal of Medical Genetics*, 87:349–353, 1999.

[21] X. H. Sun. Constitutive expression of the id1 gene impairs mouse b cell development. *Cell*, 79:893–900, 1994.

[22] R. A. Veitia. *The Biological of Genetic Dominance*. Eurekah.com/Landes Bioscience, Georgetwon, Texas, 2006.

[23] B. Papp, C. Pal, and L. D. Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424:194–197, 2003.

[24] J. Yang, R. Lusk, and W. H. Li. Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. USA*, 100:15661–15665, 2003.

[25] S. Teichmann and R. Veitia. Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: a interpretation from a dosage balance perspective. *Genetics*, 167:2121–2125, 2004.

[26] R. A. Veitia. Gene dosage balance: deletions, duplications and dominance. *Trends in Genetics*, 21:33–35, 2005.

[27] A. Goldbeter and D. E. Koshland. An amplified sensitivity arising from covalent modification in biological systems. *Proc. Natl. Acad. Sci. USA*, 78:6840–6844, 1981.

[28] J. Emsley. Very strong hydrogen bonds. *Chemical Society Reviews*, 9:91–124, 1980.

[29] R. C. Weast. *Handbook of Chemistry and Physics*. CRC, 65 edition, 1984.

[30] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA*, 37:205–211, 1951.

[31] L. Pauling and R. B. Corey. The pleated sheet. *Proc. Natl. Acad. Sci. USA*, 37:251–256, 1951.

[32] A. Fernandez and R. Scott. Dehydron: a structurally encoded signal for protein interaction. *Biophysical Journal*, 85:1914–1928, 2003.

[33] A. Fernandez and R. Scott. Adherence of packing defects in soluble proteins. *Physical Review Letters*, 91:018102, 2003.

[34] A. Fernandez, T. R. Sosnick, and A. Colubri. Dynamics of hydrogen bond desolvation in protein folding. *Journal of Molecular Biology*, 321:659–675, 2002.

[35] A. Fernandez and R. S. Berry. Molecular dimension explored in evolution to promote proteomic complexity. *Proc. Natl. Acad. Sci. USA*, 101:13460–13465, 2004.

[36] A. K. Dunker and Z. Obradovic. The protein trinity-linking function and disorder. *Nature Biotechnology*, 19:805–806, 2001.

[37] P. Romero, Z. Obradovic, X. Li, E. C. Garner, Brown C. J., and A. K. Dunker. Sequence complexity of disordered protein. *Proteins: Structure, Function, and Genetics*, 42:38–48, 2001.

[38] L. M. Iakoucheva and A. K. Dunker. Order, disorder, and flexibility: prediction from protein sequence. *Structure*, 11:1316–1317, 2003.

[39] A. Fernandez, R. Scott, and R. S. Berry. The nonconserved wrapping of conserved protein folds reveals a trend toward increasing connectivity in proteomic networks. *Proc. Natl. Acad. Sci. USA*, 101:2823–2827, 2004.

[40] B. Lebrun, R. Romi-Lebrun, M. F. Martin-Eauclaire, A. Yasuda, M. Ishiguro, Y. Oyama, O. Pongs, and T. Nakajima. A four-disulfide-bridged toxin, with high affinity towards voltage-gated k+ channels, isolated from heterometrus spinnifer (scorpionidae) venom. *Biochemical Journal*, 328:321–337, 1997.

[41] P. Savarin, R. Romi-Lebrun, S. Zinn-Justin, B. Lebrun, T. Nakajima, B. Gilquin, and A. Menez. Structural and functional consequences of the presence of a fourth disulfide bridge in the scorpion short toxins: solution structure of the potassium channel inhibitor hstx1. *Protein Science*, 8:2672–2685, 1999.

[42] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101:6062–6067, 2004.

[43] G. G. Roberts and A. P. Hudson. Transcriptome profiling of saccharomyces cerevisiae during a transition from fermentative to glycerol-based respiratory growth reveals extensive metabolic and structural remodeling. *Molecular Genetics and Genomics*, 276:170–186, 2006.

[44] P. M. Kim, L. J. Lu, Y. Xia, and M. B. Gerstein. Relating three-dimensional struc-

tures to protein networks provides evolutionary insights. *Science*, 314:1938–1941, 2006.

[45] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, J. S. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, Sonnhammer E. L., and A. Bateman. The pfam protein families database. *Nucleic Acids Research*, 36:D281–D288, 2008.

[46] R. D. Finn, M. Marshall, and A. Bateman. ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21:410–412, 2005.

[47] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. Wu, C. Orsi, and et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440:637–643, 2006.

[48] H. W. Mewes, D. Frishman, K. F. X. Mayer, M. Munsterkotter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stumpflen. Mips: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Researchearch*, 34:D169–D172, 2006.

[49] N. Eswar, B. John, N. Mirkovic, A. Fiser, V. A. Ilyin, U. Pieper, A. C. Stuart, M. A. Marti-Renom, M. S. Madhusudhan, B. Yerkovich, and A. Sali. Tools for compar-

ative protein structure modeling and analysis. *Nucleic Acids Research*, 31:3375–3380, 2003.

[50] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29:291–325, 2000.

[51] A. Sali and T. L. Blundell. Comparative protein modeling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234:779–815, 1993.

[52] C. Lange, J. H. Nett, B. L. Trumpower, and C. Hunte. Specific roles of protein-phospholipid interactions in the yeast cytochrome bc1 complex structure. *The EMBO Journal*, 20:6591–6600, 2001.

[53] C. Queitsch, T. A. Sangster, and S. Lindquist. Analysis of prion factors in yeast. *Methods in Enzymology*, 351:499–538, 2002.

[54] R. Krishnan and S. L. Lindquist. Structural insights into a yeast prion illuminate nucleation and strain diversity. *Nature*, 435:765–772, 2005.

[55] P. M. Tessier and S. Lindquist. Prion recognition elements govern nucleation, strain specificity and species barriers. *Nature*, 447:556–561, 2007.

[56] M. D. Michelitsch and J. S. Weissman. A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. *Proc. Natl. Acad. Sci. USA*, 97:11910–11915, 2000.

[57] A. Fernandez, J. Kardos, L. R. Scott, Y. Goto, and R. S. Berry. Structural defects and the diagnosis of amyloidogenic propensity. *Proc. Natl. Acad. Sci. USA*, 100:6446–6451, 2003.

[58] V. Ambros. micrornas: tiny regulators with great potential. *Cell*, 107:823–826, 2001.

[59] V. Ambros, B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun, and T. Tuschl. A uniform system for microrna annotation. *RNA*, 9:277–279, 2003.

[60] L. He and G. J. Hannon. Micrornas: small rnas with a big role in gene regulation. *Nature Reviews Genetics*, 5:522–531, 2004.

[61] S. Griffiths-Jones. The microrna registry. *Nucleic Acids Research*, 32:D109–D111, 2004.

[62] E. Wienholds and Plasterk R. H. Microrna function in animal development. *FEBS Lett*, 579:5911–5912, 2005.

[63] P. Xu, S. Y. Vernooy, M. Guo, and B. A. Hay. The drosophila microrna mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology*, 12:790–795, 2003.

[64] K. A. O'Donnell, E. A. Wentzel, K. I. Zeller, Dang C. V., and Mendell J. T. c-myc-regulated micrornas modulate e2f1 expression. *Nature*, 435:839–843, 2005.

[65] L. He, J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435:828–833, 2005.

[66] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small rnas with antisense complementarity to *lin-14*. *Cell*, 75:843–854, 1993.

[67] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, 75:855–862, 1993.

[68] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide *let-7* rna regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403:901–906, 2000.

[69] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed rnas. *Science*, 294:853–858, 2001.

[70] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny rnas with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294:858–862, 2001.

[71] R. C. Lee and V. Ambros. An extensive class of small rnas in caenorhabditis elegans. *Science*, 294:862–864, 2001.

[72] Y. Lee, M. Kim, J. Han, K. Jeon, S. Lee, S. H. Baek, and V. N. Kim. Microrna genes are transcribed by rna polymerase ii. *The EMBO Journal*, 20:4051–4060, 2004.

[73] A. M. Denli, B. B. Tops, R. H. Plasterk, R. F. Ketting, and G. J. Hannon. Processing of primary micrornas by the microprocessor complex. *Nature*, 432:231–235, 2004.

[74] D. S. Schwarz, G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore. Asymmetry in the assembly of the rnai enzyme complex. *Cell*, 115:199–208, 2003.

[75] T. Du and P. D. Zamore. microprimer: the biogenesis and function of microrna. *Development*, 132:4645–4652, 2005.

[76] J. J. Song, J. Liu, N. H. Tolia, J. Schneiderman, S. K. Smith, R. A. Martienssen, G. J. Hannon, and L. Joshua-Tor. The crystal structure of the argonaute2 paz domain reveals an rna binding motif in rnai effector complexes. *Nature Structure Biology*, 10:1026–1032, 2003.

[77] S. Yekta, I. Shih, and D. P. Bartel. Microrna-directed cleavage of hoxb8 mrna. *Science*, 304:594–596, 2004.

[78] C. P. Petersen, M. E. Bordeleau, J. Pelletier, and P. A. Sharp. Short rnas repress translation after initiation in mammalian cells. *Molecular Cell*, 21:533–542, 2006.

[79] R. S. Pillai, S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz. Inhibition of translational initiation by let-7 microrna in human cells. *Science*, 309:1573–1576, 2005.

[80] C. Y. Chu and T. M. Rana. Translation repression in human cells by microrna-induced gene silencing requires rck/p54. *PLoS Biology*, 4:e210, 2006.

[81] S. N. Bhattacharyya, R. Habermacher, U. Martine, E. I. Closs, and W. Filipowicz. Relief of microrna-mediated translational repression in human cells subjected to stress. *Cell*, 125:1111–1124, 2006.

[82] K. Forstemann, Y. Tomari, T. Du, V. V. Vagin, A. M. Denli, D. P. Bratu, C. Klattenhoff, W. E. Theurkauf, and P. D. Zamore. Normal microrna maturation and germline stem cell maintenance requires loquacious, a double-stranded rna-binding domain protein. *PLoS Biology*, 3:e236, 2005.

[83] E. Wienholds, M. J. Koudijs, F. J. van Eeden, E. Cuppen, and Plasterk R. H. The microrna producing enzyme dicer1 is essential for zebrafish development. *Nature Genetics*, 35:217–218, 2003.

[84] R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk. Dicer functions in rna interference and in synthesis of small rna involved in developmental timing in *C. elegans*. *Genes and Development*, 15:2654–2659, 2001.

[85] M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. Macdonald, S. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman, and M. Stoffel. A pancreatic islet-specific microrna regulates insulin secretion. *Nature*, 432:226–230, 2004.

[86] J. F. Abelson, K. Y. Kwan, B. J. O'Roak, D. Y. Baek, A. A. Stillman, T.M. Morgan, C. A. Mathews, D. L. Pauls, M. R. Rasin, M. Gunel, N. R. Davis, A. G. Ercan-Sencicek, D. H. Guez, J. A. Spertus, J. F. Leckman, L. S. Dure, R. Kurlan, H. S. Singer, D. L. Gilbert, A. Farhi, A. Louvi, R. P. Lifton, N. Sestan, and M. W.

State. Sequence variants in slitrk1 are associated with tourette's syndrome. *Science*, 310:317–320, 2005.

[87] G. A. Calin, C. D. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich, and C. M. Croce. Frequent deletions and down-regulation of micro- rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA*, 99:15524–15529, 2002.

[88] A. Cimmino, G. A. Calin, M. Fabbri, M. V. Iorio, M. Ferracin, M. Shimizu, S. E. Wojcik, R. I. Aqeilan, S. Zupo, M. Dono, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini, and C. M. Croce. mir-15 and mir-16 induce apoptosis by targeting bcl2. *Proc. Natl. Acad. Sci. USA*, 102:13944–13949, 2005.

[89] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microrna-target recognition. *PLoS Biology*, 3:e85, 2005.

[90] J. G. Doench and P. A. Sharp. Specificity of microrna selection in translational repression. *Genes and Development*, 18:504–511, 2004.

[91] B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115:787–798, 2003.

[92] N. Bushati and S. M. Cohen. microrna functions. *Annual Review of Cell and Developmental Biology*, 23:175, 2007.

[93] A. J. Giraldez, Y. Mishima, J. Rihel, R. J. Grocock, S. V. Dongen, K. Inoue, A. J.

Enright, and A. F. Schier. Zebrafish mir-430 promotes deadenylation and clearance of maternal mrnas. *Science*, 312:75–79, 2006.

[94] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120:15–20, 2005.

[95] Z. Du, K. Park, H. Yu, Q. Fan, and L. Li. Newly identified prion linked to the chromatin-remodeling factor swi1 in *Saccharomyces cerevisia*. *Nature Genetics*, 40:460–465, 2008.

[96] S. B. Prusiner. Novel proteinaceous infectious particles cause scrapie. *Science*, 216:136–144, 1982.

[97] S. B. Prusiner. Prions. *Proc. Natl. Acad. Sci. USA*, 95:13363–13383, 1998.

[98] S. B. Prusiner. *Prion Biology and Diseases*. Cold Spring Harbor Laboratory Press, New York, 2004.

[99] S. B. Prusiner. Molecular biology of prion diseases. *Science*, 252:1515–1522, 1991.

[100] F. E. Cohen. Protein misfolding and prion diseases. *Journal of Molecular Biology*, 293:313–320, 1999.

[101] L. Calzolai, D. A. Lysek, D. R. Perez, P. Guntert, and K. Wuthrich. Prion protein nmr structures of chickens, turtles, and frogs. *Proc. Natl. Acad. Sci. USA*, 102:651–655, 2005.

[102] F. E. Cohen and S. B. Prusiner. Pathologic conformations of prion proteins. *Annual Review of Biochemistry*, 67:793–819, 1998.

[103] M. L. DeMarco and V. Daggett. From conversion to aggregation: protofibril formation of the prion protein. *Proc. Natl. Acad. Sci. USA*, 101:2293–2298, 2004.

[104] C. Soto and J. Castilla. The controversial protein-only hypothesis of prion propagation. *Nature Medicine*, 10:S63–S67, 2004.

[105] T. Alper, D. A. Haig, and M. C. Clarke. The exceptionally small size of the scrapie agent. *Biochemical and Biophysical Research Communications*, 22:278–284, 1966.

[106] J. S. Griffith. Self-replication and scrapie. *Nature*, 215:1043–1044, 1967.

[107] D. C. Bolton, M. P. McKinley, and S. B. Prusiner. Identification of a protein that purifies with the scrapie prion. *Science*, 218:1309–1311, 1982.

[108] R. Gabizon, M. P. McKinley, D. Groth, and S. B. Prusiner. Immunoaffinity purification and neutralization of scrapie prion infectivity. *Proc. Natl. Acad. Sci. USA*, 85:6617–6621, 1988.

[109] R. B. Wickner. [ure3] as an altered ure2 protein: evidence for a prion analog in *Saccaromyces cerevisiae*. *Science*, 264:566–569, 1994.

[110] S. Lindquist. Mad cows meet psi-chotic yeast: the expansion of the prion hypothesis. *Cell*, 89:495–498, 1997.

[111] D. C. Masision, H. K. Edskes, M. Maddelein, K. L. Taylor, and R. B. Wickner. [ure3] and [psi] are prions of yeast and evidence for new fungal prions. *Current Issues in Molecular Biology*, 2:51–59, 2000.

[112] J. R. Glover, A.S. Kowal, E. C. Schirmer, M.M. Patino, J.J. Liu, and S. Lindquist. Self-seeded fibers formed by sup35, the protein determinant of [psi+], a heritable prion-like factor of s. cerevisiae. *Cell*, 89:811–819, 1997.

[113] S. M. Uptain and S. Lindquist. Prions as protein-based genetic elements. *Annual Review of Microbiology*, 56:703–741, 2002.

[114] R. B. Wickner, H. K. Edskes, E. D. Ross, M. M. Pierce, U. Baxa, A. Brachmann, and F. Shewmaker. Prion genetics: new rules for a new kind of gene. *Annual Review of Genetics*, 38:681–707, 2004.

[115] L. Partridge and N. H. Barton. Natural selection: Evolving evolvability. *Nature*, 407:457–458, 2000.

[116] H. E. Sparrer, A. Santoso, F. C. Szoka, and J. S. Weissman. Evidence for the prion hypothesis: Induction of the yeast [PSI^+] factor by in vitro- converted sup35 protein. *Science*, 289:595–599, 2000.

[117] M. L. Maddelein, S. Dos Reis, S. Duvezin-Caubet, B. Coulary-Salin, and S. J. Saupe. Amyloid aggregates of the het-s prion protein are infectious. *Proc. Natl. Acad. Sci. USA*, 99:7402–7407, 2002.

[118] C. Y. King and R. Diaz-Avalos. Protein-only transmission of three yeast prion strains. *Nature*, 428:319–323, 2004.

[119] N. Sondheimer and S. Lindquist. Rnq1: an epigenetic modifier of protein function in yeast. *Molecular Cell*, 5:163–172, 2000.

[120] A. Santoso, P. Chiien, L. Z. Oscherovich, and J. S. Weissman. Molecular basis of a yeast prion species barrier. *Cell*, 100:277–288, 2000.

[121] P. M. Harrison and M. A. Gerstein. A method to assess compositional bias in biological sequences and its application to prion-like glutamine/asparagine-rich domains in eukaryotic proteomes. *Genome Biology*, 4:R40, 2003.

[122] J. J. Liu and S. Lindquist. Oligopeptide-repeat expansions modulate 'protein-only' inheritance in yeast. *Nature*, 400:573–576, 1999.

[123] L. Z. Osherovich, B. S. Cox, M. F. Tuite, and J. S. Weissman. Dissection and design of yeast prions. *PLoS Biology*, 2:E86, 2004.

[124] B. Cox. [psi], a cytoplasmic suppressor of super-suppression in yeast. *Heredity*, 20:505–521, 1965.

[125] S. W. Liebman and F. Sherman. Extrachromosomal psi+ determinant suppresses nonsense mutations in yeast. *The Journal of Bacteriology*, 139:1068–1071, 1979.

[126] J. Shorter and S. Lindquist. Prions as adaptive conduits of memory and inheritance. *Nature Reviews Genetics*, 6:435–450, 2005.

[127] H. L. True and S. L. Lindquist. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, 407:477–483, 2000.

[128] H. L. True, I. Berlin, and S. L. Lindquist. Epigenetic regulation of translation reveals hidden genetic variation to produce complex traits. *Nature*, 431:184–187, 2004.

[129] I. L. Derkatch, M. E. Bradley, J. Y. Hong, and S. W. Liebman. Genetic and environmental factors affecting the *de novo* appearance of the [*PSI*$^+$] prion in *Saccharomyces cerevisiae*. *Genetics*, 147:507–519, 1997.

[130] I. L. Derkatch, M. E. Bradley, P. Zhou, Y. O. Chernoff, and S. W. Liebman. Prions affect the appearance of other prions: the story of [*PIN*$^+$]. *Cell*, 106:171–182, 2001.

[131] S. Ohno. Evolution by gene duplication., 1970.

[132] M. Long, E. Betran, K. Thornton, and W. Wang. The origin of new genes: glimpses from the young and old. *Nature Review Genetics*, 4:865–875, 2003.

[133] H. Liang, K. Rogale-Plazonic, J. Chen, W. H. Li, and A. Fernandez. Protein underwrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genetics*, 4:e11, 2008.

[134] E. Birney, D. Andrews, M. Caccamo, Y. Chen, and L. Clerk. Ensembl 2006. *Nucleic Acids Research*, 34:D556–D561, 2006.

[135] C. Seoighe and K. H. Wolfe. Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology*, 2:548–554, 1999.

[136] E. Marland, A. Prachumwat, N. Maltsev, Z. Gu, and W. H. Li. Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and e. coli. *Journal of Molecular Evolution*, 59:806–814, 2004.

[137] D. B. Lukatsky, B. E. Shakhnovich, J. Mintseris, and E. J. Shakhnovich. Structural similarity enhances interaction propensity of proteins. *Journal of Molecular Biology*, 365:1596–1606, 2007.

[138] M. Lynch and J. S. Conery. The origins of genome complexity. *Science*, 302:1401–1404, 2003.

[139] H. Liang and W. H. Li. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends in Genetics*, 23:375–378, 2007.

[140] E. I. Boyle, S. Weng, J. Gollub, H. Jin, and D. Botsein. Go::term-finder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20:3710–3715, 2004.