Multi-Scale Theoretical Investigations of Protein Interactions and Evolution

Multi-Scale Theoretical Investigations of Protein Interactions and Evolution

A DISSERTATION PRESENTED
BY
JEONG-MO CHOI
TO
THE DEPARTMENT OF CHEMISTRY AND CHEMICAL BIOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
CHEMISTRY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2016

Thesis advisor: Professor Eugene I. Shakhnovich Jeong-Mo Choi

Multi-Scale Theoretical Investigations of Protein Interactions and Evolution

Abstract

Evolution of biological systems requires players of multiple layers, from atoms and molecules to organisms and populations. Expression of a gene is operated by molecular machineries for transcription, translation, regulation, and maintenance, which work in concert to produce certain macroscopic and observable phenotypes. And when these phenotypes are exposed to selective pressures, more fit phenotypes (with their genes, molecular machineries, and interaction networks) survive in the population. While the relationship of a gene to its cellular consequences is not fully elucidated, it is known that molecular interactions are one of the key factors that determine the relationship.

In this dissertation, we introduce several theoretical tools to study protein interactions and evolution, and show their applications at various scales. The first tool is a coarse-grained scoring function that predicts binding free energy of a protein complex. The scoring function is a simple linear combination of exposed interface areas of different amino acids. In spite of the simplicity, it shows a reasonable predictive power, and predicts correct biochemistry qualitatively. The second is an analytical theory of a spin model on a simple graph, developed by using conventional statistical mechanics. We separated structural and energetic contributions to the free energy of the system, and also obtained a closed form of linear graph contributions. The closed form is applied to predict sequence space free energy of lattice proteins. Lastly, we introduce statistical methods to analyze cellular proteomes and transcriptomes. They can extract global responses of proteomes and transcriptomes to a perturbation, and also responses of specific gene groups. We applied the methods to *E. coli* and yeast systems to address questions on the genotype-phenotype relationship and evolution.

Contents

S. D. G.

Acknowledgments

I would like to express my deepest gratitude to people who had shaped me over my PhD years. First, I am exceptionally thankful to my advisor Prof. Eugene Shakhnovich. He always willingly provides his insight and suggestions on my research, as well as on my career as a scientist. I learned a lot from his vast knowledge and enthusiasm, and this dissertation reflects only a small part of what I learned from him. Also, I am very grateful to my committee members, Prof. Xiaowei Zhuang and Prof. Erel Levine, both of whom provided me enormous constructive feedbacks and professional support. Their comments during my progress meetings were really invaluable, helping me a lot to design and conduct my research.

I would like to thank the current and past members of the Shakhnovich group. Amy Gilson, who shares with me every moment in this group, provided me a lot by her considerate encouragement and intellectual advice. Scott Wylie, Muyoung Heo and Jiabin Xu helped me to settle down well in the group. Adrian Serohijos constantly brought me cheerful encouragement and guidance, and collaboration with him was always inspiring. During my PhD years, I was lucky enough to work with many smart scientists in the group, including Orit Peleg, Shimon Bershtein, Nicolas Chéron, Sanchari Bhattacharyya, Michael Manhart, and Joáo Rodrigues, and discussions with them furnished me with nourishment to grow up as a scientist. I am also grateful to other members, who generously shared their ideas and insights with me: Murat Çetinbaş, Pouria Dasmeh, Jaie Woodard, Ariel Weinberg, Tatyana Kuznetsova, Bharat Adkar, Nicholas Bauer, William Jacobs, Rostam Razban, and Victor Zhao. Also, without skillful administrative support from Judy Morrison and Roel Torres, my PhD years would be filled with a pile of paperwork.

I am thankful to my collaborators outside the group. Eili Klein introduced me to the fascinating world of virology. Assaf Rotem and the colleagues from the Weitz group taught me how to use microfluidics in evolution experiment. Sean Murphy, Dennis Lucarelli, and Andrew Feldman transformed my clumsy code to a beautiful program. It was always enjoyable to work with Kyungtae Kang and Jerome Fox from the Whitesides group, who led cheerful and intellectual conversations. Bogdan Budnik happily helped me with his expertise on mass spectrometry. Georgios Karras from the Lindquist group willingly shared his knowledge and intuitions whenever I was challenged by a totally new field of yeast biology. Smart undergraduates and visiting students always impressed me by their enthusiasm and talents. If allowed to count only my collaborators among many of them, I would confess that I am in debt to Leo Lofranco, Ahmee Marshall-Christensen, and Niamh Durfee.

I am also sincerely grateful to my friends. Although I enjoyed my days at Harvard University, Harvard is also a greatly challenging place to stay. I would have never survived without my awesome friends, to whom I cannot express my gratitude by a few words. If I listed all of their names, this acknowledgement section would never end; instead, I would like to mention my friend groups that

marked my life at Harvard. I shared a lot of happy and tough moments with friends who came to Harvard in 2011 with me, including those who participated in the English Language Program together. Also, my friends in the Department of Chemistry and Chemical Biology, who understand the scholarly and administrative contexts of the challenges I face, encouraged and supported me constantly. The community of Korean chemists in the Boston area also played a similar role in my life. Friends outside the department inspired me by their enthusiasm on their own research, and I gratefully realized that Harvard is one of the best places in the world to interact with scholars from other fields. Alumni groups of Seoul Science High School and KAIST, with whom I could share the memories of my old days, comforted me whenever I missed Korea. The communities of the First Korean Church in Cambridge and H2N prayer group provided emotional and spiritual support, and I would never forget my people here and their ceaseless prayers for me.

Finally, I thank my parents for their continued love, support and prayers from the day when I was born. I would also like to thank my brother, Joon-Mo, for being such a great brother to me. I thank Jessie, to whom I owe a lot intellectually, emotionally, and spiritually, for holding my hands in every challenging situation during my PhD years.

Nothing in biology makes sense except in the light of evolution.

Theodosius Dobzhansky (1973)

1 Introduction

Evolution is a fundamental process that explains a wide range of biological structures and systems, from structures of biomolecules to interactions of populations. To systematically describe evolution, scientists have been using the concept of genotype and phenotype for more than 100 years[137]. A genotype is a part of genetic information in an organism and it determines a phenotype, which is a specific and observable characteristic of the organism. Now, we know that genetic information is stored as nucleic acid sequences, whose transcription, maintenance, and regulation are all done by molecular machineries. On the other hand, phenotypic traits are usually more macroscopic and diverse; for example, according to the *Saccharomyces* Genome Database[26], reported phenotypes of yeast contain cellular morphology, cell cycles, developmental behaviors, interactions with host/environment, and growth/death. Therefore, due to its multi-scale nature from the atomistic to the cellular or higher level, there is no unique way to study the genotype-phenotype relationship, and holistic understanding on various methods and perspectives is required.

One important stepping stone between genetic information and phenotypic traits is cellular networks[8], such as protein-protein interaction networks[133], metabolic networks[120], and gene regulatory networks[128]. Genetic information is translated into protein molecules, one of the major building blocks of an organism, and proteins interact with other molecules under the laws of physical chemistry. Protein interactions, with other types of interactions, constitute cellular networks, which determine phenotypic traits of a cell (Figure 1.1). This dissertation presents several theoretical

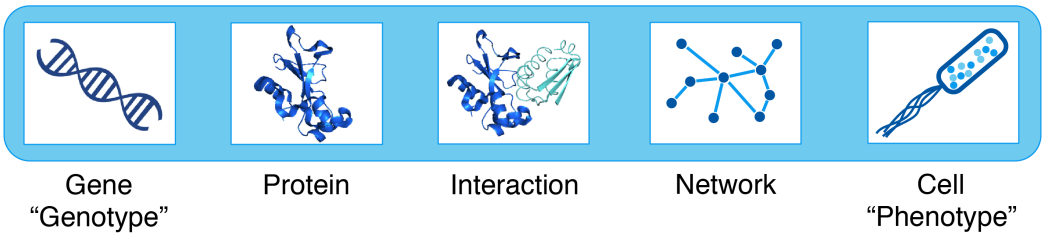


Figure 1.1: Simplified hierarchy of molecular and cellular biology.

approaches to study the interactions and networks, as well as their evolutionary consequences.

The first half of this dissertation introduces two theoretical methods to address biophysical questions about protein-protein interactions and protein evolution: scoring function development for predicting protein binding free energy (chapter 2) and analytical treatment of a spin model on a simple graph using conventional statistical mechanics (chapter 3). The latter theory is applied to predict evolution of protein structures.

In chapter 2, we developed a minimalistic model based on interface areas of protein complexes for predicting protein binding energy. The factor is described by a simple linear combination of buried surface areas according to amino-acid types. Even without structural optimization, our minimalistic model demonstrates a predictive power comparable to more complex methods, making the proposed approach the basis for high throughput applications. Application of the model to a proteomic database shows that receptor-substrate complexes involved in signaling generally have lower affinities than enzyme-inhibitor and antibody-antigen complexes, and they differ by chemical compositions on interfaces. Also, we found that protein complexes with components that come from the same genes generally have lower affinities than complexes formed by proteins from different genes, but in this case the difference originates from different interface areas. This chapter is modified from Choi *et al.*[27] with the publisher's permission.

Chapter 3 shows that there is a direct relationship between a network system's topology and its thermodynamic stability. We separated topological and enthalpic contributions to free energy of a spin model on a simple graph and found that considering the topology is sufficient to qualitatively predict its stability at high temperature, even when the energetics are not fully known. This result was applied to the metal lattice system, and we found that it partially explains why point defects are more stable than high-dimensional defects. Given the energetics, we can even quantitatively compare free energies of systems on different graph structures via a closed form of linear graph contributions. The closed form is applied to predict sequence space free energy of lattice proteins.

In the second half of this dissertation, we focus on systems-level analysis of proteomes and transcriptomes experimentally extracted from *E. coli* and yeast cells, as these are the resulting products of molecular interactions. Chapter 4 introduces statistical tools to analyze cellular proteome and transcriptome, which are here defined as relative abundance data of proteins and mRNAs, respectively. Although chapter 4 itself contains some examples on how to use the tools, chapter 5 and chapter 6 provide more applications of the developed tools to answer some interesting biological questions.

In chapter 4, we establish a quantitative relationship between the global effect of mutations on the *E. coli* proteome and bacterial fitness. We created *E. coli* strains with specific destabilizing mutations in the chromosomal *folA* gene encoding dihydrofolate reductase (DHFR) and quantified the ensuing changes in the abundances of 2,000+ *E. coli* proteins in mutant strains by a mass spectrometry-based protein identification method, which has been successfully used to compare samples in different conditions[3,80]. mRNA abundances in the same *E. coli* strains were also quantified. The proteomic effects of mutations in DHFR are quantitatively linked to phenotype: the standard deviations of the distributions of logarithms of relative-to-wildtype protein abundances anti-correlate with bacterial growth rates. Proteomes hierarchically cluster first by media conditions, and within each condition, by the severity of the perturbation to DHFR function. These results highlight the importance of a systems-level layer in the genotype-phenotype relationship. This chapter is modified from Bershtein *et al.*[11] under the CC BY copyright license.

The first application of the methods introduced in chapter 4 is horizontal gene transfer, which plays a central role in bacterial evolution, yet the molecular and cellular constraints on functional integration of the foreign genes are poorly understood. In chapter 5, we perform inter-species replacement of the chromosomal *folA* gene with orthologs from 35 other mesophilic bacteria. The orthologous inter-species replacements caused a marked drop (in the range 10-90 %) in bacterial growth rate. Serial propagation of the orthologous strains for approximately 600 generations dramatically improved growth rates. By using the statistical tools developed in chapter 4, we could compare proteomes from different strains, and found the following: by apparently distinguishing between self and non-self proteins, protein homeostasis imposes an immediate and global barrier to the functional integration of foreign genes by decreasing the intracellular abundance of their products. Once this barrier is alleviated, more fine-tuned evolution occurs to adjust the function/expression of the transferred proteins to the constraints imposed by the intracellular environment of the host organism. This chapter is modified from Bershtein *et al.*[15] under the CC BY copyright license.

Chapter 6 presents the second application, which compares proteomes of two different yeast strains when they respond to inhibition of an essential heat shock protein Hsp90. The correlations between different proteomes show that the two strains have significantly different proteome pat-

terns, regardless of the Hsp90 inhibition. We employed two different grouping methods to dissect the proteomes. First, genes are grouped by their Gene Ontology (GO) terms, and this grouping provides information on gene groups that are significantly more expressed in one of the two strains. From this information, we found that large "modules" (such as mitochondria- and vacuole-related genes) drives the proteome-level difference, suggesting that divergent evolution between the two strains adjusted modules on cellular networks, not individual genes or proteins. Also, we showed that this proteome-level difference is reduced when the two strains are treated by the inhibitor drug. The second way is grouping by phenotypes, which suggests potential experiments that discriminate the two strains upon the inhibition by their phenotypes.

The theoretical approaches introduced in this dissertation deepen our understandings on interaction networks in a cell. Protein binding affinity, which determines the structure of a protein-protein interaction network, can be estimated with a relatively low cost by the scoring function shown in chapter 2, which also provides biophysical insight that the interface area generated by association is crucial in determining binding affinity. The theory presented in chapter 3 has a direct application to the protein evolution problem, and it can be applied to interaction networks in a cell, considering the model's generality. As shown in chapters 4, 5, and 6, analysis of proteome-level differences of different cells provides a hint for an underlying network structure arising from direct and indirect interactions of cellular proteins, and it was applied to directly compare the proteomic fingerprints before and after evolution, which allows a mechanistic study on adaptation.

2

Minimalistic Predictor of Protein Binding Free Energy

2.1 Background

Protein-protein interactions (PPIs), such as those involved in signaling pathways and enzyme-inhibitor interactions, play a fundamental role in biological function and evolution. Thus significant biological insight can be gained by estimating the strength of PPIs in the whole interactome[32,37,18]. Various methods have been developed to predict binding affinities accurately and quickly, either based on physical force fields[72,57,52] or molecular dynamics[60,108]. Although several recent methods were reported to show high correlation to experiment[148,129,150], it is still challenging to estimate the precise binding energy of a specific protein complex from first principles, especially at a relatively low computational cost.

To accurately predict binding energy of two proteins, we need to identify and quantify physical factors that govern binding energy. It has been known[68,65] that major contributors to binding energy are the interface area[45,66], hot spots[100,24], conformational changes[129,56], allosteric interactions of small molecules[99,75], and even non-interacting surfaces[67]. Among them, the interface area serves as a primary factor and also it provides a playground for other factors. Usually the interface area is associated with the magnitude of hydrophobic interactions[65], but its detailed composition is also

important in determining binding affinity[66].

In this chapter, we will develop a scoring function that predicts protein binding free energy solely based on the interface area, using the mean-field approach. Here, we assume that the structure of a protein complex is given (we may or may not know about the structures of its components). Also, we will check if this simple model has a reasonable predictive power, and if it can reveal correct biochemistry.

2.2 Interface Area and Binding Free Energy

To develop a model solely based on the interface area, we use changes in accessible surface areas (ΔASAs) during association for the ingredients of the scoring function. The definition of ΔASA is simply the difference of accessible surface areas between a protein complex and its unbound components:

$$\Delta\mathrm{ASA} = \sum_i \mathrm{ASA}_{\mathrm{component}\,i} - \mathrm{ASA}_{\mathrm{complex}}. \tag{2.1}$$

However, the practical calculation of ΔASA depends on the situation, especially on availability of structures. A complex and its components can have their own different conformations. If all structures of those conformations are known, we use the (arithmetic or Boltzmann-weighted) average values for the terms in equation 2.1:

$$\Delta\mathrm{ASA} = \sum_i \langle \mathrm{ASA}_{\mathrm{component}\,i} \rangle - \langle \mathrm{ASA}_{\mathrm{complex}} \rangle. \tag{2.2}$$

This is an ideal scenario, which rarely happens. Usually, we have a single conformation (or a few) for each of a complex and its constituent components. Then the ASA can be directly calculated from each structure, using equation 2.1. A worse situation is that we do not have structural information of component conformations. In this case, we extract each component structure from the structure of a whole complex, and apply equation 2.1.

It has been reported that ΔASA provides a major contribution to binding free energy (Figure 2.1a), especially when binding is not accompanied by major conformational changes[66]. Also, they found that this correlation disappears if they include the protein complexes with large conformational changes during association. Note that this result came from the data set that contains the structures of both complex and components. Pretending the "worse" case that we do not have the component structures, we checked if this ΔASA, calculated from the structure of a complex only, can predict the binding free energy. Interestingly, this "worse" ΔASA reports a stronger correla-

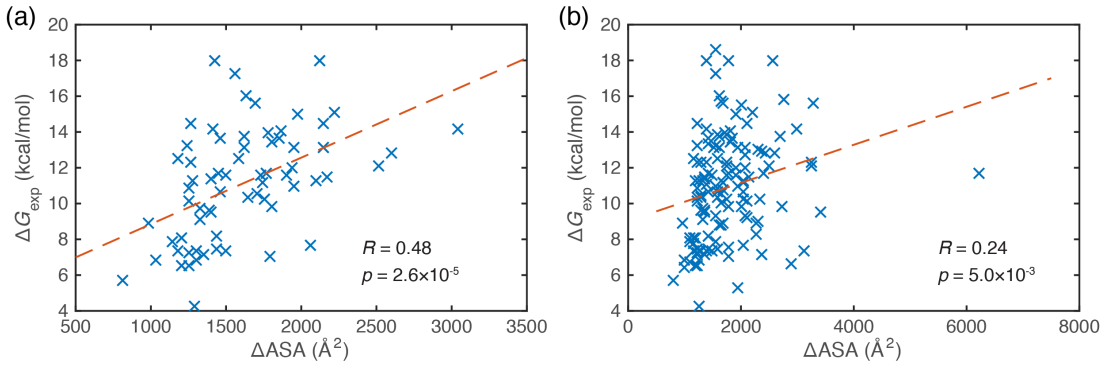


Figure 2.1: Correlations between interface areas (ΔASAs) and experimental dissociation free energies. Red lines indicate the regression lines. (a) Component ASAs are calculated from experimentally determined component structures. Shown data are for 69 protein complexes that do not accompany with large conformational changes upon association, and if other complexes are included, the trend totally disappears. Data from Kastritis *et al*[66]. (b) Component ASAs are calculated from the structures extracted from complex structures with no structure relaxation. Shown data are for all 139 protein complexes (some of which suffer large conformational changes) used in this work.

tion when we include the complexes with large conformational changes (Figure 2.1b), implying that the area generated by association is more relevant to binding affinity than the original ASAs of the components.

2.3 Model Construction

Inspired by the previous findings, we decided to construct a scoring function based on ΔASA values. The data set was constructed from 139 complexes extracted from the structure-affinity benchmark, which provides the structures and binding affinities of protein complexes, as previously described[66]. Even though the benchmark provides structural information of complex components, we only used the structures of complexes, to reach broader applicability of the model. In order to compute atomic ASAs, We employed the Shrake-Rupley algorithm[115] implemented in ASA.PY* with the probe radius of 1.4 angstroms. Here we neglected hydrogen atoms.

Based on the fact that the contribution to solvation depends on the amino acid types of residues[100], we grouped atomic ΔASAs by their amino acid types. Since ΔASAs of the backbone atoms are known to be independently crucial in stabilizing a protein structure[17], we separated the backbone atoms. Consequently, we have 20 side-chain plus one backbone ΔASA terms for each protein complex.

*http://boscoh.com/protein/asapy.html

Subset 1	1ACB, 1AK4, 1AKJ, 1AY7, 1BRS, 1BUH, 1BVK, 1CBW, 1E4K, 1E6E, 1EAW, 1EMV, 1EZU, 1FC2, 1HE8, 1I4D, 1J2J, 1KXP, 1NB5, 1NSN, 1OPH, 1PVH, 1QA9, 1RLB, 1YVB, 1ZHI, 2ABZ, 2B4J, 2BTF, 2FJU, 2I25, 2OUL, 2OZA, 2PCC, 2VIR
Subset 2	1AHW, 1AVX, 1AVZ, 1B6C, 1BJ1, 1DFJ, 1EFN, 1F6M, 1FQJ, 1FSK, 1GRN, 1JIW, 1JTG, 1JWH, 1K5D, 1KAC, 1MAH, 1MLC, 1MQ8, 1P2C, 1R0R, 1T6B, 1US7, 1VFB, 1WEJ, 1XD3, 1ZLI, 2AJF, 2B42, 2HQS, 2I9B, 2OOB, 2SIC, 2VDB, 3BP8
Subset 3	1ATN, 1FFW, 1FLE, 1GLA, 1GPW, 1GXD, 1HIA, 1I2M, 1IBR, 1IJK, 1JPS, 1KTZ, 1LFD, 1NCA, 1NVU, 1NW9, 1PPE, 1S1Q, 1UUG, 1XQS, 1XU1, 1Z0K, 2A9K, 2C0L, 2HLE, 2HRK, 2J0T, 2MTA, 2NYZ, 2OOR, 2PCB, 2UUY, 3BZD, 3CPH, 3SGB
Subset 4	1A2K, 1BVN, 1DQJ, 1E6J, 1E96, 1EER, 1EWY, 1F34, 1GCQ, 1GL1, 1H1V, 1H9D, 1HCF, 1IB1, 1IQD, 1JMO, 1KKL, 1KLU, 1KXQ, 1M10, 1OC0, 1PXV, 1R6Q, 1RV6, 1WQ1, 2AQ3, 2JEL, 2O3B, 2PTC, 2SNI, 2TGP, 2VIS, 2WPT, 4CPA

Table 2.1: Four subsets randomly constructed from the structure-affinity benchmark[66]. The interfaces are same as defined in Kastritis *et al*[66].

Following the four-fold cross-validation method[30], we divided the whole data set into four subsets of equal size (Table 2.1). Then we ran four rounds of training and testing, in each of which the union of three sets serves as a training set and the remaining subset is used as a test set. In each round, we ran a linear regression using each possible combination of 21 terms (20 side-chain ΔASA terms and 1 backbone ΔASA term), that is, $(2^{21} - 1)$ combinations were investigated for each round. For each combination, we checked if the combination is "relevant," based on the correlation coefficients of the regression. A combination is "relevant" if removing any of its constituent elements leads to a statistically significant decrease in predictive power compared to random-number terms. The details of this procedure are explained below.

Let L_0 a certain combination of terms. For example, we may choose

$$L_0 = \{\Delta\text{ASA}_{\text{Ala}}, \Delta\text{ASA}_{\text{Cys}}, \Delta\text{ASA}_{\text{Tyr}}\}. \tag{2.3}$$

For the given combination L_0, we first run the linear regression on the training set, where we assume a linear combination of the elements in L_0 as a regression model. Employing the previous example

in equation 2.3, the regression model for L_0 is

$$\Delta G_{\text{bind}} = C + w_{\text{Ala}} \times \Delta\text{ASA}_{\text{Ala}} + w_{\text{Cys}} \times \Delta\text{ASA}_{\text{Cys}} + w_{\text{Tyr}} \times \Delta\text{ASA}_{\text{Tyr}}. \tag{2.4}$$

The training set is used to determine the weights $\{w_i\}$ and constant C. The regression analysis also provides Pearson's correlation coefficient R^2, which will be denoted by $(R^2)_{0,\text{tr}}$. Using the determined weights, we also compute the R^2 value of the test set, which will be denoted by $(R^2)_{0,\text{test}}$ here.

Now, every element in L_0 is tested to check if the term is statistically significant. To that end, we remove element i from L_0 to construct a new combination L_i, and do the same regression procedure for L_i to get $(R^2)_{i,\text{tr}}$ and $(R^2)_{i,\text{test}}$. Now element i is statistically significant in L_0 only if:

1. $(R^2)_{i,\text{tr}} < (R^2)_{0,\text{tr}}$,
2. $(R^2)_{i,\text{test}} < (R^2)_{0,\text{test}}$, and
3. $\text{distance} \equiv \sqrt{[(R^2)_{i,\text{tr}} - (R^2)_{0,\text{tr}}]^2 + [(R^2)_{i,\text{test}} - (R^2)_{0,\text{test}}]^2} > c(L_0)$.

The first two conditions require L_0 to be more predictive than L_i (increase in the predictive power) on both $(R^2)_{\text{tr}}$ and $(R^2)_{\text{test}}$ axes. However, it is possible that this increase in predictive power is just by chance. Hence we added the third condition, which requires the increase to be significant compared to random numbers, which determines the criterion $c(L_0)$.

To determine $c(L_0)$, we drew random numbers from the uniform distribution on the interval (0, 1). We generated a random number set for each term (hence 21 random number sets in total), and determined R^2 values following the description above. We repeated this procedure 5,000 times and got the distribution of "distances" (Figure 2.2). We collected the data satisfying the conditions 1 and 2, and determined the standard deviation σ of its distribution. (Since a distance is nonnegative, we symmetrized the distribution to define σ.) Finally, we set $c(L_0) = 2.5\sigma$. Note that $c(L_0)$ depends only on the number of elements in L_0 (Table 2.2).

2.4 Final Model and its Predictive Power

From each round of different training and test sets, we collected four groups of "relevant" combinations, and found a common combination, which consists of only three ΔASA terms, $\Delta\text{ASA}_{\text{Tyr}}$, $\Delta\text{ASA}_{\text{Ser}}$, and $\Delta\text{ASA}_{\text{Cys}}$ (Table 2.3):

$$\Delta G_{\text{bind}} = C + w_{\text{Tyr}} \times \Delta\text{ASA}_{\text{Tyr}} + w_{\text{Ser}} \times \Delta\text{ASA}_{\text{Ser}} + w_{\text{Cys}} \times \Delta\text{ASA}_{\text{Cys}}, \tag{2.5}$$

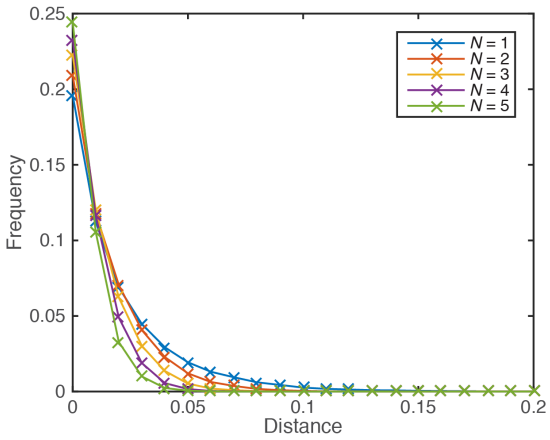


Figure 2.2: Distance distributions from random number sets. N indicates the number of terms to construct a regression model. As N increases, the standard deviation of the distribution decreases.

N	σ	N	σ	N	σ	N	σ
1	0.0255	5	0.0117	9	0.0067	13	0.0042
2	0.0223	6	0.0100	10	0.0059	14	0.0040
3	0.0182	7	0.0087	11	0.0052	15	0.0018
4	0.0147	8	0.0075	12	0.0048	16	0.0016

Table 2.2: Standard deviations (SDs) of R^2 distributions generated from random number sets. N is the number of terms in a null-model combination, and σ indicates the standard deviation. These SDs numerically define $c(L_\circ)$, a criterion used to check if a certain increase in predictive power is statistically significant. For $N > 16$, there is no data point satisfying the conditions 1 and 2 simultaneously.

First round		Second round		Third round		Fourth round	
Comb.	Dist.	Comb.	Dist.	Comb.	Dist.	Comb.	Dist.
C	0.0046	S	0.0300	R	0.0001	R	0.0006
S	0.0136	Y	0.0533	C	0.0034	H	0.0031
Y	0.0248	CY	0.1058	S	0.0133	O	0.0064
CS	0.0337	SY	0.1498	Y	0.0197	CS	0.0295
CY	0.0516	CSY	0.2072	CS	0.0317	NSY	0.0674
SY	0.0519			CY	0.0587	CSY	0.0884
NSY	0.0729			CSY	0.1016	CHSY	0.0978
CSY	0.0878			CHSY	0.1282		
CNSY	0.1103						
CKSY	0.1141						
CNSWY	0.1314						
CGNSY	0.1341						
CGNSWY	0.1506						

Table 2.3: Relevant combinations of ΔASA terms and their corresponding distances for each round, in which one subset plays a role of a test set, while the union of the remaining three subsets is a training set. For example, in the first round, the union of subsets 2, 3, and 4 is a training set, while subset 1 is a test set. Each string indicates a combination of ΔASA terms of the symbolized amino acids, where the 1-letter amino acid notation is used and letter O refers to the backbone atoms (*e.g.* string "CKSO" means $\{\Delta ASA_{Cys}, \Delta ASA_{Lys}, \Delta ASA_{Ser}, \Delta ASA_{backbone}\}$). The common combination found in all rounds is underlined.

where $C = -8.5$ kcal/mol, $w_{Tyr} = -0.0086$ kcal/mol/Å^2, $w_{Ser} = -0.014$ kcal/mol/Å^2, and $w_{Cys} = -0.032$ kcal/mol/Å^2 (these numbers have been determined by using the entire data set). We also checked if inclusion of the μ-potential, a contact-based potential that is capable of discriminating real protein complexes from decoys (Appendix A), improves the accuracy or not, but it turned out that it does not.

The final model shows Pearson's correlation coefficient R of 0.48 between predicted and observed binding affinities for the whole set of 139 protein complexes, which is unexpectedly high when compared to the known methods, especially considering its simple nature (Figure 2.3 and Table 2.4). The root mean squared error (RMSE) is 2.6 kcal/mol, comparable to the RMSE of 2.25 kcal/mol from the ZAPP calculation[135]. Equivalent error estimates for GA-PLS and BioQSAR were reported to be 0.8-1.5 kcal/mol[129,150]. It should be noted that even though the whole set contains protein complexes with large conformational changes during docking, the current model still shows a desirable performance without considering structural changes during association.

In the following two sections, we show the applicability of the simple model to discern different

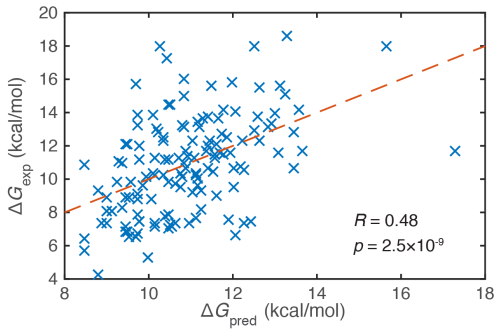


Figure 2.3: Correlation between predicted and experimental dissociation free energies. The red line indicate the regression line.

Method	R	Feature
MARS[97]	0.52	Machine learning
ZAPP[135]	0.63	Regression with 9 terms
GA-PLS[129]	0.83	Consideration of allostery
BIOQSAR[150]	0.82-0.88	Machine learning
SPA-PP[144]	0.39	Statistical potential
ROSETTADOCK[144]	0.42	Regression with 11 terms
DFIRE[97]	0.35	Statistical potential
PMF[97]	0.37	Modified statistical potential
Interface area	0.24	Regression with ΔASA_{total} only (Figure 2.1b)
This work	0.48	

Table 2.4: Comparison of Pearson's correlation coefficient R values among different methods, where the same benchmark[66] is used.

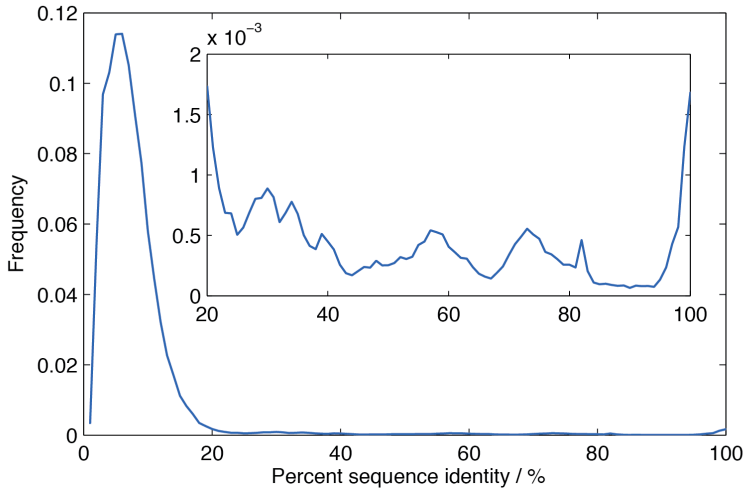


Figure 2.4: A distribution of percent sequence identities computed by pairwise comparison of 2,531 non-redundant protein heterodimers extracted from PDBePISA[73]. The inset amplifies the region between 20 % and 100 %, showing the redundancy peak around 100 %.

types of protein complexes, according to their biochemical functions and genetic origins. We collected protein complex structures from the PDBePISA database[73] to form non-redundant groups of different protein types. To construct a non-redundant group of protein complexes, we considered two protein complexes are equivalent if the percent sequence identity between the two is higher than 90 %. The distribution of percent identities of 2,531 protein heterodimers extracted from PDBePISA gives a minimum at 90 %, which can be considered as the point where the redundancy peak starts (Figure 2.4). Note that this is also consistent with one of the criteria used in a previous study[73].

2.5 Biological Functions and Binding Free Energy Distributions

First, we checked if the current model can discriminate obligatory and transient complexes, and if it can capture more subtle differences. We extracted representative non-redundant groups of various obligatory and transient protein complexes from the PDBePISA database[73]: antibody light-heavy chain recognition (LH, 367 complexes), antibody-antigen recognition (AA, 157 complexes), enzyme-inhibitor recognition (EI, 123 complexes), and receptor-substrate recognition (RS, 210 complexes). LH recognition is considered obligatory, while other three interactions are all known as transient[68,102,61,87]. It has been known that obligatory interactions are generally tighter than the tran-

Obligatory/transient	HL-AA	HL-EI	HL-RS
Current model ΔG_{pred}	2.0×10^{-59}	9.2×10^{-50}	9.9×10^{-87}
ZAPP ΔG_{pred}	4.6×10^{-74}	5.1×10^{-47}	1.9×10^{-86}
Contact number	7.1×10^{-78}	3.2×10^{-58}	2.4×10^{-77}
Among transient	AA-EI	AA-RS	EI-RS
Current model ΔG_{pred}	0.99	2.3×10^{-6}	7.4×10^{-6}
ZAPP ΔG_{pred}	4.0×10^{-5}	7.2×10^{-7}	2.2×10^{-11}
Contact number	0.013	3.0×10^{-3}	0.061

Table 2.5: p-values from the two-sample Kolmogorov-Smirnov test on each pair of free energy or contact number distributions. The abbreviations follow those in Figure 2.5.

sient interactions[68,61], but we want to quantitatively analyze the differences in binding energies.

Our model is used to calculate the binding energy distributions for all four groups (Figure 2.5a). Note that the distributions generally conform to the previously reported distribution of protein binding energies[147]. It is shown that obligatory interactions are stronger than transient ones as expected. However, among the three transient complexes, the RS complexes turned out to have significantly weaker binding than AA and EI complexes. This quantitative difference can be explained by different natures of AA/EI and RS interactions. The functions of AA/EI binding are mostly to bind to their partners as tightly as possible. In contrast, receptor-substrate binding should show a weaker interaction because binding partners should easily associate or dissociate to regulate activity[61]. Similar results were attained from more accurate ZAPP calculation (Figure 2.5b). The corresponding Kolmogorov-Smirnov p-values are given in Table 2.5.

To check whether the significant difference between the types of interactions is merely due to the differences in sizes of interfaces, we scored the binding affinity again using a simpler scoring function based solely on the contact count that reports the interface size, which is a different metric from ΔASA. A pair of atoms are considered to be in contact when the distance between them is smaller than 10 angstroms (comparable to the Debye length for a physiological concentration). The result (Figure 2.5c) shows two features: (1) HL complexes (obligatory) have significantly more contacts than the other three types of complexes (transient). (2) The difference among the other three is relatively insignificant, even though the Kolmogorov-Smirnov test shows that EI complexes have a marginally different distribution of the contact counts from AA and RS complexes. The first feature is shown in the predicted binding energy distributions, which means that the number of interface contacts essentially differentiates HL complexes from the other three types of protein complexes. To make extremely strong binders, the interface areas have to be increased, because mere

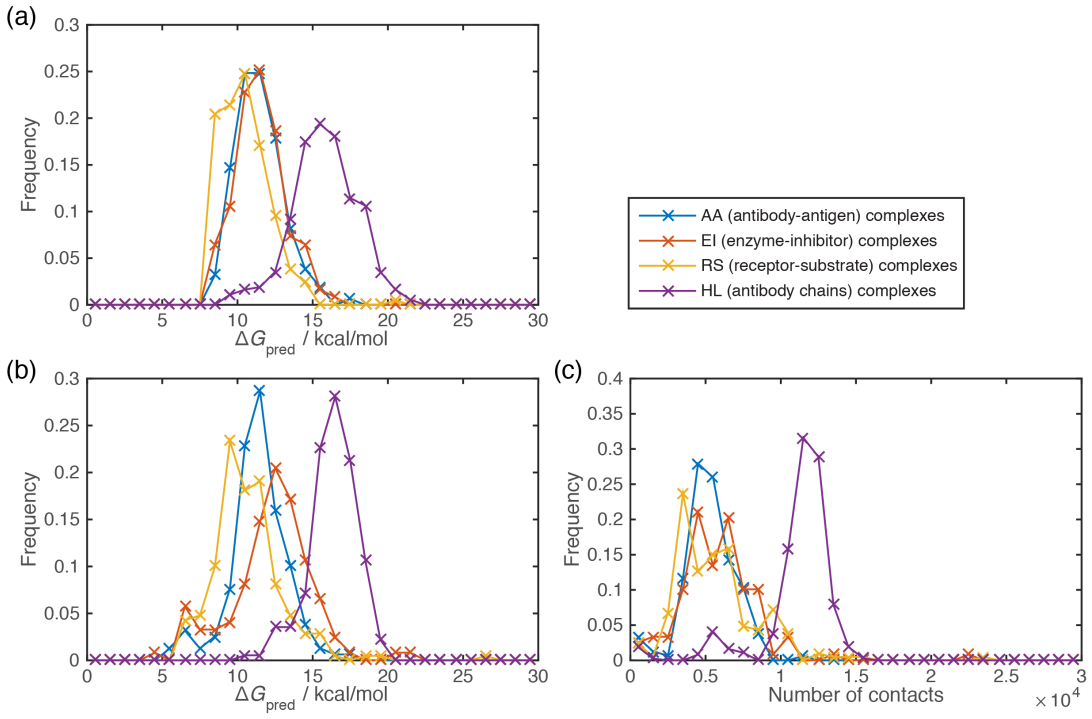


Figure 2.5: Distributions of predicted dissociation free energies and contact numbers in various biochemical contexts, represented by different colors. (a) Distributions of free energies predicted by the current model (equation 2.5). Bin size = 1 kcal/mol. (b) Distributions of free energies predicted by ZAPP [135]. Bin size = 1 kcal/mol. (c) Distributions of contact counts, which quantitatively measure the interface sizes. Bin size = 1,000. Note that this simple model can reproduce most characteristics captured by ZAPP, a more complicated model. The obligatory (HL) interactions generally have more contacts than the transient interactions, implying that the free energy difference between the two groups is due to the interface size. However, the contact count cannot explain the difference between transient interactions, which suggests that this difference originates from more subtle and qualitative differences (*e.g.* amino acid composition) of the interfaces. The Kolmogorov-Smirnov p-values are summarized in Table 2.5.

modification of chemical interactions cannot achieve this objective. However, the second feature, that AA, EI, and RS complexes are almost indistinguishable in contact number distributions, is inconsistent with previous binding energy calculations. This fact suggests that EI and AA complexes are predicted to bind more tightly than RS complexes, yet their contact interfaces are almost the same size. Also, it implies that AA and EI complexes have evolved to find stronger binding amino acids on their interfaces, leaving the total interface areas unchanged.

2.6 Role of Genetic Origin

Another interesting comparison is between complexes of protein domains coming from the same gene and those from different genes. (To our knowledge, there has been no previous study to compare their binding energy distributions quantitatively.) We focused on the dimers from *Homo sapiens*, since they comprise the largest set in the PDB. From the *H. sapiens* data set from PDBePISA[73], we collected two different non-redundant groups: intra-genic (1,213 complexes) and inter-genic (270 complexes). An interaction between two components of a dimer is considered *intra-genic* when they are from the same open reading frame, which is tagged by a unique UniProt ID[31]. Otherwise, the interaction is *inter-genic*.

The predicted binding energy distributions of these groups reveal that inter-genic interactions are significantly stronger than intra-genic ones (Figure 2.6a). The qualitative trend was also reproduced by ZAPP calculation (Figure 2.6b). This is presumably due to an entropic cost of finding their binding partners, which must be compensated by stronger binding. In other words, intra-genic complexes have their components in spatial proximity when synthesized, while inter-genic complexes need to search the subunits to be assembled. This requirement for a stronger binding affinity between inter-genic domains could be relaxed by active transport, but to the authors' knowledge, there has been no systematic study on this topic.

We also checked the interface sizes by counting interface contacts (as described in the previous section), and found that the docking difference between intra-genic and inter-genic complexes appears to be mainly driven by quantitative differences in interface size. Intra-genic complexes have, on average, a smaller number of contacts than inter-genic complexes (Figure 2.6c), implying that evolution has found that it is more efficient to synthesize small protein complexes from a single gene, while making large complexes by assembling two independent units, potentially from different origins.

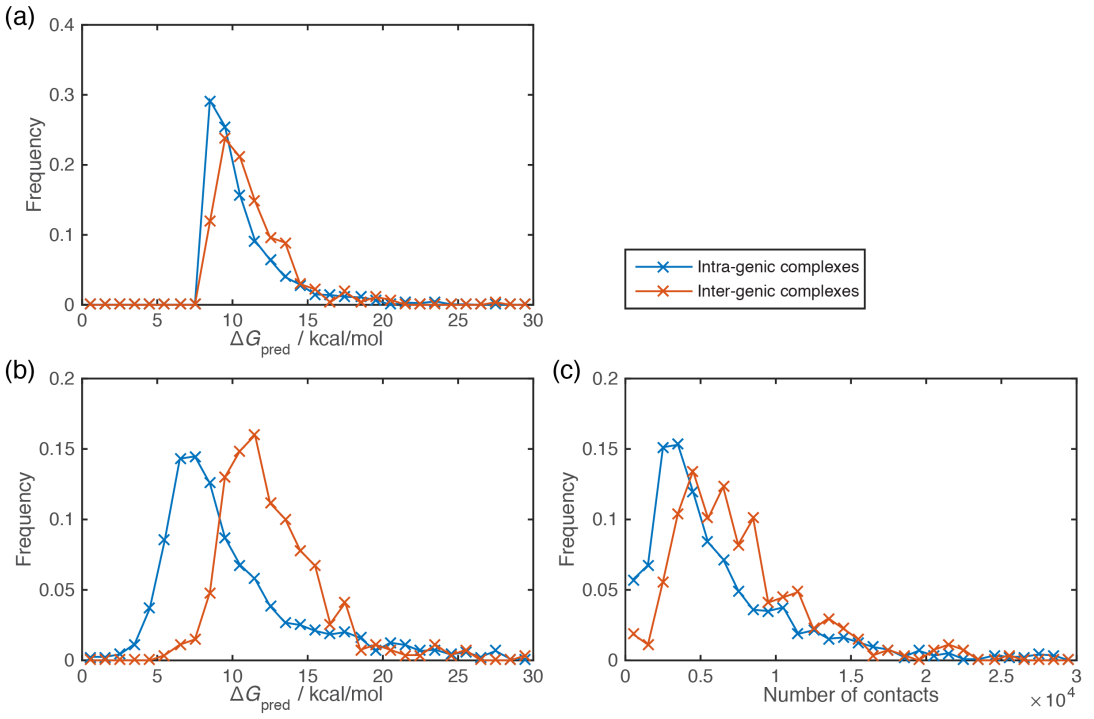


Figure 2.6: Distributions of predicted dissociation free energies and contact numbers of protein complexes with different types of genetic origins, represented by different colors. See text for the definitions of *intra-genic* and *inter-genic* complexes. (a) Distributions of free energies predicted by the current model (equation 2.5). Bin size = 1 kcal/mol. (b) Distributions of free energies predicted by ZAPP [135]. Bin size = 1 kcal/mol. (c) Distributions of contact counts, which quantitatively measure the interface sizes. Bin size = 1,000. Note that this simple model can reproduce most characteristics captured by ZAPP, a more complicated model. The inter-genic complexes generally have more contacts than the intra-genic interactions, implying that the free energy difference between the two groups is due to the interface size. The two-sample Kolmogorov-Smirnov test gives the following p-values, confirming the differences: (a) 9.2×10^{-12}, (b) 2.6×10^{-47}, and (c) 2.0×10^{-12}.

2.7 Conclusion

In this chapter, we have shown that the interface area *after association* is capable of predicting binding energy of protein complexes, even when a large conformational change occurs during association. Separation of interface areas according to amino acid types can increase the predictive power, which is comparable to that of traditional methods. Also the simple model has shown its ability to reveal important aspects of chemistry and biology of PPIs on the whole proteome scale. We expect the simple predictor of PPI binding affinity presented in this study to be used in future proteomics studies of physics and evolution of protein complexes, such as more realistic simulations of mass action dynamics in PPI networks of a variety of organisms[93].

3

Relationship between Graph Topology and System Stability

3.1 Background

Over the last thirty years, graph theory has been applied to study various networks, including protein interaction networks, neural networks, and the World Wide Web[139,1,104,40]. Especially the interplay between network structure and dynamics has attracted huge attention[41], while the equilibrium characteristics of networks, which may deepen our understanding of network phenomena, have not yet been studied thoroughly[49].

The free energy of a system on a given graph architecture is one important equilibrium characteristic. However, to our knowledge, an analytical relationship between a graph's topology and the stability of a system on the graph has not previously been reported, despite the utility such a relationship would provide in discriminating between topologies based on their stability or the insight into graph dynamics it could provide. For example, there will be a general thermodynamic tendency to prefer stable over unstable topologies even when a system is out of equilibrium.

In this chapter, we will consider a spin model on a graph, which has a wide range of applications from metal alloy behaviors to social network phenomena. Based on the model, we will study the general relationship between a system's graph topology and its free energy, without considering de-

tailed energetics, and determine the conditions where the relationship holds. Also, we will develop an approximate theory to calculate the free energy when we are given full information on interaction energies.

3.2 Hamiltonian and Free Energy

Consider a *simple graph* of N nodes (with no self-loops and no multiple links, and which may be either connected or disconnected). The graph connectivity is described by the adjacency matrix $\mathcal{A}$, whose element $\mathcal{A}_{ij} = 1$ when there is a link between nodes i and j, and $\mathcal{A}_{ij} = 0$ otherwise. Each node is in one of $\mathcal{M}$ possible states. The Hamiltonian, $\mathcal{H}$, is defined as the summation of energetic contributions of all links, each of whose energy is determined by the states of its two terminal nodes. Note that orphan nodes do not contribute energetically by definition.

The Hamiltonian can be written as

$$\mathcal{H} = \frac{1}{2}\sum_{i,j}^{N,N} \mathcal{A}_{ij} E_{s(i)s(j)}, \tag{3.1}$$

where $\mathcal{A}$ is the adjacency matrix of the given graph, E is the energy matrix, and $s(i)$ is the state of node i. The partition function $Z(\beta) = \sum_{\{s\}} \exp(-\beta\mathcal{H})$ over all possible state configurations can be expanded as follows:

$$Z(\beta) = \sum_{\{s\}} 1 - \beta \sum_{\{s\}} \mathcal{H} + \frac{\beta^2}{2!} \sum_{\{s\}} \mathcal{H}^2 - \cdots . \tag{3.2}$$

The first summation is simply the number of all possible state configurations: $\mathcal{M}^N$. This is a purely entropic term. Moving to the $\mathcal{O}(\beta)$ term, each link (*i.e.,* nonzero $\mathcal{A}_{ij}$) contributes an energetic contribution of $\sum_{s(i),s(j)} E_{s(i)s(j)}$, while the remaining nodes (other than i and j) contribute entropically as $\mathcal{M}^{N-2}$. In other words,

$$\sum_{\{s\}} \mathcal{H} = \frac{1}{2}\mathcal{M}^{N-2} \sum_{i,j} \mathcal{A}_{ij} \sum_{k,l} E_{kl} \tag{3.3}$$

$$= \frac{1}{2}\mathcal{M}^{N} \sum_{i,j} \mathcal{A}_{ij} E_0, \tag{3.4}$$

where $E_0 = \sum_{k,l} E_{kl}/\mathcal{M}^2$ is the average energy. Noting that $\sum \mathcal{A}_{ij}/2$ is the total number of links, equation 3.4 describes the mean-field energetic contribution.

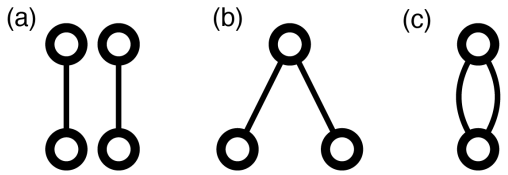


Figure 3.1: Three different types of two-link multigraphs.

Next, let us explicitly calculate the $\mathcal{O}(\beta^2)$ term in equation 3.2:

$$\frac{\beta^2}{2!}\sum_{\{s\}}\mathcal{H}^2 = \frac{\beta^2}{2!\cdot 4}\sum_{ijkl}\mathcal{A}_{ij}\mathcal{A}_{kl}\sum_{\{s\}}E_{s(i)s(j)}E_{s(k)s(l)}, \tag{3.5}$$

and there are three different types of energetic contribution, depending on the relationship between the two node pairs (i,j) and (k,l). The contribution of each type is represented diagrammatically in Figure 3.1. Here, each link represents a single energy term $E_{s(i)s(j)}$. Note that the graphs are no more simple in these diagrams; they are *multigraphs*, which allow multiple links (see Figure 3.1c).

To be specific, Figure 3.1 describes three different cases: (a) the pairs are totally disconnected (no nodes are same). The contribution is $\mathcal{M}^{N-4}\sum_{m,n,p,q}E_{mn}E_{pq}$. (b) They share only one of their nodes ($i = k$ or $i = l$ or $j = k$ or $j = l$; other nodes are all different). The contribution is $\mathcal{M}^{N-3}\sum_{m,n,p}E_{mn}E_{np}$. (c) The two pairs are identical (either $i = k$ and $j = l$, or $i = l$ and $j = k$). The contribution to the partition function is $\mathcal{M}^{N-2}\sum_{m,n}E_{mn}^2$. To generalize this energetic contribution, let us define $E(g)$ for graph g as follows:

$$E(g) = \mathcal{M}^{-n(\text{nodes})}\sum_{\text{nodes}}\prod_{k=1}^{n(g)}E_{l_k}, \tag{3.6}$$

where l_k indicates a link of index k, $n(g)$ is the number of links in graph g, and $n(\text{nodes})$ is the number of nodes in graph g. This is the energetic contribution of each multigraph normalized by the entropic contribution $\mathcal{M}^N$.

The number of possible (i,j,k,l) combinations for each multigraph type should be calculated. Let us define another variable, $[g]$:

$$[g] = \sum_{\text{nodes}}\prod_{k=1}^{n(g)}\mathcal{A}_{l_k}, \tag{3.7}$$

and the three different two-link multigraphs in Figure 3.1 give

$$\left[\;\right] = \sum_{i,j,k,l} A_{ij}A_{kl} \tag{3.8}$$

$$\left[\;\right] = \sum_{i,j,k} A_{ij}A_{jk} \tag{3.9}$$

$$\left[\;\right] = \sum_{i,j} A_{ij}^2 \tag{3.10}$$

However, they do not count the exact numbers, because equation 3.8 contains contributions from equations 3.9 and 3.10, and equation 3.9 includes those from equation 3.10. The exact number of contributing combinations for graph type g, which will be denoted by $W(g)$, is given as follows:

$$W\left(\;\right) = 2\left[\;\right] \tag{3.11}$$

$$W\left(\;\right) = 4\left\{\left[\;\right] - W\left(\;\right)\right\} \tag{3.12}$$

$$W\left(\;\right) = \left[\;\right] - W\left(\;\right) - W\left(\;\right) \tag{3.13}$$

The factors of 2 and 4 respectively in equations 3.11 and 3.12 come from the symmetry counting. For two pairs of indices (i, j) and (k, l), we have four possibilities to get graph 3.1b: $i = k, j = k, i = l$, and $j = l$, and two possibilities to get graph 3.1c: $(i = k) \wedge (j = l)$, and $(j = k) \wedge (i = l)$.

Furthermore, the energy contribution can be decomposed by defining energy deviation $\varepsilon_{kl} = E_{kl} - E_0$. For the linear term,

$$\sum_{k,l} E_{kl} = \mathcal{M}^2 E_0, \tag{3.14}$$

as we have seen above. Considering that $\sum_{k,l} \varepsilon_{kl} = 0$, arithmetics leads to

$$\sum_{k,l} E_{kl}^2 = \mathcal{M}^2 E_0^2 + \sum_{k,l} \varepsilon_{kl}^2, \tag{3.15}$$

$$\sum_{k,l,m} E_{kl}E_{lm} = \mathcal{M}^3 E_0^2 + \sum_{k,l,m} \varepsilon_{kl}\varepsilon_{lm}, \tag{3.16}$$

and

$$\sum_{k,l,m,n} E_{kl}E_{mn} \quad = \quad M^4 E_o^2. \tag{3.17}$$

Therefore, the explicit form of the $\mathcal{O}(\beta^2)$ term (equation 3.5) is

$$\frac{\beta^2}{2!\cdot 4}\left[W\left(\right) M^{N-4}\cdot M^4 E_o^2 + W\left(\right) M^{N-3}\left(M^3 E_o^2 + \sum_{k,l,m} \varepsilon_{kl}\varepsilon_{lm}\right) + W\left(\right) M^{N-2}\left(M^2 E_o^2 + \sum_{k,l} \varepsilon_{kl}^2\right)\right], \tag{3.18}$$

which becomes

$$\frac{\beta^2}{2!\cdot 4} M^N \left\{ E_o^2\left[W\left(\right) + W\left(\right) + W\left(\right)\right] + W\left(\right) M^{-3}\sum_{k,l,m}\varepsilon_{kl}\varepsilon_{lm} + W\left(\right) M^{-2}\sum_{k,l}\varepsilon_{kl}^2 \right\}, \tag{3.19}$$

or using equations 3.11, 3.12, and 3.13,

$$\frac{\beta^2}{2!\cdot 4} M^N \left\{ E_o^2\left[\right] + 4M^{-3}\sum_{k,l,m}\varepsilon_{kl}\varepsilon_{lm}\left[\right] + \left(2M^{-2}\sum_{k,l}\varepsilon_{kl}^2 - 8M^{-3}\sum_{k,l,m}\varepsilon_{kl}\varepsilon_{lm}\right)\left[\right] \right\}. \tag{3.20}$$

Note that the first term in parentheses is indeed

$$E_o^2\left[\right] = E_o^2\left(\sum A_{ij}\right)^2 = E_o^2\left[\right]^2 = \left\{ E_o W\left(\right)\right\}^2. \tag{3.21}$$

For higher-order terms, it is convenient to use the language of graph topology. As shown in equation 3.5, generally higher-order terms contain summations of products of A and E matrix elements, and we will systematically investigate them by using definitions 3.6 and 3.7.

Calculation of $W(g)$, the degeneracy of graph g, provides the partition function in the form of

$$Z(\beta) = M^N \left\{ 1 + \sum_g \frac{(-\beta/2)^{n(g)}}{n(g)!} W(g) E(g)\right\}, \tag{3.22}$$

where $n(g)$ is the number of links in graph g and the summation is over all possible (connected and disconnected) graphs g. To calculate $W(g)$, let us define a *child* graph and a *parent* graph. A child graph h of graph g is obtained by (non-simple) contraction[58] of unconnected nodes in graph g, and g is called a parent graph of h. For example, in Figure 3.1, graph b is a child graph of a, and a parent graph of c. Note that a child graph always has the same number of links as its parent.

Generally the following equation holds:

$$W(g) = K(g)\left\{[g] + \sum_{g' \in \mathcal{C}(g)} (-1)^{m(g',g)} K(g',g)[g']\right\}, \tag{3.23}$$

where $K(g)$ is the combinatoric factor to construct graph g from $n(g)$ links, $K(g',g)$ is the combinatoric factor to generate graph g' from graph g by node contraction, $m(g',g)$ is the number of contraction operations required to construct g' from g, and $\mathcal{C}(g)$ is the set containing all child graphs of graph g. For example, see equations 3.11, 3.12, and 3.13 for the case of two-link graphs. Since the number of links is same for parent and child graphs, we can write equation 3.22 in terms of $[g]$ by using the following formula: for the given number of links n,

$$\sum_{\substack{g \text{ s.t.} \\ n(g)=n}} W(g)E(g) = \sum_{\substack{g \text{ s.t.} \\ n(g)=n}} H(g)[g], \tag{3.24}$$

where

$$H(g) = K(g)E(g) + \sum_{g' \in \mathcal{P}(g)} (-1)^{m(g,g')} K(g,g')K(g')E(g'), \tag{3.25}$$

and $\mathcal{P}(g)$ is a set containing all parents of graph g, and the partition function is now

$$Z(\beta) = \mathcal{M}^N \left\{1 + \sum_g \frac{(-\beta/2)^{n(g)}}{n(g)!} H(g)[g]\right\}. \tag{3.26}$$

Note that since

$$K(g) + \sum_{g'} (-1)^{m(g,g')} K(g,g')K(g') = 0 \tag{3.27}$$

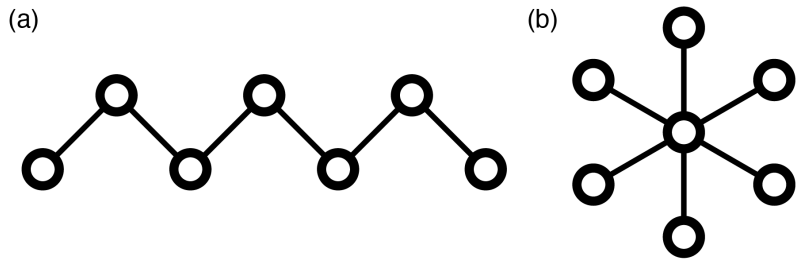


Figure 3.2: Examples of chain and star graphs. (a) Chain graph of length 6. (b) Star graph with 6 leaves.

unless g is a one-link graph, we can always replace $E(g)$ in equation 3.25 by $\varepsilon(g)$, which is defined as

$$\varepsilon(g) = \mathcal{M}^{-n(\text{nodes})} \sum_{\text{nodes}} \prod_{k=1}^{n(g)} (E_{l_k} - E_0). \tag{3.28}$$

For an unconnected graph g consisting of connected graphs $\{g_k\}$, $E(g) = \prod_k E(g_k)$, and $[g] = \prod_k [g_k]$. Thus, equation 3.22 can be written as

$$Z(\beta) = \mathcal{M}^N \exp\left\{ \sum_{\text{connected } g} \frac{(-\beta/2)^{n(g)}}{n(g)!} H(g)[g] \right\}, \tag{3.29}$$

and the free energy is

$$F(\beta) = -Nk_B T \ln \mathcal{M} + \sum_{\text{connected } g} \tilde{F}(g, \beta), \tag{3.30}$$

where

$$\tilde{F}(g, \beta) = -\frac{1}{\beta} \frac{(-\beta/2)^{n(g)}}{n(g)!} H(g)[g]. \tag{3.31}$$

One advantage of equation 3.30 is that the graph topology, contained in $[g]$, is now separated from detailed energetics, contained in $H(g)$. Hence, even without knowing the exact energy matrix, it is possible to compare $[g]$ values from different structures and, in some cases, we can determine which structure provides a more stable system. To illustrate this, let us consider two different graph systems, a chain graph of length N and a star graph with N leaves (Figure 3.2). They have the same numbers of nodes and links, so they have the same free energy up to the order of $\mathcal{O}(\beta^0)$ in equation 3.30.

The elements of their adjacency matrices for the chain and star graph systems, denoted by A^{chain}

and $\mathcal{A}^{\text{star}}$ respectively, are given below:

$$\mathcal{A}^{\text{chain}}_{ij} = \delta_{i,j+1} + \delta_{i,j-1} \tag{3.32}$$

$$\mathcal{A}^{\text{star}}_{ij} = \delta_{i,1} + \delta_{j,1} - 2\delta_{i,1}\delta_{j,1}, \tag{3.33}$$

where δ_{ij} is the Kronecker delta and index 1 for the star graph indicates the center. In calculating $[g]$ for a multiple graph g, we have a summation over various products of $\mathcal{A}_{ij}$, and when there is a common index (*e.g.* j in $\mathcal{A}_{ij}\mathcal{A}_{jk}$), the star graph gives a larger contribution to $[g]$ than the chain graph, since the two indices are disentangled in the former system. For example, let us calculate $S = \sum_{ijk} \mathcal{A}_{ij}\mathcal{A}_{jk}$ for both graphs:

$$\begin{aligned} S^{\text{chain}} &= \sum_{ijk}^{N} (\delta_{i-1,j} + \delta_{i+1,j})(\delta_{j,k+1} + \delta_{j,k-1}) \\ &= \sum_{i,k} (\delta_{i-1,k+1} + \delta_{i-1,k-1} + \delta_{i+1,k+1} + \delta_{i+1,k-1}) \\ &= (N-2) + (N-1) + (N-1) + (N-2) \\ &= 4N - 6, \end{aligned} \tag{3.34}$$

where $(N-k)$ terms are obtained by considering the boundary conditions. Also,

$$\begin{aligned} S^{\text{star}} &= \sum_{ijk}^{N} (\delta_{i,1} + \delta_{j,1} - 2\delta_{i,1}\delta_{j,1})(\delta_{j,1} + \delta_{k,1} - 2\delta_{j,1}\delta_{k,1}) \\ &= \sum_{ijk} (\delta_{i,1}\delta_{j,1} + \delta_{i,1}\delta_{k,1} + \delta^2_{j,1} + \delta_{j,1}\delta_{k,1} + \mathcal{O}(\delta^3)) \\ &= N^2 + 3N - 4N \\ &= N^2 - N. \end{aligned} \tag{3.35}$$

Here, since $\delta^2_{ij} = \delta_{ij}$, we get a quadratic dependence on N, which does not appear in the first case. Hence, $S^{\text{chain}} < S^{\text{star}}$ for $N > 3$. The same logic can be employed to conclude that

$$[g]^{\text{chain}} < [g]^{\text{star}}. \tag{3.36}$$

At the temperature high enough that the infinite sum in equation 3.30 for the star graph does not diverge and if it is stable (negative), we can conclude that the star graph system is more stable than

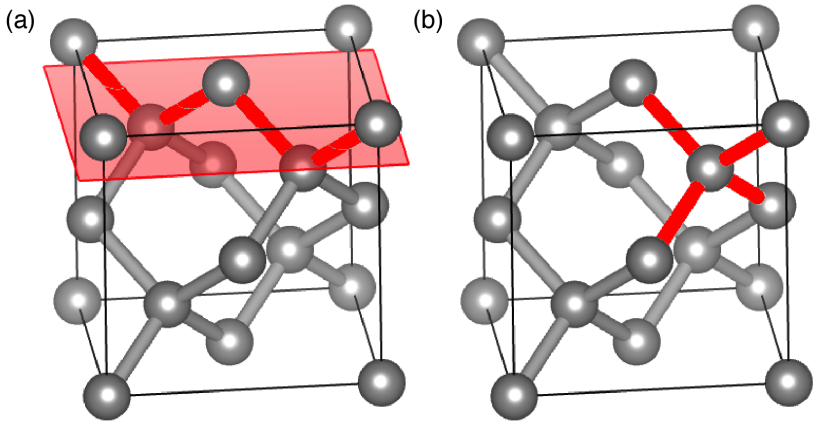


Figure 3.3: Two different types of defects. Broken bonds are marked in red. (a) Diamond structure with a grain boundary indicated by a red plane. (b) Same structure with a vacancy defect. Drawn by VESTA 3 [98].

the chain graph system. Note that this qualitative result is independent of the details of the energy matrix, which quantitatively determines the range of temperatures valid for the previous conclusion and the difference in free energies (*e.g.*, for the Ising model, the star graph system is always more stable than the chain graph system at any temperature). This explains why a graph system usually prefers a branched structure to a linear structure, if there are no other factors than link energies that determine the system energy.

3.3 Lattice Systems with Defects

A realistic application of this general conclusion is to a lattice system with defects. We will consider two types of defects. One is a planar defect (grain boundary), and another is a point defect (point vacancy). We constructed a 3-dimensional diamond-like lattice structure in $3 \times 3 \times 3$ unit cells with the periodic boundary condition (216 lattice points). The first system contains the grain boundary modeled by a discontinuity on the (001) plane (see Figure 3.3a). The second system simulating vacancy defects was constructed by removing lattice points randomly (see Figure 3.3b) until the number of broken bonds was equal to the number of bonds broken by the grain boundary in the first system (the number of remaining bonds = 324). Using a similar argument as above, it can be shown that $[g]$ is generally greater for the point vacancy system than for the grain boundary system, and a Monte Carlo (MC) simulation was conducted to check that the system with point vacancies is indeed more stable than that with a grain boundary.

A binary alloy of silicon and germanium is considered to make different states $s(i) = \text{Si, Ge}$ on

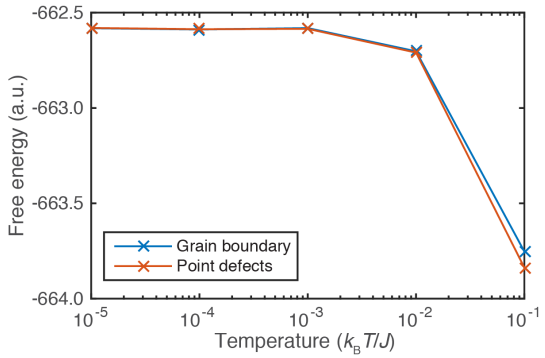


Figure 3.4: MC free energies of two Si-Ge alloy lattice systems with different types of defects, as a function of temperature (blue: grain boundary, red: point vacancies). The two systems have the same number of bonds.

lattice site i, and the interatomic potential developed in previous works[79,23] was employed, assuming equal bond lengths: $E(\text{Si, Si}) = -2.17J$, $E(\text{Ge, Ge}) = -1.93J$, $E(\text{Si, Ge}) = -2.04J$, where J is a constant. We tested five different temperatures, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5} in units of βJ. We compiled 1,000 independent MC simulations for each temperature, and in each simulation we performed 1.1 million MC steps and neglected the first 0.1 million steps.

The result is summarized in Figure 3.4. As inverse temperature βJ increases, the free energy difference between the two systems increases as well. As expected, the point defect system is more stable, which is consistent with the known fact that point vacancies have a thermal equilibrium concentration whereas higher-dimensional defects do not and they require external sources[106]. The entropic effect of multiple vacancy configurations and the stabilizing effect of structure relaxation have previously been used to explain this difference[125], but both factors were fixed in our simulations so they cannot account for the stability differences observed here. Also, note that the numbers of broken bonds are equal, meaning that the "surface areas" are the same. Thus, this result implies that the stability of point defects (compared to line and planar defects) is partially due to the lattice topology itself.

3.4 High-Temperature Expansion

Until now, we have considered qualitative differences between different graph systems. If an energy matrix is fully known, can we make a quantitative prediction? Among the multigraphs under consideration, each of which contributes independently to free energy (equation 3.30), we have *linear graphs*, defined as graphs where two (terminal) nodes have vertex degree 1 and the other nodes have

degree 2. A linear graph can be a parent graph of other connected graphs, but it does not have any connected parent graph. Since $[g] \geq [h]$ if g is a parent graph of h, the linear graph provides one of the largest $[g]$ values among the connected graphs with the same number of links. Explicitly, for a linear graph g of length $n(g)$,

$$[g] = \sum \mathcal{A}_{i_0 i_1} \mathcal{A}_{i_1 i_2} \cdots \mathcal{A}_{i_{n(g)-1} i_{n(g)}} = \mathrm{su}\, \mathcal{A}^{n(g)}, \tag{3.37}$$

where $\mathrm{su}\, \mathcal{A} \equiv \sum_{ij} \mathcal{A}_{ij}$. For energetic contributions, it can be shown similarly that

$$\varepsilon(g) = \mathcal{M}^{-n(g)-1} \mathrm{su}\, \varepsilon^{n(g)}. \tag{3.38}$$

Also,

$$K(g) = 2^{n(g)-1} \cdot n(g)!, \tag{3.39}$$

where $2^n \cdot n!$ is a combinatoric factor and due to symmetry we have the double-counting correction of 1/2. Considering only the first term in equation 3.25, we can denote this contribution to the free energy by

$$\tilde{F}_{\mathrm{lin}}(\beta) = -\frac{1}{2\beta\mathcal{M}} \sum_{\substack{\text{linear } g \\ \text{s.t. } n(g)\geq 2}} \left(-\frac{\beta}{\mathcal{M}}\right)^{n(g)} \mathrm{su}\, \varepsilon^{n(g)} \mathrm{su}\, \mathcal{A}^{n(g)}. \tag{3.40}$$

Diagonalization helps to get a closed form of this equation. For eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ of square matrix B of size N, we have $\mathrm{su}\, B^k = \sum_{i=1}^{N} |a_i|^2 \lambda_i^k$, where a_i is the inner product of the eigenvector corresponding to eigenvalue λ_i and an all-ones vector of size N[95]. Using this, we can write

$$\mathrm{su}\, \mathcal{A}^k = \sum_{i=1}^{N} |c_i|^2 \lambda_i^k \tag{3.41}$$

$$\mathrm{su}\, \varepsilon^k = \sum_{j=1}^{M} |d_j|^2 \mu_j^k, \tag{3.42}$$

where $\{\lambda_i\}$ and $\{\mu_i\}$ represent the spectra of $\mathcal{A}$ and ε respectively, and $\{c_i\}$ and $\{d_i\}$ correspond to $\{a_i\}$ above for $\mathcal{A}$ and ε respectively. Thus, equation 3.40 becomes

$$\tilde{F}_{\mathrm{lin}}(\beta) = \frac{1}{2\mathcal{M}^2} \sum_{i,j}^{N,M} \frac{|c_i|^2 |d_j|^2 \lambda_i \mu_j}{1 + \beta\lambda_i \mu_j / \mathcal{M}}, \tag{3.43}$$

for $Mk_BT > \max(|\lambda_i|)\max(|\mu_j|)$. The total free energy is now

$$F(\beta) = -Nk_BT\ln M + \frac{1}{2}E_0 \cdot \mathrm{su}\, A - \frac{\beta}{4M^2}\left(\mathrm{tr}\,\varepsilon^2 - \frac{2}{M}\mathrm{su}\,\varepsilon^2\right)\mathrm{tr}\,A^2 + \frac{1}{2M^2}\sum_{i,j}\frac{|c_i|^2|d_j|^2\lambda_i\mu_j}{1+\beta\lambda_i\mu_j/M} + \mathcal{O}(\beta^2), \tag{3.44}$$

while a mere linear approximation (from equation 3.20) gives

$$F(\beta) = -Nk_BT\ln M + \frac{1}{2}E_0 \cdot \mathrm{su}\, A - \frac{\beta}{4M^2}\left\{\left(\mathrm{tr}\,\varepsilon^2 - \frac{2}{M}\mathrm{su}\,\varepsilon^2\right)\mathrm{tr}\,A^2 + \frac{2}{M}\mathrm{su}\,\varepsilon^2\mathrm{su}\,A^2\right\} + \mathcal{O}(\beta^2). \tag{3.45}$$

3.5 Sequence Space Free Energy of Heteropolymer

To illustrate the utility of the quantitative formula (equation 3.44), let us consider an example from biophysics. As previously investigated [48], the sequence space free energy of a heteropolymer is closely related to protein evolution, which is governed by sequence space dynamics through mutations and is therefore of deep interest to protein biophysicists. This free energy can be quantitatively predicted by the formulae above, and equation 3.44 shows a better performance than equation 3.45 for predicting free energies of lattice proteins, which can be described by small graphs and hence have degeneracies in low-order structural terms such as $\mathrm{su}\, A^2$.

The Hamiltonian of a 3-dimensional lattice protein is given by

$$\mathcal{H} = \frac{1}{2}\sum_{i,j}^{N,N} A_{ij}E_{\mathrm{AA}(i)\mathrm{AA}(j)}. \tag{3.46}$$

Here A is called a contact matrix in the protein structure literature, whose element A_{ij} is 1 when residues i and j are in contact, and $A_{ij} = 0$ otherwise. E is an interaction matrix that contains information about interaction energy between two amino acid types. N is the chain length, and $\mathrm{AA}(k)$ indicates the amino acid type of residue k. Unlike the previous work [48], we do not need to assume any special form of the interaction matrix.

We studied the sequence space of a $3\times3\times3$ lattice protein structure, whose graph representation consists of only 27 nodes. There are 103,346 maximally compact structures of a $3\times3\times3$ lattice protein [113], but we used 10,000 representative structures to reduce the computational cost, following Heo *et al* [59]. We also used two-letter alphabet, whose corresponding interaction matrix was chosen

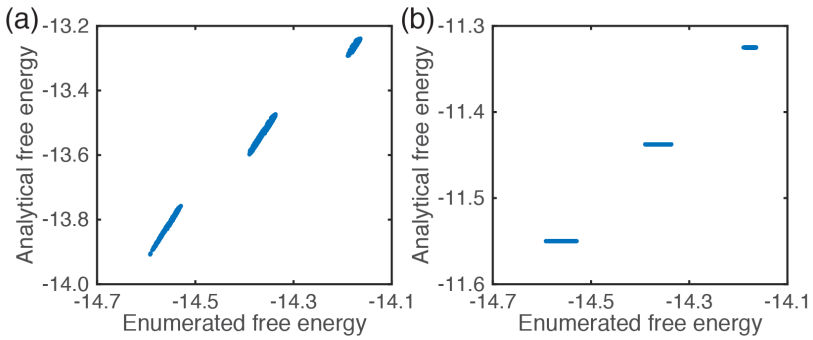


Figure 3.5: Scatter plots of sequence-space free energy distributions for 10,000 lattice protein structures, comparing enumerated (exact) free energy and analytical (approximate) free energy. To compute analytical free energy, equation 3.44 was used for (a) and equation 3.45 for (b). $\beta = 0.1$ for both panels.

to represent hydrophobic-polar interactions in proteins:

$$E = \begin{bmatrix} -3 & 1 \\ 1 & 0 \end{bmatrix} \tag{3.47}$$

Then we scanned all 2^{27} (about 1.3×10^8) possible sequences to compute $Z(\beta) = \sum_{\text{sequences}} \exp(-\beta\mathcal{H})$ and corresponding $F(\beta)$ for given β. We denote this $F(\beta)$ as "enumerated free energy," to distinguish it from the theoretical prediction, which will be called "analytical free energy."

Figure 3.5 shows the sequence space free energy distributions at $\beta = 0.1$. Equation 3.44 can perfectly discriminate different structures (one-to-one correspondence; panel a), but equation 3.45 is not capable of discriminating among structures with different free energies (panel b), due to the degeneracies in tr $\mathcal{A}^2$ and su $\mathcal{A}^2$. Note that even in the former case, the structures are mainly separated by three different su $\mathcal{A}^2$ values, implying that they are still in the high-temperature regime.

3.6 Conclusion

In this chapter, we demonstrate an analytical method for systematically calculating the free energy of a spin model on a simple graph. Through this approach, we find that the topology, realized by $[g]$, contributes to the free energy, independently of energetic or other factors. Thus, it can be used to qualitatively predict a more stable structure among different ones at high temperature. The theory was illuminated by comparison between chain and star graphs; without specifying the interaction matrix, we showed that the star graphs are more stable than chain graphs at the high-temperature limit. The approach was applied to lattice models with different defect types, which lead to different

free energies, even with the same surface areas of defects. We also showed that linear graphs are special in a sense that their infinite sum can be computed exactly, and this approach was applied to the protein design problem. The relative order of sequence space free energies of lattice proteins were perfectly predicted by the formula containing the linear graph contribution, whereas a mere linear approximation barely discriminate three groups with different su A^2 values. We believe that this theory can be expanded and applied to other graph-related problems in physics, from more complex spin systems (*e.g.* spin glass model) to biological systems (*e.g.* protein-protein interaction networks and neural networks) and also social networks.

4

Systems-Level Responses of *Escherichia coli* to Perturbations

4.1 Background

To understand how a small genetic variation can lead to a drastic effect on cellular behaviors, it is required to consider a complex interplay between various scales, from the molecular level, through the systems level, to the cellular level[82]. Several studies demonstrated that mutations in metabolic enzymes have local effects on fitness through changes in metabolic flux[6,38,119]. Mutations that change protein stability can also affect fitness through modulation of the number of functional folded proteins[14,51,142] or by affecting the number of toxic unfolded species[39,42]. However, in most cases, a direct link between the mutational effects on protein function and organismal phenotype is not obvious due to pleiotropic effects, such as protein aggregation[42] and formation of functional and non-functional multimers[13,89,147]. Furthermore, recent studies have shown that partial inhibition of an enzyme can cause broad changes in the metabolic profile of the cell, extending far beyond the immediate products of enzymes in question[78,77].

The systems-level proteomic response to a genetic variation is an important stepping stone to understand the relationship of genetic variations to cellular responses. Earlier studies showed that bulk characteristics of the macromolecular composition in the cell cytoplasm (*e.g.*, the total protein

concentration or the ratio of proteins to RNAs) are sensitive to changes in growth conditions, such as the availability of nutrients[44,69]. However, the effect of mutations or changed growth conditions on the abundances of individual proteins in the cytoplasm is not known. The key objective of the present study is to understand to what extent point mutations in a metabolic enzyme and/or variations in the media affect the proteome composition in the bacterial cytoplasm and how these changes are related to the fitness effects of such mutations.

The Shakhnovich group has been experimentally studying the effect of mutations in the chromosomal copy of the *folA* gene, which encodes the core metabolic enzyme dihydrofolate reductase (DHFR). DHFR is one of key enzymes in *E. coli* metabolic pathways. It catalyzes the reduction reaction of dihydrofolate into tetrahydrofolate, which is a crucial substrate of the one-carbon metabolic pathway. This pathway is linked to *de novo* synthesis of purine, as well as methionine and glycine biosynthetic cycles[112]. Therefore, a perturbation in DHFR may impact a relatively large number of pathways and their constituents, so that the perturbation eventually leads to cellular behavior changes that are experimentally detectable. Also, DHFR has a low copy number in *E. coli* cytoplasm (approximately 40 copies per cell)[124], and hence, the possibility of protein aggregation after a perturbation is relatively small.

In this chapter, we will study how a perturbation in DHFR alters the expression levels of other proteins, and show that the systems-level analysis can broaden our understanding on the genotype-phenotype relationship.

4.2 Perturbations

We employed two types of perturbations, point mutations and drug inhibition, and two types of growth media for the mutant strains, as explained below.

4.2.1 Point mutations

In the previous studies[13,12], a set of chromosomal missense point mutations was introduced in the open reading frame of the *E. coli folA* gene, and their effects were evaluated in terms of the biophysical and biochemical properties of the encoded DHFR molecule, as well as the cellular growth rate, which is a proxy to the fitness[20,53,54]. Among a wide range of mutations, we found several mutations that led to a noticeable drop of growth rate, even though they form soluble oligomers, implying that the drop is not simply due to aggregation-associated toxicity[13].

In this study, we selected four mutant strains with estimated $\Delta\Delta G$ values ranging from 2.8 to 6.4 kcal/mol (these values are estimated upon the assumption of additivity of stability effects of single-

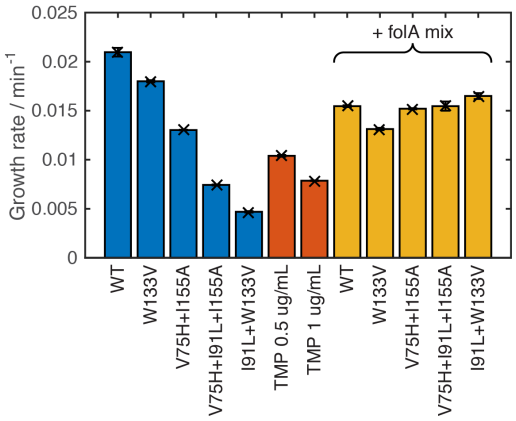


Figure 4.1: Growth rates of all studied strains under various conditions. Growth rates were determined from the exponential phase of growth using the three-parameter fit of $\log(OD/OD_0)$ versus time curves proposed in Zwietering *et al* [151]. The folA mix brought the growth rates of the mutant strains very close to the WT level. Error bars represent the SDs of three independent growth measurements.

point mutations). Mutants W133V and V75H+I155A showed a slight drop in growth rates, while the growth of V75H+I91L+I155A and I91L+W133V strains was severely compromised (Figure 4.1). We determined that the observed loss of fitness stems primarily from the loss-of-function effect of the destabilizing mutation that renders DHFR molecules susceptible to rampant aggregation or degradation in the cell [12].

4.2.2 Drug inhibition

Trimethoprim (TMP) is a well-known competitive inhibitor of DHFR, and it shows a strong specificity for DHFR [131]. TMP also has a high degree of specificity for bacterial DHFRs over eukaryotic DHFRs, which allows it to have been widely used as an antibiotic, sometimes combined with other drugs such as sulfamethoxazole [5]. In this study, we used two different concentrations of TMP, 0.5 ug/mL and 1.0 ug/mL, and as expected, the growth rate drops as the TMP concentration increases (Figure 4.1).

4.2.3 Standard medium and the FolA mix

The standard growth medium is the M9 minimal medium supplemented with 0.2 % glucose, 1-mM $MgSO_4$, 0.1 % casamino acids, and 0.5-ug/mL thiamine. However, it has been known that supplementing the growth media with a combination of purine, thymidine, pantothenate, glycine, and

Proteome (- folA mix)		Proteome (+ folA mix)		Transcriptome	
W133V	2,195	WT	1,849	W133V	4,192
V75H+I155A	2,194	W133V	1,850	V75H+I155A	4,189
TMP 0.5 ug/mL	2,195	V75H+I155A	1,850	TMP 0.5 ug/mL	4,189
V75H+I91L+I155A	2,194	V75H+I91L+I155A	1,849	V75H+I91L+I155A	4,191
I91L+W133V	2,195	I91L+W133V	1,847	I91L+W133V	4,189

Table 4.1: Numbers of genes and protein products detected and quantified in the quantitative proteomics and transcriptomics experiments.

methionine sustains the growth of *E. coli* that lacks the *folA* gene[116]. We found that the growth rate differences between the WT and mutants are equalized when the cells grow on this media (Figure 4.1). We will call this additional nutrient combination the "folA mix."

4.3 Global Effects of Perturbations on the Proteome and Transcriptome

To determine the relationship between the fitness of the selected mutant strains and the systems-level responses to the DHFR mutations, we quantified changes in the protein abundances in the *E. coli* proteome. To this end, we applied chemical labeling based on isobaric TMT technology with subsequent LC-MS/MS quantification[4,117,127]. This method allowed us to obtain relative protein abundances (RPAs) between each strain/condition in question and a reference system. As a reference, we chose WT *E. coli* in our standard growth media ("unperturbed system"):

$$\mathrm{RPA}(g;c) = \frac{N_{\mathrm{protein}}(g;c)}{N_{\mathrm{protein}}(g;\varnothing)}, \tag{4.1}$$

where $N_{\mathrm{protein}}(g;c)$ indicates the protein expression level of gene g in the system under condition c and $N_{\mathrm{protein}}(g;\varnothing)$ the protein expression level of g in the reference system. In total, we quantified 11 proteomes that included all conditions listed in Figure 4.1. To control for natural biological variation at different stages of growth, we also collected the RPA data for WT strains grown to different optical density (OD) levels. We were able to detect and quantify approximately 2,000 proteins (Table 4.1) available for direct comparison between all 11 proteomes.

To assess the relationship of the proteome changes to the transcriptome, we also obtained, under identical experimental conditions, transcripts of the *folA* mutant strains and the WT strain treated with 0.5 ug/mL of TMP. Here we similarly define relative mRNA abundances (RMAs):

$$\mathrm{RMA}(g;c) = \frac{N_{\mathrm{mRNA}}(g;c)}{N_{\mathrm{mRNA}}(g;\varnothing)}, \tag{4.2}$$

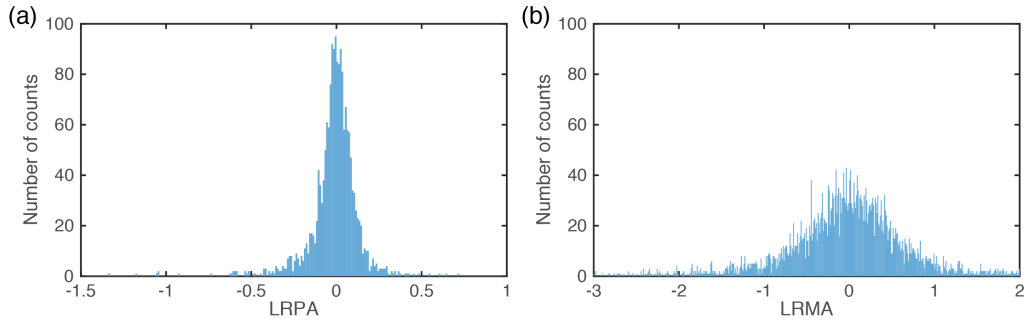


Figure 4.2: Typical distributions of LRPAs and LRMAs. Data shown are for one repeat of the W133V strain. Other distributions are shown in Figure 4.12 at the end of the chapter.

where $N_{\mathrm{mRNA}}(g; c)$ indicates the mRNA expression level of gene g in the system under condition c and $N_{\mathrm{mRNA}}(g; \varnothing)$ the mRNA expression level of g in the unperturbed system. Total mRNA was extracted and the reads were aligned to the reference genome so that we could measure the absolute transcript levels N_{mRNA}, from which the ratio was calculated according to equation 4.2.

We use logarithms of RPAs and RMAs (denoted by LRPAs and LRMAs, respectively) to see the fold changes of expression levels easily:

$$\mathrm{LRPA}(g; c) = \log N_{\mathrm{protein}}(g; c) - \log N_{\mathrm{protein}}(g; \varnothing), \tag{4.3}$$

$$\mathrm{LRMA}(g; c) = \log N_{\mathrm{mRNA}}(g; c) - \log N_{\mathrm{mRNA}}(g; \varnothing). \tag{4.4}$$

Note that the sign of LRPA or LRMA indicates the direction of regulation. If the logarithmic value is positive, the condition increases the expression level of the gene with respect to the reference system, and this can be considered as up-regulation. The opposite case will be down-regulation. The typical LRPA and LRMA distributions are shown in Figures 4.2a and 4.2b, respectively, and other distributions are summarized in Figure 4.12 at the end of this chapter.

We found that there exists a robust and statistically significant anti-correlation between the standard deviations (SDs) of LRPA distributions and the growth rates (Figure 4.3a). Generally, the SDs of LRMA distributions are about twice as big as those of LRPA distributions (Figure 4.3b), suggesting that mRNA abundances are more sensitive to genetic variation, probably due to the lower copy numbers of mRNAs compared with the proteins that they encode. (Note that another basic statistical quantity, the mean of an LRPA/LRMA distribution may vary from sample to sample due to slight variation of final OD of samples, and thus it cannot be a reliable measure of the systems-level response.)

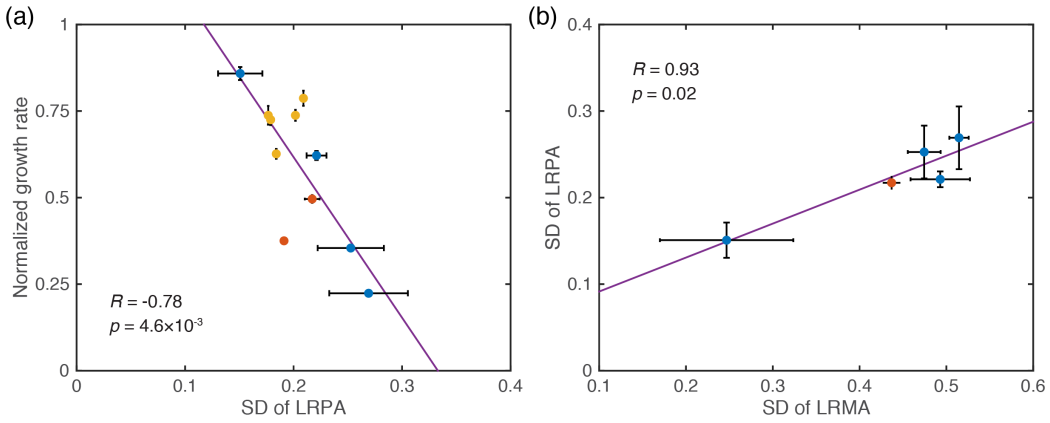


Figure 4.3: Global statistical properties of proteomes and transcriptomes. Colors indicate different experimental conditions (blue: mutation, red: TMP inhibition, yellow: folA mix), and purple lines refer to the regression lines. (a) The SD of LRPA distribution is anti-correlated with the growth rate. Error bars correspond to the SDs of three independent experiments. (b) The SD of LRMA distribution is correlated with that of LRPA distribution. The slope is close to 2, suggesting that transcriptomes are more readily perturbed than proteomes. Error bars correspond to the SDs of two independent transcriptomics experiments (x axis) and three independent proteomics experiments (y axis).

Importantly, the variation of SD of LRPA between strains and conditions is not a mere consequence of natural biological variation between growth stages: the SD of LRPA for the WT strain grown to different OD remain remarkably constant (Figure 4.4a). In addition, when comparing two proteomes extracted independently from the WT strain grown up to entrance into stationary phase under identical conditions (biological repeats), the correlation of LRPA between them is very high (Pearson's correlation coefficient $R = 0.94$), indicating that the TMT-labeling-based proteome quantification technique is highly reproducible (Figure 4.4b).

4.4 Comparison between Biological Repeats: Reproducibility

The broad distributions of LRPA and LRMA might indicate that variations in protein and mRNA abundances are just a consequence of stochastic sample-to-sample variation between colony founder cells. If this were the case, we could not see strong reproducibility from sample to sample. Another possibility is that the broad distributions of LRPA and LRMA are due to long-time intrinsic stochasticity in gene expression [46], which extends beyond a mere cell-to-cell variation to affect the total abundances in the bulk. In that case, we might still find that the overall statistical properties of the proteome response to a perturbation (such as the SD of an LRPA/LRMA distribution) are robust and reproducible between samples from biological repeats.

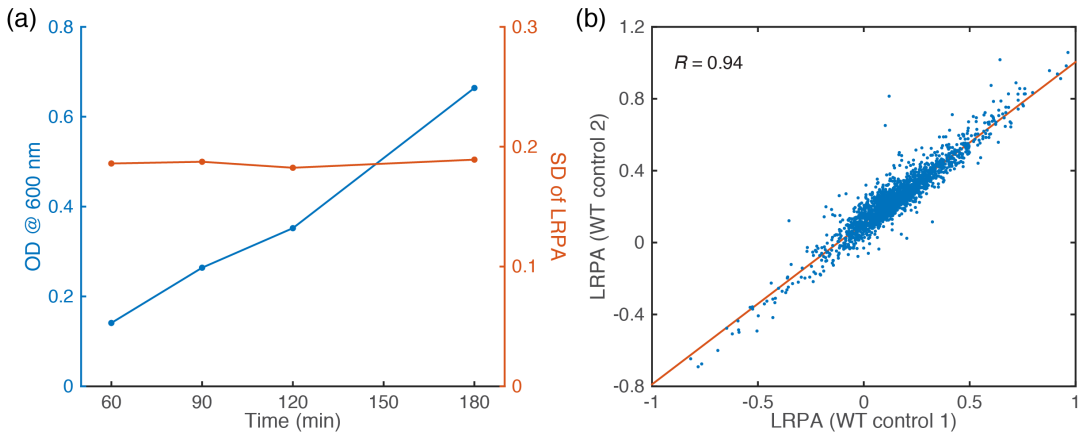


Figure 4.4: Reproducibility of global statistical properties. (a) The SDs of LRPA distributions determined for wild-type strains grown to different OD are remarkably constant. This result indicates that observed variability of SDs between strains and conditions is not due to natural variation of the biomass. (b) Scatter plot of LRPAs between two independent proteome data of the wild-type strain. The red line indicates the regression line. Both proteomes were obtained independently from the wild-type strain grown to the same OD levels under identical conditions, and both are normalized to the proteomics data obtained at t=60 min. The TMT-labeling based LC-MS/MS quantification technique shows a very high correlation between highly complex protein mixtures (over 2,000+ proteins) extracted from identical biological repeats.

An extreme scenario of the latter case is that each protein abundance consistently responses to genetic or media variation. In other words, the LRPA/LRMA of each protein is always reproducible (apart from the experimental noise) from sample to sample at the identical condition. We note that a mere analysis of the LRPA/LRMA distribution from an individual experiment does not allow us to distinguish between randomly and consistently varying quantities since the LRPAs or LRMAs for all genes, whether random or consistent, appear to be drawn from the same distributions, as shown in Figures 4.2 and 4.12. Only comparison of LRPA/LRMA distributions between biological repeats can reveal the degrees of randomness and consistency in the proteomics and transcriptomics responses to the perturbations.

For further analysis, we separated the strain-to-strain variation of global statistical properties – means and SDs – from the variation of the abundances of individual proteins. To that end, we normalized the LRPA and LRMA for gene g in experimental condition c to obtain the z-score:

$$z(g;c) = \frac{L(g;c) - \langle L(c) \rangle_g}{\sigma_{L(c)}}, \tag{4.5}$$

where $L(g;c)$ is the LRPA or LRMA of gene g in condition c, $\langle L(c) \rangle_g$ and $\sigma_{L(c)}$ indicate the arithmetic mean and standard deviation of the LRPA or LRMA distribution in condition c, respectively.

We modeled the gene set as a set consisting of two different types of genes. The genes of one type consistently response to the change of experimental condition, while the other gene group responses randomly. Assume that the total number of genes is N and the number of genes in the former group is K. For the "consistent" genes, the z-scores from two biological repeats A and B in the same experimental condition c are identical up to the experimental noise:

$$z_{\mathrm{B}}(i;c) = z_{\mathrm{A}}(i;c) + \eta(i), \tag{4.6}$$

where i is the gene index ($i = 1, 2, \cdots, K$) and $\eta(i)$ indicates the experimental noise on gene i. On the contrary, the z-scores of "random" genes (whose indices are from $K+1$ to N) are statistically independent between biological repeats. Then, we can write Pearson's correlation coefficient $R_{\mathrm{AB}}(c)$ between the z-scores of gene sets from two biological repeats A and B in experimental condition c as

follows:

$$R_{AB}(c) = \frac{\sum_{i=1}^{N} z_A(i;c) z_B(i;c)}{\sqrt{\sum_{i=1}^{N} [z_A(i;c)]^2} \sqrt{\sum_{i=1}^{N} [z_B(i;c)]^2}} \tag{4.7}$$

$$= \frac{\sum_{i=1}^{K} [z_A(i;c)]^2 + \sum_{i=1}^{K} z_A(i;c)\eta(i) + \sum_{i=K+1}^{N} z_A(i;c) z_B(i;c)}{\sqrt{\sum_{i=1}^{N} [z_A(i;c)]^2} \sqrt{\sum_{i=1}^{N} [z_B(i;c)]^2}} \tag{4.8}$$

$$\approx \frac{K\langle [z_A(i;c)]^2 \rangle_i + \mathcal{O}\left(\sqrt{K}\right) + \mathcal{O}\left(\sqrt{N-K}\right)}{N\langle [z_A(i;c)]^2 \rangle_i} \tag{4.9}$$

$$\approx \frac{K}{N}. \tag{4.10}$$

Here we apply the central limit theorem to approximate the sum of n random variables with accuracy up to $\mathcal{O}(\sqrt{n})$, and assume that N and K are both large numbers. The final result is obtained by keeping linear terms only in both the numerator and denominator and omitting all square-root terms. Note that the error is of the order of $1/\sqrt{N}$, as indicated in equation 4.9.

From equation 4.10, we can estimate the number of "consistent" genes in the gene set by comparing two biological repeats and checking Pearson's correlation coefficient. We prepared three biological repeats for proteomics analysis and two biological repeats for transcriptomics analysis, in five different experimental conditions (4 mutants + 1 TMP inhibition). As shown in Figure 4.5, the correlation coefficients are generally high, ranging from 0.56 to 0.92. Note that the correlation coefficients are overall higher for mRNAs than for proteins. This implies that the stochasticity at the mRNA level is smaller than that at the protein level.

This simple analysis suggests that a good portion of the observed LRMA and LRPA distributions in different experimental conditions are not just simple manifestation of a noisy gene expression or an epigenetic sample-to-sample variation in the founder clones. Rather, we observed that in each case more than 1,000 genes vary their mRNA and protein abundances in a consistent manner in response to DHFR mutations and drug inhibition. It is important to note that this conclusion does not depend on the assumption about the amplitude of the experimental noise, since the error is of the order of $1/\sqrt{N}$ and $N \approx 2,000 \gg 1$.

Lastly, we checked if the consistent behavior of genes is due to variation between the growth stages and culture densities for different experimental conditions. We compared the WT proteomes at different OD values to the proteomes of different perturbations, and found generally low correlations at all OD values (Figure 4.6). This indicates that the variation of proteomes at different growth stages does not account for the LRPAs in different experimental conditions. Consequently,

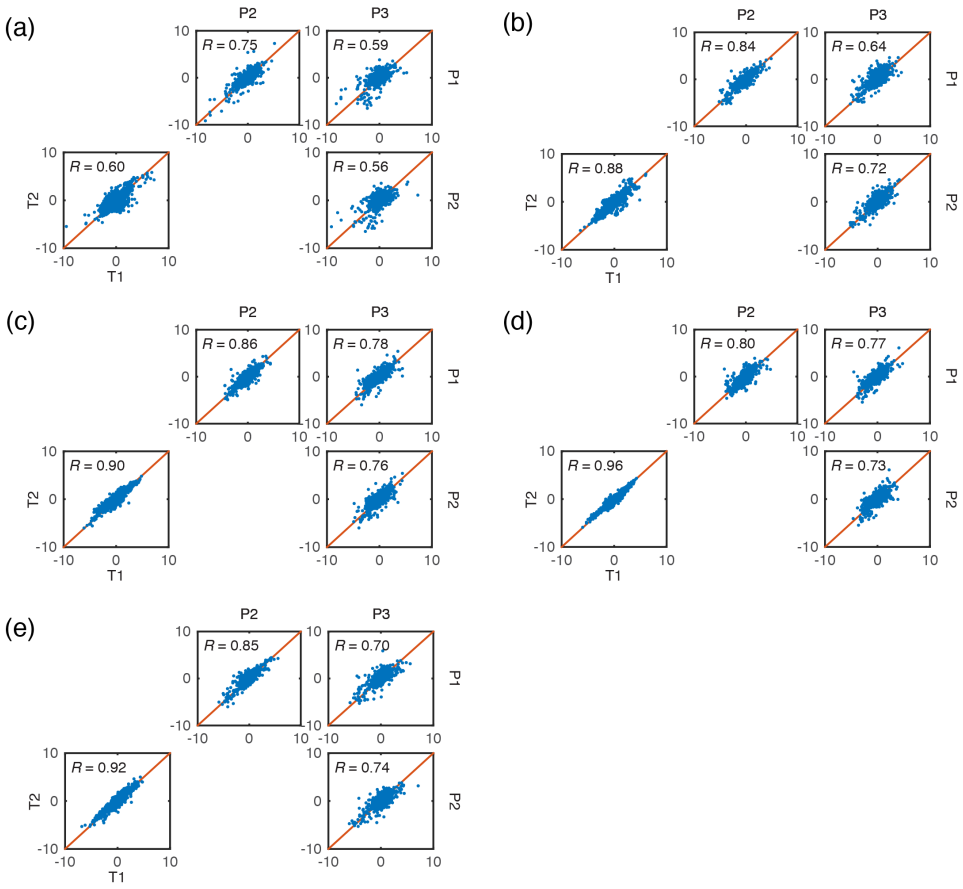


Figure 4.5: Correlations between LRPA/LRMA z-scores in biological repeats. T1 and T2 denote transcriptomics repeats and P1, P2 and P3 denote repeated proteomics experiments. (a) W133V mutant. (b) V75H+I155A mutant. (c) V75H+I91L+I155A mutant. (d) I91L+W133V mutant. (e) WT strain treated with TMP 0.5 ug/mL.

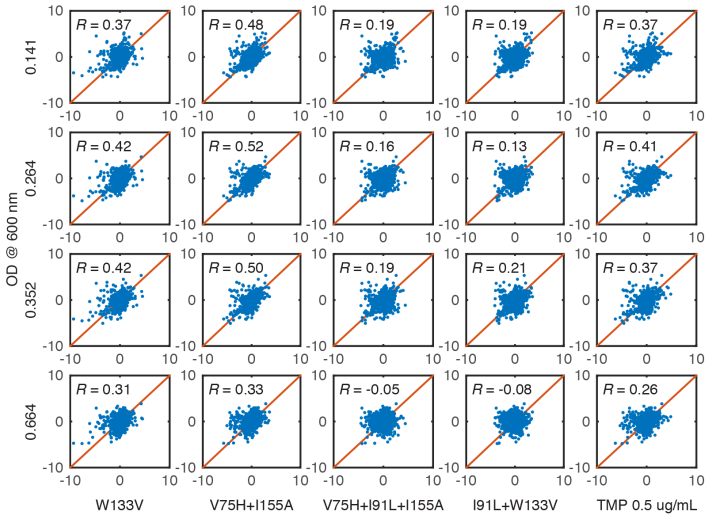


Figure 4.6: Scatter plots of the z-scores of LRPA between proteomes of the WT strain grown to various OD levels versus those of the mutant strains and TMP-treated WT strain. Red lines mark the y=x lines. Low correlations indicate that perturbations observed in response to DHFR mutations or functional inhibition by TMP is largely unrelated to natural variation rooted in different stages of the growth cycle.

we conclude that the *E. coli* proteome and transcriptome are highly sensitive to perturbations in the metabolic enzyme DHFR; a vast number (in the range of 1,000-2,000) of genes consistently vary their transcription levels and protein abundances in response to mutations in the *folA* gene or TMP inhibition.

4.5 Comparison between Different Types of Perturbations

The z-score of a specific gene indicates the relative intensity of regulation of the gene in the proteome/transcriptome. For example, a higher z-score of LRPA indicates that the gene is more up-regulated upon the given perturbation than other genes in the whole proteome. Comparison of the z-score distributions of two different experimental conditions provides a measure for how much their proteome/transcriptome responses to the perturbations are similar. We compared the proteomes from 6 different conditions (4 mutants + 2 TMP inhibition systems) in the standard medium, the proteomes from 5 different conditions (4 mutants + 1 TMP inhibition) in the folA mix medium, and the transcriptomes from 5 different conditions (4 mutants + 1 TMP inhibition) in the standard medium.

There is a remarkable pattern in the correlations between proteomes of different conditions.

Proteomes that show a moderate decrease in growth (W133V, V75H+I155A, and WT treated with 0.5 ug/mL of TMP) are closely correlated between themselves, as are the proteomes of cells with a severe decrease in growth rates (I91L+W133V, V75H+I91L+I155A, and WT treated with 1 ug/mL of TMP). The correlation between members of these two groups is considerably weaker, albeit still highly statistically significant (Figure 4.7a, upper right). The addition of the folA mix, which nearly equalizes the growth between WT and even the most detrimental mutants (Figure 4.1), significantly reduces this separation of two classes, making correlations between all proteomes uniformly high (Figure 4.7a, lower left). A similar but less pronounced pattern of correlations is observed for LRMA data (Figure 4.7b). The observation that strains having similar growth rates tend to have similar proteomes might suggest that the growth rate is the single determinant of the proteome composition. However, a more careful analysis shows that this is not the case: the growth rate is not the sole determinant of the proteome composition.

We clustered the LRPA z-scores using the Ward clustering algorithm[138]. The z-score set from experimental condition c is considered as a vector $\vec{x}_c$, whose elements are z-scores of different genes, in a multi-dimensional space (dimensionality $\approx$ 2,000). The distance between a pair of vectors is measured by the typical Euclidean distance metric. Based on this set of metric data, Ward's hierarchical clustering method is used to determine which conditions are clustered. Here, clusters are generated in order to minimize the error sum of squares, or the variance. Consider two clusters A and B, each of which respectively contains n_A and n_B elements and $A \cap B = \varnothing$. The merging cost $\Delta(A, B)$ of combining A and B is defined as

$$\Delta(A, B) = \sum_{c \in A \cup B}^{n_A+n_B} ||\vec{x}_c - \vec{m}_{A\cup B}||^2 - \sum_{c \in A}^{n_A} ||\vec{x}_c - \vec{m}_A||^2 - \sum_{c \in B}^{n_B} ||\vec{x}_c - \vec{m}_B||^2, \tag{4.11}$$

where $||\vec{x} - \vec{y}||$ indicates the distance between $\vec{x}$ and $\vec{y}$, and $\vec{m}_K = \sum_{c \in K}^{n_K} \vec{x}_c / n_K$ is the center of cluster K of size n_K. Note that each summation corresponds to the error sum of squares, *i.e.* the variance, of each cluster. Arithmetics leads to the following simple formula:

$$\Delta(A, B) = \frac{n_A n_B}{n_A + n_B} ||\vec{m}_A - \vec{m}_B||^2. \tag{4.12}$$

To get the same dimension with the Euclidean distance, the following metric (Ward's linkage) is used to determine the distance between clusters A and B:

$$d(A, B) = \alpha \sqrt{\Delta(A, B)}, \tag{4.13}$$

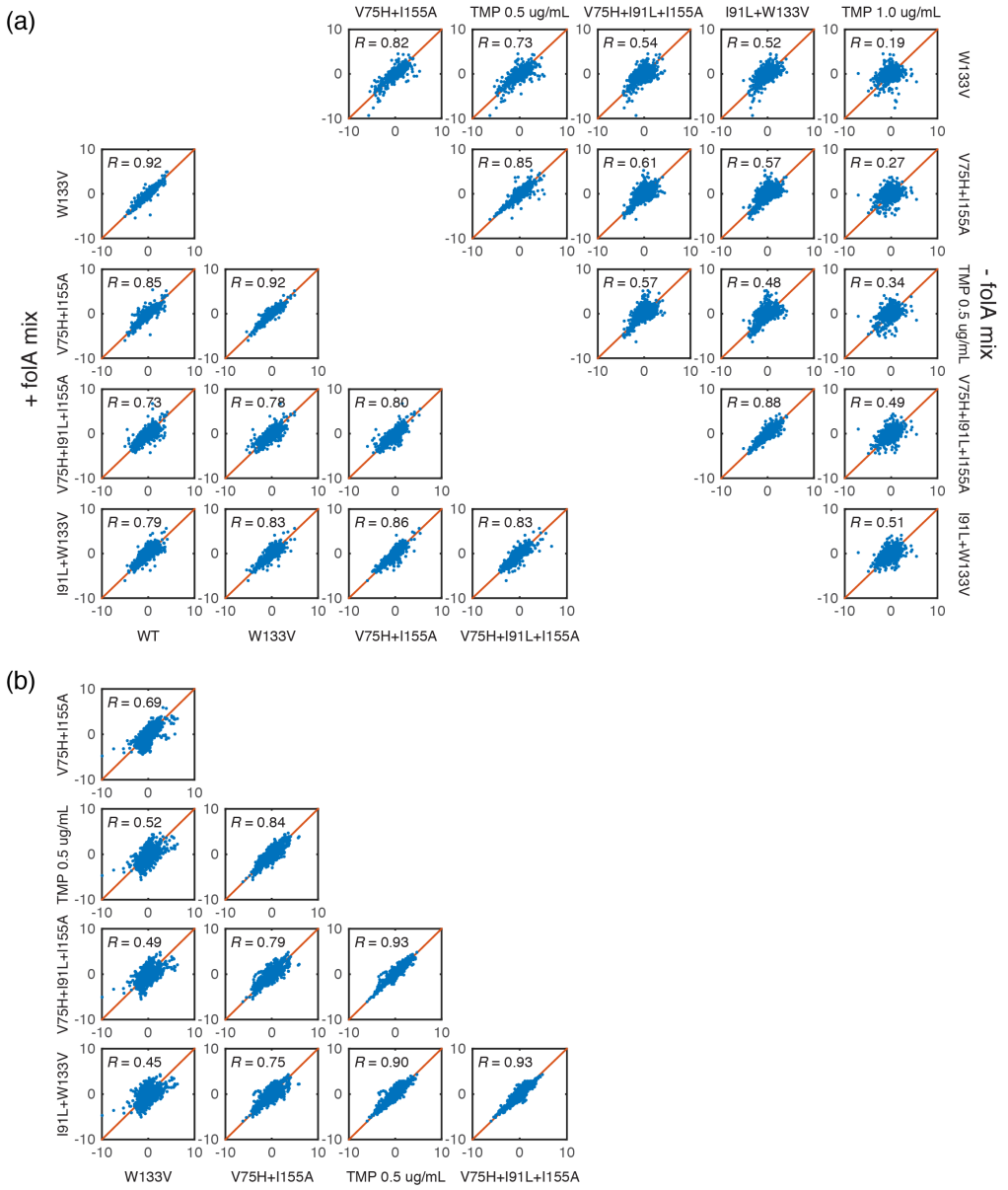


Figure 4.7: Correlations between proteomes and transcriptomes. Red lines mark the y=x lines. (a) Scatter plots between proteomes of all systems under the standard growth condition (upper right) and in presence of the folA mix (lower left). The addition of the folA mix minimizes the variations between different proteomes. (b) Transcriptomics data obtained for systems grown under the standard condition. Correlations are overall higher for mRNA abundances, but similar classes of transcriptomes are discernible.

and constant α is set to $\sqrt{2}$ so that the Euclidean distance between two singleton clusters A and B is identical with the Euclidean distance $\sqrt{||\vec{m}_A - \vec{m}_B||^2}$. Beginning with each vector consisting of its own cluster, Ward's hierarchical clustering merges clusters to minimize the growth of the sum of $d(A, B)$. In practice, this method is known to be comparatively effective to reconstruct the original cluster structure[50]. One advantage of Ward's method is that it is not extremely sensitive to outliers, and in this sense we think that this method is better than other methods for our purpose, since we are in the situation with a relatively strong background biological noise.

It should be noted here that the Euclidean distance between two vectors has a one-to-one correspondence to Pearson's correlation coefficient R. For two vectors $\vec{x}$ and $\vec{y}$ on the multi-dimensional space, their euclidean distance is defined as

$$||\vec{x} - \vec{y}|| = \sqrt{\sum_i (x_i - y_i)^2} = \sqrt{\sum_i x_i^2 + \sum_i y_i^2 - 2\sum_i x_i y_i}, \tag{4.14}$$

and their correlation coefficient is defined as

$$R(\vec{x}, \vec{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}. \tag{4.15}$$

Note that x_i and y_i are z-scores of a certain proteome/transcriptome data set, meaning that they have distributions of mean 0 and SD 1, and since the dimensionality is huge, we can safely assume that the variance $V \equiv \sum_i x_i^2 \approx N$, where N is the dimensionality. Substitution leads to

$$R(\vec{x}, \vec{y}) = 1 - \frac{||\vec{x} - \vec{y}||^2}{2N}, \tag{4.16}$$

and the actual data are consistent with the prediction of equation 4.16 (Figure 4.8).

The clustering result (Figure 4.9) shows that proteomes cluster hierarchically in a systematic, biologically meaningful manner. At the first level of the hierarchy, proteomes separate into two classes depending on the growth media: Proteomes from *E. coli* cells grown in the presence of the folA mix tend to cluster together as do those from the cells grown in the standard medium without the folA mix. At the next levels of the hierarchy, *i.e.*, within each media condition, different experimental setups cluster according to their growth rates. This result suggests a peculiar interplay of media conditions and the internal state of the cells (growth rate) in sculpting their proteomes.

To evaluate the significance of this finding, we generated hypothetical null model proteomes (NMPs) whose correlations are determined exclusively by their assigned growth rates and clustered them by applying the same Ward algorithm. In this null model, proteomes of two systems A and B

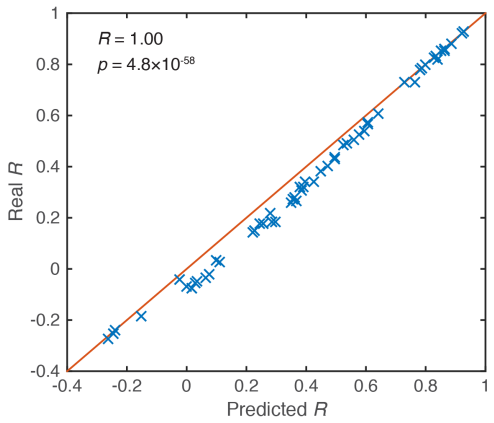


Figure 4.8: Correlation between predicted and actual correlation coefficients. The red line indicates the $y=x$ line. The prediction values are calculated according to equation 4.16, and small deviation is due to a significant difference in gene numbers of two proteomes (see Table 4.1).

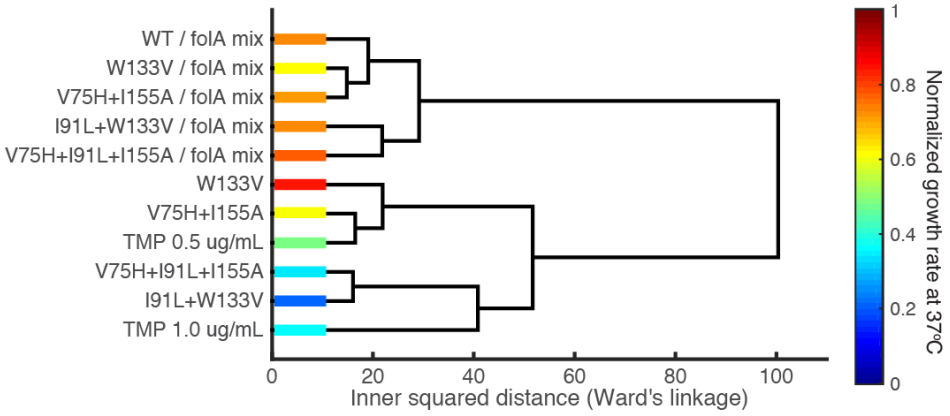


Figure 4.9: Ward hierarchical clustering of proteomes. Colors show the normalized growth rates, and the values of the horizontal axis at split points indicate Ward distances between corresponding clusters. Proteomes cluster hierarchically according to media conditions and growth rates.

are correlated according to a simple relationship

$$R_{\mathrm{AB}}^{\mathrm{null}} = \begin{cases} G_A/G_B, & \text{if } G_A < G_B \\ G_B/G_A, & \text{otherwise} \end{cases} \tag{4.17}$$

where G_i indicates the growth rate of system i. Note that when two *E. coli* systems grow at the same rate (no matter how slow or fast) their proteomes are fully correlated at $R_{\mathrm{AB}}^{\mathrm{null}} = 1$. The NMPs are generated as eleven 2000-dimensional vectors of z-scores by assigning to each NMP a growth rate of one of the 11 strains and conditions probed in this analysis and randomly generating each z-score to satisfy the correlation condition 4.17. To that end, we applied a Metropolis Monte-Carlo algorithm where each step corresponds to replacement of an individual z-score component in one of 2000-element z-score vectors by a random number drawn from the Gaussian distribution with mean 0 and SD 1, and accepting or rejecting the change according to the Metropolis criterion with the following "Hamiltonian":

$$\mathcal{H}\left(\{R_{ij}\}, \{R_{ij}^{\mathrm{null}}\}\right) = \frac{1}{2} \sum_{i \neq j = 1}^{11} \left(R_{ij} - R_{ij}^{\mathrm{null}}\right)^2, \tag{4.18}$$

where i and j indicate the indices of two hypothetical proteomes, R_{ij} is the current correlation coefficient of z-score vectors between NMPs i and j, and R_{ij}^{null} is the "target" correlation coefficient based on strains growth rates as given by equation 4.17. Iterating the MC steps with the Metropolis temperature $(k_{\mathrm{B}}T)_{\mathrm{null}} = 10^{-5}$, we obtained the NMPs where pairwise correlations satisfy the condition 4.17 with better than 5 % accuracy and clustered them using the same Ward clustering to obtain the control tree (Figure 4.10). We stochastically generated numerous NMPs and found the same tree for each realization.

The NMP tree in Figure 4.10 is qualitatively different from the real data (Figure 4.9), thereby rejecting the null hypothesis that the growth rate is the sole determinant of the correlation between the proteomes. The differences between the real proteomes and NMPs are further highlighted by the observation that real proteomes cluster hierarchically while NMPs do not. Each branch point on the tree represents the root of a cluster, which has two properties, the Ward distance at the branch point (*i.e.*, branch point on the x-axis coordinate) and the number of leaves – the number of proteomes that belong to it. For hierarchical trees, these two properties are correlated, while for simple trees, they are not. Indeed, the analysis shows that real proteomes cluster hierarchically, while NMPs do not (Figure 4.11).

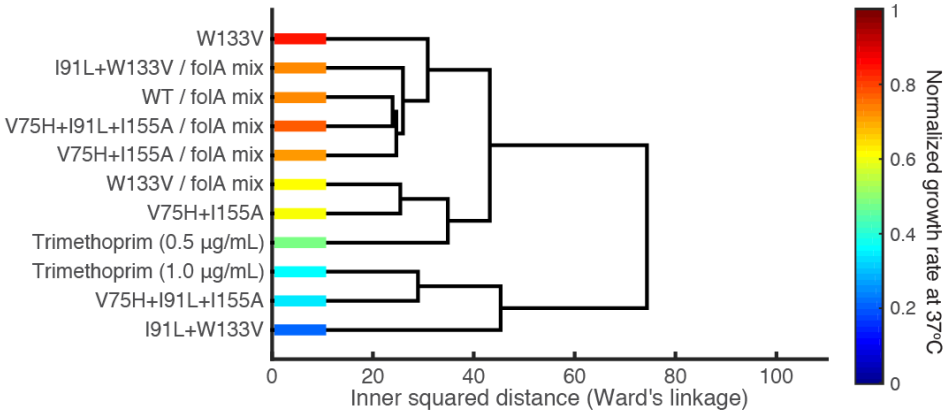


Figure 4.10: Ward hierarchical clustering of NMPs generated to correlate according to growth rates. The hierarchical structure shown in Figure 4.9 is lost.

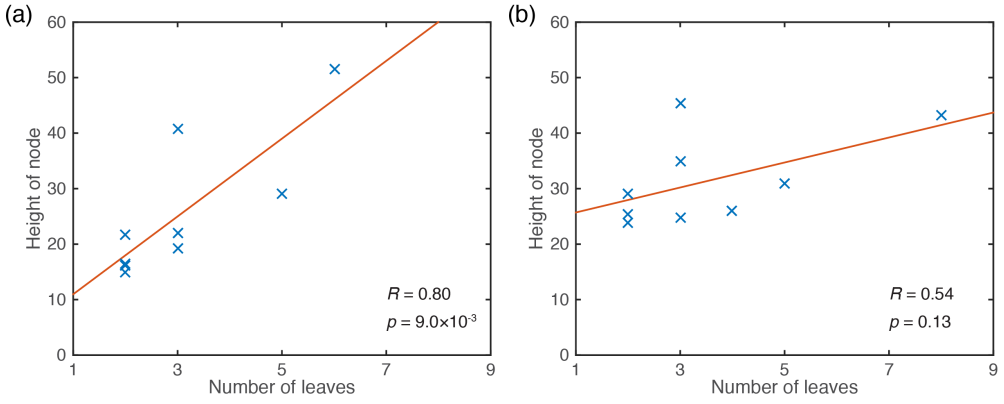


Figure 4.11: Correlation between the Ward distance at each branch point of a cluster and the number of members of the cluster for (a) real proteomes and (b) NMPs. The red lines indicate the regression lines.

4.6 Specific Effects of Perturbations on Functional Groups

To study specific biological processes rather than the whole proteome, we grouped the genes into 480 overlapping functional classes introduced by Sangurdekar *et al*[111]. We evaluated the cumulative z-score z_c^p as an average among all proteins belonging to functional class c at specific perturbation p. A large absolute value of z_c^p indicates that LRPAs or LRMAs for all proteins within the functional class in concert shift up or down due to the perturbation.

We focused on several interesting functional groups of genes, especially the ones that show statistically significant shifts in the whole proteome or transcriptome. The statistical significance p-value, that show whether a group of genes is significantly up-regulated or down-regulated (either in the proteome or transcriptome), can be estimated based on a simple null model of independence of LRPAs or LRMAs of genes within a class. In this null model, each z_c^p represents an average of a large number of independent random numbers, and the expected distribution of z_c^p is Gaussian, according to the central limit theorem:

$$P_{\text{null}}(z_c^p) = \frac{1}{\sqrt{2\pi N_c}} \exp\left[-\frac{N_c (z_c^p)^2}{2}\right], \tag{4.19}$$

where N_c is the number of genes in functional class c. Therefore, the probability, or p-value, of observed $|z_c^p|$ under the null hypothesis is

$$p\text{-value}(|z_c^p|) = 2\int_{|z_c^p|}^{\infty} P_{\text{null}}(z)dz = \sqrt{\frac{2}{\pi N_c}} \int_{|z_c^p|}^{\infty} e^{-N_c z^2/2} dz, \tag{4.20}$$

where the factor of 2 accounts for the possibility of both up- and down-regulation of genes from the given functional group. This value can be numerically calculated by using the error function.

Figure 4.13 at the end of this chapter shows the p-values for variation of average LRPA/LRMA for genes grouped by function (upper) and by operon (lower). Besides shifts in *folA* expression and DHFR abundances, significant variations were found for many important functional groups of genes (Figure 4.13, upper) First, the genes responsible for motility shut down across the mutant strains with a concomitant drop in their protein abundances (see the *fliA* operon in Figure 4.13, lower). Interestingly, the addition of the folA mix completely reverses this trend (except partial reversal for the I91V+W133V mutant). Also, while a broad set of SOS response genes is transcriptionally up-regulated (in contrast to the RpoS-regulated subset of stress-induced genes), the protein abundances of these gene products are highly elevated only in the slowest growing strains, I91L+W133V and V75H+I91V+I155A. Addition of the folA mix alleviates the SOS response in all

strains. Moreover, TMP does not trigger the SOS response at either 0.5 or 1.0 ug/mL, nor does it trigger DNA repair genes. Possibly, the depletion of precursor purines and pyrimidines might not lead to overall DNA damage that triggers the SOS response. Expression of genes belonging to the pyrimidine biosynthesis pathway is significantly up-regulated, but the abundances of their protein products drop in all strains, with the most significant impact on the slower growing I91L+W133V and V75H+I91V+I155A strains and WT strain treated with a high concentration of TMP. Addition of the folA mix again reverses this proteomic trend, giving rise to increased abundances of all the gene products belonging to this pathway.

Additional systematic insights come from the analysis of the variation of genes grouped by common transcriptional units regulated by operons (Figure 4.13, lower). For example, the genes responsible for the uptake of ferric ions (under the Fur regulator) exhibit major transcriptional down-regulation and a concomitant drop in protein abundance. For some genes, however, variations of transcript numbers and protein abundances do not exactly go hand in hand. For example, arginine catabolism genes (ArgR operon) are transcriptionally up-regulated, but their protein abundances significantly drop in the mutant strains in the M9 medium and slightly drop in the presence of the folA mix. This effect is probably common to the genes in the nitrogen metabolism pathway, as seen for the RpoN and NtrC operons. Other pathways like catabolite activation (CRP) and fumarate/nitrate reduction (FNR) show concerted transcriptome and proteome changes (up-regulation in both cases) for the *folA* mutants that moderately affect growth rates (W133V and V75H+I155A). However, a reversal of this trend is observed for the mutants that exhibit severely compromised growth (V75H+I91L+I155A and I91L+W133V), where the abundances of CRP- and FNR-regulated proteins drop significantly. An interesting insight comes from the analysis of RpoS-dependent genes. It has been shown that the phosphorylated response regulator ArcA is a direct suppressor of RpoS transcription[96]. Indeed, we observed transcriptional up-regulation of ArcA and down-regulation of RpoS. However, at the proteome level, there is down-regulation of ArcA for the V75H+I91L+I155A and I91L+W133V strains, while there is a small but noticeable increase in the abundance of proteins controlled by RpoS for the same mutants. This also holds true for the WT treated with a high concentration of TMP.

4.7 Conclusion

Quantitative transcriptomics and proteomics are powerful tools in systems biology. They have been widely used to analyze systems-level changes associated with disease phenotypes in mammalian cells[134]. Other applications include the study of the systems-level response to major perturbations

such as whole genome duplication[36], osmolarity and oxidative stresses[90,134], and loss of function mutations in the RNA degradosome in *E. coli*, which affect global RNA turnover and regulation[10,149]. Also, quantitative proteomics was used to explore the general relationship between cellular proteomes and growth rates[20,53,117]. In particular, Geiler-Samerotte and collegues established the relationship between growth rates and the total numbers of soluble and insoluble proteins in yeast[53]. In contrast to earlier studies, the focus of the present work is on the systems-level proteome and transcriptome response to the minimal and most fundamental genetic perturbations – missense point mutations introduced through genome editing into a core metabolic enzyme.

A popular conceptual view in systems biology postulates that modularity and stability of transcriptional networks had evolved to confer robustness to biological systems[19,136]. In particular, an effect of a point mutation in a robust biological system should be limited to genes and their protein products that physically, genetically, or metabolically interact with a perturbed protein. However, we found that local perturbations of DHFR function reproducibly affect transcription and protein abundances of a huge number of genes that are apparently unrelated to the folate pathway, which highlights a highly pleiotropic systems-level effect of mutations in DHFR. A detailed analysis of gene groups provided a rationale for some but not all of these shifts. All mutant and TMP-treated WT strains shut down motility, presumably as a way to conserve resources. However, for many pathways, an intuitive explanation of the changes is not obvious. For example, the genes responsible for nitrogen metabolism and ferric ion uptake are significantly affected. Moreover, for these genes, mRNA and protein abundances change in the opposite directions in a statistically significant way, indicating the importance of regulation at the level of protein turnover. Another striking example of the turnover effect is DHFR itself. Both destabilizing DHFR mutations and TMP treatment caused activation of the *folA* promoter, but the abundance of DHFR proteins increases only upon TMP treatment. Up-regulation of the gene does not save the destabilized mutants. This effect can be attributed to protein quality control, which detects and degrades partly folded mutant DHFR[12]. It should be noted that the overall increase in DHFR abundance upon TMP treatment cannot alleviate the detrimental fitness effect of TMP; the number of active DHFR molecules would still decrease upon addition of TMP due to the inhibition of DHFR by the antibiotic.

The key finding of this study is that point mutations in an essential enzyme have a profound pleiotropic effect extending to the level of the whole proteome and transcriptome. Moreover, the SD of an LRPA/LRMA distribution appears to provide a reliable global quantification of the degree of the pleiotropic effects associated with the corresponding perturbation. “Narrow” distributions (low SD) indicate that the mutations do not induce widespread systems-level perturbations and their fitness effects are minimal, whereas “wide” distributions (high SD) reveal a comprehensive

systems-level response with ensuing pronounced fitness effects. While we do not have a full mechanistic explanation for this finding, some reasons can be speculated. In particular, we note that partial loss of DHFR function has a profound effect on the pool of cell metabolites[77]. Such a global change may affect biophysical properties (such as stability and binding affinity) and the ensuing degradation rates of multiple proteins, thus causing changes in the protein turnover balance. Indirect support for this view comes from the hierarchical clustering of proteomes, which shows that media composition rather than mere growth rate determines the crucial segregation between proteomes at the top of the hierarchy. Mutations in DHFR cause a domino-like effect leading to transcriptional activation of the *folA* gene, the changes in abundance for the whole *E. coli* proteome, and finally, changes of growth rates in the perturbed cellular systems. The quantitative measures of these effects on all scales strongly correlate, suggesting the existence of a common underlying cause that drives these changes. Future studies will reveal the existence and exact nature of this cause.

Figure 4.12 ***(following page):*** Distributions of LRPAs and LRMAs for all strains and conditions for which MS data were obtained. Three repeats for the mutant strains and the WT strain treated with TMP 0.5 ug/mL are shown.

Figure 4.12: (continued)

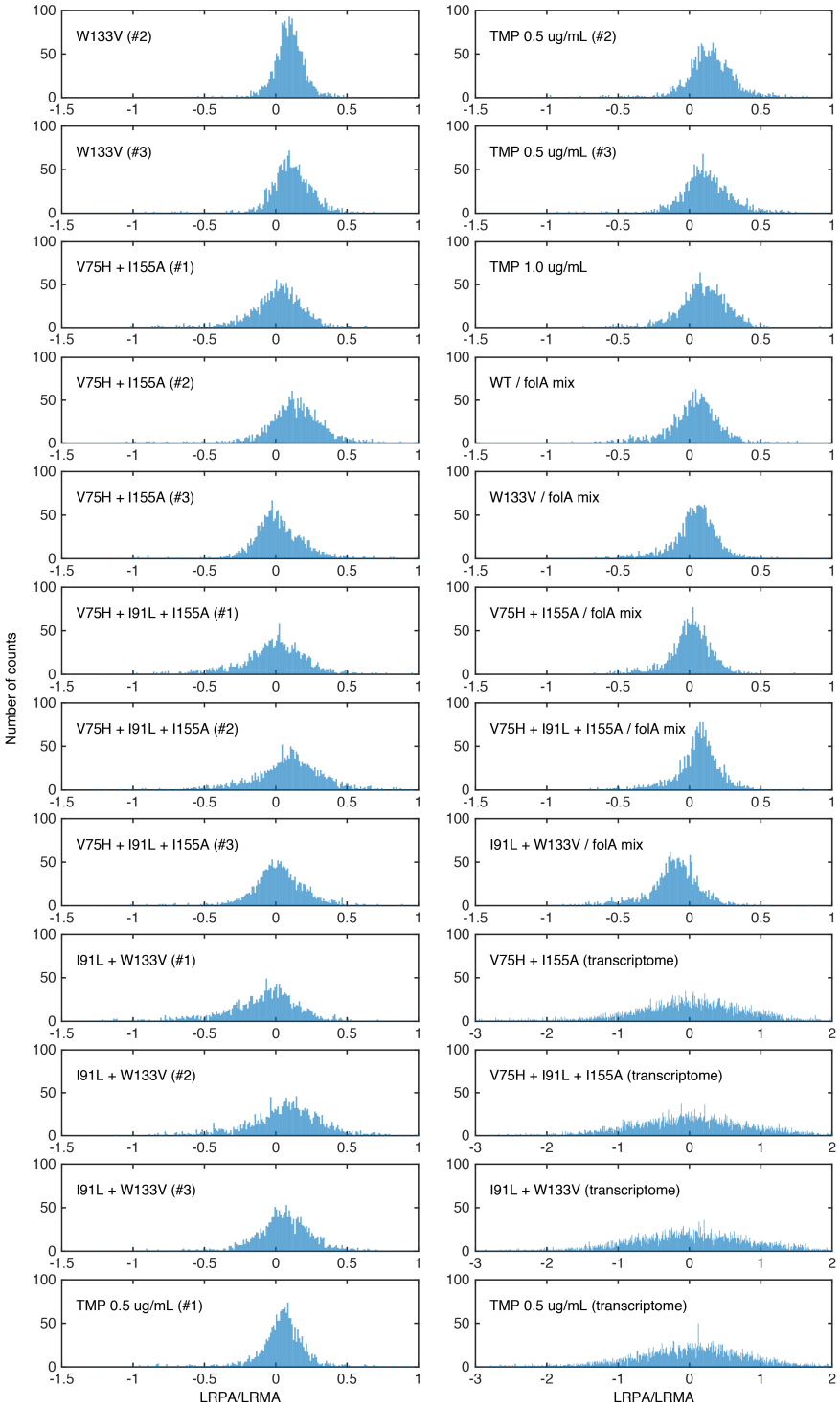
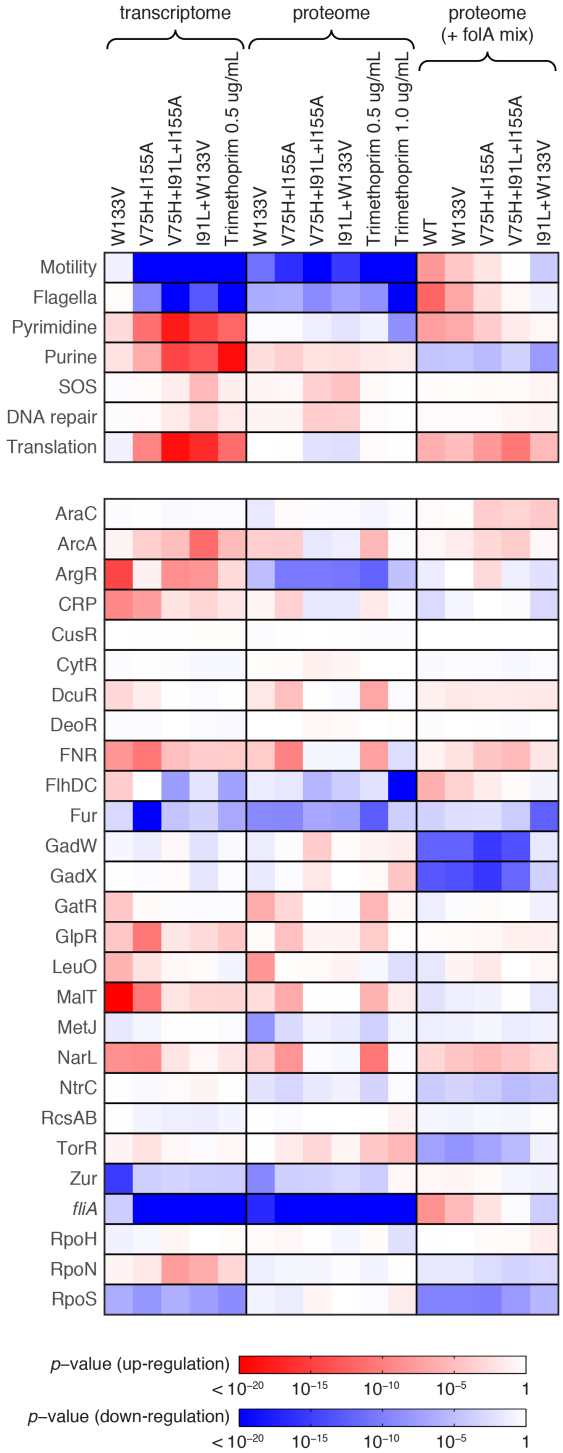


Figure 4.13 ***(following page):*** Statistical significance of variation for specific gene groups. Clustering global variation in mRNA and protein abundances as belonging to functional classes (upper) or co-regulated by a specific operon (lower) reveals the highly statistically significant variation of several functional groups. The color code indicates the direction of change (blue: down-regulation, red: up-regulation), and the color depth indicates the logarithm of p-values against the null model of independent variation within a group of genes.

Figure 4.13: (continued)



This page intentionally left blank

5

Response and Adaptation of *E. coli* to Horizontal Gene Transfer

5.1 Background

Horizontal gene transfer (HGT) is a major force in bacterial evolution[71,81,92]. Comparative genomic analyses show that HGT events can be broadly classified into three types: a) acquisition of a new gene not present in the taxa, b) acquisition of an orthologous gene in addition to the endogenous chromosomal copy, and c) direct chromosomal replacement of a gene by its ortholog from other species (also known as xenologous horizontal gene transfer)[71]. Koonin *et al.* also found that all three HGT types are approximately equally common and represent an efficient mechanism for rapid evolution and/or adaptation to new niches[71].

The genetic mechanisms responsible for the horizontal transfer of foreign genes (*i.e.*, transformation of naked DNA, conjugation, and viral transduction) are well characterized[55,105,126]. However, the material transfer of DNA from other bacterial species is only an initial step. The evolutionary fate of an HGT event (fixation, elimination by purifying selection, or persistence as a subdominant clone) depends on the fitness benefit or cost of the newly acquired gene. Previous studies on these fitness effects have arrived at apparently inconsistent conclusions. For example, Sorek *et al.* expressed multiple proteins from 79 prokaryotic genomes in an expression vector under control of an

inducible promoter and measured the ensuing fitness effects in *E. coli*. They found that expression of many foreign proteins is detrimental to the *E. coli* host and attributed the fitness cost to a gene dosage-related toxicity[118]. Lind *et al.* found that inter-species chromosomal replacement of three native genes encoding ribosomal proteins in *S. typhimurium* was detrimental to fitness, apparently due to low expression of transferred proteins[84]. Although these studies showed that many HGT events incur fitness costs, they did not provide mechanistic or molecular explanations of why this was the case. Meanwhile, other studies have argued that HGT is predominantly neutral rather than deleterious. Insertion of random DNA fragments from other bacteria in the Salmonella chromosome showed no significant fitness effect for about 90 % of the inserts[70]. Introduction of foreign and complex subunits in *E. coli* also showed no loss in fitness[141].

The apparent controversies on the nature of the fitness landscape of HGT events can be attributed to several challenges:

1. Pleiotropy at the molecular level. The starting genetic material has a broad distribution of molecular and sequence properties that are not entirely independent (*e.g.*, potential effect of GC-content on RNA stability[74] that could affect transcription/translation, and of protein folding stability and activity[16] that could affect function).

2. Pleiotropy at the cellular level. Beyond the foreign gene's immediate functional context, HGT may affect or be affected by other cellular factors, such as protein-protein, metabolic, or regulatory interaction networks[103,34,83,147]. Another example is a protein homeostasis (proteostasis) machinery, which maintains the integrity of the proteome through assisted folding and degradation and is known to buffer against the deleterious effects of mutations[107,130]. However, the actual effect of proteostasis on horizontal gene transfer is not yet known.

3. Time and length scales in evolution. Similarly to mutations, HGT events can be accompanied by immediate and transient responses of the cell that are particularly hard to detect using comparative genomics, because it analyzes HGT that has survived selection over long evolutionary time scales.

Altogether, these challenges need to be addressed to understand the fitness landscape of HGT and the cellular responses that lead to the subsequent accommodation or rejection of a foreign gene.

In this chapter, we sought to develop an experimental system that allows full control over the molecular properties of the transferred gene (Figure 5.1a). Our focus is on the functional barriers to HGT emerging at the protein level rather than genomic barriers affecting transcription and translation. To this end, we used the essential gene *folA* encoding dihydrofolate reductase (DHFR) as a

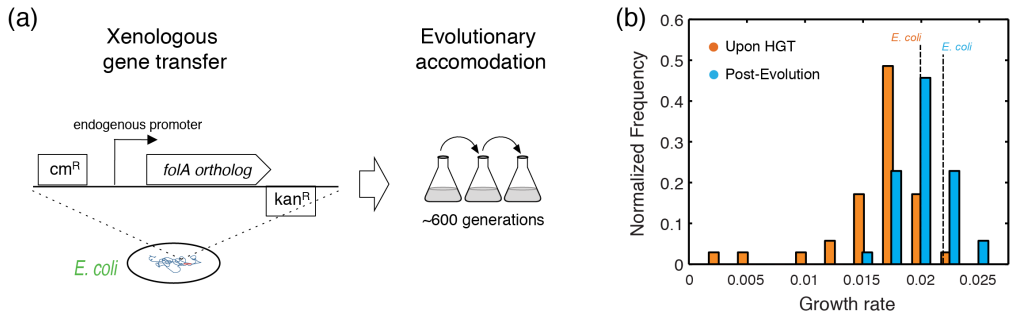


Figure 5.1: Experimental scheme for mimicking HGT and its consecutive evolution with the resultant growth rate distribution. (a) The open reading frame of the *folA* gene encoding DHFR in the *E. coli* chromosome is replaced with orthologs from 35 other mesophiles, while preserving the endogenous promoter. The strains carrying the orthologous DHFR replacements are evolved for 31 serial passages (approx. 600 generations) under standard conditions. (b) Distribution of the growth rates before (immediately upon HGT) and after evolution experiment. The growth rates of the WT strain with *E. coli* DHFR are indicated by dashed lines. The Kolmogorov-Smirnov (KS) test indicates that pre- and post-evolution populations are significantly different in terms of their growth rates (p-value < 10^{-10}). While 31 out of the 35 naive strains (88 %) have lower growth rates than WT, 30 % of the post-evolution strains have higher growth rates than WT.

model. DHFR catalyzes an electron-transfer reaction to form tetrahydrofolate, a carrier of single-carbon functional groups utilized in central metabolism, including *de novo* purine biosynthesis, dTTP formation, and methionine and glycine production[112]. DHFR is also an important target of antifolate therapy by trimethoprim (TMP), a competitive inhibitor that binds with high specificity to the active site of bacterial enzymes[22]. Additionally, comparative genomics studies have demonstrated that HGT plays an important role in the evolution of the folate metabolic pathway, including the spread of antifolate resistance[35,121]. Moreover, DHFR is an essential enzyme in *E. coli* with a relatively low basal expression level (approximately 40 copies per cell on average[124]), and its activity is linked to bacterial fitness in a dosage-dependent manner[147]. As such, DHFR is a convenient model to study the HGT-related fitness effects. We experimentally mimicked multiple HGT events by replacing the *folA* gene on the *E. coli* chromosome with its orthologs from 35 phylogenetically diverse mesophiles. This collection of orthologs explores a broad distribution of protein sequences and biophysical properties.

5.2 Growth Rates Before and After Evolution

We initially identified 290 orthologous DHFR sequences from mesophilic bacteria and selected 35 diverse sequences with amino acid identity to *E. coli* DHFR ranging from 29 % to 96 % (Figure 5.2). First, we sought to minimize the contributions from confounding factors that mostly affect

transcription and translation of replaced genes, such as GC content[101], codon-usage pattern[94,132], specific loci at which chromosomal incorporations occur, and the copy number of the transferred genes[126,74]. The amino acid sequences of the chosen 35 orthologous DHFRs were converted into DNA sequences using the codon signature of *E. coli*'s *folA* gene. We used the λ-red recombination system[33] to replace the open reading frame (ORF) of *folA* with the synthetic DNA sequences, while preserving *E. coli*'s wild-type *folA* promoter. Thus, the resulting 35 strains carrying the orthologous DHFR gene replacements are identical with respect to the chromosomal location of the *folA* gene and the mode of regulation of their DHFR expression. In addition, they have similar GC content and codon usage signature.

We assayed the fitness of the resulting HGT strains by measuring their growth rates at 37 °C (this condition was consistently used throughout the work). As shown in Figure 5.1b and Figure 5.2, We found that *E. coli* fitness (here and below we use the terms fitness and growth rate interchangeably) is very sensitive to the orthologous replacements of its DHFR. Growth rates are lower than wild-type (WT) *E. coli* in 31 out of 35 strains, with six strains (DHFR-23, 35, 36, 37, 38 and 43; highlighted in Figure 5.2) exhibiting a severe fitness loss of 70-85 %. DHFR-21 (from *W. paramesenteroides*) did not grow at all under the conditions of the experiments. Surprisingly, we found no significant correlation between growth rates of the HGT strain and the evolutionary distance between DHFR orthologs, measured as % of amino acid identity relative to *E. coli* DHFR (Spearman $R = 0.16$; p-value = 0.4) (Figure 5.3), thus, challenging the notion that sequence similarity between endogenous and transferred genes facilitates horizontal gene transfer.

The high fitness cost of the orthologous replacements of *E. coli* DHFR demonstrates the existence of a molecular constraint ("a barrier") to HGT. To determine whether the evolutionary process can traverse this barrier, we conducted high-throughput serial passaging of the HGT strains (Figure 5.1a). Overall, we performed 31 passages which amount to approximately 600 generations for the WT strain. Growth rate measurements after the evolution experiment show that orthologous strains have substantially improved their fitness (Figure 5.2). Moreover, about 30 % of the strains grew as well as or better than WT after the evolution experiment (Figure 5.1b and 5.2). The improvement in growth rates was especially dramatic among strains that experienced the most severe fitness loss upon HGT (DHFR-23, 35, 36, 37, 38 and 43; highlighted in Figure 5.2). Thus, the molecular constraints to horizontal transfer of the DHFR coding genes were largely alleviated during experimental evolution.

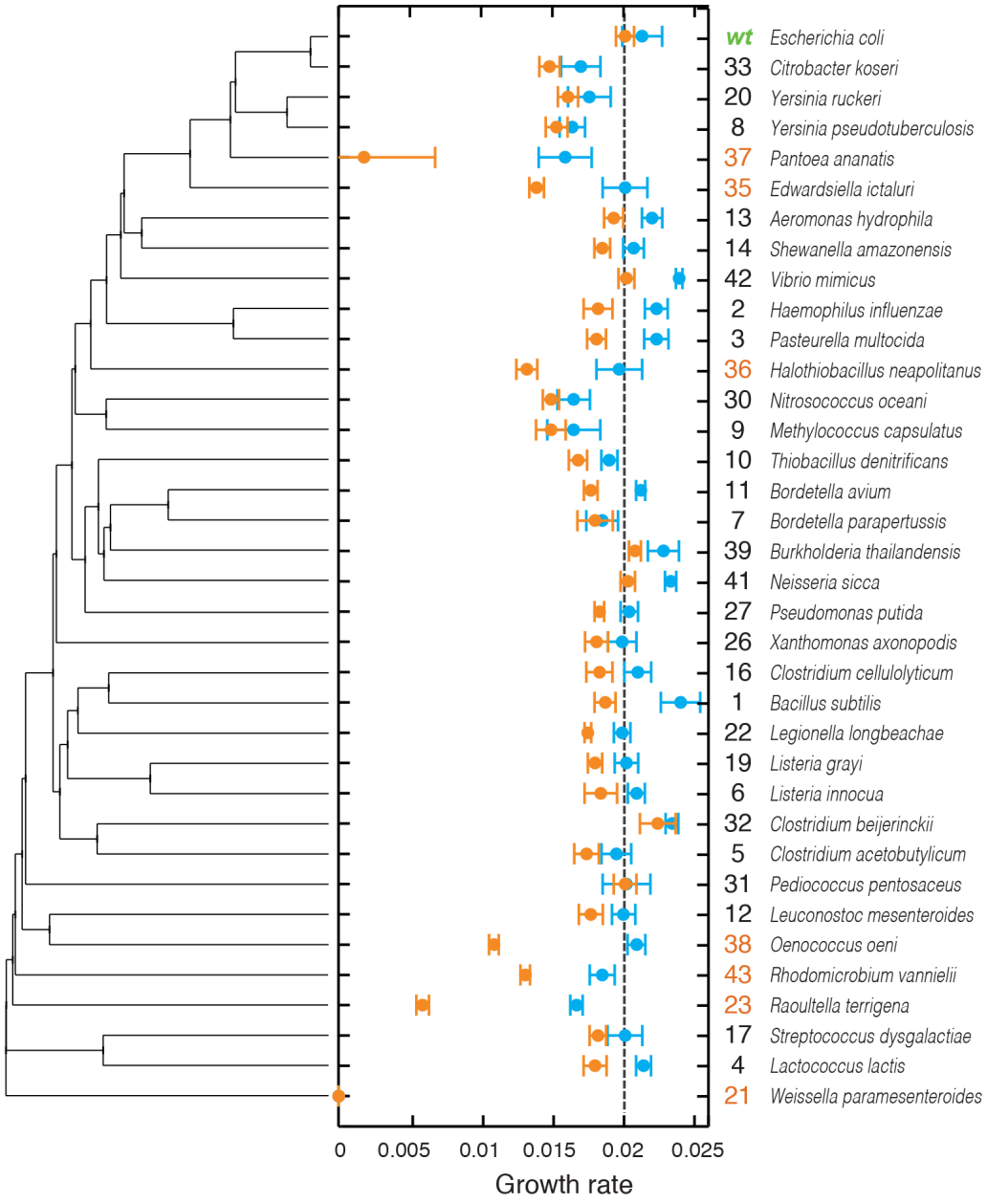


Figure 5.2: Growth rates before and after evolution for each strain as a function of its DHFR's position in the phylogeny. Color scheme is similar to Figure 5.1b. Strains are sorted according to the phylogenetic tree on the left. On the right we show an ID number for each strain (used throughout the text) and the original species carrying the DHFR ortholog. We highlight in orange the ID numbers of strains that experience severe fitness drop (30 % and lower) upon DHFR replacement. Error bars represent standard deviations of 4 independent measurements.

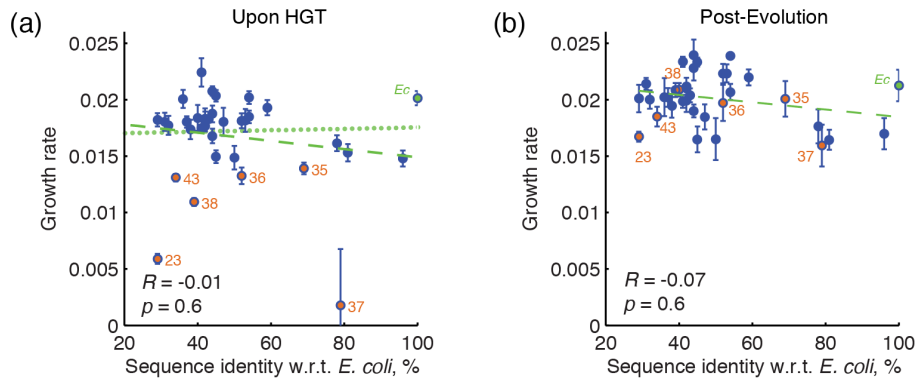


Figure 5.3: Evolutionary distances of orthologous DHFRs, measured as % of amino acid sequence identities with respect to WT *E. coli* DHFR. Sequence identities does not correlate with growth rates (a) before or (b) after evolution (Spearman R and p-values are indicated). *Ec* (green) denotes the WT *E. coli* strain. Strains with severe fitness effects are colored in orange. Dashed green lines are regression fits to all points. Excluding DHFR-37 in (a), there is still no correlation between identify and growth rate (R = 0.06, p-value = 0.72; dotted line is the regression fit excluding DHFR-37). Error bars represent standard deviations of 4 independent measurements.

5.3 Expression Levels

To determine how the effect of HGT percolates throughout the entire *E. coli* proteome, we analyzed the systems-level effect of inter-species DHFR replacements before and after the evolution experiment. To that end, we quantified relative (to WT) abundances of approx. 2000 proteins in the cytoplasm using tandem mass tags (TMT) with subsequent LC-MS/MS analysis, as described in chapter 4. We picked five strains for proteomic characterization based on their fitness effect upon HGT (Figure 5.1b): DHFR-23, 35 and 38 (severely deleterious); DHFR-22 (mildly deleterious); and DHFR-39 (beneficial). For reference, we compared the proteomic effects of orthologous replacements with the proteomic effect of treating *E. coli* with 1 ug/mL of trimethoprim (TMP).

Following the analysis in chapter 4, we first checked the correlation between growth rates and LRPA SDs. As shown in Figure 5.4, the correlation holds for most systems, except the DHFR-23 system upon HGT that shows a dramatic drop in growth rate but a mild perturbation level measured by the SD. Other nine systems, regardless of whether they are upon HGT or post-evolutionary, follow the same trend well. Hence, we confirm that if a cell needs to change expression levels of its genes more radically to buffer the impact of perturbation, the reduction of growth is larger, implying that the unperturbed *E. coli* system has an optimized cellular network to maximize its growth rate.

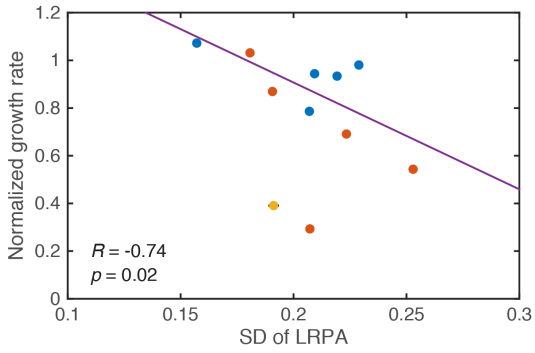


Figure 5.4: The SD of LRPA distribution is anti-correlated with the growth rate. Error bars on the y-axis correspond to the SDs of four independent experiments. No error bar indicates that the corresponding SD is negligible. Red dots indicate the data upon HGT, blue dots indicate the post-evolution data, and a yellow dot indicates the TMP data. The purple line refers to the regression line, except the outliers (DHFR-23 upon HGT and TMP-treated WT).

5.4 Similarities of Proteomes

For each of the > 2000 proteins detected, we quantified their enrichment using the logarithms of relative protein abundances (LRPA) that are expressed as z-scores (see chapter 4). In Figure 5.5, we show the correlation plots of the z-scores between proteomes. As expected for DHFR-35 and 38, where HGT is severely deleterious, their proteomes strongly resemble the proteome of TMP-treated WT strain (R = 0.42, p-value = 9.5×10^{-74} and R = 0.50, p-value = 7.1×10^{-105}, respectively). This result suggests that the systems-level response to HGT of the DHFR genes is akin to response to inactivation of the endogenous DHFR protein by TMP. Interestingly, despite significant evolutionary distance between the DHFR alleles from strains DHFR-35 and 38 (Figure 5.2), the correlation between their proteomic profiles is significant (R = 0.84, p-value < 10^{-300}). However, the proteome of DHFR-23, another orthologous strain with a severely reduced growth, was not similar to the TMP-treated WT proteome (R = 0.07, p-value = 0.0041), suggesting that, at least for some strains, the systems-level response to the partial loss of DHFR function follows a different pattern. The proteome of DHFR-22, a strain with moderately reduced fitness, was much less similar to the proteome of TMP-treated WT (R = 0.30, p-value = 8.6×10^{-35}). The proteome of DHFR-39, one of the few strains that grew better than WT upon HGT, bore no resemblance to TMP treatment (R = -0.05, p-value = 0.026).

After the evolution experiment, the proteomic profiles of the strains lose their resemblance to TMP-treated WT cells (Figure 5.5), which reflects the alleviation of the detrimental effects of HGT. Additionally, after the evolution experiment the proteomic profiles of the strains become more sim-

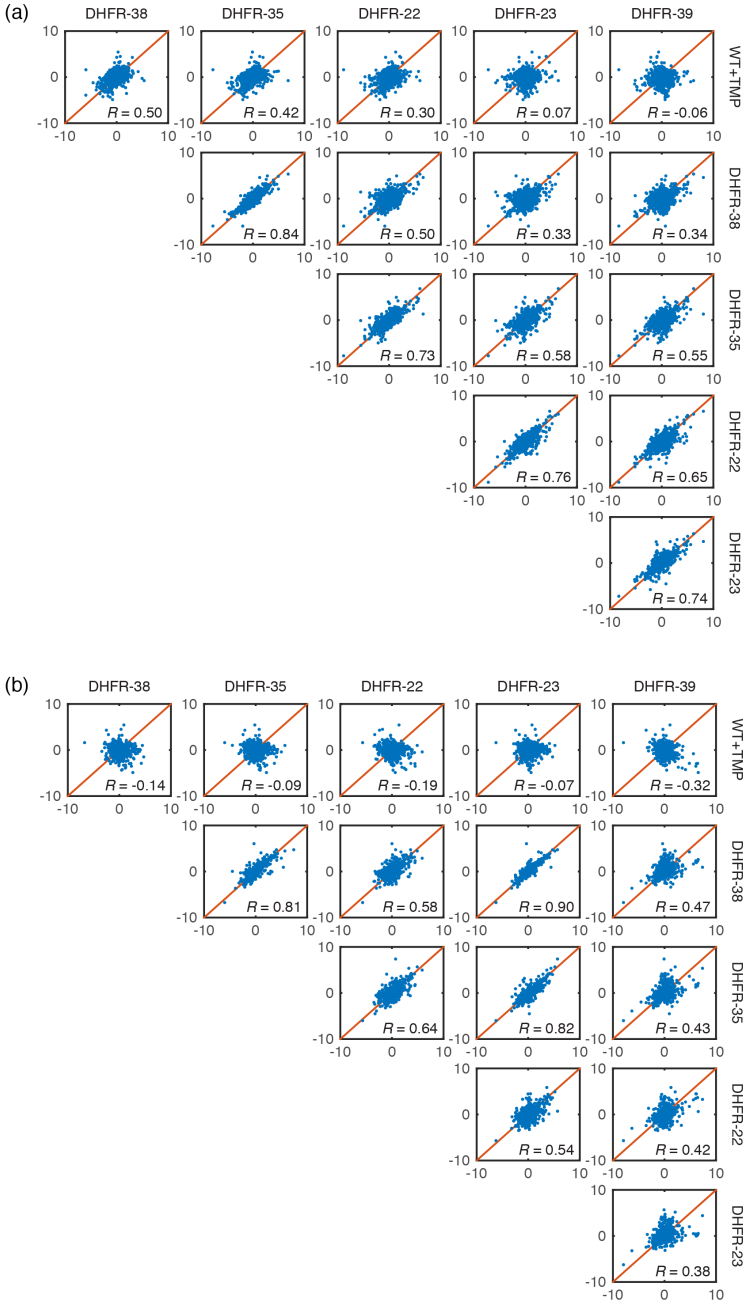


Figure 5.5: z-scores correlation plots between proteomes of indicated DHFR orthologous strains (a) upon HGT and (b) after evolution and WT strain treated with 1 ug/mL trimethoprim (TMP). The strains are representative of the fitness effects upon HGT: DHFR-23, 35 and 38 are severely deleterious; DHFR-22 is mildly deleterious; and DHFR-39 is beneficial (Figure 5.2). Global proteome quantification was obtained using LC-MS/MS analysis of TMT-labeled proteomes (see chapter 4). The strains are sorted left to right according to decreasing similarity of their proteomes upon HGT with that of TMP-treated WT cells.

ilar (the increased correlations in inter-strain comparison in Figure 5.5). In particular, although the proteome of DHFR-23 is barely similar to DHFR-35 and 38 before the evolution experiment, it becomes much more similar to DHFR-35 and 38 after it.

5.5 Functional Pathways and Operons

Next, we carried out a comparative analysis at the level of functional pathways and operons. Using the functional and regulatory classification of genes by Khodursky and co-workers[111] (see chapter 4), we collected the gene groups that show significant regulations upon HGT (Figure 5.6a) and after evolution (Figure 5.6b), compared to WT. As shown in chapter 4, upon a perturbation, cells first reduce expression of the motility-related gene groups (motility, flagellum, flagella biosynthesis, and *fliA* groups in Figure 5.6a). However, DHFR-23 and 39, whose proteomes show little similarities to that of TMP-treated WT, regulate those groups in the opposite direction (up-regulation). Also, upon severely deleterious point mutations, *E. coli* cells show down-regulation of translation-related genes (see chapter 4), but it is apparently not the case for DHFR-22, 23, and 39. Rather, DHFR-23 and 39 showed strong up-regulation of translation-related genes. However, after serial evolution, this strong distinction is lost (Figure 5.6b); DHFR-38, 35, and 22, which drastically repress expression of motility-related genes, now show up-regulation of those genes. Note that the proteomic responses of DHFR-35 and 38 upon HGT and post-evolution are similar despite their evolutionary distance (see Figure 5.2).

To systematically study the effect of evolution, we screened those that collectively changed their abundances significantly during the evolution experiment, by employing the Kolmogorov-Smirnov (KS) test. First, we determined the direction of adaptation by comparing the average z-values of the naive and evolved proteome sets of a given gene group. If the average z-value increases, it means that the cell evolved to increase expression levels of the gene group; otherwise, it decreased the expression levels. Then, we applied the two-sample KS test on the two sets, which provides the p-value for the null hypothesis that the two sets were drawn from the same distribution. Hence, we can quantitatively interpret a lower p-value as an indication that the two sets have more significantly different distributions of z-values.

As shown in Figure 5.7, we found that the genes responsible for cell motility show the largest increase in their abundances, which is notable, because it suggests that through the adaptation process, HGT strains have eliminated the energetic burden which caused shutting down of cell motility in the first place (Figure 5.6a). In the same vein, genes responsible for a number of metabolic processes such as synthesis of amino acids and nucleotides as well as turnover of several metals show highly

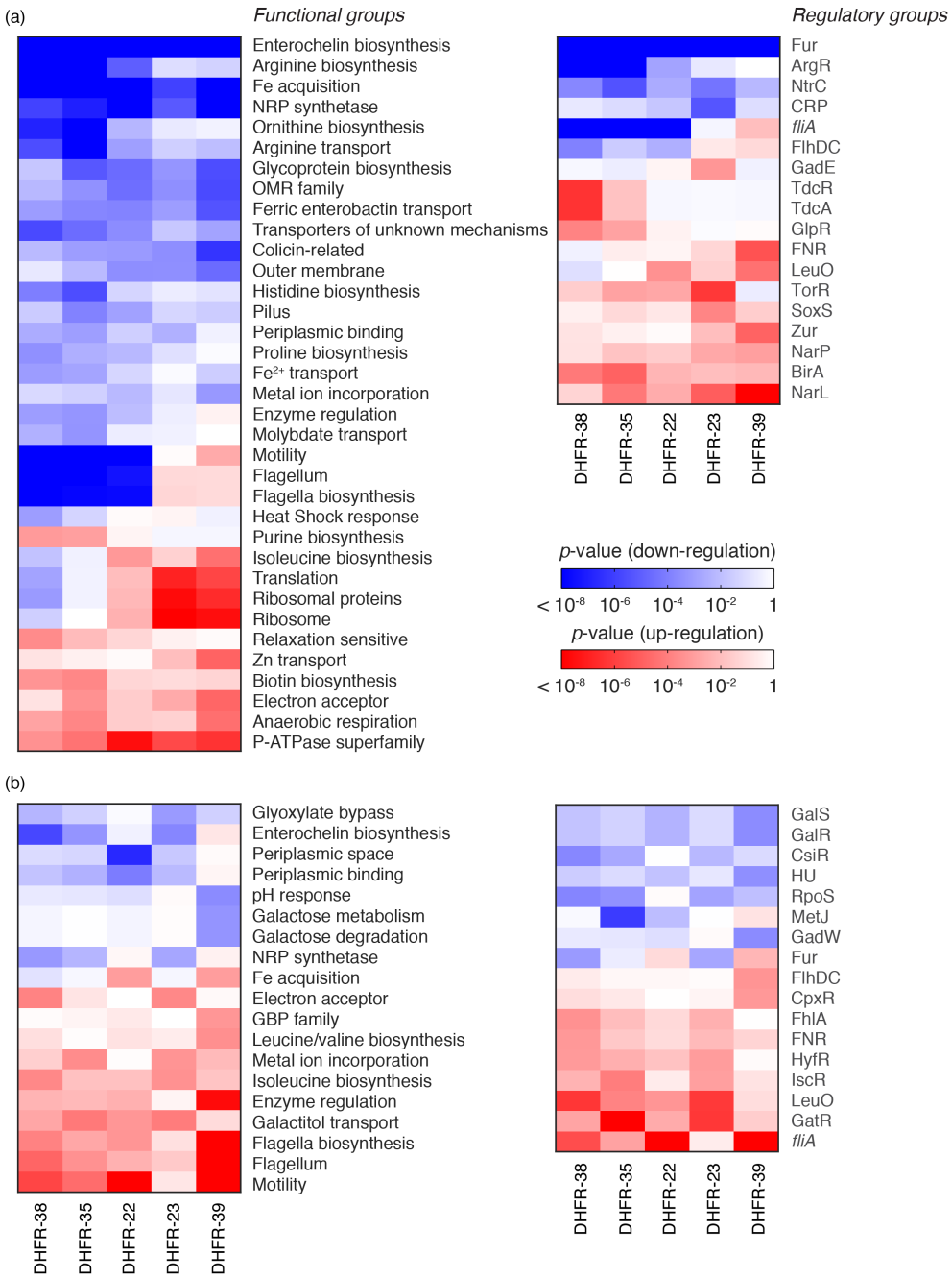


Figure 5.6: Statistical significance of variation for functional and regulatory classes of genes (a) upon HGT and (b) after serial evolution. The color code indicates the direction of change (blue: down-regulation, red: up-regulation), and the color depth indicates the logarithm of p-values against the null model of independent variation within a group of genes. We showed only the gene groups whose change upon evolution experiment is significant for any strain ($p < 10^{-3}$).

significant changes between the naive and evolved strains.

5.6 Conclusion

In this chapter, we provided systems-level insight into the impact of HGT events and adaptation of *E. coli* to overcome the impact. We focused here on xenologous transfer of DHFR coding genes, one of common modes of HGT[71]. The chromosomal gene replacements were done while preserving the endogenous promoter. Such an experimental design provided us with a direct control over conditions of expression, enabling us to focus on the link between variation of sequence and the proteome pattern due to HGT. We were able to purify 33 orthologous DHFRs and measure their growth rates, which showed most of the replacements are detrimental immediately upon HGT. However, the serial evolution experiment mostly recovers their growth rates.

The key finding of this work is that *E. coli* cells apparently discriminate between own and foreign proteins. Upon HGT, the cells response to this event as a perturbation and show immediate adjustment of protein expression levels, as they do in the case of DHFR point mutations and inhibition by TMP (see chapter 4). However, when selective pressures act on the growth rates of the cells, after several hundred generations, the cells adapt themselves to the transferred gene by rewiring their cellular networks. Restoration of growth rates show that this adaptation process is quite successful, and the proteome patterns after evolution show that the reconstructed network structures are significantly different from the original *E. coli* network structure, which is optimized to maximize the growth rate of a normal *E. coli* cell.

Besides being relevant to understanding the evolutionary dynamics of HGT, our approach is broadly applicable to the study of the genotype-phenotype relationship. While the concept of a fitness landscape is dominant in evolutionary biology, it remains highly metaphoric as its "axes" remain unlabeled. A promising approach to map fitness landscape is by introducing "bottom up," controllable genomic variations that cause known changes of the molecular properties of proteins[12,13,43,140,88,28,29,63]. However, point mutations and/or random mutagenesis are limited in their ability to generate a broad variation of catalytic activities and other physical properties of proteins. In contrast, "borrowing" highly diverged yet catalytically active orthologous proteins from other species allows us to cover a broad range of variation of molecular properties of proteins. By systematically exploring the relationship between molecular properties of xenologously replaced proteins and the fitness of corresponding cells, this approach provides an opportunity to quantitatively characterize the global properties of fitness landscapes.

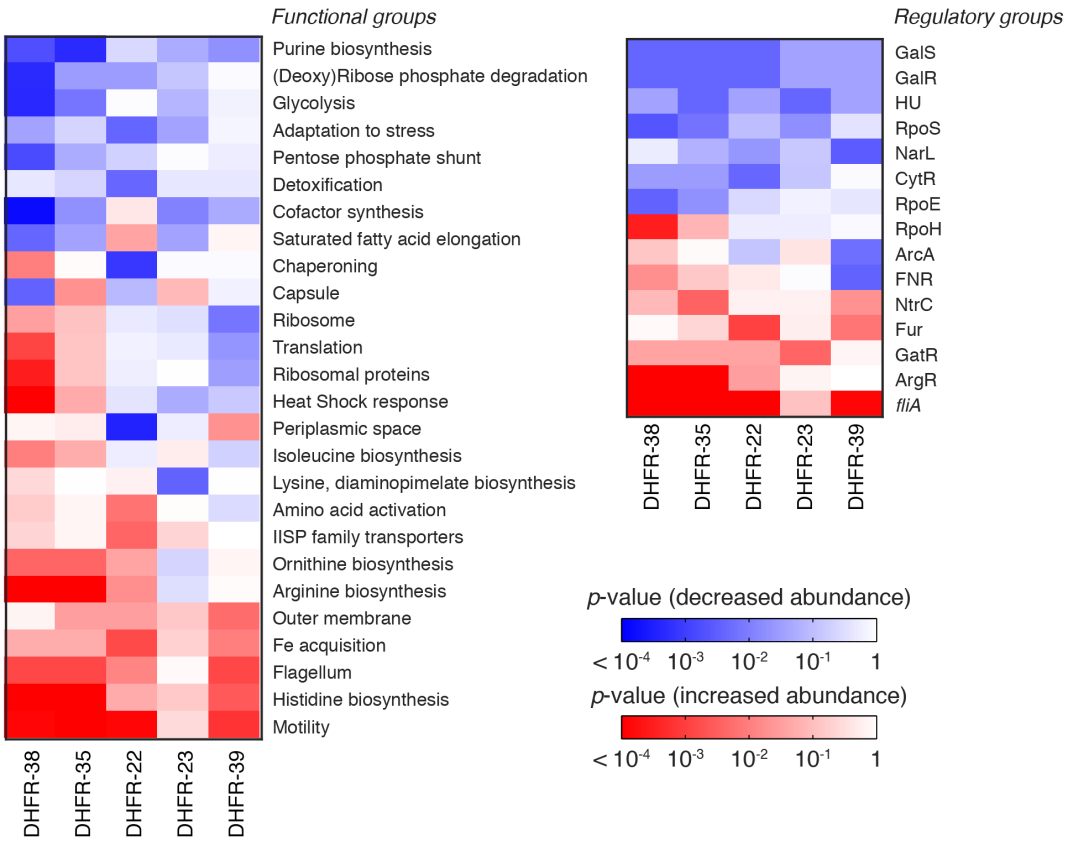


Figure 5.7: Change in global variation in protein abundances induced by experimental evolution of the indicated orthologous strains calculated for functional and regulatory classes of genes. Color code indicates the logarithms of p-values from the two-sample KS tests performed on pre- and post-evolution proteomes along with the direction of change (blue: drop in abundance, red: increase in abundance). We plotted only the gene groups whose change upon evolution experiment is significant for any strain (KS p-value less than 0.01).

6

Multi-Level Responses of Different Yeast Strains to Heat Shock Protein Inhibition

6.1 Background

Genetic variation within a species produces various strains, whose differences can be identified at the genetic, proteomic, and cellular levels. Comparative study provides insight on the detailed mechanism of how a small change at the genetic level can propagate through cellular networks to drastically different phenotypes of organisms. The primitive comparison comes from genomic sequence comparison [2,122], but there is a gap in our understanding at the systems level.

In chapter 4, we developed statistical tools to analyze the systems-level responses of an *E. coli* cell to various perturbations, and it was applied to investigate adaptation of *E. coli* to reduce the impact of horizontal gene transfer in chapter 5. These works are focused on how a perturbation affects mRNA and protein abundance profiles in a cell to change the growth rate. However, this method can be used to compare two different strains.

In this chapter, we will use two different strains of yeast *Saccharomyces cerevisiae* as a model system, and inhibit an essential heat shock protein to obtain their proteome fingerprints upon this perturbation. Then, we will use different grouping methods to extract biological meanings and to systematically analyze the difference in responses of different strains to the inhibition.

6.2 Relative Protein Abundance Distribution Statistics

The two different *S. cerevisiae* strains used in this work are a standard lab strain S288C (denoted by BY) and a vineyard isolate RM11-1a (denoted by RM)[21], which differ at approximately 0.5 % of their nucleotide sequences[62]. We treated both strains by radicicol of 5 ug/mL, which is dissolved in dimethyl sulfoxide (DMSO). Radicicol (RAD) is known as a competitive inhibitor of Hsp90 family proteins[114]. The Hsp90 family proteins, among major heat-shock proteins, play an important role as an evolutionary capacitor[109] and a hub regulator of protein homeostasis[123]. Hence, inhibition of Hsp90 is expected to have a huge impact on the proteome, leading to a drastic change in cellular traits. We prepared three independent replicates for each of RAD-treated BY and RM strains, and as a control experiment, we also prepared two independent replicates for each of BY and RM strains treated only with the solvent DMSO. We measured their relative protein abundances (RPAs) by the TMT-MS techniques used in the previous chapters, and converted each PRA into its logarithm (LRPA). The total number of detected proteins is 4,310.

As described in chapter 4, the standard deviation (SD) of an LRPA distribution (generated from comparison between two proteomes) quantitatively captures how different one proteome distribution is from another. We first compared how different the replicates are from each other, to set the natural noise level due to cell-to-cell variations and experimental errors (Figure 6.1, gray bars). The average value of the noise levels is 0.042 (Figure 6.1, gray line).

Next, we checked the difference between the systems with and without RAD, and also the difference between two different strains. The LRPA values of each gene are averaged over biological replicates to give a representative value, which was used to calculate the LRPA SD. As shown in Figure 6.1 (blue bars), the SD values are comparable to the average noise level, which implies that the global effect of RAD on each proteome is negligible. In contrast, the difference between the BY and RM strains is remarkable (Figure 6.1, red bars); the SD values are about three times as much as the noise level.

To avoid possible confusion, let us here define the 4 different LRPA values for each gene g (see

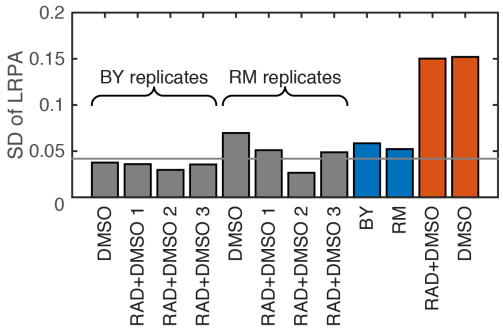


Figure 6.1: Standard deviations of LRPA distributions. Gray bars indicate comparisons between biological repeats of one strain in the same experimental condition (*e.g.*, the first bar compares two replicates of the BY strain without RAD treatment), and they provide the noise level. The average noise level is shown as a gray line. Blue bars show the effect of RAD on two strains, and red bars show the strain differences at the given treatment. For blue and red bars, the average LRPA value (over replicates) for each gene is used to calculate LRPA SDs. (For rigorous definitions of LRPAs, see equations 6.1-6.4 in the text.) The LRPA SDs between RAD-treated and untreated strains (either BY or RM) are close to the noise level (blue bars), implying that the global impact of RAD treatment on the proteome is not significant. However, the LRPA SDs between different strains (regardless of RAD treatment) are significantly higher than the noise level (red bars), and it suggests that there is a global difference between the proteomes of the two strains.

equation 4.3 in chapter 4):

$$\mathrm{LRPA}(g;\mathrm{BY}) = \log \frac{N_{\mathrm{protein}}(g;\mathrm{BY},\mathrm{RAD{+}DMSO})}{N_{\mathrm{protein}}(g;\mathrm{BY},\mathrm{DMSO})}, \tag{6.1}$$

$$\mathrm{LRPA}(g;\mathrm{RM}) = \log \frac{N_{\mathrm{protein}}(g;\mathrm{RM},\mathrm{RAD{+}DMSO})}{N_{\mathrm{protein}}(g;\mathrm{RM},\mathrm{DMSO})}, \tag{6.2}$$

$$\mathrm{LRPA}(g;\mathrm{RAD{+}DMSO}) = \log \frac{N_{\mathrm{protein}}(g;\mathrm{BY},\mathrm{RAD{+}DMSO})}{N_{\mathrm{protein}}(g;\mathrm{RM},\mathrm{RAD{+}DMSO})}, \tag{6.3}$$

$$\mathrm{LRPA}(g;\mathrm{DMSO}) = \log \frac{N_{\mathrm{protein}}(g;\mathrm{BY},\mathrm{DMSO})}{N_{\mathrm{protein}}(g;\mathrm{RM},\mathrm{DMSO})}. \tag{6.4}$$

In other words, $\mathrm{LRPA}(g;\mathrm{BY})$ represents the expression level difference of gene g between the BY strains with and without RAD, $\mathrm{LRPA}(g;\mathrm{RM})$ that of gene g between the RM strains with and without RAD, $\mathrm{LRPA}(g;\mathrm{RAD})$ that of gene g between the RAD-treated BY and RM strains, and $\mathrm{LRPA}(g;\mathrm{DMSO})$ that of gene g between the BY and RM strains grown in the media containing DMSO without RAD.

6.3 z-Score Distributions

In order to compare different proteomes, we transformed the LRPAs into their corresponding z-scores, according to the following equation (see equation 4.5 in chapter 4):

$$z(g; c) = \frac{\mathrm{LRPA}(g; c) - \langle \mathrm{LRPA}(c) \rangle_g}{\sigma_{\mathrm{LRPA}(c)}}, \tag{6.5}$$

where $\mathrm{LRPA}(g; c)$ is the LRPA of gene g with control variable c (as defined in the previous section), $\langle \mathrm{LRPA}(c) \rangle_g$ and $\sigma_{\mathrm{LRPA}(c)}$ respectively indicate the arithmetic mean and standard deviation of the LRPA distribution with control variable c. Note that z-scores are extracted from comparison between two different systems. For example, $z(g; \mathrm{BY})$ indicates the relative expression level difference of gene g between the BY strains with and without RAD treatment (equation 6.1), and $z(g; \mathrm{DMSO})$ indicates the relative expression level difference of gene g between the BY and RM strains grown in the media containing DMSO but not RAD (equation 6.4).

Figure 6.2a shows the correlation between the $z(g; \mathrm{BY})$ distribution and $z(g; \mathrm{RM})$ distribution, which compares the reactions of the two strains to the RAD inhibition. Pearson's correlation coefficient R is 0.50, which is not extremely high, implying that there are some different systems-level responses to the drug between the BY and RM strains. However, if we compare two experimental conditions (RAD+DMSO and DMSO only), R becomes 0.96, which suggests that most of the proteome-level differences between the two strains are conserved even after addition of RAD (Figure 6.2b).

We checked the $z(g; \mathrm{DMSO{+}RAD})$ distribution to study differences between the BY and RM strains under the same experimental condition, which is the RAD treatment. (As Figure 6.2b implies, the $z(g; \mathrm{DMSO})$ and $z(g; \mathrm{DMSO{+}RAD})$ distributions have very similar profiles, so it would be sufficient to investigate the DMSO+RAD case only.) We grouped the genes according to their Gene Ontology (GO) terms, which contain detailed information about biological process, molecular function, and/or cellular component each gene is involved in[7]. Each GO term group has the number of members (genes) and their average z-score, from both of which we can calculate the p-value that shows if the difference in the gene group between the two strains is significant or not (see chapter 4). We removed redundant groups by considering the hierarchical structure of GO term annotation. If two GO term groups share the same gene members and one GO term among the two is a broader term that contains another, we remove the former, to obtain GO terms as specific as possible. If there is no direct hierarchical relationship between the two groups with the same members, we keep both, since the two GO terms would deliver different information.

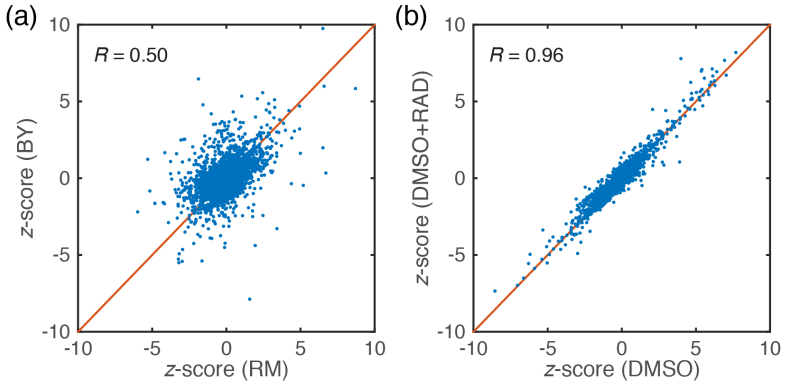


Figure 6.2: z-Score correlations. Red lines indicate the y=x lines. (a) Correlation between BY and RM strains in terms of their responses to RAD (comparing the RAD+DMSO and DMSO-only conditions). (b) Correlation between two experimental conditions (RAD+DMSO, DMSO only) in terms of strain differences (comparing BY and RM). See equations 6.1-6.4 in the text.

Figure 6.3 shows top 30 GO term groups that are significantly more expressed in the BY strain than in the RM strain when both strains are under the RAD treatment, according to their p-values. Similarly, Figure 6.4 shows top 30 GO term groups more expressed in the RM strain. We filtered out the groups with 10 or less genes to reduce biological noises. Notably, proteins involved in crucial pathways (*e.g.* translation, electron transport chain, secretion pathway) and organelles (*e.g.* mitochondria, ribosome, vacuole) have drastically different expression levels in the two strains, suggesting that the unexpectedly large variation between the proteomes of the BY and RM strains (Figure 6.2a) originates from rewiring of "modules" on cellular networks, not individual genes or proteins.

6.4 Proteome-Level Differences in Responses to Radicicol

Considering the significant systems-level difference between the BY and RM strains, there might be some pathways that show different dynamics upon perturbations between the two strains. To investigate this, we studied the genes in the $z(g;\mathrm{BY})$ and $z(g;\mathrm{RM})$ distributions by grouping them according to their GO terms and calculating the p-values of the groups, as done in the previous section.

Since we are interested in the difference between the two strains, we define a single measure $D(G)$, that shows how much different the behaviors of a GO term group G are in the BY and RM strains:

$$D(G) = -\mathrm{sgn}(\langle z\rangle_{G,\mathrm{BY}})\log(p_{G,\mathrm{BY}}) + \mathrm{sgn}(\langle z\rangle_{G,\mathrm{RM}})\log(p_{G,\mathrm{RM}}), \tag{6.6}$$

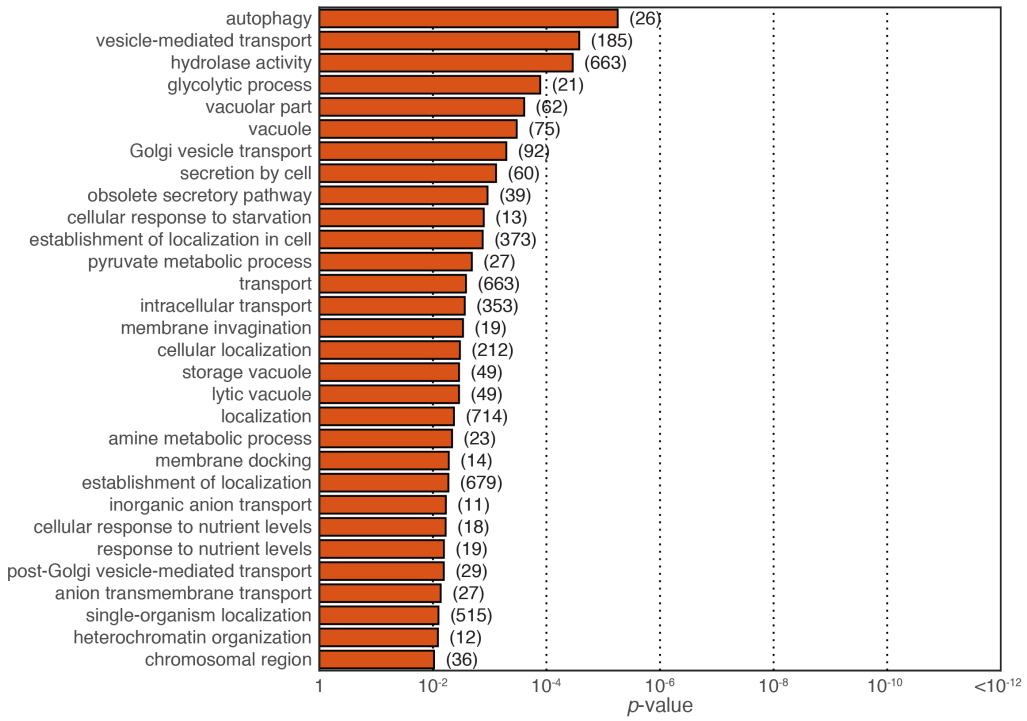


Figure 6.3: Top 30 GO term gene groups that are significantly more expressed in the BY strain than in the RM strain when both strains are treated by RAD. Each bar indicates the p-value of each GO term group, and the number of members is given in parentheses. The groups with at least 11 genes are only shown.

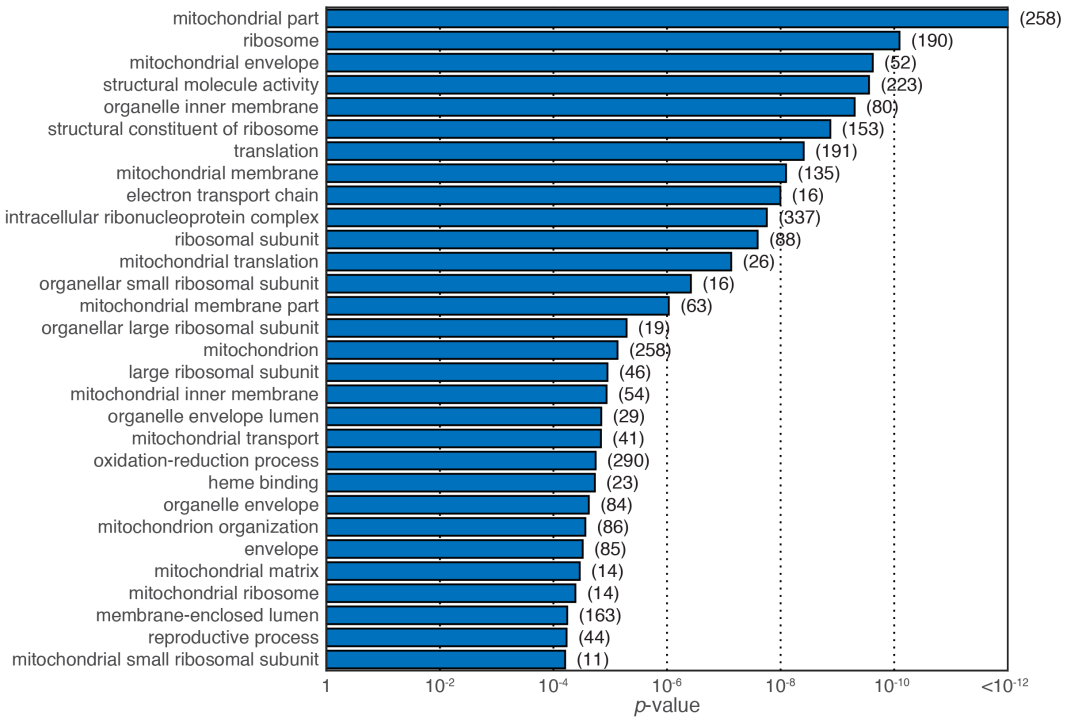


Figure 6.4: Top 30 GO term gene groups that are significantly more expressed in the RM strain than in the BY strain when both strains are treated by RAD. Each bar indicates the p-value of each GO term group, and the number of members is given in parentheses. The groups with at least 11 genes are only shown. The first GO term group "mitochondrial part" has the p-value of 2.5×10^{-21}, which exceeds the upper limit of the x-axis.

where $\mathrm{sgn}(x)$ indicates the sign of x, and $\langle z \rangle_{G,S}$ indicates the arithmetic mean of $z(g; S)$ values in group G, whose p-value is denoted by $p_{G,S}$. Hence, if the two strains regulate the gene group G significantly in different directions, $|D(G)|$ is amplified, but if they move in concert, their significances cancel each other. The sign of $D(G)$ shows which strain regulates the gene group more significantly (positive: BY, negative: RM).

Comparing this $D(G)$ and the "unperturbed" difference (difference between the BY and RM strains with no RAD treatment)

$$D_0(G) = -\mathrm{sgn}(\langle z \rangle_{G,\mathrm{DMSO}}) \log(p_{G,\mathrm{DMSO}}), \tag{6.7}$$

we found an interesting anti-correlation (Figure 6.5), where we only used non-redundant GO term groups whose sizes are greater than 10. This significant anti-correlation indicates that the BY and RM strains react to the inhibition of an essential heat shock protein by reducing the proteomic difference between the two strains. To obtain a more detailed picture, we collected four types of GO term groups:

1. the gene groups with $p_{G,\mathrm{BY}} < 0.05$, $p_{G,\mathrm{RM}} < 0.05$, and the same signs of average z-scores (significant and consistent regulations upon the RAD treatment for both strains),

2. the gene groups with $p_{G,\mathrm{BY}} < 0.05$, $p_{G,\mathrm{RM}} < 0.05$, and the opposite signs of average z-scores (significant and opposite regulations upon the RAD treatment for the two strains),

3. the gene groups with $p_{G,\mathrm{BY}} < 0.05$ and $p_{G,\mathrm{RM}} > 0.05$ (significant regulations upon the RAD treatment only for the BY strains), and

4. the gene groups with $p_{G,\mathrm{BY}} > 0.05$ and $p_{G,\mathrm{RM}} < 0.05$ (significant regulations upon the RAD treatment only for the RM strains).

The gene groups of type 1, on whose genes there are significant and consistent regulations upon the RAD treatment for both BY and RM strains, are shown in Figure 6.6. Unsurprisingly, the genes related to the function of Hsp90 are all up-regulated: protein folding, response to temperature stimulus, response to heat, unfolded protein binding, and protein refolding. Especially, the *HSP82* and *HSC82* genes, which encode the Hsp90 proteins in *S. cerevisiae*, are annotated with GO terms "protein folding" and "unfolded protein binding" but not the other three, showing that the drive is not solely due to the strong up-regulation of *HSP82* or *HSC82*.

There is no group of type 2, and the gene groups of type 3 are shown in Figure 6.7a. The most significant regulation is observed for the genes responsible for mitochondrial constituents. It has

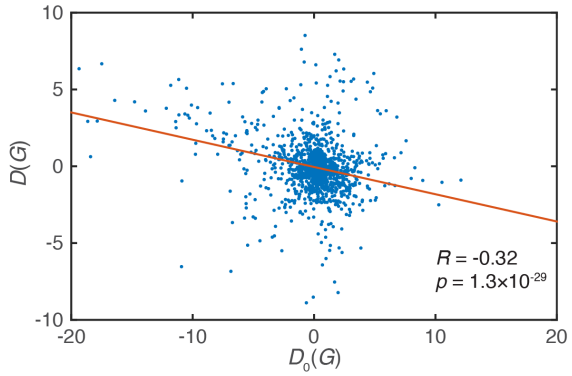


Figure 6.5: The correlation between the unperturbed inter-strain difference $D_0(G)$ and the difference in the reactions of the BY and RM strains to the RAD treatment $D(G)$ (see text for the definitions). The red line indicates the regression line. If a gene group is significantly more expressed in the BY strain, it is more likely that upon the RAD treatment, the RM strain up-regulates the group (or the BY strain down-regulates it, or both) to reduce the difference of the average expression levels of the gene group in the two strains.

been known that in human cancer cells the genes responsible for mitochondrial constituents are highly expressed when Hsp90 proteins are inhibited[91,64]. Here, only the BY strain shows this behavior, and this is presumably because the RM strain already has a sufficient number of mitochondrial genes (Figure 6.4) that it can absorb the impact of the RAD inhibition.

As shown in Figure 6.7b, among the gene groups of type 4, the most significantly down-regulated are the genes responsible for various aspects of membrane components, which are more expressed in the RM strain than the BY strain (Figure 6.4). Again, the membrane-related genes have higher expression levels in the RM strain than in the BY strain, and upon the perturbation, the RM strain reduces their expression levels while the BY strain shows no significant regulation on them, so that the difference between the two strains is reduced. This also happens to the ribosome-related gene group (compare Figures 6.4 and 6.7b).

In summary, if a gene group is significantly more expressed in the BY strain than in the RM strain, it is more likely that upon the RAD treatment, either the RM strain up-regulates the group or the BY strain down-regulates it, to apparently reduce the difference of the average expression levels of the gene group in the two strains. It also holds for the gene groups that are significantly more expressed in the RM strain.

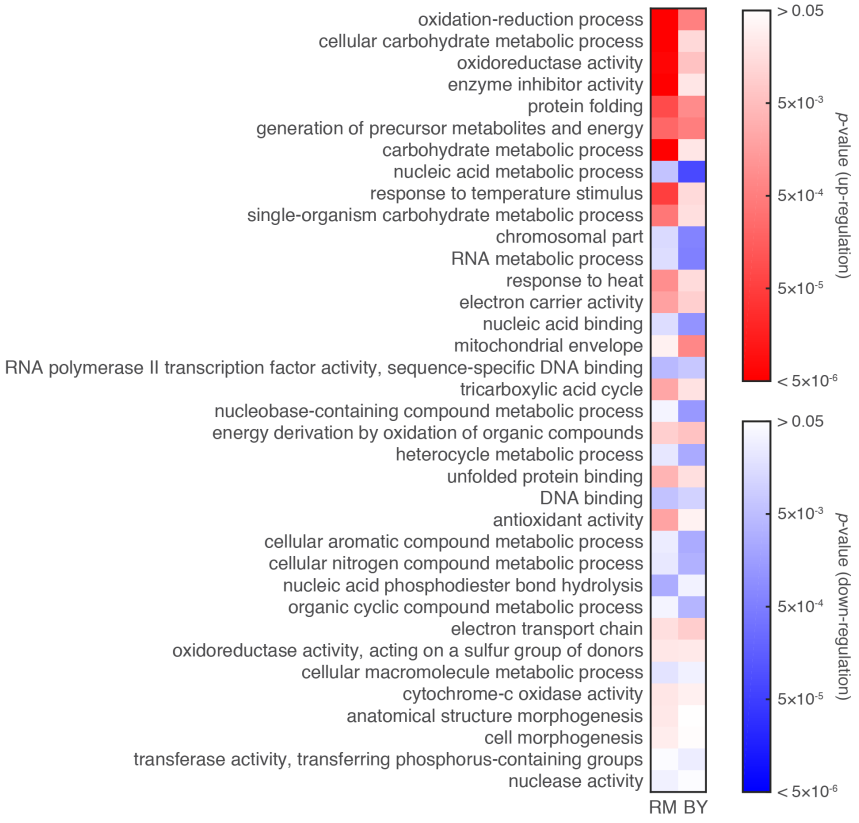


Figure 6.6: The GO term gene groups that show significant and consistent regulations upon the RAD treatment for both BY and RM strain. The hue of color indicates the direction of regulation upon the RAD treatment (red: up-regulation, blue: down-regulation), and the saturation of color provides the p-value. The groups is sorted according to the geometric mean of two p-values.

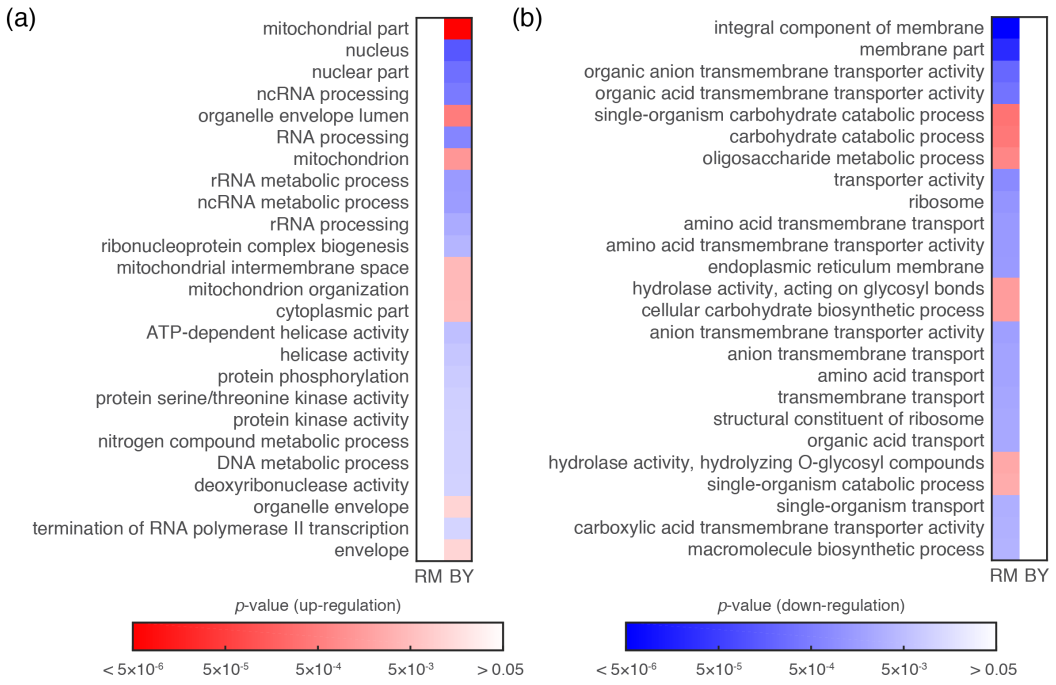


Figure 6.7: The GO term gene groups that show significant regulations upon the RAD treatment (a) only for the BY strain and (b) only for the RM strain. The hue of color indicates the direction of regulation upon the RAD treatment (red: up-regulation, blue: down-regulation), and the saturation of color provides the p-value. The groups are sorted according to the p-values, and only top 25 groups are shown for each panel.

6.5 Phenotype-Level Differences in Responses to Radicicol

As discussed so far, the BY and RM strains show non-identical responses to the RAD treatment at the systems level (due to their different "unperturbed" proteome patterns), so as a consequence, the two strains might show some different cellular level traits under certain conditions. To investigate this, we grouped the genes according to their phenotypes. We employed the phenotypic curation data provided by the *Saccharomyces* Genome Database (SGD)[47], which provides information about phenotypes due to single point mutations on the target gene. For example, the *HSP82* gene is annotated with the phenotypes such as decreased budding index, decreased competitive fitness, and increased heat sensitivity.

Similarly to the previous section, we used the $z(g; \mathrm{BY})$ and $z(g; \mathrm{RM})$ data and calculated the average z-score and p-value for each phenotypic gene group. Again, the sign of the average z-score indicates the direction of regulation due to the RAD treatment (up- or down-regulation), and the p-value indicates the statistical significance of the difference. Also, using the classification scheme for gene groups introduced in the previous section, we collected phenotypic gene groups of types 3 and 4; the former corresponds to the gene groups whose regulation upon the RAD treatment is significant only for the BY strain, and the latter those whose regulation is significant only for the RM strain. Note that there was no phenotypic gene group of type 2, whose members the two strains significantly regulate in opposite directions. Figure 6.8 shows the phenotypic gene groups of type 3, while the phenotypic gene groups of type 4 are shown in Figure 6.9. Here, we only present the gene groups whose sizes are greater than 10.

Although it is not easy to explain all these behaviors in terms of the systems-level changes, there are a few points to note. First, the genes responsible for phenotype "mitochondrial genome maintenance" are strongly up-regulated after the RAD treatment in the BY strain, while this behavior is invisible in the RM strain. This is consistent with the systems-level behavior of mitochondrial genes, as discussed in the previous section. The phenotypic group for "autophagy" is slightly up-regulated only in the RM strain, and this can be explained by already high expression levels of autophagy-related gene groups in the BY strain (Figure 6.3).

6.6 Conclusion

Despite the small difference at the genomic level, the BY and RM strains of *S. cerevisiae* show rather significant differences at the proteomic level in responses to inhibition of a crucial heat shock protein. The inhibition itself does not perturb the proteomes much (the SD of LRPA does not increase much), but some functional and constitutional gene groups are drastically affected by the pertur-

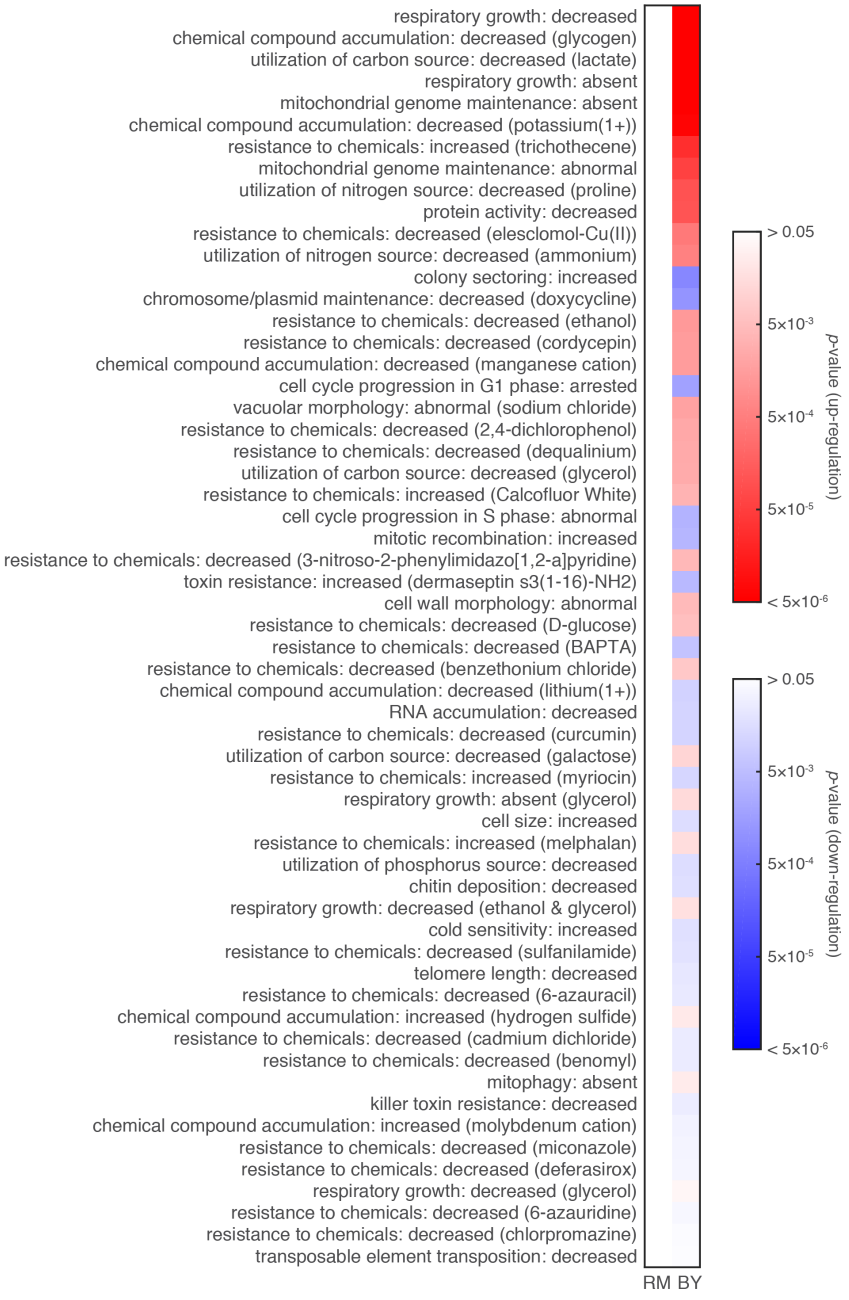


Figure 6.8: The phenotypic gene groups that show significant regulations upon the RAD treatment only for the BY strain. The hue of color indicates the direction of regulation upon the RAD treatment (red: up-regulation, blue: down-regulation), and the saturation of color provides the p-value. The chemicals added to the experimental system are noted in parentheses. The groups are sorted according to their p-values.

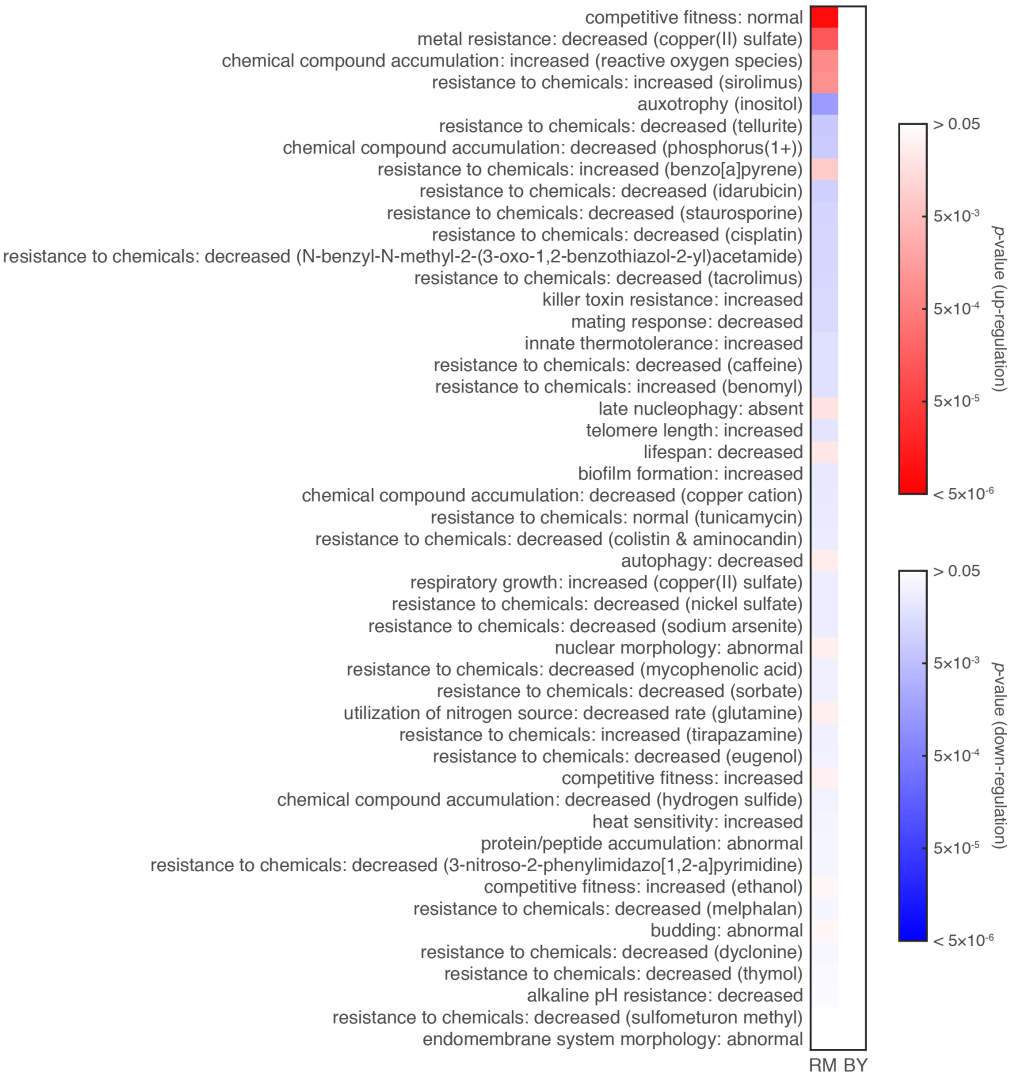


Figure 6.9: The phenotypic gene groups that show significant regulations upon the RAD treatment only for the RM strain. The hue of color indicates the direction of regulation upon the RAD treatment (red: up-regulation, blue: down-regulation), and the saturation of color provides the p-value. The chemicals added to the experimental system are noted in parentheses. The groups are sorted according to their p-values.

bation. In order to investigate their behaviors systematically, we employed two different grouping methods, one using the GO annotation and another using the SGD phenotype annotation.

The GO annotation grouping provides insight on various gene groups at diverse levels of biology. The most striking feature found by this analysis is that the response of each strain to the perturbation tends to decrease the difference between the two strains. It implies that the proteomic patterns of different strains to address the stress are rather similar, suggesting that these patterns were conserved well during evolutionary adaptation processes. This is presumably because the stress response modules are crucial for survival of biological systems, as cell stress response pathways are well conserved among different biological kingdoms [86,9].

Although the relationship between the GO term groups and phenotype term groups is not trivial except for a few cases, the phenotypic grouping complements the bottom-up approach of the proteomic analysis, helping the design of experiment to reveal the difference between the two strains, in responses to the crucial perturbation. We hope that the current analysis and its data will be used to inspire new experiments that provide full understanding on strain differences.

This page intentionally left blank

A

The μ-Potential for Protein-Protein Interactions

The μ-potential is a mean-field knowledge-based potential that assigns a negative (attractive) energy score for a frequent contact and a positive (repulsive) score for an infrequent one, where the frequency is determined statistically from a large data set[76]. This potential captures the elements of positive and negative design into the scoring function, and it has been reported several times that the potential successfully predicts folded protein structures[25,145,146,143]. In this appendix, we will show that the μ-potential can be used to predict binding, since the positive and negative design principle is also intrinsic in a protein-protein interaction (PPI).

The μ-potential contact energy of atom types A and B is defined as

$$E_{AB} = \frac{-\mu \sum_p N_{AB}^p + (1-\mu) \sum_p \tilde{N}_{AB}^p}{\mu \sum_p N_{AB}^p + (1-\mu) \sum_p \tilde{N}_{AB}^p}, \tag{A.1}$$

where N_{AB}^p and $\tilde{N}_{AB}^p$ are the numbers of AB pairs in protein complex p that are in contact and that are not in contact, respectively. Atoms A and B are counted only when they are on the surfaces of different chains. To determine whether a residue is on the surface of the protein chain, we calculated its accessible surface area (ASA) using the Shrake-Rupley algorithm[115] implemented in ASA.PY (see chapter 2). This ASA was normalized to the ASA of the same residue in an Ala-X-Ala motif[110]. The

residue and its constituent heavy atoms are considered to be on the surface when the normalized ASA value exceeds 20 %.

We used the atom-typing scheme developed in a previous study[25], augmented by 5 more atom types for backbone N, α-C, carbonyl C and O atoms, as well as glycine α-C atoms. Two atoms are considered to be in contact if the interatomic distance is less than 10 Å. This choice of cutoff is motivated by the Debye screening length in physiological salt concentrations of 100 mM. Hydrogen atoms are ignored. We used 888 non-redundant protein heterodimers extracted from PDBePISA[73] to compute $\sum_p N^p_{\mathrm{AB}}$ and $\sum_p \tilde{N}^p_{\mathrm{AB}}$. The parameter μ is set to be 0.9906 to make the average interaction over all different types of atomic pairs to be zero, in order to maximize the energy gap between a native docked state and other unstable states. The total contact energy of a PPI system is computed from a simple sum of pairwise μ-potential energies,

$$E_{\mathrm{contact}} = \sum_{\substack{i<j \\ \text{in contact}}} E_{A^1_i A^2_j}, \tag{A.2}$$

where A^k_i is the atom type of surface atom i on component k.

An immediate sanity check is the ability of the potential to distinguish between real protein complexes from decoys. We prepared a non-redundant test set of 292 protein dimers from PDBePISA[73], and two decoy sets by 41 protein dimers from DOCKGROUND[85]. One of the decoy sets consists of dimers whose fraction of correct contacts, f, is close to 0.5, while the other decoy set has $f \approx 0$. The trained μ-potential contact energy is shown to have a discriminative power between native protein complexes and protein complex decoys (Figure A.1). The two-sample Kolmogorov-Smirnov test gives a p-value against the null hypothesis that the data from two sample sets originate from the same distribution; comparison of the training and test sets gives $p = 0.31$ (implying that the two sets are from the same distribution), while comparison of the training set and other two decoy sets yields $p = 4.1 \times 10^{-6}$, and $p = 5.3 \times 10^{-18}$ for the decoy sets with $f = 0.5$ and $f = 0.0$, respectively. Therefore, the μ-potential is capable of discerning decoy protein complexes from real protein complexes.

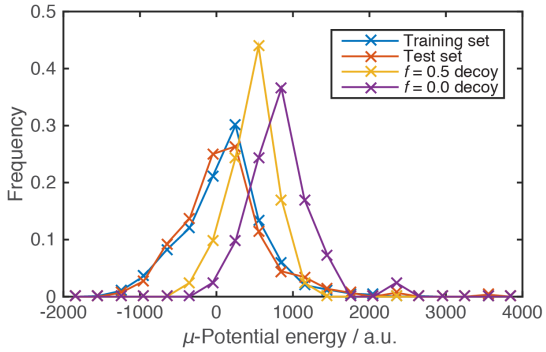


Figure A.1: μ-Potential energy distributions of native and decoy protein sets. Two non-redundant sets of native protein dimers, the training (888 dimers) and the test (292 dimers) sets, are independent to each other. They show almost perfect convergence (p-value = 0.31). Two decoy sets are different by their f values, a fraction of the number of correct contacts to a total number of contacts, showing that the distribution deviates from the native protein distribution as the number of incorrect contacts increases. The bin size is 300 arbitrary units (a.u.).

This page intentionally left blank

References

[1] Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47–97.

[2] Alm, R. A., Ling, L.-S. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uria-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vovis, G. F., & Trust, T. J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen helicobacter pylori. *Nature*, 397(6715), 176–180.

[3] Altelaar, A. F. M., Munoz, J., & Heck, A. J. R. (2013a). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1), 35–48.

[4] Altelaar, A. M., Frese, C. K., Preisinger, C., Hennrich, M. L., Schram, A. W., Timmers, H. M., Heck, A. J., & Mohammed, S. (2013b). Benchmarking stable isotope labeling based quantitative proteomics. *Journal of Proteomics*, 88, 14–26.

[5] Amyes, S. G. (1982). Bactericidal activity of trimethoprim alone and in combination with sulfamethoxazole on susceptible and resistant escherichia coli k-12. *Antimicrobial Agents and Chemotherapy*, 21(2), 288–293.

[6] Applebee, M. K., Joyce, A. R., Conrad, T., Pettigrew, D. W., & Palsson, B. Ø. (2011). Functional and metabolic effects of adaptive glycerol kinase (glpk) mutants in escherichia coli. *Journal of Biological Chemistry*, 286, 23150–23159.

[7] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.

[8] Barabasi, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2), 101–113.

[9] Bayles, K. W. (2014). Bacterial programmed cell death: making sense of a paradox. *Nature Reviews Microbiology*, 12(1), 63–69.

[10] Bernstein, J. A., Lin, P.-H., Cohen, S. N., & Lin-Chao, S. (2004). Global analysis of escherichia coli rna degradosome function using dna microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), 2758–2763.

[11] Bershtein, S., Choi, J.-M., Bhattacharyya, S., Budnik, B., & Shakhnovich, E. (2015a). Systems-level response to point mutations in a core metabolic enzyme modulates genotype-phenotype relationship. *Cell Reports*, 11(4), 645–656.

[12] Bershtein, S., Mu, W., Serohijos, A. W. R., Zhou, J., & Shakhnovich, E. I. (2013). Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Molecular Cell*, 49(1), 133–144.

[13] Bershtein, S., Mu, W., & Shakhnovich, E. I. (2012). Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13), 4857–4862.

[14] Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., & Tawfik, D. S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121), 929–932.

[15] Bershtein, S., Serohijos, A. W. R., Bhattacharyya, S., Manhart, M., Choi, J.-M., Mu, W., Zhou, J., & Shakhnovich, E. I. (2015b). Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. *PLoS Genetics*, 11(10), 1–25.

[16] Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15), 5869–5874.

[17] Bolen, D. W. & Rose, G. D. (2008). Structure and energetics of the hydrogen-bonded backbone in protein folding. *Annual Review of Biochemistry*, 77(1), 339–362.

[18] Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. *Nature*, 468(7325), 851–854.

[19] Bornholdt, S. & Sneppen, K. (2000). Robustness as an evolutionary principle. *Proceedings of the Royal Society of London B: Biological Sciences*, 267(1459), 2281–2286.

[20] Brauer, M. J., Huttenhower, C., Airoldi, E. M., Rosenstein, R., Matese, J. C., Gresham, D., Boer, V. M., Troyanskaya, O. G., & Botstein, D. (2008). Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Molecular Biology of the Cell*, 19(1), 352–367.

[21] Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568), 752–755.

[22] Burchall, J. J. & Hitchings, G. H. (1965). Inhibitor binding analysis of dihydrofolate reductases from various species. *Molecular Pharmacology*, 1(2), 126–136.

[23] Cannavacciuolo, L. & Landau, D. P. (2005). Critical behavior of the three-dimensional compressible ising antiferromagnet at constant volume: A monte carlo study. *Physical Review B*, 71, 134104.

[24] Carbonell, P., Nussinov, R., & del Sol, A. (2009). Energetic determinants of protein binding specificity: Insights into protein interaction networks. *Proteomics*, 9(7), 1744–1753.

[25] Chen, W. W. & Shakhnovich, E. I. (2005). Lessons from the design of a novel atomic potential for protein folding. *Protein Science*, 14(7), 1741–1752.

[26] Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., & Wong, E. D. (2012). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue), D700–5.

[27] Choi, J.-M., Serohijos, A. W., Murphy, S., Lucarelli, D., Lofranco, L. L., Feldman, A., & Shakhnovich, E. I. (2015). Minimalistic predictor of protein binding energy: Contribution of solvation factor to protein binding. *Biophysical Journal*, 108(4), 795–798.

[28] Chou, H.-H., Delaney, N. F., Draghi, J. A., & Marx, C. J. (2014). Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. *PLoS Genetics*, 10(2), e1004149.

[29] Chou, H.-H., Marx, C. J., & Sauer, U. (2015). Transhydrogenase promotes the robustness and evolvability of e. coli deficient in nadph production. *PLoS Genetics*, 11(2), e1005007.

[30] Clarke, B., Fokoué, E., & Zhang, H. (2009). *Principles and Theory for Data Mining and Machine Learning*. New York: Springer.

[31] Consortium, T. U. (2015). Uniprot: a hub for protein information. *Nucleic Acids Research*, 43(D1), D204–D212.

[32] Cusick, M. E., Klitgord, N., Vidal, M., & Hill, D. E. (2005). Interactome: gateway into systems biology. *Human Molecular Genetics*, 14(suppl 2), R171–R181.

[33] Datsenko, K. A. & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in escherichia coli k-12 using pcr products. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), 6640–6645.

[34] Davids, W. & Zhang, Z. (2008). The impact of horizontal gene transfer in shaping operons and protein interaction networks–direct evidence of preferential attachment. *BMC Evolutionary Biology*, 8, 23.

[35] de Crecy-Lagard, V., El Yacoubi, B., de la Garza, R. D., Noiriel, A., & Hanson, A. D. (2007). Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics*, 8, 245.

[36] de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Frohlich, F., Walther, T. C., & Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217), 1251–1254.

[37] De Las Rivas, J. & Fontanillo, C. (2010). Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology*, 6(6), e1000807.

[38] Dean, A. M., Dykhuizen, D. E., & Hartl, D. L. (1986). Fitness as a function of beta-galactosidase activity in escherichia coli. *Genetical Research*, 48(1), 1–8.

[39] Dobson, C. M. (2003). Protein folding and disease: a view from the first horizon symposium. *Nature Reviews Drug Discovery*, 2(2), 154–160.

[40] Dorogovtsev, S. & Mendes, J. (2013). *Evolution of Networks: From Biological Nets to the Internet and WWW*. OUP Oxford.

[41] Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2008). Critical phenomena in complex networks. *Reviews of Modern Physics*, 80, 1275–1335.

[42] Drummond, D. A. & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2), 341–352.

[43] Dykhuizen, D. E., Dean, A. M., & Hartl, D. L. (1987). Metabolic flux and fitness. *Genetics*, 115(1), 25–31.

[44] Ehrenberg, M., Bremer, H., & Dennis, P. P. (2013). Medium-dependent control of the bacterial growth rate. *Biochimie*, 95(4), 643–658.

[45] Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319(6050), 199–203.

[46] Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584), 1183–1186.

[47] Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R., Oughtred, R., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Dolinski, K., Botstein, D., & Cherry, J. M. (2010). Saccharomyces genome database provides mutant phenotype data. *Nucleic Acids Research*, 38(suppl 1), D433–D436.

[48] England, J. L. & Shakhnovich, E. I. (2003). Structural determinant of protein designability. *Physical Review Letters*, 90, 218101.

[49] Farkas, I., Derényi, I., Palla, G., & Vicsek, T. (2004). Equilibrium statistical mechanics of network structures. *Lecture Notes in Physics*, 650, 163–187.

[50] Ferreira, L. & Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Communications in Statistics - Simulation and Computation*, 38(9), 1925–1949.

[51] Firnberg, E., Labonte, J. W., Gray, J. J., & Ostermeier, M. (2014). A comprehensive, high-resolution map of a gene's fitness landscape. *Molecular Biology and Evolution*, 31(6), 1581–1592.

[52] Gao, Y., Wang, R., & Lai, L. (2003). Structure-based method for analyzing protein–protein interfaces. *Journal of Molecular Modeling*, 10(1), 44–54.

[53] Geiler-Samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M., Hartl, D. L., & Drummond, D. A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 108(2), 680–685.

[54] Geiler-Samerotte, K. A., Hashimoto, T., Dion, M. F., Budnik, B. A., Airoldi, E. M., & Drummond, D. A. (2013). Quantifying condition-dependent intracellular protein levels enables high-precision fitness estimates. *PLoS ONE*, 8(9), e75320.

[55] Gogarten, J. P. & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9), 679–687.

[56] Goh, C.-S., Milburn, D., & Gerstein, M. (2004). Conformational changes associated with protein–protein interactions. *Current Opinion in Structural Biology*, 14(1), 104–109.

[57] Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology*, 320(2), 369–387.

[58] Hartung, S. & Talmon, N. (2015). The complexity of degree anonymization by graph contractions. In R. Jain, S. Jain, & F. Stephan (Eds.), *Theory and Applications of Models of Computation: 12th Annual Conference, TAMC 2015, Singapore, May 18-20, 2015, Proceedings*, Lecture Notes in Computer Science (pp. 260–271). Springer International Publishing.

[59] Heo, M., Maslov, S., & Shakhnovich, E. (2011). Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10), 4258–4263.

[60] Huo, S., Massova, I., & Kollman, P. A. (2002). Computational alanine scanning of the 1:1 human growth hormone–receptor complex. *Journal of Computational Chemistry*, 23(1), 15–27.

[61] Janin, J., Rodier, F., Chakrabarti, P., & Bahadur, R. P. (2007). Macromolecular recognition in the protein data bank. *Acta Crystallographica Section D*, 63(1), 1–8.

[62] Jarosz, D. F. & Lindquist, S. (2010). Hsp90 and environmental stress transform the adaptive value of natural genetic variation. *Science*, 330(6012), 1820–1824.

[63] Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B., & Bolon, D. N. A. (2013). Latent effects of hsp90 mutants revealed at reduced expression levels. *PLoS Genetics*, 9(6), e1003600.

[64] Kang, B. H., Plescia, J., Dohi, T., Rosa, J., Doxsey, S. J., & Altieri, D. C. (2007). Regulation of tumor cell mitochondrial homeostasis by an organelle-specific hsp90 chaperone network. *Cell*, 131(2), 257–270.

[65] Kastritis, P. L. & Bonvin, A. M. J. J. (2012). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79), 20120835.

[66] Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. J. J., & Janin, J. (2011). A structure-based benchmark for protein–protein binding affinity. *Protein Science*, 20(3), 482–491.

[67] Kastritis, P. L., Rodrigues, J. P., Folkers, G. E., Boelens, R., & Bonvin, A. M. (2014). Proteins feel more than they see: Fine-tuning of binding affinity by properties of the non-interacting surface. *Journal of Molecular Biology*, 426(14), 2632–2652.

[68] Keskin, O., Gursoy, A., Ma, B., , & Nussinov, R. (2008). Principles of protein-protein interactions: What are the preferred ways for proteins to interact? *Chemical Reviews*, 108(4), 1225–1244.

[69] Klumpp, S., Zhang, Z., & Hwa, T. (2009). Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139(7), 1366–1375.

[70] Knoppel, A., Lind, P. A., Lustig, U., Nasvall, J., & Andersson, D. I. (2014). Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Molecular Biology and Evolution*, 31(5), 1220–1227.

[71] Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology*, 55, 709–742.

[72] Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), 14116–14121.

[73] Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3), 774–797.

[74] Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in escherichia coli. *Science*, 324(5924), 255–258.

[75] Kuriyan, J. & Eisenberg, D. (2007). The origin of protein interactions and allostery in colocalization. *Nature*, 450(7172), 983–990.

[76] Kussell, E., Shimada, J., & Shakhnovich, E. I. (2002). A structure-based method for derivation of all-atom potentials for protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8), 5343–5348.

[77] Kwon, Y. K., Higgins, M. B., & Rabinowitz, J. D. (2010). Antifolate-induced depletion of intracellular glycine and purines inhibits thymineless death in e. coli. *ACS Chemical Biology*, 5(8), 787–795.

[78] Kwon, Y. K., Lu, W., Melamud, E., Khanam, N., Bognar, A., & Rabinowitz, J. D. (2008). A domino effect in antifolate drug action in escherichia coli. *Nature Chemical Biology*, 4(10), 602–608.

[79] Laradji, M., Landau, D. P., & Dünweg, B. (1995). Structural properties of $si_{1-x}ge_x$ alloys: A monte carlo simulation with the stillinger-weber potential. *Physical Review B*, 51, 4894–4902.

[80] Larance, M. & Lamond, A. I. (2015). Multidimensional proteomics for cell biology. *Nature Reviews Molecular Cell Biology*, 16(5), 269–280.

[81] Lawrence, J. G. & Ochman, H. (1998). Molecular archaeology of the escherichia coli genome. *Proceedings of the National Academy of Sciences of the United States of America*, 95(16), 9413–9417.

[82] Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*, 14(3), 168–178.

[83] Lercher, M. J. & Pal, C. (2008). Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular Biology and Evolution*, 25(3), 559–567.

[84] Lind, P. A., Tobin, C., Berg, O. G., Kurland, C. G., & Andersson, D. I. (2010). Compensatory gene amplification restores fitness after inter-species gene replacements. *Molecular Microbiology*, 75(5), 1078–1089.

[85] Liu, S., Gao, Y., & Vakser, I. A. (2008). Dockground: protein-protein docking decoy set. *Bioinformatics*, 24(22), 2634–2635.

[86] Liu, X. D., Liu, P. C., Santoro, N., & Thiele, D. J. (1997). Conservation of a stress response: human heat shock transcription factors functionally substitute for yeast hsf. *The EMBO Journal*, 16(21), 6466–6477.

[87] Lukatsky, D., Shakhnovich, B., Mintseris, J., & Shakhnovich, E. (2007). Structural similarity enhances interaction propensity of proteins. *Journal of Molecular Biology*, 365(5), 1596–1606.

[88] Lunzer, M., Miller, S. P., Felsheim, R., & Dean, A. M. (2005). The biochemical architecture of an ancient adaptive landscape. *Science*, 310(5747), 499–501.

[89] Lynch, M. (2013). Evolutionary diversification of the multimeric states of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30), E2821–E2828.

[90] Maier, T., Schmidt, A., Güell, M., Kühner, S., Gavin, A.-C., Aebersold, R., & Serrano, L. (2011). Quantification of mrna and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7(1), 511.

[91] Margineantu, D. H., Emerson, C. B., Diaz, D., & Hockenbery, D. M. (2007). Hsp90 inhibition decreases mitochondrial protein turnover. *PLoS ONE*, 2(10), 1–14.

[92] Marri, P. R., Hao, W., & Golding, G. B. (2007). The role of laterally transferred genes in adaptive evolution. *BMC Evolutionary Biology*, 7(Suppl 1), S8.

[93] Maslov, S. & Ispolatov, I. (2007). Propagation of large concentration changes in reversible protein-binding networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34), 13655–13660.

[94] Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A., & Collado-Vides, J. (2004). Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Molecular Biology and Evolution*, 21(10), 1884–1894.

[95] Merikoski, J. K. (1984). On the trace and the sum of elements of a matrix. *Linear Algebra and its Applications*, 60, 177–185.

[96] Mika, F. & Hengge, R. (2005). A two-component phosphotransfer network involving arcb, arca, and rssb coordinates synthesis and proteolysis of σs (rpos) in e. coli. *Genes & Development*, 19(22), 2770–2781.

[97] Moal, I. H., Agius, R., & Bates, P. A. (2011). Protein-protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, 27(21), 3002–3009.

[98] Momma, K. & Izumi, F. (2011). Vesta3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of Applied Crystallography*, 44(6), 1272–1276.

[99] Monod, J., Changeux, J.-P., & Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology*, 6(4), 306–329.

[100] Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots: A review of the protein-protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4), 803–812.

[101] Navarre, W. W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S. J., & Fang, F. C. (2006). Selective silencing of foreign dna with low gc content by the h-ns protein in salmonella. *Science*, 313(5784), 236–238.

[102] Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions. *The EMBO Journal*, 22(14), 3486–3492.

[103] Pal, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, 37(12), 1372–1375.

[104] Pastor-Satorras, R. & Vespignani, A. (2007). *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press.

[105] Popa, O. & Dagan, T. (2011). Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology*, 14(5), 615–623.

[106] Priester, L. (2013). *Grain Boundaries: From Theory to Engineering*, chapter Defects in the Grain Boundary Structure, (pp. 135–146). Springer Netherlands: Dordrecht.

[107] Queitsch, C., Sangster, T. A., & Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889), 618–624.

[108] Rajamani, D., Thiel, S., Vajda, S., & Camacho, C. J. (2004). Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), 11287–11292.

[109] Rutherford, S. L. & Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709), 336–342.

[110] Samanta, U., Bahadur, R. P., & Chakrabarti, P. (2002). Quantifying the accessible surface area of protein residues in their local environment. *Protein Engineering*, 15(8), 659–667.

[111] Sangurdekar, D. P., Zhang, Z., & Khodursky, A. B. (2011). The association of dna damage response and nucleotide level modulation with the antibacterial mechanism of the anti-folate drug trimethoprim. *BMC Genomics*, 12(1), 1–14.

[112] Schnell, J. R., Dyson, H. J., & Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annual Review of Biophysics and Biomolecular Structure*, 33(1), 119–140.

[113] Shakhnovich, E. & Gutin, A. (1990). Enumeration of all compact conformations of copolymers with random sequence of links. *Journal of Chemical Physics*, 93(8), 5967–5971.

[114] Sharma, S. V., Agatsuma, T., & Nakano, H. (1998). Targeting of the protein chaperone, hsp90, by the transformation suppressing agent, radicicol. *Oncogene*, 16(20), 2639–2645.

[115] Shrake, A. & Rupley, J. (1973). Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2), 351–371.

[116] Singer, S., Ferone, R., Walton, L., & Elwell, L. (1985). Isolation of a dihydrofolate reductase-deficient mutant of escherichia coli. *Journal of Bacteriology*, 164(1), 470–472.

[117] Slavov, N., Budnik, B. A., Schwab, D., Airoldi, E. M., & van Oudenaarden, A. (2014). Constant growth rate can be supported by decreasing energy flux and increasing aerobic glycolysis. *Cell Reports*, 7(3), 705–714.

[118] Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., & Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, 318(5855), 1449–1452.

[119] Soskine, M. & Tawfik, D. S. (2010). Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics*, 11(8), 572–582.

[120] Spirin, V., Gelfand, M. S., Mironov, A. A., & Mirny, L. A. (2006). A metabolic network in the evolutionary context: Multiscale structure and modularity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23), 8774–8779.

[121] Stern, A., Mayrose, I., Penn, O., Shaul, S., Gophna, U., & Pupko, T. (2010). An evolutionary analysis of lateral gene transfer in thymidylate synthase enzymes. *Systematic Biology*, 59(2), 212–225.

[122] Studier, F. W., Daegelen, P., Lenski, R. E., Maslov, S., & Kim, J. F. (2009). Understanding the differences between genome sequences of escherichia coli b strains rel606 and bl21(de3) and comparison of the e. coli b and k-12 genomes. *Journal of Molecular Biology*, 394(4), 653–680.

[123] Taipale, M., Jarosz, D. F., & Lindquist, S. (2010). Hsp90 at the hub of protein homeostasis: emerging mechanistic insights. *Nature Reviews Molecular Cell Biology*, 11(7), 515–528.

[124] Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., & Xie, X. S. (2010). Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), 533–538.

[125] Tateyama, Y. & Ohno, T. (2003). Stability and clusterization of hydrogen-vacancy complexes in $\alpha - \mathrm{Fe}$: an *ab initio* study. *Physical Review B*, 67, 174105.

[126] Thomas, C. M. & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature Reviews Microbiology*, 3(9), 711–721.

[127] Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., & Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical Chemistry*, 75(8), 1895–1904.

[128] Thompson, D., Regev, A., & Roy, S. (2015). Comparative analysis of gene regulatory networks: From network reconstruction to evolution. *Annual Review of Cell and Developmental Biology*, 31(1), 399–428.

[129] Tian, F., Lv, Y., & Yang, L. (2011). Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. *Amino Acids*, 43(2), 531–543.

[130] Tokuriki, N. & Tawfik, D. S. (2009). Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, 459(7247), 668–673.

[131] Toprak, E., Veres, A., Michel, J.-B., Chait, R., Hartl, D. L., & Kishony, R. (2012). Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics*, 44(1), 101–105.

[132] Tuller, T., Girshovich, Y., Sella, Y., Kreimer, A., Freilich, S., Kupiec, M., Gophna, U., & Ruppin, E. (2011). Association between translation efficiency and horizontal gene transfer within microbial communities. *Nucleic Acids Research*, 39(11), 4743–4755.

[133] Vo, T. V., Das, J., Meyer, M. J., Cordero, N. A., Akturk, N., Wei, X., Fair, B. J., Degatano, A. G., Fragoza, R., Liu, L. G., Matsuyama, A., Trickey, M., Horibata, S., Grimson, A., Yamano, H., Yoshida, M., Roth, F. P., Pleiss, J. A., Xia, Y., & Yu, H. (2016). A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell*, 164(1), 310–323.

[134] Vogel, C. & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4), 227–232.

[135] Vreven, T., Hwang, H., Pierce, B. G., & Weng, Z. (2012). Prediction of protein-protein binding free energies. *Protein Science*, 21(3), 396–404.

[136] Wagner, A. & Wright, J. (2007). Alternative routes and mutational robustness in complex regulatory networks. *Biosystems*, 88(1–2), 163–172.

[137] Wanscher, J. H. (1975). An analysis of wilhelm johannsen's genetical genotype "term" 1909–26. *Hereditas*, 79(1), 1–4.

[138] Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

[139] Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684), 440–442.

[140] Weinreich, D. M., Delaney, N. F., Depristo, M. A., & Hartl, D. L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770), 111–114.

[141] Wellner, A. & Gophna, U. (2008). Neutrality of foreign complex subunits in an experimental model of lateral gene transfer. *Molecular Biology and Evolution*, 25(9), 1835–1840.

[142] Wylie, C. S. & Shakhnovich, E. I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24), 9916–9921.

[143] Xu, J., Huang, L., & Shakhnovich, E. I. (2011). The ensemble folding kinetics of the fbp28 ww domain revealed by an all-atom monte carlo simulation in a knowledge-based potential. *Proteins: Structure, Function, and Bioinformatics*, 79(6), 1704–1714.

[144] Yan, Z., Guo, L., Hu, L., & Wang, J. (2013). Specificity and affinity quantification of protein-protein interactions. *Bioinformatics*, 29(9), 1127–1133.

[145] Yang, J. S., Chen, W. W., Skolnick, J., & Shakhnovich, E. I. (2007). All-atom ab initio folding of a diverse set of proteins. *Structure*, 15(1), 53–63.

[146] Yang, J. S., Wallin, S., & Shakhnovich, E. I. (2008). Universality and diversity of folding mechanics for three-helix bundle proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 105(3), 895–900.

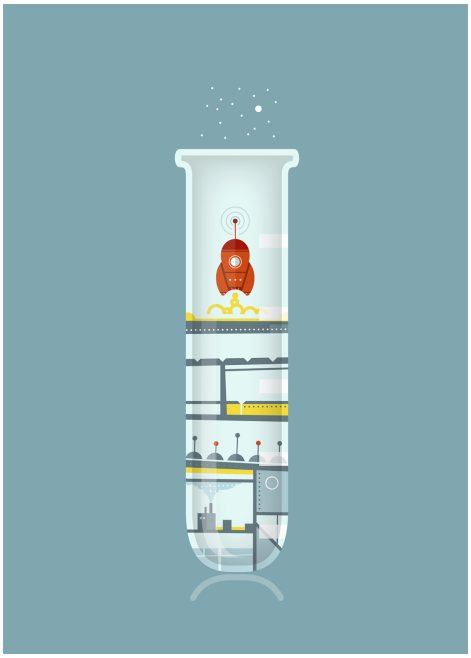
[147] Zhang, J., Maslov, S., & Shakhnovich, E. I. (2008). Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Molecular Systems Biology*, 4(1), 210.

[148] Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., & Honig, B. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421), 556–560.

[149] Zhou, L., Zhang, A. B., Wang, R., Marcotte, E. M., & Vogel, C. (2013a). The proteomic response to mutants of the escherichia coli rna degradosome. *Molecular BioSystems*, 9, 750–757.

[150] Zhou, P., Wang, C., Tian, F., Ren, Y., Yang, C., & Huang, J. (2013b). Biomacromolecular quantitative structure-activity relationship (bioqsar): a proof-of-concept study on the modeling, prediction and interpretation of protein–protein binding affinity. *Journal of Computer-Aided Molecular Design*, 27(1), 67–78.

[151] Zwietering, M. H., Jongenburger, I., Rombouts, F. M., & van 't Riet, K. (1990). Modeling of the bacterial growth curve. *Applied and Environmental Microbiology*, 56(6), 1875–1881.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.