massachusetts institute of technology — artificial intelligence laboratory

Understanding Subsystems in Biology through Dimensionality Reduction, Graph Partitioning and Analytical Modeling

Philip Mjong-Hyon Shin Kim

AI Technical Report 2003-001 February 2003

Understanding Subsystems in Biology through Dimensionality Reduction, Graph Partitioning and Analytical Modeling

by

Philip Mjong-Hyon Shin Kim

VORDIPLOM IN PHYSICS
VORDIPLOM IN BIOCHEMISTRY
UNIVERSITY OF TÜBINGEN, 1997

Submitted to the Department of Chemistry in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physical Chemistry

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2003

Certified by: Bruce Tidor
Associate Professor of Bioengineering and Computer Science
Thesis Supervisor

Accepted by: Robert W. Field
Chairman, Department Committee on Graduate Students

This thesis has been examined by a committee of the
Department of Chemistry as follows:

Jianshu Cao: Thesis Committee Chair

Bruce Tidor: Thesis Supervisor

Robert T. Sauer:

Alexander van Oudenaarden:

Understanding Subsystems in Biology through Dimensionality Reduction, Graph Partitioning and Analytical Modeling

by

Philip Mjong-Hyon Shin Kim

Submitted to the Department of Chemistry on January 21, 2003, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Physical Chemistry

Abstract

Biological systems exhibit rich and complex behavior through the orchestrated interplay of a large array of components. It is hypothesized that separable subsystems with some degree of functional autonomy exist; deciphering their independent behavior and functionality would greatly facilitate understanding the system as a whole. Discovering and analyzing such subsystems are hence pivotal problems in the quest to gain a quantitative understanding of complex biological systems. In this work, using approaches from machine learning, physics and graph theory, methods for the identification and analysis of such subsystems were developed. A novel methodology, based on a recent machine learning algorithm known as non-negative matrix factorization (NMF), was developed to discover such subsystems in a set of large-scale gene expression data. This set of subsystems was then used to predict functional relationships between genes, and this approach was shown to score significantly higher than conventional methods when benchmarking them against existing databases. Moreover, a mathematical treatment was developed to treat simple network subsystems based only on their topology (independent of particular parameter values). Application to a problem of experimental interest demonstrated the need for extentions to the conventional model to fully explain the experimental data. Finally, the notion of a subsystem was evaluated from a topological perspective. A number of different protein networks were examined to analyze their topological properties with respect to separability, seeking to find separable subsystems. These networks were shown to exhibit separability in a nonintuitive fashion, while the separable subsystems were of strong biological significance. It was demonstrated that the separability property found was not due to incomplete or biased data, but is likely to reflect biological structure.

Thesis Supervisor: Bruce Tidor
Title: Associate Professor of Bioengineering and Computer Science

Acknowledgements

First, I would like to thank my thesis advisor, Bruce Tidor, for his guidance, support and the freedom he gave me to pursue my own interests. It has been a great experience to work in his laboratory and I have learned a great deal.

I thank my peers in our research group for helpful discussions and a lot of technical advice and help. Especially, I am indebted to Michael D. Altman, who has been a great resource in everything relating to computers and computer programming. It is fair to say that much of my work was made possible through his tireless efforts. On a related note, I would like to thank David F. Green, for computer help and administering our computer system as well as helpful scientific suggestions. Moreover, I would like to acknowledge Bambang S. Adiwijaya and Caitlin A. Bever for many fruitful discussions and Zak Hendsch, Justin A. Caravella, Karl J. Hanf, Brian (Woody) H. Sherman, Alexander Akhiezer, Alessandro Senes and Shaun Lippow for helpful suggestions.

A special thank you I extend to H. Sebastian Seung, who got me initially interested in the field of systems biology at the start of my graduate career and who has provided me with various kinds of mentorship in many areas.

Furthermore, I thank Trey Ideker whose experience and insight had much impact on my thinking especially in my work on the separability of protein networks.

I gratefully acknowledge the Boehringer-Ingelheim Fonds and the Merck/MIT collaboration for generous funding throughout my graduate career.

Finally, I would like to thank my parents, Bjong-Ro and Suh-Young Kim for much encouragement and support in all stages of my life.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

> "He who learns but does not think, is lost! He who thinks but does not learn is in great danger."
> *Confucius, 551 BC - 479 BC*

1.1 Paradigm changes in biology, a brief overview

Biology as a scientific discipline has seen a tremendous development over the past century. In the time before 1900, it was largely a descriptive science, with its main branches, zoology and botany, focusing on cataloguing and categorizing lifeforms according to phenotypic characteristics that were elucidated by visual inspection. Different classification schemes arose, for instance categorizing the animal kingdom (which is the subject of zoology) into taxonomical categories: *phylum*, *classis*, *ordo*, *familia*, *subfamilia*, *genus* and *species*. The focus of research was on different whole organisms, their morphology and behavior, inspectable by the human eye [131].

Changes of paradigms and scientific thinking were brought about by advances in technology. The invention of the light microscope in 1665 by Robert Hooke gave rise to the notion of a *cell* and ultimately spawned the sub-discipline of cellular biology [91]. Here still, the discipline remained a descriptive one for some time, classifying different cell types by their morphological and functional characteristics and describing cellular subcomponents. The focus shifted from whole organisms to whole cells or tissues, inspectable by the light microscope.

In the early 1900's, advances from organic chemistry brought about the revolution of biochemistry. It became possible to isolate, identify, charac-

terize and later even synthesize biological agents and chemicals of strong biological efficacy. This development had tremendous impact not only on the medical fields, but it also transformed biology from a descriptive into an analytical discipline. Scientists began to apply a reductionist approach, which had proven so successful in physics and chemistry, to biological systems; arguably the most important in a long line of revolutionary achievements was the discovery of deoxy-ribonucleic acid as the agent of biological inheritance. In the mid 1900's, another revolution was brought forth by the powerful techniques of molecular biology, which enabled researchers to duplicate, amplify, isolate and alter biological agents, namely DNA/RNA and proteins with hitherto unimaginable ease. The scientific focus again went to a smaller scope, shifting from cells and tissues to macromolecules, proteins and lipids, carbohydrates and ribonucleic acids [128].

Finally, in the late 1900's advances in computer technology and a number of physico-chemical techniques, namely the development of protein crystallography, made structural biology a reality [117]. It was now possible to study biomolecules at the *atomic* level, predict and elucidate their dynamics and relate atomic structural properties to biological function.

So the scope in biology over the past century went from the macroscopic (on a scale of $\approx 10^0 m$) to the microscopic ($\approx 10^{-5} m$), to the macromolecular ($\approx 10^{-8} m$) and finally to the atomic ($\approx 10^{-10} m$) level. This reductionist approach, just as in the physical sciences, has proven tremendously successful and has greatly improved our understanding of biological phenomena.

1.2 Systems biology: Experimental methodologies

The end of the last century saw the beginning of a new paradigm in biology. An array of very powerful new experimental methods, as well as the advent of ample computation, data storage and retrieval technologies (namely in form of the internet), led to an explosive increase in the amount of data available to bioscientists. This data is a mixture of results from new, so-called high-throughput experiments, which are capable of measuring thousands or even tens of thousands of datapoints at a time, and the collection of curated data from the literature. With this large collection of data at hand and the promise of the high-throughput technologies to deliver new data at an unprecedented pace, efforts are underway to reverse the shift in scope mentioned above; complementing the "one-gene, one-protein" approach, that has proven so successful, genome-wide and cell-wide approaches are emerging [52, 72, 74].

Probably the most visible advance was made in gene sequencing technology, which ultimately led to the Human Genome Project and the elucidation

Species	Abundance levels	Interactions
Small molecules	Mass spectrometry NMR	KEGG, MetaCyc databases Primary literature
Proteins	Mass spectrometry 2D Gel-electrophoresis GFP fluorescence Protein arrays	Yeast-two hybrid assays Co-immunoprecipitation and tandem mass spectrometry BIND, DIP, MIPS Databases
RNA	Gene expression	
DNA	Gene sequencing	Chromatin immunoprecipitation microarray analysis REGULONDB, TRANSFAC, BIND databases

Table 1.1: An overview over existing high-throughput experiments and large-scale databases.

of the entire sequence of the human genome [75, 126]. This development marked the beginning of the so-called "post-genomic era". However, while the Human Genome Project stirred up a lot of publicity and certainly will have a large impact on bioscience in the future, it was one of many steps in the advancement of biology; for instance, the genome sequence of some model organisms, most notably of *Saccharomyces cerevisiae*, baker's yeast, had already been sequenced. Moreover, while a wealth of invaluable information is contained in the genome sequence — it is often referred to as the parts list of the human body — a quantitative understanding of the dynamics and the interplay of those components is needed.

While recent advances have provided us with a good understanding of many of the processes governing single molecules and their function and dynamics, a single molecule, as large or biologically relevant it may be, is not "alive" [117]. Life comes about, as we now start to grasp, as a dynamic phenomenon through the interplay of vast arrays of complex biomolecules and molecular machines. It is the promise of the emerging discipline of systems biology to deliver a quantitative understanding of biology. In other words it is the goal to understand the dynamic interplay of biomolecules that brings about life. The scope is shifting back from the atomic level to the whole cell or even whole organism level as we start to integrate the vast amounts of data.

Table 1.1 representatively lists an array of different high-throughput experimental technologies and databases that were recently developed. While the systems-level study of metabolites and their interactions and dynamics has had a longstanding tradition [107], high-throughput measurement technologies of metabolite concentration (*metabonomics*) are still under development [43, 90]. It is not limited by detection difficulties (using nuclear magnetic

resonance (NMR) , thousands of peaks are discernible), but by identification difficulties — even though NMR is a very mature experiment, it is still difficult to assign each peak to a chemical species. While there are currently no high-throughput techniques to measure interactions between small molecules and proteins directly, protein array technology is promising to be able to carry out that task [81, 137]. Past research has led to an already fairly comprehensive catalog of existing metabolic pathways, which have been summarized in databases such as MetaCyc [69], WIT [94] and KEGG [92].

Protein abundance as well as their phosphorylation state can be measured on a large-scale using mass-spectrometry and NMR techniques [48, 136]. Those techniques are still being developed and, mostly because of the high level of expertise and expense needed to carry out those experiments, not yet adopted by mainstream researchers. For quantitative dynamic models, time-resolved data is necessary; to this end, approaches using the green fluorescent protein (GFP) have been developed [103]. Using fluorescence microscopy it is possible to assess protein abundance with high accuracy and a time resolution of about 1 minute. For measurements of protein–protein interactions, two now mature technologies have emerged: yeast two-hybrid assays [28, 123] and protein co-immunoprecipitation in conjunction with tandem mass-spectrometry [33, 57]. Most large-scale screens tend to focus on yeast, but other model organisms, such as *H. pylori* [96] and *C. elegans* [130] have also been targeted. Protein–protein interactions are cataloged in large-scale databases such as DIP [133], BIND [6] or MIPS [86].

The most mature of the high-throughput technologies discussed are DNA microarrays, which measure the relative abundance of RNA in the cell, thereby giving the changes in the global gene expression profile [31, 108]. As of today, they are very comprehensive (for yeast and other model organisms, complete microarrays, covering every RNA species, exist), literally high-throughput (a single researcher can carry out several experiments per week, generating thousands of datapoints at a time) and well-characterized (while they are known to be noisy, models to assess and evaluate the noise exist [79]). Several databases, with large catalogs of expression profiles under different conditions such as TransciptionDB [1] have subsequently been assembled.

Finally, aside from the advances in gene sequencing technology mentioned above, protein–DNA interactions can be measured systematically using the technique of chromatin immunoprecipitation coupled with subsequent microarray measurements [78, 101]. Furthermore, extensive curation of the literature has led to databases of protein–DNA interactions such as TRANSFAC [132] or RegulonDB [105].

A vast availability of data and the development of new high-throughput experimental techniques are the prerequisite for the development of computa-

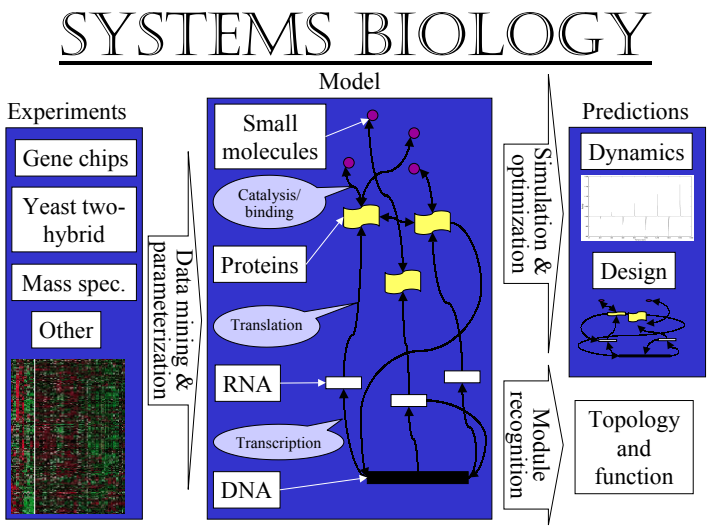

Figure 1.1: Systems biology - an overview

tional systems biology. While biologists used to study single genes and single proteins, these novel technologies allow us to carry out genome-wide experiments. The term "-omics" has been adopted as common suffix; the new fields are known as genomics, proteomics, physiomics, transcriptomics etc. All the data can be integrated to yield a complete description of the system, which can be at the sub-cellular, cellular or organism level.

1.3 Computational approaches in systems biology

While advances in experimental technology were the driving force in the field, computational and modeling methods do remain central in that they are required to integrate, organize and interpret the vast amounts of data generated by the above methods. Biological and medical sciences have largely remained empirical disciplines; the immense complexity of biological systems inhibited the emergence of theories or models with predictive power such as exist in the physical sciences. Despite tremendous advances, our understanding of

many macroscopic biological phenomena is still quite poor. It is the vision of computational systems biology to change that — to create models and theories that will not only be able to explain current data, but also to be able to predict the behavior of a biological system on a macroscopic scale. Applications in the biological or medical fields are numerous. For instance it is conceivable, with a model description of a human body at hand, to predict the effects of newly developed drugs thereby reducing the development time significantly. Of course, such large-scale models are still far in the future, the current focus of research is still to understand much of the cellular machinery and build models on a cellular level [72].

When envisioning such a model of even just a single cell, the incredible complexity of biological systems comes to mind immediately. Is it conceivable to grasp life with its great multitude of biomolecules, numerous processes and delicate detail in space and time? Most current models do not even come close to this regime of complexity; clearly a method for abstraction is needed. A currently widespread notion is that of so-called modular biology [52]. It is hypothesized that biological systems consist of separable, functionally independent subsystems in analogy to e.g. electronic systems; a modern microcomputer, despite its complexity with millions of transistors, can be understood by first separating it into its largely independent, but interacting components, using a divide-and-conquer approach. Translating this analogy to biological systems, it can be envisioned that a cell can be understood by first separating it into its subsystems, examining them one piece at a time and subsequently reintegrating them to rebuild the whole system. This approach may be a necessary intermediate step in scale; while going from the Ångstrom (10^{-10}) to the cellular (10^{-4}) scale, separate subsystems or modules have to understood on an intermediate scale.

Among computational approaches in systems biology, there exist three distinctly different branches (see Figure 1.1):

Data mining. The first branch focuses on extracting insight from the experimental data, which involves separating signal from noise and recognizing patterns and motifs. What is known as classical bioinformatics, such as sequence analysis and database structure work can be considered to be part of this branch. A particularly active subfield has been the development of gene expression microarray mining techniques [3, 15, 23, 119, 121]. Moreover, so-called data integration efforts have started recently, aiming to combine different kinds of data (from the sources mentioned in Section 1.2) to gain important biological insight [34, 35, 42, 67, 95, 121]. As depicted in Figure 1.1, from the modelers point of view, the focus of this branch is to construct the network model, preferably complete with kinetic rate constants for all its reactions. Also, it can be conceived that functional subsystems with some level

of functional autonomy might be recognizable from characteristic patterns in large-scale data. In Chapter 2 a novel data mining method for finding such biological subsystems in large-scale gene expression data is discussed.

Modeling – prediction and design. The second branch focuses on what is traditionally known as modeling. Given the data and knowledge gained from data mining approaches, network models are constructed and their behavior predicted using analytical methods or computer simulations. Also, insight gained from a working model can be used to design synthetic biological networks or optimize existing ones. Models can have different levels of abstraction, from the detailed molecular level model, describing each chemical species, to more abstract models, describing only select groups of molecules and averaging other more detailed effects [24, 26, 84, 97, 109, 125]. As this branch is a rapidly growing field, the need for a standardized language arose. Efforts have been made to standardize model descriptions to facilitate the sharing and exchange between research groups, through standards such as the systems biology markup language (SBML) [29, 73]. As mentioned above, current research focuses on smaller systems with the goal of later reintegrating the models. Hence, understanding the behavior of simple subsystems is pivotal to the field. In Chapter 3 an analytical and computational treatment to predict the behavior of simple genetic network subsystems is discussed and some limits of currently widespread models are demonstrated.

Topology and function. Modern science is obsessed with networks. While they correspond to vastly different systems in reality, ranging from protein interaction networks to networks of the world-wide-web, they all can be represented as graphs, consisting of nodes and edges [9]. In the network graphs studied in the case of systems biology, nodes usually correspond to proteins or genes whereas edges correspond to interactions. This subfield is driven by the hypothesis that just as the essential information about any given gene can in principle be read from its sequence, some of the information about the behavior and function of a network is encoded in its topological structure [2, 44, 68, 83, 98, 110]. The notion of biological subsystems can be translated easily into topological terms; intuitively, functional subsystems of a networks can be imagined to consist of strongly connected subnetworks. Understanding the network topology in terms of its separability could prove very valuable, as it could uncover functional subsystems. A study to this end is described in Chapter 4.

Chapter 2

Subsystem Identification through Dimensionality Reduction of Large-scale Gene Expression Data[1]

> "According to our experiences up till now, the assumption that nature is the realisation of what is mathematically simplest is justified."
> *Albert Einstein, 1879 - 1955*

Abstract

The availability of parallel, high-throughput biological experiments that simultaneously monitor thousands of cellular observables provides an opportunity for investigating cellular behavior in a highly quantitative manner at multiple levels of resolution. One challenge to more fully exploit new experimental advances is the need to develop algorithms to provide an analysis at each of the relevant levels of detail. Here the data analysis method non-negative matrix factorization has been applied to the analysis of gene array experiments. While current algorithms identify relationships based on large-scale similarity between expression patterns, non-negative matrix factorization is a recently developed machine learning technique capable of recognizing similarity between subportions of the data corresponding to localized features in

[1] P.M. Kim and B. Tidor, *Genome Research* (under revision)

expression space. A large data set consisting of 300 genome-wide expression measurements of yeast was used as sample data to illustrate the performance of the new approach. Local features detected are shown to map well to functional cellular subsystems. Functional relationships predicted by the new analysis are compared with those predicted using standard approaches; validation using bioinformatic databases suggests predictions using the new approach are roughly twice as accurate as conventional approaches.

2.1 Introduction

Gene expression microarrays are a recently developed technology that allows genome-wide measurement of RNA expression levels in a highly quantitative fashion [31, 41, 108]. Studies with microarrays generally produce large 2-D data sets (e.g., simultaneous monitoring of thousands of genes measured in up to hundreds of different experiments [16, 17, 60, 64, 71, 115]). The promise of this type of highly parallel and quantitative data is that they contain detailed and subtle information about relationships among cellular, biochemical, and genetic components that underlie the behavior of cells; the difficulty is that current approaches lead to data that are somewhat noisy [14, 18, 79], and the development of methods for exploring and extracting relationships within the data is still in its infancy.

The collection, processing, and analysis of microarray data present many challenges. Appropriate treatment of noise and systematic error is necessary to ensure that further analysis is not clouded by data inaccuracy, and some approaches have been proposed [13, 14, 79]. Methods of analysis must be developed that answer particular and relevant questions. Often these questions involve seeking and identifying patterns of similarity (correlation or anti-correlation) within the data. An array of methods capable of recognizing different types of similarity and similarity at different levels of resolution is needed. Moreover, the development of approaches to test individual hypotheses given a particular set of data and to more fully incorporate pre-existing models [36, 51, 61, 120] or other sources of information in the analysis is an important research area [11, 15, 39].

One productive use of expression data is to propose and to study relationships between genetic, cellular, or environmental components. Examples include the elucidation of metabolic [22, 27] or regulatory [59, 101, 121] networks. The standard methodology involves clustering of expression patterns based on similarity [3, 17, 23, 55, 112, 119, 121, 138].

The main assumption generally applied is that similar gene expression profiles imply related function. There are other techniques, many of which come from the machine learning community, capable of detecting similar-

ity or partially repeated patterns in large data sets. In principle, these techniques provide alternative approaches for recognizing potential relationships within large biological data samples, including expression arrays, that may complement existing methods. Here one such machine learning algorithm, non-negative matrix factorization (NMF), has been applied to the analysis of microarray data. One characteristic of NMF is that, using dimensionality reduction, it is capable of identifying patterns that exist in only a subset of the data [76]. For example, the application of clustering to recognize experimental conditions with similar patterns of gene expression focuses attention on conditions for which similarity extends across all genes. Another data analysis approach, singular value decomposition (SVD), also bases its description of the underlying data on global relationships that extend across essentially all the data has been recently applied to microarray data [89]. By contrast, NMF recognizes sets of experimental conditions in which smaller sets of genes behave in strongly correlated fashion. Thus, while other analysis methods examine global patterns in search of similarity and correlation, NMF is capable of finding smaller, more localized patterns as well as global patterns. Such an approach might be particularly useful in identifying biological subsystems (i.e., sets of genes that function in concert in a relatively tightly regulated manner) and might be an especially sensitive means for detecting functional genetic relationships.

Here the potential usefulness of NMF for the analysis of high-dimensional biological data was evaluated using a publicly available compendium microarray data set for *Saccharomyces cerevisiae* in which 6316 genes were monitored in each of 300 experiments [60]. Most of the experiments (276 of the 300) corresponded to deletion mutants of individual genes. In addition, 13 involved mutants with individual genes overexpressed using tetracycline regulated alleles and 11 involved wild-type cells treated with specific drugs. This data set spans a relatively wide set of significant cellular perturbations. The size of the data set is large by current standards, which presents a challenge for computational approaches but also an opportunity find patterns in what appears to be a particularly rich set of experiments. Analysis using NMF suggested that reduction of the data to a 50-dimensional subspace is appropriate. The lower dimensional subspace was capable of reconstructing the original data to high fidelity. The 50 vectors describing the subspace were relatively insensitive to moderate amounts of noise added to the original data set. The vectors described the local feature space detected by NMF and showed that each set of features was dominated by a few functional categories, indicating that they represent a grouping of genetic components based on cellular function. Individual pairwise functional relationships were scored based on standard approaches and, alternatively, using the similarity as measured by NMF. Scoring the relationships using the Munich Information Center for Pro-

tein Sequences functional categories (MIPS categories; http://mips.gsf.de/; [85]) and the Yeast Proteome Database (YPD; Proteome, Inc., Beverly, MA; http://www.proteome.com/; [20]) indicated that the new approach is significantly more reliable at predicting relationships than standard approaches. NMF appears to be a promising methodology, complementary to current approaches, for the analysis of high-dimensional biological data.

2.2 Results

The compendium data set contained expression patterns monitored for 6316 *S. cerevisiae* genes in 300 experiments involving a variety of strains and conditions. The expression of each gene in each experiment was represented as a ratio of the expression in the experiment to that in a control experiment of wild type grown under standard conditions. Genes whose expression in the control was not measurable were removed from the data set to prevent division by zero, leaving 5346 genes, and the natural logarithm of each ratio was taken. Data analysis involved using non-negative matrix factorization (NMF) to reduce the dimensionality of the data and to extract common features repeated in correlated fashion throughout the data (see Methods). These common feature elements were represented as basis vectors resulting from the technique. In typical usage, each basis vector represented an "experiment" in that it contained a relative expression for each gene comprising the feature represented.

Selection of NMF dimensionality. An essential feature of the NMF approach is that it reduces the data set from its full dimensionality (original data space) to a lower dimensional NMF space. Initial calculations were performed to select an appropriate size for the lower dimensional NMF space. Trial calculations carried out with NMF dimension of size ten to eighty suggested that fifty represented a good compromise that provided an adequate reconstruction of the experimental data while giving basis vectors that appeared to recognize repetitive features. The RMS error between the original and NMF reconstructed data is shown as a function of the size (dimensionality) of the NMF space in Figure 2.1.

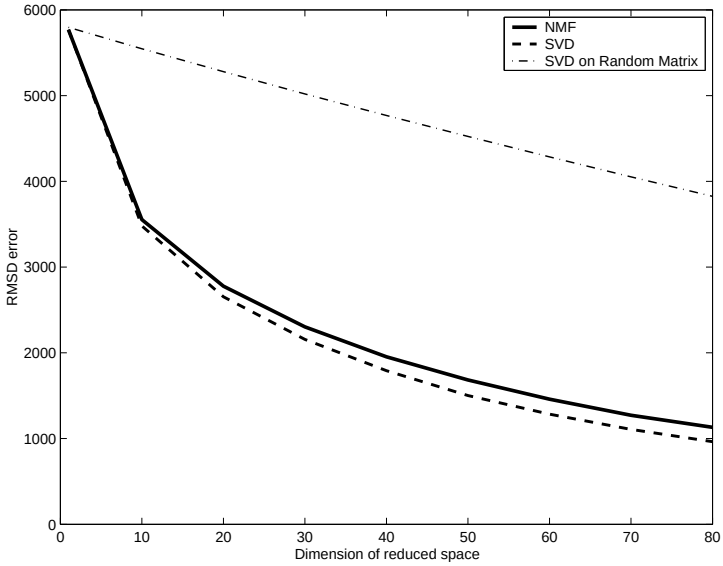

Figure 2.1: The RMS error of NMF and SVD factorizations of the original data as a function of the number of dimensions in the reduced space. For comparison, SVD factorization was also carried out on a random matrix based on the data matrix. The results show that NMF is nearly as good as SVD at reproducing the original data at for any dimensionality and that near a dimensionality of about 50 the marginal increase (slope) in NMF's ability to describe the original data is similar to SVD's ability to match random (unstructured) data. Thus, an NMF dimensionality of 50 is appropriate to describe the structure in the data.

Also shown in the Figure is the RMS error for singular value decomposition (SVD), which is another matrix factorization approach that is guaranteed to produce the minimum error for a given dimensionality (but does not generally extract localized features from complex data sets). SVD was applied both to the actual data matrix and to a random matrix of values selected from a gaussian distribution with the same mean and variance as the data matrix and subject to the non-negativity constraint. The close similarity between the error for NMF and SVD on the actual data indicated that the computational procedures used for NMF were effective (details given in Methods). The random matrix could be viewed as one without correlated features to be detected through factorization; the slope of the RMS error plot for this matrix represents the added ability to reproduce unstructured data with additional basis vectors. Below a dimensionality of fifty, the NMF factorization curve had a steeper slope than the random matrix line, which indicated improvements due to capturing organization and structure within the data. This further justified the choice of fifty for the NMF dimension. Interestingly, a previous study using expression arrays to study yeast also found an inherent dimensionality of fifty [4].

Basis vectors ("basis experiments") obtained from NMF factorization with a dimensionality of fifty were sparse and reproducible. One measure of sparsity is the fraction of non-zero entries per basis vector, which averaged 5.5% over the fifty vectors. The factorization produced somewhat different results each time it was started from a different random starting point. When the basis vectors from different factorizations using the same dimensionality were compared, the correlation coefficient was found to be greater than 0.9 between pairs. This indicates that results of NMF are robust with respect to the mathematical procedures used here to perform the calculations. The RMS error of the reconstructed data (through NMF dimensional reduction) compared to the original data was only about 7.8% of the RMS error of a random permutation of the original data, in which the experiments (columns) of the original data matrix were permuted.

Figure 2.2 illustrates six examples of expression experiments in the original gene expression space, in the fifty-dimensional NMF space, and reconstructed from the NMF space back into the original space. This shows the ability of the dimensionality reduction to still capture many of the details of the original data. Also, it is demonstrated that every experiment is represented by the combination of only a few important basisvectors. They correspond to similarities across many but not all genes.

To examine the robustness of the algorithm to noise, gaussian noise was added to the original data to produce "corrupted" data vectors. Table 2.1 lists the average correlation between results of the analysis performed on the original and corrupted data. Noise was added in progressively larger increments

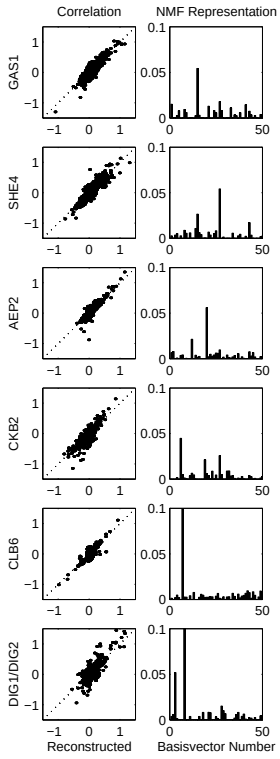

Figure 2.2: Representation of gene expression data in full and NMF-reduced spaces. In the left-hand column the log-ratios of the original (y-axis) are plotted against the log-ratios of the reconstruction from the NMF representation back to the original experimental space (using $\overleftrightarrow{W} \cdot \overleftrightarrow{H}$, x-axis) for six individual experiments in the space of 5346 genes. The 50-dimensional NMF representation of the data is shown in the right-hand column. The data show that the NMF reduction is capable of regenerating the experiments to relatively high fidelity and that the NMF representation of an experiment is often dominated by one or a small number of features (basis vectors).

of the standard deviation of the data. At low noise (0.2 times the standard deviation) there was very little change in the results. The correlation of NMF vectors was better than 0.90 as was that for the data reconstructed from the dimensionality reduction. This is not surprising, because the original data vectors and corrupted vectors also showed a correlation coefficient of greater

Noise Added	NMF Basis vectors	Reconstructed Data	Original Data
0.2	0.933	0.93	0.943
0.5	0.879	0.893	0.781
1	0.865	0.816	0.573
5	0.368	0.313	0.159

Table 2.1: Robustness of NMF basis vectors to noise. Gaussian noise was added to the original data and was quantified as a multiplier of the standard deviation of the original data set. NMF basis vectors: The average correlation of basis vectors from the original data to the basis vectors from original data with added noise. Reconstructed data: The average correlation of the reconstructed data from the basis vectors with and without noise. Original data: Average correlation of the original data to the data with added noise.

than 0.90. However, when adding more noise (equal to the full standard deviation) both the NMF basis vectors as well as the reconstructed data were still very similar after adding noise (correlation of better than 0.80), whereas the original data was changed substantially more (correlation of 0.57). This fact shows the high robustness of NMF to noise in the data, and suggests that NMF might be useful as a noise-reduction filter in certain applications.

Annotation of basis vectors. Each of the fifty basis vectors (basis experiments) contained many genes with zero expression and others with non-zero expression. The genes with non-zero expression were used to assign sets of functional categories to basis vectors using the MIPS classification scheme (see Methods [85]), and the results are listed in Table 2.2. Each basis vector appeared to be dominated by only a few functional categories, with some categories showing increased and others decreased expression relative to wild-type, untreated cells. Basis vector 17, for example, showed increased expression of genes associated with amino-acid metabolism and metabolism of energy reserves together with decreased expression of genes involved in rRNA transcription. Basis vector 20 showed increased expression of genes involved in ion transport, homeostasis of cations, and mitochondrial organization with decreased expression of genes for amino-acid metabolism, ribosomal proteins, translation, and organization of cytoplasm. Basis vector 9 showed increased expression of genes associated with carbon compound (C-compound) and carbohydrate metabolism and transporters as well as metabolism of energy reserves, and at the same time decreased expression of amino-acid metabolism genes. In some cases, specific metabolic pathways could be seen in the basis vectors. For instance most elements of

the TCA-cycle were upregulated in basis vector 43. Furthermore, this basis vector, which seemed mostly responsible for energy metabolism, contained all but two of the genes involved in the pentose-phosphate shunt. Of these two genes, one is a transketolase that is highly homologous to another transketolase found in basis vector 43, and the other is the ribose-5-phosphate ketol isomerase. In 16 of the basis vectors, no single MIPS category was significantly enriched, which is partly due to the lack of sparsity (i.e., too many genes occur in a basis vector so no single category was significant) and partly due to an abundance of as yet uncategorized genes.

1	+1 amino-acid metabolism (204 ORFs) [82] +2 nitrogen and sulphur metabolism (74 ORFs) [27] +81 stress response (169 ORFs) [43] –34 ribosomal proteins (206 ORFs) [98] –35 translation (62 ORFs) [22] –92 organization of cytoplasm (557 ORFs) [163]
3	+1 amino-acid metabolism (204 ORFs) [21]
4	+81 stress response (169 ORFs) [9]
8	+21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [24] –4 phosphate metabolism (31 ORFs) [3]
9	+5 C-compound and carbohydrate metabolism (413 ORFs) [115] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [21] +47 C-compound and carbohydrate transporters (46 ORFs) [23] –1 amino-acid metabolism (204 ORFs) [53]
17	+1 amino-acid metabolism (204 ORFs) [40] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14] –29 rRNA transcription (104 ORFs) [17]
19	+13 respiration (85 ORFs) [18] +100 mitochondrial organization (364 ORFs) [27] –1 amino-acid metabolism (204 ORFs) [18]
20	+34 ribosomal proteins (206 ORFs) [29] +46 ion transporters (76 ORFs) [13] +88 homeostasis of cations (112 ORFs) [17] +100 mitochondrial organization (364 ORFs) [53] –1 amino-acid metabolism (204 ORFs) [18] –34 ribosomal proteins (206 ORFs) [24] –35 translation (62 ORFs) [7] –92 organization of cytoplasm (557 ORFs) [40]
23	+11 tricarboxylic-acid pathway (23 ORFs) [4] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [6] –4 phosphate metabolism (31 ORFs) [3]
36	+29 rRNA transcription (104 ORFs) [41] –11 tricarboxylic-acid pathway (23 ORFs) [10] –15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14] –5 C-compound and carbohydrate metabolism (413 ORFs) [77]
42	+1 amino-acid metabolism (204 ORFs) [40] +6 lipid, fatty-acid and isoprenoid metabolism (210 ORFs) [21] +81 stress response (169 ORFs) [22] –21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [11]
43	+5 C-compound and carbohydrate metabolism (413 ORFs) [126] +10 pentose-phosphate pathway (9 ORFs) [7] +11 tricarboxylic-acid pathway (23 ORFs) [17] +81 stress response (169 ORFs) [64]
	Continued on next page

	Continued from previous page
	–29 rRNA transcription (104 ORFs) [58] –34 ribosomal proteins (206 ORFs) [87]

Table 2.2: Annotation of 12 of the 50 NMF basis vectors based on the MIPS functional categories. Each annotation includes a plus or minus sign (indicating whether expression is enhanced or decreased compared to control experiments), an integer number indexing the MIPS category, the name of the MIPS category, the number of ORFs belonging to the MIPS category, and the number of genes in the basis vector belonging to the MIPS category (in square brackets). The full set of 50 basis vectors is provided as supplementary information.

Independent of the classification scheme proposed by MIPS, the occurrence of well-characterized gene groups was examined in basis vectors. The processed data set contained 9 histone genes, which were all present together in basis vector 1. This enrichment was over 5σ higher than what would occur by chance. Aside from histone genes, basis vector 1 was also strongly enriched in ribosomal genes, genes related to translation, and genes involved in amino-acid and nitrogen metabolism. Similarly, the data set contained 109 ribosomal genes, of which 70 appeared in basis vector 1 and 52 in basis vector 43. Basis vector 43 was involved in energy metabolism, stress response, and rRNA transcription. The enrichment of 70 ribosomal genes in basis vector 1 was 26σ higher than what would occur by chance. Between basis vectors 1 and 43, all but 17 ribosomal genes were found.

Next, the occurrence of genes in both the *GAL4* and the *STE12* pathway was examined. These pathways were recently studied extensively by Ren et al. [101]. No deletion mutant of any of the genes involved in the *GAL4* pathways was present in the compendium data set; therefore, no significant enrichment of those genes might be expected. However, of the 9 genes present in the data, 5 were enriched in basis vector 9. This enrichment was 5σ higher than would be expected by chance. Basis vector 9 was also involved in C-compound and carbohydrate metabolism. It seemed that basis vector 9 was responsible for a broad range of functions relating to carbohydrate metabolism, the degradation of galactose being a subset of those.

There was a deletion mutant of *STE12* in the data, along with several mutants of related genes, including *FUS3*, *KSS1*, and *STE5*. A total of 25 genes, forming a subset of those identified by Ren et al. [101] as members of the *STE12* pathway, were present in the processed data. A majority, 16 out of the 25, were present in basis vector 8 (a significance of 16σ higher than expected by chance). The genes included *PRM1* (linked to membrane biosynthesis), *FIG2*, *AGA1*, *FUS1* (cell fusion), *GIC2* (mating projection formation), *CIK1*, *KAR2* (nuclear fusion), *FUS3*, *STE12*, and *HYM1* (mating signaling) along with other genes of yet unknown relevance (*YOR0343C*, *PEP1*,

SCH9, YIL036C, YIL083C, YOL155C, CIK1). It should be noted that *all* genes to which Ste12p binds (as identified by Ren et al. [101]) before and after α factor addition were included in this list (a total of 8 genes; 17 additional genes which were in the preprocessed data set were shown to bind Ste12p only after α factor addition). Basis vector 8 also included large contributions from genes represented in the MIPS database as involved in mating signaling and pheromone response, indicating related cellular functions for the genes dominating this basis vector. Another MIPS category which was found to be enriched in basis vector 8 with 4.5σ (just below the cutoff implemented of 5σ) was membrane biosynthesis, which is consistent with the appearance of *PRM1*, classified as an effector in membrane biosynthesis.

Basis vector 8 (the mating basis vector) was then examined more closely and the function of all its member genes examined using information from the Yeast Proteome Database (YPD), constructed by Proteome, Inc. (Beverly, MA; http://www.proteome.com/; [20]). The YPD is a compilation of published results of yeast genes (*S. cerevisiae*) and their functions, including functional relationships reported in the literature. Aside from the 16 genes described above, it contained 15 other genes involved in mating or pheromone response (*TEC1, KAR4, PRM3, PGU1, YLR042C, DDR48, PRM5, SAG1, HAP4, SST2, MSG5, AGA1, PRM4, SAG1, KSS1*), 6 of which (underlined) were annotated in YPD as directly induced by *STE12*. Furthermore, this vector contained 10 genes (*ECM18, SPI1, CHS7, GFA1, KTR2, SCW10, WSC3, STR2, GSC2, PHD1*) involved in cell wall or cell membrane biosynthesis or maintenance. Among its other members were several genes involved in carbohydrate metabolism (*GLK1, SOL4, GPH1, GLC3*), heat shock or stress response (*HSP26, HSP30, PRY2, DDR48*), and many ORFs of yet unknown function (*YDR124W, YDR537C, PTI1, YGR250C, YHR213W, YIL060W, YIL082W, YIL083C, YJR026W, YJR027W, YJR028W, SRL3, YLR177W, YLR334C, YLR422W, YOL106W, YOR296W, SVS1*) as well as a few genes of other functionality (*ADR1, BNA1, FRE2*). Besides examining the genes that contribute strongly to the basis vector, it is informative to examine which of the 300 experiments in the compendium were described using a large contribution from this basis vector. Basis vector 8 was mostly used to describe experiments of deletion mutants of *DIG1/DIG2* (double deletion; *DIG1* is a known *STE12* repressor), *DIG1* (single deletion), and *FUS3* (linked to mating and pheromone response). Note that the data set did not contain a *DIG2* single deletion mutant.

Prediction of functional relationships. The compendium data set analyzed here was dominated by measurements of gene expression in deletion strains of yeast compared to wild type (276 of 300 experiments). Of the remaining experiments, 13 were measurements of single-gene overexpression

relative to wild type. Thus, all but 11 experiments involved direct manipulation of a single gene in a common background. (The 11 exceptions involved measurements of wild-type yeast treated with a single drug relative to untreated wild type.) Experiments showing similar (correlated or anti-correlated) changes in gene expression at some level might be expected to be functionally related. In particular, the two genes different in the two experiments could be expected to be part of the same or related cellular function. To test this hypothesis, predictions of functional relationships were made and scored against available database information. Moreover, predictions based on correlations in the entire gene space were compared to those from dimensionally reduced spaces, such as that produced by NMF, to understand whether dimensionality reduction can enhance the detection of known genetic relationships. One difficulty with any approach of this type is that available database information is likely to be incomplete and may be partially inaccurate. Nevertheless, a method's ability to recapitulate current knowledge is a good indicator of its ability to predict new relationships. Thus, the score these methods achieve in validated functional relationships should only be interpreted relative to each other, since many true functional relationships may be missing from current databases.

Predictions of functional relationships were made using the pairwise correlations between experiments measured in each of six spaces — the original data space, the 50-dimensional NMF space, and four other 50-dimensional spaces chosen for comparison. The six spaces are (i) the original space in which the data was collected, corresponding to 5346 genes used in the analysis, (ii) the 50-dimensional space resulting from NMF data reduction, (iii) the 50-dimensional space spanned by 50 genes whose expression varied the most across the 300 experiments, (iv) the 50-dimensional space spanned by the 50 genes whose expression varied the least across the 300 experiments, (v) the 50-dimensional space spanned by the 50 genes whose expression variance was closest to average across the 300 experiments, and (vi) the 50-dimensional space explaining the largest variation in the experimental data as found by singular value decomposition (SVD). For each case the pairwise correlations were sorted by magnitude, with the higher magnitude correlations corresponding to stronger predictions. Predictions were checked against the MIPS database (see Methods), and the results are shown in Figure 2.3.

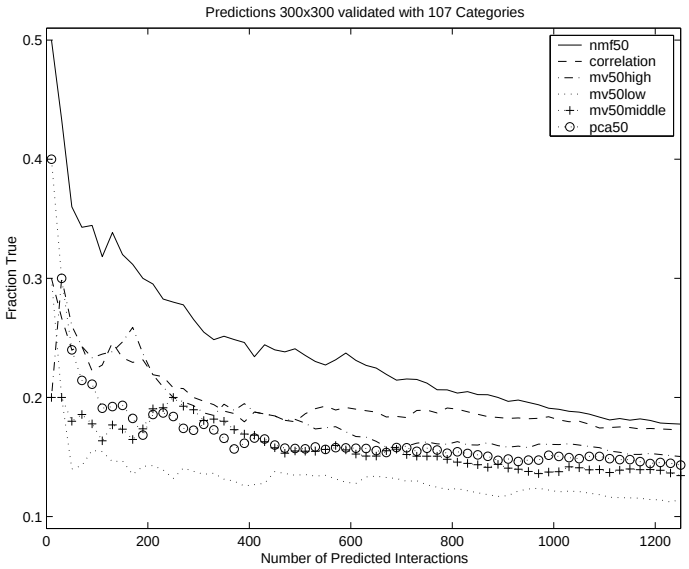

Figure 2.3: Performance of different spaces at predicting functional relationships between experiments with comparison to the MIPS classification of the deleted genes. NMF50: NMF space with 50 basis vectors. Original Space: Original gene expression space. SVD50: SVD space with 50 eigenvectors. MV50high, MV50middle, MV50low: Space of the 50 most, middle and least varying genes.

The Figure shows for each of the six methods, the percentage of predictions validated by MIPS as a function of the number of predictions made (when ordered from strongest to weakest correlation). In general the methods exhibited the highest validation for their strongest predictions (except for MV50high). For up to 600 predicted relationships (four per gene, on average), NMF far outperformed all other methods. For instance, for the 100 strongest predictions the reliability in the NMF space was about 35% whereas for all other spaces, including the original gene expression space, only 15% to 25% of the predictions were validated. Beyond 800 predicted relationships, correlations in the original space did almost as well as NMF. However, the false positive rate at this level of prediction is likely to be too high to be useful. The reliability of predictions dropped off sharply for all spaces and eventually reached 9%, which was the probability of making a true prediction from the data set by chance.

A second and independent method was used to evaluate the predictions of functional relationships produced by NMF by comparing to data compiled in the Yeast Proteome Database (YPD; [20]). For the purposes of this study the relationships reported by YPD were categorized as "hard" (indicating a direct measure of interaction, such as binding, or participation in the same pathway) and "soft" (indicating an indirect detection, such as co-expression). Examination of the strongest 100 predictions from NMF found that 58% were validated by querying YPD (38% were hard and 20% were soft functional relationships). This compared with about 35% of the same set that were verified through the MIPS database. The 58 validated predictions are listed in Table 2.3, and the 42 predictions that were not validated (but are testable predictions nonetheless) are listed in Table 2.4. Applying the same procedure to the strongest 100 predictions from the original gene expression space produced only 31% that could be verified by YPD (19% were hard and 12% soft). Thus, using the Yeast Proteome Database, dimensionality reduction through NMF appeared to be roughly twice as productive in predicting functional relationships as correlation (essentially clustering) in the original space of the data.

Co-regulated	
dfr1	ecm34
gyp1	yap7
ade16	sir1
hpt1	sir1
rml2	ymr293c
cbp2	mrpl33
mrpl33	rml2
cnb1	yor072w
ade16	ymr041c
gfd1	utr4
cla4 (haploid)	KAR2 (tet promoter)
yel001c	ymr141c
ckb2	gcn4
arg5,6	rpl8a
mrt4	rpl12a
clb6	whi2
erp2	ymr141c
erp2	yel001c
erp2	yor015w
rpl12a	yel033w
ckb2	rtg1
eca39	ras1
Identical Genes	
isw1	isw1, isw2
dig1, dig2	dig1, dig2 (haploid)
fks1 (haploid)	FKS1 (tet promoter)
bub3	bub3 (haploid)
Binding	
cla4 (haploid)	CDC42 (tet promoter)
qcr2 (haploid)	rip1
far1 (haploid)	ste4 (haploid)
bub1 (haploid)	bub3
bub1 (haploid)	bub3 (haploid)
Cell Wall	
fks1 (haploid)	2-deoxy-D-glucose
2-deoxy-D-glucose	Glucosamine
gas1	Tunicamycin
fks1 (haploid)	Glucosamine
yer083c	Tunicamycin
	Continued on next page

	Continued from previous page
ste12 (haploid)	ste5 (haploid)
Mating	
ste5 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste5 (haploid)
ste18 (haploid)	ste5 (haploid)
ste12 (haploid)	ste18 (haploid)
ste18 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste18 (haploid)
fus3, kss1 (haploid)	ste7 (haploid)
fus3, kss1 (haploid)	ste12 (haploid)
ste12 (haploid)	ste7 (haploid)
Ergosterol Pathway	
erg3 (haploid)	Itraconazole
erg2	Itraconazole
yer044c (haploid)	ERG11 (tet promoter)
ERG11 (tet promoter)	Itraconazole
erg3 (haploid)	ERG11 (tet promoter)
erg3 (haploid)	yer044c (haploid)
erg2	erg3 (haploid)
erg2	yer044c (haploid)
erg2	ERG11 (tet promoter)
Vacuolar ATPase	
cup5	mac1
mac1	vma8
cup5	vma8

Table 2.3: The 58 predictions that could be validated by YPD of the 100 strongest functional relationships detected by NMF. “Co-regulated” genes were found to be co-regulated by other functional genomics studies. “Binding” refers to genes whose proteins have been shown to bind each other. “Cell Wall,” “Mating,” and “Ergosterol Pathway” are all genes that have been experimentally shown to be involved in the named cellular function.

2.3 Discussion

Here non-negative matrix factorization (NMF), a new machine learning approach capable of identifying localized features in complex data sets, was applied to the analysis of microarray data from a series of 300 yeast experiments (of which 276 were deletion strains; [60]). The essence of NMF is that

the algorithm must choose a small number of features (basis vectors) to act as building blocks that can be scaled and added together in various combinations to best reconstruct the original data. Restriction to a small number of basis vectors causes the algorithm to select patterns of genes that occur frequently in the data. The application of a data analysis approach that extracts localized data features from a set of experiments that span a wide range of genetic variation holds the potential to be a particularly powerful method to detect functional cellular subsystems (the features encoded in the basis vectors) as well as individual pairwise functional genetic relationships.

The experimental variation sampled by the 300 experiments could be well represented with just 50 features. Moreover, this set of 50 features encoded in the basis vectors tended to correspond to sets of known functional genetic groupings of genes. Large numbers of genes involved in similar or related cell functions appeared together due to a local similarity in their expression profiles. It should be noted that because of limited data (i.e., not all yeast deletion strains were sampled) not all cellular functions were identified. Some cellular systems were sampled more in the experiments than others. For example, the mating and pheromone grouping is particularly well identified. Basis vector 8 consisted mostly of genes involved in mating and even contained 6 verified targets of *STE12* that were not identified by previous studies.

Pairwise relationships between experiments were evaluated by locating pairs of experiments that were constructed from the same NMF building blocks (basis vectors). With this detection scheme, NMF was far superior to any other method examined, including other sets of 50 basis vectors constructed from other procedures as well as standard correlations in the full gene expression space of the original experiments. The initial analysis of this same compendium data set reported by Hughes et al. [60] used conventional clustering methods and found a series of very interesting and useful relationships between genes. Many of the genes that were clustered together in that study were also scored as related in the current work. For example, the sections headed “Ergosterol Pathway,” “Cell Wall,” “Mating,” and “vacuolar ATPase” in Table 2.3 contain many relationships also detected by Hughes et al. [60] using more standard techniques; however, most of the other validated relationships evaded detection by conventional techniques. This includes the section headed “Binding” in the Table, for which particularly strong experimental validation is available.

Figure 2.4 illustrates the increased similarity seen in the NMF feature space compared to that seen in the original data space for 4 pairwise functional relationships from Tables 2.3 and 2.4. Two of these (yer084w:*SBH2* and ymr025w:ymr029c) were not corroborated by YPD, whereas the other two (*STE5*:*STE11* and *RTS1*:*RTG1*) are each known to be functional relationships. As the numerical values in the Figure indicate, the correlation in NMF

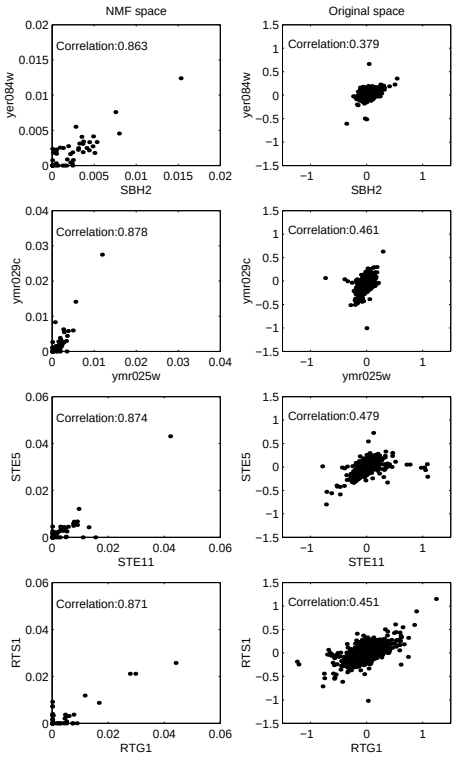

Figure 2.4: Correlation for four illustrative pairwise functional genetic relationships. For comparison the correlation plot of the pair of experiments in NMF space is shown on the left and in the original gene space on the right.

space was significantly higher than in the original gene expression space. Essentially, this stems from the fact that NMF recognized the expression patterns of strains deleted for the genes in question as being constructed from very similar sets of building blocks, and the correlation in the expression pattern was larger for the genes comprising these building blocks. For instance, the expression profiles for strains deleted in *STE11* and in *STE5* were each dominated by basis vector 8 (the building block consisting largely of mating genes) and had relatively small (but still correlated) contributions from other basis vectors. NMF recognized this local similarity across some genes, whereas most clustering algorithms would focus only on the global similarity of the expression profile. Comparing the same two strains in the original data space shows that their gene expression patterns were highly correlated

for some genes but not for others. Therefore, NMF is a way to focus on the functionally important parts of gene expression profiles.

	Gene
rtg1	vps8
are1, are2 (haploid)	yor015w
pex12	yea4
ckb2	yel008w
yer002w	ymr034c
mrt4	yel033w
ckb2	rts1
mrpl33	ymr293c
imp2	yer050c
cbp2	pet111
cyt1	pet111
yer034w	ynd1
rps24a	ymr014w
yel001c	yor015w
ymr014w	yor006c
aep2	rml2
aep2	mrpl33
ymr014w	yor078w
rml2	yer050c
mrpl33	yer050c
aep2	imp2
sir1	ymr041c
ymr034c	yor015w
pfd2	yor051c
ymr025w	ymr029c
ckb2	vps8
msu1	ymr293c
sbh2	yer084w
mrpl33	msu1
imp2	ymr293c
rtg1	rts1
msu1	yer050c
msu1	rml2
yml003w	ymr034c
aep2	msu1
CDC42 (tet promoter)	KAR2 (tet promoter)
rps24a	yor078w
pfd2	yel044w
gcn4	yel008w
yer050c	ymr293c
	Continued on next page

	Continued from previous page
aep2	yer050c
aep2	ymr293c

Table 2.4: The 42 predictions of functional relationships that could not be verified on YPD from the 100 strongest relationships detected.

Table 2.4 lists 42 predictions of functional relationships detected by NMF but not present in YPD. Some predicted relationships are between genes classified as mitochondrial (e.g., *AEP2*:*YER050C* and *MSU1*:*MRPL33*) just as some of the verified relationships are between mitochondrial genes (e.g., *RML2*:*YMR293C*). Moreover, a number of small networks of mitochondrial genes occur in the strongest 100 NMF relationships; most genes in these networks were clustered together in the original analysis of the data by Hughes et al. [60]. Another tight network of functional relationships can be seen among *CKB2*, *YEL008W*, *GCN4*, *RTS1*, *RTG1*, and *VPS8*, some of which were and some of which were not verified by YPD. The existence of these tight interconnected relationships adds to the likelihood that the predictions are, indeed, correct.

While this manuscript was in preperation, two studies [33, 56] which focused on the large scale identification of protein–protein interactions in yeast were published. For the study by Ho et al. [56] data was readily available online and we used it as additional means of verifying predictions. There was little overlap in scope with our study, as of a total of identified 8114 interactions only 74 fall within the set of 276 gene deletions in our study. Allowing for one connecting link between interacting proteins, this number increases to 2001. Out of our best 100 predictions an additional 2 are verified (between *CKB2* and *VPS8* and between *CDC42* and *KAR2*). It should be noted that both come from the set of 42 interactions where no direct link was found on YPD, i.e. that none of the 58 predictions which were verified on YPD were found by Tyers and Coworkers. The little overlap with our predictions may hence stem from the sparseness of both datasets. When scoring our predictions using the interactions found by Tyers et al., we still find that predictions made using NMF have a higher likelihood of being correct than ones made from pure correlation (data not shown).

One feature of the approach taken here is that pairwise relationships were only scored for genes that had been directly manipulated in the experiments (deleted or overexpressed). As described in Methods, NMF can also be applied to detect relationships between genes that have been monitored using expression arrays but not directly manipulated experimentally. Preliminary studies using NMF in this mode suggest that it is again superior in detecting functional genetic relationships compared to approaches that apply clustering

or correlation directly in the original data space. A further shortcoming that remains, however, is the elimination from analysis of genes whose expression is undetectable in the control experiments (to avoid division by zero). Functional relationships involving such genes (comprising roughly one-sixth of the genome for the current data set) can not be scored. In future studies it may be possible to insert a minimal expression level for such genes in the control experiment, although further work is necessary to see whether this introduces other problems, such as feature mis-scaling.

In the current study no separate attempt was made to smooth or filter the data set to reduce or eliminate the effects of experimental noise or error. In some sense, NMF itself performs a smoothing function on the data through factorization and reconstruction. Features that appear consistently in the data set are selected out to become basis vectors, whereas features that appear inconsistently in the data due to experimental variability or other factors tend to be smoothed. For the results reported here, only genes with no detectable expression in the control experiment were removed. When more stringent significance filters were applied to the data, the results remained similar (data not shown).

Of the 50 basis vectors resulting from this analysis, many were sparse (that is, they represented features consisting of a relatively small number of genes). However, some basis vectors were not sparse and contained too many genes to be easily annotated as associated with a small number of cellular functions. The NMF algorithm could be modified to enforce sparser basis vectors; alternatively, it is anticipated that larger data sets will result in basis vectors that are more uniformly sparse and may correspond to smaller features. Indeed, an advantage of NMF is that it is expected to be a better detector of features when confronted with larger data sets.

2.4 Methods

General approach. Data from a set of expression array experiments were represented as a single matrix $\overleftrightarrow{V}$. Each column corresponded to the processed intensities from one experiment; each element of a column was derived from the intensity for one gene probe in the corresponding experiment. A row of the matrix corresponded to the processed intensity for a single gene probe across all experiments. An $n \times m$ matrix $\overleftrightarrow{V}$ corresponded to m arrays (i.e., experiments) in which measurements were made for the same n genes in each. The major analysis method applied here, non-negative matrix factorization, corresponded to an approximate factorization of the matrix $\overleftrightarrow{V}$ into a pair of

matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$.

$$\overleftrightarrow{V} \approx \overleftrightarrow{W} \cdot \overleftrightarrow{H} \tag{2.1}$$

The factorization was chosen with a particular rank, k, so that $\overleftrightarrow{W}$ was of dimension $n \times k$ and $\overleftrightarrow{H}$ was $k \times m$. In the work described here, k was chosen to be relatively small compared to the dimensions of the original data $\overleftrightarrow{V}$ (that is, $k \cdot (m+n) < n \cdot m$) so the factorization was approximate and corresponded to a compression of the data. Moreover, the factorization could be viewed as a representation of the data in a new space of lower dimensionality (k). There are two equally valid interpretations of the dimensionality reduction. One is that the columns of $\overleftrightarrow{W}$ were "basis experiments" (having the dimensionality of a single array or experiment), and each row of $\overleftrightarrow{H}$ was the representation of a particular experiment in the new k-dimensional space. Alternatively, the rows of $\overleftrightarrow{H}$ were "basis genes," and each column of $\overleftrightarrow{W}$ then corresponded to a representation of a particular gene in the new space. The unique feature of non-negative matrix factorization is that none of the matrices in equation (2.1) ($\overleftrightarrow{V}$, $\overleftrightarrow{W}$, or $\overleftrightarrow{H}$) are permitted to have negative entries [76].

Implementation of NMF. The NMF algorithm was coded using the mathematics and matrix algebra package MATLAB version 6 (R12) (Mathworks Inc., Waltham, MA). The key features of the algorithm involved iteratively improving matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$ to improve the approximation to $\overleftrightarrow{V}$ while maintaining non-negative matrix entries throughout. This was achieved using using an update-rule approach [76]. For a given value of the NMF dimensionality k, the algorithm was started with random matrices $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$. The random initial seed was a uniform distribution of real numbers from 0 to 1 for all matrix elements of $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$. The two matrices were iteratively updated using the rules,

$$\overleftrightarrow{H}_{a\mu} \leftarrow \overleftrightarrow{H}_{a\mu} \frac{(\overleftrightarrow{W}^T \overleftrightarrow{V})_{a\mu}}{(\overleftrightarrow{W}^T \overleftrightarrow{W} \overleftrightarrow{H})_{a\mu}} \tag{2.2}$$

$$\overleftrightarrow{W}_{ia} \leftarrow \frac{(\overleftrightarrow{V} \overleftrightarrow{H}^T)_{ia}}{(\overleftrightarrow{W} \overleftrightarrow{H} \overleftrightarrow{H}^T)_{ia}} \tag{2.3}$$

which minimize the root-mean-square (RMS) error ($E = \| \overleftrightarrow{V} - \overleftrightarrow{W} \cdot \overleftrightarrow{H} \|_2$) between the actual data $\overleftrightarrow{V}$ and the reduced-dimension reconstruction of the data ($\overleftrightarrow{W} \cdot \overleftrightarrow{H}$; [77]). Because the update rules were multiplicative, initial non-negative matrices remained non-negative for all future iterations. Iterations

were continued until the RMS error change in an iteration was less than 0.1 in absolute RMS error, which corresponded to roughly 0.005% of the final RMS error.

The update rules corresponded to a form of gradient descent and thus found only a local minimum. To address this limitation, the procedure was repeated 100 times starting with different initial matrices. The factorization leading to the lowest RMS error was used in further analysis. Studies were carried out for values of the NMF dimensionality (k) ranging from 10 to 80. The solutions found were reproducible; basis vectors from factorizations that differed in the initial matrices showed correlation coefficients of over 0.90.

A single NMF factorization for a 5346×300 data set required approximately 30 minutes of CPU time on a 500 MHz Pentium III workstation and occupied roughly 70 MB of memory. The current implementation was dominated by matrix multiplication, leading to computation times that scaled as the number of matrix entries raised to roughly the power 1.35 (typical of matrix multiplication in MATLAB and other modern packages). The relative simplicity of the update-rule implementation does not require first- or second-derivative information, which would add significantly to memory usage. Memory requirements scaled linearly with data set size due to the need to store data and factor matrices.

Trial implementations on smaller test problems were also carried out with non-linear optimizers CONOPT2 version 2.071G (ARKI Consulting & Development A/S, Bagsvaerd, Denmark) and LOQO version 4.01 (Princeton University, Princeton, NJ); the values of matrix elements in $\overleftrightarrow{W}$ and $\overleftrightarrow{H}$ were optimized directly and subject to non-negativity constraints in order to minimize the RMS error. While quite successful on small problems, these methods require additional memory for storage of the gradient and were thus not feasible for the data set analyzed here.

To ensure sparsity of the resulting basis vectors, the most significant genes for every basis experiment were selected using a fixed percentage of the maximum gene (9.7%) in every basis vector. After selecting of the most significant genes, all other genes were constrained to zero and the resulting "sparsified" basis vectors were re-optimized to convergence using the update rules in Equation (2.3).

In separate calculations, singular value decomposition (SVD) of data matrices was carried out using the built-in functionality in MATLAB. A representation of an SVD factorization of rank k corresponded to using only the k highest eigenvalues. Absolute RMS error values were calculated for the same data set and the same ranks as for NMF. Furthermore, as a control SVD was carried out on random matrices comprised of vectors of the same mean and standard deviation as the sample data.

Annotating basis vectors. The functional categorizations available at the Munich Information Center for Protein Sequences (the MIPS categories) were used to assign genes to biochemical pathways or cellular function. There are a total of 107 MIPS categories that cover different metabolic pathways, such as the TCA cycle and glycolysis, as well as different cellular functions, such as cell membrane biosynthesis or mating [85]. Some of the categories overlap (for example the category *glycolysis* is a subset of the category *energy metabolism*), and one gene can be assigned to more than one category.

Each basis vector (basis experiment) was annotated with the MIPS categories that dominated its makeup by comparing the frequency with which genes from each category appeared in a basis vector with that expected from a random distribution. One million genes were selected at random from the same set of genes present in the experimental data. The corresponding MIPS categories were identified and the mean and the standard deviation of occurrence was calculated for every category. This procedure was carried out twice to ensure convergence of the random distribution. If the occurrence of a particular MIPS category in a basis vector exceeded the mean of the random occurrence by more than 5 times its standard deviation (a 5σ cutoff), this particular category was assigned to the basis vector as enriched. As a negative control, basis vectors were generated from random numbers and subjected to the same significance cutoffs and annotation procedure. In 1000 random basis vectors, no category was ever assigned as being enriched.

Predicting functional relationships. An important test of the utility of data reduction using non-negative matrix factorization was to assess its ability to predict functional relationships between genes. To predict functional relationships between genes or experiments based on expression data, it is typical to assume that similarity in expression suggests a functional relationship between genes or experiments. Here the same assumption was made both in the original space of the data and in the reduced dimensional spaces, such as that computed by non-negative matrix factorization. The Pearson correlation coefficient was calculated between all genes (or all experiments), and the absolute value of the correlation coefficient was used as a predicted score for the relationship. This method scored positive and negative correlations equally strongly.

In the data set used, most experiments corresponded to deletion mutants of a specific gene, so that functional relationships between experiments in turn implied functional relationships of the deleted genes. Other experiments corresponded to the overexpression of genes, which again linked the experiments directly to the gene in question. The rest of the experiments corresponded to treatment with a well-characterized drug. Those experiments then linked the response in expression pattern to the functional mechanism of this

particular drug.

To judge the predicted functional relationships between genes required some set of "true relationships." For this purpose, existing bioinformatic databases were used, though clearly such data are largely incomplete and may not be fully verified. The two databases used were the MIPS categorization [85] and the YPD [20]. Two genes appearing in the same MIPS category were scored as functionally related (e.g., two genes encoding ribosomal proteins). The MIPS categorization was checked for every gene in the data set, as well as for every gene for which there was a deletion or overexpressed mutant in our data set, yielding a list of validated interactions. Predictions from the correlation score in NMF space were compared to this list, starting with the predictions of highest correlation.

The functional relationships predicted from gene expression data using NMF were compared to functional relationships predicted from other approaches. The same analysis and validation procedure was applied to the correlation score in five other spaces: the original full-dimensionality of the experimental space, reduced dimensionality using SVD with the fifty most significant dimensions, reduced dimensionality using only the fifty most variable genes in the data set, reduced dimensionality using only the fifty least variable genes in the data set, and reduced dimensionality using only the fifty genes whose variance is closest to average in the data set. The value of fifty was chosen to compare different same-sized reduced-dimension representations of the data to that from NMF.

A second and independent method of scoring predicted functional relationships used the Yeast Proteome Database (YPD; [20]). YPD contains detailed information about genetic or physical interaction, functional relationships and co-regulation of all genes in yeast. The information in YPD is based on a large number of papers from the scientific literature. Results catalogued in YPD include those from biophysical, molecular biological, genetic, and functional genomic experiments. The strongest 100 predictions from NMF and from correlations in the original experimental data space were scored against YPD. Any link in YPD between two genes (e.g. co-regulation, genetic interaction, or binding) was viewed as a validation of the prediction. Moreover, for the "soft" validations, one linking gene was permitted. That is, if gene A interacted with gene Z and gene B was co-regulated with gene Z in YPD, then gene A and B were scored as co-regulated. For this purpose, at least one of the two relations was required to be a "hard" interaction.

Data source and preprocessing. This study is based on analysis of a large, publicly available microarray data set from Rosetta Inpharmatics Inc. encompassing genome-wide expression data of *Saccharomyces cerevisiae* in 276 deletion mutants, 11 tetracycline regulated alleles of essential genes (overex-

pression) and 13 wt-strains treated with well-characterized drugs (a total of 300 experiments; [60]). All experiments used the Saccharomyces Genome Deletions Consortium strain background. Most of the deletion mutants were diploid mutants (i.e., both alleles were deleted from the genome). For some essential genes, haploid mutants were made. This impaired but did not remove the gene function. The strains were grown according to a standard protocol and in parallel with corresponding wild-type control cultures.

Gene expression was measured using spotted microarrays, giving the ratio of expression in the mutant (or drug-treated) strain relative to the gene expression in the control (wild-type) experiments. The spotted arrays measured expression for a total of 6316 ORFs; the data set was 6316 genes by 300 experiments (data available from Rosetta Inpharmatics Inc. at http://www.rii.com/register/cell2000102Hughes/EULA.htm). It is likely that much of the "yeast gene expression space" is sampled in this data set, which spans very different conditions; thus, it appears a good data source in which to seek gene expression features.

The log-transformed ratios were used as input data for our algorithm; the transformed ratios ranged from approximately –3 (1000 times downregulated with respect to the control experiment) to +3 (1000 times upregulated). Some genes had no detectable expression in the control experiment and were removed from further analysis to prevent division by zero. The resulting data set contained 5346 genes. To make the data fit the constraint of non-negativity, the data were "folded." Every gene was represented in two rows of the matrix, the first occurrence to indicate positive expression relative to wild-type and the second to indicate negative. This effectively doubled the size of the data set (to 10692 genes). In any one experiment the log-expression ratio for every gene was either positive (i.e., the gene was upregulated with respect to the control experiment) or negative. The resulting data matrix was of size 10692×300 and half its entries were equal to zero.

2.5 Acknowledgments

We thank H. Sebastian Seung, Michael D. Altman, Justin A. Caravella, Gerald R. Fink, David F. Green, Chris Kaiser, Sriram Kosuri, Douglas A. Lauffenburger, Robert T. Sauer, Anthony J. Sinskey, Peter K. Sorger, and Shari Spector for helpful discussions and suggestions. This work was partially supported by the Alfred P. Sloan Foundation. PMK was a Merck/MIT Graduate Fellow and is a Ph.D. Fellow of the Boehringer Ingelheim Fonds.

2.6 Appendix

	Supplementary Information
1	+1 amino-acid metabolism (204 ORFs) [82] +2 nitrogen and sulphur metabolism (74 ORFs) [27] +81 stress response (169 ORFs) [43] –34 ribosomal proteins (206 ORFs) [98] –35 translation (62 ORFs) [22] –92 organization of cytoplasm (557 ORFs) [163]
2	+81 stress response (169 ORFs) [7] –21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [12] –80 intracellular communication (131 ORFs) [7] –91 organization of plasma membrane (143 ORFs) [7]
3	+1 amino-acid metabolism (204 ORFs) [21]
4	+81 stress response (169 ORFs) [9]
5	
6	+15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [5] –1 amino-acid metabolism (204 ORFs) [78] –2 nitrogen and sulphur metabolism (74 ORFs) [16] –48 amino-acid transporters (25 ORFs) [7]
7	+1 amino-acid metabolism (204 ORFs) [14]
8	+21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [24] –4 phosphate metabolism (31 ORFs) [3]
9	+5 C-compound and carbohydrate metabolism (413 ORFs) [115] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [21] +47 C-compound and carbohydrate transporters (46 ORFs) [23] –1 amino-acid metabolism (204 ORFs) [53]
10	+46 ion transporters (76 ORFs) [11] +88 homeostasis of cations (112 ORFs) [15] –11 tricarboxylic-acid pathway (23 ORFs) [3] –13 respiration (85 ORFs) [6] –46 ion transporters (76 ORFs) [5] –88 homeostasis of cations (112 ORFs) [6]
	Continued on next page

	Continued from previous page
	–100 mitochondrial organization (364 ORFs) [9]
11	–1 amino-acid metabolism (204 ORFs) [49] –2 nitrogen and sulphur metabolism (74 ORFs) [20] –10 pentose-phosphate pathway (9 ORFs) [5]
12	+1 amino-acid metabolism (204 ORFs) [36] +5 C-compound and carbohydrate metabolism (413 ORFs) [60] +11 tricarboxylic-acid pathway (23 ORFs) [9] +13 respiration (85 ORFs) [21] +90 organization of cell wall (33 ORFs) [10] +100 mitochondrial organization (364 ORFs) [53] +105 extracellular/secretion proteins (20 ORFs) [8] –34 ribosomal proteins (206 ORFs) [59] –92 organization of cytoplasm (557 ORFs) [92]
13	–34 ribosomal proteins (206 ORFs) [13]
14	+5 C-compound and carbohydrate metabolism (413 ORFs) [77] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [16] +47 C-compound and carbohydrate transporters (46 ORFs) [17] +55 drug transporters (35 ORFs) [15] +91 organization of plasma membrane (143 ORFs) [31] –1 amino-acid metabolism (204 ORFs) [22] –48 amino-acid transporters (25 ORFs) [6]
15	+1 amino-acid metabolism (204 ORFs) [48] +2 nitrogen and sulphur metabolism (74 ORFs) [16] –34 ribosomal proteins (206 ORFs) [33] –100 mitochondrial organization (364 ORFs) [33]
16	
17	+1 amino-acid metabolism (204 ORFs) [40] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14] –29 rRNA transcription (104 ORFs) [17]
18	
19	+13 respiration (85 ORFs) [18] +100 mitochondrial organization (364 ORFs) [27] –1 amino-acid metabolism (204 ORFs) [18]
20	+34 ribosomal proteins (206 ORFs) [29] +46 ion transporters (76 ORFs) [13] +88 homeostasis of cations (112 ORFs) [17] +100 mitochondrial organization (364 ORFs) [53] –1 amino-acid metabolism (204 ORFs) [18] –34 ribosomal proteins (206 ORFs) [24] –35 translation (62 ORFs) [7] –92 organization of cytoplasm (557 ORFs) [40]
21	+34 ribosomal proteins (206 ORFs) [39] +92 organization of cytoplasm (557 ORFs) [46]
	Continued on next page

	Continued from previous page
22	
23	+11 tricarboxylic-acid pathway (23 ORFs) [4] +15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [6] –4 phosphate metabolism (31 ORFs) [3]
24	-34 ribosomal proteins (206 ORFs) [41]
25	
26	+81 stress response (169 ORFs) [33] –34 ribosomal proteins (206 ORFs) [23]
27	+5 C-compound and carbohydrate metabolism (413 ORFs) [59] +13 respiration (85 ORFs) [36] +90 organization of cell wall (33 ORFs) [11] +100 mitochondrial organization (364 ORFs) [52] –1 amino-acid metabolism (204 ORFs) [23] –4 phosphate metabolism (31 ORFs) [7]
28	
29	
30	+29 rRNA transcription (104 ORFs) [20] –11 tricarboxylic-acid pathway (23 ORFs) [7] –46 ion transporters (76 ORFs) [15] –100 mitochondrial organization (364 ORFs) [52]
31	–52 allantoin and allantoate transporters (9 ORFs) [3]
32	–34 ribosomal proteins (206 ORFs) [48] –47 C-compound and carbohydrate transporters (46 ORFs) [18]
33	
34	+5 C-compound and carbohydrate metabolism (413 ORFs) [50] +47 C-compound and carbohydrate transporters (46 ORFs) [10] +81 stress response (169 ORFs) [28]
35	
36	+29 rRNA transcription (104 ORFs) [41] –5 C-compound and carbohydrate metabolism (413 ORFs) [77] –11 tricarboxylic-acid pathway (23 ORFs) [10] –15 metabolism of energy reserves (glycogen, trehalose) (37 ORFs) [14]
37	
38	+1 amino-acid metabolism (204 ORFs) [32]
39	
40	
41	+1 amino-acid metabolism (204 ORFs) [76] +2 nitrogen and sulphur metabolism (74 ORFs) [16] –6 lipid, fatty-acid and isoprenoid metabolism (210 ORFs) [10] –105 extracellular/secretion proteins (20 ORFs) [3]
42	+1 amino-acid metabolism (204 ORFs) [40] +6 lipid, fatty-acid and isoprenoid metabolism (210 ORFs) [21] +81 stress response (169 ORFs) [22]
	Continued on next page

	Continued from previous page
	–21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [11]
43	+5 C-compound and carbohydrate metabolism (413 ORFs) [126] +10 pentose-phosphate pathway (9 ORFs) [7] +11 tricarboxylic-acid pathway (23 ORFs) [17] +81 stress response (169 ORFs) [64] –29 rRNA transcription (104 ORFs) [58] –34 ribosomal proteins (206 ORFs) [87]
44	–52 allantoin and allantoate transporters (9 ORFs) [3]
45	
46	
47	
48	–21 pheromone response, mating-type det., sex-spec. proteins (159 ORFs) [22] –80 intracellular communication (131 ORFs) [13]
49	
50	

Table 2.5: Annotations of the 50 basis vectors as determined by NMF based on the MIPS categories. Each annotation includes a plus or minus sign (indicating whether expression is enhanced or decreased compared to control experiments), an integer number indexing the MIPS category, the name of the MIPS category, the number of ORFs belonging to the MIPS category, and the number of genes in the basis vector belonging to the MIPS category (in square brackets).

Chapter 3

The Topology and Behavior of Genetic Networks: A Mathematical Treatment[1]

> "Plurality should not be assumed without necessity."
> *William of Occam, 1285 - 1347*

Abstract

Predicting the behavior of a genetic network upon perturbation from its topological structure is a central problem in the field of systems biology. Recent work by Leibler and co-workers showed that topologically identical networks can exhibit qualitatively very different behavior to a symmetric set of perturbations [45]. Using a set of three genes (LacI, TetR and lambda CI) and five promoters, they constructed different genetic networks. The networks can act a logic gates since they have two inputs through the effector molecules of LacI (IPTG) and TetR (aTc). GFP, which is regulated by cI, acts as an output. They show that two symmetric networks, having the role of LacI and TetR exchanged, exhibit non-symmetric behavior to the same input.

Here we develop a general and rigorous mathematical framework to analyze the behavior of this type of synthetic network. We start from the following commonly used assumptions: 1. We assume that there is no spatial dependence of any protein or RNA concentration (i.e., the inside of the cell is well-stirred). 2. We assume that there is transcription level control only.

[1] P.M. Kim and B. Tidor, *Genome Research* (submitted)

3. We assume that there is no cross-talk between the promoters. 4. Central to our treatment is the steady-state assumption, which we relax in the subsequent analysis. 5. The dependence of the transcription and translation rate on protein and RNA concentration, respectively, is assumed to be strictly monotonic.

We prove that, despite the generality of the model, which accounts for all imaginable parameters and nonlinear functional dependencies of rates on molecule concentrations, the behavior observed by Guet et al. [45] in some networks can not be reconciled with it. However, the assumptions used to construct the model are widely used, and it is important to understand possible sources for the discrepancy. We explore relaxing model assumptions to explain the observed behavior, allowing for both dynamic and stochastic phenomena, and propose an alternative model. Our alternative model includes the suggestion of a new mechanism by which the counterintuitive behavior could be achieved; central to the model is the assumption that the Clp protein degradation system, which is responsible for the regulatory proteins used in this study, becomes saturated.

Moreover, the framework developed here is general and independent of rate constants and can be applied to other systems.

3.1 Introduction

It is a central problem in systems biology to predict the behavior of a genetic network. Various papers in recent years have made predictions about network behavior, either through computational modeling or analytical theory [10, 26, 84, 97, 109, 113, 114, 125]. Also, several levels of modeling have been carried out, from molecular level simulations up to very abstract models of cellular scale. While models have varied greatly in style and scope, virtually all attempts to predict behavior has depended on a (sometimes very large) number of parameters, which were mostly adapted from the literature (sometimes from in-vitro experiments, i.e. the parameter had to be adjusted), inferred or fitted. Since measurements of biochemical kinetic or thermodynamic parameters in vivo is quite cumbersome and difficult, it is undesirable for model predictions to be strongly dependent on fitted parameters.

In this work we do not assume nor fit any numerical values for any parameters, nor do we assume any explicit functional dependence. Rather, mathematical relationships are constructed that permit qualitative predictions of network behavior. Moreover, we suggest an extension to a commonly used model, that could explain the observed behavior.

Here we explore the behavior of a set of genetic circuits constructed and studied by Guet et al. [45]. Our goal is to understand how two networks with the same topology but interchanged roles of two regulatory elements can exhibit different behaviors. Our original hypothesis, that parameter differences for the two regulatory elements could explain behavior differences, is shown not to be true, at least in the context of models usually employed to describe genetic circuitry.

The synthetic networks constructed by Guet et al. [45] consist of the three genes LacI, cI and TetR, which are combinatorially assigned promoters to which the three repressors each bind. This allows for 27 different network topologies when restricting the promoters to be all of repressing nature. These systems were then perturbed by adding saturating amounts of anhydrotetracycline (aTc) and isopropyl-β-thiogalactopyranoside (IPTG), the effectors of TetR and LacI, respectively. As read out, green fluorescent protein (GFP), which was under repression of the third gene, cI, was used. The GFP level was measured under the 4 different conditions, i.e. without effector, with ITPG, with aTc, and with both effectors added.

3.2 Methods

Assumptions and general model. We begin with the following assumptions, which are frequently used in this field of research [5, 24, 122].

- We assume no spatial dependence of the molecule concentrations or rate constants, i.e. the inside of the cell is well-stirred. This enables us to treat the system with ordinary differential equations.
- We assume no crosstalk between promoters, i.e. every repressor only binds to the promoter it is designed for.
- We assume that control of expression takes place at the transcription level only.
- We will assume steady state at first, but later show that relaxing the assumption does not affect results.
- The dependence of translation and transcription rates on protein and RNA concentration, respectively, is assumed to be strictly monotonic. However, we do not further constrain the dependence of translation or transcription rates on protein or RNA concentration.

The most general model for gene expression looks as follows:

$$\dot{r}_i = \quad \text{deg}_{r_i}(r_i) + \text{tr}_{p_i}(p_{y_i}) \tag{3.1}$$

$$\dot{p}_i = \quad \text{deg}_{p_i}(p_i) + \text{tl}_{p_i}(r_i) \tag{3.2}$$

p_i and r_i stand for protein and RNA concentrations, respectively; $\text{deg}_{r_i}(r_i)$ stands for the RNA degradation rate of RNA r_i, $\text{tr}_{r_i}(p_{y_i})$ stands for the transcription rate of RNA r_i as a function of the repressor concentration p_{y_i} that controls RNA r_i expression, $\text{deg}_{p_i}(p_i)$ stands for protein degradation rate and $\text{tl}_{p_i}(r_i)$ stands for the rate of translation of r_i into p_i. Note that $\text{deg}_{r_i}(r_i)$ and $\text{deg}_{p_i}(p_i)$ will always be negative, as they correspond to the reduction of RNA/protein concentration through degradation. Most of the assumptions are implicit in the notation; we assume further that $\text{tr}(p)$ is strictly monotonically decreasing for every repressor p, $\text{tl}(r)$ strictly monotonically increasing and all degradation rates are strictly monotonically decreasing. The network topology is encoded in the y_i. They determine which transcription factor represses which gene, i.e. each p_{y_i} is one of the p_i.

Simplification of the equations. Now we assume steady state and show properties of the relationship between transcription factor concentration and steady-state protein expression:

$$0 = \quad \text{deg}_{r_i}(r_i) + \text{tr}_{y_i}(p_{y_i}) \tag{3.3}$$

$$0 = \quad \text{deg}_{p_i}(p_i) + \text{tl}_{p_i}(r_i) \tag{3.4}$$

Since we assume the degradation rates to have strictly [2] monotonic dependence on concentration we can invert Equation 3.3 and then eliminate r_i from Equation 3.4:

$$r_i = \quad \text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i})) \tag{3.5}$$

$$0 = \quad \text{deg}_{p_i}(p_i) + \text{tl}_{p_i}(\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i}))) \tag{3.6}$$

$$\text{Finally:} \quad p_i = \quad \text{deg}_{p_i}^{-1}(-\text{tl}_{p_i}(\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i}))) \tag{3.7}$$

Now note that $\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i}))$ is strictly monotonically decreasing, since with rising p_{y_i},$-\text{tr}_{y_i}(p_{y_i})$ will get larger (smaller in absolute value) and $\text{deg}_{r_i}^{-1}$ is strictly monotonically decreasing. Likewise, $-\text{tl}_{p_i}(\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i})))$ is strictly monotonically increasing and finally, $\text{deg}_{p_i}^{-1}(-\text{tl}_{p_i}(\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_{y_i}))))$ is strictly monotonically decreasing. Since we do not assume any further information about the degradation or translation and transcription functions aside from their strict monotonicity, we can replace the right-hand side of Equation 3.7 with one function that is strictly monotonically decreasing:

$$p_i = f_{iy_i}(p_{y_i}) \tag{3.8}$$

We can now describe any system of genetic networks built from genes that repress one another with a system of equations similar to Equation 3.8, the network topology is encoded in the label y_i. These equations will define the steady-state behavior of the system, if it is in fact fully defined. In other words, the steady-state level of any given protein has monotonically decreasing dependence of the concentration of a repressor controlling its expression . This is also an intuitive result, but here it has been determined assuming only strict monotonicity and a simple but commonly applied model. The formalism can be extended to include activators.

Validity of the steady-state assumption. For any gene with an autoregulatory loop the following equations hold for its expression:

$$\dot{r} \quad = \text{deg}(r) + \text{tr}(p) \tag{3.9}$$

$$\dot{p} \quad = \text{deg}(p) + \text{tl}(r) \tag{3.10}$$

The jacobian will then be:

$$\begin{pmatrix} \frac{\partial \text{deg}(r)}{\partial r} & \frac{\partial \text{tr}(p)}{\partial p} \\ \frac{\partial \text{tl}(r)}{\partial r} & \frac{\partial \text{deg}(p)}{\partial p} \end{pmatrix}$$

[2] We need the strict monotonic property of the degradation rates in order to be able to compute the inverses, $\text{deg}_{r_i}^{-1}$ and $\text{deg}_{p_i}^{-1}$.

From monotonicity restrictions we get that $\frac{\partial \text{deg}(r)}{\partial r}, \frac{\partial \text{tr}(p)}{\partial p}, \frac{\partial \text{deg}(p)}{\partial p} < 0$ and $\frac{\partial \text{tl}(r)}{\partial r} > 0$. For eigenvalues λ_1 and λ_2,

$$\frac{\partial \text{deg}(r)}{\partial r} \cdot \frac{\partial \text{deg}(p)}{\partial p} - \frac{\partial \text{tr}(r)}{\partial p} \cdot \frac{\partial \text{tl}(r)}{\partial r} = \lambda_1 \cdot \lambda_2$$

From restrictions above, $\lambda_1 \cdot \lambda_2 > 0$.

$$\frac{\partial \text{deg}(r)}{\partial r} + \frac{\partial \text{deg}(p)}{\partial p} = \lambda_1 + \lambda_2$$

From here we get $\lambda_1 + \lambda_2 < 0$; therefore $\lambda_1, \lambda_2 < 0$ for all p, r. This fact makes the steady-state values for r_0 and p_0 an asymptotically stable point according to the Ljapunov-Poincare theorem [116]. Furthermore, if r_0 and p_0 define the steady state, for any $r_1 > r_0$ and $p_1 > p_0$, one can see from Equations 3.9 and 3.10 that $\dot{r}_1$ and $\dot{p}_1$ will be negative (since $\dot{r}_0 = \dot{p}_0 = 0$), therefore bringing the system to the steady state. The same arguments can be applied to the cases of $r_1 < r_0$, $p_1 > p_0$ etc. so that from every r,p steady state will eventually be reached in this model.

Monte Carlo simulations. Monte Carlo simulations were carried out using the algorithms developed by Gillespie [37]. The algorithm was implemented in the mathematics programming package MATLAB (The Mathworks, Natick, MA). Simulations were run on an AMD Athlon 1200-MP workstation and took less than 20 minutes to carry out. Initial parameters for the simulations were adapted from Elowitz et al. [24], where a very similar system was modeled. However, while Elowitz et al. [24] modelled both TetR and LacI as dimers and emulated a Hill model by sequential binding of two proteins, we use a Hill model (perfect cooperativity) with Hill coefficients of 2 and 4 for TetR and LacI, respectively. Binding to promoter was modeled with 1 $\text{nM}^{-1}\text{sec}^{-1}$, unbinding from promoter with 9 sec^{-1}, transcription from occupied promoter with 0.0005 sec^{-1}, transcription from unoccupied promoter with 0.5 sec^{-1}, translation with 0.167 $\text{mRNA}^{-1}\text{sec}^{-1}$, protein degradation with a half-life of 10 minutes and mRNA degradation with a half-life of 2 minutes. A model according to Elowitz et al. [24] yielded similar results (data not shown). We sampled parameter space stochastically and ran simulations for 500 different parameter sets. For every new set of parameters, each parameter was allowed to vary over a range of 3 orders of magnitude. The parameters were randomized by taking 10 to the power of a random number from a uniform distribution from -1.5 to 1.5 and multiplying the result with the parameter value. We carried out each simulation separately with and without effector addition. Each simulation was run four times to steady state.

Simulations for the extended model were carried out using the same methodology, with addition the binding and degradation reactions of Clp

3.3 Results

Application to the case of combinatorially synthesized networks by Guet et al. [45] yields three equations as follows:

$$p_{LacI} = f_{LacI,y_{LacI}}(p_{y_{LacI}}) \tag{3.11}$$
$$p_{TetR} = f_{TetR,y_{TetR}}(p_{y_{TetR}}) \tag{3.12}$$
$$p_{cI} = f_{cI,y_{cI}}(p_{y_{cI}}) \tag{3.13}$$
$$y_i \in cI, LacI, TetR \tag{3.14}$$

It is important to note that while the functions f_{iy_i} can all be different, they also all share the property that they are monotonically decreasing in their argument. When only considering networks consisting of mutual repressors, there are a total of 27 different topologies possible. Also, in this case the symmetry between the three different genes is broken, since in the experiments designed by Guet et al. [45] LacI and TetR are used as inputs (through its effectors, IPTG and aTc, respectively). Lambda cI is used as output, by its coupling to GFP, which can then be read out. To model the behavior observed in the experiments, we predict the change in p_{cI} when adding IPTG, aTc, or both. Again, we can use the very general framework to model this perturbation of the system. The following equation describes the behavior when the effector which acts on p_{y_i} (i.e. aTc for TetR and IPTG for LacI) is added:

$$p_i = \text{deg}_{p_i}^{-1}(-\text{tl}_{p_i}(\text{deg}_{r_i}^{-1}(-\text{tr}_{y_i}(p_i^e))) \tag{3.15}$$

So we merely have to replace the function $tr_{y_i}(p)$ with a term $tr_{y_i}(p^e)$, where p^e is the concentration of active repressor in the presence of effector. Because of the high effector concentration used in the experiments, here we treat the effector as inactivating its repressor ($p^e \ll p$). [3] Instead of Equation 3.8 we then get the following:

$$p_i = f_{iy_i}(p_{y_i}^e) \tag{3.16}$$

From Equation 3.15 we can see that $f(p^e) > f(p)$ will hold [4]. We shall write that as:

$$p_i = f_{iy_i}^e > f_{iy_i}(p_i) \tag{3.17}$$

And we treat $f_{iy_i}^e$ as a constant without dependence on p.

In 9 out of the 27 cases, the following equation will hold.

$$p_{cI} = f_{cI,y_{cI}}(p_{cI}) \tag{3.18}$$

[3] Guet et al. [45] added aTc at 100ng/ml and IPTG at 1mM, which is known to be about 10-fold higher than the saturating concentration of each effector [80]. Note that our treatment becomes somewhat less general by assuming effector saturation.

[4] We can apply simple arguments about monotonicity here. Since we know that $tr(p^e) > tr(p)$ and all functions in 3.16 are monotonic, we can deduce that $f(p^e) > f(p)$.

This equation defines a steady-state in terms of the read out variable. This is the case where λ cI is repressing itself and hence in the current model, addition of either effector is not expected to change the level of cI at all. The predicted behavior for all networks is shown in Table 1. While for most networks, our model does make a prediction of the change in GFP level upon effector addition, it cannot predict changes for some cases. In those cases, the monotonicity constraint is not strong enough to make predictions; further specification of functional form and parameters are needed. While Guet et al. [45] employed two LacI promoters of different characteristics, no case is shown where the difference in promoter affected the qualitative behavior of the network. This in agreement with our model, for which results are independent of promoter characteristics.

For two particular networks we will show the results of our modeling framework. For the network D038 (Number 27 in Table 1) we get the following equations:

$$p_{LacI} = f_{LacI,TetR}(p_{TetR}) \tag{3.19}$$
$$p_{cI} = f_{cI,LacI}(p_{LacI}) \tag{3.20}$$
$$p_{TetR} = f_{TetR,TetR}(p_{TetR}) \tag{3.21}$$

Equation 3.21 implicitly defines a steady-state value p^0_{TetR}. This steady-state will be assumed eventually (see Methods). Then p^0_{cI} and p^0_{LacI} are also defined:

$$p^0_{LacI} = f_{LacI,TetR}(p^0_{TetR}) \tag{3.22}$$
$$p^0_{cI} = f_{cI,LacI}(p^0_{LacI}) \tag{3.23}$$

When now adding aTc, Equation 3.21 will become (because of the inactivation through binding of the protein product of TetR):

$$p_{TetR} = f^{aTc}_{TetR,TetR} \tag{3.24}$$

This equation defines another steady-state, p^{aTc}_{TetR} which can be shown to be always larger than p^0_{TetR}[5]. Due to a similar argument it follows that $p^{aTc}_{LacI} > p^0_{LacI}$ and $p^{aTc}_{cI} < p^0_{cI}$. Hence, we expect GFP levels to be higher than without effector, since a lower concentration of p_{cI} will lead to less repression of GFP production.

Upon addition of IPTG, Equation 3.21 becomes:

$$p_{cI} = f^{IPTG}_{cI,LacI} \tag{3.25}$$

[5]We have seen in Equation 3.17 that $f^e = f(p^e) > f(p)$, therefore also $p^{aTc}_{TetR} = f^{aTc,e}_{TetR,TetR} > f_{TetR,TetR}(p^0_{TetR})$

And it is easy to see that the following holds:

$$p_{LacI}^{IPTG} = p_{LacI}^{0} \tag{3.26}$$
$$p_{cI}^{IPTG} > p_{cI}^{0} \tag{3.27}$$
$$p_{TetR}^{IPTG} = p_{TetR}^{0} \tag{3.28}$$

Thus, we expect GFP levels to be lower than without IPTG.

Adding both IPTG and aTc we can see following analogous arguments:

$$p_{LacI}^{IPTG/aTc} > p_{LacI}^{0} \tag{3.29}$$
$$p_{cI}^{IPTG/aTc} > p_{cI}^{0} \tag{3.30}$$
$$p_{TetR}^{IPTG/aTc} > p_{TetR}^{0} \tag{3.31}$$

Again, we expect lower GFP levels.

In summary, we expect the GFP levels to be the highest in the case with aTc added which is in accordance with experimental findings [45]. However, upon IPTG addition, the model predicts a GFP level less than the level without IPTG,whereas the experiments reported no change. Note that the model does allow for the GFP level change to be very small, and possibly undetected by experiments. It does rule out, however, a higher GFP level.

However, for a topologically equivalent network, the experimental results cannot be reconciled with current the model: In the network D052 (No. 18 in Table 1) the following equations describe the behavior of gene expression:

$$p_{LacI} = f_{LacI,LacI}(p_{LacI}) \tag{3.32}$$
$$p_{cI} = f_{cI,TetR}(p_{TetR}) \tag{3.33}$$
$$p_{TetR} = f_{TetR,LacI}(p_{LacI}) \tag{3.34}$$

Here Equation 3.33 implicitly defines a solution p_{LacI}^{0}. We then can also solve the other equations:

$$p_{cI}^{0} = f_{cI,TetR}(p_{TetR}^{0}) \tag{3.35}$$
$$p_{TetR}^{0} = f_{TetR,LacI}(p_{LacI}^{0}) \tag{3.36}$$

When now adding IPTG, Equation 3.33 will become

$$p_{LacI} = f_{LacI,LacI}^{IPTG} \tag{3.37}$$

The solution to this equation (p_{LacI}^{IPTG}) can be shown to be always larger than p_{LacI}^{0}. Also:

$$p_{LacI}^{IPTG} > p_{LacI}^{0} \tag{3.38}$$
$$p_{cI}^{IPTG} < p_{cI}^{0} \tag{3.39}$$
$$p_{TetR}^{IPTG} > p_{TetR}^{0} \tag{3.40}$$

Hence, GFP levels should be higher than without effector. In the presence of aTc, Equation 3.34 becomes:

$$p_{cI} = f^{aTc}_{cI,TetR} \tag{3.41}$$

Its hence easy to see that the following holds:

$$p^{aTc}_{LacI} = p^0_{LacI} \tag{3.42}$$
$$p^{aTc}_{cI} > p^0_{cI} \tag{3.43}$$
$$p^{aTc}_{TetR} = p^0_{TetR} \tag{3.44}$$

Thus, we expect the GFP levels to be lower than without effector. Adding both IPTG and aTc we can see following analogous arguments:

$$p^{IPTG/aTc}_{LacI} > p^0_{LacI} \tag{3.45}$$
$$p^{IPTG/aTc}_{cI} > p^0_{cI} \tag{3.46}$$
$$p^{IPTG/aTc}_{TetR} < p^0_{TetR} \tag{3.47}$$

We expect again lower GFP levels than without effector. In summary, we expect the GFP levels to be the highest in the case with IPTG added. Experimentally it is found that they are the highest when no effector is added, which is in clear contradiction with the predictions of the model.

The network D052 is topologically equivalent to D038, the roles of TetR and LacI are switched. However, the symmetry is broken through the facts that the promoters used (P_{LtetO1} and P_{LlacO1}) have different repression thresholds and Lac is a tetramer whereas $TetR$ is a dimer [80]. Therefore it might be supposed that the difference in parameters and Hill coefficients could lead to the fundamentally different behavior that was observed. Our results show that this is not case, since for our general model, the two networks are perfectly equivalent and are predicted to behave similarly (network D038 perturbed with aTc should act like D052 perturbed with IPTG and vice versa).

The behavior of all other networks can be derived using analogous arguments and the resulting predicted behavior is shown in Table 1.

3.4 Discussion

It is a surprising finding that a model as general as the one used cannot be reconciled with experimental findings. It should be noted here that the model not only allows for any possible combination of parameters but also any functional dependencies within the monotonicity constraint, so encompasses a

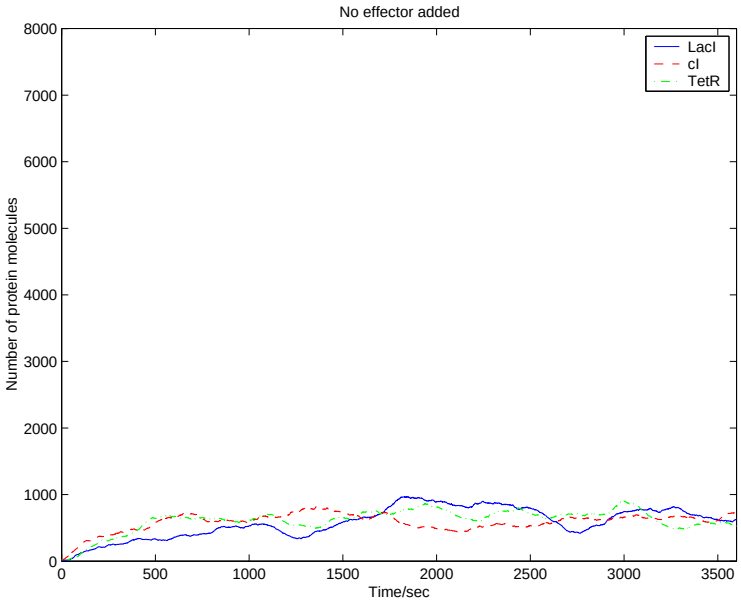

Figure 3.1: The time course of the protein numbers in the network D052 in an example Monte Carlo simulation with a randomly perturbed parameter set. The simulation was run for the system without the addition of effectors.

great variety of different models. The observation that all such models should behave similarly for cases in which a prediction is possible permits model simplification.

Because topologically equivalent networks don't behave similarly, it is reasonable to assume that one of the assumptions we made has to be incorrect. The first assumption we challenge is the assumption of steady-state behavior, since it is known that many biological networks do not necessarily exhibit this kind of behavior. However, it is easy to show (see Methods) that at least in all network models of the type applied here and with autoregulatory loops, steady-state will be reached. This includes the two networks D038 and D052, which were the focus of the study. While we show this independent of parameter values, we have to ask the question whether on the time scale on which the experiments were carried out, steady-state could be achieved. Several experimental studies have measured the time course of the GFP level in systems very similar to the one used in the experiments modeled here [24, 32]. It was found that steady-state is achieved within several hours.

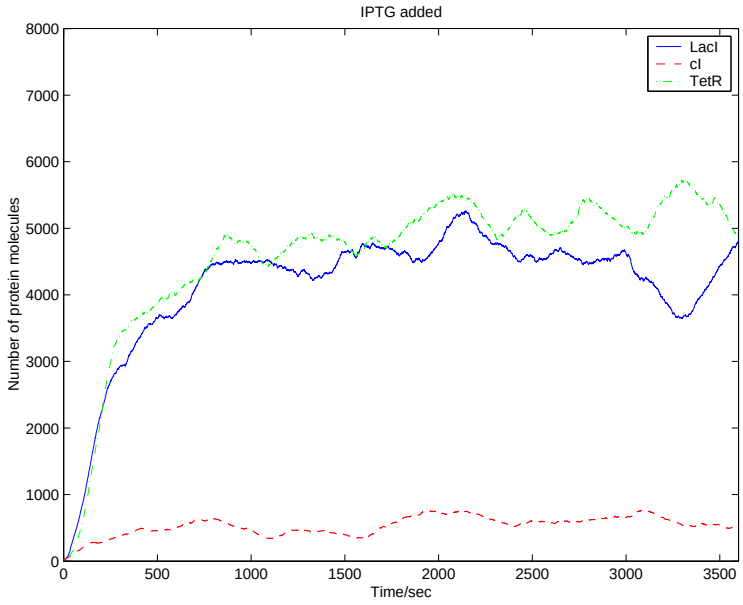

Figure 3.2: The time course of the protein numbers in the network D052 in an example Monte Carlo simulation with a randomly perturbed parameter set. This simulation was run for the system with the addition of IPTG, using the same parameters as Figure 3.1

Guet et al. [45] measured GFP levels after overnight growth, so that there was enough time allotted for steady state to be reached.

Assumptions of transcription level control only and no cross talk are certainly widely used and there is no biological evidence which would put those in question. Although it is easy to imagine cases in which the dependence of transcription or translation rate on protein or RNA concentration, respectively, would not be monotonic, such cases tend to be the result of intentional design.

No.	LacI	cI	TetR	IPTG	aTc	IPTG/aTc
1	LacI	cI	TetR	0/0	0/0	0/0
2	LacI	cI	LacI	0/0	0/0	0/0
3	LacI	cI	cI	0/x	0/x	0/x
4	TetR	cI	TetR	0/x	0/x	0/x
5	cI	cI	TetR	0/x	0/x	0/x
6	TetR	cI	cI	0/x	0/x	0/x
7	TetR	cI	LacI	0/x	0/x	0/x
8	cI	cI	cI	0/x	0/x	0/x
9	cI	cI	LacI	0/x	0/x	0/x
10	TetR	TetR	cI	0/x	-/x	-/x
11	TetR	TetR	LacI	0/0	-/-	-/-
12	TetR	TetR	TetR	0/x	-/x	-/x
13	cI	TetR	cI	0/x	x/x	x/x
14	cI	TetR	LacI	0/-	x/-	x/-
15	cI	TetR	TetR	0/0	x/-	x/-
16	LacI	TetR	TetR	x/x	x/x	x/x
17	LacI	TetR	cI	+/x	-/x	x/x
18	LacI	TetR	LacI	+/-	-/-	-/-
19	LacI	LacI	cI	-/x	0/x	-/x
20	LacI	LacI	TetR	-/-	0/-	-/-
21	LacI	LacI	LacI	-/-	0/-	-/-
22	cI	LacI	cI	x/x	0/x	x/x
23	cI	LacI	TetR	x/-	0/0	x/-
24	cI	LacI	LacI	x/0	0/0	x/0
25	TetR	LacI	LacI	x/0	x/0	x/0
26	TetR	LacI	cI	-/x	+/x	x/x
27	TetR	LacI	TetR	-/0	+/+	-/0

Table 3.1: Predicted behavior of all 27 synthetic repressor networks consisting of the genes LacI, TetR and cI. In the columns titled LacI, cI and TetR, its repressor gene is given in the table. In the columns titled IPTG, aTc and IPTC/aTc, GFP level changes are are given (predicted/observed by Guet et al. [45]) as +,-,0, or x (unknown).

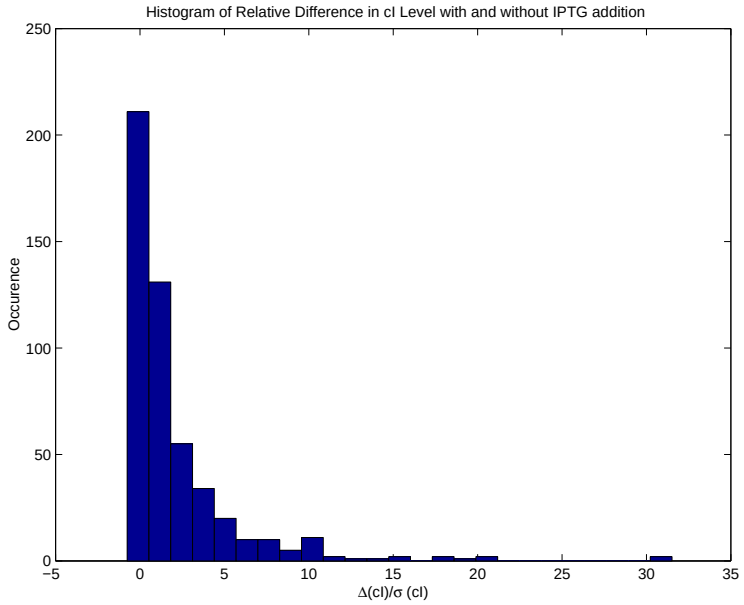

Figure 3.3: A histogram of the difference between the Monte Carlo simulations of D052 without effector and with ITPG added are shown for the stochastic parameter sampling. The difference in protein number of cI relative to the standard deviation of both simulations is shown.

Often neglected is the dilution effect of growing cell cultures. Since cells grow and divide, any given protein concentration, even if neither degradation nor translation take place, will decrease by pure dilution. At stationary phase, the dilution effect can be neglected since cell division and growth is very limited. In general, the dilution effect has a monotonically decreasing dependence on molecule concentration itself, so it can be viewed and treated mathematically as part of the degradation term, and we do not need to treat it separately [104].

Another possibility is that the existence of noise may affect the results of our study. In other words, as has been pointed out recently [5, 25, 122], gene expression is a stochastic process and the ordinary differential equation model is merely an approximation to it. To address the question whether stochastic effects would affect the results, we used an approach based on the Master equation [122]. The model shows that at steady state, stochastic effects will induce a deviation from the steady state, which disappears by averaging large numbers (see appendix). Experimental data is given in terms of population

averages, hence stochastic effects are unlikely to be the cause of large differences in this particular case. Furthermore, Monte Carlo simulations based on the method introduced by Gillespie [37] were carried out for the network D052. Parameter space was sampled stochastically for 500 parameter sets. In each of the 500 simulations, the steady-state level of cI without effector was lower than or approximately equal to the level of cI with ITPG. A significant increase in cI concentration, which is implied by the lower observed GFP level, could not be found in 500 simulations. Hence stochastic effects do not seem to bring our model into agreement with experiments. If Figure 3.3 we show a histogram of the relative difference in protein number of the simulations run with and without IPTG addition with respect to the standard deviation in the simulation.

Having challenged and validated the basic assumptions used in applying the model, we next consider limitations of the model itself. The synthetic network employed in the study by Guet et al. [45] has one property not generally found in genetic circuits - all regulatory proteins carry an ssrA Tag and are, thus, degraded by special cellular machinery, the Clp system [70]. The cellular concentrations of ClpX and ClpP are thought to be fairly low and it is possible that the system can be saturated. Because three proteins carry an ssrA Tag in the network of Guet et al. [45] they may have to compete for binding to limited ClpX. Therefore, the degradation of a given protein may become dependent on the concentrations of the other proteins. In other words, if one particular protein becomes abundant, it would slow down the degradation of the other proteins because it would outcompete them for binding ClpX.

The extended model looks as follows:

$$\dot{r}_i = \quad \text{deg}_{r_i}(r_i) + \text{tr}_{y_i}(p_{y_i}) \tag{3.48}$$

$$\dot{p}_i = \quad \text{deg}_{p_i}(p_i, p_{tot}) + \text{tl}_{p_i}(r_i) \tag{3.49}$$

The main difference to the simple model in Equation 3.3 is that here the protein degradation rate also depends on the total protein concentration, which reflects saturation effects. The degradation rate will be monotonically decreasing with rising total protein concentration. It can then be solved analogously to the simple model to:

$$0 = \text{deg}_i(p_i, p_{tot}) + \text{tl}(\text{deg}^{-1}(-\text{ts}_{y_i}(p_{y_i}))) \tag{3.50}$$

It can be shown using arguments analogous to the ones used above, that the simple addition of saturation effects would allow for the observed behavior (see appendix). In Figures 3.4 and 3.5 we show the graph of MC simulations of the extended model of the network D052 with and without IPTG. It can account for the observed behavior (GFP levels are lower and cI levels are higher after IPTG addition).

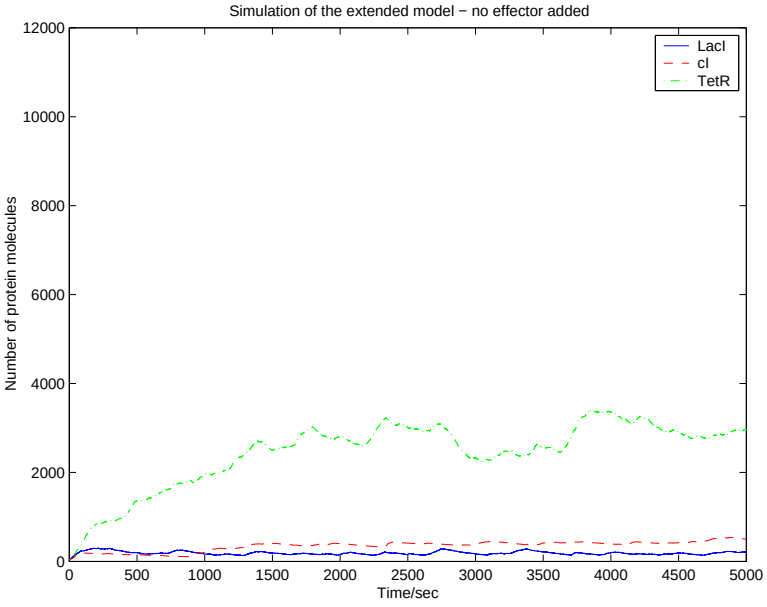

Figure 3.4: A Monte Carlo simulation is shown for the extended model. The development of protein numbers is given for all three proteins without effector added.

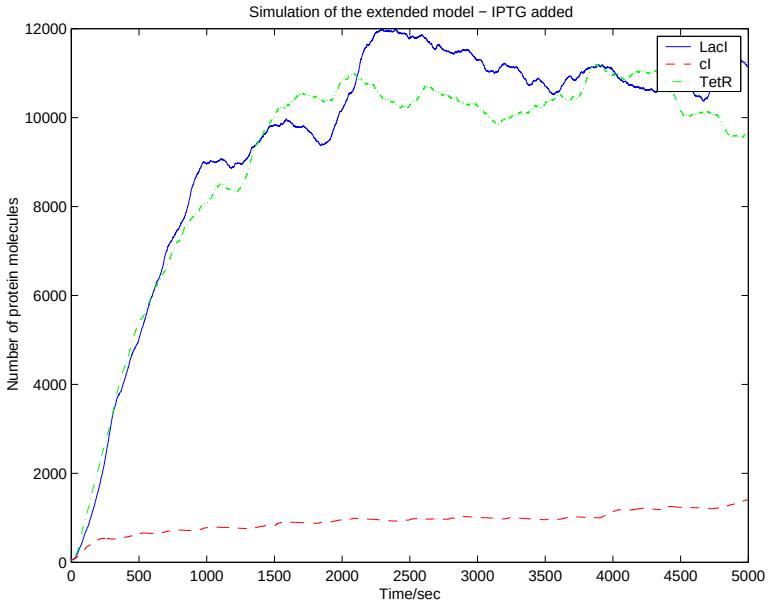

Figure 3.5: Monte Carlo simulation for the extended model with the addition of ITPG. The protein numbers for both TetR and LacI rise rapidly and outcompete cI for degradation by the Clp system. Therefore, the concentration of cI increases beyond the level it had without IPTG addition.

Intuitively, adding IPTG releases the repression of both LacI and TetR. Both proteins are expressed strongly, reaching high cellular concentrations and subsequently outcompete cI for degradation. This effect can then lead to an overall increase in cI concentration if the basal transcription rate of the repressed cI promoter is higher than the resulting degradation rate of cI. There are several options to test this model experimentally. The same experimental setup as used by Guet et al. [45] could be reused by removing the ssrA tag to avoid clp related effects. This approach, however, could lead to problems with achievement of steady state, since the removal of the ssrA tags would significantly increase the lifetime of the proteins. Also, since the measurements do not have to be taken at high time resolution, it is conceivable to quantitate LacI, TetR and cI directly using a conventional biochemical technique.

It is important to note here that our model extension is only one of many possibilities. A different possibility is that spatial effects will affect the behavior of the network [118]. In other words, the low number of ClpX and

ClpP molecules in the cell can make protein degradation into an essentially diffusion controlled process. In this case, the well-stirred assumption breaks down and spatial effects could lead to similar saturation and competition effects as in our proposed model.

The possibility of saturation of the clp system is particularly simple in that it does not invoke new or obscure cellular phenomena, but only allows for effects which have been observed experimentally. If this is limiting to the development of synthetic biological networks, it may be advantageous to overexpress the clp system components.

3.5 Acknowledgments

We thank Igor Levchenko and Chris Hayes for useful information on the Clp system and Bambang S. Adiwijaya, Drew Endy, and Caitlin A. Bever for helpful discussions and suggestions. PMK is a Ph.D. Fellow of the Boehringer Ingelheim Fonds. This work was partially supported by the National Institutes of Health (MH62344).

3.6 Appendix

The Master equation model. In a stochastic formulation of the model, each term in Equations 3.3 and 3.4, which contribute to the rate of change in RNA or protein concentration, becomes a probabilistic variable. In general, the Master equation, which describes the time evolution of probability densities looks as follows for our system:

$$\dot{P_{r_i}} = (E_i^{-1} - 1)(\sum_j \text{tr}_{y_j}(p_{y_j}))P_{r_i} + (E_i^{+1} - 1)(\sum_j \text{deg}_{r_j}(r_j))P_{r_i} \quad (3.51)$$

$$\dot{P_{p_i}} = (E_i^{-1} - 1)(\sum_j \text{tl}_{y_j}(r_j))P_{p_i} + (E_i^{+1} - 1)(\sum_j \text{deg}_{p_j}(p_j))P_{p_i} \quad (3.52)$$

Where E is the step operator, $E_i^k P(q_i, ...) = P(q_i + 1)$. The Master equation can only be solved in its linear form, after linearization and replacing the single p_i and r_i with one q_i we obtain (Matrix A contains the linearized transcription and translation rates and Matrix B the linearized degradation rates):

$$\dot{P_{q_i}} = (E_i^{-1} - 1)(\sum_j A_{ij} q_j)P_{q_i} + (E_i^{+1} - 1)(\sum_j B_{ij} q_j)P_{q_i} \quad (3.53)$$

This treatment follows that of Thattai et al. [122] and Reichl [100]. This can be solved using the moment generating function:

$$F(z_j, t) = \sum_{q_l=1}^{\infty} (\prod_{l=1}^{n} z_l^{q_l})P_{q_l} \quad (3.54)$$

We can assume that the degradation rate of each species (the linearized form of which is contained in Matrix B) only depends on its own concentration, so that Matrix B is diagonal. We then obtain:

$$\dot{F} = \sum_j (1 - z_j)(B_i \frac{\partial F}{\partial z_i} - \sum_j A_{ij} z_j \frac{\partial F}{\partial z_j}) \quad (3.55)$$

Now, since $\frac{\partial F}{\partial z_j} = < q_j >$ and $\frac{\partial^2 F}{\partial z_j^2} = < q_j^2 > - < q_j >$ and $\dot{F} = 0$ we obtain the following linear equations:

$$(A - B)J = 0 \quad (3.56)$$

$$(A - B)K + L = -((A - B)K + L)^T \quad (3.57)$$

where $J_i = \frac{\partial F}{\partial z_i}$ are the means and $K_{ij} = \frac{\partial^2 F}{\partial z_i \partial z_j}$ are the variances and $L_{ij} = A_{ij} J_i$. From Equation 3.57 it can be seen that the means (which is what is measured when taking a measurement from a population average) are given as simple linear equation of transcription, translation and degradation rates, just as in the non-stochastic case.

The extended model. In the extended model, we can write Equation 3.49 as:

$$0 = \deg_i(p_i, p_{tot}) + g_{y_i}(p_{y_i}) \tag{3.58}$$

Here, the function $g_{y_i}(p_{y_i})$ is monotonically decreasing with p_{y_i}. Also, $\deg_i(p_i, p_{tot})$ is monotonically increasing in its first argument but monotonically decreasing in the second one (thereby taking into account the saturation of Clp). Specifically, for the D052 network:

$$0 = \deg_{LacI}(p_{LacI}, p_{tot}) + g_{TetR}(p_{TetR}) \tag{3.59}$$

$$0 = \deg_{cI}(p_{cI}, p_{tot}) + g_{LacI}(p_{LacI}) \tag{3.60}$$

$$0 = \deg_{TetR}(p_{TetR}, p_{tot}) + g_{TetR}(p_{TetR}) \tag{3.61}$$

From monotonicity restrictions we know that the equation $y = \deg_i(p_i, p_t)$ implicitly defines a function ϕ such that:

$$p_i = \phi_i(y, p_t) \tag{3.62}$$

We can also deduce that $\phi_i(y, p_t)$ is monotonically increasing in y and also monotonically increasing in p_t [6]. We can then define $p_{cI} = \phi_{cI}(g_{LacI}(p_{LacI}), p_t)$ and write Equation 3.61 as:

$$0 = \deg_{TetR}(p_{TetR}, p_{LacI} + p_{TetR} + \phi_{cI}(g_{LacI}(p_{LacI}), p_t)) + g_{TetR}(p_{TetR}) \tag{3.63}$$

From here we can see that any change in the g_{TetR} term (due to addition of effector) can be accommodated with changes in both directions from the protein concentrations (As g_{TetR} increases, $\deg_{TetR}$ has to decrease, which can either be achieved by decreasing p_{TetR} or increasing p_{LacI} or increasing ϕ_{cI}). In other words, simple monotonicity restrictions are not sufficient anymore to define the behavior of the system after effector addition.

[6] Let $z = f(x, y)$ be monotonically increasing in x and decreasing in y. Then $z = \phi(z, y)$ monotonically increasing in z. Let $z = z^*$ be fixed. Any increase of x now leads to an increase in $f(x, y)$. To satisfy $z^* = f(x, y)$ y has to increase. Therefore $x = \phi(z, y)$ is increasing in y.

Chapter 4

Analysis of the Separability of Protein Networks According to Four Different Measures[1]

> "Nature is constructed in a way that it can be understood. Or maybe I should state more correctly: Our thinking is constructed in such a way that it can understand nature."
> *Werner Heisenberg, 1901 - 1976*

Abstract

Motivation: Biological systems carry out many control and effector functions in an integrated fashion. Effective design may dictate topological features inherent in network structure that, if understood, could facilitate analysis and modeling. A common preconception is that biological networks should possess a feature of separability, in that groups of nodes with many intra-group connections are themselves joined by relatively few inter-group connections. In this view, groups of tightly interconnected nodes, which may participate in carrying out the same overall function, can be separated from each other by severing a relatively small number of inter-group connections. Isolating such separable pieces may be particularly useful for purposes ranging from assigning gene function to logical analysis of biochemical circuits.

[1] P.M. Kim, T. Ideker and B. Tidor, *Intelligent Systems for Molecular Biology* (submitted)

In this work we examine a number of different protein networks to analyze their topological properties with respect to separability. Current interaction databases are incomplete as well as likely to contain a large number of false positives; simulation studies are carried out to determine whether our results are likely to reflect biases in current interaction databases or properties inherent in biological network structure.

Results: We find that whether protein networks exhibit separability depends strongly on how separability is defined. According to the most intuitive measure, maximum flow, which separates components with more connections within them than to the outside, the examined networks do not exhibit separable structure. However, using the measures of geodesic distance, topological overlap and betweenness, networks appear separable, with significant similarity between the separable units found by different measures. Furthermore, statistical under- and oversampling of the data indicates that the differences in separability are due to biological structure rather than merely reflecting a data bias.

4.1 Introduction

The recent advent of novel high-throughput experimental technologies such as large-scale yeast-two-hybrid methods [63, 123], mass-spectrometry methods [33, 57] or co-immunoprecipitation methods [65, 78] has led to a tremendous increase in the availability of protein–protein and protein–DNA interaction data. This trend has been amplified by extensive curation efforts, tapping into the vast resource of the past literature. Several large protein interaction databases exist, defining a network of over 5,000 proteins and over 15,000 interactions in the case of yeast [6, 86, 133, 134]. This wealth of information carries the promise of providing valuable insight into the functional and physical features of the cellular and biochemical components of the cell. However, especially high-throughput approaches have been shown to be inherently noisy and to yield a large number of false positives [21, 83]. Also, poorly understood biases in the underlying experimental technology further complicate interpretation of interaction data.

The notion of separability has received much attention in the field of systems biology [8, 45, 52, 93, 97, 125]. The hypothesis is that separable network structures might possess a certain degree of functional autonomy and perhaps even form recurring units that can substitute for one another in plug-and-play fashion. With such ideas in mind, they are sometimes referred to as modules. It is believed that understanding the behavior of such autonomous subunits will facilitate the understanding of the whole organism. Also, separability has been shown to occur in different types of non-biological networks, such as social networks or the world-wide-web [30, 38]. In a recent study Ravasz et al. [98] showed indication that metabolic networks adopt a separable hierarchy in terms of structure and function. Furthermore, Milo and others [44, 88, 111] showed the existence of recurring network motifs in protein–DNA interaction networks. This led to the question of whether protein–protein networks are inherently separable, i.e. whether there exist separable subnetworks.

We start by defining a separable unit from a topological perspective; a separable unit is a subgraph consisting of connected nodes, the nodes are required to have a higher number of connections within the subgraph than to the rest of the graph. Intuitively one might suspect that there are indeed such subpieces, as there are functional complexes consisting of highly connected protein members, such as the ribosome, proteasome or DNA polymerase. Moreover, if there do exist separable units, finding and dissecting them can serve as a useful tool for biologists for viewing and exploring the otherwise inaccessibly large and complex networks.

Here, in an attempt to better understand the separable nature of molecular interaction networks, we survey four different graph theoretic measures: maximum flow, geodesic distance, topological overlap and edge betweenness.

Maximum flow is an intuitive measure derived from the study of pipelines. It is defined as the number of distinct paths between two nodes, which do not share any edges - intuitively corresponding to the "maximum flow" of a liquid between them [19]. It can find separable subgraphs as defined by the strict definition above. Geodesic distance is a particularly simple measure, it is defined as the number of edges that separate two nodes by shortest path. Both of these measures are basic independent topological measures - while maximum flow is not affected by distance, distance is not affected by maximum flow. The measure of topological overlap was introduced by Ravasz et al. [98] and shown to find separable structure in metabolic networks. It is defined for each node pair as the number of common neighbors (plus one for adjacent nodes) divided by the minimum number of neighbors of the two nodes. It thereby combines a simple notion of maximum flow with a simple notion of distance. Edge betweenness has been shown by Girvan and Newman [38] to find community structure in social networks; In contrast to the other three measures, it is not defined for pairs of nodes, but rather, it is defined for every edge as the number of shortest paths that run through it.

Our results indicate that, surprisingly, protein interaction networks generated from current networks are not separable according to the maximum flow measure. However, according to the three other measures, subgraphs are found that are shown to be of functional relevance and conserved across the three measures. We compare the relative performance of the three measures according to topological and functional scoring functions. Finally, results from simulation studies of under- and oversampling the current datasets suggest that the separability properties found reflect biological structures and are unlikely to be due to incomplete or noisy data.

4.2 Methods

Network datasets. Several different datasets, all giving interactions in *Saccharomyces Cerevisiae* were examined. The yeast subsection of the database of interacting proteins (DIP) given by Xenarios et al. [133] was analyzed as well as the so-called DIP core by Deana et al. [21]. The DIP core is a subset of DIP consisting of interactions that were cross validated using expression data, orthologous interactions or the literature. Also, the dataset in the biomolecular interaction network database (BIND) [6] was examined.

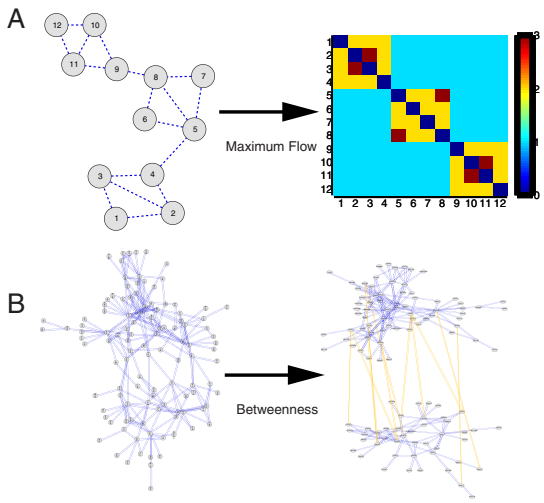

Figure 4.1: Maximum flow and betweenness measures of separability. (A) In a sample network three separable groupings of four nodes each are linked in a chain by single inter-group connections. (B) In a subsection of the DIP core (bottom), which contains more complex structure, whether the network is viewed as separable depends on the measure used. The network in (A) is defined as separable by maximum flow and betweenness measures, whereas the network in (B) is only separable using the betweenness definition.

Four measures of separability. The maximum flow between two nodes in a graph is the number of distinct paths that exist between them, where distinct paths are paths that do not share any edges. A fundamental theorem in graph theory states that the maximum flow between two nodes is equivalent to the minimum cut between them, i.e. the number of edges that have to be cut to separate the two nodes [19]. For a single maximum-flow calculation the algorithm given by Ford and Fulkerson [19] was used. Figure 4.1 (A) illustrates the notion of maximum flow. The all-pairs maximum flow (i.e. the matrix of the maximum flow from every node to every other node) for each graph was calculated using the algorithm given by Gomory and Hu [40] implemented both in the programming package MATLAB version 6 (R12) (Mathworks Inc., Waltham, MA) and in Java version 1.4 (Sun Microsystems Inc., Sunnyvale, CA) using the yFiles graph algorithm and visualization library version 1.3 (yWorks GMBH, Tübingen, Germany). The simpler Gusfield [46] variant of the Gomory-Hu algorithm was also implemented in Java using yFiles. Both variants solve the all-pairs maximum flow problem exactly and efficiently. Briefly, the algorithm proceeds by first computing a so-called Gomory-Hu cut-tree which is a spanning tree of the input graph. It is calculated by $n - 1$ (n being the number of nodes in the graph) rather than n^2 successive maximum flow calculations and successive node contractions and expansions; for details refer to [40]. The tree is weighted and each edge weight corresponds to the capacity of maximum flow between the two nodes separated by this edge. From this tree it is trivial to create an all-pair maximum flow matrix. This tree (also called a cut-tree) can also be used as an approximation to solve the minimum *k*-cut problem [47, 50, 106], however in the datasets we examined, only very unbalanced partitionings could be obtained which is likely a result of the monocentric structure of the data. The implementation in Java ran for roughly 5 minutes on the core DIP dataset on a 1200 MHz Athlon workstation.

Partitioning of the graphs using the measure of betweenness was implemented as described by Girvan and Newman [38] in Java 1.4 using the yFiles library version 1.3. Briefly, for every pair of nodes in a connected graph, there is one (in some cases more than one) shortest path to connect the two. Then each graph contains a total of $S \geq n^2$ shortest paths. A number $k < S$ of those paths will run through any given edge. This number is called the edge betweenness. It is calculated using a modified version of breadth first search. A graph is partitioned by subsequently removing each edge with the highest edge betweenness associated with it. After each edge removal, the betweenness for all edges contained in the affected subgraph is recalculated. With the removal of some edges, pieces of the graph are disconnected, thereby building a hierarchical tree. A pairwise measure for all nodes can be derived from this procedure by assigning each pair of nodes the edge betweenness number

of the last edge that has to be removed to separate it. Partitioning the DIP core network took about 10 minutes on a 1200 MHz Athlon workstation.

Topological overlap was calculated as described by Ravasz et al. [8] in the programming package MATLAB. The resulting matrices were partitioned using an average linkage hierarchical clustering algorithm.

Finally, geodesic distance (shortest distance) between any given two nodes is defined as the length (number of edges) of the shortest path connecting them. It was computed using a simple breadth-first search algorithm as found in [19]. While more efficient algorithms exist, this calculation took less than 1 minute for most datasets examined using an unoptimized implementation in MATLAB.

Generating Partitions from Pairwise Matrices. In the cases of maximum flow, topological overlap and geodesic distance, an average linkage hierarchical clustering algorithm was used on the matrices of the pairwise measures. The resulting dendrogram was cut at a chosen cutoff value to generate the partitioning.

Scoring different partitions. Topological validity of a partitioning was assessed by computing both the number of edges that were cut and the variability in size of each subgraph. The latter was calculated as the ratio of the standard deviation to the mean in the number of nodes of each subgraph. To calculate the score of biological significance we used the functional annotations available from MIPS [86]. Using the hypergeometric probability distribution function, for each subgraph and each category we calculated the probability of occurrence of this particular category when choosing genes randomly from the network. The logarithm with base 10 of the probability was used as the functional significance score.

Subgraph scores were compared to those of random subgraphs generated using to two different approaches. One approach was to simply disregard the topology and pick genes from the pool of genes in the network at random (Called “Random Nodes” in Figure 4.3). The second approach was to subsequently remove random edges until a given number of subgraphs was separated from the graph (Called “Random Cuts” in Figure 4.3). This second approach takes the graph topology into account, in that it will only pool genes that are topologically close together; however, by construction it tends to generate fairly uneven random cuts.

Similarity between different partitionings was computed according to an asymmetric similarity function: Each subgraph in partitioning 1 was assigned a subgraph in partitioning 2, with which it shared the most nodes. The percent similarity score is then the percentage of subgraphs in partitioning 1 that share more than half of its nodes with their most similar pendant in partitioning 2.

Visualization. Network graphs were visualized using the gene network visualization package CYTOSCAPE [62]. All calculations aside from the all-pairs maximum flow and betweenness calculations were carried out in MATLAB.

4.3 Results

BIND contains 3,470 proteins and 5,003 interactions (Status: December 2002), whereas the full DIP contains 15,132 interactions among 4,720 proteins (Status: December 2002). The DIP core contains 3,003 interactions among 1,788 proteins. We examined the largest connected component in each network only, which in every case contains more than 90% of all the interactions in the dataset.[2]

Four measures of separability. Our strictest definition of a separable unit is a connected subgraph that contains more intra-subgraph connections than interconnections to the rest of the graph. This corresponds to the maximum flow definition of separability. Intuitively, calculating the maximum flow matrix and using a clustering algorithm on it will reveal this type of separable unit. We computed the maximum flow matrix for each dataset as described in methods and used an average linkage hierarchical clustering algorithm on it. As is illustrated in Figure 4.1, it can neatly separate subgraphs fit the strict criterion of a separable unit as defined above.

Surprisingly, for any of the networks examined, we observed a maximum flow distribution without apparent separable subgraphs; the graph consists of one highly connected center and a much less connected periphery. Roughly half of the nodes appear to be "leaves", i.e. connected to the rest of the graph through only one edge. Therefore, in the datasets examined, including BIND and DIP no separable units as defined above seem to exist. All graphs have one highly connected center; that is, they are "monocentric". In the DIP core a small number (31) of maximum flow clusters were found. However, they are fairly small and contain only 8.3% of the proteins (See Figure 4.2 and Table 4.1), so that more than 90% of the nodes in the DIP core also exhibit the aforementioned monocentric topology.

While maximum flow is capable of finding separable units that follow our above strict definition, other methodologies can uncover separable or community structure following a less strict notion of separability. A particularly

[2]Other datasets were analyzed as well, including the munich information center for protein sequences database (MIPS) [86], the protein–DNA data given by Lee et al. [78], the networks from high-throughput experiments given by Uetz et al. [123] and Ito et al. [63] as well as the data given by Gavin et al. [33]. Results were similar to those shown here (not shown).

simple method is to use the geodesic distance between two nodes and use a clustering algorithm to partition the graph given a distance matrix. Intuitively, this approach can separate units of nodes with large inter-group but small intra-group distances. However, since distance has no notion of connectivity, the clusters found are not optimized to be tightly connected, nor are they optimized for size. As mentioned above, two measures that combine some of the characteristics of both maximum flow and distance are edge betweenness and topological overlap. Edge betweenness can separate groups of nodes based on the number of shortest paths that do separate them; it therefore has notions of both. Furthermore, a recent study by Girvan and Newman [38] showed that it can uncover separable structure in other types of networks. Topological overlap was introduced by Ravasz et al. [8] to uncover hierarchical separable structure in metabolic networks. It is defined as the number of neighbors common to two nodes divided by the minimum of the connectivity of both nodes. Hence, it also combines a very simplistic notion of distance (it is zero except for nodes with distance 2 or 1) with a notion of maximum flow (but only counting paths with distance 2).

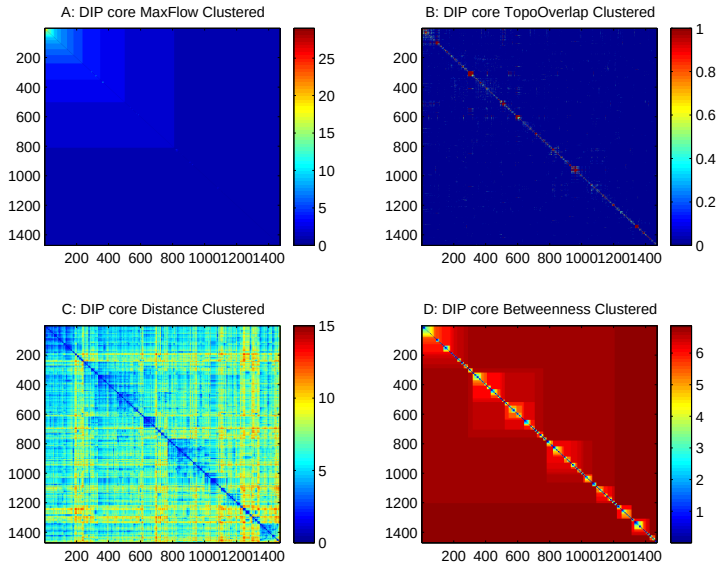

Figure 4.2: Results from the 4 partitioning methods on the DIP core in matrix form. In every matrix, both axis correspond to node numbers, while the order of nodes is determined by the respective partitioning algorithm. A colorscale shows the pairwise measure for every pair of nodes. (A) The maximum flow matrix shows no clear partitionable structure. (B) Topological overlap generates a block diagonal structure. (C) Distance clustering shows a roughly block diagonal structure. (D) Betweenness shows block diagonal structure with roughly equal sized subgraphs.

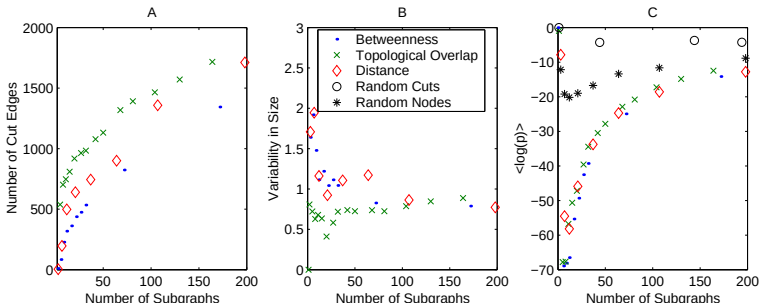

Figure 4.3: Comparison of the different partitioning methods. (A) The number of edges that are cut in every partitioning is shown; betweenness requires the fewest number of cuts per number of subgraph produced. (B) The variability in size (standard deviation divided by the mean of the subgraph sizes) is shown. At low number of subgraphs, topological overlap generates the most even cuts. (C) The functional relevance score (the log of the hypergeometric probability score of enrichment in functional categories) is shown. All methods score similarly well, with betweenness generating subgraphs that may be slightly more functionally significant.

Results of the four different partitioning methods are shown in Figure 4.2. It can be seen that betweenness shows the clearest block diagonal structure. One drawback of all the approaches chosen is that a cutoff value has to be chosen arbitrarily to produce partitions from the pairwise matrix of relations. Whereas in the maximum flow case one can make a clearcut a priori definition of a separable subgraph, in all these cases an ad-hoc choice of a cutoff value has to be made. However, it is possible to optimize the cutoff choice to produce more topologically and biologically sensible partitions.

Quantitative performance of different measures. In Figure 4.3 a comparison of the different partitioning methods when applied with different cutoffs to the DIP core dataset is shown. In Figure 4.3 (A), the parameter used to compare the partitions is the number of edges that have to be cut to separate each piece from the rest of the graph. In other words, how strongly connected was each piece to the graph? Intuitively, minimizing this value brings the subgraphs found closer to the strict definition of a separable unit given above. Naturally, with an increasing number of subgraphs the number of cut edges has to increase. At the same time, the relative number of cut edges starts to decrease, which is due to the fact that at higher numbers of subgraphs, many subgraphs are “leaves”, i.e. consisting of only one node. Since these

	Maximum Flow	Betweenness	Topological Overlap	Distance
# of Clusters	31	72	68	64
% of Nodes in Clusters	8.3	100	100	100
% of Edges Cut	370	824	1318	900
Biological Score	-6	-25	-23	-24
% Similarity w/ MF	-	2	2	0
% Similarity w/ Bet	54	-	82	84
% Similarity w/ TO	49	89	-	76
% Similarity w/ Dist	57	82	73	-

Table 4.1: Comparison of the 4 measures of separability and the similarity of their clusters

subgraphs are biologically less interesting, it would be desirable to obtain approximately even partitionings of the network. One possible measure to estimate how evenly the network was partitioned is to calculate the ratio of the standard deviation to the average of the number of nodes of each subgraph. This measure is plotted in Figure 4.3 (B). All partitioning techniques tend to produce fairly uneven partitionings at low number of subgraphs, while they tend to stabilize at roughly 50 subgraphs. Figure 4.3 (B) shows that partitioning using topological overlap produces the most uniform size groupings, whereas edge betweenness cuts the least number of edges and thus produces separable units closest to our strict definition.

While topological measures can recognize tight clusters, i.e. are capable of measuring how close the subgraphs are to the strict definition of a separable unit, it is more important to validate the partitions with respect to biological functionality. The enrichment of MIPS functional categories is measured for that purpose. In Figure 4.3 (C) the hypergeometric probability score is shown for the three partitioning methods with respect to the number of subgraphs. The hypergeometric probability score corresponds to the logarithm of the probability of the given enrichment in functional categories by chance. It can be seen that all three measures exhibit a sharp minimum in their probability score at a relatively low number of subgraphs, reflecting the fact that functionally tightly correlated subgraphs are discovered and separated early in the process, whereas the later separated subgraphs have less functional relevance and lower the overall score. As in the topological scoring schemes, it can also be seen here that edge betweenness partitioning performs slightly better than the other two methodologies. Moreover, all methods perform significantly better than two methods to randomly pick clusters from the data.[3]

[3]Our hypergeometric scoring function treats different functional categories as independent.

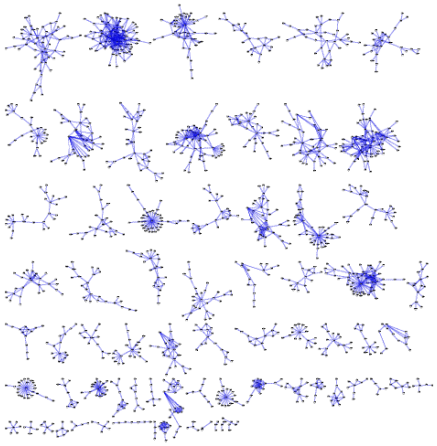

Figure 4.4: A sample partitioning of the DIP core using betweenness partitioning, generating 72 subgraphs.

Figure 4.4 shows the clusters that edge betweenness partitioning finds when partitioning the DIP core dataset to 72 clusters. As Table 4.1 shows, the clusters found by the three different methods are quite similar. For the most part, there is a one-to-one correspondence among subgraphs, which represents the fact that, while most of the clusters do differ between the different methodologies, there is a core that remains essentially conserved across methods. The specific edges that are cut do differ, but most core nodes are the same.

The effects of imperfect data. Thus, separable subgraphs do exist according to three of the measures, but not to the (intuitive) maximum flow measure. Is this due to real biological structure or to error or bias in the interactions network? On the one hand, we do not know how much of the real interaction space has been sampled by the current data. In other words, current databases are incomplete, so addition of more edges to the graph could lead to the formation of intuitively separable units. On the other hand, especially data from high-throughput approaches is known to be inherently noisy and contains a large number of false positives [21]. Hence it is possible that subgraphs which would be separable given only the true connections, are not seen as separable

---

As there is some overlap between categories and thus the assumption of independence is not quite correct, both types of randomization do get a nonzero hypergeometric score

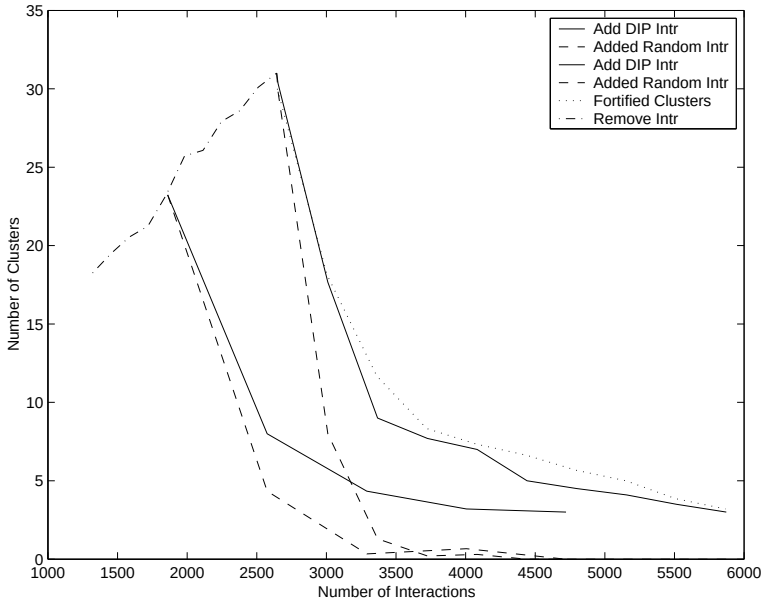

Figure 4.5: Disappearance of maximum flow clusters in the DIP core upon the addition of DIP and random interactions. The blue line represents the disappearance of clusters when adding DIP interactions, the green line represents the slightly slower disappearance when the clusters are fortified to be completely connected, the red line represents the faster appearance when adding random interactions. The effect of removing random interactions is also shown.

because many false positives interconnect them.

In Figure 4.5 the number of maximum flow clusters in the DIP core is shown while adding additional interactions which are missing from the core but are contained in full DIP. It can be seen that the number of discernible clusters drops rapidly with adding new interactions. To address the effects of potential undersampling, we artificially added interactions to each of the found separable units in the DIP core, thereby turning them all into completely connected subgraphs, i.e. every node in the subgraph is connected directly to every other node (fortified). In other words, every cluster was made as strongly interconnected as it possibly could be, a total of 303 interactions were added. Then we again observed the disappearance of clusters while adding interactions until reaching the full DIP dataset as shown in Figure 4.5. While fortifying the clusters to be completely connected subgraphs

does make them more robust to the addition of false positives, the effect is weak and the disappearance curve is hardly distinguishable from the original one. In the full DIP dataset the interactions are so numerous that even those completely connected subgraphs can no longer be recognized as separable units. We hence conclude that the effect of false positives is stronger than the one of undersampling of data. When adding completely random interactions (among nodes contained in the network, however) to the dataset, the drop-off in the number of clusters is more pronounced and it reaches zero quickly. This may reflect the fact that only part of the additional interactions in full DIP are false positives.

A different way to address the issue of undersampling is to remove edges at random and again examine the appearance and disappearance of clusters as shown in Figure 4.5. As can be seen, the effects are quite minor. Even the removal of half the interactions reduces the number of clusters by only a third. When adding interactions from full DIP to a network with 30% of interactions removed, the effect is quite similar to when adding interactions to the original data.

4.4 Discussion

Separability of protein networks. In this study, we examined different protein networks with respect to their separability. Attempting to keep our investigation unbiased, we surveyed four different graph theoretic measures of separability and also examined a range of different datasets.

It is a surprising result that none of the examined large datasets showed signs of separability in the strict (maximum flow) sense. Rather, all exhibit a topology which could be expected from a randomized scale-free network, as a random scale-free network would exhibit a monocentric topology as seen in the networks examined above (data not shown). However, scale-freeness by definition merely describes the node connectivity distribution as being governed by a power law rather than an exponential [7]. It is hence not at all in contradiction with strict separable structure; many ways in which strict separability is compatible with a scale-free topology can be imagined. For instance, connecting many disjoint scale-free networks by one link each would create a scale free separable network: the separable units correspond to the formerly disjoint networks and the node connectivity distribution is still governed by a power-law. However, the evolutionary genesis of existing biological networks and their topological features are still subject for speculation.

Still, since tightly connected protein complexes are known and separability has been the recent focus of research, it is surprising that large-scale dataset do not seem to consist of strictly separable units, as could be found

by maximum flow. But as Table 4.1 shows, they do exhibit a separability according to the other three measures; moreover, the subunits found are quite similar to each other, suggesting that the separability found is not merely due to a peculiarity in the partitioning algorithm used. Furthermore, the subgraphs found all score quite highly when measuring their enrichment in functional categories. Hence, protein networks do not seem to follow the notion of intuitive separability as shown in Figure 4.1 (A), rather, the functional subunits are highly intertwined and interconnected, making them hard to recognize by visual inspection, such as shown in Figure 4.1 (B).

Effects of oversampling and undersampling the real network. Further puzzling is the finding that smaller, sparser datasets do contain strict separable units, even though they are small and not plentiful. Three possibilities are to be distinguished: Firstly, the existence of strict separable units could be an artifact due to undersampling of the real network in small datasets. This possibility is unlikely, since removing random interactions from the data, i.e. making the data more undersampled, decreases the cluster count. Secondly, the disappearance of separability in larger datasets could also be an artifact of undersampling of the real network, in this case the undersampling of the real interactions within the separable units. This possibility is also unlikely, since artificial fortification of the clusters as shown in Figure 4.5 has only a very slight effect on their disapperance when adding DIP interactions. Thirdly the disappearance of separability in larger datasets could be an artifact of oversampling (i.e. false positives) of the real network in large datasets. Our analysis suggests this as the most likely possibility. The effect of additional DIP interactions on the number of clusters is similar to that of random interactions.

A particularly surprising fact is that both the removal and the addition of interactions at random seem to decrease the cluster count. While it can be understood in hindsight - the addition of edges can make the cluster disappear because of new connections to the main graph and removal of edges within the cluster can make it disappear - it does indicate that the existing number of clusters in the DIP core are real topological features of the network and not a product of under- or oversampled data.

Conclusions. In this study we examine notions about separability of protein networks. Our results suggest that protein networks are separable in a largely non-intuitive manner; the separable pieces do not simply correspond to highly interconnected subgraphs which are easily split by cutting a low number of edges as could be expected from an intuitive maximum flow notion. Rather, the separable units are highly intertwined and not intuitively recognizable. Also the partitions according to the three measures distance, betweenness and topological overlap are found to be quite similar and have

high functional relevance. Betweenness appears to be best suited for partitioning protein networks in a meaningful fashion. However, our results also suggest that the complete disappearance of maximum flow clusters is likely the result of the false positives in current datasets. In other words, cleaner datasets are likely to be more easily partitioned.

4.5 Acknowledgments

We thank Bambang S. Adiwijaya, Michael D. Altman, Caitlin A. Bever, Owen W. Ozier and Nikolay St. Stoyanov for insightful discussions and technical assistance. This work was partially supported by the National Institutes of Health (MH62344). PMK was supported by a Ph.D. Fellowship from the Boehringer Ingelheim Fonds. TI was supported by a research fellowship from Pfizer.

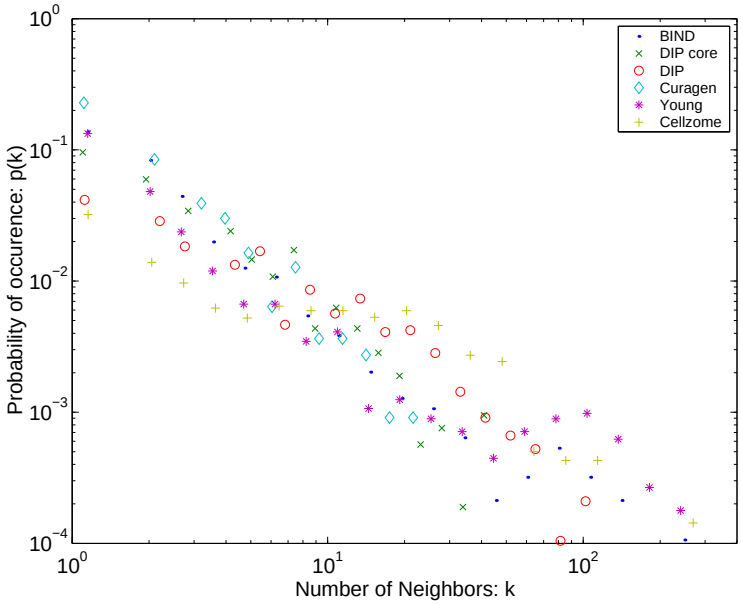

Figure 4.6: The Node Connectivity Distribution for all examined datasets follows a power-law

4.6 Appendix

Scale-free topology of protein networks. We could verify earlier claims [2, 7, 68]; all the networks examined exhibit a scale-free topology. In other words, the node connectivity distribution follows a power law — interestingly all datasets do follow roughly the same power law, which is suprising at least for the protein–DNA dataset which consists of a different kind of interactions. In essence the power law distribution of the node connectivity reflects the fact that the topology of the network is dominated by a small number of highly connected nodes, so-called hubs.

Maximum Flow, Connectivity and Distance. To further investigate the topology of the graph we examined correlations between the node connectivity, distance between nodes and the maximum flow. Interestingly, as shown in Figure 4.7, there is a high degree of correlation between maximum of the maximum flow of a given node (i.e. the maximum number of path by which it is connected to any node in the graph), and its degree of connectivity. Maybe

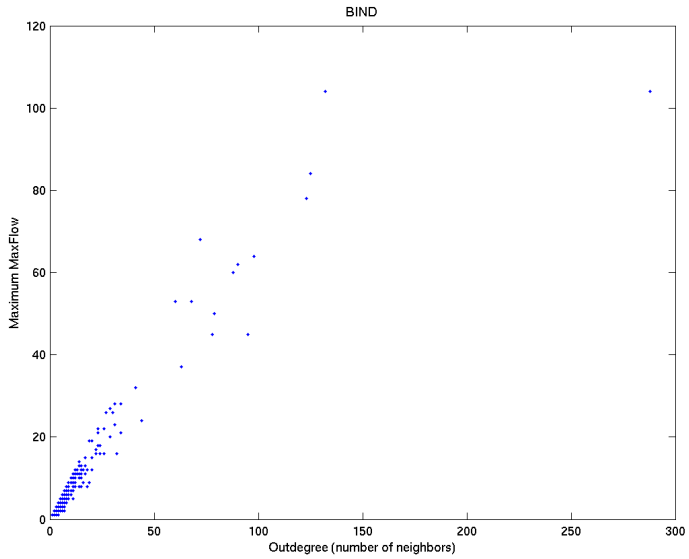

Figure 4.7: The relation between node connectivity and its maximum maximum flow

not surprisingly, the nodes with the highest connectivity degree are also the ones with the highest maximum of the maximum flow. However, as has been observed before [9], nodes with high connectivity do not necessarily correspond to obvious biological importance. For instance the 10 genes which have both the highest maximum flow as well as the highest number of neighbors, do not seem to exhibit particular biological significance.

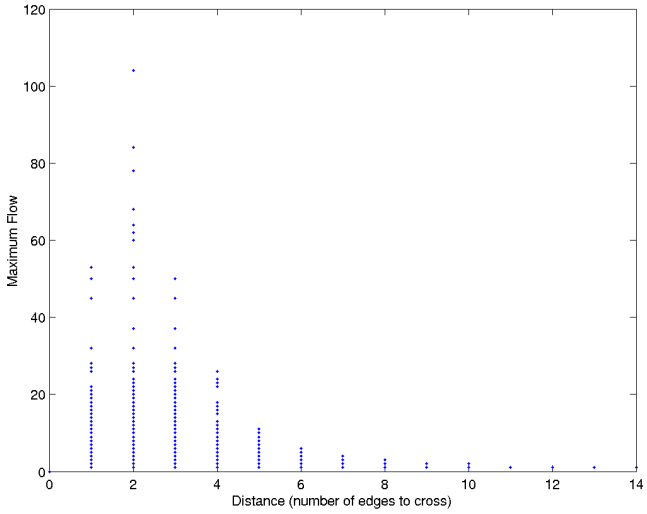

Figure 4.8: The relation between maximum flow and distance

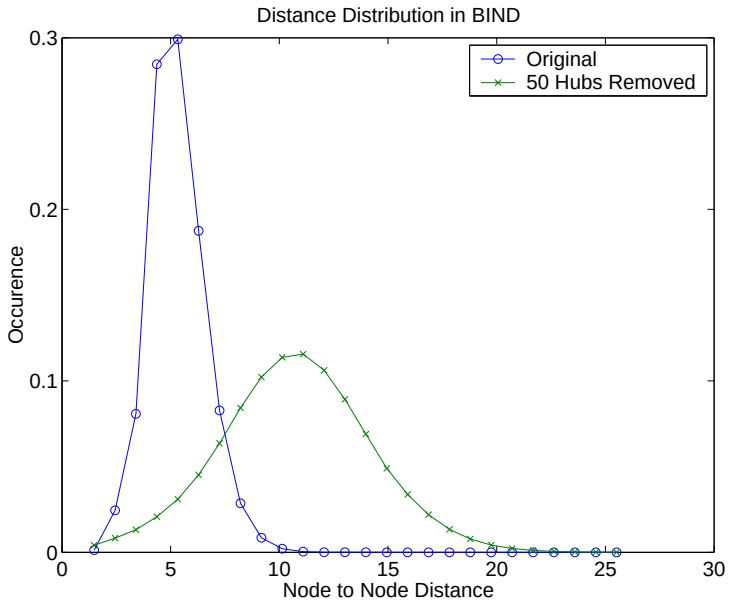

Figure 4.9: The change in diameter of the network due to the removal of hubs

Also, there is a relationship between the distance and the maximum flow of two nodes as shown in Figure 4.8. While both measures are inherently independent, we do observe that at large distances between two nodes the maximum flow is small — intuitively that corresponds to that fact that in that these nodes tend to be at the weakly connected periphery of the graph, in accordance with our findings of a monocentric topology.

Topology change upon hub–removal. Since hubs are thought to dominate the topology of scale-free networks [7], subsequent removal of hubs could be expected to affect the monocentric topology of the networks. Specifically, since the center of the graph is dominated by those highly connected hubs with more than 200 neighbors and maximum flows of over 100, it could be imagined that removal of those hubs of the graph could reveal other, seperable centers in the graph. However, the maximum flow matrix of the BIND dataset after the removal of the 50 most highly connected hubs still exhibits the same structure as before the hub-removal, suggesting a monocentric topology. While the most highly connected node has 288 neighbors (JSN1) the least highly connected of those 50 has merely 15 neighbors. Removal of

hubs has drastic effects on the overall topology of the network, affecting the node-connectivity distribution as seen in Figure 4.9 and the diameter which increases from 6 to 11. This is somewhat counterintuitive, as the number of nodes actually decreases.[4]

[4]In addition to the removed hubs, the giant component of the network decreases in size, the giant component of BIND after the removal of 50 hubs is merely 1700 nodes in contrast to 3050 nodes with all hubs

Chapter 5

General Conclusions

> "By three methods we may learn wisdom: First, by deduction, which is the noblest; second, by imitation, which is the easiest; third, by experience, which is the bitterest." [1]
> *Confucius, 551 BC - 479 BC*

In this thesis, the much-publicized notion of a biological subsystem was examined from different perspectives. Guided by the vision of bringing quantitative understanding to complex cellular phenomena, rigorous approaches from mathematics, physics and computer science were used to solve a number of related problems focusing on the discovery, behavior and topology of biological subsystems. Throughout this work, preexisting notions and models were challenged and their validity reevaluated with respect to existing experimental data.

A novel approach to identify subsystems from gene expression data was developed. These subsystems were shown to be of biological significance, they were discovered by way of reducing the dimensionality of the dataset and thereby focusing on local similarities in the data. While it was the belief in the field that approaches focusing on global similarities in gene expression profiles, such as clustering, would reveal the most reliable predictions from the data, it was shown rather that *local* similarities can reveal relationships that are better predictors of biological function.

Gene expression modeling studies generally use a simple, but widely accepted classical kinetic gene expression model. Here, it was proven rigorously that, this model can not be reconciled with a current set of data. An only slightly more complex model, which can explain the data, was developed. Thereby, a framework to predict the behavior of simple subsystems

[1] As in most scientific endeavors, all three methods were used in this work.

under perturbation was developed and an experimentally testable hypothesis was generated.

The notion of a subsystem was also examined from a topological perspective. Results show that, contrary to what might intuitively be expected, protein interaction networks do not consist of easily separable subsystems, tightly intra-connected subgraphs with only weak inter-connectivity. Rather, it was shown that, while those subsystems do exist, they are highly intertwined and only separable in a non-intuitive fashion. Methods to uncover these subgraphs were developed and it was shown that they are of biological significance.

The results presented here emphasize the importance and value of rigorous theoretical and computational analysis in systems biology. While the need for experimental data and advances in experimental technology is obvious, it was demonstrated that computational approaches are needed to mine the data, build models complementing the experiments and abstract higher principles that can shape our thinking. The establishment of a general framework to discover biological subsystems and predict their behavior will be future work.

Appendix A

Thermal Stability of Proteins and the Role of Electrostatics

A.1 Introduction

Protein stability is defined as the free energy difference between the denatured and the native states. Even though the importance of several physical factors - such as hydrophobic effect, hydrogen bonds and electrostatics - has been examined, our understanding on an atomic level is still limited. The recent discovery of hyperthermophilic organisms, which can thrive at temperatures of up to 113^{o}C, has brought attention to the issue of thermal stability [99]. The factors contributing to protein stability at those high temperatures are subject of recent research efforts [58, 66, 82].

Several studies have emphasized the importance of electrostatic interactions for thermal stability [49, 127, 135] and salt bridges are found more frequently in hyperthermophilic proteins than in their mesophilic counterparts. At the same time, recent theoretical and experimental findings indicate that electrostatic interactions are not major stabilizing factors in mesophilic proteins and are less stabilizing than hydrophobic interactions [53, 54, 129]. These facts seem to be contradicting each other. However, the strength of the hydrophobic effect is known to decrease with temperature, which in turn leads to a weaker contribution to stability by hydrophobic interactions. It could hence be hypothesized that electrostatic interactions are stronger than hydrophobic ones at high temperatures. Moreover, the stability contribution for each salt bridge can be increased by formation of extended ion-pair net-

works. Those networks have been observed in many hyperthermophilic proteins [66].

Here we investigate the role of salt bridges for protein stability at higher temperatures in bacteriophage P22 Arc repressor, a well-established model system for protein folding [12, 87]. Arc is a member of the ribbon-helix-helix family of transcription factors. The native dimer forms a single globular domain by intertwining the β-sheets and α-helices of both monomers. As a result, folding and subunit association are tightly coupled reactions, and the overall folding reaction is a cooperative bimolecular reaction [102]. Therefore, the apparent thermal stability and melting temperature of Arc depend on the protein concentration. This effect enables us to measure stability at varying temperatures using the melting temperature as an indicator (see Section A.3).

The wild-type protein features a buried salt bridge triad comprised of the residues Arg31, Glu36 and Arg40 (wild-type protein will be referred to as RER). In comparison we examine a mutant in which the salt-bridge triad is replaced by hydrophobic residues (Ala31, Tyr36, Leu40 - will be referred to as AYL). RER and AYL have very similar circular dichroism spectra and their stability at 25°C, as measured by urea denaturation experiments, is comparable (RER: 10.0 ± 0.5 kcal and AYL: 10.5 ± 0.5 kcal). A comparison of RER and AYL allows us to investigate the relative importance of hydrophobic and electrostatic interactions on protein stability at varying temperatures.

A.2 Results and Discussion

Figure A.1 shows thermal denaturations of both wild-type Arc (RER) and the R31A, E36Y, R40L mutant (AYL), monitored by circular dichroism spectroscopy. In both proteins thermal stability increases with increasing concentration. However, for concentrations beyond 100 μM, the thermal stability of AYL does not seem to increase significantly; the thermal denaturation curves are quite similar. By contrast, the thermal stability of RER does continue to increase beyond that point.

From the thermal denaturation curves thermodynamic parameters were estimated using a fitting procedure, and in Figure A.2 protein concentration is plotted against melting temperature. As explained in Section A.3, protein concentration is related to the free energy of unfolding in a bimolecular folding system.

Arc does not follow the simple (and widely used) two-state model of protein folding, because there is a dimeric folding intermediate state (termed the denatured dimer). The folding equilibrium follows the reaction:

$$2[D] \rightleftharpoons [D_2] \rightleftharpoons [N_2]$$

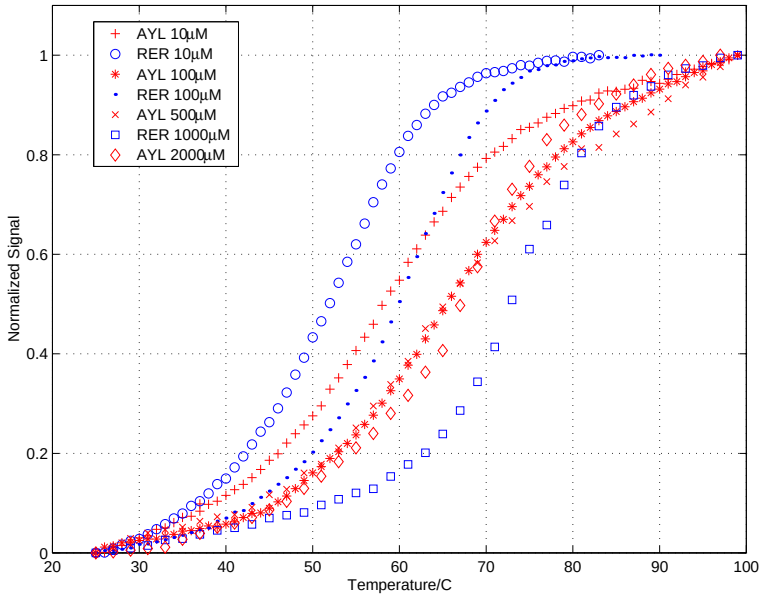

Figure A.1: Thermal denaturations of RER and AYL

(D referring to denatured states and N to the native state). At lower temperatures the folding intermediate state D_2 is minimally populated so that the equilibrium behaves very similar to the two-state model $2[D] \rightleftharpoons [N_2]$. However, at high temperatures the denatured dimer state is more highly populated. Hence, the folding equilibrium is more similar to $[D_2] \rightleftharpoons [N_2]$, which is a monomolecular reaction. Therefore, the dependence of melting temperature on concentration disappears. This effect limits the thermal stability accessible by increasing the protein concentration [102]. The limit in melting temperature was determined to be about 70°C. In the case of AYL this limit is lower (Figure A.2), the T_m is limited to about 66° C. This decrease is due to an increased stability of the denatured dimer, i.e. this state is populated at lower temperatures than in RER.

As can be seen in Figure A.2, at lower temperatures AYL exhibits higher stability than RER. This order switches at temperatures higher than 64°C. To further emphasize this point we show CD wavelength scans of the two proteins at 25°C at a concentration of $10\mu M$ in Figure A.3. The absolute CD signal can be used as an indicator of folding status (see Section A.3). In this graph, both signals are approximately equal which indicates the fact that at

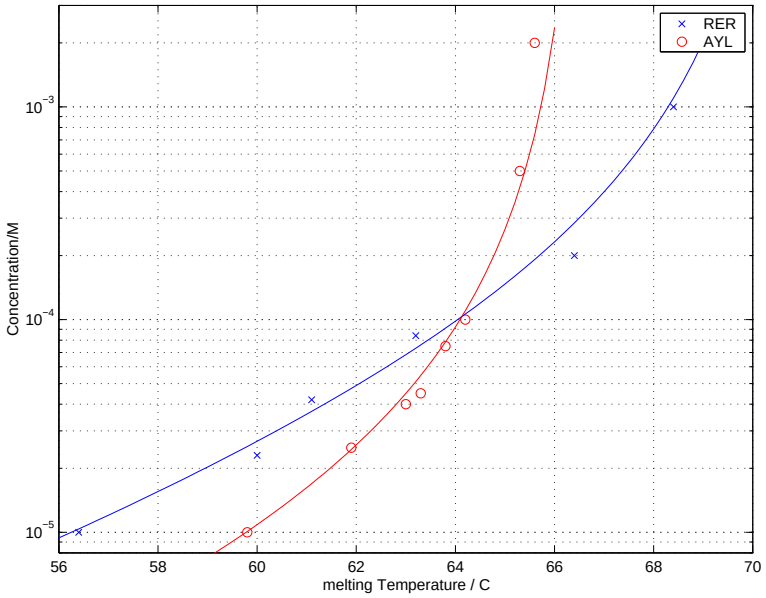

Figure A.2: Dependence of melting temperature on concentration.

room temperature both proteins are fully folded - the slightly higher stability of AYL does not have a measurable effect under this conditions. In Figure A.4 a CD scan of AYL and RER at 65^{o}C and 1.2mM concentration is shown. As can be seen, the RER signal is significantly stronger, indicating that a larger fraction of this protein is folded under this conditions. Therefore, under these conditions, RER exhibits higher thermal stability than AYL.

We have shown a case study in which the relative apparent stability of two proteins switches at higher temperatures. Since the main difference between RER and AYL is the buried salt-bridge triad which is replaced with hydrophobic interactions in AYL, one could infer that this switch is due to a switch of the relative strength of electrostatic versus hydrophobic interactions at higher temperature. However, at such high concentrations the three state nature of Arc's folding reaction has to be taken into account; the shift in apparent stability is not necessarily due to a switch in relative stability contributions of electrostatic versus hydrophobic interactions at higher temperatures. The folding intermediate can be neglected at low concentrations and melting temperatures, but its importance increases with concentration. It is more stable in AYL ($\Delta G = 6.5$ kcal/mol) than in RER ($\Delta G = 5.2$ kcal/mol),

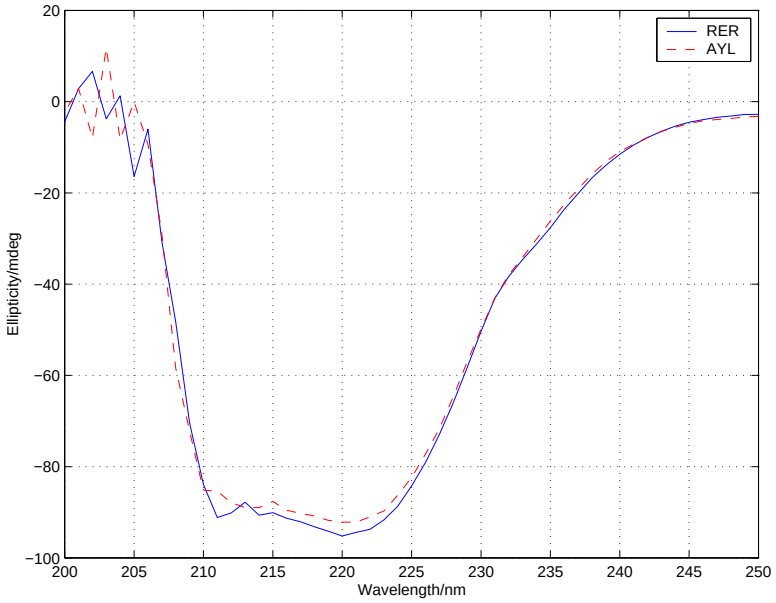

Figure A.3: CD spectra at 25°C and 10μM

therefore the two state model of folding breaks down at lower concentrations of AYL than of RER. In other words, when the denatured dimer state is highly populated (at high concentrations), the bimolecular folding reaction behaves like a monomolecular reaction and hence protein stability can no longer be increased through increasing concentration. Because of the relative stability of their respective denatured dimers, this effect occurs at lower concentrations for AYL than for RER.

This effect could be one of the reasons for the prevalence of salt bridges in hyperthermophilic proteins. Hyperthermophilic proteins do feature more salt bridges than their mesophilic counterparts [66]. Since electrostatic interactions have been implicated in providing specificity to a protein fold, they can be introduced to destabilize a possible folding intermediate. In oligomeric proteins it would thus be possible to increase the thermal stability limit accessible by increasing apparent protein concentration. At higher concentrations, proteins can attain higher thermal stability by making use of this phenomenon.

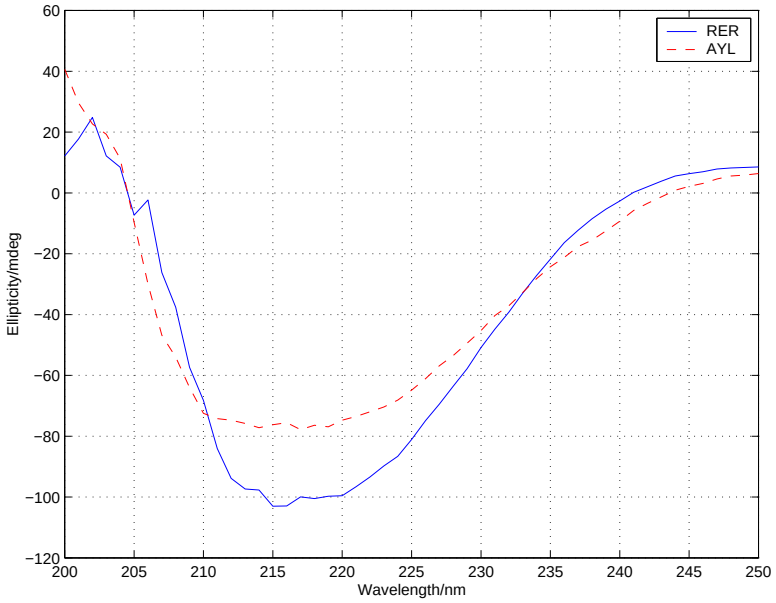

Figure A.4: CD spectra at 65°C and 1.2mM

A.3 Methods

Circular dichroism spectroscopy and denaturation experiments

Circular dichroism (CD) spectra were obtained using an AVIV CD60S spectrometer. Depending on concentration, different quartz cuvettes with pathlength (0.1mm to 1cm) were used. In thermal denaturation experiments, samples were equilibrated at the desired temperature for 2 minutes (or longer if necessary) and CD signal at 222nm was recorded as a function of temperature. Thermodynamic parameters were estimated by fitting denaturation data using the minimization routine in the program package *Matlab* to the two-state model which is explained below. The ratio of fraction unfolded protein is given by the following:

$$F_u = \frac{y - b_l}{b_u - b_l} \tag{A.1}$$

Where $b_{u,l}$ are the upper/lower baselines of the melting transition and y is the absolute CD signal. The free energy is given by:

$$\Delta G = \Delta H_0 + \frac{T}{T_m}(-\Delta H_0 - RT_m \ln(P_{tot})) + \Delta C_p(T - T_m - T\ln(\frac{T}{T_m})) \tag{A.2}$$

Here ΔG is the free energy of unfolding, $\Delta H_0, T_m$ are the reference enthalpy and Temperature (melting Temperature), P_{tot} is the total protein concentration. This can easily be derived from Eq. (A.3). The equilibrium constant is given by:

$$K = e^{\frac{-\Delta G}{TR}}$$

And finally the fraction unfolded can be calculated by:

$$F_u = \frac{1}{4P_{tot}}(-K + \sqrt{K^2 + 8KP_{tot}})$$

This equation is substituted in Eq. (A.1) and the resulting function is fit to the data. Values for the thermodynamic parameters are obtained from this fit.

Protein thermodynamics

In a monomeric folding system, the protein folding reaction is described by the simple equilibrium (D is the denatured and N the native state of the protein):

$$[D] \rightleftharpoons [N]$$

The melting temperature is defined that at T_m $[N] = [D]$. Therefore, at T_m $K = 1$ and $\Delta G = 0$.

Now consider a folding reaction $2[D] \rightleftharpoons [N_2]$. At the melting temperature we have by definition: $[D] = P_{tot}/2$ and $[N_2] = P_{tot}/4$, then

$$\Delta G = -RT \ln K = -RT \ln \frac{[D]^2}{[N_2]} = -RT \ln P_{tot} \tag{A.3}$$

In other words, the free energy unfolding is no longer zero at the melting temperature, but directly depends on concentration. The thermal stability can be increased by increasing concentration. The dependence of protein concentration and melting temperature can be seen as follows:

$$RT_m \ln P_{tot} = T_m \Delta S_0 - \Delta H_0 - \Delta C_p(T_m - T_0 - T_m \ln \frac{T_m}{T_0}) \tag{A.4}$$

Where $\Delta H_0, \Delta S_0, T_0$ are reference Enthalpy, Entropy and melting temperature.

For the three state model $2[D] \rightleftharpoons [D_2] \rightleftharpoons [N_2]$ we have $P_{tot} = D + 2D_2 + 2N_2$ and at T_m we have $N_2 = P_{tot}/4$. Therefore, we get:

$$P_{tot} = D + \frac{2D^2}{K_1} + \frac{2D^2}{K_1 K_2} = \frac{K_1 K_2}{(1 - K_1)^2} \tag{A.5}$$

and the free energies are of course given by:

$$\Delta G_{i,m} = \Delta H_{i,0} - \frac{T_m}{T_0}(\Delta H_{i,0} - \Delta G_{i,0}) + \Delta C_{p,i}(T_m - T_0 - T\ln(\frac{T_m}{T_0})) \qquad i = 1, 2 \tag{A.6}$$

Three state T_m vs. protein concentration data was fitted to these equations.

Appendix B

Survival Probability for Single Tumor Cells, a Master Equation Model

Abstract

Tumors can enter a dormant state, termed tumor dormancy, in which its population size remains constant, but it retains its full malignant potential. This state is achieved by a balance of apoptosis and cell division. In this study it is examined whether this balance is due to a causal linkage of those two processes. Based on experimental studies measuring the survival time of a single tumor cell, a model based on the Master equation was constructed and fit to experimental data. The results indicate that there is no direct causal linkage between apoptosis and cell division in the system examined.

B.1 Introduction

In recent years cancer dormancy has emerged as a common phenomenon. Tumor cells are defined as being dormant if they have full malignant potency but are under growth control. A tumor can remain dormant for years and suddenly revert to malignancy [124]. For instance, patients with a specific form of non-Hodgkin's lymphoma may have remissions many years, but the disease is eventually lethal. Since cellular factors such as the antigen receptor is unique, it can be shown that the secondary tumor, leading to lethality, is recurring and not a new one. Among other factors, humoral and cellular immunity are known to be involved in the induction of dormancy in tumors.

In this study, mouse B-cell lymphoma cells were examined. Tumor dormancy can be induced by treating them with an antibody via the B-cell antigen receptor.

B.2 Experimental results

Studies carried out by Scheuermann and co-workers show that in culture, the population size remains constant for weeks after the induction of the dormant state[1]. Cell division can be observed in these cultures, hence the absence of population growth can not be due to cell cycle arrest, but has to be due to a careful balance between apoptosis and cell division. Furthermore, the rate of apoptosis and cell division have to be equal to maintain constant population.

Such balance between apoptosis and cell division rates could be the result of a direct linkage of both processes, i.e. one process is required for the other one to occur. An asymmetric cell division could be hypothesized: One of the daughter cells undergoes apoptosis, whereas the other one survives. However, it is also conceivable that there is no direct linkage, but rather, both processes have matched rate constants. To address this question Scheuermann and co-workers did several experiments. Dormant cultures were observed at different temperatures, reasoning that cell cycle time has to change, but apoptosis rate would not necessarily change at the same rate. In all those experiments population stability was retained. Moreover, experiments with single cells were carried out. A single cell was plated in a culture well and dormancy was induced. In the case of linked apoptosis and cell division, survival of the single cell would be expected, whereas in the latter case of independent processes, a short survival probability could be expected. A large portion of those cells (around 80%) dies within several days. These experiments are modeled in this study.

B.3 Modeling the survival probability

Both kinds of models mentioned above were considered. In the first one, cell division is required for apoptosis. This implies that at any given time, one cell is observed (except for the short period of time when one of cells is executing its apoptotic program). The second model requires balance of the apoptosis rate and division rate but otherwise the processes are unlinked. There might

[1]These studies together with the work discussed here are in submission for publication: A.K. Hammill, R.C. Hsueh, P.M. Kim, J.W. Uhr and R.H. Scheuermann, "Asymmetric cell division results in differential apoptotic cell fates following stimulation through the B cell antigen receptor in a lymphoma model of tumor dormancy"

t/days	0 cells	1 cell	2 cells
0	0	1	0
1	0.22	0.5	0.27
2	0.44	0.35	0.2
4	0.5	0.37	0.09
6	0.76	0.19	0.04
0	0	1	0
1	0.22	0.55	0.21
3	0.41	0.38	0.19
5	0.62	0.3	0.06
7	0.8	0.16	0.03

Table B.1: Data from single cell experiments. The percentage of wells with no, 1 and 2 cells is given respectively, two different experiments.

be a (so far unknown) mechanism which regulates the two rates to be equal but there is no order in the occurrence of the two processes.

To match the data, the probabilities of having n cells in a well was calculated (P(n,t)). In the first model P(1,t) is trivially equal to one. From the data it is clear that this is not the case. Therefore we focused our attention on the second case. We chose to use the simplest model, viewing both processes as strictly markovian random processes. Especially for cell division this might be not be an optimal approximation. In this case, however, the phase of the cell cycle in which the observed cell is unknown and thus for the occurrence of first cell division is indeed random. Since experimental data only observes several cell cycles, a simple stochastic model is reasonable.

The Master equation

The Master equation governs the stochastic dynamics of Markov processes (i.e. processes which have no "memory", which only depend on the current state of the system). In our case, it takes on the fairly intuitive form:

$$\frac{\partial P(n,t)}{\partial t} = d_{n-1}(t)P(n-1,t)+a_{n+1}(t)P(n+1,t)-(d_n(t)+a_n(t))P(n,t) \tag{B.1}$$

Here, $P(n,t)$ is the probability that the system is in state n at time t, i.e. that at time t there are n cells. d_n is the division rate at state n and a_n is the apoptosis rate.

We assume the the rates to be equal and proportional to the number of cells (so $d_n = a_n = rn$) and get:

$$\frac{\partial P(n,t)}{\partial t} = r(n-1)P(n-1,t) + r(n+1)P(n+1,t) - 2nrP(n,t) \quad \text{(B.2)}$$

Solution of the Master equation for our specific system

Using the generating function and the methods of characteristics, it is possible to solve this equation exactly and we get for the probabilities that n cells exist in the well:

$$P(0,t) = \frac{rt}{1+rt}$$

$$P(n,t) = \frac{r^{n+1}t^{n+1}}{(1+rt)^{n+1}} + \frac{r^{n-1}t^{n-1}(1-rt)^n}{(1+rt)^n}$$

$P(0,t)$ (Probability of no cells) $P(1,t)$ and $P(2,t)$ are plotted.

Step by step derivation

Introduce the generating function. Useful for solving the Master equation is the generating function:

$$G(z,t) = \sum_{n=-\infty}^{\infty} z^n P(n,t) \quad \text{(B.3)}$$

The generating function is the summation of all the moments of the probability distribution and is often useful to calculate it. We have that

$$\frac{\partial G(z,t)}{\partial t} = \sum_{n=-\infty}^{\infty} z^n \frac{\partial P(n,t)}{\partial t} \quad \text{(B.4)}$$

and

$$\frac{\partial G(z,t)}{\partial z} = \sum_{n=-\infty}^{\infty} n z^{n-1} P(n,t) \quad \text{(B.5)}$$

It is easy to see that the following is equivalent to Equation B.3

$$\frac{\partial G}{\partial t} = r(z-1)^2 \frac{\partial G}{\partial z} \quad \text{(B.6)}$$

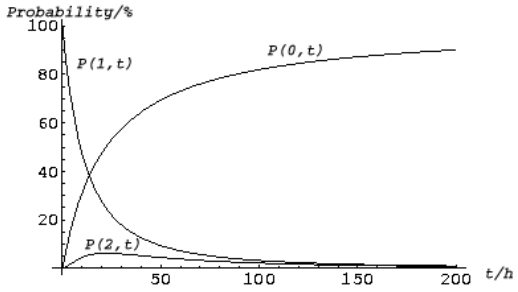

Figure B.1: Plot of P(0,t), P(1,t) and P(2,t) for the case of division and apoptosis rate of 1/22h

It becomes clear if we substitute P into this equation:

$$\sum_{n=-\infty}^{\infty} z^n \frac{\partial P(n,t)}{\partial t} = r(z^2 - 2z + 1) \sum_{n=-\infty}^{\infty} n z^{n-1} P(n,t) =$$

$$r \sum_{n=-\infty}^{\infty} (n z^{n+1} P(n,t) - 2nz P(n,t) + r n z^{n-1} P(n,t)$$

Now we can equate the same powers of z on each side of the equation and obtain Equation B.3.

The characteristic. The characteristics of a differential equation are the lines in the z-t plane where G is constant. They are described by the equation $r(z-1)^2 dt = dz$. One can see that a function $f(\int \frac{dz}{r(z-1)^2} - t)$ will satisfy Equation B.3. Therefore we have:

$$G(z,t) = G(e^{\int \frac{dz}{r(z-1)^2} - t}) \tag{B.7}$$

That means, our solution G is any function with an implicit dependence on $f(\int \frac{dz}{r(z-1)^2} - t)$. To obtain an explicit solution we apply the initial condition.

Apply initial condition. We know that at $t = 0$ there is one cell, therefore $P(n,0) = \delta_{n,1}$ (i.e. P(n,0)=1 for n=1 and P(n,0)=0 for all other n). From there it follows that G(z,0)=z. We obtain the relation:

$$G(e^{\int \frac{dz}{r(z-1)^2} - t}) = z$$

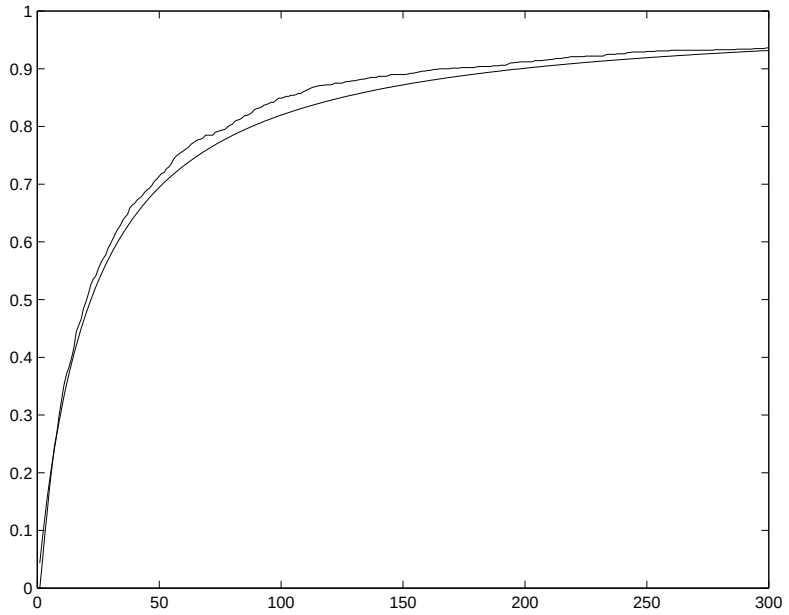

Figure B.2: Predicted and simulated P(0,t)

with $x = \frac{1}{e^{r(1-z)}}$ and $z = 1 - \frac{1}{rln(u)}$ we get

$$G(x) = 1 - \frac{1}{rln(x)}$$

and finally

$$G(z,t) = 1 - \frac{1}{rln(e^{\frac{1}{r(1-z)}-t})} = 1 - \frac{1}{\frac{1}{1-z} - rt} = \frac{z - rt(1-z)}{1 - rt(1-z)} \quad \text{(B.8)}$$

Looking back at Equation B.3 we can expand $G(z,t)$ in terms of powers of z and obtain P(0,t), P(1,t) etc.

Taylor expansion. The result for $G(z,t)$ expanded in a Taylor series looks as follows:

$$G(z,t) = \frac{a\,t}{1+a\,t} + \left(\frac{a^2\,t^2}{(1+a\,t)^2} + \frac{1-a\,t}{1+a\,t}\right) z + \left(\frac{a^3\,t^3}{(1+a\,t)^3} + \frac{a\,t\,(1-a\,t)}{(1+a\,t)^2}\right) z^2 + O(z^3) \quad \text{(B.9)}$$

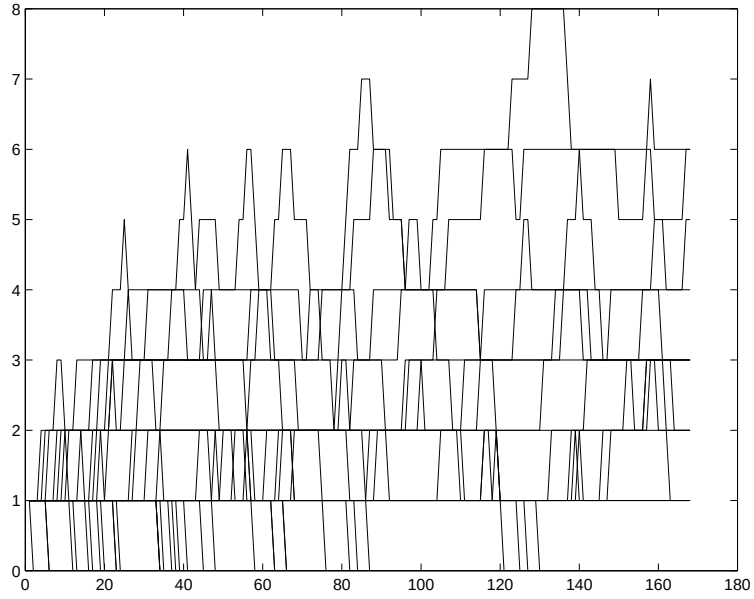

Figure B.3: Simulated run of 50 wells. The number of cells in each well is shown, simulation was run for 168h or 7 days.

From here we can obtain the solution for $P(n,t)$ by substituting it back into Equation B.3 and again, equate powers of z, i.e. for $P(0,t)$ we take the coefficient of z^0, for $P(1,t)$ the coefficient of z^1 and so on. then we obtain the solution given in Section B.2.

Comparison to experimental data

First we tested the predicted curves against curves we obtained from numerical simulations. The predicted $P(0,t)$ and the simulated $P(0,t)$ are shown in Figure B.2. The agreement is very good. In Figure B.3 a sample simulation run is shown which shows cell occupancies in 20 different wells for a simulation run of 6 days.

The calculated $P(n,t)$ was fitted to the data in Table B.1. A global fitting procedure which fits all three curves at once was used in the program package Mathematica. The fits are reasonably good, taking into account a sizable experimental error and the simplicity of the model. Fits to the other dataset worked equally well.

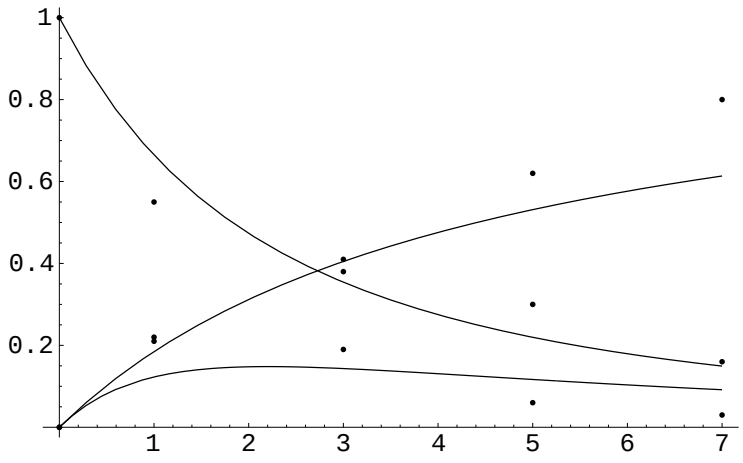

Figure B.4: The observed data and the predicted curves. A division time of 3.1 days resulted from the fit with a large Confidence interval (5.2 days to 2.1 days.

B.4 Conclusion

We developed a simplistic stochastic model for the survival of a single cell under tumor dormant conditions. Despite of its simplicity and strong assumptions, it seems to fit the given data well. On the other hand, a model which links apoptosis and cell division directly can not fit the experimental data. Therefore, our results suggest that apoptosis and cell division are indeed unlinked processes and that in the dormant state, their rates are matched to achieve population stability. However, single cell and culture conditions are quite different and it is quite conceivable that apoptosis and cell division are affected. Still, in the tumor dormant state, the B-lymphoma cells are likely to have no linkage between apoptosis and cell division, as a simple unlinked model matches the data far better than a model with direct linkage.

Bibliography

1. J. Aach, W. Rindone, and G. Church. Systematic management and analysis of yeast gene expression data. *Genome Res.*, **10**:431–445, 2000.

2. R. Albert, H. Jeong, and A. Barabasi. Error and attack tolerance of complex networks. *Nature*, **406**:378–382, 2000.

3. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.*, **96**:6745–6750, 1999.

4. O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.*, **97**:10101–10106, 2000.

5. A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *Escherichia coli* cells. *Genetics*, **149**:1633–1648, 1998.

6. G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue. BIND - the biomolecular interaction network database. *Nucleic Acids Res.*, **29**:242–245, 2001.

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

8. A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, **311**:590–614, 2002.

9. A. L. Barabasi. *Linked: The New Science of Networks*. Perseus Publishing, Cambridge, Massachusetts, 2002.

10. A. Becskel and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, **405**:590–593, 2000.

11. M. Bittner, P. Meltzer, Y. Chen, J. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, M. Hayward, and J. Trent. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**:536–540, 2000.

12. J. U. Bowie and R. T. Sauer. Identifying determinants of folding and activity for a protein of unknown structure. *Proc. Natl. Acad. Sci. U.S.A.*, **86**:2152–2156, 1989.

13. P. Broet, S. Richardson, and F. Radvanyi. Bayesian hierarchical model for identifying changes in gene expression from microarray experiments. *J. Comput. Biol.*, **9**:671–683, 2002.

14. C. S. Brown, P. C. Goodwin, and P. K. Sorger. Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:8944–8949, 2001.

15. M. P. S. Brown, W. N. Grundy, D. Lin, N. Christiani, C. W. Sugnet, T. S. Furey, M. A. Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.*, **97**:262–267, 2000.

16. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, **2**:65–73, 1998.

17. S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transciptional program of sporulation in budding yeast. *Science*, **282**:699–705, 1998.

18. H. A. Coller, C. Grandori, P. Tamayo, T. Colbert, E. S. Lander, R. N. Eisenman, and T. R. Golub. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. U.S.A.*, **97**:3260–3265, 2000.

19. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts, 2nd edition, 2001.

20. M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrells. YPD, PombePD, and WormPD: Model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**:75–79, 2001.

21. C. Deana, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics*, **5**:349–356, 2002.

22. J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**:680–686, 1997.

23. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, **95**:14863–14868, 1998.

24. M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**:335–338, 2000.

25. M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, **297**:1183–1186, 2002.

26. D. Endy, L. You, J. Yin, and I. J. Molineux. Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acd. Sci. U.S.A.*, **97**:5375–5380, 2000.

27. T. L. Ferea, D. Botstein, P. O. Brown, and R. F. Rosenzweig. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **96**:9721–9726, 1999.

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

29. A. Finney, M. Hucka, H. Sauro, J. Doyle, H. Kitano, and H. Bolouri. The systems biology markup language. *Mol. Biol. Cell*, **12**:708–61, 2001.

30. G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communitites. *Computer*, **53**:66–71, 2002.

31. S. P. A. Fodor, R. P. Rava, X. H. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams. Multiplexed biochemical assays with biological chips. *Nature*, **364**:555–556, 1993.

32. T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**:339–342, 2000.

33. A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**:141–147, 2002.

34. H. Ge, Z. Liu, G. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.*, **29**:482–486, 2001.

35. M. Gerstein, N. Lan, and R. Jansen. Proteomics - Integrating interactomes. *Science*, **295**:284–+, 2002.

36. G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, **97**:12079–12084, 2000.

37. D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**:2340–2361, 1977.

38. M. Girvan and M. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, **99**:7821–7826, 2002.

39. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**:531–537, 1999.

40. R. Gomory and T. Hu. Multi-terminal network flows. *J. Soc. Indust. Appl. Math.*, **9**:551–570, 1961.

41. S. Granjeaud, F. Bertucci, and B. R. Jordan. Expression profiling: DNA arrays in many guises. *BioEssays*, **21**:781–790, 1999.

42. D. Greenbaum, N. Luscombe, R. Jansen, J. Qian, and M. Gerstein. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.*, **11**:1463–1468, 2001.

43. J. Griffin, C. Mann, J. Scott, C. Shoulders, and J. Nicholson. Choline containing metabolites during cell transfection: an insight into magnetic resonance spectroscopy detectable changes. *FEBS Lett.*, **509**:263–266, 2001.

44. N. Guelzim, S. Bottani, P. Bourgine, and F. Kepes. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genet.*, **31**:60–63, 2002.

45. C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler. Combinatorial synthesis of genetic networks. *Science*, **296**:1466–1470, 2002.

46. D. Gusfield. Very simple methods for all pairs network flow analysis. *SIAM J. Comput.*, **19**:143–155, 1990.

47. N. Guttmann-Beck and R. Hassin. Approximation algorithms for minimum *K*-cut. *Algorthmica*, **27**:198–207, 2000.

48. S. Gygi, B. Rist, S. Gerber, F. Turecek, M. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.*, **17**:994–999, 1999.

49. P. Haney, J. Badger, G. Buldak, C. Reich, C. Woese, and G. Olsen. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extermely thermophilic *Methanococcus species*. *Proc. Natl. Acad. Sci. U.S.A.*, **96**:3578–3583, 1999.

50. J. Hao and J. B. Orlin. A faster algorithm for finding the minimum cut in a directed graph. *J. Algorithms*, **17**:424–446, 1994.

51. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomp.*, **6**:422–433, 2001.

52. L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. From molecular to modular cell biology. *Nature*, **402**:C47–C52, 1999.

53. Z. S. Hendsch, T. Jonsson, R. T. Sauer, and B. Tidor. Protein stabilization by removal of unsatisfied polar groups: Computational approaches and experimental tests. *Biochemistry*, **35**:7621–7625, 1996.

54. Z. S. Hendsch and B. Tidor. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.*, **3**:211–226, 1994.

55. L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Res.*, **9**:1106, 1999.

56. Y. Ho, A. Gruhler, G. D. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**:180–183, 2002.

57. Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. Willems, H. Sassi, P. Nielsen, K. Rasmussen, J. Andersen, L. Johansen, L. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. Sorensen, J. Matthiesen, R. Hendrickson, F. Gleeson, T. Pawson, M. Moran, D. Durocher, M. Mann, C. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**:180–183, 2002.

58. J. Hollien and S. Marqusee. A thermodynamic comparison of mesophilic and thermophilic ribonucleases H. *Biochemistry*, **38**:3831–3836, 1999.

59. F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**:717–728, 1998.

60. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennet, E. Coffey, H. Dai, Y. He, M. Kidd, A. King, M. Meyer, D. Slade, P. Lum, S. Stepaniants, D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. Friend. Functional discovery via a compendium of expression profiles. *Cell*, **102**:109–126, 2000.

61. T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Baumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**:929–934, 2001.

62. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**:S233–S240, 2002.

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

64. V. R. Iyer, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. F. Lee, J. M. Trent, L. M. Staudt, J. Hudson, M. S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P. O. Brown. The transcriptional program in the response to human fibroblasts to serum. *Science*, **283**:83–87, 1999.

65. V. Iyer, C. Horak, C. Scafe, D. Botstein, M. Snyder, and P. Brown. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**:533–538, 2001.

66. R. Jaenicke and G. Böhm. The stability of proteins in extreme environments. *Curr. Op. Struct. Biol.*, **8**:738–748, 1998.

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

68. H. Jeong, B. Tombor, R. Albert, Z. Oltval, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, **407**:651–654, 2000.

69. P. Karp, M. Riley, S. Paley, and A. Pellegrini-Toole. The MetaCyc database. *Nucleic Acids Res.*, **30**:59–61, 2002.

70. K. C. Keiler, P. R. Waller, and R. T. Sauer. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, **271**:990–993, 1996.

71. S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson. A gene expression map for *Caenorhabditis elegans*. *Science*, **293**:2087–2092, 2001.

72. H. Kitano. Computational systems biology. *Nature*, **420**:206–210, 2002.

73. H. Kitano. Standards for modeling. *Nature Biotechnol.*, **20**:337–337, 2002.

74. H. Kitano. Systems biology: A brief overview. *Science*, **295**:1662–1664, 2002.

75. E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. Waterston, R. Wilson, L. Hillier, J. McPherson, M. Marra, E. Mardis, L. Fulton, A. Chinwalla, K. Pepin, W. Gish, S. Chissoe, M. Wendl, K. Delehaunty, T. Miner, A. Delehaunty, J. Kramer, L. Cook, R. Fulton, D. Johnson, P. Minx, S. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. Gibbs, D. Muzny, S. Scherer, J. Bouck, E. Sodergren, K. Worley, C. Rives, J. Gorrell, M. Metzker, S. Naylor, R. Kucherlapati, D. Nelson, G. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. Davis, N. Federspiel, A. Abola, M. Proctor, R. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. Cox, M. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. Evans, M. Athanasiou, R. Schultz, B. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. Brown, C. Burge, L. Cerutti, H. Chen, D. Church, M. Clamp, R. Copley, T. Doerks, S. Eddy, E. Eichler, T. Furey, J. Galagan, J. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. Johnson, T. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. Kent, P. Kitts, E. Koonin, I. Korf, D. Kulp, D. Lancet, T. Lowe, A. McLysaght, T. Mikkelsen, J. Moran,

N. Mulder, V. Pollara, C. Ponting, G. Schuler, J. Schultz, G. Slater, A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. Wolf, K. Wolfe, S. Yang, R. Yeh, F. Collins, M. Guyer, J. Peterson, A. Felsenfeld, K. Wetterstrand, A. Patrinos, and M. Morgan. Initial sequencing and analysis of the human genome. *Nature*, **409**:860–921, 2001.

76. D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**:788–791, 1999.

77. D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Info. Proc. Syst.*, **13**:556–562, 2001.

78. T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, T. Volkert, E. Fraenkel, D. Gifford, and R. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**:799–804, 2002.

79. C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:31–36, 2001.

80. R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/l_1-l_1 regulatory elements. *Nucleic Acids Res.*, **25**:1203–1210, 1997.

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

82. S. M. Malakauskas and S. L. Mayo. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.*, **5**:470–475, 1998.

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

84. H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **94**:814–819, 1997.

85. H. W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.*, **28**:37–40, 2000.

86. H. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.*, **30**:31–34, 2002.

87. M. E. Milla, B. M. Brown, and R. T. Sauer. Protein stability effects of a complete set of alanine substitutions in Arc repressor. *Nature Struct. Biol.*, **1**:518–523, 1994.

88. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, **298**:824–827, 2002.

89. J. Misra, W. Schmitt, D. Hwang, L. L. Hsiao, S. Gullans, G. Stephanopoulos, and G. Stephanopoulos. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Research*, **12**:1112–1120, 2002.

90. J. Nicholson, J. Connelly, J. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Rev. Drug Discov.*, **1**:153–161, 2002.

91. W. Nultsch. *Allgemeine Botanik*. Thieme Verlag, Stuttgart, Germany, 10th edition, 2001.

92. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**:29–34, 1999.

93. Z. Oltvai and A. Barabasi. Life's complexity pyramid. *Science*, **298**:763–764, 2002.

94. R. Overbeek, N. Larsen, G. Pusch, M. D'Souza, E. Selkov, N. Kyrpides, M. Fonstein, N. Maltsev, and E. Selkov. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**:123–125, 2000.

95. Y. Pilpel, P. Sudarsanam, and G. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**:153–159, 2001.

96. J. Rain, L. Selig, H. D. Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. The protein-protein interaction map of helicobacter pylori. *Nature*, **409**:211–215, 2001.

97. C. V. Rao and A. P. Arkin. Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.*, **3**:391–419, 2001.

98. E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, **297**:1551–1555, 2002.

99. D. Rees and M. Adams. Hyperthermophiles: Taking the heat and loving it. *Structure*, **3**:251–254, 1995.

100. L. E. Reichl. *Statistical Physics*. John Wiley and Sons, New York, 1997.

101. B. Ren, F. Robert, J. Wyrick, O. Aparicio, E. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. Volkert, C. Wilson, S. Bell, and R. Young. Genome-wide location and function of DNA binding proteins. *Science*, **290**:2306–+, 2000.

102. C. Robinson, D. Rentzeperis, J. Silva, and R. Sauer. Formation of a denatured dimer limits the thermal stability of Arc repressor. *J. Mol. Biol.*, **273**:692–700, 1997.

103. M. Ronen, R. Rosenberg, B. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U. S. A.*, **99**:10555–10560, 2002.

104. N. Rosenfeld, M. B. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.*, **323**:785–793, 2002.

105. H. Salgado, A. Santos-Zavaleta, S. Gama-Castro, D. Millen-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides. RegulonDB (version 3.2): Transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**:72–74, 2001.

106. H. Saran and V. V. Vazirani. Finding k cuts within twice the optimal. *SIAM J. Comput.*, **24**:101–108, 1995.

107. M. A. Savageau. *Biochemical Systems Theory*. Addison-Wesley, Reading, Massachusetts, 1978.

108. M. Schena, D. Schalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, **270**:467–470, 1995.

109. B. Schoeberl, C. Eicher-Johnsson, E. D. Gilles, and G. Mueller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnol.*, **20**:370–375, 2002.

110. B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnol.*, **18**:1257–1261, 2000.

111. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.*, **31**:64–68, 2002.

112. G. Sherlock. Analysis of large-scale expression data. *Curr. Op. in Immun.*, **12**:201–205, 2000.

113. S. Y. Shvartsman, C. B. Muratov, and D. A. Lauffenburger. Modeling and computational analysis of EGF receptor-mediated cell communication in drosophila oogenesis. *Development*, **128**:2577–2589, 2002.

114. R. Somogyi and C. A. Sniegoski. Modeling the complexity of genetic networks: Understanding multigenetic and pleitropic regulation. *Complexity*, **1**:45–63, 1996.

115. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**:3273–3297, 1998.

116. S. Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1994.

117. L. Stryer. *Biochemistry*. WH Freeman and Co., New York, NY, 4th edition, 1995.

118. J. Sung and S. Lee. Nonequilibrium distribution function formalism for diffusion-influenced bimolecular reactions: Beyond the superposition approximation. *J. Chem. Phys*, **111**:796–803, 1999.

119. P. Tamayo, D. Slomin, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. U.S.A.*, **96**:2907–2912, 1999.

120. A. Tanay and R. Shamir. Computational expansion of genetic networks. *Bioinformatics*, **17**:S270–S278, 2001.

121. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genet.*, **22**:281–285, 1999.

122. M. Thattai and A. van Oudenaarden. Intrinic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.*, **98**:8614–8619, 2001.

123. P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. J. Yang, M. Johnston, S. Fields, and J. Rothberg. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**:623–627, 2000.

124. J. Uhr, R. Scheuermann, N. Street, and E. Vitetta. Cancer dormancy: Opportunities for new therapeutic approaches. *Nat. Med.*, **3**:505–509, 1997.

125. G. van Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, **406**:188–192, 2000.

126. J. Venter, M. Adams, E. Myers, P. Li, R. Mural, G. Sutton, H. Smith, M. Yandell, C. Evans, R. Holt, J. Gocayne, P. Amanatides, R. Ballew, D. Huson, J. Wortman, Q. Zhang, C. Kodira, X. Zheng, L. Chen, M. Skupski, G. Subramanian, P. Thomas, J. Zhang, G. Miklos, C. Nelson, S. Broder, A. Clark, C. Nadeau, V. McKusick, N. Zinder, A. Levine, R. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. Heiman, M. Higgins, R. Ji, Z. Ke, K. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. Merkulov, N. Milshina, H. Moore, A. Naik, V. Narayan, B. Neelam, D. Nusskern, D. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik,

T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. Abril, R. Guigo, M. Campbell, K. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, **291**:1304–+, 2001.

127. C. Vetriani, D. Maeder, N. Tolliday, K. Yip, T. Stillman, K. Britton, D. Rice, H. Klump, and F. Robb. Protein thermostability above 100 C: A key role for ionic interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **95**:12300–12305, 1998.

128. D. Voet and J. G. Voet. *Biochemistry*. John Wiley and Sons, New York, NY, 2nd edition, 1995.

129. C. D. Waldburger, J. F. Schildbach, and R. T. Sauer. Are buried salt bridges important for protein stability and conformational specificity? *Nature Struct. Biol.*, **2**:122–128, 1995.

130. A. Walhout, R. Sordella, X. Lu, J. Hartley, G. Temple, M. Brasch, N. Thierry-Mieg, and M. Vidal. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**:116–122, 2000.

131. R. Wehner and W. Gehring. *Zoologie*. Thieme Verlag, Stuttgart, Germany, 22nd edition, 1990.

132. E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**:281–283, 2001.

133. I. Xenarios, L. Salwinski, X. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**:303–305, 2002.

134. I. Xenarios and D. Eisenberg. Protein interaction databases. *Curr. Op. Biotech.*, **12**:334–339, 2001.

135. L. Xiao and B. Honig. Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.*, **289**:1435–1444, 1999.

136. H. Zhou, J. Watts, and R. Aebersold. A systematic approach to the analysis of protein phosphorylation. *Nature Biotechnol.*, **19**:375–378, 2001.

137. H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. Dean, M. Gerstein, and M. Snyder. Global analysis of protein activities using proteome chips. *Science*, **293**:2101–2105, 2001.

138. Z. Zhu, Y. Pilpel, and G. M. Church. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.*, **318**:71–81, 2002.