# Inferring the influence of cultivation parameters on transcriptional regulation

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J.T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op vrijdag 20 maart 2009 om 12:30 uur
door

Theo Arjan KNIJNENBURG

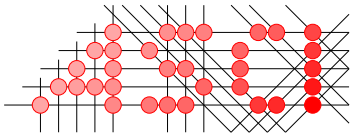
elektrotechnisch ingenieur
geboren te Leidschendam.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. M.J.T. Reinders

Co-promotor:
Dr. L.F.A. Wessels

Samenstelling promotiecommissie:

| | |
|---|---|
| Rector Magnificus | voorzitter |
| Prof. dr. ir. M.J.T. Reinders | Technische Universiteit Delft, promotor |
| Dr. L.F.A. Wessels | Technische Universiteit Delft, co-promotor |
| Prof. dr. J.T. Pronk | Technische Universiteit Delft |
| Prof. dr. R.C. Jansen | Rijksuniversiteit Groningen |
| Prof. dr. F.C.P. Holstege | Universiteit Utrecht |
| Prof. dr. C.J.F. ter Braak | Wageningen Universiteit en Researchcentrum |
| Prof. dr. T.M. Heskes | Radboud Universiteit Nijmegen |
| Prof. dr. ir. R.L. Lagendijk | Technische Universiteit Delft, reservelid |

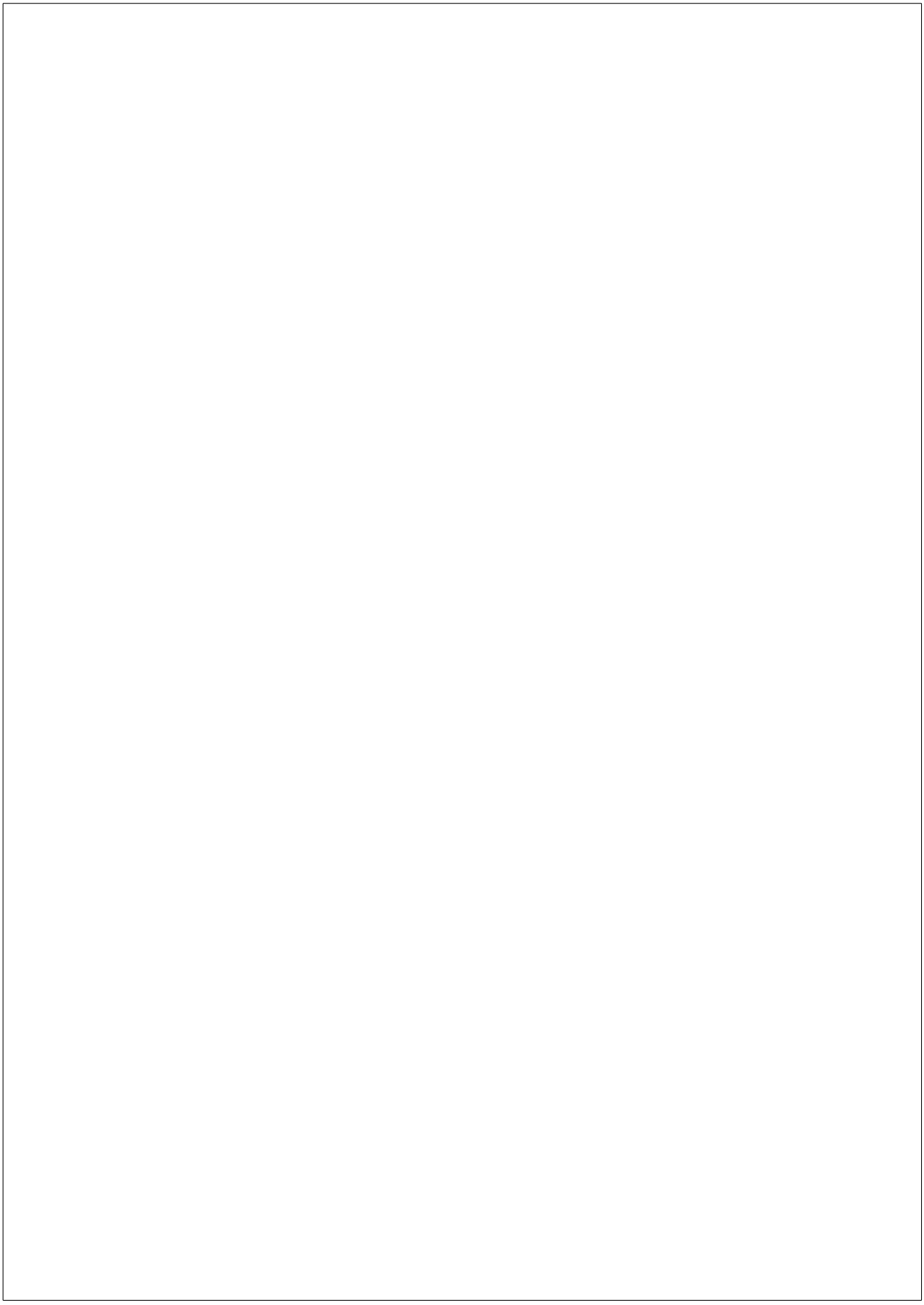



This work was carried out in the ASCI graduate school.
ASCI dissertation series number 174.

This work forms a part of the research performed in the Kluyver Centre for Genomics of Industrial Fermentation and was financially supported by the Netherlands Genomics Initiative (NGI).

ISBN 978-90-9024013-8

*to my parents*



# CONTENTS

# CHAPTER 1

# INTRODUCTION

## 1.1 Scope of the thesis

Cell biology seeks to understand the cell, life's fundamental building block. It studies the structure and function of the cell, its (intracellular) constituents and its interaction with the (extracellular) environment. It studies the cell's life cycle and, in multi-cellular organisms, it studies cellular differentiation, which is the process by which cells become specific types of cells, such as skin cells, localized in specific parts of the organism.
The development of measurement techniques has provided the means to observe intracellular components down to the molecular level. By far the most important discovery was that of DNA, the molecule that contains the genetic instructions used in the development and functioning of all known living organisms. It gave rise to the central dogma of molecular biology that presents the genetic information flow within the cell from DNA to RNA to proteins, of which the latter can be seen as both the structural and functional units of the cell. Nowadays, many aspects of the intracellular molecules and subunits can be measured, enabling cell biologists to hypothesize about the interaction between the cell's different components and elucidate cellular mechanisms.
Besides the acquisition of fundamental knowledge, cell biology research finds direct applications in the medical and the industrial domain, mainly focusing on (human) disease and food production using microorganisms. Among the many things learned about the cell the past decades, is the reality that (despite the simplicity of genetic information flow) the cell forms an overwhelmingly complex control system, which is far from being understood and for which much remains to be discovered.

The modern 'high-throughput' measurement techniques could provide a valuable data source for such discoveries. These techniques are termed high-throughput, because they enable one to quickly conduct thousands or even millions of biochemical measurements. One example of such a measuring technique (which will play a large role in this thesis) is the gene expression microarray. A gene is a region on the DNA, corresponding to a unit of inheritance, which can be transcribed into an RNA and later translated into a protein.

The microarray technology provides a score (or measurement) of the concentration (or quantity) of RNAs in a cell (or sample) for each gene in the genome, which, for most organisms, is on the order of thousands. The obtained snapshot of the genome-wide gene expression (also called transcriptome) forms an unbiased and complete starting point for data analysis. This is in contrast to small scale gene expression experiments, where only the expression of presumably relevant genes is investigated. Another example of a high-throughput technique is genome (or DNA) sequencing. The latest genome sequencers can determine the order of the four nucleotide bases, adenine (A), guanine (G), cytosine (C), and thymine (T) that make up the DNA sequence of any organism with unprecedented speed, enabling scientists to sequence a complete genome (typically billions of bases) in a reasonable amount of time.
Analysis and interpretation of the enormous amounts of data generated with the high-throughput techniques necessitates the use of computational tools. The interface between cell biology and informatics, where computer science techniques, machine learning and statistics are employed to analyze cell biology data, can be labeled with the partially redundant names bioinformatics, biostatistics and computational biology. To successfully address a biological question using large amounts of data, the main challenge is to apply, design or modify computer techniques while taking into account 1) the properties of the biological system or entities upon which the measurements were performed, 2) the properties of the employed measurement device and 3) the properties of the employed algorithms themselves.

This thesis embodies an example of the interface between cell biology and informatics. Our research activities are focused on the transcriptional program of yeast, in particular *Saccharomyces cerevisiae* (or baker's yeast). This unicellular microorganism is known for its age-old application in alcohol fermentation (mostly beer and wine) and the baking of bread. Besides obtaining a fundamental understanding of this organism on a molecular level, applications are indeed found in the food and drink industry. Another, more recent, application is that of biofuel, where yeast is used to convert the sugars in biological material, such as plants and crops, to ethanol.
A substantial part of current yeast research is aimed at unraveling the transcriptional program of *S. cerevisiae*. This yeast has about 6400 genes, of which each gene product (or protein) fulfills one or more specific functions in the cell. Although yeast is among the best-studied organisms and for many genes the function or process of involvement is roughly known, the interaction between the different gene products and precise information on how or when genes are transcribed remains to be uncovered. More specifically, little is known on how yeast integrates the multiple chemical and physical signals from its environment to adapt its transcriptome. Also, the mechanisms behind the activation of the proteins that control the rate at which a gene is transcribed, i.e. the transcription factors (TFs) and chromatin remodeling proteins (CRPs), are yet to be elucidated.
To unravel the transcriptional program of baker's yeast, we analyze the transcriptional response of this yeast to different environmental conditions. In contrast to the commonly used shake-flask cultures, the microarray data employed in this thesis originates from yeast grown in chemostat cultures. In a chemostat, culture broth (including biomass) is continuously replaced by fresh medium at a fixed and accurately determined dilution rate. These steady-state chemostat cultures enable the accurate control, measurement and manipulation of individual cultivation parameters, such as growth rate, temperature and nutrient concentrations. A growth condition can thus be characterized by the com-

bined settings of several cultivation parameters. Analysis of the gene expression levels as obtained with microarray measurements allows for identification of the influence of these cultivation parameters on the transcriptome of the yeast cell. The main contents of this thesis consists of exploring methods that incorporate the cultivation parameters into the computational model in order to identify the influence of these cultivation parameters on gene expression and the activity of TFs.

## 1.2 Incorporating the growth conditions in the computational model

In order for yeast to adapt its transcriptome to changes in the extracellular environment, the information on the environment has to be transmitted to the cell nucleus, where the DNA is situated and where the information carriers can alter gene expression. Here, we divide this system into two parts. First, the sensing, importing and subsequent signaling of the environment that alters the activity of the TFs and CRPs. Second, the activity of TFs and CRPs that are able to manipulate gene transcription rates. Figure 1.1 graphically depicts this model of the cell's transcriptional response to its extracellular environment.

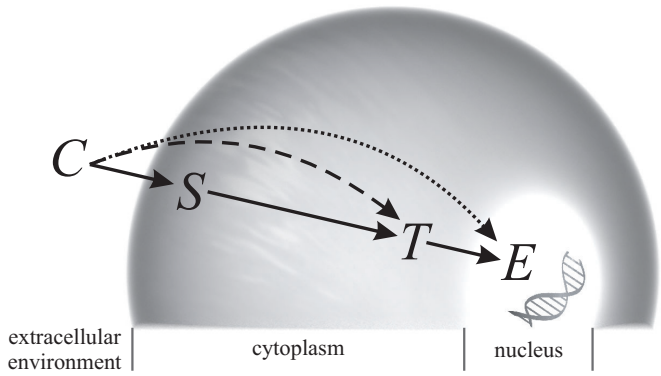



**Figure 1.1 – Schematic overview of the four factors involved in the cell's transcriptional response to its extracellular environment.**
The black arrows indicate the complete flow from environmental conditions to gene expression: Extracellular conditions ($C$) are imported or sensed, resulting in intracellular signaling ($S$). These signals alter the activity of TFs and CRPs ($T$), which manipulate the rate at which genes are expressed ($E$). Dashed and dotted arrows indicate shortcuts to model the influence of $C$ on $E$, i.e. by leaving some of the factors out of the model.

*Models of transcriptional regulation as employed in this thesis*

In this thesis we focus on how extracellular conditions ($C$) affect gene expression ($E$) and the activity of TFs and CRPs ($T$). The diverse mechanisms that the cell uses to sense or import the myriad of different external stimuli or molecules and their subsequent signaling ($S$) are not part of the employed models. Two different models are employed:

In the first, gene expression is directly modeled as a function of the extracellular condi-

tions:

$$E = f(C) \tag{1.1}$$

This scenario is illustrated by the dotted arrow in Figure 1.1. Here, the aim is to infer the influence of the different environmental stimuli to which the cell is exposed, on gene transcription rates.

The second model incorporates $T$, the activity of transcription factors (TFs) and chromatin remodeling proteins (CRPs). In general, the transcription rate (of which the measured expression level is a score) is the net result of the highly non-linear and complex interplay between these regulatory proteins. The TFs (some of which bind the DNA near the gene) form the general transcription machinery that copies the gene into its mRNA equivalent. This process can be manipulated by gene specific TFs (also some of which can bind the DNA) that interact with the transcription machinery and thereby manipulate (enhance or repress) the rate at which a gene is transcribed. The CRPs package the DNA and thereby influence the accessibility of the TFs to the DNA. Modification of the activity of these proteins is the means by which the cell dynamically regulates its transcriptome in order to carry out cellular processes or adapt to changes in the extracellular environment. In the second approach, gene expression is modeled as a function of the activity of TFs and CRPs, which is itself a function of the extracellular conditions:

$$E = g(T) \tag{2.1}$$

$$T = h(C) \tag{2.2}$$

This scenario is illustrated by the dashed arrow in Figure 1.1. (In this thesis, $T$ is restricted to represent the activity of TFs, since data is lacking for CRPs.)

The cell is a complex control system with feedback mechanisms on and between many levels. Therefore, one would expect to also see arrows from $E$ to $T$ and $S$ in Figure 1.1. (Obviously, the proteins that participate in $T$ and $S$ are gene products resulting from $E$.) However, the microarray data that is employed in this thesis originates from steady-state cultures, where the intracellular and extracellular parameters that encompass $C$, $S$, $T$ and $E$ are assumed to be constant. Therefore, the model employs the linear chain of cause and effect that runs from the ultimate cause $C$ via $S$ and $T$ to $E$.

*Multifactorial descriptions of the extracellular conditions (C)*

In order to gain a thorough understanding of yeast's transcriptome in response to different environments, it is crucial to systematically identify the different parameters that characterize the cell's environment. Only in this way can individual and combined effects of the environmental parameters on gene expression and TF activity be analyzed. A condition for which a microarray experiment has been performed should thus be described by a multifactorial variable, where the factors represent different environmental parameters. Figure 1.2 provides a graphic description of microarray datasets with single-factor and multifactorial variables.

Obviously, chemostat steady-state cultivation forms an excellent platform to incorporate multi-factorial growth conditions into the computational model. Chemostats allow one to accurately control and measure many of the extracellular parameters, such that

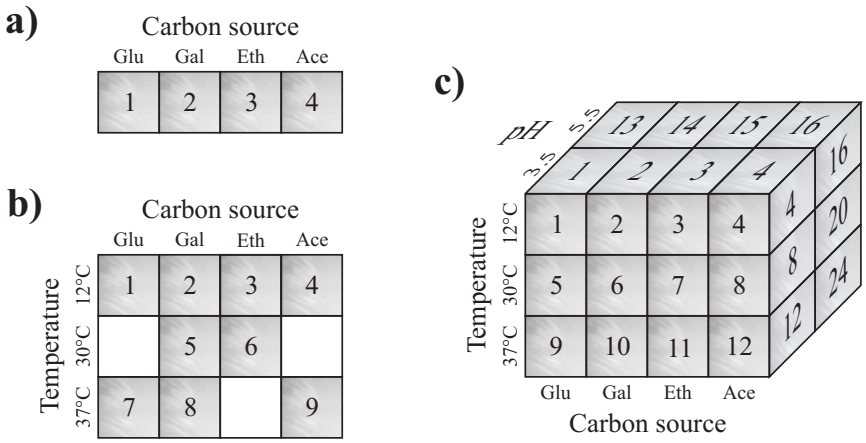



**Figure 1.2 – Schematic representation of the conditions within three microarray datasets.**
Each numbered cell represents one microarray experiment/condition. **a)** A dataset consisting of four microarrays. The conditions are described by a single factor, namely 'Carbon source' which can assume four different values ('Glu', 'Gal', 'Eth' and 'Ace'). **b)** A dataset consisting of nine microarrays. In this case, the condition is a multi-factorial variable. That is, a condition is described by multiple (two) factors, namely 'Carbon source' and 'Temperature'. Note that not all possible combinations of these factors are measured (indicated by the white unnumbered cells). **c)** A dataset consisting of 24 microarrays. A condition is described by three factors. This dataset has a full combinatorial (or factorial) design. Although, in this example all factors are categorical (or nominal) variables, in principle, the factors can also be ordinal, interval or ratio variables.

one can characterize a growth condition by the settings or values of multiple cultivation parameters. For example, the yeast cell's environment can be characterized using parameters, like nutrient concentrations, temperature, oxygen availability, etc. The interrelations between the different growth conditions (in terms of comparable cultivation parameters) can be exploited in order to infer the influence of not only single cultivation parameters, but also of combinations of cultivation parameters (interaction effects) on the gene expression levels. (In contrast, for datasets generated with shake-flasks, the multi-factorial space cannot be systematically explored, since not all parameters that span this space can be strictly controlled and accurately measured.) Furthermore, including information on TF binding allows for the investigation of the effect of cultivation parameters on the activity of TFs. This could offer great insight in the aforementioned interplay between these proteins, which regulate the copying process of the genes.

## 1.3 Related work

Microarray measurements are performed to compare an organism's gene expression levels between different conditions. Up to this point, these 'conditions' were mainly seen as growth environments of the yeast cell. However, in a broader context, these conditions can also be different types of human tissue, such as healthy liver cells or cancerous skin cells, different strains of the *E. coli* bacterium, e.g. strains with or without a slight genetic modification (mutant and wild-type, respectively) or white blood cells of black

mice with different diets, etc.

*Single-factor conditions*

For most (and especially the earliest) microarray experiments the conditions are characterized by a single categorical variable that, in most cases, assumes one of two values. In other words, the microarray experiments were performed to compare two conditions, and these conditions were not described by different parameters. For example, in DeRisi *et al.* [DeRi 96] two different human melanoma cell lines were compared to each other; in Alon *et al.* [Alon 99] colon tumor tissue samples were compared to healthy colon tissue samples. Usually, one condition is seen as the condition of interest (in medical settings this is commonly the diseased sample) and the other as the reference condition (commonly the healthy sample). The main goal of these approaches is to identify the genes, which are differentially expressed (upregulated or downregulated) in the condition of interest with respect to the reference condition. This strategy is implicitly incorporated into two-channel microarray systems, which output the ($\log_2$) ratio between the condition of interest and the reference condition[1]. Some approaches use the conditions as class labels to define two-class classification problems, e.g. Golub *et al.* [Golu 99]. Both in detection of differentially expressed genes and these classification problems, a class-dependent representation of the gene expression levels is built. For example, in a T-test the gene expression levels of each class are represented by a normal distribution. Thus, implicitly, these strategies implement Eq. 1.1, where gene expression is modeled as a function of the conditions (class labels).
Also in yeast research the dual channel system is most often employed to measure transcript levels. Typically, the reference condition is yeast growing under rich medium conditions in exponential phase (in a shake-flask). For example, the stress response microarray compendium by Gasch *et al.* [Gasc 00] compares this condition to thirteen different stress conditions, such as temperature shocks, addition of hydrogen peroxide and amino acid starvation, most of which are followed over time, leading to a total of 142 microarray experiments. One of these thirteen stresses is 'hydrogen peroxide treatment', which is formed by the microarray measurements for samples taken at 10, 20, 30, 40, 50, 60, 80, 100 and 120 minutes after addition of hydrogen peroxide to the medium. Although it might seem that the condition is described as a multi-factorial variable, where the different stresses and time would form the factors, it cannot be treated as such. First, all stresses are applied independently of one another; factors are not combined (in a systematic fashion). But more importantly, the stresses like 'hydrogen peroxide treatment' are merely terms that provide a global description of the applied stress; they are not measurable parameters that define the cell's environment. This is not even possible with (the often used) shake-flask cultures, because the parameters cannot be controlled (nor measured), but are continuously changing. Several approaches have described and used the experimental conditions as a single-factor categorical variable that can assume thirteen values; however not to explain gene expression as in Eq. 1.1, but in the post-analysis stage for data interpretation purposes. In Gasch *et al.* [Gasc 00] visual inspection of the expression patterns reveals how clusters of genes respond to the thirteen condition groups. (The clustering was performed without taking the grouping into account.) In Segal *et al.* [Sega 03], which also uses this gene expression data, co-

[1] The microarray data used in this thesis originates from single-channel scanners, which output scores that represent absolute mRNA concentrations.

regulated experiments are consulted for overrepresentation in one of the thirteen groups in order to link these conditions to the inferred regulatory program. (This method is explained in more detail below.)

In general, when analyzing a microarray dataset in terms of experimental conditions, it is crucial to characterize all the cultivation parameters (or other factors) that differ between the experiments, such that differential expression can be fully explained in terms of these parameters. This is the reason why the great majority of computational approaches that employ large amounts of transcript data (i.e. many microarrays) cannot take into account the conditions. Besides unquantified cultivation parameters due to shake-flask cultivation, most of these approaches combine microarray measurements from different studies, different laboratories (each with their own microarray protocol) and different yeast strains. These differences result in large confounding effects on gene expression (as we demonstrate in Chapter 6), since there are many unspecified factors contributing to the differences in gene expression levels. For example, in Bar-Joseph *et al.* [Bar 03] expression data from 23 studies is combined, totaling over 500 microarrays, and integrated with TF binding information to find co-regulated gene modules. These modules are uncovered using only the similarity (co-expression) between genes across all microarrays; not the conditions under which these microarray measurements were taken.

*Multifactorial conditions*

Different approaches have already demonstrated that if microarray conditions are multifactorial (i.e. they can be decomposed into multiple factors) or if additional information on the conditions is available, incorporating this into the computational model is useful. For example, Pittman *et al.* [Pitt 04] demonstrates that combining expression data of patients with their clinical factors, such as lymph node status and tumor size, improves the prediction accuracy of the disease outcome and survival time. In Matsui *et al.* [Mats 07] multivariate linear regression is used to model gene expression as a linear function of multiple clinical phenotypes of bladder cancer patients, such as pathological stage and grade. Also, normalization procedures of microarray experiments successfully employ multi-factorial models to reflect experimental design (e.g. array and dye effects) and thereby correct for potential confounding effects [Kerr 00]. Several recent approaches have focused more specifically on the combinatorial effects of experimental parameters on transcriptional regulation. These effects can only be uncovered using multifactorial conditions. For example, in Smith and Kruglyak [Smit 08] two different yeast strains were grown on two different carbon sources to identify strain-environment interactions on gene expression. Odom *et al.* [Odom 07] studied the tissue-specific differences between transcriptional regulation in human and mouse.
Chemostat cultivation, where distinct features of the cell's environment are measured, naturally provides multi-factorial conditions. Early chemostat analyses mainly used pair-wise comparisons that aim to identify genes that are differentially expressed between two conditions [Boer 03, Tai 05], e.g. between aerobic and anaerobic growth at 30°C. These sets can then be analyzed for the individual gene content or functional overrepresentation. Furthermore, these gene sets can be compared to other gene sets derived from similar pair-wise experiments, e.g. between aerobic and anaerobic growth at 12°C. This allows for the identification of context-specific (or combinatorial) effects of these

environmental parameters on gene expression. However, pair-wise comparisons provide only limited computational possibilities when analyzing many different conditions described by many parameters, since it is not straightforward to combine all pair-wise comparisons.

*Inferring TF activity*

Besides the (combinatorial) effects of cultivation parameters on gene expression, this thesis also focuses on the (combinatorial) regulation of gene expression by TFs. Many of the approaches in this area first cluster genes based on expression data and then try to infer the regulatory program of TFs or their DNA binding sites (motifs) for each cluster. These methods are based on the assumption that the similarly expressed genes in a cluster are regulated in the same way. For example, in Beer and Tavazoie [Beer 04] a Bayesian model is employed to explain cluster membership in terms of motif presence including their orientation and distance to each other and to the transcription start site. Other types of methods employ regression to relate TFs and expression. Bussemaker *et al.* [Buss 01] employed a forward step-wise regression strategy to explain the measured gene expression levels. The predictors were formed by motif counts in the promoter regions of individual genes. Later, other sources of TF binding potential were used as predictors, such as ChIP-chip TF binding data and scores obtained by scanning promoter regions with TF binding site information. In stead of modeling the TF activity as a hidden variable that needs to be estimated, Segal *et al.* [Sega 03] inferred the activity of a TF from the measured expression level of the gene encoding the TF. In this work, genes are grouped into regulatory modules, which are defined by a hierarchical decision tree, where the decisions at the nodes of the tree are based on the expression levels of regulators, such as TFs. A major drawback of this approach is the fact that most TFs are post-transcriptionally regulated, resulting in a poor correlation between the TF in its active form and the expression of the gene encoding the TF.
All these approaches, in one way or another, implement Eq. 2.1 by modeling the expression levels as a function of the activity of TFs. However, neither these approaches nor any other aim to model the activity of TFs as a function of the cultivation parameters (Eq. 2.2).

## 1.4 Contribution of this thesis

The computational methods that will be presented in this thesis integrate gene expression data with the growth conditions under which the microarrays were performed. One important element of some of these approaches is the discretization of continuous expression levels (Chapters 2, 3 and 4). The discretized representations of the gene expression patterns indicate up- and downregulation under the cultivation parameters that comprise the conditions of the microarray dataset. For example, a gene is characterized as upregulated under zinc limitation irrespective of oxygen presence. Similarly, clusters, which are formed by grouping genes with (nearly) identical discretized expression representations, can also be described in terms of their transcriptional response to particular cultivation parameters. In these discretization procedures, information on gene expression levels is sacrificed, while interpretability is gained. That is, discretizing the continuous gene expression levels results in a very compact and crude representation of the expression behavior of the genes. On the other hand, the discretized representations,

which are described in terms of the employed growth conditions, allow one to understand the expression behavior in terms of the growth conditions. This facilitates one to generate hypotheses about the influence of growth parameters on transcriptional regulation. In standard gene clustering approaches, the 'functionality' of a cluster is often determined using enrichment tests that identify overrepresented functional categories (e.g. heavy metal ion transport) amongst the cluster's genes. For clusters that are characterized in terms of growth conditions (by the discretization procedure on chemostat data), enrichment tests can point to the influence of particular growth conditions on TF activity, cellular functions, biological processes, etc. For example, based on the cluster of genes that is upregulated under zinc limitation irrespective of oxygen presence and enriched for heavy metal ion transport, one can hypothesize that zinc limitation affects heavy metal ion transport.

Yet, the main contribution of this thesis is formed by the computational approaches that model gene expression and TF activity as a function of the cultivation parameters. Here, the combinatorial setup of cultivation parameters within different growth conditions is used to investigate the (combinatorial) influence of these cultivation parameters on gene expression and TF activity. These methods are implementations of Eqs. 1.1, 2.1 and 2.2. In one approach, we model oxygen presence as a linear effect (having both an additive and multiplicative component) on gene expression (Chapter 4). Here, we demonstrate that exploiting the interrelatedness between growth conditions increases the interpretability and functional enrichment of uncovered gene clusters. In another approach, a linear regression strategy was applied to reconstruct measured gene expression patterns by selecting significant (combinations of) cultivation parameters as predictors in the regression model (Chapter 6). This is an implementation of Eq. 1.1, where gene expression is modeled as a linearly weighted sum of the contribution of significant cultivation parameters. Here, we show that including combinatorial effects leads to more sensible clusters in terms of enrichment of functional categories. Also, more variance within the gene expression patterns can be explained when taking the interaction effects between cultivation parameters into account.

Most techniques presented in this thesis employ hypergeometric tests to the infer the activity of TFs (Chapters 2-6). The test assesses the significance of the overlap between a cluster of genes and the regulon of a TF, i.e. all genes that can be bound (upstream) by the TF. Since genes are clustered based on their shared discretized representation (Chapters 2-5) or on the shared response to a cultivation parameter (Chapter 6), it is possible to link cultivation parameters to TFs. More specifically, one can hypothesize that a TF is activated in response to a particular cultivation parameter. However, the statistical test does not model TF activity as the causal relationship given in Eqs. 2.1 and 2.2. In one of the final methodologies presented in this thesis, we do model the activity of TFs as a function of cultivation parameters by inferring which (combination of) cultivation parameters activate which TFs (Eq. 2.2). Simultaneously, the model infers how activated TFs interact with each other to upregulate or downregulate the expression of a gene (Eq. 2.1), thereby elucidating the interplay of the TFs on the upstream regions of genes (Chapter 7).

The main motivation behind the incorporation of the growth conditions into the com-

putational model is to enable the interpretation of the results in terms of the growth conditions. The ability to identify the influence of (combinatorial) cultivation parameters on gene expression provides detailed clues towards the functionality of individual (uncharacterized) genes and pathways as well the activation of TFs. Validation of the results is achieved by consulting literature and yeast micro-biologists. Additional validation is provided by enrichment tests. In these tests, groups of genes obtained with the applied methods are compared to functionally related groups from gene annotation databases for significant overlap (Chapter 8). Higher enrichment indicates a larger ability of the method to capture functional association between genes.

## 1.5 Outline of the thesis

The theme of this thesis is the incorporation of the growth conditions in the computational model used to analyze the gene expression data. The goal is to interpret or understand the results in terms of the cultivation parameters that characterize the growth conditions. The thesis advances from analyses on small microarray datasets with simple cultivation parameter integration strategies to very large gene expression datasets approached with a more complex integration of both the growth conditions as well as TF binding data.

In Chapter 2 the transcriptional response of *S. cerevisiae* to four different weak organic acids is studied. Here, an unstressed condition (no addition of organic acids) is used as a reference condition. A discretization procedure is designed, such that each gene is represented by a tertiary (-1, 0, 1) vector of length four, indicating up- or downregulation of a gene when yeast is exposed to each of the four acids with respect to the unstressed condition. Based on these discrete representations genes are grouped into clusters. The clusters, which are readily interpretable with respect to the four organic acids, are consulted for enrichment of functional categories and TF binding. This study reveals that *S. cerevisiae* exhibits a minimal generic transcriptional response to weak organic acids. The consequences of these findings are that the often-used term 'weak-organic acid stress' should preferably be avoided and that the use of individual organic acids as 'model compounds' for general responses to organic acids should be treated with caution.

This chapter was published in FEMS Yeast Research, 2007.

Chapter 3 reports on a similar analysis. Here, microarray data is employed of *S. cerevisiae* grown under six different conditions, i.e. three different nutrient limitations; carbon, nitrogen and zinc, grown both aerobically and anaerobically. Discretization is used to build a tertiary representation of the genes. In this case, however, there is no reference condition. This makes it non-trivial to decide upon up- and downregulation. The discretization procedure uses a k-means clustering procedure for each gene individually; the six conditions are clustered to decide, which of these conditions are labeled upregulated, downregulated or not differentially expressed. In this work, genes are clustered together when their discretized expression patterns satisfy certain constraints. For example, genes that have a higher discretized expression value under zinc limitation than under the other two limitations in both the aerobic and anaerobic case are grouped together. The results from this analysis were used to redefine the zinc-specific Zap1p

regulon. Also, the study reveals a more important role for zinc in mitochondrial function and biogenesis than so far assumed.

This chapter was published in Applied and Environmental Microbiology, 2007.

In Chapter 4 a microarray dataset of eight conditions is analyzed. In this case, there are four different nutrient limitations; carbon, nitrogen and phosphorus and sulfur, grown both aerobically and anaerobically. Using a regression strategy the effect of oxygen presence on the expression of each gene is modeled as a linear effect (having both an additive and multiplicative component). The estimated parameters (offset and slope) are employed to 'correct for' the oxygen effect in the expression pattern. A discretization procedure is designed to represent each gene with a tertiary vector of length nine, where the last entry is that of the oxygen effect. Genes are clustered based on their discretized representations and related to TF binding data to infer the (combinatorial) effect of oxygen availability and nutrient limitations on TF activity. The inclusion of the cultivation parameters in uncovering regulatory modules and TF activity leads to a more valuable regulatory network that resultantly provides detailed insight in yeasts respiration and metabolism. The power of this approach in recognizing the individual and combinatorial effects of nutrient-limitations and oxygen presence is reflected in the results that strengthen and broaden the existing knowledge on regulatory mechanisms. For example, our results confirm the established role of TF Hap4 in both aerobic regulation and glucose derepression.

This chapter was published in BMC Genomics, 2007.

Chapter 5 uses the results of Chapter 4 to focus on the oxygen-specific effects within this dataset. The eight conditions are described as states. The activity of TFs is assessed for the different state transitions. Special attention is devoted to TFs that seem to perform a regulatory role under aerobic conditions, but not under anaerobic growth (or vice versa). The resulting regulatory network reveals nutrient-limitation-specific effects of oxygen presence on expression behavior and TF activity. The analysis identifies many TFs that seem to play a very specific and subtle regulatory role at the nutrient and oxygen availability transitions.

This chapter was published in Computational Methods in Systems Biology, 2006.

Chapter 6 presents a large chemostat microarray compendium consisting of 170 microarray measurements with 55 unique conditions. These conditions are characterized by the settings of ten different cultivation parameters. Using a regression strategy the influence of cultivation parameters on gene expression is investigated. Here, the main focus is on the influence of combinations of cultivation parameters on gene expression. The explained variance of gene expression patterns and functional enrichment of gene clusters is evaluated for regression models both including and excluding these combinatorial effects. Also, the influence of cultivation parameters on gene expression is used in the interpretation of shake-flask-based transcriptome studies and for guiding functional analysis of (uncharacterized) genes and pathways. This study demonstrates that modeling the combinatorial effects of environmental parameters on the transcriptome is crucial for understanding transcriptional regulation. In this way, the goal of systems biology to investigate and understand the interactions between different components and/or levels in biological systems can be complemented by an equally integrative approach towards

the complex environmental context in which cells grow and survive.

This chapter was published in BMC Genomics, 2009.

In Chapter 7 the regression results from Chapter 6 are used to construct regulatory transcription networks. Here, TF binding data is employed to 'explain' the influence of cultivation parameters on gene expression. The method described in this chapter aims to estimate under which cultivation parameters a TF becomes active as an enhancer or a repressor to (co-)regulate the expression of a gene. The interplay between activated enhancers and repressors that bind a gene promoter determine the possible up- or downregulation of the gene. The model is translated into a linear integer optimization problem and solved accordingly. This study is the first to demonstrate how environmental parameters can be employed to derive transcriptional regulation networks.

This chapter was published in Bioinformatics, 2008.

Chapter 8 presents an alternative to the hypergeometric test procedure used to test gene groups for functional enrichment. The test described in this chapter is based on the central limit theorem. In contrast to the rest of the thesis, the method is applied to time series microarray data in order to create gene set activity profiles, which represent the enrichment of a gene set over time. Since for each gene set a unique activity profile can be derived, differences in the activity of e.g. biological processes or transcription factors in terms of the degree of enrichment and timing can be analyzed, thereby offering profound insight in (the hierarchy of) regulatory mechanisms.

This chapter was published in International Journal of Bioinformatics Research and Applications, 2008.

# CHAPTER 2

# TRANSCRIPTIONAL RESPONSE TO ORGANIC ACID STRESS

In this chapter the transcriptional response of *S. cerevisiae* to four different weak organic acids is studied. Here, an unstressed condition (no addition of organic acids) is used as a reference condition. A discretization procedure is designed, such that each gene is represented by a tertiary (-1, 0, 1) vector of length four, indicating up- or downregulation of a gene when yeast is exposed to each of the four acids with respect to the unstressed condition. Based on these discrete representations genes are grouped into clusters. The clusters, which are readily interpretable with respect to the four organic acids, are consulted for enrichment of functional categories and TF binding. This study reveals that *S. cerevisiae* exhibits a minimal generic transcriptional response to weak organic acids. The consequences of these findings are that the often-used term 'weak-organic acid stress' should preferably be avoided and that the use of individual organic acids as 'model compounds' for general responses to organic acids should be treated with caution.

This chapter is published as:

'Generic and specific transcriptional responses to different weak organic acids in anaerobic chemostat cultures of *Saccharomyces cerevisiae*'

Derek A. Abbott, Theo A. Knijnenburg, Linda M.I. de Poorter, Marcel J.T. Reinders, Jack T. Pronk and Antonius J.A. van Maris

FEMS Yeast Research, Volume 7 Issue 6 p. 819-833, September 2007

Note: TAK's contribution to this chapter is limited to the computational analysis of the microarray data.

## 2.1 Abstract

Transcriptional responses to four weak organic acids (benzoate, sorbate, acetate and propionate) were investigated in anaerobic, glucose-limited chemostat cultures of *Saccharomyces cerevisiae*. To enable quantitative comparison of the responses to the acids, their concentrations were chosen such that they caused a 50% decrease of the biomass yield on glucose. The concentration of each acid required to achieve this yield was negatively correlated with membrane affinity. Microarray analysis revealed that each acid caused hundreds of transcripts to change by over 2-fold relative to reference cultures without added organic acids. However, only 14 genes were consistently upregulated in response to all acids. The moderately strongly lipophilic compounds benzoate and sorbate and, to a lesser extent, the less lipophilic acids acetate and propionate, showed overlapping transcriptional responses. Statistical analysis for overrepresented functional categories and upstream regulatory elements indicated that responses to the strongly lipophilic acids were focused on genes related to the cell wall, while acetate and propionate had a stronger impact on membrane-associated transport processes. The fact that *S. cerevisiae* exhibits a minimal generic transcriptional response to weak organic acids along with extensive specific responses is relevant for interpreting and controlling weak acid toxicity in food products and in industrial fermentation processes.

## 2.2 Introduction

Short-chain weak organic acids are potent inhibitors of microbial growth that are widely applied as preservatives in food and beverages. At low extracellular pH, weak acids occur predominantly in the undissociated form, which has relatively high membrane permeability. After entry into the cell via passive diffusion, the higher pH of the cytosol causes dissociation of the acid, thus acidifying the cell and triggering the ATP-dependent efflux of protons [Pamp 89]. Consequently, weak acids can cause, at the very least, a transient reduction of intracellular ATP levels [Holy 96]. At high concentrations, ATP exhaustion, acidification of the cytoplasm and dissipation of the proton-motive force may occur [Imai 95]. This 'weak-acid uncoupling' mechanism is customarily cited as the major mechanism underlying weak organic acid toxicity [Kreb 83, Russ 91, Salm 84]. In addition, the anion of the weak acid, which is much less membrane permeable than the undissociated acid, accumulates intracellularly, where it may reach toxic concentrations [Pamp 90, Russ 92]. Membrane disruption [Holy 99, Kreb 83] and enzyme inhibition have been proposed as possible mechanisms of anion toxicity. Furthermore, benzoate and acetate have been implicated in inhibition of autophagy and induction of apoptosis, respectively [Haza 04, Ludo 01].

Despite the negative impact of weak organic acids on growth, many spoilage organisms, including yeasts, can adapt and proliferate at the maximum legal dosage of preservatives. As such, considerable economic loss is imparted in combination with consumer concern [Thom 93, Tudo 93]). A better understanding of the underlying molecular and regulatory responses is crucial for the development of preservation strategies to prevent microbial-mediated food spoilage.

Short-chain organic acids also occur as inhibitory compounds in industrial fermentation processes. One important example is the detrimental effect of acetic acid and other weak acids on the production of bio-ethanol with the yeast *Saccharomyces cerevisiae*

[Nare 01]. The presence of these naturally occurring metabolites results in substantial economic losses just as observed in the food industry. However, in contrast to applications in food preservation, a greater understanding of weak organic acid toxicity would serve to increase the robustness of bio-catalysts under process conditions.
The adaptive response to weak organic acids has been extensively studied in *S. cerevisiae*. For example, activity of the plasma membrane $H^+$-ATP-ase, Pma1, has been shown to be modulated in the presence of weak acids [Holy 96]. It has also been shown that many genes upregulated in cells exposed to organic acids are regulated by Msn2/Msn4 of the general stress response pathway [Schu 04]. Furthermore, Pdr12 which is regulated by War1 [Kren 03] and facilitates ATP-dependent efflux of moderately lipophilic short-chain acid anions, has been identified as a key determinant in organic acid tolerance [Pipe 98]. More recently, additional subsets of genes, which appear to be independent of Msn2/4 and War1, have been discovered. Schueller *et al.* [Schu 04] identified a group of 21 genes, including *HSP30*, that were regulated independently of War1 and Msn2/4 in response to sorbate. In addition, Haa1 has been shown to regulate the expression of a small set of genes that, upon their deletion, confer hypersensitivity to acetic, propionic and butyric acid, but not to the more lipophilic compounds, benzoic and octanoic acid [Fern 05].
Although the currently available literature suggests a relation between lipid solubility of weak organic acids and the physiological responses of *S. cerevisiae*, a quantitative comparison of the physiological and transcriptional responses to different weak organic acids has not been performed. Indeed, it is unclear whether a 'generic' transcriptional response to weak organic acids exists in this important industrial microorganism.
The aim of the present study was to quantify and compare unity and diversity in the physiological and transcriptional responses of *S. cerevisiae* to four organic acids: benzoate and sorbate, two moderately lipophilic weak acids, and acetate and propionate, two acids that are much less lipophilic. Anaerobic, glucose-limited chemostat cultures were utilized to quantitatively compare the physiological effects and transcriptional regulation induced by these four acids. This experimental setup has a number of benefits: (i) Chemostat cultures, in contrast to batch or shake-flask cultivation, offer the possibility to study the effect of constant and defined environmental stimuli (concentrations, pH, etc.) at a fixed specific growth rate. (ii) In contrast to shake-flasks, chemostat cultures allow for control of the pH which is crucial in studies on weak acids. (iii) Anaerobic conditions eliminate consumption of these organic acids and are relevant to many industrial applications where organic acids are present. (iv) Although irrelevant for the anaerobic conditions of choice, experiments in shake flask, an often used experimental system in weak-acid studies, often progress to oxygen limitation. Besides the choice for anaerobic chemostat cultivation, the comparison was further facilitated by choosing the concentration of each acid such that equivalent biomass yields on glucose were obtained. While this study does not strive for an exhaustive comparison between each individual acid, the data generated during this study has been made publicly available to facilitate such studies.

**Table 2.1 – Properties of weak organic acids used in this study.**
The concentrations required to reduce the biomass yield to 50% of the reference condition ($YRC_{50}$) and the predicted concentration of undissociated acid at pH 5.0 (based on the Henderson-Hasselbach equation) are indicated along with the most commonly cited $pK_a$ and partition coefficient.

| | Formula | $pK_a$ | Octanol Water Partition Coefficient (logP) | $YRC_{50}$ | Concentration Undissociated |
|---|---|---|---|---|---|
| Acetic | $CH_3COOH$ | 4.75 | -0.31 | 105.0 mM | 37.7 mM |
| Propionic | $CH_3CH_2COOH$ | 4.88 | 0.33 | 20.0 mM | 8.6 mM |
| Sorbic | $CH_3CH{=}CHCH{=}CHCOOH$ | 4.76 | 1.33 | 1.3 mM | 0.47 mM |
| Benzoic | $C_6H_5COOH$ | 4.19 | 1.87 | 2.0 mM | 0.27 mM |

## 2.3 Materials and methods

### 2.3.1 Yeast strain and growth conditions

The laboratory reference strain CEN.PK 113-7D (MATa) was grown at 30°C in 2-L chemostat fermentors (Applikon, Schiedam, The Netherlands) with a working volume of 1 L using an electronic level sensor to maintain a constant volume. All cultures, including the reference, were fed with minimal medium as described by Verduyn *et al.* [Verd 92] with 25 g $L^{-1}$ glucose as the limiting nutrient and 0.15 ml $L^{-1}$ silicone antifoam (BDH, Poole, England) to prevent excessive foaming. The dilution rate was set to 0.10 $h^{-1}$ and the pH was controlled at 5.0 with the automatic addition (ADI 1031 bio controller, Applikon) of 2 M KOH. The stirrer speed was set at 800 RPM and anaerobicity was maintained by sparging the fermentor with $N_2$ gas at 500 ml $min^{-1}$. To prevent diffusion of oxygen, the fermentor was equipped with Norprene tubing and Viton O-rings and the medium vessel was also flushed with $N_2$ gas. A comparable degree of weak acid stress was ensured by decreasing the biomass yield to approximately 50% of the reference condition (no organic acids added) with the addition of the appropriate concentration of acetic acid, sodium benzoate, propionic acid or potassium sorbate to the reservoir media (Table 2.1).

### 2.3.2 Analytical methods

Chemostat cultures were assumed to be in steady state when, after at least five volume changes, the culture dry weight and specific carbon dioxide production rate changed by less than 2% over 2 volume changes. Steady state samples were taken between 10 and 14 volume changes after inoculation to avoid possible evolutionary adaptation during long-term cultivation. Culture dry-weights were determined in duplicate via filtration onto dry, pre-weighed nitrocellulose membranes. Samples were dried in a microwave oven for 20 minutes at 360 W. Culture supernatants were obtained after centrifugation of chemostat broth or by a rapid sampling method using pre-cooled (-20°C) steel beads [Mash 03]. For the purpose of flux determination and carbon recovery, supernatants and media were analyzed via HPLC using an AMINEX HPX-87H ion exchange column with 5 mM $H_2SO_4$ as the mobile phase. Off-gas was first cooled with a condenser (2°C) and then dried with a Perma Pure dryer (PD-625-12P). $CO_2$ and $O_2$ concentrations in the off-gas were measured with an NGA 2000 Rosemount gas analyzer.

### 2.3.3 Microarray analysis

Sampling of chemostat cultures, probe preparation and hybridization to Affymetrix GeneChip microarrays was performed as described previously [Pipe 02], but with the following modifications. Double-stranded cDNA synthesis was carried out using 15 $\mu$g of total RNA and the components of the One Cycle cDNA Synthesis Kit (Affymetrix). The double-stranded cDNA was purified (Genechip Sample Cleanup Module, Qiagen) before *in vitro* transcription and labeling (GeneChip IVT Labeling Kit, Affymetrix). Finally, labeled cRNA was purified (GeneChip Sample Cleanup Module) prior to fragmentation and hybridization of 15 $\mu$g of biotinylated cRNA.
Data acquisition, quantification of array images and data filtering were performed with the Affymetrix software packages Microarray Suite v5.0, MicroDB v3.0 and Data Mining Tool v3.0. All arrays were scaled to a target value of 150 using the average signal from all genes. Expression values below 12 are considered insignificant variations in unexpressed genes and were consequently set to 12 as previously described [Pipe 02]. To enable further study of this data by other researchers in the field of organic acid toxicity/tolerance the data of the Affymetrix GeneChip microarrays used in this study are available via Gene Expression omnibus series accession number GSE5926.

### 2.3.4 Differential expression

To assess which genes exhibit statistically significant up- or downregulation as a consequence of the different organic acid challenges, pairwise tests between each condition and the reference situation were performed. Thus, the gene expression levels as measured in the presence of each of the four organic acids were compared with the expression levels of the reference anaerobic cultures. For this, we employed the framework of Significance Analysis of Microarrays [Tush 01]. In an effort to reduce biological noise, a gene was called differentially expressed only if there was at least a two-fold difference in average expression and its Q-value was lower than the stringent median false discovery rate (FDR) of 0.5%.
As a result, each gene was represented by a discretized expression pattern of length four, indicating whether the gene, was not differentially expressed (0), upregulated (1) or downregulated (-1) under each of the four test conditions. For example, a gene that had the following discretized expression pattern

| A | B | P | S |
|---|---|---|---|
| 1 | 0 | -1 | 0 |

was upregulated due to acetate exposure (A) and downregulated in response to propionate (P), while the two other conditions, benzoate (B) and sorbate (S), did not significantly change the expression of this gene compared to the reference situation. This discretized representation of the expression behavior of a gene was used for further analysis. Although information density is reduced when going from the continuous expression levels to the discretized representation, much interpretability is gained when analyzing the outcomes of the stringent statistical tests [Knij 07]. Furthermore, the discretized representation allows for a simple and meaningful way to cluster genes into functionally coherent groups.

### 2.3.5 Clustering genes

Gene clusters were created in three different ways to identify groups of genes that exhibited both overall and specific response to the different acids:

- Genes that had identical discretized expression patterns form clusters. Thus, the overall response of the genes in a cluster to the four organic acids is identical.
- Additionally, genes were grouped into clusters, when they were up- or downregulated in response to one specific organic acid, regardless of their expression behavior under the other three test conditions. This led to four clusters of genes that exhibited upregulation under one of the conditions, and similarly, four clusters of genes that were downregulated due to one particular stimulus.
- To investigate more thoroughly the acetate-propionate relationship and the benzoate-sorbate relationship, genes were clustered when they were either upregulated or downregulated under both acetate and propionate exposure, and similar for the benzoate and sorbate conditions.

### 2.3.6 Hypergeometric tests

The (overlapping) gene clusters were consulted for enrichment in functional annotation (Munich Information Center for Protein Sequences, MIPS [Mewe 97]) and significant transcription factor (TF) binding. To test for significant relations the hypergeometric test was employed. In the case of the TF binding data, the largest available TF binding dataset for yeast in its most conservative setting (highest binding confidence) was used [Harb 04]. This dataset, which originally indicates the number of binding sites for each of 102 TFs in the promoter region of each gene, was binarized, such that the data indicates whether a TF can bind a gene (upstream) or not. Then, the hypergeometric test assesses if a TF (or a TF pair) can bind the promoter region of the genes in a cluster much more frequently than in a random set of genes. In case of the employed gene annotation information it assesses if the number of genes in a cluster that belongs to a particular functional category within the MIPS database is much larger than would be expected by chance. The $P$-value cutoff to decide whether a relation is significant is $P \leq 1/(n_c n_x)$, where $n_c$ is the number of clusters and $n_x$ is the number of TFs (or TF pairs) or the number of MIPS annotation categories. Consequently, $P$-value cutoffs were different for assigning significance to functional categorization, TF binding and binding of TF pairs. This adjustment for multiple testing, corresponds with a per comparison error rate (PCER) of one [Ge 03], resulting in $P$-value cutoffs around $10^{-5}$.

## 2.4 Results

### 2.4.1 Effects of different organic acids on biomass yields in anaerobic chemostat cultures

Prior to performing steady-state chemostat cultures, trial runs were performed in which the concentration of each acid in the medium reservoir was titrated to reduce the biomass yield to 50% of the reference condition. For acetate, propionate, benzoate and sorbate, different concentrations were required to achieve this reduction of the biomass yield,

even when the concentration of undissociated acid in the cultures was calculated from their respective $pK_a$ values (Table 2.1). A strong correlation was observed between the amount of acid required to reduce the biomass yield on glucose by 50% and their octanol/water partitioning coefficient, consistent with the notion that membrane permeability of the undissociated acid is a key factor in weak-acid toxicity.
Using the concentrations of the acids deduced from the trial runs, triplicate anaerobic, glucose-limited chemostat fermentations were performed for each organic acid and compared to triplicate glucose-limited reference cultures without organic acids. The reduced biomass yield of the cultures grown with added organic acids was mirrored by an approximately two-fold increase in the specific rates of ethanol and carbon dioxide production. In addition, the low but significant rates of lactate production observed in the reference cultures were approximately doubled in the cultures to which organic acids had been added. In *S. cerevisiae*, D-lactate is formed via the methylglyoxal bypass. Activity of this bypass of glycolysis has been shown to be correlated with glycolytic flux [Mart 01], probably via the intracellular concentrations of dihydroxyacetone phosphate, the immediate precursor of methylglyoxal formation.
In anaerobic cultures of *S. cerevisiae*, glycerol formation serves as a redox sink for re-oxidation of excess NADH that is formed in biosynthetic reactions [Dijk 86]. Biomass-specific rates of glycerol formation were the same in all cultures, except for those with acetate addition, which showed a markedly reduced specific rate of glycerol production. In anaerobic, glucose-limited cultures, acetate can be converted to acetyl-coenzyme A by the acetyl-CoA synthetase Acs2 [Berg 96]. Formation of this key precursor for the synthesis of amino acids and lipids from glucose is an oxidative process that yields NADH. The reduced production of glycerol by the acetate cultures probably reflects a previously proposed NADH-sparing effect of acetate cometabolism [Tahe 96]. Since, under anaerobic conditions, dissimilation of acetate does not occur, only a small fraction of the acetate added to the reservoir media was consumed.
The residual concentrations of glucose in cultures grown with organic acids were higher than in the reference cultures. In micro-organisms, the specific rate of consumption of the growth-limiting substrate $q_s$ often exhibits saturation kinetics with respect to its concentration $C_s$. These kinetics can be described by the modified Monod equation ($q_s = q_s^{max} \frac{C_s}{C_s + Ks}$). Thus, the increased rate of glucose consumption by the cultures may be at least partially responsible for the increased residual glucose concentration. However, despite the essentially identical rates of glucose consumption that were observed in the cultures to which organic acids had been added, the residual glucose concentrations were different for the four acids (Table 2.2). This suggests the involvement of acid-specific effects on the expression and/or activity of genes/proteins involved in glucose consumption. In fact, uptake of $^{14}$C-labeled glucose has been shown to decrease in response to benzoate and lactate challenges [Thom 06]. Since the residual glucose concentrations remained well below 5 mM in all cultures, no substantial impact of glucose repression on gene expression was anticipated [Walk 98]. Moreover, with possibly the exception of genes involved in fatty acid oxidation, significant transcriptional changes in glucose repressible gene expression were not observed in the current study.

### 2.4.2 Transcriptome analysis: data quality and overall responses

The physiological analysis presented in the previous paragraph suggested that, although the dose-response relationships differed, the physiological effects on *S. cerevisiae* were

**Table 2.2 – Physiology of *S. cerevisiae* grown in the presence of weak organic acids.**
The acids were added to C-limited, anaerobic chemostat cultures at various concentrations to reduce the biomass yield to approximately 50% of the reference condition. The corresponding steady state fluxes ($q$: mmol/g/h) from triplicate chemostats performed at pH 5.0 at a dilution rate of 0.10 $h^{-1}$ are indicated along with the standard deviation.
[a] The concentrations of benzoic, propionic and sorbic acid were determined by HPLC to be equal in the feed medium and culture supernatant (data not shown). Consequently, fluxes for these compounds were not included and they were not used in the calculation of C recovery.

| | Reference (No Stress) | 105 mM Acetic acid | 2 mM Benzoic acid[a] | 20 mM Propionic acid[a] | 1.3 mM Sorbic acid[a] |
|---|---|---|---|---|---|
| $q$ Glucose | -6.03 ± 0.10 | -12.17 ± 0.20 | -12.12 ± 0.58 | -12.98 ± 0.48 | -12.08 ± 0.20 |
| $q$ $CO_2$ | 10.40 ± 0.45 | 22.97 ± 0.50 | 22.84 ± 1.45 | 23.73 ± 0.94 | 21.12 ± 0.28 |
| $q$ Ethanol | 9.52 ± 0.16 | 21.45 ± 0.35 | 21.19 ± 1.06 | 21.41 ± 1.32 | 21.40 ± 0.47 |
| $q$ Glycerol | 0.79 ± 0.02 | 0.54 ± 0.01 | 0.96 ± 0.06 | 1.00 ± 0.03 | 0.83 ± 0.01 |
| $q$ Lactate | 0.05 ± 0.01 | 0.09 ± 0.00 | 0.10 ± 0.00 | 0.11 ± 0.01 | 0.09 ± 0.01 |
| $q$ Acetate | 0.02 ± 0.00 | -0.57 ± 0.02 | 0.08 ± 0.01 | 0.03 ± 0.01 | 0.02 ± 0.01 |
| Biomass (g/L) | 2.25 ± 0.02 | 1.13 ± 0.02 | 1.17 ± 0.03 | 1.06 ± 0.03 | 1.22 ± 0.02 |
| Yield ($g_x/g_s$) | 0.09 ± 0.00 | 0.05 ± 0.01 | 0.05 ± 0.01 | 0.04 ± 0.01 | 0.05 ± 0.00 |
| C Recovery (%) | 99.4 ± 0.8 | 95.1 ± 0.6 | 95.6 ± 1.0 | 96.5 ± 1.3 | 93.4 ± 1.0 |
| Residual Glucose (mM) | 0.2 ± 0.0 | 2.1 ± 0.1 | 1.7 ± 0.8 | 3.4 ± 0.3 | 0.7 ± 0.3 |

similar for the four organic acids. To investigate whether this also held for the transcriptional responses to the four acids, the chemostat cultures were subjected to a full transcriptome analysis.
To obtain statistically reliable transcriptome data, triplicate chemostat cultivations and microarray analyses were carried out for each condition. The average coefficient of variation for triplicate arrays in each condition was below 18%, which is indicative of reliable, reproducible analyses [Pipe 02]. A fold-change threshold of 2, combined with a false-discovery rate of 0.5% was used to assess significance of changes in transcript levels. Using these criteria, a comparison of the acid-exposed cultures to the reference condition yielded 4059 genes that did not exhibit a significantly changed transcript level (Figure 2.1). An additional 902 genes were not transcribed (average expression $< 12$) under any of the conditions tested. This left 1422 genes (22% of the genome) whose transcript levels were significantly modulated in response to at least one weak acid. Transcripts with identical discretized patterns (see Methods section) were grouped together prior to further analysis. For example, transcripts downregulated by all acids were represented by a discretized pattern [-1 -1 -1 -1]. The 1422 transcripts whose level was modulated in response to weak acids yielded 45 distinct discretized patterns, 25 of which contained 10 or more genes.

### 2.4.3 Identification of a minimal generic transcriptional response to weak organic acids

Only 14 genes were identified whose transcript levels were significantly upregulated in response to all four acids [1 1 1 1]: *CWP2*, *PIR1*, *CCR4*, *PAN1*, *TIM44*, *IMP2*, *RRD1*, *YHR087W*, *SOD2*, *WSC4*, *SPI1*, *RNQ1*, *YGP1* and *SML1*. While five of these genes (*SPI1*, *CWP2*, *PIR1*, *YGP1* and *WSC4*) are related to cell wall structure and organization, no statistically significant overrepresentation of MIPS functional categories or

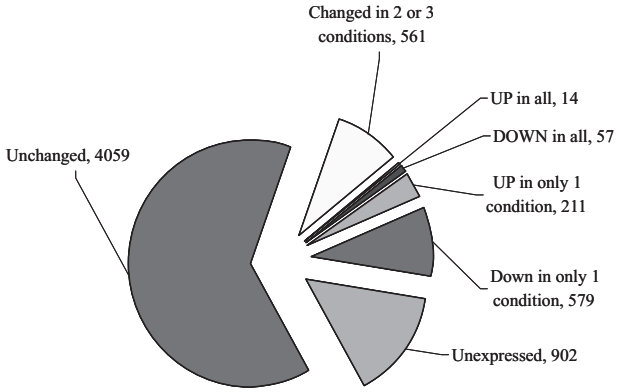



**Figure 2.1** – The global transcriptional response of *S. cerevisiae* to anaerobic chemostat growth in the presence of weak organic acids. Significantly changed transcripts were identified for each acid following Affymetrix transcriptional profiling and data analysis using a FDR of 0.5% with a fold-change of 2. The data clearly indicates a large acid-specific response in combination with a very small response shared by cultures grown in the presence of acetate, benzoate, propionate or sorbate.

promoter elements (based on the data of [Harb 04]) was identified. See Materials and Methods. Although the MIPS category of cell wall structure and organization was highly enriched ($P = 8.9{\cdot}10^{-4}$), this is not deemed significant with the stringent multiple testing correction. Interestingly, *SOD2*, which encodes the Mn-containing mitochondrial superoxide dismutase and *TIM44* which is involved in mitochondrial protein import [Geis 00], and possibly in removal of mitochondrial superoxide [Mats 05], were both upregulated in all conditions. Finally, a number of genes involved in DNA synthesis (*SML1*) and repair (*RRD1*, *IMP2*) were present within this group, suggesting that organic acids are capable of inducing DNA damage.

Similar to the common upregulated gene set, a set of 57 genes that showed consistent transcriptional downregulation for the four organic acids also failed to reveal overrepresented promoter elements. However, a hypergeometric distribution analysis of functional categories indicated a significant overrepresentation of genes involved in fatty acid oxidation (*ECI1*, *POT1*, *SPS19* and *YGR207C*, $P = 6.97 \cdot 10^{-7}$). As described above for *SOD2* and *TIM44*, the oxygen-dependency of lipid oxidation, combined with the anaerobic cultivation conditions, makes the physiological significance of this transcriptional response difficult to explain. Since genes involved in fatty acid oxidation are very sensitive to glucose repression [Ganc 98, Veen 87], the downregulation of this set may instead reflect the slightly elevated residual glucose concentrations in the acid-exposed cultures. However, raw expression values did not indicate any correlation between residual glucose concentrations and transcript levels.

A number of genes involved in various transport processes at the plasma membrane level were also among the common downregulated genes. In particular, five of the genes, *TAT1*, *MMP1*, *DIP5*, *AQR1* and *MEP3* are involved in transport of amino acids and ammonium. *TAT1*, *MMP1* and *DIP5* function in uptake of amino acids and *MEP3* encodes an ammonium permease while *AQR1* has been implicated in amino acid excretion [Vela 04]. In addition, transport of zinc (*ZRT1*), copper (*CTR3*) and even sterols (*AUS1*) is downregulated.

Rather unexpectedly, only a small number of genes pertaining to the translational ma-

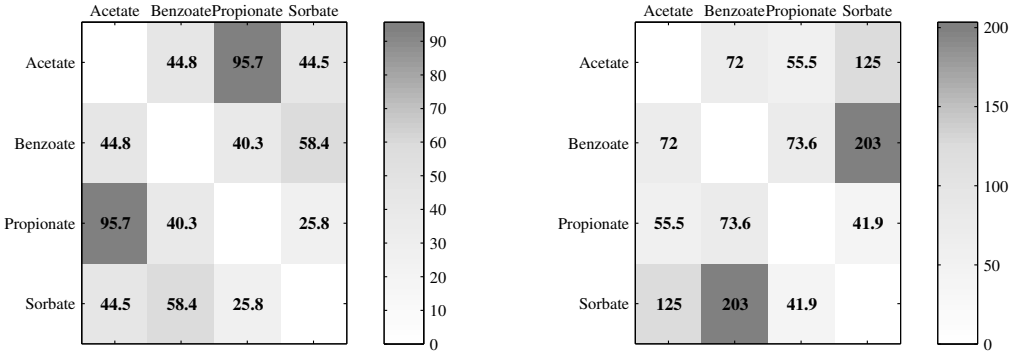



**Figure 2.2 – The significance of the correlation (-log *P*-value) between conditions was predicted using a hypergeometric distribution.**
Analysis of the upregulated genes (left panel) showed that the overlap between acetate and propionate is of most significance, followed by the overlap between benzoate and sorbate. With respect to the downregulated genes (right panel), the correlation between acetate and propionate is much lower, while the degree of benzoate/sorbate co-regulation is highlighted by an extremely low P-value.

chinery were identified in the common downregulated gene set. Of the 57 genes sharing common downregulation, only seven (*MIP6*, *MRPL25*, *RPR1*, *SSF2*, *NIP7*, *SPP2*, *LSM7*) were related to ribosome biogenesis or RNA processing.

### 2.4.4 Unique responses and co-responses to different organic acids: correlation with lipid solubility

The limited generic transcriptional response to the four organic acids, along with the lack of a regulon defined by a common known promoter element or upregulation of genes belonging to one or a few functional categories, is intriguing. Based on the discretized expression profiles (>2-fold change at 0.5% FDR) 561 genes showed a significant transcriptional response to two or three of the four acids. Hypergeometric distribution analysis was applied to statistically evaluate co-responses to all possible combinations of two acids. This approach established a clear correlation between the occurrence of common transcriptional responses to individual organic acids and their membrane solubility (octanol/water partition coefficients, Table 2.1). In addition to the strongly overlapping transcriptional responses to benzoate and sorbate, a highly significant overlap of the transcriptional responses to acetate and propionate was also identified (Figure 2.2).
Despite the overlap of the transcriptional responses to subsets of the four organic acids, many transcripts were uniquely regulated in response to the four acids used in this study. In total, 211 genes were identified as being uniquely upregulated by a single acid, while 579 genes indicated downregulation by a single acid. For each of the four acids, the unique transcriptional response was comprised of over 100 genes. Propionate had the largest (395 transcripts) set of uniquely regulated genes, which corresponds to approximately half of propionate-responsive genes. In contrast, only 104 genes were specifically modulated in the presence of acetate, corresponding to slightly less than 25% of the overall response to acetate. Similarly, the specific changes in sorbate and benzoate exposed cultures also comprised approximately 25% of the complete response (Table 2.3).

**Table 2.3 – The overall transcriptional response to weak organic acids.**
Significantly changed transcripts were identified for each acid following Affymetrix transcriptional profiling and data analysis using a FDR of 0.5% with a fold-change of 2. The complete response to each acid is represented by the total and transcriptional responses, which were specifically regulated under only one condition are denoted as unique.

| | Up | | | Down | | |
|---|---|---|---|---|---|---|
| | Total | Unique | % Unique | Total | Unique | % Unique |
| Acetic | 168 | 42 | 25 | 291 | 62 | 21.3 |
| Benzoic | 103 | 22 | 21.4 | 439 | 106 | 24.1 |
| Propionic | 252 | 114 | 45.2 | 528 | 281 | 53.2 |
| Sorbic | 118 | 33 | 28 | 480 | 130 | 27.1 |

Further evidence for acid-specific responses was found in the overrepresentation of functional categories (Table 2.4) and transcription-factor binding sites (Table 2.5) among the genes that responded to the four acids.

### 2.4.5 Benzoate and sorbate responsive transcripts

The genes that were transcriptionally upregulated in response to benzoate showed a significant enrichment for the MIPS functional category "cell rescue, defense and virulence" and, more specifically, for the subcategories "stress response" and "cell wall" (Table 2.4). The sorbate-upregulated genes were enriched for the MIPS functional category "interaction with the cellular environment" (subcategories "cellular sensing and response" and "chemoperception and response").
Analysis of overrepresented transcription factor binding among the benzoate-upregulated transcripts (Table 2.5) corroborated a cell wall-related response to benzoate. The MAPK cascade transcription factors Dig1 and Ste12 are directly linked to pseudohyphal growth and cell wall processes. Moreover, Msn2 and Msn4 regulatory sites were also identified along with the enrichment of Skn7 and Swi5 binding. Skn7 has been implicated in the control of cell wall biosynthesis, cell cycle, and the osmotic stress response in addition to its role in oxidative stress [Lee 99]. Overexpression of *SKN7* suppresses the cell wall assembly mutation *kre9* [Brow 93] and the growth defect associated with deletion of a regulator involved in cell surface assembly [Brow 94].
The sorbate upregulated genes showed an overrepresentation of many transcription factors (Table 2.5). Although no overrepresentation of cell wall-related functional categories was observed for sorbate, several of the enriched transcription factors and transcription factor pairs were cell wall-related. Sok2 has been implicated in cell wall stress [Lago 03], while Ste12 and Tec1 of the MAPK cascade are associated with the regulation of cell wall integrity [Qi 05] and pseudohyphal growth [Ganc 01]. Dig1 (Rst1) acts as a regulator of Ste12 [Tedf 97]. Previous studies by Mollapour *et al.* [Moll 04] and de Nobel *et al.* [Nobe 01] corroborate the importance of cell wall proteins in the response of *S. cerevisiae* to sorbate and benzoate, although the identity of the genes found in their studies is different. However, discrepancies in individual genes may be reflective of the difference between transient adaptive changes in gene expression and steady state responses.
For the sake of brevity, we will not discuss the overrepresented functional categories and transcription factor binding (Tables 2.4 and 2.5) for the downregulated genes in the benzoate and sorbate comparison. This, however, does not mean that the downregulation

**Table 2.4 – Overview of MIPS functional categories identified in the presence of organic acids in upregulated (shades of red) and downregulated (shades of green) gene sets.**
Overrepresentation is indicated by the dark red (upregulated) and dark green (downregulated) boxes. The similarities between responses are also indicated for acetate/propionate and benzoate/sorbate. The significance of each category is numerically indicated as -log10 *P*-value.
O (overall) columns indicate the analysis of all genes responding to each condition regardless of their expression in other conditions. S (specific) columns indicate the analysis of genes which solely respond to the indicated condition(s).

| Functional Category | Acetate | | Propionate | | Acetate/ Propionate | | Benzoate | | Sorbate | | Benzoate/ Sorbate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | O | S | O | S | O | S | O | S | O | S | O | S |
| CELL RESCUE, DEFENSE AND VIRULENCE | 8.11 | 1.5 | 6.63 | 1.03 | 9.4 | 3.5 | 4.51 | 0.25 | 2.77 | 1.28 | 1.78 | 0.28 |
| stress response | 7.13 | 0.38 | 4.82 | 0.6 | 7.15 | 1.17 | 5.94 | 0.35 | 3.01 | 0.73 | 2.34 | 0.23 |
| osmotic and salt stress response | 3.86 | 0.36 | 2.09 | 0.31 | 3.63 | 1.07 | 2.6 | 0.2 | 1.64 | 0.15 | 1.14 | 0.91 |
| cell wall | 0.59 | 1.27 | 1.41 | 0.14 | 1.01 | 0.15 | 3.87 | 0.28 | 2.26 | 1.03 | 2.33 | 0.57 |
| METABOLISM | 2.61 | 0.82 | 5.87 | 4.97 | 0.86 | 0.34 | 1.22 | 0.11 | 1.85 | 0.46 | 1.76 | 0.79 |
| amino acid metabolism | 2.24 | 2.1 | 5.72 | 5.38 | 0.34 | 0 | 1.42 | 0.25 | 1.07 | 0.68 | 0.99 | 0.38 |
| assimilation of ammonia, metabolism of the glutamate group | 1.25 | 1.14 | 5.05 | 3.83 | 0.24 | 0 | 0.77 | 0.27 | 0.34 | 0 | 0.54 | 0.76 |
| metabolism of arginine | 0.64 | 0 | 7.63 | 6.05 | 0 | 0 | 1.35 | 0.55 | 0.73 | 0 | 0.71 | 1.38 |
| biosynthesis of arginine | 0 | 0 | 5.31 | 5.72 | 0 | 0 | 1.26 | 0.71 | 0.19 | 0 | 0.38 | 0.71 |
| metabolism of urea (urea cycle) | 1.56 | 0 | 4.48 | 0.99 | 0 | 0 | 1.03 | 0 | 1.16 | 0.93 | 0 | 0 |
| metabolism of the pyruvate family and D-alanine | 0.23 | 0 | 4.2 | 3.52 | 0 | 0 | 0.57 | 0 | 0.52 | 0 | 0.87 | 0 |
| oxidation of fatty acids | 3.35 | 0 | 2.38 | 0 | 4.87 | 0 | 2.68 | 0 | 3.64 | 0.77 | 3.54 | 0 |
| secondary metabolism | 1.35 | 2.25 | 4.56 | 3.49 | 1 | 0.9 | 0.4 | 0 | 0.49 | 0 | 0.45 | 0.46 |
| INTERACTION WITH THE CELLULAR ENVIRONMENT | 1.33 | 0.85 | 1.02 | 0.74 | 0.89 | 0.61 | 1.91 | 0.1 | 3.75 | 3.49 | 1.32 | 1.15 |
| cellular sensing and response | 0.52 | 0.54 | 1.23 | 1.16 | 0.57 | 0.32 | 2.75 | 0.2 | 4.87 | 4.17 | 1.75 | 1.67 |
| chemoperception and response | 0.41 | 0.6 | 1.37 | 1.39 | 0.44 | 0.11 | 2.02 | 0.03 | 4.05 | 4.37 | 1.33 | 1.75 |
| CELLULAR TRANSPORT, FACILITATION AND ROUTES | 3.95 | 1.26 | 0.53 | 0.26 | 1.78 | 0.47 | 1.3 | 0.65 | 2.22 | 0.64 | 0.88 | 1.46 |
| transported compounds (substrates) | 6.72 | 2.59 | 0.66 | 0.1 | 3.19 | 0.68 | 1.3 | 0.33 | 2.91 | 0.44 | 1.11 | 0.87 |
| anion transport (Cl, SO4, PO4, etc.) | 4.89 | 2.71 | 0.44 | 0 | 1.92 | 0 | 0.46 | 0 | 1.96 | 1.01 | 0.54 | 0 |
| amino acid transport | 5.64 | 1.19 | 0.85 | 0.75 | 2.95 | 0.52 | 0.47 | 0 | 3.61 | 0.23 | 1.02 | 0.29 |
| drug transport | 5.3 | 2.24 | 1.8 | 0.83 | 1.48 | 1.41 | 0.58 | 0 | 1.21 | 0.26 | 0.6 | 0.33 |
| RNA processing | 0.1 | 0.14 | 1.63 | 0.28 | 0.24 | 0.2 | 4.23 | 1.28 | 1.33 | 0.8 | 1.89 | 1.64 |
| rRNA processing | 0.02 | 0.09 | 2.32 | 0.71 | 0.09 | 0 | 4.25 | 2.21 | 1.63 | 0.91 | 1.42 | 0.86 |

**Table 2.5 – Overview of transcription factors (TFs) involved in the response to organic acids in upregulated (shades of red) and downregulated (shades of green) gene sets.**
Overrepresentation of binding sites for each TF is indicated by the dark red (upregulated) and dark green (downregulated) boxes. TF pairs which were significantly enriched in each condition are also listed within the table. The similarities between responses are also indicated for acetate/propionate and benzoate/sorbate. The significance of each category is numerically indicated as -log10 $P$-value.
O (overall) columns indicate the analysis of all genes responding to each condition regardless of their expression in other conditions. S (specific) columns indicate the analysis of genes which solely respond to the indicated condition(s).
Note: TF Cin5/Yap4 is not only enriched in the downregulated O Acetate cluster ($P = 10^{-5.69}$), but also enriched in the upregulated cluster ($P = 10^{-4.47}$). This is not visible in the table.

| Transcription Factor | Acetate | | Propionate | | Acetate/ Propionate | | Benzoate | | Sorbate | | Benzoate/ Sorbate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | O | S | O | S | O | S | O | S | O | S | O | S |
| Aft2 | 2.97 | 0.13 | 2.19 | 1.49 | 1.29 | 1 | 3.12 | 0.43 | 1.96 | 0.3 | 2.98 | 0.75 |
| Cad1/Yap2 | 3.32 | 3.63 | 2.87 | 1.72 | 1.08 | 0 | 1.03 | 0 | 0.93 | 0 | 0.67 | 0.38 |
| Cin5/Yap4 | 5.69 | 0.65 | 1.67 | 0.33 | 2.38 | 1.84 | 1.26 | 0.35 | 3.11 | 0.66 | 1.08 | 0 |
| Dig1 | 1.86 | 2.59 | 1.84 | 0.99 | 1.29 | 0.25 | 4.75 | 1.06 | 5.68 | 5.17 | 1.65 | 1.41 |
| Gcn4 | 2.56 | 2.62 | 8.23 | 9.95 | 0.15 | 0.32 | 0.77 | 0.21 | 1.29 | 0.71 | 1.04 | 0.68 |
| Gln3 | 3.05 | 1.22 | 1.55 | 1.62 | 1.04 | 0.39 | 1.3 | 0.71 | 1.38 | 1.09 | 1.1 | 0.35 |
| Hap1 | 1.92 | 0 | 3.93 | 1.39 | 3.13 | 3.5 | 0.16 | 0.92 | 0.41 | 0.5 | 0.02 | 0.15 |
| Hsf1 | 3.74 | 0 | 3.66 | 1.05 | 3.8 | 3.12 | 1.2 | 0.43 | 0.35 | 0.3 | 0.01 | 0.19 |
| Mac1 | 1.47 | 0 | 1.42 | 0 | 2.53 | 4.06 | 1.03 | 0 | 0.94 | 0 | 1.6 | 1.56 |
| Mcm1 | 0.69 | 0.35 | 3.07 | 2.77 | 0.83 | 0 | 2.52 | 0.57 | 1.62 | 0.43 | 1.54 | 0.11 |
| Msn2 | 3.56 | 0.33 | 1.87 | 0.49 | 2.55 | 0.51 | 3.23 | 0 | 2.92 | 0.42 | 2.87 | 0.87 |
| Msn4 | 5.82 | 0 | 4.26 | 0.58 | 4.75 | 1.11 | 6.68 | 0.75 | 2.45 | 0 | 3.11 | 0 |
| Nrg1 | 2.27 | 0.69 | 3.32 | 1.11 | 1.04 | 0.54 | 1.77 | 0.44 | 2.51 | 0.58 | 1.83 | 0.2 |
| Rcs1 | 2.14 | 0.36 | 5.26 | 1.73 | 2.75 | 1.47 | 1.16 | 0.38 | 1.94 | 0.28 | 0.86 | 1.28 |
| Skn7 | 2.59 | 1.08 | 2.48 | 1.22 | 0.68 | 0.43 | 3.72 | 1 | 2.64 | 0.25 | 2.99 | 1.35 |
| Sok2 | 2.34 | 0 | 1.91 | 0.38 | 2.19 | 0.86 | 2.06 | 0 | 3.22 | 1.2 | 1.67 | 0.14 |
| Stb4 | 0.67 | 0 | 3.59 | 4.94 | 0 | 0 | 0.87 | 0 | 0.81 | 0 | 1.18 | 0.86 |
| Ste12 | 2.38 | 2.17 | 2.34 | 1.89 | 0.86 | 0.13 | 6.16 | 0.73 | 12.47 | 7.98 | 4.35 | 4.04 |
| Swi4 | 1.59 | 0.54 | 0.84 | 0.96 | 0.3 | 0 | 2.72 | 0 | 6.16 | 1.73 | 3.81 | 1.63 |
| Swi5 | 1.35 | 0 | 4.56 | 2 | 1 | 0.46 | 5.62 | 1.58 | 1.32 | 0 | 2.66 | 0.15 |
| Tec1 | 0.82 | 0 | 1.17 | 0.87 | 0.96 | 0.53 | 2.51 | 0.71 | 4.8 | 4.85 | 1.1 | 0.89 |
| Yap1 | 4.17 | 2.88 | 1.34 | 1.23 | 0.21 | 0.61 | 0.59 | 0.7 | 1.11 | 0.19 | 0.5 | 0 |
| Aft2 - Msn4 | 1.88 | 0 | 1.54 | 0 | 1 | 0 | 5.67 | 1.62 | 0.91 | 0 | 1.29 | 0 |
| Cad1 - Yap1 | 4.52 | 4.99 | 1.71 | 1.54 | 0 | 0 | 0.18 | 0 | 0.61 | 0 | 0.33 | 0 |
| Dig1 - Hsf1 | 2.18 | 0 | 0 | 0 | 0 | 0 | 4.4 | 1.77 | 0 | 0 | 0 | 0 |
| Dig1 - Ste12 | 1.05 | 1.57 | 2.11 | 1.12 | 1.16 | 0.31 | 5.59 | 1.21 | 6.69 | 5.75 | 1.94 | 1.57 |
| Dig1 - Tec1 | 0.61 | 0 | 1.14 | 0.87 | 0.84 | 0.73 | 3.6 | 0.93 | 5.48 | 6.04 | 1.53 | 1.12 |
| Gln3 - Swi5 | 0.9 | 0 | 0.74 | 0 | 0 | 0 | 4.4 | 1.77 | 1.05 | 0 | 1.43 | 0 |
| Mcm1 - Swi5 | 3.76 | 0 | 4.94 | 0 | 4.49 | 0 | 6.5 | 0 | 2.49 | 0 | 3.26 | 0 |
| Mcm1 - Ste12 | 0.8 | 0.79 | 4.5 | 3.91 | 0.47 | 0 | 1.16 | 0 | 2.85 | 0.88 | 1.76 | 0 |
| Msn4 - Swi5 | 3.04 | 0 | 0.56 | 0 | 0.94 | 0 | 5.37 | 1.56 | 2.05 | 0 | 2.82 | 0 |
| Ste12 - Tec1 | 0.86 | 0 | 1.59 | 0.9 | 1.32 | 0.66 | 3.18 | 0.85 | 5.96 | 5.6 | 1.37 | 1.03 |
| Ste12 - Swi5 | 1.66 | 0 | 4.99 | 3.37 | 0.89 | 0 | 2.06 | 0 | 0.81 | 0 | 1.18 | 0 |
| Ste12 - Swi6 | 0.63 | 0.68 | 0.8 | 0.89 | 0.32 | 0 | 2.64 | 0 | 4.43 | 3.13 | 2.66 | 2.59 |
| Ste12 - Swi4 | 0.71 | 0.58 | 0.91 | 0.98 | 0.35 | 0 | 3.92 | 0 | 7.23 | 3.29 | 4.12 | 2.7 |
| Yap1 - Yap7 | 5.66 | 3.47 | 0.8 | 0.89 | 0.32 | 0.74 | 1.05 | 0.94 | 1.36 | 0.29 | 0.77 | 0 |

of the expression of these genes is not important for the response of *S. cerevisiae* to the various organic acids.

### 2.4.6 Acetate and propionate responsive transcripts

The acetate-upregulated transcripts revealed a significant overrepresentation of the MIPS functional category "cell rescue, defense and virulence" and, more specifically, the subcategories "stress response" and "osmotic and salt stress response" (Table 2.4). The latter subcategory was also overrepresented among the propionate-upregulated genes and among the genes that were upregulated by both acetate and propionate (Table 2.4). The propionate-upregulated transcripts showed a strong overexpression of the MIPS category "metabolism" and, in particular, several subcategories involved in nitrogen and amino acid metabolism. A detailed investigation of the transcripts involved revealed many genes pertaining to biosynthesis and degradation of nitrogenous compounds, suggesting an overall up-regulation of nitrogen turnover. Consistently, binding of the Gcn4 transcription factor, which is involved in the general control of nitrogen metabolism, was very strongly overrepresented among the 252 genes upregulated in response to propionate (Table 2.5).
The 168 acetate-upregulated genes showed an overrepresentation of binding for the general stress response regulators Msn2 and Msn4. In addition, the consensus sequence of Hsf1, a regulator of heat shock proteins and possibly cell wall remodeling [Imaz 05] was abundant within the genes responding to acetate. Interestingly, a number of genes bound by Msn2/4 also respond to heat shock, which is consistent with the reported cross tolerance observed between mild acid stress and thermo-tolerance [Carm 98]. Binding sites for Cin5, a regulator involved in chitosan resistance [Zakr 05], were also overrepresented among the acetate-upregulated genes.
Intriguingly, the same transcription factor was found to be overrepresented among the acetate downregulated transcripts (Table 2.5). The 62 acetate downregulated genes showed an overrepresentation of the MIPS category "cellular transport, facilitation and routes". Closer inspection revealed many genes belonging to the major-facilitator superfamily (MFS), indicating that reduced transcription of membrane-transporter genes is an integral part of the response of *S. cerevisiae* to acetate. In sharp contrast to the propionate upregulated gene set, in which a plethora of transcription factors were overrepresented, the 528 genes downregulated on propionate showed no significant enrichment in functional categories and only a slight over-representation of a single transcription factor, Mcm1, which plays a central role in formation of repressor and activator complexes [Elbl 91]. Given the fact that Mcm1 is involved in a number of different repressor and activator complexes, it was not surprising that this subset failed to display a discernable functional grouping (Table 2.4).

## 2.5 Discussion

### 2.5.1 Methodology

This study represents the first attempt to compare cellular responses of *S. cerevisiae* to different organic acids at concentrations of the acids resulting in an identical decrease of the biomass yield on glucose. This indicates that the amount of ATP required for

maintenance of the intracellular pH ($pH_i$) and/or export of the anions increased drastically in comparison to the reference condition. Although the experimental setup used in this study does not provide insight on the transient changes in gene expression, which is reflective of the dynamic adaptive response to sudden changes in growth conditions, findings from this study can be used to facilitate functional analysis and increase the understanding under such conditions.
The relevance of this study might be challenged by stating that the concentrations of the weak acids did not result in complete growth arrest and that, therefore, the concentration of acids was not sufficient to induce any generic transcriptional responses. Although further dose-response work is definitely of interest, it is relevant to note that the concentrations of weak acids used in the present study were sufficient to (i) result in an over 2-fold change of the transcript level of more than 450 genes for each of the organic acids and (ii) induce specific response mechanisms to the organic acids studied. Examples of the latter include the *PDR12* gene which, in agreement with previous studies [Hatz 03, Pipe 98], was strongly upregulated in response to propionate, benzoate and sorbate but not acetate, and induction by acetate of *TPO2*, which encodes a trans-membrane protein implicated in the active efflux of poorly lipophilic acidic anions [Fern 05].
The present study was confined to the transcriptional level. However, it is known that relevant adaptations to organic acids (such as the activation of the plasma membrane ATPase, Pma1, [Serr 83]) may also occur posttranscriptionally. For example, such post-transcriptional responses may play a key role in the benzoate-induced tolerance to acetate and propionate [Wart 89]. Therefore, care should be taken to extrapolate the conclusions from the present study beyond the transcriptional level.

### 2.5.2 Comparison with known responses to organic acids and implications for current models of weak acid toxicity

A number of genes and regulons which are of utmost importance to tolerance to organic acids have been extensively described in previous research. For example, Pma1, which is responsible for maintaining intracellular pH via ATP-dependent efflux of protons is paramount for growth in the presence of weak acids [Holy 96, Pipe 97, Vieg 98]. The fact that the expression of *PMA1* is unchanged in the current study (Table 2.6) does not contradict this since the basal expression level may be sufficient to provide the necessary Pma1 activity to prevent intracellular acidification. Alternatively, unchanged transcriptional expression may be reflective of posttranscriptional regulation (as described above). Strikingly, the gene encoding the negative regulator of Pma1, *HSP30*, is differentially transcribed when the lipophilic and less-lipophilic acids are compared. In the presence of acetate and propionate, transcript levels of *HSP30* are upregulated (Table 2.6), whereas *HSP30* is not upregulated in response to benzoate and is actually down-regulated in the presence of sorbate. Since maintaining the proton-motive force is more challenging for more lipophilic weak acids, these observations give further indication to the different modes of toxicity of these two groups of acids. Moreover, *YRO2*, which is a homolog of *HSP30*, is upregulated on sorbate, acetate and propionate (Table 2.6), indicating a gap in the current understanding of the (posttranscriptional) regulation of Pma1, Yro2 and Hsp30p.
Aside from changes in $pH_i$ mediated by intracellular dissociation of free organic acids, the anion itself and the induction of energy-dependent efflux can influence energetics

**Table 2.6 – Discretized expression patterns of genes and regulons which have been previously described as important determinants of organic acid tolerance in *S. cerevisiae*.**
The last two columns indicate genes containing the consensus binding sequence for War1, which were induced in response to sorbate [Schu 04] and genes of the Haa1 regulon, which has previously been described as being particularly important for resistance to poorly lipophilic acids [Fern 05].

| | Acetate | Propionate | Benzoate | Sorbate | War1 regulon | Haa1 regulon |
|---|---|---|---|---|---|---|
| *PMA1* | 0 | 0 | 0 | 0 | | |
| *HSP30* | 1 | 1 | 0 | -1 | | |
| *WAR1* | 0 | 0 | 0 | 0 | | |
| *PDR12* | -1 | 1 | 1 | 1 | x | |
| *FUN34* | -1 | 1 | 1 | 1 | x | |
| *ALG12* | 0 | 1 | 1 | 1 | x | |
| *HXK1* | 0 | 0 | 0 | 0 | x | |
| *TFS1* | 0 | 1 | 1 | 0 | x | |
| *ACH1* | 0 | 1 | 0 | 0 | x | |
| *GAT1* | 0 | 1 | 1 | 1 | x | |
| *ALD4* | -1 | 0 | 0 | 0 | x | |
| *TPO1* | 0 | 0 | 0 | 0 | x | |
| *HAA1* | 0 | 0 | 0 | 0 | | |
| *TPO2* | 1 | 1 | 0 | 0 | | x |
| *TPO3* | 0 | 0 | 0 | 0 | | x |
| *YRO2* | 1 | 1 | 0 | 1 | | x |
| *YGP1* | 1 | 1 | 1 | 1 | | x |
| *GRE1* | 0 | 0 | 0 | 0 | | x |
| *PHM8* | 0 | 1 | 0 | 0 | | x |
| *YIR035C* | 1 | 1 | 0 | 0 | | x |
| *YLR297W* | 0 | 0 | 0 | 0 | | x |
| *YPR157W* | 1 | 0 | 0 | 0 | | x |
| *YER130C* | 1 | 0 | 0 | 0 | | x |

and other cellular processes. Previous research has outlined the toxic mechanisms of the anion and the export mechanisms which are activated to counteract the toxic accumulation of these anions. Namely, Pdr12, belonging to the War1 regulon, has been implicated as a key determinant of resistance to moderately lipophilic weak organic acids [Hatz 03, Pipe 98]. The protein encoded by *PDR12* functions in the energy-dependent export of moderately lipophilic organic acid anions from the cytosol [Holy 99]. Indeed, *PDR12* and a number of other genes that are dependent on *WAR1* were upregulated in response to benzoate, sorbate and propionate (Table 2.6). Interestingly, the same genes are either unchanged or downregulated upon exposure to acetate (the least lipophilic acid in the present study). Consequently, the importance of *PDR12* and other genes of the *WAR1* regulon is once again highlighted for moderately lipophilic weak organic acids.

Another regulon which has recently been identified as an important determinant of organic acid tolerance is the *HAA1* regulon [Fern 05]. The expression pattern of this regulon is clearly distinct from that of the *WAR1* regulon (Table 2.6) and the distinction appears to be correlated to membrane affinity. The involvement of both the *HAA1* and *WAR1* regulons in response to propionate is especially intriguing considering that the membrane affinity of propionate is an intermediate between the poorly lipophilic acetate and the moderately lipophilic compounds benzoate and sorbate. Particular genes,

which are regulated by Haa1 may encode proteins that facilitate the export of poorly lipophilic anions. The most likely candidate is *TPO2*, which is upregulated upon exposure to acetate and propionate in this study. *TPO2* (and *TPO3*) encode plasma membrane multidrug transporters that are known to promote the export of spermine [Albe 03, Uemu 05]. However, deletion of *HAA1* or *TPO3* in the presence of acetic acid resulted in increased lag times which were correlated to increased levels of intracellular acetate [Fern 05]. Therefore, analogous to War1 and moderately lipophilic acids, Haa1 may represent a key activator of defense mechanisms required for resistance to poorly lipophilic acids.

### 2.5.3 Transcriptional responses to weak acids: leads for functional analysis

Although the aim of this work was not to investigate the molecular mechanisms involved in the cellular responses to weak acids, the dataset generated in this study may be applied to direct future functional analysis studies. For instance, the common upregulation of *SOD2* in these anaerobic chemostats merits further exploration of the role of superoxide dismutases in anaerobic conditions. Although weak acids have been associated with the formation of reactive oxygen species (ROS) in aerobic cultures [Pipe 99], it is difficult to envisage such a link under the anaerobic conditions used in the present study. However, the identification of genes involved in protection against ROS is perhaps indicative of a physical interaction between the mitochondria and weak acids, which results in damage or disruption of the mitochondrial membrane and ultimately leads to increased ROS production in the presence of oxygen.

The importance of the cell wall has also been highlighted in this study. Although the cell wall is not generally considered to be a protective barrier to small molecules due to its porous nature [Nobe 91], the consistent identification of cell wall-related genes in response to organic acids suggests otherwise. For example, *SPI1* has been shown to have a prominent role in weak acid tolerance [Simo 06], while *YGP1* and *SPI1* showed increased expression in response to low pH (pH 3.5 vs pH 5.5), in conjunction with increased expression and immobilization of Pir-related cell wall proteins [Kapt 01]. Furthermore, increased presence of mannoproteins in the cell wall was correlated to decreased cell wall porosity, a characteristic that has been mainly attributed to the bulky mannan side-chains [Nobe 90]. Indeed, deletion of several mannosyltransferase-encoding genes has been shown to result in hypersensitivity to sorbic acid [Moll 04]. However, the transcriptional regulator Rlm1, a key regulator of cell wall integrity [Levi 05], was not among the enriched transcription factors in this study.

In the presence of sorbate, Ste12 and Tec1, which are both final targets of the Ras2-activated signaling cascade that regulates pseudohyphal growth [Ganc 01], were amongst the upregulated transcription factors. Although nitrogen starvation [Gime 92] and various environmental stresses [Zara 00] have been shown to induce such morphological changes, microscopic inspection did not show pseudohyphal growth in the reference or acid-challenged cultivations in this study. Consequently, this may represent a previously uncharacterized relationship between regulators of pseudohyphal growth and weak organic acid tolerance.

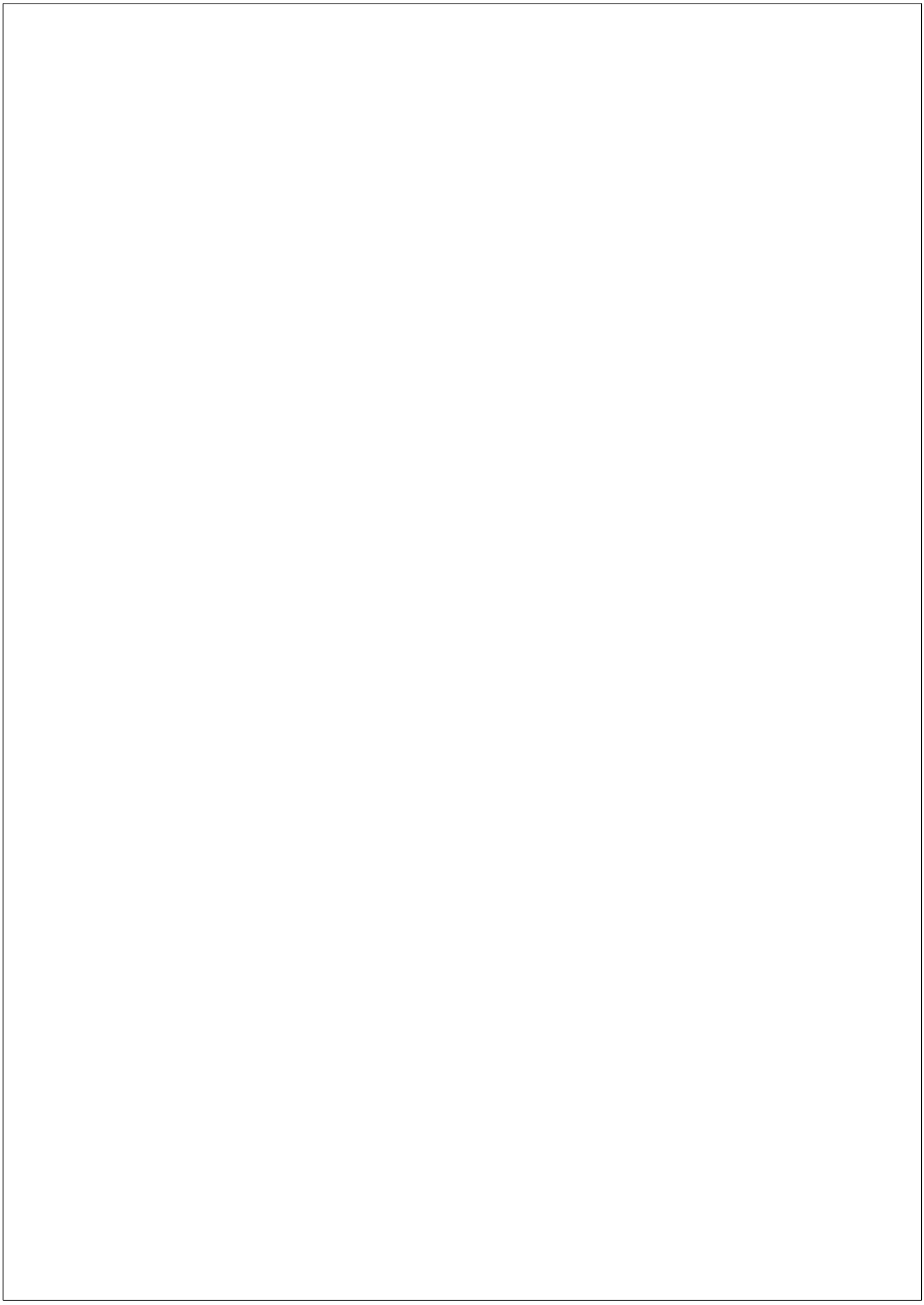
Although the identity and function of the affected genes was different, the less lipophilic amino acids caused significant transcriptional responses of genes involved in nitrogen metabolism or transport (Table 2.4). Propionate induced upregulation of genes involved

in biosynthesis and degradation of various nitrogen containing compounds including organic acids, while acetate caused a downregulation of the transport of several amino acids. These observations provide a strong incentive for further studies on the relationship between these weak acid and central nitrogen metabolism. This is especially intriguing given the fact that the *ACS1* gene, which encodes an acetyl-CoA synthetase that can also activate propionate [Berg 96], was only upregulated in the presence of propionate. Consequently, the effect on central nitrogen metabolism is either mediated by propionate itself or by an, as yet, unknown metabolite. However, propionate concentrations were identical in the growth medium reservoir and the culture supernatant, which suggests that propionate itself is the cause. Further work is required to investigate whether the effect of propionate and acetate on amino acid biosynthesis are due to specific effects on individual metabolic pathways (such as the acetate inhibition of methionine biosynthesis in *E. coli* [Roe 02]) or to general effects on regulatory networks involved in nitrogen metabolism.
Finally, the common downregulation of membrane transport processes is noteworthy. Downregulation of amino acid transport is consistent with the observations of Bauer and Kuster [Baue 03] indicating a general disruption of aromatic acid uptake. However, the current data indicates a more general limitation of membrane transport processes which is likely initiated in an attempt to reduce the diffusional entrance of weak acids. Such an aspecific response is somewhat counterintuitive as the reduced uptake of nitrogenous compounds along with sterols and heavy metals is bound to have far reaching, and possibly negative, effects on yeast metabolism. Consequently, detailed studies of the signaling mechanisms which trigger such a seemingly broad response and the secondary effects of the resulting reduction in nutrient uptake are imminent.

### 2.5.4 A minimal generic transcriptional response to weak acids: implications for applied research

Although a comparison of physiological parameters (yields, fluxes) suggested a similar response to benzoate, sorbate, acetate and propionate, large differences were found with respect to the transcriptional responses to these weak acids. Indeed, when challenged with different weak acids under the experimental conditions applied in the current investigation, *S. cerevisiae* does not exhibit extensive similarities in transcriptional modulation that can be characterized by a common functional category or transcription factor activation/repression. The consequences of these findings are that the often-used term "weak-organic acid stress" should preferably be avoided and that the use of individual organic acids as "model compounds" for general responses to organic acids should be treated with caution. Instead, molecular analysis of the response to weak acids should take into account the unique responses to individual acids.
Although care should be taken to extrapolate from transcript profiles to in vivo cellular processes (a changed transcript level is not necessarily indicative for a changed in vivo activity of the encoded protein [Dara 04, Kolk 06, Ross 06]), our observations strongly suggest the possibility that the toxicity of weak acids involves overlapping but unique sets of cellular targets. Although the synergistic interactions and physiological responses have previously been investigated for various combinations of acids [Nare 01, Sava 02], the underlying transcriptional changes have yet to be determined. From an applied point of view, this study suggests the likelihood that different weak acids may act synergistically due to the fact that they induce dissimilar transcriptional responses. Therefore,

understanding the interaction between acids at the transcriptional level could lead to improved strategies for growth inhibition at reduced concentrations of these preservatives.



# CHAPTER 3

# TRANSCRIPTIONAL RESPONSE TO ZINC LIMITATION

In this chapter microarray data is employed of *S. cerevisiae* grown under six different conditions, i.e. three different nutrient limitations; carbon, nitrogen and zinc, grown both aerobically and anaerobically. Discretization is used to build a tertiary representation of the genes. In this case, however, there is no reference condition. This makes it non-trivial to decide upon up- and downregulation. The discretization procedure uses a k-means clustering procedure for each gene individually; the six conditions are clustered to decide, which of these conditions are labeled upregulated, downregulated or not differentially expressed. In this work, genes are clustered together when their discretized expression patterns satisfy certain constraints. For example, genes that have a higher discretized expression value under zinc limitation than under the other two limitations in both the aerobic and anaerobic case are grouped together. The results from this analysis were used to redefine the zinc-specific Zap1 regulon. Also, the study reveals a more important role for zinc in mitochondrial function and biogenesis than so far assumed.

This chapter is published as:

'Physiological and transcriptional responses of *Saccharomyces cerevisiae* to zinc limitation in chemostat cultures'

Raffaele De Nicola, Lucie A. Hazelwood, Erik A. F. De Hulster, Michael C. Walsh, Theo A. Knijnenburg, Marcel J.T. Reinders, Graeme M. Walker, Jack T. Pronk, Jean-Marc Daran and Pascale Daran-Lapujade

Applied and Environmental Microbiology, Volume 73 No 23 p. 7680-7692, December 2007

Note: TAK's contribution to this chapter is limited to the computational analysis of the microarray data.

## 3.1 Abstract

Transcriptional responses of *Saccharomyces cerevisiae* to Zn availability were investigated at a fixed specific growth rate under limiting and abundant Zn concentrations in chemostat culture. To investigate the context-dependency of this transcriptional response and eliminate growth rate-dependent variations in transcription, yeast was grown under several chemostat regimes resulting in various carbon (glucose), nitrogen (ammonium), zinc and oxygen supplies. A robust set of genes that responded consistently to Zn limitation was identified and enabled the definition of the Zn-specific Zap1 regulon comprising of 26 genes and characterized by a broader ZRE consensus (MHHAACCBYNMRGGT) than so far described. Most surprising was the Zn-dependent regulation of genes involved in storage carbohydrate metabolism. Their concerted downregulation was physiologically relevant as revealed by a substantial decrease in glycogen and trehalose cellular content under Zn limitation. An unexpectedly large amount of genes were synergistically or antagonistically regulated by oxygen and Zn availability. This combinatorial regulation suggested a more prominent involvement of Zn in mitochondrial biogenesis and function than hitherto identified.

## 3.2 Introduction

Zinc is a cofactor of many proteins and is indispensable for their catalytic activity and/or structural stability. Zn is also a ubiquitous component of enzymes involved in transcription and of the Zn finger proteins that regulate gene expression [Bohm 97]. In the yeast *Saccharomyces cerevisiae*, zinc is estimated to be required for the function of nearly 3% of the proteome [Bohm 97]. Besides its involvement in protein structure and function [Vall 90, Mago 92], interaction of zinc with lipids contributes to regulation of membrane fluidity [Bind 01] and its interaction with nucleic acids helps to prevent deleterious radical reactions [Berg 96]. Deficiency of this essential trace element can have severe consequences. For example, in beer fermentation, zinc depletion in wort leads to "sluggish" fermentation and thus to deterioration of beer quality [Jaco 79]. While accurate monitoring of the zinc concentration in such industrial fermentations is important, formation of complexes with polyphenols, proteins and other compounds [Kred 99] implies that the concentration of zinc per se does not always accurately predict its bioavailability to yeast.

Excess zinc is toxic. It can compete with other metal ions for the active sites of enzymes or intracellular transport proteins [Gita 98, Kami 89, Miya 00, Mart 03, Rega 06]. For this reason, organisms have evolved mechanisms that tightly control intracellular zinc levels. Zinc homeostasis in yeast can be mediated via i) control of zinc uptake, ii) storage of zinc in vacuoles, iii) intracellular binding of zinc by metallothioneins and iv) efflux of zinc from the cells. In *S. cerevisiae*, various proteins involved in zinc uptake and storage have been identified in the last decade. Zinc uptake across the plasma membrane mainly occurs via the transporters Zrt1 and Zrt2 [Zhao 96b, Zhao 96a]. Fet4 and Pho84, low-affinity and broad substrate range transporters of heavy metals, can also transport zinc [Wate 02]. Zinc storage occurs in the vacuole and transport of zinc into this compartment is mediated by Cot1 and Zrc1 [Li 01, Miya 01], while release of zinc from vacuolar storages is mediated by Zrt3 [MacD 00, MacD 02]. Msc2 [Li 01] and Yke4 [Kuma 06] are implicated in transport of Zn into the lumen of the endoplasmic reticulum

and perhaps an additional organelle involved in the secretory pathway. The genes encoding these transporters are transcriptionally induced by Zap1 (Zinc Activated Protein) under conditions of zinc limitation or deficiency [Zhao 97]. Contrary to the situation in mammalian cells, no plasma membrane transporter dedicated to zinc export from yeast cells has been identified so far [Palm 95]. Two cytosolic metallothioneins (Cup1-1 and Cup1-2) involved in copper chelation can also bind zinc [Wing 85]. However, the expression of these proteins is not zinc-dependent, and involvement in zinc detoxification has not yet been demonstrated [Wing 85].
In order to better define the Zap1 regulon, Lyons *et al.* analyzed the genome-wide transcriptional response of a *S. cerevisiae* Zap1 mutant strain and a control strain to zinc abundance or depletion [Lyon 00]. A combinatorial analysis identified a subset of 46 zinc-responsive genes whose expression was reduced in the Zap1 mutant and that possessed a Zinc-Responsive Element (ZRE, 5'-ACCYYNAAGGT-3'). Among the members of this updated defined Zap1 regulon were the well-characterized plasma membrane, vacuolar and endoplasmic reticulum zinc transporters. However, involvement of many of the proposed Zap1 targets in zinc homeostasis was difficult to interpret and, as suggested by the authors, may be due to contribution of factors other than zinc depletion. Indeed, these experiments were performed in shake flask in which the growth conditions cannot be strictly monitored and maintained at constant level as the pH, the dissolved oxygen and nutrient concentrations change during growth. Furthermore, zinc depletion and ZAP1 deletions are bound to reduce the specific growth rate as compared to zinc sufficient cultures of a wild-type strain. The regulation of gene expression is therefore affected not only by the difference in growth conditions but also by the specific growth rate [Rege 06]. This variation in gene regulation can obscure the interpretation of the results.
The goal of the present study was to investigate physiological and transcriptional responses of *S. cerevisiae* to zinc limitation, while minimizing the impact of secondary effects of zinc limitation. To this end, *S. cerevisiae* was grown at a fixed specific growth rate, oxygen availability, temperature and pH under zinc limitation in chemostat cultures. Comparing the transcriptome of zinc-limited cultures to those of carbon and nitrogen limited cultures identified sets of genes that responded uniquely to zinc limitation. Furthermore, these cultures were grown both in the presence and the complete absence of oxygen, in order to identify genes that are subjected to combinatorial control by oxygen and zinc availability.

## 3.3 Materials and methods

### 3.3.1 Yeast strain and maintenance

The haploid prototrophic *S. cerevisiae* strain CEN.PK 113-7D (MATa) was obtained from Dr. P. Kötter, Frankfurt, Germany. Zinc-depleted cells were obtained by four serial transfers of yeast cells in shake flasks containing synthetic medium [Verd 92] from which zinc was omitted, and subsequently mixed with glycerol (final concentration 20%), aliquoted and stored at -80°C.

**Table 3.1 – Composition of the media used to perform carbon, nitrogen and zinc limitation under aerobic and anaerobic environment.**
Numbers in bold indicate the modifications introduced to the synthetic media described by Verduyn *et al.* [Verd 92] in order to obtain the relevant nutrient limitations.

| | Limiting nutrient | Glucose g/l | $(NH_4)_2SO_4$ g/l | $K_2SO_4$ g/l | $ZnSO_4{\cdot}7H_2O$ mg/l |
|---|---|---|---|---|---|
| Aerobic | Carbon | **7.5** | 5 | - | 4.5 |
| | Nitrogen | 59 | **1** | **5.3** | 4.5 |
| | Zinc | 66 | 5 | - | **0.014** |
| Anaerobic | Carbon | **25** | 5 | - | 4.5 |
| | Nitrogen | 46 | **1** | **5.3** | 4.5 |
| | Zinc | 58 | 5 | - | **0** |

**Table 3.2 – Physiological characteristics of CEN.PK113-7D grown in aerobic and anaerobic carbon-, nitrogen-, or zinc-limited chemostat cultures (dilution rate of 0.1 $h^{-1}$).**
DW: biomass dry weight; NA: not applicable; ND: not determined; BD: below detection
[a] Biomass yield on glucose
[b] Respiratory quotient: $q_{CO_2}/q_{O_2}$
[c] Specific consumption rates of glucose and oxygen and specific production rates of ethanol, glycerol, acetate and carbon dioxide

| Growth condition | | Residual glucose (mM) | Zn in biomass $\mu mol{\cdot}g_{DW^{-1}}$ | $Y_{SX}$[a] $g_{DW}{\cdot}g_{glucose^{-1}}$ | RQ[b] | Carbon recovery (%) |
|---|---|---|---|---|---|---|
| Aer | C | BD | 2.36 ± 0.4 | 0.49 ± 0.00 | 1.0± 0.0 | 98 ± 3 |
| | N | 92.7 ± 5.5 | 2.74 ± 1.3 | 0.09 ± 0.00 | 4.5 ±0.2 | 96 ± 1 |
| | Zn | 102.4 ± 6.4 | 0.9 ± 0.2 | 0.10 ± 0.00 | 4.5 ±0.1 | 105 ± 2 |
| Ana | C | BD | 2.1 ± 0.4 | 0.09 ± 0.00 | NA | 101 ± 2 |
| | N | 100.8 ± 8.6 | 2.74 ± 1.3 | 0.07 ± 0.00 | NA | 101 ± 2 |
| | Zn | 110.4 ± 3.9 | 0.52 ± 0.03 | 0.07 ± 0.00 | NA | 100 ± 1 |

| Growth condition | | Rates[c], $mmol{\cdot}g_{DW^{-1}}{\cdot}h^{-1}$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $q_{Glucose}$ | $q_{Ethanol}$ | $q_{Glycerol}$ | $q_{Acetate}$ | $q_{O_2}$ | $q_{CO_2}$ |
| Aer | C | - 1.1 ± 0.0 | 0.0 ± 0.0 | BD | BD | - 2.8 ± 0.3 | 2.8 ± 0.3 |
| | N | - 5.8 ± 0.1 | 8.0 ± 0.1 | 0.08 ± 0.01 | 0.03 ± 0.01 | - 2.7 ± 0.1 | 12.1 ± 0.2 |
| | Zn | - 5.3 ± 0.0 | 8.1 ± 0.2 | 0.08 ± 0.01 | 0.07 ± 0.01 | - 2.8 ± 0.0 | 12.3 ± 0.2 |
| Ana | C | - 6.0 ± 0.0 | 9.6 ± 0.1 | 0.81 ± 0.06 | 0.01 ± 0.00 | NA | 10.3 ± 0.4 |
| | N | - 8.4 ± 0.0 | 13.5 ± 0.6 | 0.76 ± 0.04 | 0.06 ± 0.05 | NA | 14.8 ± 0.3 |
| | Zn | - 8.4 ± 0.0 | 13.7 ± 0.2 | 1.09 ± 0.01 | 0.16 ± 0.02 | NA | 15.5 ± 0.5 |

### 3.3.2 Minimizing Zn contamination of culture vessels

To minimize zinc contamination, all glassware (including shake flasks for pre-cultivation), tubing and fermentors were subjected to an overnight soak in 2 % nitric acid, followed by two washes with deionised water, one wash with 0.1 M EDTA and four further washes with deionised water.

### 3.3.3 Media for chemostat cultivation

The synthetic medium composition was based on that described by Verduyn *et al.* [Verd 92]. The modifications introduced for carbon, nitrogen and zinc limited growth are listed in Table 3.1. In all chemostats except for those limited by carbon, the residual glucose concentration was targeted to 17 g/l (95 mM) in order to have the same degree of glucose repression (Table 3.2). Under anaerobic glucose-limited conditions, the glucose concentration was increased to compensate for a low biomass yield. The decreased sulfate concentration (resulting from the reduced $(NH_4)_2SO_4$ concentration under nitrogen limitation) was compensated by $K_2SO_4$ addition. The zinc replete cultures (carbon and nitrogen-limited) contained excess zinc concentration, but at sub-toxic levels [Jone 84]. In anaerobic zinc-limited cultures, a minute zinc contamination (probably leaking from the metal fermentor parts) was enough to sustain growth. Conversely, aerobic zinc-limited cultures could not grow at a dilution rate of 0.10 $h^{-1}$ without the addition of zinc as 0.05 $\mu$M zinc sulfate. For anaerobic cultivations, the reservoir medium was supplemented with the anaerobic growth factors Tween-80 and ergosterol [Verd 90].

### 3.3.4 Chemostat cultivation

Zinc-depleted pre-cultures were obtained by inoculating shake flasks that contained 100 ml zinc-free synthetic medium with zinc-depleted cells (obtained as described above). After overnight cultivation, these zinc depleted precultures were inoculated in 2-liter fermentors (Applikon) with a working volume of 1 l [Berg 96]. Chemostat cultures were fed with synthetic medium (as described in the previous section) that limited growth by carbon, nitrogen or zinc with all other growth requirements in excess and at constant residual concentration [Boer 03]. The dilution rate was set at 0.10 $h^{-1}$. Cultures were assumed to be in steady-state when, after at least five volume changes, culture dry-weight, glucose concentration, carbon-dioxide production rate and oxygen consumption rate varied by less than 2% during one additional volume change [Fere 99]. Steady-state samples were taken after 10 generations at the latest to avoid strain adaptation due to long-term cultivation [Jans 04]. Each cultivation condition was performed in triplicate. The pH was measured on-line and kept constant at 5.0 by the automatic addition of 2 M KOH using an Applikon ADI 1030 Biocontroller. The stirrer speed was set at 800 rpm. Anaerobic conditions were maintained by sparging the medium reservoir (0.05 liter·$min^{-1}$) and the fermentor (0.5 liter·$min^{-1}$) with pure nitrogen gas. Norprene tubing and butyl rubber septa were used to minimize oxygen diffusion into the anaerobic cultures [Viss 90]. The off-gas was cooled by a condenser connected to a cryostat set at 2°C. Oxygen and carbon dioxide were measured off-line with an NGA 2000 Rosemont gas analyzer.

### 3.3.5 Analytical methods

Culture supernatants were obtained after centrifugation of samples from the chemostats. For the purpose of glucose determination and carbon recovery, culture supernatant and media were analyzed by high performance liquid chromatography (HPLC) on an AMINEX HPX-87H ion exchange column using 5 mM $H_2SO_4$ as the mobile phase. Culture dry weights were determined via filtration as described by Postma *et al.* [Post 89]. Trehalose and glycogen measurements were adapted according to François *et al.* [Fran 01]. Trehalose was determined in triplicate measurements for each chemostat. Glycogen was determined in duplicate for each chemostat. Glucose was determined using the UV-method based on Roche kit no. 0716251.

### 3.3.6 Microarray analysis

Sampling of cells from chemostats and total RNA extraction was performed as previously described [Abbo 07]. Probe preparation and hybridization to Affymetrix Genechip microarrays were performed following Affymetrix instructions. The one-cycle eukaryotic target labeling assay was used, starting with 15$\mu$g of total RNA. The quality of total RNA, cDNA, cRNA and fragmented cRNA were checked using the Agilent Bioanalyzer 2100 (Agilent Technologies). Results for each growth condition were derived from three independent culture replicates.

### 3.3.7 Transcriptomics data acquisition and statistical analysis

Acquisition and quantification of array images and data filtering were performed using Affymetrix GeneChip Operating Software version 1.2. Before comparison, all arrays were globally scaled to a target value of 150 using the average signal from all gene features using GeneChip Operating Software (GCOS), version 1.2. To eliminate insignificant variations, genes with expression values below 12 were set to 12 as previously described [Boer 03].
To detect genes that exhibited differential expression in at least one of the experimental conditions, an in-house version of SAM (Significance Analysis of Microarrays) [Tush 01] was employed using the multiclass setting. Genes with a Q-value below the median FDR (false discovery rate) of $1.5 \cdot 10^{-4}$ were considered differentially expressed.
Transcript data can be downloaded from GEO under the following series accession numbers: zinc-limited chemostats GSE8035; carbon-limited chemostats GSE8088 and GSE5326; nitrogen-limited chemostats GSE8089.

### 3.3.8 Grouping of genes into modules

The continuous expression levels of all (1500) differentially expressed genes were discretized, as described in Knijnenburg *et al.* [Knij 07]. Resultantly, each gene is represented by a discretized expression pattern of length six, indicating whether the gene is not differentially expressed (0), upregulated (1) or downregulated (-1) under each of the six cultivation conditions. For example, a gene that has the following discretized expression pattern:

| C-Ana | N-Ana | Zn-Ana | C-Aer | N-Aer | Zn-Aer |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | -1 |

is upregulated when grown anaerobically under a nitrogen limitation (N-Ana) and downregulated when grown aerobically under a zinc limitation (Zn-Aer), while the four other conditions do not exhibit differential expression. Genes are grouped into modules based on this discretized representation by imposing certain constraints on the discretized expression pattern of a gene in order for it to be part of a particular module. For example, a module could be formed by grouping all genes that have a higher discretized expression level under the zinc limitation, when compared to the other two limitations, both for aerobic and anaerobic growth. This approach provides a coherent and meaningful way to create modules of genes, since the expression behavior of the genes in a module is directly related to the cultivation conditions, allowing for a straightforward interpretation. In our study, six modules were created. The exact constraints on the discretized expression pattern of a gene to be included in one of the six modules are found in the Appendix. Table 3.3 gives a short verbal description for each of the modules.

### 3.3.9 Hypergeometric tests

The six modules were consulted for enrichment in functional annotation and significant transcription factor (TF) binding. To test for significant relations the hypergeometric test was employed. In the case of the TF binding data, the largest available TF binding dataset for yeast in its most conservative setting (highest binding confidence) was used [Harb 04]. This dataset, which originally indicates the number of binding sites for each of 102 TFs in the promoter region of each gene, was binarized, such that the data indicates whether a TF can bind a gene (upstream) or not. Then, the hypergeometric test assesses if a TF (or a TF pair) can bind the promoter region of the genes in a module much more frequently than in a randomly selected set of genes. In case of the employed gene annotation information (MIPS [Mewe 97] and KEGG [Kane 00]) it assesses if the number of genes in a module that belongs to a particular functional category is much larger than would be expected by chance. The $P$-value cut-off to decide whether a relation is significant is $P \leq 1/(n_c n_x)$, where $n_c$ is the number of modules and $n_x$ is the number of TFs (or TF pairs) or the number of MIPS or KEGG annotation categories. This adjustment for multiple testing, corresponds with a per comparison error rate (PCER) of one [Ge 03].

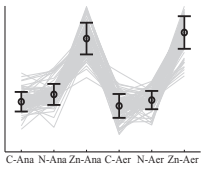
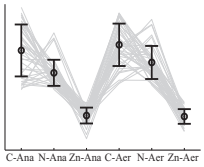
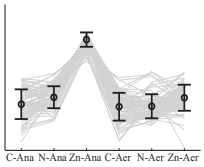
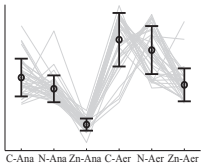
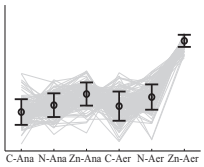
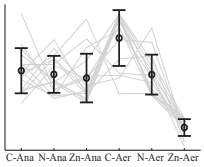
### 3.3.10 Motif discovery

The promoters (from -800 to -1) of the genes in each module were analyzed for overrepresented regulatory motifs using the web-based software MEME [Bail 94]. The $P$-value cut-off to consider a motif significant was $10^{-4}$. Other parameter settings included a motif width from 6 to 15 nucleotides, that could be repeated any number of times.

### 3.3.11 Comparison with the transcriptome study from Lyons *et al.*

The data from the Lyons *et al.* [Lyon 00] were downloaded from http://genome-www.stanford.edu/zinc/rawdata.html. As this website only provides raw data, the array data were processed following the instructions described in their publication and 496 genes that were upregulated in response to zinc depletion were thus

**Table 3.3 – Clustering of the zinc-responsive genes.**
[a] Expression pattern of genes of each module along with their averaged expression and standard deviation. Here, the expression levels of each gene are normalized to have zero mean and unit variance. (y-axis: normalized expression, x-axis: culture condition, from left to right: anaerobic carbon, nitrogen, zinc limitation, and aerobic counterparts).
[b] Each data-set was analyzed individually for enrichment of transcription factor binding and functional categories as described in the Material and Methods section.

| Module (no. of genes) | Description | Expression pattern[a] | Overrepresented category(ies)[b] | No. in: Genome | No. in: Module | *P*value |
|---|---|---|---|---|---|---|
| 1 (93) | Genes up-regulated under the zinc limitation regardless of aeration |   | TF: Zap1p | 8 | 6 | $2.2 \times 10^{-10}$ |
| | | | MIPS: heavy metal ion transport (Cu, Fe, etc.) | 52 | 7 | $8.7 \times 10^{-6}$ |
| 2 (40) | Genes down-regulated under the zinc limitation regardless of aeration |   | TF: Msn2 | 65 | 11 | $6.8 \times 10^{-4}$ |
| | | | MIPS: C-compound and carbohydrate metabolism | 508 | 3 | $2.0 \times 10^{-4}$ |
| | | | Metabolism of energy reserves (e.g. glycogen, trehalose) | 53 | 5 | $1.7 \times 10^{-5}$ |
| | | | Metabolism of Leu and Val | 7 | 3 | $7.8 \times 10^{-6}$ |
| | | | KEGG: Val, Leu and Ile biosynthesis | 16 | 3 | $1.2 \times 10^{-4}$ |
| | | | Panthotenate and CoA biosynthesis | 10 | 3 | $2.7 \times 10^{-5}$ |
| 3 (77) | Genes up-regulated only under anaerobic zinc limitation |   | TF: Hsf1 | 133 | 7 | $1.0 \times 10^{-3}$ |
| | | | Skn7 | 156 | 8 | $5.4 \times 10^{-4}$ |
| | | | MIPS: homeostasis of cations | 162 | 9 | $1.3 \times 10^{-4}$ |
| 4 (36) | Genes down-regulated only under anaerobic zinc limitation |   | MIPS: proteasomal degradation (ubiquitin/proteasamol pathway) | 191 | 7 | $7.7 \times 10^{-5}$ |
| | | | KEGG: TCA cycle | 30 | 3 | $6.0 \times 10^{-4}$ |
| | | | Proteasome | 34 | 4 | $3.5 \times 10^{-5}$ |
| 5 (119) | Genes up-regulated only under aerobic zinc limitation |   | TF: Yap7 | 158 | 12 | $3.4 \times 10^{-5}$ |
| 6 (16) | Genes down-regulated only under aerobic zinc limitation |   | MIPS: Metabolism | 1526 | 11 | $1.8 \times 10^{-4}$ |

isolated. The slightly larger size of this gene set compared to the one isolated by Lyons *et al.* (458 genes) probably results from a few differences in data handling.

## 3.4 Results

### 3.4.1 Establishing Zn-limited chemostat cultures of *S. cerevisiae*

While macronutrient limitation in chemostats can be achieved in a straightforward manner, establishing micronutrient limitation still presents an experimental challenge. This holds especially for metals (Zn, Fe, Cu) that are present in laboratory equipment and that can sustain growth at extremely low concentrations (typically in the micromolar range). Despite thorough and repeated washing steps and use of high-grade medium components, we did not achieve completely Zn-free cultivation conditions, presumably due to Zn leakage from the metal parts (fermentor lid, pipes and connections). This contamination was sufficient to allow for anaerobic Zn-limited growth at a steady-state biomass concentration of 2.5 $g{\cdot}L^{-1}$. However, 0.05 $\mu$M $ZnSO_4$ had to be added to the Zn-deficient medium to enable aerobic Zn-limited growth (steady-state biomass concentration 4.2 $g{\cdot}L^{-1}$). Addition of 15 $\mu$M Zn to anaerobic and aerobic Zn-limited cultures resulted in a large increase of the biomass concentration, thus confirming that growth was solely limited by Zn availability (data not shown). The Zn content of biomass from Zn-limited cultures was up to five-fold lower than that of carbon- and nitrogen-limited cultures (Table 3.2). Consistent with a higher Zn requirement for aerobic cultivation, the Zn content of biomass from aerobic Zn-limited cultures was two-fold higher than that of anaerobic Zn-limited cultures (Table 3.2). Since genes encoding Zn transporters were not differentially transcribed in the presence and absence of oxygen, this difference is unlikely to be due to a different affinity for Zn uptake.

### 3.4.2 Physiology of Zn, glucose- and ammonia-limited chemostat cultures

Zn-limited cultures were grown at a high residual glucose concentration. Comparison of their physiology and transcriptome with those of glucose-limited cultures will therefore also identify changes caused by the different glucose concentrations in the cultures. Therefore, nitrogen-limited cultures, grown at the same residual glucose concentration as the Zn-limited cultures, were included as an additional reference situation. The combination of three nutrient limitations under aerobic and anaerobic conditions resulted in six unique physiological situations (Table 3.2).
Only in the carbon-limited aerobic cultures, a completely respiratory sugar metabolism was observed, resulting in a high biomass yield on glucose (Table 3.2). In the anaerobic cultures, glucose metabolism was fully fermentative, the main products of glucose dissimilation being ethanol and carbon dioxide. Finally, in glucose-sufficient (i.e. N- or Zn-limited) aerobic cultures, a mixed respiro-fermentative metabolism was observed. The Zn-limited cultures strongly resembled the nitrogen-limited cultures with respect to biomass yields and rates of product formation. Even under anaerobic conditions, the biomass yield on glucose of these glucose sufficient cultures was lower than that of glucose-limited cultures, indicating a partial uncoupling of dissimilation and biomass formation under these 'energy excess' conditions. The only notable difference was a slightly

**Table 3.4 – Identity and expression levels of the genes from Module 1 consistently upregulated in response to zinc limitation and containing ZRE sequences (Zinc Responsive Element).**
Genes indicated in bold were also part of the regulon defined by Lyons *et al.*.

| Gene | Description | no of of ZREs | Transcript level Anaerobic C | Anaerobic N | Anaerobic Zn | Aerobic C | Aerobic N | Aerobic Zn |
|---|---|---|---|---|---|---|---|---|
| ***ZAP1*** | Zn-responsive TF | 1 | 23.2 | 41.5 | 479.9 | 33.8 | 43.5 | 356.6 |
| ***ZRT1*** | High affinity Zn transporter | 3 | 175.6 | 286 | 2004.1 | 68.3 | 240.6 | 1867.5 |
| ***ZRT2*** | Low affinity Zn transporter | 2 | 108.6 | 124 | 739.4 | 102.8 | 162.8 | 551.5 |
| ***ZRT3*** | Vacuolar Zn efflux | 1 | 236.9 | 257.2 | 1665.5 | 313.4 | 282.9 | 1708.5 |
| ***ZRC1*** | Vacuolar Zn influx | 1 | 237.8 | 367.3 | 712.2 | 261.3 | 337.8 | 811.6 |
| ***ZRG17*** | Putative Zn transporter | 1 | 84.7 | 91.9 | 446.3 | 142.6 | 97.9 | 527.4 |
| ***FET4*** | Low affinity Fe transporter | 1 | 235 | 223.8 | 743.3 | 12 | 89.9 | 443.5 |
| ***ADH4*** | Alcohol dehydrogenase | 3 | 153.3 | 206.3 | 2889.8 | 76.6 | 118.1 | 2872.1 |
| *HOR2* | Glycerol-P phosphatase | 1 | 45.1 | 63.3 | 141.3 | 97.3 | 84.9 | 198.5 |
| ***DPP1*** | DAGPP phosphatase | 1 | 307.3 | 517.8 | 1019.4 | 294.4 | 636.8 | 1233.3 |
| ***URA10*** | Pyrimidine biosynthesis | 1 | 18.7 | 25.1 | 81.3 | 30.4 | 16.7 | 50.9 |
| *FLO11* | Cell surface flocculin | 1 | 1150.8 | 1293.3 | 1935.4 | 42 | 60.2 | 2105 |
| ***ZPS1*** | Cell surface mannoprotein | 2 | 74.7 | 225.8 | 3385.4 | 139.5 | 119 | 3349.2 |
| ***MNT2*** | Mannosyl transferase | 3 | 18.1 | 12 | 51.6 | 12 | 12 | 40.9 |
| *KTR6* | Mannosyl transferase | 1 | 449.7 | 388.4 | 537.8 | 314 | 253.5 | 620.4 |
| ***MCD4*** | Transferase for GPI anchor synthesis | 1 | 337 | 323.3 | 996.5 | 232.2 | 279.2 | 1190.3 |
| ***ZIP1*** | Synaptonemal complex | 1 | 12 | 12 | 46.2 | 12.5 | 12 | 24.6 |
| *KTI12* | tRNA modification | 1 | 157.9 | 134.1 | 218.4 | 120.3 | 121.5 | 278.3 |
| *VTC3* | Vacuolar transporter chaperone | 1 | 66.4 | 95.7 | 175 | 51.1 | 85.8 | 272.8 |
| *TEX1* | TREX complex | 1 | 29.4 | 29.9 | 91.5 | 25.1 | 24.4 | 177.9 |
| *MUP1* | Methionine transporter | 1 | 69.6 | 38.7 | 636.5 | 128.8 | 211.9 | 912.5 |
| ***YNL254C*** | Unknown | 1 | 22.2 | 27.5 | 342 | 17.4 | 32.3 | 354.6 |
| *YER130C* | Unknown | 1 | 32.2 | 34.5 | 179.1 | 30.7 | 28.1 | 66.8 |
| ***ICY2*** | Unknown | 2 | 290.5 | 204.4 | 1939.2 | 568.3 | 143.8 | 1436.1 |
| ***VEL1*** | Unknown similar to *YOR387C* | 3 | 12 | 12 | 1047.3 | 12 | 12 | 858.6 |
| ***YOR387C*** | Unknown similar to VEL1 | 3 | 12 | 12 | 2803 | 12 | 12 | 2612 |

higher specific rate of acetate and glycerol production in the Zn-limited cultures, which may be related to a reduced in vivo activity of Zn-dependent alcohol dehydrogenases.

### 3.4.3 Overall transcriptional responses to Zn limitation

For all six culture conditions described above, microarray analysis was performed on three independent replicate cultures. Statistical analysis (see Material and Methods section) identified 1500 genes that were differentially transcribed in at least one cultivation condition. 381 of these genes responded specifically to Zn-limited growth. Of these Zn-responsive genes, 81 proteins do not yet have an assigned cellular function. The 381 Zn-responsive genes were subjected to a further analysis to identify combinatorial effects of Zn and oxygen availability (Table 3.3). A majority of the genes that showed a transcriptional response to Zn-limitation (248 genes, Modules 3-6 in Table 3.2) did so in an oxygen-dependent manner. The remainder (133 genes, Modules 1-2 in Table 3.3) of the Zn-responsive genes showed a consistent response to Zn limitation that was independent of oxygen availability. The identity and transcript levels of the genes contained in the six modules are available in Supplementary Material 1 of [Nico 07] online. Below, we will analyze these sets of Zn-responsive genes for overrepresentation of genes involved in specific functional categories and/or controlled by specific transcription factors (see Material and Methods section).

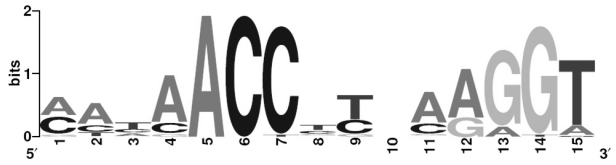



**Figure 3.1 – Consensus ZRE sequence identified by MEME using Module 1 as input.**

### 3.4.4 Zinc homeostasis and the Zap1 regulon

The MIPS functional category "heavy metal transport" was overrepresented among the 93 genes that were transcriptionally upregulated in response to Zn limitation irrespective of oxygen availability (Table 3.3, Module 1). Of the seven genes belonging to this category found in Module 1, six are directly involved in Zn homeostasis. *ZRT1*, encoding the plasma-membrane high-affinity Zn transporter, was strongly induced (average fold-change of 13, Table 3.4). Transcript levels of *ZRT2*, *ZRT3*, *ZRC1* and *ZRG17*, involved in Zn transport and homeostasis, were also increased but to a lesser extent than those of *ZRT1* (fold-changes ranging from two to seven). *FET4* (upregulated 3 to 43 fold under Zn limitation) encodes a protein involved in iron transport that has been demonstrated to also be a physiologically relevant Zn carrier [Wate 02]. The comparison of aerobic and anaerobic cultures confirmed the previously described combinatorial regulation of *FET4* by Zn and oxygen availability [Wate 02]. In addition, a clear hierarchy was observed: while *FET4* was strongly regulated by oxygen availability under Zn sufficient conditions [Wate 02], its transcript level in Zn-limited cultures was consistently high regardless of oxygen supply (Figure 3.2). The transcriptional regulation of these six genes was in agreement with previous studies [Higg 03, Lyon 00], and so was the upregulation of *ZAP1*, the transcriptional activator of these six transporters (8 to 20 fold increase relative to Zn-sufficient cultures). *FRE1*, which also belongs to the "heavy metal transport" category, encodes a protein specifically involved in ferric iron transport [Geor 99, Yun 01]. *FRE1* does not contain a ZRE and its increased transcript levels under Zn-limited conditions suggest an indirect effect.

Previous reports have investigated the role of *MSC2* in Zn transport into the endoplasmic reticulum [Elli 04, Li 01] and have found that mutations in the latter affect the cellular distribution of zinc [Li 01]. In our study, *MSC2* was not found among the genes that were transcriptionally induced under Zn limitation. Instead, its transcript levels remained low under the conditions tested. Consistent with this observation, transcription of *MSC2* was not affected in a *zap1* mutant [Lyon 00]. *ZRG17* encodes a protein that has been proposed to act as a complex with Msc2 [Elli 05, Li 01]. The promoter of *ZRG17* does contain a ZRE and its transcript levels were increased in Zn-limited cultures, suggesting that this protein could be the regulatory sub-unit of the complex.

In an attempt to further define the Zap1 regulon, the promoter regions of the 93 genes that showed a robust, oxygen-independent response to Zn limitation (Module 1, Table 3.3) were searched for overrepresented motifs. The web-based software MEME [Bail 94], which enables unbiased probability-based motif discovery, identified 26 genes with a 15-nucleotide motif that strongly resembled the previously published ZRE Zap1-binding consensus sequence (Figure 3.1, Table 3.4). In agreement with previous reports on Zap1 regulation, all six Zn transporters in Module 1, as well as *ZAP1* itself, harbored this

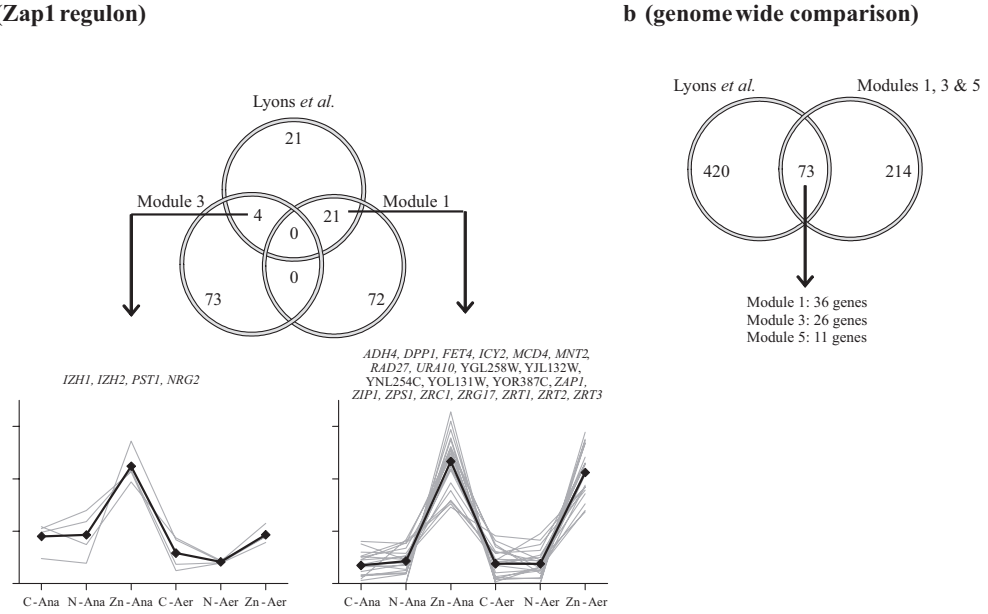



**Figure 3.2 – Venn diagram of chemostat based transcriptome data in comparison with data obtained by Lyons *et al.*.**
**a:** Zap1-regulon (Modules 1 and 3 in comparison with 46 genes from Lyons *et al.*). **b:** genome-wide comparison (Modules 1, 3 and 5 in comparison with all upregulated genes from Lyons *et al.*)

element. Twelve additional genes (Table 3.4) have been previously proven or proposed to be Zap1 targets. An additional 7 genes that harbored the 15-nucleotide motif had not previously been implicated as Zap1 targets [Lyon 00] (Table 3.4). The detailed ZRE sequences and positions are listed in Supplementary Material 2 of [Nico 07] online.

### 3.4.5 Comparison with previous Zn-related transcriptome studies

Two previous transcriptome studies investigated yeast adaptation to Zn depletion in batch cultures of an industrial [Higg 03] and a laboratory strain of *S. cerevisiae* [Lyon 00]. Using maltose-grown cultures, Higgins *et al.* observed a downregulation of maltose-permease and maltase genes (*MAL12*, *MAL32* and *MAL31*) in Zn-depleted cultures. In the present study, growth on glucose resulted in the absence of *MAL* gene transcripts, thus masking transcriptional responses of these genes to Zn availability. Lyons *et al.* identified a Zap1 regulon consisting of 46 genes by comparing the transcriptional responses to Zn depletion of a *zap1Δ* mutant and its parental strain. Three of these 46 genes (*COS2*, *COS4* and *COS6*) were not represented on the microarrays used in our study. Of the remaining 43 genes, 25 showed increased transcript levels in Zn-limited chemostat cultures (Figure 3.2a). The large majority of these (21 genes) were consistently induced in response to Zn limitation irrespective of oxygen availability (Module 1, Table 3.3; Figure 3.2a). MEME failed to identify a ZRE sequence in 3 of these 21 genes (*RAD27*, *YJL132W* and *YOL131W*), which are therefore absent from Table 3.4. Four

genes from the Zap1 regulon defined by Lyons *et al.* (*IZH1*, *IZH2*, *NRG2* and *PST1*) were found in Module 3 (Table 3.3), indicating that their transcription was induced under Zn limitation but only when oxygen was absent. Their identification by Lyons *et al.* may have been caused by the poor oxygen transfer characteristics of shake flask cultures [Schu 64, Gupt 03, McDa 65]. Two additional genes (*ADE17* and *GPG1*) identified as Zap1-targets by Lyons *et al.* were upregulated in Zn-limited chemostat cultures, however their expression resulted from an intricate regulation by Zn, glucose and oxygen availability. Both genes responded to zinc-limitation under aerobic and anaerobic conditions. However, they also responded to limiting glucose supply, but this response was oxygen specific; while *ADE17* was upregulated under glucose-limitation in the presence of oxygen, *GPG1* expression increased under glucose limited anaerobic growth. The remaining 16 of the 43 genes identified as Zap1 targets by Lyons *et al.* and included on our microarrays did not respond to Zn availability in our chemostat study.

Eight potential Zap1-targets identified in the present study (Table 3.4) were not found in the study of Lyons *et al.*. However, of these 8 genes, *HOR2* and *TEX1* were found to be transcriptionally induced by Zn depletion in their study. Furthermore, Zap1 was shown to bind *TEX1* on ChIP on chip experiments [Harb 04]. Seven genes (*HOR2*, *FLO11*, *KTR6*, *KTI12*, *VTC3*, *MUP1* and *YER130C*; Table 3.4) are here for the first time proposed to be Zap1 targets. *HOR2* encodes a glycerol-3-phosphate phosphatase involved in glycerol biosynthesis [Pahl 01], which may account for the slightly, but significantly (T-test $P$-value $< 0.05$) elevated glycerol production observed under zinc-limited growth. *VTC3* encodes a vacuolar transport chaperone involved in inorganic ion transport [Cohe 99]. Although it has been shown to be involved in polyphosphate transport, it may also participate in vacuolar Zn transport [Ogaw 00]. Alternatively, Zn may be involved in polyphosphate accumulation or react with polyphosphates. Like the previously identified Zap1 target *MNT2* [Lyon 00] (Table 3.4), *KTR6* encodes a mannosyl transferase involved in glycosylation of cell wall proteins [Luss 97]. It can be speculated that they play a role in mannosylation of Zn-scavenging cell wall proteins. For instance *ZPS1*, a Zap1 target also upregulated under Zn limitation (Table 3.4), encodes a cell wall mannoprotein with high similarity to Zn metalloproteinases from filamentous fungi [Lamb 03, Lyon 00]. The yeast cell wall, and more specifically mannoproteins, has been shown to fix a substantial fraction of the cellular zinc [DeNi 06]. Zinc fixation by mannoproteins may represent an efficient mechanism to scavenge low zinc concentrations [Moch 96]. The upregulation of mannoproteins such as *ZPS1* under zinc limitation would support this zinc scavenging function of the cell wall. The consistent upregulation of *FLO11*, *KTI12*, *MUP1* and *YER130C* in Zn-limited cultures and the presence of a ZRE-like motif in their promoters suggest that the encoded proteins have some as yet unknown role under Zn-limited conditions, too. For example, Flo11 is known to play an essential role in biofilm formation, filamentation and invasive growth [Lo 98]. In addition, studies on *Candida albicans* have demonstrated that dimorphic switching from budding growth to mycelium formation is regulated by zinc [Soll 81, Bede 79]. However, in the present study, we did not observe any difference in morphology between the different culture conditions.

When the 289 genes in Modules 1, 3 and 5 that were induced under Zn limitation in chemostat cultures, either in an oxygen-dependent or in an oxygen-independent manner, were compared to the 493 genes that were induced upon Zn depletion in shake flasks [Lyon 00], 73 genes overlapped between the two studies. These were for the most part clustered in Modules 1 and 3 (Figure 3.2b, Supplementary Material 3 of [Nico 07]

online). Only a small overlap was observed with Module 5 (representing only 8% of the genes in this module), which includes genes that are only induced by Zn limitation under aerobic conditions. As mentioned above, this small overlap may reflect a limiting oxygen supply in the shake flask studies.

### 3.4.6 Transcriptional regulation of structural genes for zincdependent proteins

*S. cerevisiae* contains multiple alcohol dehydrogenases. While the enzymes encoded by *ADH1*, *2*, *3* and *5* all require Zn as a cofactor, Adh4 uses Mg. *ADH4* has been shown to be regulated by Zap1, while expression of the Zn-requiring isoenzymes has been reported to be decreased upon Zn depletion (presumably via Rap1) [Bird 06]. In agreement with earlier findings, *ADH4* was strongly upregulated in response to Zn limitation irrespective of the aeration conditions. Transcript levels of other, Zn-dependent alcohol dehydrogenase genes were either unchanged or reduced. In addition to alcohol dehydrogenases, many other yeast proteins use Zn as structural component or cofactor. Regalla and Lyons [Mart 03, Rega 06] separated the Zn dependent protein in two distinct classes, i) the proteins that use zinc in a catalytic capacity (105 genes) and ii) the proteins with a structural Zn binding domain (360 genes). Of 105 *S. cerevisiae* proteins that use Zn as a cofactor [Mart 03, Rega 06], none of the structural genes were found to be transcriptionally regulated in response to Zn availability in chemostat cultures (with the clear exception of alcohol dehydrogenases). On the other hand, out of the 360 *S. cerevisiae* proteins that contain a structural Zn binding domain, 16 genes were upregulated in response to Zn limitation (Modules 1, 2 and 5) while 7 were downregulated (Modules 2, 4 and 6). Most of these Zn-responsive genes encoded proteins that have a function in nucleic acid binding (transcription factors, chromatin reorganizing activity, mRNA binding). The two homologous transcription factors Met31 and Met32 that induce the expression of genes involved in methionine biosynthesis were only affected by Zn availability in the presence of oxygen. While *MET32* expression increased two-fold, *MET31* expression decreased two-fold. These changes in gene expression probably resulted in modifications of the transcriptional regulation of these transcriptional activators as their target genes displayed a slightly higher expression under conditions of aerobic Zn limitation. This antagonistic regulation of *MET31* and *MET32* remains difficult to relate to Zn supply as both proteins contain two Zn finger domains and do not have a different Zn content.
In agreement with previous reports [Wu 07], *SOD1*, which encodes the cytosolic Zn-Cu superoxide dismutase, showed a two fold reduction of its transcript level under conditions of low Zn supply. However, *SOD2*, which encodes mitochondrial manganese-containing superoxide dismutase, did not show an increased transcript level in Zn-limited cultures. In fact, *SOD2*, which was only transcribed in aerobic cultures, was also downregulated by ca. two-fold under Zn limitation. As proposed previously [Wu 07], reduced expression of superoxide dismutase may affect resistance to oxidative stress. A more direct involvement of zinc in oxidative stress resistance was previously suggested via the transcriptional regulation of *TSA1*, encoding a Zn-dependent peroxiredoxin, by Zap1 [Wu 07]. Unfortunately, in our experiments *TSA1* expression was independent of zinc and oxygen availability. This difference with earlier work may be attributed either to the difference between complete Zn depletion [Wu 07] and Zn-limited growth (this study) or to a different strain background. However, close scrutiny of the transcript levels revealed

no oxidative stress response (*AAD3*, *AAD6*, *AAD10*, *AAD14*, *AAD15*, *ATR1*, *CCP1*, *GTT2*, *GRE2*, *LYS20*, *OYE2*, *OYE3*, *TRR1*, *TRX2*, *YDR453C*, *YLR460C*, *YNL134C*, *YMR318C* and *YML131W*) [Koer 02]. Although the transcript levels of both *SOD1* and *SOD2* were reduced, their levels (748 and 295 respectively under aerobic zinc-limitation) may still be high enough to enable efficient processing of ROS and thereby to prevent oxidative stress.
Finally, while we cannot exclude the possibility that Zn sparing and/or mobilisation mechanisms occurs at (a) post-transcriptional level(s), these results indicate that a general 'Zn sparing' regulation at the transcriptional level is most probably absent in *S. cerevisiae*. The exceptions of alcohol dehydrogenase and superoxide dismutase may be related to the relative abundance of these proteins and their pivotal role in fermentative and respiratory metabolism, respectively.

### 3.4.7 Combinatorial response of mitochondrial function to oxygen and zinc availability

Aerobic Zn-limitation of *S. cerevisiae* resulted in the upregulation of 119 genes and the downregulation of 16 genes (Table 3.3, Modules 5 and 6). However, hypergeometric distribution analysis did not reveal clear trends in the identity and function of these oxygen-responsive proteins. In order to better investigate the potential synergetic effects between oxygen and zinc availability, different discretized patterns were considered. As described in Figure 3.4 for the aerobically upregulated genes, the applied constraints selected genes for which the expression under carbon limitation was unaffected by oxygen, the expression under nitrogen limitation was also oxygen-insensitive, but for which the response to zinc limitation was oxygen-dependent. 196 genes respecting these constraints were identified, 130 being upregulated in the presence of oxygen in a Zn-dependent manner and 66 downregulated (given in Supplementary Material 4 of [Nico 07] online). Fisher's exact statistics was then applied to search for overrepresentation of genes involved in specific functional categories and/or controlled by specific transcription factors. While no enrichment was found within the genes that were downregulated, the module containing the upregulated showed interesting trends. This module was characterized by enrichment for two functional categories: 'respiration' (10 genes) and 'mitochondrial biogenesis' (14 genes). The category of 'respiration' comprised genes encoding various subunits of the $F_o$ (*ATP4*, *ATP14*, *ATP18* and *ATP20*) and $F_1$ (*ATP3* and *ATP15*) domains of mitochondrial ATP synthase [Deve 00] but also *COX23*, *COX14*, *MAM33* and *MBA1* involved in the assembly of respiratory complexes in mitochondria [Barr 04, Gler 95, Muta 97, Rep 96]. The relation between Zn availability and these proteins remains unclear, although cytochrome *c* oxydase activity has been shown to be inhibited by Zn. Most of the genes in the 'mitochondrial biogenesis' category encoded mitochondrial ribosomal proteins (*MRPL10*, *MRPL11*, *MRPL37*, *MNP1*, *RSM19* and *MRPS16*), but also *MSS116*, a gene involved in the splicing of mitochondrial group I and II introns [Huan 05]. Finally, also *TIM10*/*MRS11* responded synergistically to Zn and oxygen availability. *TIM10* encodes a protein involved in the translocation of mitochondrial proteins from the cytoplasm to the mitochondria. For instance Aac1 and Aac2, encoding ADP/ATP mitochondrial carrier cannot be translocated in a *tim10* mutant [Vasi 04]. This translocation process, also identified in plant [Bhus 03], requires Zn [Lu 05]. The present study reveals a more important role for Zn in mitochondrial function and biogenesis than so far assumed. Although still not clearly understood this

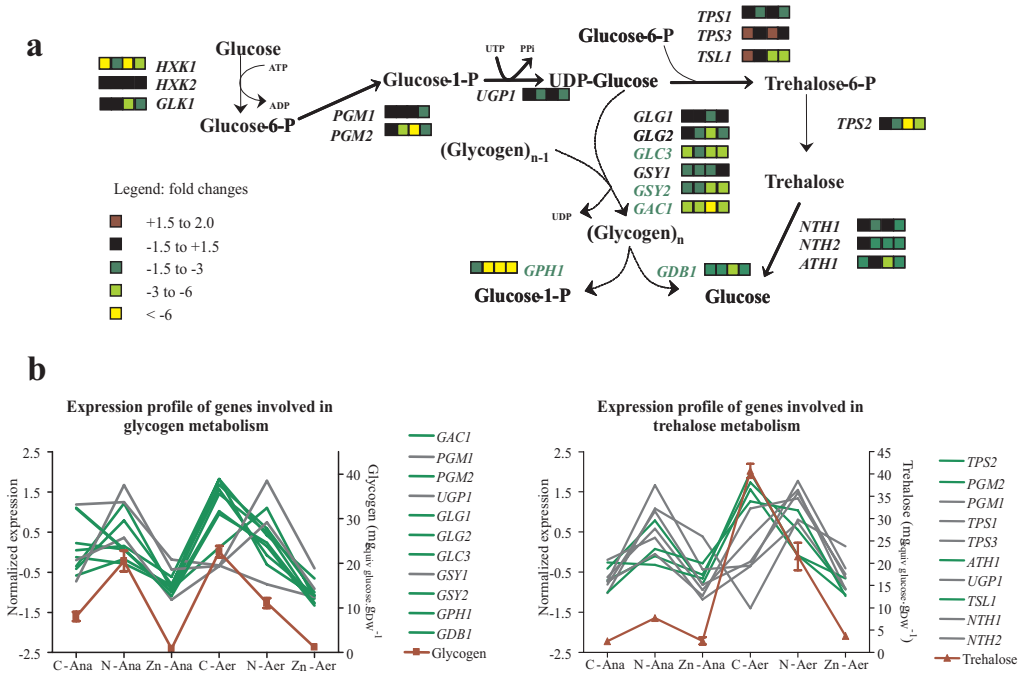



**Figure 3.3 – Transcriptional response of glycogen and trehalose metabolism genes.**
**a**: Glycogen and trehalose metabolism in *S. cerevisiae*. Genes indicated in green are clustered in Module 2. The four boxes indicate the following fold-changes from left to right: zinc vs carbon anaerobic, zinc vs nitrogen anaerobic, zinc vs carbon aerobic, zinc vs nitrogen aerobic. Intensities of fold changes are indicated by the color map in the legend. **b**: Normalized expression profile of genes involved in glycogen and trehalose metabolism and intracellular glycogen and trehalose concentrations.

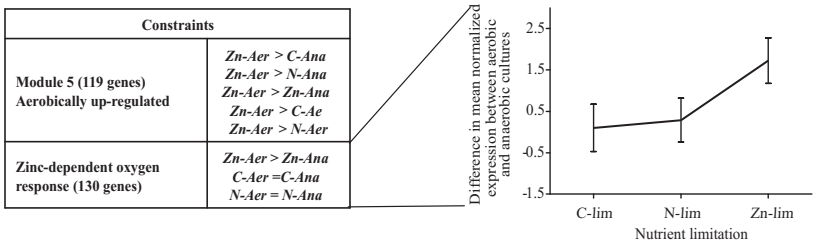



**Figure 3.4 – Combinatorial regulation of gene expression by Zn and oxygen availability.**
Given are the constraints for the selection of the discretized patterns and average expression profile of these selected patterns. Only the oxygen-induced genes are represented.

role could, at least in part, explain the higher Zn requirement for cells grown in the presence of oxygen, condition where mitochondria are essential for respiration.

### 3.4.8 Zn limitation and storage carbohydrate metabolism

Genes from both glycogen biosynthesis (*GSY2*, *GAC1*, *GLC3*) and degradation pathways (*GDB1*, *GPH1*) were downregulated by up to 22-fold in Zn-limited chemostat cultures, regardless of oxygen availability (Figure 3.3a). Several additional genes involved in glycogen metabolism that did not pass the very stringent statistical test used in genome-wide analysis, displayed a decreased expression upon closer inspection (Figure 3.3a). To investigate whether these transcriptional modifications resulted in phenotypic differences, glycogen contents were analyzed in the chemostat cultures on which the transcriptome analyses had been performed (Figure 3.3b). Indeed, glycogen accumulation was strongly (10 to 20-fold) reduced in Zn-limited cultures. Genes involved in glycogen metabolism are known to be transcriptionally regulated in response to a wide variety of environmental conditions and signaling pathways [Enja 04] (temperature, nutrient supply, oxidative stress). This regulation is mediated by the general environmental stress response (ESR) and HOG pathways [Fran 01, Gasc 00]. However, no other target genes of these signaling pathways were found to be differentially transcribed in response to Zn limitation. This suggests that the regulation of glycogen metabolism by Zn occurs via another, hitherto unknown signal transduction mechanism.
Several genes involved in trehalose metabolism were also significantly downregulated in Zn limited cultures (*PGM1*, *PGM2*, *TPS1*, *TPS2* and *TPS3*, see Figure 3.3a). These downregulations coincided with substantially lower trehalose biomass contents, an effect that was most pronounced in aerobic cultures (Figure 3.3b). These results clearly demonstrated, for the first time the impact of Zn availability on reserve carbohydrate accumulation. As the genes involved in glycogen and trehalose metabolism do not contain ZREs, their transcriptional regulation is unlikely to be directly mediated by Zap1. In addition, their downregulation probably occurs via a STRE-independent mechanism (we did not find overrepresented STRE in the promoters of downregulated genes). Among the above-mentioned Zap1 regulon, *YER130C*, encoding a protein of unknown function containing two tandem Zn-finger domains, was upregulated under zinc limitation. This putative transcription factor may be involved in a Zap1-dependent regulation of genes involved in trehalose and glycogen and is an interesting candidate for further functional analysis.

## 3.5 Discussion

### 3.5.1 Analysis of Zn limitation in chemostat cultures

The unique option of chemostat cultures to control specific growth rate prevented occurrence of specific-growth-rate-related responses. For example, in a previous study in batch cultures of *S. cerevisiae* [Higg 03], the observed downregulation of ribosomal proteins in low-Zn cultures is likely to have been caused by a decrease in specific growth rate rather than directly by Zn depletion.
The use of different aeration regimes showed that yeast responses to Zn limitation are strongly context dependent. This notwithstanding, a set of genes was identified whose specific transcriptional regulation by Zn availability was independent of the oxygen supply. This enabled us to propose a more precise definition of the Zap1 regulon. Most of these 26 potential Zap1 targets overlapped with those proposed in a previous batch-cultivation study [Lyon 00]. The present study demonstrated that responses of several of

the previously identified putative Zap1 targets were not Zn-specific. Instead, they were synergistically or antagonistically regulated by carbon, nitrogen and/or oxygen supply. As compared to the transcriptional responses observed in chemostat cultures under other nutrient limitations [Boer 03, Boer 07, Dara 03, Tai 05], transcriptional responses to Zn limitation were strikingly pleiotropic. Genes involved in a large variety of cellular functions, apparently unrelated to Zn availability, showed marked differences to Zn limitation. Statistical analysis of coregulated genes identified only a very limited number of overrepresented functional categories or DNA binding proteins, with the clear exception of the Zap1 regulon. These observations suggest that the only direct effect of Zn limitation on transcriptional regulation is mediated by Zap1. Although no concerted transcriptional regulation was observed for genes encoding proteins that contain Zn as a catalytic or structural component, Zn availability is likely to influence the in vivo activity of such proteins, many of which are transcription factors. The apparently 'scattered' transcriptional responses to Zn limitation may further be due to the fact that new roles of Zn in yeast physiology continue to be discovered. For instance, the involvement of Zn in protein translocation by the Tim10/Tim9 complex has only been recently revealed [Lu 05].

### 3.5.2 Effects of Zn limitation on storage carbohydrate accumulation: a possible cause for stuck fermentations in beer fermentation?

Zn used by yeast during the beer fermentation process comes from barley malt and is extracted during the mashing procedure (starch conversion and extraction). However, Zn content varies largely between fermentations as its concentration is dependent on the crop quality [Houg 82] and is partly removed from wort during lautering [Kred 99] or wort separation. Insufficient Zn supply during brewing results in 'sluggish' fermentations characterized by a slow fermentation rate [Brom 97]. The metabolic and/or regulatory processes in yeast cells that underlie such retarded fermentations are incompletely understood. Yeast crops are commonly re-used four to ten times for inoculating succeeding brews and are generally stored around 2°C under starvation [Mart 03]. Under such conditions, high reserve carbohydrates contents have been shown to be critical for the survival and recovery of metabolic activity of yeast [Mart 03]. To our knowledge, no published study has investigated how storage carbohydrate metabolism might be affected by Zn deficiency. This present study demonstrates for the first time that Zn limitation causes a strong transcriptional downregulation of genes involved in reserve carbohydrate accumulation. The physiological relevance of this response was verified by analysis of intracellular glycogen and trehalose contents, which were strongly reduced in Zn limited cultures. Comparative studies with nitrogen-limited cultures showed that the decreased accumulation of storage carbohydrates was specific for Zn limitation and not merely a consequence of glucose-excess conditions. Furthermore, this effect was independent of the aeration of the cultures and the expression profiles of several genes involved in reserve carbohydrate metabolism perfectly matched the profile of trehalose and glycogen accumulation (Figure 3.3b).

While our hypothesis remains to be tested under brewing conditions and with brewing strains of *S. cerevisiae*, it seems highly probable that the fermentation performance of Zn-limited brewers' yeast will be strongly compromised. Additionally, follow-up research should focus on the molecular mechanisms that link reserve carbohydrate metabolism

and Zn availability.

### 3.5.3 Potential implication of Zn-limitation for flavor formation

Another consequence of limiting Zn supply during the course of beer fermentation might be related to flavor formation. Indeed, three genes involved in the biosynthesis of the branched-chain amino acids leucine, valine and isoleucine (*ILV2*, *ILV3* and *BAT2*) were consistently downregulated under Zn-limited growth, both in the presence and in the absence of oxygen (Table 3.3). The flux through the branched chain amino acids synthetic pathways has been shown to have a positive impact on desirable flavor compound production, such as isoamyl acetate and isobutyl acetate [Lee 95] and Zn supplementation to wort results in increased production of the acetate esters of higher alcohols [Hodg 90]. The present data suggests that this effect of zinc availability on flavor formation may be mediated by the transcriptional regulation of *ILV2*, *ILV3* and *BAT2*. Maintaining a sufficiently high zinc level during beer fermentation is clearly critical to maintain the desired balance between several flavor compounds.

### 3.5.4 Signature transcripts for diagnosing Zn bio-availability in industrial media

In complex industrial fermentation media such as wort or other plant biomass hydrolysates, Zn can form complexes with several medium components, thereby reducing its bioavailability for yeast [Kred 99, Jaco 79, Kuhb 06]. This limits the relevance of chemical analyses of the Zn content to test the bioavailability of zinc in wort and other industrial media. Addition of Zn in the form of salt or trub is a common practice to prevent Zn depletion during the brewing process [Taid 00]. Especially in beer brewing, this is not risk-free as excess Zn leads to the modification of flavor compound formation [Dufo 03]. Molecular markers can be used to monitor fermentation processes through transcript profiling [Higg 03]. For such diagnostic purposes, it would be preferable to construct small, cost-effective microarrays that contain a limited number of 'signature transcripts'. A prerequisite of these signature transcripts is that they are specific to one environmental parameter and show a robust response in various environmental (process) contexts. Comparison of multiple chemostat regimes enabled the identification of such Zn-specific signature transcripts. For instance, *ZAP1* and *ZRT1* would be very good signature transcripts. Also *YOR387C* and *YGL258W*, encoding proteins that have not been characterized yet and that have been previously proposed as potential signature transcripts for Zn depletion [Lyon 00, Higg 03], were specifically and consistently induced under Zn limitation in chemostat cultures. Conversely *NRG2* and *PST1*, potential Zap1-targets [Lyon 00] were here shown to be also regulated by oxygen availability and are therefore not recommended for diagnostic purposes.

## 3.6 Appendix

*Constraints imposed to group zinc responsive genes into modules.*

Let the discretized expression pattern of a gene be denoted by vector $\mathbf{x}$ of length six. The values of the elements of $\mathbf{x}$ can either be 0 (no differential expression), 1 (up-regulated)

or -1 (down-regulated). The elements of $\mathbf{x}$ correspond to the cultivation conditions as follows:

| $\mathbf{x}(1)$ | $\mathbf{x}(2)$ | $\mathbf{x}(3)$ | $\mathbf{x}(4)$ | $\mathbf{x}(5)$ | $\mathbf{x}(6)$ |
|---|---|---|---|---|---|
| C-Ana | N-Ana | Zn-Ana | C-Aer | N-Aer | Zn-Aer |

Below, we state the constraints on $\mathbf{x}$ that must be satisfied in order for a gene to be part of a particular module. Note that all constraints must be met to suffice.

| Module 1 | Upregulated regardless of aeration |
|---|---|
| constraints: | $\mathbf{x}(3) > \mathbf{x}(1)$ |
| | $\mathbf{x}(3) > \mathbf{x}(2)$ |
| | $\mathbf{x}(3) > \mathbf{x}(4)$ |
| | $\mathbf{x}(3) > \mathbf{x}(5)$ |
| | $\mathbf{x}(6) > \mathbf{x}(1)$ |
| | $\mathbf{x}(6) > \mathbf{x}(2)$ |
| | $\mathbf{x}(6) > \mathbf{x}(4)$ |
| | $\mathbf{x}(6) > \mathbf{x}(5)$ |

| Module 2 | Downregulated regardless of aeration |
|---|---|
| constraints: | $\mathbf{x}(3) < \mathbf{x}(1)$ |
| | $\mathbf{x}(3) < \mathbf{x}(2)$ |
| | $\mathbf{x}(3) < \mathbf{x}(4)$ |
| | $\mathbf{x}(3) < \mathbf{x}(5)$ |
| | $\mathbf{x}(6) < \mathbf{x}(1)$ |
| | $\mathbf{x}(6) < \mathbf{x}(2)$ |
| | $\mathbf{x}(6) < \mathbf{x}(4)$ |
| | $\mathbf{x}(6) < \mathbf{x}(5)$ |

| Module 3 | Anaerobically upregulated |
|---|---|
| constraints: | $\mathbf{x}(3) > \mathbf{x}(1)$ |
| | $\mathbf{x}(3) > \mathbf{x}(2)$ |
| | $\mathbf{x}(3) > \mathbf{x}(4)$ |
| | $\mathbf{x}(3) > \mathbf{x}(5)$ |
| | $\mathbf{x}(3) > \mathbf{x}(6)$ |

| Module 4 | Anaerobically downregulated |
|---|---|
| constraints: | $\mathbf{x}(3) < \mathbf{x}(1)$ |
| | $\mathbf{x}(3) < \mathbf{x}(2)$ |
| | $\mathbf{x}(3) < \mathbf{x}(4)$ |
| | $\mathbf{x}(3) < \mathbf{x}(5)$ |
| | $\mathbf{x}(3) < \mathbf{x}(6)$ |

| Module 5 | Aerobically upregulated |
|---|---|
| constraints: | $\mathbf{x}(6) > \mathbf{x}(1)$ |
| | $\mathbf{x}(6) > \mathbf{x}(2)$ |
| | $\mathbf{x}(6) > \mathbf{x}(3)$ |
| | $\mathbf{x}(6) > \mathbf{x}(4)$ |
| | $\mathbf{x}(6) > \mathbf{x}(5)$ |

| Module 6 | Aerobically downregulated |
|---|---|
| constraints: | $\mathbf{x}(6) < \mathbf{x}(1)$ |
| | $\mathbf{x}(6) < \mathbf{x}(2)$ |
| | $\mathbf{x}(6) < \mathbf{x}(3)$ |
| | $\mathbf{x}(6) < \mathbf{x}(4)$ |
| | $\mathbf{x}(6) < \mathbf{x}(5)$ |

# CHAPTER 4

# EXPLOITING THE COMBINATORIAL SETUP

In this chapter a microarray dataset of eight conditions is analyzed. In this case, there are four different nutrient limitations; carbon, nitrogen and phosphorus and sulfur, grown both aerobically and anaerobically. Using a regression strategy the effect of oxygen presence on the expression of each gene is modeled as a linear effect (having both an additive and multiplicative component). The estimated parameters (offset and slope) are employed to 'correct for' the oxygen effect in the expression pattern. A discretization procedure is designed to represent each gene with a tertiary vector of length nine, where the last entry is that of the oxygen effect. Genes are clustered based on their discretized representations and related to TF binding data to infer the (combinatorial) effect of oxygen availability and nutrient limitations on TF activity. The inclusion of the cultivation parameters in uncovering regulatory modules and TF activity leads to a more valuable regulatory network that resultantly provides detailed insight in yeasts respiration and metabolism. The power of this approach in recognizing the individual and combinatorial effects of nutrient-limitations and oxygen presence is reflected in the results that strengthen and broaden the existing knowledge on regulatory mechanisms. For example, our results confirm the established role of TF Hap4 in both aerobic regulation and glucose derepression.

This chapter is published as:

'Exploiting combinatorial cultivation conditions to infer transcriptional regulation'

Theo A. Knijnenburg, Johannes H. de Winde, Jean-Marc Daran, Pascale Daran-Lapujade, Jack T. Pronk, Marcel J. T. Reinders and Lodewyk F. A. Wessels

BMC Genomics, Volume 8 No 25, January 2007

## 4.1 Abstract

Regulatory networks often employ the model that attributes changes in gene expression levels, as observed across different cellular conditions, to changes in the activity of transcription factors (TFs). Although the actual conditions that trigger a change in TF activity should form an integral part of the generated regulatory network, they are usually lacking. This is due to the fact that the large heterogeneity in the employed conditions and the continuous changes in environmental parameters in the often used shake-flask cultures, prevent the unambiguous modeling of the cultivation conditions within the computational framework.
We designed an experimental setup that allows us to explicitly model the cultivation conditions and use these to infer the activity of TFs. The yeast *Saccharomyces cerevisiae* was cultivated under four different nutrient limitations in both aerobic and anaerobic chemostat cultures. In the chemostats, environmental and growth parameters are accurately controlled. Consequently, the measured transcriptional response can be directly correlated with changes in the limited nutrient or oxygen concentration. We devised a tailor-made computational approach that exploits the systematic setup of the cultivation conditions in order to identify the individual and combined effects of nutrient limitations and oxygen availability on expression behavior and TF activity.
Incorporating the actual growth conditions when inferring regulatory relationships provides detailed insight in the functionality of the TFs that are triggered by changes in the employed cultivation conditions. For example, our results confirm the established role of TF Hap4 in both aerobic regulation and glucose derepression. Among the numerous inferred condition-specific regulatory associations between gene sets and TFs, also many novel putative regulatory mechanisms, such as the possible role of Tye7 in sulfur metabolism, were identified.

## 4.2 Introduction

The simple and often used biological model to unravel transcriptional regulation ascribes the change in gene expression levels, as observed between different cellular conditions, to changes in the activity of transcription factors (TFs). Change of the transcriptional activity of a TF is one of the means by which an organism adapts to changes in the extracellular environment. A substantial amount of research has employed this model to infer regulatory networks by integrating gene expression data, sequence data (to detect the *cis*-regulatory binding sites of TFs), e.g. [Roth 98, Buss 01, Kell 03], and/or TF binding data, e.g. [Lee 02, Bar 03, Lusc 04]. For an overview see Banerjee and Zhang [Bane 02], Siggia [Sigg 05] and Blais and Dynlacht [Blai 05]. In most cases, the generated regulatory networks are derived from large microarray compendia. Notwithstanding the many advantages of such approaches, two main drawbacks can be identified. Firstly, these compendia gather very heterogeneous gene expression data derived from various culture conditions (media, pH, temperature, etc.) that, in a large majority of the cases, solely compare the culture conditions to their direct condition-specific references. Different cultivation conditions within the compendium can, therefore, hardly be compared. Secondly, the interpretation of transcriptome data obtained from the generally employed shake-flask cultivations is likely to be complicated by differences in specific growth rate, carbon catabolite repression, nitrogen catabolite repression, and more generally contin-

uous changes in environmental conditions. This prevents the establishment of a direct link between the activity of TFs and specific growth conditions.
A frequently employed approach links a TF to a module, i.e. a set of co-expressed genes, based on TF binding data or promoter analysis. Enrichment of functional categories (such as GO [Ashb 00] and MIPS [Mewe 97]) within the module provides clues about the function of the TFs associated with the module. Although this can provide a global view of the transcriptional role of a TF, we are convinced that the precise conditions or perturbations that trigger a change in the activity of TFs should be an integral part of the generated regulatory network.
To this end, we designed an experimental setup that allowed us to explicitly model the cultivation conditions and use these to infer the activity of TFs. To achieve this, we employed chemostat cultures that enable the cultivation of micro-organisms under tightly defined environmental conditions. Chemostat cultures are superior to the shake-flask cultures in both accuracy and reproducibility [Pipe 02]. In a chemostat, culture broth (including biomass) is continuously replaced by fresh medium at a fixed and accurately determined dilution rate. When the dilution rate is lower than $\mu_{max}$, the maximal specific growth rate of the micro-organism, a steady-state situation will be established in which the specific growth rate equals the dilution rate. In such a steady-state chemostat culture, $\mu$ is controlled by the (low) residual concentration of a single growth-limiting nutrient. In this research, microarrays were employed to measure the genome-wide transcriptional response of the yeast *Saccharomyces cerevisiae* to growth limitation by four different macronutrients (carbon, nitrogen, phosphorus, and sulfur) in both aerobic and anaerobic chemostat cultures (Figure 4.1) [Tai 05]. Except for the different nutrient

| Experiment | Nutrient Limitation | | | | Oxygen supply | |
|---|---|---|---|---|---|---|
| | Carbon | Nitrogen | Phosphorus | Sulfur | Aerobic | Anaerobic |
| 1. ClimAer | ■ | | | | ■ | |
| 2. NlimAer | | ■ | | | ■ | |
| 3. PlimAer | | | ■ | | ■ | |
| 4. SlimAer | | | | ■ | ■ | |
| 5. ClimAna | ■ | | | | | ■ |
| 6. NlimAna | | ■ | | | | ■ |
| 7. PlimAna | | | ■ | | | ■ |
| 8. SlimAna | | | | ■ | | ■ |

**Figure 4.1 – Schematic overview of the combinatorial cultivation conditions.** Black squares indicate the employed nutrient limitation and oxygen supply.

limitations and oxygen availability, all other culture parameters (such as growth rate, pH, temperature, etc.) were kept constant throughout the different experiments. Thus, changes in gene expression levels can solely be attributed to the different nutrient limitations and the oxygen regime. We devised a computational approach that exploits the interrelatedness between the conditions in order to identify the individual and combined effects of nutrient limitations and oxygen availability on expression behavior and TF activity. The inclusion of the growth conditions in the analysis allows for the identification of direct links between the cultivation conditions, TFs triggered by specific cultivation conditions and the targets of these TFs.

## 4.3 Results

### 4.3.1 Overview of the computational approach

From the continuous expression levels measured across the cultivation conditions we derive a discretized representation of the expression behavior for each gene. This representation indicates up- or downregulation as a consequence of the individual or combined effects of the nutrient limitations and oxygen availability. Here, we exploit the combinatorial setup of the cultivation conditions to recognize and dissect the effect of the presence of oxygen on the expression levels of a gene. More specifically, we employ a regression strategy to detect, model and correct for the effect of oxygen presence. This procedure is outlined in Figure 4.2 and explained in detail in the Methods section.
Modules are generated by clustering genes with identical expression representations (Figure 4.3). Next, we integrate TF binding data [Harb 04] to assess whether a TF or a pair of TFs binds the promoter regions of a module much more frequently than would be expected by chance. A significant relationship between a module and a TF suggests that the TF is (partly) responsible for the expression behavior of that particular module. Since the expression behavior of a module reveals under which combination of cultivation conditions the genes are up- or downregulated, we are not only able to relate TFs to the groups of genes that they presumably regulate, but also to the precise environmental conditions that trigger their activity to perform their regulatory role.

### 4.3.2 Overview of the uncovered regulatory relationships

The TF circle (Figure 4.4) depicts an overview of all the TFs, which are significantly related to one or more modules. In addition, pairs of TFs that can bind the promoter region of the genes in a module significantly often, are connected by a solid line. In the TF circle, the modules and their associated TFs are categorized according to the cultivation parameters under which the genes in the module are differentially regulated, i.e. where the discretized representation differs from zero. This arrangement is given by the color coding of the segments in the circle. From this it is clear which cultivation parameters affect the activity of a TF. Additional information concerning enrichment of gene annotation categories and results of motif discovery in promoter regions of the genes within the modules can be found in Table 4.1 and more comprehensively in Additional file 1 of [Knij 07] online.

**Legend to Figure 4.2. Procedure to derive the discretized representation of a gene.**
**a:** Examination of the expression levels under the eight cultivation conditions led to the observation that for many genes the expression pattern across the four nutrient limitations when grown aerobically is a scaled and offset version of its anaerobic counterpart. (Permutation tests were performed to confirm this notion (Additional file 3 of [Knij 07] online)). **b:** This "global oxygen effect", i.e. the effect that presence of oxygen has on the expression levels across all or most of the nutrient limitations, is modeled as a linear relationship and estimated using a regression strategy. **c:** The estimated regression parameters (slope and offset) are employed to isolate the oxygen effect by transforming the aerobic expression values. Discretization of this pattern allows for identification of up- or downregulation as a consequence of specific nutrient limitations and possible nutrient-limitation-specific effects of oxygen presence. **d:** Pairwise T-tests are performed to compare the original aerobic and anaerobic expression values for each nutrient limitation. The results of these tests are combined to detect possible consistent and significant higher or lower expression as a consequence of oxygen presence. **e:** The derived discrete representation of the expression of a gene is visualized in a nine-bit ternary (-1,0,1) vector.

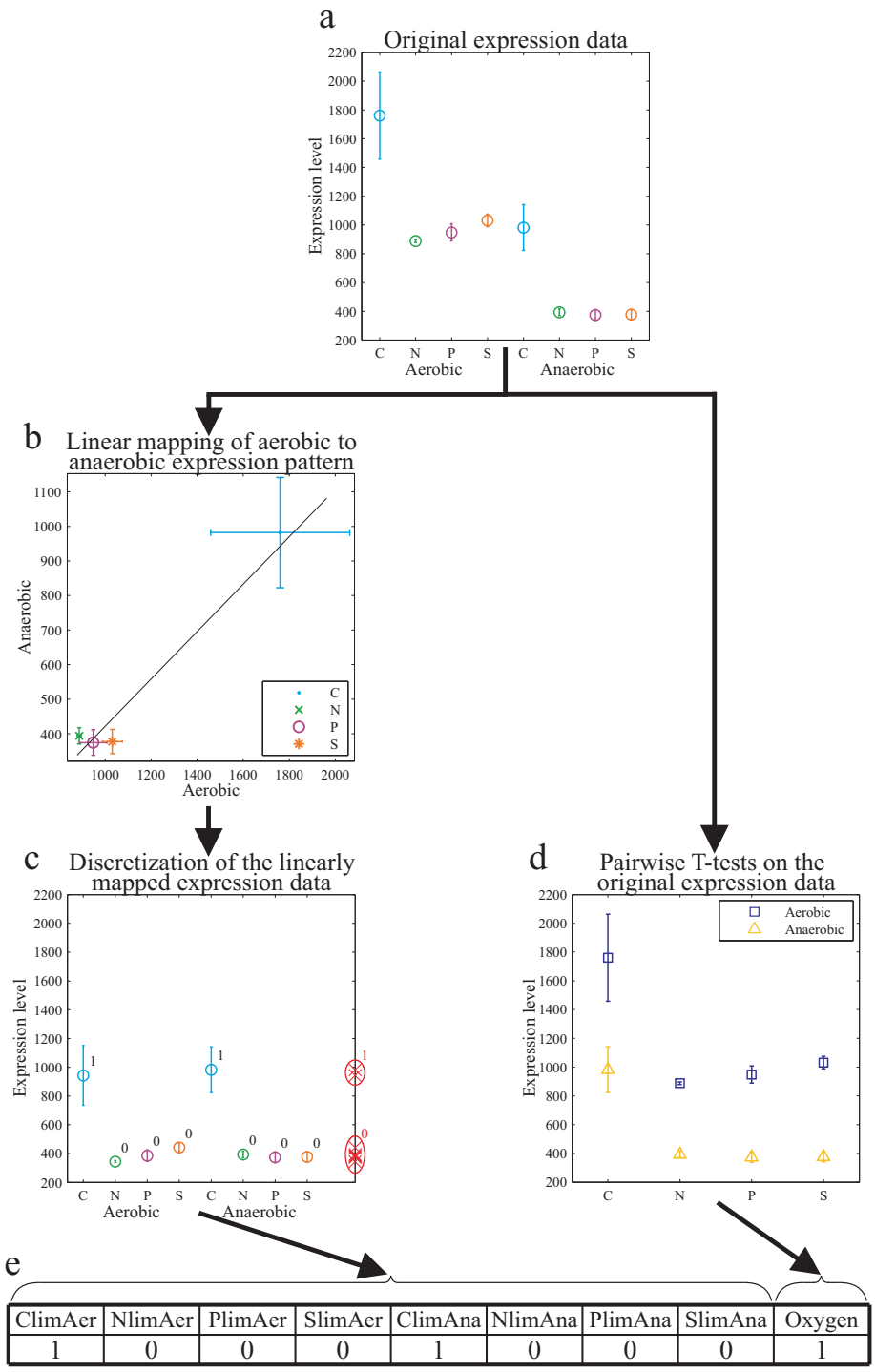



**Figure 4.2** – See page 56 for legend.

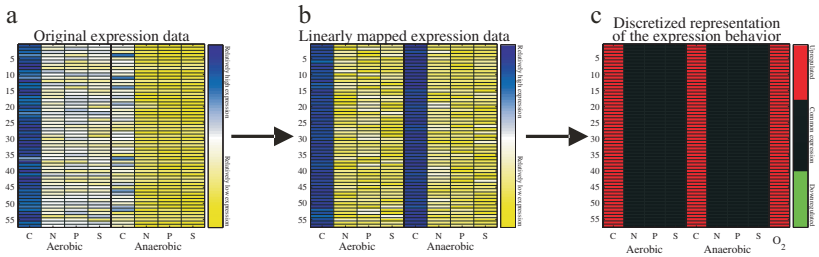



**Figure 4.3 – Heatmaps of a module.**
**a:** Normalized expression pattern of all (57) genes that share the same discretized representation, namely 100010001, and consequently, form a module. This representation, which indicates upregulation under carbon limitation and higher expression when grown within the presence of oxygen, is identical to the one derived in Figure 4.2. The expression patterns of the genes in this heatmap are comparable to the expression pattern in Figure 4.2a. **b:** Normalized expression pattern of the genes after the linear mapping is applied. Isolation of the oxygen effect clearly reveals upregulation under the carbon limitation. The linearly mapped expression patterns are comparable to the one in Figure 4.2c. **c:** The (identical) discretized expression pattern for the 57 genes. Note that our discretization procedure assigns a 0 to the cultivation conditions that form the most common expression level. For these 57 genes this common expression level is represented in **b** by the dark yellow, which occurs in six of the eight conditions. The ninth entry of this representation, i.e. the oxygen effect, is also characterized as upregulated, since the original expression levels in **a** are consistently higher under aerobic growth when compared to anaerobic growth.

In the remainder of this section, modules connected to anaerobiosis, aerobiosis and sulfur metabolism, are discussed in more detail. However, first we consider Module 13 (grey segment in Figure 4.4) that contains all genes that do not exhibit differential expression between the eight experimental conditions. (The discretized expression pattern consists of all zeros.) Three regulators have been assigned to this module, Fhl1, Sfp1 and Rap1. All three TFs are known to play an essential role in the regulation of ribosomal protein genes [Yeas, Mari 04, Moeh 91]. Although the strains were grown under different nutrient limitations and oxygen regime, the dilution rate (in other words the growth rate) of *Saccharomyces cerevisiae* was kept equal (0.1 $h^{-1}$) during the chemostat steady state in all the fermentation conditions tested [Tai 05, Pipe 02]). Given that expression regulation of ribosomal protein genes is one of the end targets of the Tor (target of rapamycin) signaling pathway, our results suggest that the regulation through the Tor signaling cascade is independent of the applied nutrient limitation and oxygen availability, but would rather reflect how the cell senses the limiting nutrient to maintain a determined growth rate.

### 4.3.3 Controlling Anaerobiosis

Module 12 (yellow segment in Figure 4.4) comprises all (383) genes that show consistent upregulation under anaerobic conditions, irrespective of any nutrient condition. Note that our strategy enables us to isolate the effect that the presence of oxygen has on the expression level of a gene. This offers the obvious advantage to independently analyze this effect. The irrelevance of the nutrient limitations is indicated by 'x's in the discretized representation of Module 12 in Figure 4.4. Several TFs and TF pairs were found to be able to bind the genes of this anaerobiosis module significantly often. Cur-

rent knowledge on gene expression regulation under anaerobic conditions cannot explain all the regulatory relationships and related TFs. The anaerobic growth conditions within our systematic experiments can therefore contribute to elucidate the role of several regulators in the absence of oxygen.

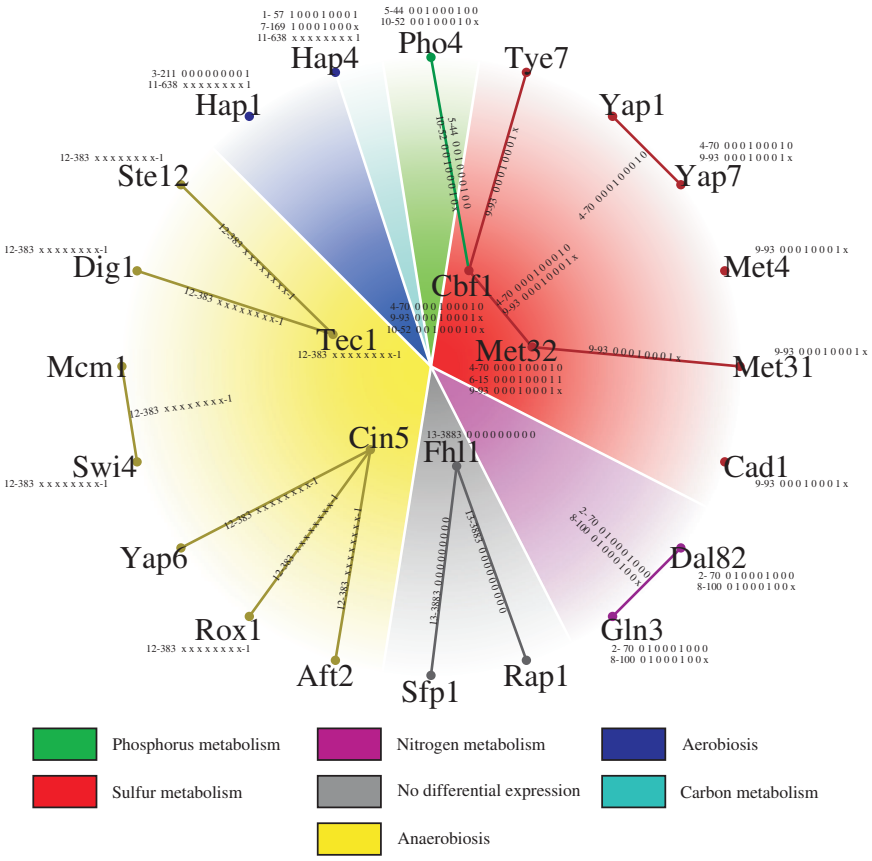



**Figure 4.4 – The TF circle.**
The TF circle depicts all the TFs and TF pairs, which are significantly related to at least one module. Related modules are represented by strings in the vicinity of the relevant TF or, in the case of a TF pair, in the vicinity of the line connecting both TFs. The strings are made up out of three parts. The first number represents the number that was assigned to the module. The second number indicates the number of genes in the module. The third part is the discretized expression pattern of the genes in the module. Here, an 'x' indicates the irrelevance (don't care) of a particular cultivation parameter. The color coding of the circle is based on the discretized expression representation of the modules. The placement of the TFs (near the center or the edge) is for reasons of visibility only.

The identification of Rox1, already known to play a role in low oxygen processes, objectively validates the truthfulness of this analysis. According to [Lee 02], this heme-dependent transcriptional repressor of hypoxic genes [Tai 05, Zito 92] constitutes a multi-component transcription factor loop together with Yap6 and Cin5, i.e. these three TFs form a regulatory circuit in which they regulate each other. Although our algorithm does not explore these kind of network structures, we identify the concerted regulation amongst these three TFs and based on our results can hypothesize that this loop is active

**Table 4.1 – Overview of the uncovered modules.**
Detailed information for all modules that are significantly related to at least one TF(-pair). Besides the discretized expression pattern and the significant TFs from binding data, the table reports overrepresented motifs through motif discovery as well as TFs associated to these motifs. Also, the most highly enriched GO, MIPS and KEGG category for each module is given (if significant).

| Module | | Disc.Expr.Pattern | | | | | | | | | | TF binding | | Motif Discovery | | Annotation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Aer | | | | Ana | | | | | | | | | | | | |
| no. | # genes | C | N | P | S | C | N | P | S | Ox | | TF's | TF pairs | Motif | Ass. TF's | GO | MIPS | KEGG |
| 1 | 57 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | | Hap4 | | CCAATCA<br>ATTGG | Hap5, Hap2/3/4, Mcm1<br>Hap5, Hap2/3/4, Mcm1, ... | GO: Oxidative phosphorylation | MIPS: Respiration | KEGG: Oxidative phosphorylation |
| 2 | 70 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | | Dal82<br>Gln3 | Gln3-Dal82 | AGATAAG<br>CTTATC | Gzf3, Dal80, Gat1<br>Gat1, Gzf3, Dal82, ... | GO: Catabolism | MIPS: Nitrogen and sulfur utilization | KEGG: Cyanoamino acid metabolism |
| 3 | 211 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | Hap1 | | | | | | |
| 4 | 70 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | | Cbf1<br>Met32<br>Yap7 | Met32-Cbf1<br>Yap7-Yap1 | CACGTGA<br>GCCACA | Cbf1, Tye7, Ino4, ...<br>Met4, Rpn4 | GO: Sulfur metabolism | MIPS: Amino acid metabolism | KEGG: Sulfur metabolism |
| 5 | 44 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | | Pho4 | Pho4-Cbf1 | ACGTGC<br>CACGTGG | Pho4, Cbf1, Ino2, ...<br>Pho4, Tye7, Cbf1, ... | GO: Anion transport | | |
| 6 | 15 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | | Met32 | | GCCAC<br>CTGTGGC | Rpn4, Met4, R. car1, ...<br>Met4, Rfx1 | | | |
| 7 | 169 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | x | | Hap4 | | GGGGTA<br>ACCCC | Mig1, Rap1<br>Mig1, Adr1, Msn4, ... | GO: Oxidative phosphorylation | MIPS: Respiration | KEGG: Oxidative phosphorylation |
| 8 | 100 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | x | | Dal82<br>Gln3 | Gln3-Dal82 | CTTATC<br>AGATAAG | Gat1, Gzf3, Dal82, ...<br>Gzf3, Dal80, Gat1 | GO: Amine transport | MIPS: Nitrogen and sulfur utilization | |
| 9 | 93 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | x | | Cad1<br>Cbf1<br>Met31<br>Met32<br>Met4<br>Yap7 | Met32-Cbf1<br>Met32-Met31<br>Tye7-Cbf1 | GCCACA<br>CACGTGA<br>CTGTGGC | Met4, Rpn4<br>Cbf1, Tye7, Ino4, ...<br>Met4, Rfx1 | GO: Sulfur metabolism | MIPS: Metabolism of methionine | KEGG: Sulfur metabolism |
| 10 | 52 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | x | | Cbf1<br>Pho4 | Pho4-Cbf1 | ACGTGC<br>CACGTGG | Pho4, Cbf1, Ino2, ...<br>Pho4, Tye7, Cbf1, ... | GO: Anion transport | | |
| 11 | 638 | x | x | x | x | x | x | x | x | 1 | | Hap1<br>Hap4 | | CCGATA | Hap1 | GO: Oxidative phosphorylation | MIPS: Respiration | KEGG: Oxidative phosphorylation |
| 12 | 383 | x | x | x | x | x | x | x | x | -1 | | Dig1<br>Rox1<br>Ste12<br>Swi4<br>Tec1 | Cin5-Aft2<br>Rox1-Cin5<br>Swi4-Mcm1<br>Tec1-Dig1<br>Tec1-Ste12<br>Yap6-Cin5 | ACAATAG<br>TGCTTT | Yox1, Rox1<br>Upc2 | GO: Lipid metabolism | MIPS: Metabolism | |
| 13 | 3883 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | Fhl1 | Rap1-Fhl1<br>Sfp1-Fhl1 | AAAAT<br>GAAAA<br>AAAAA<br>TGAAA<br>AAATA<br>AAATT | Rlr1, Spt23<br>Rlr1, Ume1, Azf1, ...<br>Azf1, Sig1, Met4<br>Ste12, Dig1, Ume1, ...<br>Smp1, Rlm1, Azf1, ...<br>Pho2, Spt23 | | | |

under anaerobic conditions. Additionally, we find the pair Ste12 and Tec1 which is known to activate genes associated with pseudohyphal growth, as well as Dig1, which conversely is involved in the negative regulation of genes involved in pseudohyphal growth [Norm 99]. (We observed a large overlap between the genes in the regulon of Tec1-Dig1 and those in the "conjugation with cellular fusion" GO-category ($P = 6.7 \cdot 10^{-8}$ according to the hypergeometric test)).
Finally, the TF pair Mcm1 and Swi4 is connected to anaerobiosis, although both are known to be involved in controlling cell cycle [Simo 01]. Moreover, Mcm1 (also named PRTF for "Pheromone Receptor Transcription Factor" [Haye 88]) is also involved in mating and response to pheromone, relating it to the cluster of Ste12, Tec1 and Dig1. These results correlate with the observation that *Saccharomyces cerevisiae* grown under anaerobic conditions exhibits elongated cell-shape irrespective of the applied nutrient limitation (See Additional file 6 of [Knij 07] online). Further investigation is needed to gain more insight into the role of these regulators in control of anaerobiosis.
Missing from the TFs significantly related to the anaerobiosis module is Upc2, which together with Rox1 is involved in regulating the expression of many genes induced under anaerobic conditions [Tai 05, Kwas 02]. The reason for not retrieving Upc2 is simply

the absence of this TF in the genome-wide location analysis employed to build the TF database. Employing motif discovery, however, the aerobic regulator 1 (AR1) binding motif of Upc2 (TCGTT [Kwas 02]) was found 244 times in the upstream regions of the 383 genes ($P = 2.4 \cdot 10^{-13}$) (See Table 4.1).

### 4.3.4 Controlling Aerobiosis

The TFs Hap1 and Hap4 are associated with the regulation of aerobiosis (dark blue segment in Figure 4.4). Hap1 is solely connected to the presence of oxygen (Modules 3 and 11), while Hap4 is also connected to carbon-limitation (Modules 1 and 7). This is in agreement with a role for Hap4 in both aerobic regulation and glucose derepression [Fors 89]. Amongst the targets of Hap1, which are overrepresented in Modules 3 and 11, we find well-known oxygen specific Hap1 regulated genes such as *CTT1*, *CYB2* and *CYC1*, confirming that its regulatory role is linked to the presence of oxygen irrespective of limited or high glucose availability.
The presence of Hap4 as part of the Hap2/Hap3/Hap4/Hap5 complex fits with the enrichment in energy categories in the aerobic genes (see Table 4.1 and Additional file 1 of [Knij 07] online). This is in line with the involvement of the Hap complex in the regulation of mitochondrial functions such as TCA cycle, electron transport chain and respiration. However, overrepresentation of only Hap4 targets from the location analysis dataset may appear as a surprise. Overrepresentation of Hap2 or Hap3 may be expected, because these two subunits of Hap2/Hap3/Hap4/Hap5 actually bind the DNA, while Hap4, as a regulatory subunit, does not. Furthermore, a clear-cut discrepancy exists between the location analysis data of the separate members of the Hap complex. The results of this study imply that the TF binding data of Hap4 is the more relevant one. This would then suggest that in order to monitor the DNA binding of a transcriptional complex, e.g. Hap2/Hap3/Hap4/Hap5, it would be more suitable to tag the subunits that do not bind the DNA template, speculating that tagging DNA binding subunits may alter the structure of the complex and, consequently, the affinity and the specificity of the interaction with the DNA.

### 4.3.5 Sulfur metabolism

The systematic combinatorial setup of cultivation conditions used to generate the transcript data allows us to extract specific information on genes regulated in response to a certain nutrient limitation. Modules 9, 6 and 4 and 82 form prime examples. Module 9 (red segment of the circle) contains all (93) sulfur-limitation-upregulated genes, regardless of the effect that the presence of oxygen might have on the expression of the genes. Modules 6, 82 and 4 consist of the sulfur-limitation-upregulated genes for which oxygen presence leads to higher expression (15 genes), lower expression (8 genes, not in Figure 4.4) and no significant change in expression (70 genes). Thus, Module 9 is the union of Modules 6, 4 and 82. Figure 4.5 displays genes from Module 9 that are bound by the TFs, which are significantly related to the set of sulfur regulated genes. In this map, genes are subdivided into groups based on their response to oxygen presence.
Several genes that show either a higher or lower expression level depending on oxygen presence, i.e. genes from Module 6 and 82 respectively, also have a binding site for the significant TFs. For example, *MET22*, involved in methionine biosynthesis, exhibits higher expression when grown anaerobically. This can be related to the fact that the

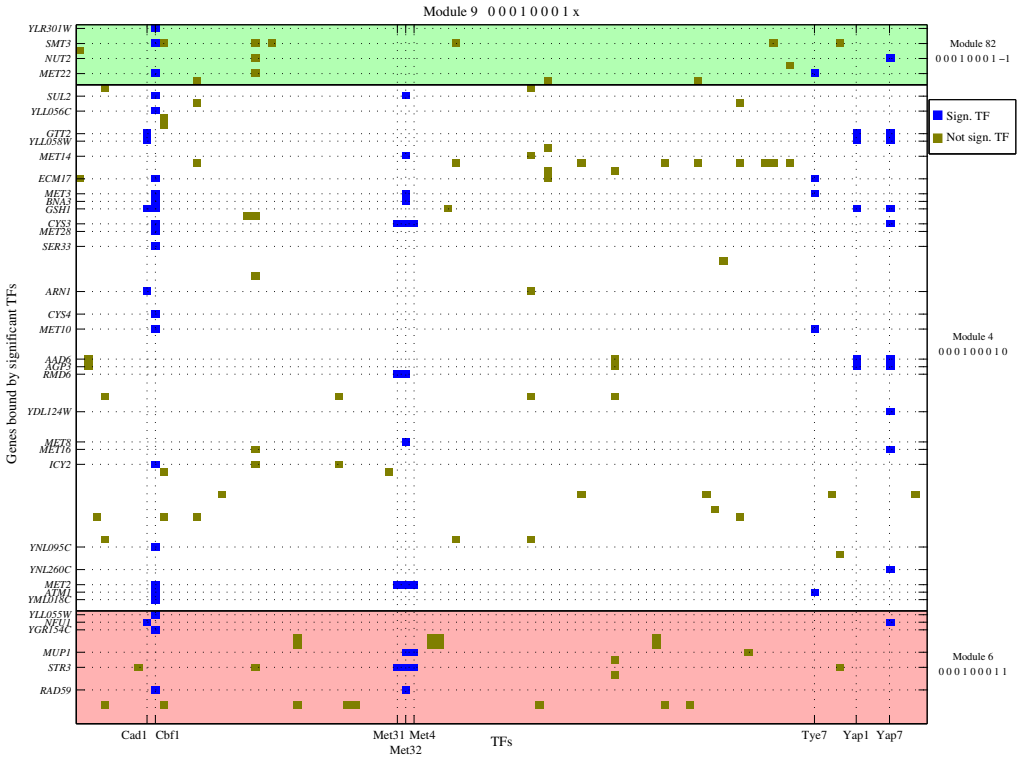



**Figure 4.5 – TF-Gene Map for Module 9.**
The TF-Gene map indicates which genes in the module can be bound (upstream) by the TFs that are significantly related to this module. Only those genes that have a binding motif in their upstream region for one of these significant TFs are annotated along the vertical axis. For these genes a dotted horizontal line is drawn. The significant TFs are annotated along the horizontal axis. For these TFs a dotted vertical line is drawn. This module, which contains all genes upregulated under sulfur limitation irrespective of the oxygen effect, can be subdivided into groups characterized by their response to oxygen presence. Genes at the top of the map (with green background) have a significantly lower expression when grown without the presence of oxygen. This group corresponds to Module 82. The middle part of the map (with white background) displays genes, which are not affected by the presence of oxygen. This group corresponds to Module 4. Genes in the bottom of the map (with red background) have higher expression when grown aerobically. This group corresponds to Module 6.

promoter sequence of *MET22* contains a LORE (low oxygen response element) motif [Vasc 01], which provides clues for future research to elucidate the functionality of this gene. Amongst the genes that have a higher expression when grown aerobically and that are bound by significant TFs, is *STR3*, involved in homocysteine and cysteine interconversion that is part of the sulfur amino acid biosynthesis and sulfur degradation pathway. Currently no relationship is known between sulfur- and oxygen-dependent regulation of this gene.

The regulatory network constructed from our analysis reveals a complex interplay between six individual transcription factors (Met4, Met31, Met32, Cbf1, Yap7 and Cad1) and four pairs of regulators (Tye7-Cbf1, Cbf1-Met32, Met32-Met31 and Yap1-Yap7) connected to sulfur metabolism. Met4, Met31, Met32 and Cbf1 constitute an internal

validation of the analysis, since these four factors are indeed known as members of the Met regulatory complex [Roui 00] that also includes the regulatory subunit Met28. More interestingly, our data provide new insight into sulfur metabolism regulation by implicating new regulators as Tye7 and the members of the fungal-specific family of basic leucine zipper (bZIP) proteins Yap1, Cad1 (Yap2) and Yap7. Literature reports available so far concerning Tye7 limit its role to cell cycle [Hora 02]. Our results, however, would implicate that Tye7 in combination with Cbf1 would participate in the regulation of the genes encoding the upper part of the sulfur assimilation pathway including *MET3*, *MET10*, *ECM17*, *MET22* and *ATM1*, who's gene products are involved in maturation of cytosolic Fe/S (iron-sulfur) proteins [Sipo 02]. Even more interesting is the possible cross-coupling with phosphate metabolism. As indicated in Figure 4.4, Cbf1 was also found to bind the upstream regions of phosphorus regulated genes significantly often. Given that Cbf1, Pho4 and Tye7 recognize similar binding sites, our results could shed new light on the possible cross-regulation of phosphate and sulfate metabolism that centers around Pho4 and Cbf1 [OCon 92].
In the case of Cad1 and Yap1 the link to sulfur metabolism may correlate to their reported role in mediating resistance to cadmium ($Cd^{2+}$), which leads to changes in the sulfate assimilation pathway and to sulfur sparing [Fauc 02]. When *Saccharomyces cerevisiae* is exposed to $Cd^{2+}$ most of the sulfur assimilated by the cells is converted into glutathione, a thiol-metabolite essential for detoxification. Yeast adapts to this vital metabolite requirement by globally modifying its proteome to reduce the production of abundant sulfur-rich proteins.

## 4.4 Discussions and Conclusions

We observed and successfully modeled that the presence of oxygen leads to an offset (addition) and/or scaling (multiplication) of the expression levels of many genes, corroborating the existence of various types of regulation on various levels. The uncovered results find their origin in the systematic combinatorial setup of the well-defined cultivation conditions within the experiment. Our tailored approach exploits the interrelatedness between the conditions and links the cultivation parameters to TF activity and gene expression behavior.
We compared our strategy to an approach that follows the exact same steps, but which does *not* exploit the systematic setup of the cultivation conditions. In short, when the interrelatedness between the conditions is not used, the original continuous expression levels are discretized without modeling the oxygen effect. Results of this comparison indicate that more genes can be related to a particular cultivation parameter when incorporating the relations between the cultivation conditions. See Table 4.2. Additionally, we can relate more TFs and TF pairs to the generated modules and achieve higher functional annotation enrichment. See Additional files 4 and 5 of [Knij 07] online (as well as Additional files 1 and 2 for a more in depth comparison). These results provide additional evidence for the validity of the adopted approach.
Moreover, the inclusion of the conditions within the computational framework accommodates the assessment of the direct effect of these conditions on gene expression, TF activity and other biological processes in the cell. This is in contrast to the currently used compendium approaches, where the relation between the cultivation conditions is ambiguous and can not be modeled. There, large heterogeneity in terms of the strain,

**Table 4.2 – Effect of the linear mapping on module size and enrichment.**
For the modules that are most straightforwardly related to one of the cultivation parameters (the four nutrient limitations and the oxygen availability) this table indicates the size of the respective module, the number of associated TFs, TF pairs and annotation categories; both with and without appliance of the linear mapping. (Note that when no linear mapping is applied the original continuous expression levels are discretized and no oxygen effect can be computed, resulting in a discretized expression pattern of length eight.

| Cultivation parameter | Linear mapping applied | | | | No linear mapping | | | |
|---|---|---|---|---|---|---|---|---|
| | Disc.Expr.Pattern | # genes | # TF(pairs) | # Ann.cat. | Disc.Expr.Pattern | # genes | # TF(pairs) | # Ann.cat. |
| Carbon | 1 0 0 0 1 0 0 0 x<br>-1 0 0 0 -1 0 0 0 x | 169<br>69 | 1<br>- | 48<br>- | 1 0 0 0 1 0 0 0<br>-1 0 0 0 -1 0 0 0 | 59<br>23 | -<br>- | 5<br>- |
| Nitrogen | 0 1 0 0 0 1 0 0 x<br>0 -1 0 0 0 -1 0 0 x | 100<br>2 | 2(1)<br>- | 8<br>- | 0 1 0 0 0 1 0 0<br>0 -1 0 0 0 -1 0 0 | 42<br>0 | 2(1)<br>- | 8<br>- |
| Sulfur | 0 0 1 0 0 0 1 0 x<br>0 0 -1 0 0 0 -1 0 x | 52<br>2 | 2(1)<br>- | 1<br>- | 0 0 1 0 0 0 1 0<br>0 0 -1 0 0 0 -1 0 | 39<br>1 | 2(1)<br>- | 6<br>- |
| Phosphorus | 0 0 0 1 0 0 0 1 x<br>0 0 0 -1 0 0 0 -1 x | 93<br>4 | 6(3)<br>- | 27<br>- | 0 0 0 1 0 0 0 1<br>0 0 0 -1 0 0 0 -1 | 59<br>1 | 5(3)<br>- | 27<br>- |
| Oxygen | x x x x x x x x 1<br>x x x x x x x x -1 | 638<br>383 | 2<br>5(6) | 75<br>13 | 1 1 1 1 0 0 0 0<br>0 0 0 0 1 1 1 1 | 115<br>76 | 2<br>- | 19<br>5 |

growth rate, growth conditions, measuring technique and other environmental or measurement parameters may have a profound, but undetermined impact on the behavior of the cell and the resulting dataset. Consequently, these approaches often resort to annotation databases to determine the functionality of a module or TF. For example, in the GRAM method [Bar 03], where the functionality of a module is based on enrichment in MIPS functional categories, the TF Hap4 was only related to respiration. We could, on other hand, not only demonstrate that oxygen plays an important role, but also identified the known effect of the extracellular glucose concentration on Hap4 and its regulon.

In this study we identified many novel putative regulatory relationships. Examples include the role of Tye7 in regulating sulfur metabolism and the cross-regulation between phosphate and sulfate metabolism. Given the quality and uniqueness of the dataset, many other clues about regulation mechanisms related to yeast's metabolism and respiration can still be extracted.

We believe that quantification of the complex relationships that control cellular adaptation to different environments necessitates well-designed and carefully controlled experiments. In this respect, the design of experimental setups, where interrelated cultivation conditions are systematically combined, is especially important. The analysis of the individual and combined effects of the cultivation parameters in such experiments will help to reveal the multi-faceted nature of cellular regulatory mechanisms.

## 4.5 Methodology

### 4.5.1 Selection of differentially expressed genes

Genes that show differential expression across the experimental conditions are selected. For this purpose, we employed a multi-class SAM analysis [Tush 01]. Here, the classes are the eight different experimental conditions. The 2500 most significantly changed genes are selected (median false discovery rate of 0.01%). This is an estimate of the number of genes involved in the metabolic processes of yeast grown under the experimental conditions [Tai 05].

### 4.5.2 Isolation of the global oxygen effect

To investigate the linear relationship between the aerobic and anaerobic expression values of a gene, we perform the following steps: First, we compute the mean and standard deviation across the replicates, $\mu_{ij}$ and $\sigma_{ij}$, for the nutrient limitations $i = 1...4$ and both aerobic ($j = 1$) and anaerobic ($j = 2$) growth. We model the joint aerobic-anaerobic expression distribution for each nutrient limitation $i$ as a normal distribution $N(\mu_i, \Sigma_i)$, with $\mu_i = [\mu_{i1}, \mu_{i2}]$ and $\Sigma_i = \begin{pmatrix} \sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2 \end{pmatrix}$. This is graphically depicted in Figure 4.2b. Next, we estimate the parameters of a linear model (slope and offset) by fitting a straight line through the four normal distributions. This heteroscedastic regression problem is solved as described in [Leed 00]. As a goodness-of-fit criterion for the regression, a $P$-value was computed by employing the Student's T cumulative distribution function with the ratio between the slope and the standard deviation of the slope. The $P$-value cut-off was set at $10^{-4}$. When no significant linear relationship ($P > 10^{-4}$) is found employing the four nutrient limitations, we successively leave one of the nutrient limitations out, thus employing only three normal distributions to find a linear relationship. If $P \leq 10^{-4}$ for the best of the resulting four fits, this fit is used. This strategy handles genes with one nutrient-limitation-specific reaction to oxygen presence. See Additional file 7 of [Knij 07] online. When again no good linear relationship is found, the slope is fixed to one and only the offset (i.e. the difference between the mean aerobic and anaerobic expression level) is computed. See Additional file 8 of [Knij 07] online. The three different regression strategies (use of four nutrient limitations, use of three nutrient limitations, only compute the offset) were applied to 1190, 518 and 792 genes, respectively. For each gene, we apply the estimated parameters (slope $a$ and offset $b$) to map the original aerobic expression values $\mathbf{x}$ to their linearly mapped values $\mathbf{x}'$, via $\mathbf{x}' = a \cdot \mathbf{x} + b$, thereby aligning the aerobic and anaerobic expression patterns, such that the differences in the resulting expression pattern are not caused by the oxygen effect. See for example, Figure 4.2c.

### 4.5.3 Construction of the discretized representation

A gene is represented by a ternary expression pattern of length nine. The first eight entries represent the discretized representation of the linearly mapped continuous expression data, which can be either 0, -1 or 1, indicating the most common expression level, downregulation or upregulation, respectively. Since the linear mapping changes the continuous expression pattern of a gene, SAM is run again on the linearly mapped data. Genes that now drop out of the top 2500 most differentially expressed genes are assigned a value of zero in the first eight entries of the expression pattern. Genes, that remain in the top 2500 (2062 genes) are discretized by employing k-means clustering for each gene separately, i.e. in an one-dimensional space on the eight mean expression levels associated with the eight experimental conditions. (Red crosses on the right vertical axis in Figure 4.2c). The Davies-Bouldin index [Davi 79] was used to choose between $k = 2$ (most common level and down- or upregulation) and $k = 3$ (all three quantized levels). Genes for which no compact and well-separated clusters could be found, i.e. for which the Davies-Bouldin index for both $k = 2$ and $k = 3$ exceeded 0.5, were removed. The most common level (0) was assigned to the experimental conditions that formed the largest cluster. The clusters with higher or lower gene expression levels with respect to the most common level cluster are labeled as upregulated (1) or downregulated (-1) respectively.

The ninth entry of the discretized expression pattern of a gene represents the global oxygen effect. This can either be 0, -1 or 1. No significant difference between expression under aerobic and anaerobic growth is indicated by a zero (0). A consistent significantly lower or higher expression level when grown anaerobically is indicated by -1 and 1, respectively. The global oxygen effect is determined by performing pairwise T-tests for all nutrient limitations, comparing the original expression levels when grown aerobically with those when grown anaerobically. See Figure 4.2d. When at least three of the four nutrient limitations have a significantly ($P \leq 5 \cdot 10^{-2}$) higher expression when grown aerobically (or anaerobically) we assign a 1 (or -1 respectively). (In the case where only three nutrient limitations were used in regression only two of these three should be significantly higher (or lower) to pass the test.)

### 4.5.4 Generation of the modules

Modules are formed by grouping genes with identical discretized expression patterns, i.e. by performing a hierarchical clustering on the discretized data with Hamming distance as dissimilarity measure and then forming clusters by cutting the dendrogram at a distance of zero (linkage is irrelevant). Additionally, modules are formed with the global oxygen effect being irrelevant, i.e. genes are clustered together when only the first eight entries of the expression pattern are identical. Similarly, modules are created based solely on the oxygen effect. This strategy creates overlapping clusters of genes, that represent different characterizations based on the global oxygen effect.

### 4.5.5 Identification of significant TFs and enrichment of annotation categories

Modules are related to TFs by the hypergeometric test, which assesses the probability that the observed frequency that the genes in a module are bound by a TF would occur by chance. The $P$-value cutoff to decide whether a relation is significant is $P \leq 1/(n_m n_x)$, where $n_m$ is the number of modules consisting of more than ten genes and $n_x$ is the number of TFs or TF pairs that bind to more than ten genes. This Bonferroni correction for multiple testing results in a per-family error rate (PFER) of one [Ge 03]. Considering the stringency of the Bonferroni correction and the fact that the tests are not independent, the $P$-value correction is quite conservative. The same procedure is employed to assess the overrepresentation of GO, MIPS and KEGG annotation categories.

### 4.5.6 Motif discovery

RSAT motif discovery [Held 03] was applied to modules, which are significantly related to at least one TF or TF pair. An oligonucleotide analysis was run with motif sizes ranging from five to eight. Significant (RSAT occurrence significance score larger than one) and dissimilar motifs for each module were manually extracted. Published PWM/PSSM matrices for known TFs [Harb 04, Tran, SCPD] are captured in the weight matrix form as described in [Hert 99]. A simple similarity score between a motif and a weight matrix, i.e. the sum of the weights of the matrix for the letters of the aligned motif, was employed to relate the uncovered motifs to known TFs.

# CHAPTER 5

# CONDITION TRANSITION ANALYSIS

This chapter uses the results of Chapter 4 to focus on the oxygen-specific effects within this dataset. The eight conditions are described as states. The activity of TFs is assessed for the different state transitions. Special attention is devoted to TFs that seem to perform a regulatory role under aerobic conditions, but not under anaerobic growth (or vice versa). The resulting regulatory network reveals nutrient-limitation-specific effects of oxygen presence on expression behavior and TF activity. The analysis identifies many TFs that seem to play a very specific and subtle regulatory role at the nutrient and oxygen availability transitions.

This chapter is published as:

'Condition transition analysis reveals TF activity related to nutrient-limitation-specific effects of oxygen presence in yeast'

Theo A. Knijnenburg, Lodewyk F.A. Wessels and Marcel J.T. Reinders

Computational Methods in Systems Biology (CMSB), International Conference, Trento, Italy, Proceedings, p. 271-284, 18-19 October 2006

## 5.1 Abstract

Regulatory networks are usually presented as graph structures showing the (combinatorial) regulatory effect of transcription factors (TFs) on modules of similarly expressed or otherwise related genes. However, from these networks it is not clear when and how TFs are activated. The actual conditions or perturbations that trigger a change in the activity of TFs should be a crucial part of the generated regulatory network.
Here, we demonstrate the power to uncover TF activity by focusing on a small, homogeneous, yet well defined set of chemostat cultivation experiments, where the transcriptional response of yeast grown under four different nutrient limitations, both aerobically as well as anaerobically was measured. We define a condition transition as an instant change in yeast's extracellular environment by comparing two cultivation conditions, where either the limited nutrient or the oxygen availability is different. Differential gene expression as a consequence of such a condition transition is represented in a tertiary matrix, where zero indicates no change in expression; 1 and -1 respectively indicate an increase and decrease in expression as a consequence of a condition transition. We uncover TF activity by assessing significant TF binding in the promoter region of genes that behave accordingly at a condition transition. The interrelatedness of the conditions in the combinatorial setup is exploited by performing specific hypergeometric tests that allow for the discovery of both individual and combined effects of the cultivation parameters on TF activity. Additionally, we create a weight-matrix indicating the involvement of each TF in each of the condition transitions by posing our problem as an orthogonal Procrustes problem. We show that the Procrustes analysis strengthens and broadens the uncovered relationships.
The resulting regulatory network reveals nutrient-limitation-specific effects of oxygen presence on expression behavior and TF activity. Our analysis identifies many TFs that seem to play a very specific regulatory role at the nutrient and oxygen availability transitions.

## 5.2 Introduction

The systems biology view of an organism as an interacting network of genes, proteins and biochemical reactions seems very promising for revealing the underlying networks of transcriptional regulation in *Saccharomyces cerevisiae*. For this yeast enormous amounts of different intracellular data have been measured, enabling the integration of multiple data sources [Bane 02]. In inferring regulatory networks common approaches focus on integration of microarray gene expression data, ChIP-chip TF binding data and sequence data (to detect cis regulatory elements) [Chua 04]. The resulting networks are usually presented as graph structures showing the (combinatorial) regulatory effect of TFs on modules of similarly expressed or otherwise related genes (e.g [Pilp 01, Bar 03, Wang 05]). However, from these networks it not clear when and how TFs are activated. This is quite strange, since the actual conditions or perturbations that trigger a change in the activity of TFs should be a crucial part of the generated regulatory network. Three main reasons for this exclusion can be identified: Firstly, the present inability to directly measure protein levels in vivo prevents direct assessment of the presence of a TF in a particular condition. Secondly, in most cases post-transcriptional and/or post-translational regulation prevent deriving TF activity from gene expression, although an

attempt was made based on this assumption [Sega 03]. Thirdly, the trend of employing increasingly large compendia of heterogeneous microarray data, where yeast is grown under a wide variety of very different and unrelated conditions, makes it impossible to incorporate all these conditions in a regulatory program. Hence, the functionality of modules and TFs is assigned based on enrichment in annotation categories (e.g. Gene Ontology [Ashb 00]). This means that the functionality purely depends on the result of clustering, i.e. the grouping of genes, and not specifically on the cultivation conditions under which the expression behavior is characteristic for a module. This approach can only provide a global overview of TF activity and obstructs novel knowledge discovery, since an existing body of knowledge, i.e. the ontologies, is taken as a golden standard. Here, we demonstrate the power in uncovering TF activity by focusing on a small, homogeneous, yet well defined set of chemostat cultivation experiments, where the transcriptional response of yeast grown under four different nutrient limitations, both aerobically as well as anaerobically was measured (See Table 5.1 and Figure 5.1) [Tai 05]. In this research we focus on condition *transitions* by comparing gene expression profiles of two cultivation conditions and evaluate whether genes are differentially expressed between these two conditions. TF activity is inferred by assessing significant TF binding in the promoter region of genes that behave accordingly at the transitions. For this, we use the largest available TF binding dataset [Harb 04]. The interrelatedness of the cultivation conditions within the systematic combinatorial setup is exploited by performing specific hypergeometric tests. This enabled us to reveal nutrient-limitation-specific effects of oxygen presence on expression behavior and TF activity. Additionally, we create a weight-matrix indicating the involvement of TFs in each of the condition transitions by posing our problem as an orthogonal Procrustes problem. Analysis of this weight matrix broadens the significant relations found by the hypergeometric test. The uncovered regulatory mechanisms offer valuable clues of how yeast changes its metabolism and respiration as a result of specific changes in nutrient and oxygen availability.

## 5.3 Methods

### 5.3.1 Data and preprocessing

The employed microarray gene expression data consists of the measured transcriptional response of the yeast *Saccharomyces cerevisiae* to growth limitation by four different macronutrients in both aerobic and anaerobic media. See Table 5.1. Three independently cultured replicates were performed per experimental condition. For more information see Tai *et al.* [Tai 05]. SAM [Tush 01] was employed (with median false discovery rate of 0.01%) to select genes that are differentially expressed amongst the eight conditions. Next, we remove the observed global effect that the presence of oxygen has on the expression level of each gene under all nutrient limitations by a linear regression strategy as described in Knijnenburg *et al.* [Knij 07]. Then, for each gene individually the expression levels are discretized by employing a k-means clustering algorithm on the eight mean expression levels (corresponding to the eight conditions) in a one-dimensional space [Knij 05]. Here, the Davies-Bouldin index [Davi 79] was employed to select between $k = 2$ and $k = 3$. The conditions that comprise the largest cluster are said to have common expression level, while conditions that form a cluster with a higher or lower expression level when compared to the largest cluster are called up- or downregulated, respectively. (In the case that $k = 2$ one cluster has common expres-

**Table 5.1 – Experimental conditions.**
The black squares indicate the employed nutrient limitation and oxygen supply.

| Experimental condition | Nutrient limitation | | | | Oxygen supply | |
|---|---|---|---|---|---|---|
| | Carbon | Nitrogen | Phosphorus | Sulfur | Aerobic | Anaerobic |
| 1. ClimAer | ■ | | | | ■ | |
| 2. NlimAer | | ■ | | | ■ | |
| 3. PlimAer | | | ■ | | ■ | |
| 4. SlimAer | | | | ■ | ■ | |
| 5. ClimAna | ■ | | | | | ■ |
| 6. NlimAna | | ■ | | | | ■ |
| 7. PlimAna | | | ■ | | | ■ |
| 8. SlimAna | | | | ■ | | ■ |

sion level and the other is either upregulated or downregulated.) Hence, the expression behavior of a gene is defined in terms of up- and/or down regulation under the eight cultivation conditions. Discretized expression patterns of all genes are captured in $\mathbf{G}$, a tertiary matrix of $6383 \times 2 \times 4$. In $\mathbf{G}_{g,o,n}$, $g = \{1 \dots 6383\}$ are the different genes in the genome, $o = \{1, 2\}$ represents oxygen supply (aerobic and anaerobic respectively) and $n = \{1 \dots 4\}$ are the four nutrient limitations (carbon, nitrogen, phosphorus and sulfur respectively). Zero indicates common expression level; 1 and -1 indicate upregulation and downregulation respectively. An example:

$$\mathbf{G}_{453,:,:} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

This gene (*MTD1*, indexed as no. 453) is thus upregulated under the phosphorus limitation (both aerobically and anaerobically) and downregulated under the nitrogen limitation in anaerobic growth. Note that genes that are not differentially expressed are assigned zeros in all cultivation conditions.
The TF binding data indicates the number of motifs in the promoter region of each gene for 102 TFs [Harb 04]. In this study we have employed motifs that are bound at high confidence ($P \leq 0.001$); not taking into account conservation among other *sensu stricto Saccharomyces* species. The $6383 \times 102$ matrix, denoted by $\mathbf{F}$, is binarized, such that $\mathbf{F}_{g,f}$ indicates whether the promoter region of gene $g$ can be bound by TF $f$.

### 5.3.2 Condition transition analysis

From expression matrix $\mathbf{G}$ we derive the condition transition matrix $\mathbf{T}$. We define a condition transition as an instant change in yeast's extracellular environment by comparing two cultivation conditions and assess whether genes exhibit change in expression level when "going" from one cultivation condition to the other. In total we define sixteen condition transitions. These are only the transitions, where either the nutrient limitation or the oxygen availability changes; not both. The transitions are indicated by the edges in the cube of Figure 5.1. The six nutrient limitation transitions, both aerobically and anaerobically, (edges in the upper and lower face of the cube) are computed by:

$$\mathbf{T}_{g,I(n_1,n_2,o)} = sign(\mathbf{G}_{g,o,n_1} - \mathbf{G}_{g,o,n_2}) \quad \forall\{g, o, n_1 > n_2\} \tag{5.1}$$

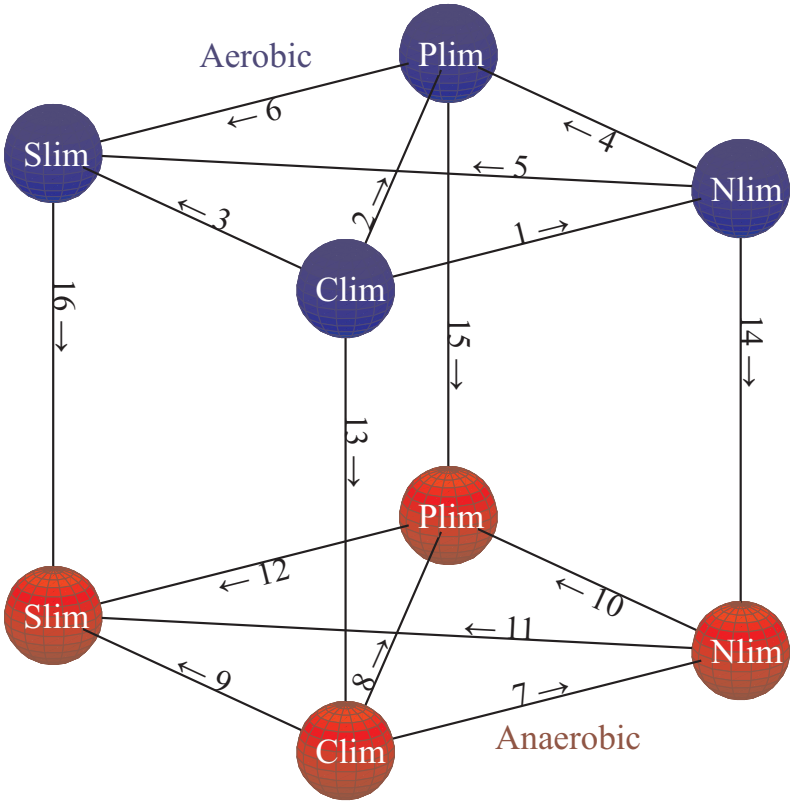



**Figure 5.1 – Cube representing the eight cultivation conditions.**
Edges indicate defined condition transitions.

The four oxygen availability transitions (vertical edges) are computed by:

$$\mathbf{T}_{g,12+n} = sign(\mathbf{G}_{g,1,n} - \mathbf{G}_{g,2,n}) \quad \forall\{g,n\} \tag{5.2}$$

Here, $I(n_1, n_2, o) = [6 * (o-1) + n_1 + 4 \cdot (n_2 - 1) - \frac{n_2 \cdot (n_2+1)}{2}]$, such that the indices of the different transitions in $\mathbf{T}$ correspond to the numbers assigned to the edges in the cube of Figure 5.1. $\mathbf{T}$ ($6383 \times 16$) is again a tertiary matrix, where zero indicates no change in expression; 1 and -1 respectively indicate an increase and decrease in expression as a consequence of a condition transition. Now, by consulting the TF binding matrix $\mathbf{F}$, a hypergeometric test can be employed to assess if genes that are up- and/or downregulated at a condition transition are bound (upstream) by a TF much more frequently than would be expected by chance. In more general terms, by employing the hypergeometric distribution we compute the probability of the observed (or more extreme) overlap between two sets of genes under the assumption that these sets of genes were randomly chosen from all genes [Bara 01]. Small probabilities ($P$-values) indicate that the hypothesis that these sets are randomly drawn must be dismissed, thereby acknowledging a significant relation between the two sets. In our setting, one set is constituted

**Table 5.2 – Conditions on T which define nine groups for each nutrient limitation transition ($t = \{1 \dots 6\}$).**
The last two columns indicate the vertical placement (vp) and color of TFs that are significantly related to these groups as visualized in Figure 5.2.

| no. | $\mathbf{T}_{g,t}$ | $\mathbf{T}_{g,t+6}$ | Description | vp | color |
|---|---|---|---|---|---|
| I | 1 | 0,-1 | Only up under aerobic growth | top | orange |
| II | 0,-1 | 1 | Only up under anaerobic growth | bottom | orange |
| III | 1 | 1 | Up under both aerobic and anaerobic growth | center | orange |
| IV | -1 | 0,1 | Only down under aerobic growth | top | green |
| V | 0,1 | -1 | Only down under anaerobic growth | bottom | green |
| VI | -1 | -1 | Down under both aerobic and anaerobic growth | center | green |
| VII | 1,-1 | 0 | Only diff. expressed under aerobic growth | top | black |
| VIII | 0 | 1,-1 | Only diff. expressed under anaerobic growth | bottom | black |
| IX | 1,-1 | 1,-1 | Diff. expressed under both aerobic and anaerobic growth | center | black |

of all genes that can be bound by a particular TF, while the other set consists of e.g. all genes upregulated at a particular condition transition.
However, the systematic setup of the cultivation conditions in this dataset, allows for selection of more interesting groups of genes to input into the hypergeometric test. For example, genes that are upregulated at an aerobic nutrient limitation transition, yet not upregulated at the same nutrient limitation transition without the presence of oxygen. More specifically, for each of the six nutrient limitation transitions we define nine different groups of genes allowing us to focus on upregulation (1), downregulation (-1) and differential expression (-1 or 1), both specifically for aerobic or anaerobic growth as well as regardless of the oxygen supply. See Table 5.2. The 54 groups, augmented with groups of genes up-, downregulated or differentially expressed under the four oxygen availability transitions (Transitions 13-16), are tested for significant association with TFs by employing the hypergeometric test. (To adjust for multiple testing, the $P$-value cutoff was set, such that we expect one false positive (per-comparison error rate (PCER) of one [Ge 03]), corresponding with $P \leq 1.5 \cdot 10^{-4}$.) Figure 5.2 displays the significant relations in (for reasons of visibility) a part of the cube. We now have a regulatory network, which associates a TF with a cluster of genes that shows specific gene expression changes when a transition is made from one condition to the next.
In an attempt to gain more insight into the dynamics and combinatorial effects within the complete generated regulatory network, in stead of performing stringent tests of individual hypotheses, we add an additional step to our analysis. Here, we aim at modeling the expression behavior at all condition transitions $\mathbf{T}$ by employing binding matrix $\mathbf{F}$ and assess the activity of each TF at a condition transition. This approach is based on the simple biological model that ascribes the change of gene expression levels as observed at a condition transition to changes in TF activity; the means by which the organism adapts to the changed extracellular environment. In contrast to the landmark article by Bussemaker *et al.* [Buss 01], where expression was explained using cis-regulatory elements, we thus explain expression behavior at transitions by using TF binding data. In a more recent article from Bussemaker's group [Gao 04] also TF binding data was used to explain expression. However, they used a continuous score (the logarithm of the binding $P$-value) to represent the degree of TF binding, while we use the binary one, which indicates simply whether there is the ability to bind or not. Furthermore, we do not employ continuous expression levels, which are a measure of absolute mRNA quantities. We use the discrete elements of $\mathbf{T}$ that represent relative up- and downregulation, since

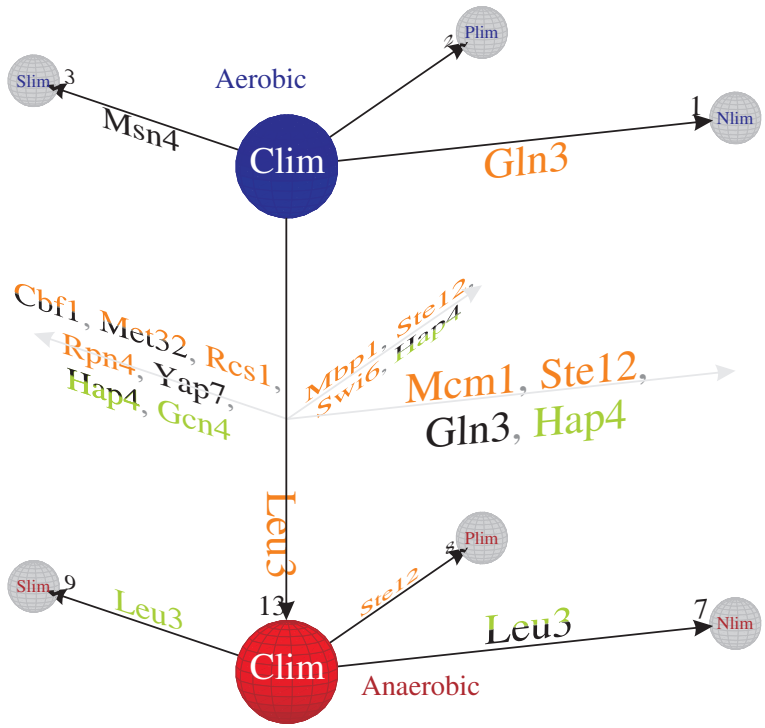



**Figure 5.2 – TF activity for part of the transitions.**
Green, orange and/or black TFs are significantly related to genes that are downregulated, upregulated or differentially expressed respectively when going from one cultivation condition to the other (in the direction of the arrows). TFs on the top and bottom edges are activated only under aerobic or anaerobic growth respectively; TFs in the center of a surface indicate activation independent of the presence of oxygen. For example, Mcm1, Ste12, Gln3 and Hap4 are associated with transitions from carbon limitation to nitrogen limitation, independent of the presence of oxygen.

we find this more robust and informative compared to (the difference between) absolute mRNA levels. Another big difference is that we do not use an iterative procedure to solve the problem, but aim at explaining all the transitions using all TFs in one time. Our problem finds its mathematical formulation in the orthogonal Procrustes problem, where we explore the possibility that $\mathbf{F}$ can be rotated into $\mathbf{T}$ by solving:

$$\min \|\mathbf{T}' - \mathbf{W}\mathbf{F}'\|_{Fro} \qquad \text{subject to } \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{5.3}$$

In principle, this is a linear transformation of the points in $\mathbf{F}$ to best conform them to the points in $\mathbf{T}$. In our setting, the change in expression of a gene at a condition transition (as given in $\mathbf{T}$) is approximated by a weighted sum of ones. These ones correspond to the TFs that can bind the upstream region of that particular gene (as given in $\mathbf{F}$). Thus, the elements in $\mathbf{W}$ represent a measure of the activity of a TF at a condition transition. Properties of the Procrustes rotation are the closed solution (via a SVD decomposition [Golu 96]) in minimizing the *Frobenius norm* (sum of squared errors) and the orthonormality of weight matrix $\mathbf{W}$. A prerequisite for this rotation is that the num-

ber of columns (TFs) in $\mathbf{F}'$ should match the number of columns (condition transitions) in $\mathbf{T}'$. Since our main focus is on TFs that regulate differently at nutrient limitation transitions as a consequence of oxygen supply, we select only the first twelve columns from $\mathbf{T}$. The twelve selected TFs are those that (according to the hypergeometric test) are most significantly related to up- or downregulation, specifically under aerobic or anaerobic growth (i.e. related to groups I, II, IV and V in Table 5.2). Furthermore, we only employ those genes which exhibit different expression between aerobic and anaerobic growth for at least one of the six nutrient limitation transitions. These adjustments on $\mathbf{F}$ and $\mathbf{T}$ yield $\mathbf{F}'$ and $\mathbf{T}'$ (both $1493 \times 12$), which are employed in Eq. 5.3. Figure 5.3 visualizes the resulting $\mathbf{W}$.
Permutation tests were performed to assess the statistical significance of these weights. The rows (genes) of $\mathbf{T}'$ were randomly permuted after which the Procrustes rotation (Eq. 5.3) was recomputed. This was done 10,000 times. The Wilcoxon signed rank test was applied to check if the original weights could be the medians of the distributions of weights generated by the permutations. The extremely low $P$-values for almost all weights indicated that this hypothesis should be dismissed. (Results not shown.) This attaches, at least, a statistical meaning to the derived weight matrix. More interestingly, for each of the twelve TFs and each of this six nutrient limitation transitions we assessed the significance of the difference between the assigned weight under aerobic growth and the weight under anaerobic growth. A $P$-value was computed by determining the fraction of permutations in which the difference between the aerobic and anaerobic weight was larger than for the original (non-permuted) data. Significant differences ($P \leq 0.05$) point towards oxygen specific regulation of a TF at a specific nutrient limitation transition.

## 5.4 Results

The network of TF activity, as partly presented in Figure 5.2, provides many very specific clues towards the transcriptional regulation of yeast's metabolism and respiration. Some of these can be linked to existing biological knowledge quite easily. One obvious example is the TF Hap4, of which the mRNA abundance is decreased by the presence of glucose [Fors 89]. This explains downregulation of the regulon of Hap4 in the three nutrient transitions moving away from the carbon limitation. Furthermore, in the carbon to sulfur limitation transition, we find Met32, a known transcriptional regulator of methionine metabolism [Yeas], as well as Cbf1, which is part of the transcription activation complex Cbfl-Met4-Met28 [Blai 98]. To find TF Gln3 at the transition from carbon limited growth to growth where nitrogen becomes the limiting nutrient is also not surprising. Ammonium, the nitrogen source used in these experiments and generally considered to be the preferred nitrogen source for *S. cerevisiae*, is in excess under carbon-limited growth, while absent under nitrogen-limited growth. It is well known that high concentrations of ammonium lead to nitrogen catabolite repression (NCR), a transcriptional regulation mechanism that represses pathways for the use of alternative nitrogen sources [Maga 02]. Gln3 is one of the four so-called GATA factors active in NCR to adapt to the change in need of alternative nitrogen sources at this transition. It is however surprising that Gln3 is significantly related to genes upregulated, especially under aerobic conditions. Also unexpectedly, Leu3, a regulator for genes of branched-chain amino acid biosynthesis pathways, is significantly related to genes downregulated,

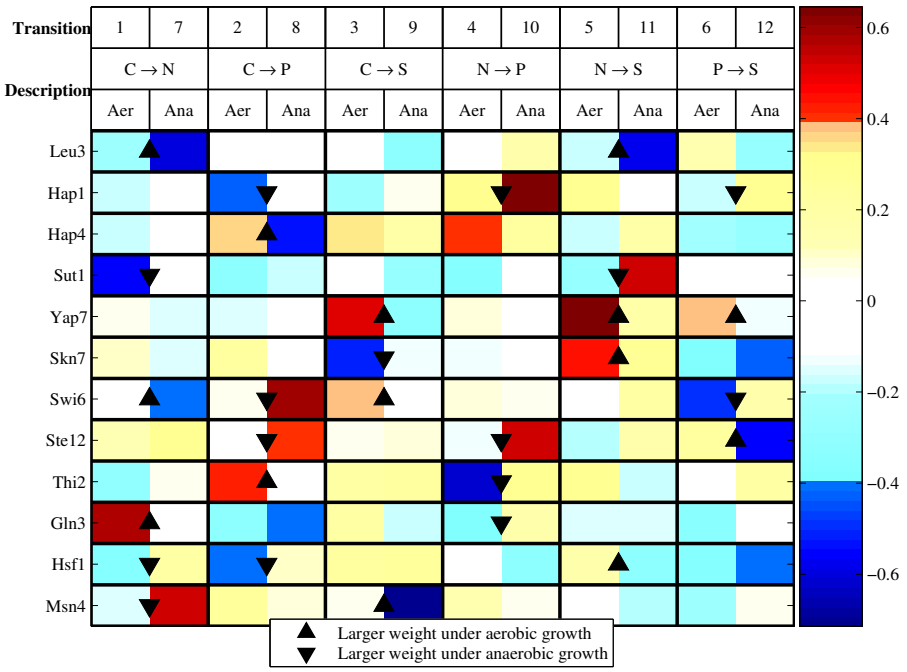



**Figure 5.3 – Visualization of W, representing the TF activity of twelve TFs under the six nutrient limitation transitions, both aerobically and anaerobically.**
Large positive weights (red) indicate involvement in upregulation, negative weights (blue) refer to downregulation. Triangles indicate a significance difference in weights ($P \leq 0.05$) for a nutrient limitation transition between the aerobic and anaerobic case.

especially at the anaerobic transition from carbon to nitrogen limitation.
Here, we come to the crux of our work. Our approach is able to infer TF activity related to very specific changes in combinatorial cultivation parameters. The algorithm that is especially designed for the combinatorial setup of nutrient limitations and oxygen supply in the employed microarray dataset, not only provides unprecedented detailed insight into the behavior of yeast's metabolism and respiration at the transcriptional level, but also in terms of TF activity. Thus, we do not find many TFs that are globally related to particular nutrients. (These have already been identified in previous studies, e.g. [Bar 03, Knij 05]). More specifically, we identify lots of TFs that are not primarily related to the metabolism of a particular nutrient, yet seem to play a more specific and subtle (and as of yet unknown) regulatory role at these transitions between nutrient limitations. The involvement of these TFs demonstrate the complex and multiple regulatory roles that they exhibit in transcriptional regulation in different processes.
The involvement of a particular TF in different processes has of course been established by many independent studies. For example, Mcm1 is a known multifunctional protein which plays a role both in the initiation of DNA replication (cell-cycle) and in the transcriptional regulation of diverse genes [Pass 89]. A more recent study also suggests that

**Table 5.3 – Significantly enriched ($P \leq 5 \cdot 10^{-5}$) MIPS and GO functional categories for the nine groups defined at the carbon to nitrogen limitation transition.**
Processes other than metabolism, energy and cellular transport are underlined.

| no. | MIPS | GO |
|---|---|---|
| I | metabolism | |
| II | energy, oxidative stress response | response to stress |
| III | metabolism, complex cofactor/cosubstrate binding | |
| IV | tetracyclic and pentacyclic triterpenes biosynthesis | lipid metabolism, steroid metabolism and biosynthesis |
| V | metabolism of the pyruvate family and D-alanine, mitochondrion | cellular biosynthesis, nitrogen compound biosynthesis, a.o. |
| VI | | lipid metabolism, steroid metabolism and biosynthesis |
| VII | mitotic cell cycle and cell cycle control, cellular transport, a.o. | |
| VIII | energy, respiration, a.o. | aerobic respiration, generation of precursor metabolites, a.o. |
| IX | energy, respiration, transported compounds, a.o. | oxidative phosphorylation, transport, a.o. |

in response to changes in their nutritional states, yeast cells modulate the activity of global regulators like Mcm1 via posttranscriptional regulation induced by the flux of glycolysis [Chen 95]. The identification of Mcm1 as a regulator in the carbon to nitrogen limitation transition, where the glycolysis flux changes dramatically, thus strengthens and even broadens this hypothesis.
In general, the results provide new regulatory roles for many TFs in metabolism and respiration. Additionally, the results underline the complexity of transcriptional regulation in the cell, especially when taking into account the fact that changes in nutrient and oxygen availability can not be seen in isolation from (or even modulate) cell-cycle (e.g. [Newc 03]) and energy processes (e.g. [Wu 04]) and is even known to evoke stress responses (e.g [Jami 98, Gasc 02]). To strengthen this notion, enrichment in MIPS [Mewe 97] and GO [Ashb 00] functional categories was computed. Table 5.3 displays the results for the transition from carbon to nitrogen limitation. These results also indicate that many non-metabolic processes play a role in the nutrient and oxygen availability transitions.
In the remainder of this section we focus on three identified TFs and hypothesize about their putative role in regulation at specific transitions. Here, we also demonstrate the power of the Procrustes approach in clarifying more subtle patterns of regulation.

*Leu3*

Leu3 is the main transcriptional regulator of branched-chain amino acid metabolism and has been extensively studied [Kohl 03, Boer 05]. To exactly meet the demands of protein synthesis, the activity of Leu3 is modulated by $\alpha$-isopropylmalate ($\alpha$-IPM), an intermediate of the branched-chain amino acid pathway. As a result Leu3 can act as both an activator and a repressor. Our findings indicate an oxygen-specific role of Leu3 in several nutrient limitation transitions. Figure 5.4 displays the expression behavior at transitions for the regulon of Leu3. Many genes are downregulated at the $C \rightarrow N$ and $C \rightarrow S$ transitions under anaerobic conditions in comparison to the same transitions grown under aerobic conditions. (This can be seen by the much larger number of green boxes in transition 7 w.r.t. transition 1 and similarly for transitions 9 and 3). Furthermore, when going from aerobic to anaerobic carbon-limited growth many genes are upregulated. (All above mentioned relations were found significant in the hypergeometric tests, as can be seen in Figure 5.2.) The involvement of Leu3 as a repressor and activator at these transitions has not been established before. Personal communication with the first author of Boer *et al.* [Boer 05] lead to the observation that the expression pattern of Leu3's regulon under anaerobic growth is quite remarkable. If it were

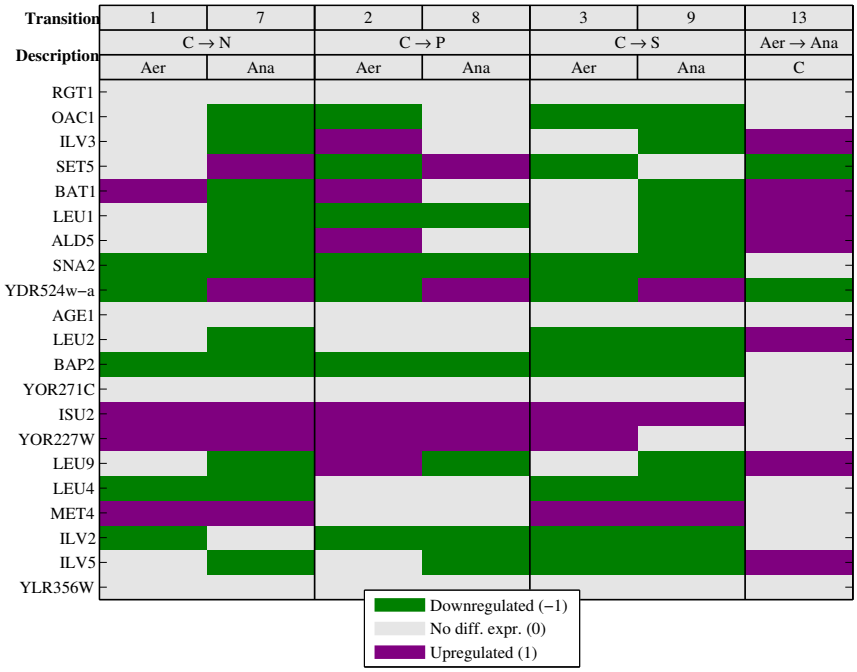



**Figure 5.4 – Expression behavior at the condition transitions for the Leu3 regulon.**
Part of the transition matrix **T**, indicating the expression behavior of all genes to which TF Leu3 can bind (upstream) for the transitions that are displayed in Figure 5.2.

the case that also at the anaerobic $C \rightarrow P$ transition many genes were downregulated, one could associate this to mitochondrial capacity [Tai 05], since the synthesis route of $\alpha$-IPM is mainly located in the mitochondrion. However, this is not the case. Possibly, regulation of Leu3 under anaerobic growth can be linked to different concentrations of $\alpha$-IPM, caused by different concentrations of Acetyl-CoA and ATP/ADP that change at the transitions. However, this is not more than speculation at this point.

*Yap7*

The TF Yap7 was only significantly associated with upregulation of genes when going from nitrogen to sulfur limited ($N \rightarrow S$) aerobic growth. (This result is not visible in Figure 5.2.) The Procrustes analysis, however, shows a more interesting pattern of regulation. In Procrustes, the TF binding data set is employed to explain the different expression behavior between *all* the aerobic and anaerobic nutrient limitation transitions simultaneously. (This in contrast to employing the hypergeometric distribution, where hypotheses can only be tested individually.) Furthermore, the orthonormality constraint emphasizes the difference in activity of a TF at different transitions. When investigating the weights assigned to Yap7, we see that not only in $N \rightarrow S$ the weight is significantly larger under aerobic growth, but also in the case of the other transitions moving towards sulfur limited growth; $C \rightarrow S$ and $P \rightarrow S$. (See Figure 5.3.) For the other transitions the weights are near zero. Thus, we can hypothesize that Yap7 (a

member of the yeast bZip family of proteins, of which two other members can only be linked indirectly to sulfur metabolism [Mend 05]) is involved in regulation under aerobic sulfur-limited growth, thereby assigning a very specific putative regulatory role for this poorly studied TF.

*Ste12*

Also in the case of Ste12, the Procrustes rotation confirms and broadens the relationships as established by the hypergeometric tests. From the literature it follows that Ste12 is a transcription factor that binds to the pheromone response element (PRE) to regulate genes required for mating and also functions with Tec1 to regulate genes required for pseudohyphal growth [Yeas]. Additionally to these functionalities, we find it to upregulate genes when entering a phosphorus-limited state, especially when no oxygen is present. See the condition transition weights for Ste12 in Figure 5.3. (Note that the $S \rightarrow P$ is not in the table, but it is justified to expect that these weights will be the complement of the $P \rightarrow S$ transition.)

## 5.5 Discussion

Today's main use and strength of bioinformatics tools is generating hypotheses on all types of relationships and functionalities of and between quantifiable parameters inside and outside the cell. Specific biological experiments are, however, still required to validate the automatically generated hypotheses before accepting them as newly discovered knowledge. The common trend of focusing on large compendia of intracellular measurement datasets is often in contrast with the biologist's very specific field of research. These broad approaches are able to recognize global patterns in the data, but miss specific and subtle effects that characterize the complex reality of the cell.
In this research we applied a tailor-made informatics approach on a small, well defined dataset. This enabled us to provide the biologist with very detailed hypotheses about the specific biological processes of interest. The basis for this work is the systematic combinatorial setup of the cultivation conditions under which yeast was grown in highly controllable chemostats. Incorporation of TF binding data through stringent statistical tests as well as a Procrustes rotation, led us to infer the activity of TFs at transitions between the different cultivation conditions. In contrast to common approaches the generated regulatory network thus shows the actual changes in conditions that lead to the activation of TFs.
Incorporation of (changes in) conditions is a crucial part of regulatory networks and in the quest for simulation of the complete regulatory mechanisms within the cell, will be part of more elaborate future analysis. Additionally, future work will aim at interpreting the uncovered results, not only by literature, but also by performing specific follow-up experiments. Furthermore, the uncovered results have proved to be very interesting, and therefore encourage application of similar techniques to other systematically setup datasets.

# CHAPTER 6

# CHEMOSTAT STEADY-STATE MICROARRAY COMPENDIUM

This chapter presents a chemostat microarray compendium consisting of 170 microarray measurements with 55 unique conditions. These conditions are characterized by the settings of ten different cultivation parameters. Using a regression strategy the influence of cultivation parameters on gene expression is investigated. Here, the main focus is on the influence of combinations of cultivation parameters on gene expression. The explained variance of gene expression patterns and functional enrichment of gene clusters is evaluated for regression models both including and excluding these combinatorial effects. Also, the influence of cultivation parameters on gene expression is used in the interpretation of shake-flask-based transcriptome studies and for guiding functional analysis of (uncharacterized) genes and pathways. This study demonstrates that modeling the combinatorial effects of environmental parameters on the transcriptome is crucial for understanding transcriptional regulation. In this way, the goal of systems biology to investigate and understand the interactions between different components and/or levels in biological systems can be complemented by an equally integrative approach towards the complex environmental context in which cells grow and survive.

This chapter is published as:

'Combinatorial effects of environmental parameters on transcriptional regulation in *Saccharomyces cerevisiae*: A quantitative analysis of a compendium of chemostat-based transcriptome data'

Theo A. Knijnenburg, Jean-Marc Daran, Marcel A. van den Broek, Pascale Daran-Lapujade, Johannes H. de Winde, Jack T. Pronk, Marcel J.T. Reinders and Lodewyk F.A. Wessels

BMC Genomics, Volume 10 No 53, January 2009

## 6.1 Abstract

Microorganisms adapt their transcriptome by integrating multiple chemical and physical signals from their environment. Shake-flask cultivation does not allow precise manipulation of individual culture parameters and therefore precludes a quantitative analysis of the (combinatorial) influence of these parameters on transcriptional regulation. Steady-state chemostat cultures, which do enable accurate control, measurement and manipulation of individual cultivation parameters (e.g. specific growth rate, temperature, identity of the growth-limiting nutrient) appear to provide a promising experimental platform for such a combinatorial analysis.
A microarray compendium of 170 steady-state chemostat cultures of the yeast *Saccharomyces cerevisiae* is presented and analyzed. The 170 microarrays encompass 55 unique conditions, which can be characterized by the combined settings of 10 different cultivation parameters. By applying a regression model to assess the impact of (combinations of) cultivation parameters on the transcriptome, most *S. cerevisiae* genes were shown to be influenced by multiple cultivation parameters, and in many cases by combinatorial effects of cultivation parameters. The inclusion of these combinatorial effects in the regression model led to higher explained variance of the gene expression patterns and resulted in higher function enrichment in subsequent analysis. We further demonstrate the usefulness of the compendium and regression analysis for interpretation of shake-flask-based transcriptome studies and for guiding functional analysis of (uncharacterized) genes and pathways.
Modeling the combinatorial effects of environmental parameters on the transcriptome is crucial for understanding transcriptional regulation. Chemostat cultivation offers a powerful tool for such an approach.

## 6.2 Introduction

The transcriptional program of a cell is to a large extent determined by its extracellular environment. Signaling pathways, transcription factors (TFs) and chromatin remodeling mediate the transcriptional response that enables the organism to adapt to changed conditions. In order to understand the transcriptional response to changes in the extracellular environment, a large majority of the transcriptome analysis studies are based on the comparison of a single "reference" condition against a different condition. Genes that show a different transcript level between the two situations are often labeled "upregulated" or "downregulated" in the non-reference situation. This binary mode of analysis does not take into account the fact that many genes are influenced by multiple environmental stimuli and regulated by multiple TFs. The rate of transcription of a gene is, in general, the net result of the integration of multiple inputs. Consequently, transcriptional responses to individual environmental stimuli may be strongly dependent on the experimental context in which they are studied.
While the context dependency of transcriptional responses has been acknowledged as an important factor by several authors (e.g. [Bar 03, Lusc 04]), it is only rarely considered in experimental design and in data interpretation. Three main reasons can be identified for this omission. First, most transcriptome studies on micro-organisms are based on shake-flask cultivation, in which key physiological parameters such as the specific growth rate and nutrient availability change continuously and cannot be adequately controlled.

This makes it impossible to quantify the context dependency of transcriptional responses. Secondly, research questions are often approached from a one-dimensional perspective, in which differential gene expression is completely attributed to the difference between a condition of interest and a reference condition. This strategy is implicitly incorporated into the two-channel microarray experimental design, where the ratio of intensities from the channels represents the gene expression ratio between the condition of interest and the reference condition. A final factor that complicates meaningful combinatorial analyses of transcriptional regulation is that integration of data from different studies and laboratories may be hampered by differences in experimental procedures for microarray experiments (including the use of different microarray platforms, mRNA extraction, normalization and summarization algorithms [Tan 03, Bamm 05]).
The "one-dimensional" design of transcriptome studies, as outlined above, ignores combinatorial effects of growth parameters, i.e., the possibility that repetition of the measurements in, for example, a different medium composition or temperature, might yield a different transcriptional response to the same change in the parameter of interest. Recently, a relatively small number of studies have quantitatively explored the context dependency of transcriptional regulation in chemostat cultures of the yeast *Saccharomyces cerevisiae* [Tai 05, Knij 07, Nico 07, Cast 07]. In steady-state chemostat cultures, individual environmental parameters can be manipulated in a controlled manner and at a fixed specific growth rate [Hosk 05, Dara 09]. This forms an important advantage over the use of shake flasks and other batch cultivation procedures, in which changes in environmental parameters affect specific growth rate, thus precluding the dissection of primary responses to environmental parameters and indirect effects of a different specific growth rate. Recent chemostat-based studies have demonstrated that, indeed, specific growth rate itself has a strong effect on transcriptional regulation in *S. cerevisiae* [Rege 06, Cast 07, Brau 08]. Additionally, chemostat experiments on combinatorial effects of macronutrient limitation, oxygen availability and temperature provided compelling evidence for the impact of context dependency [Tai 05, Knij 07, Tai 07].
The goal of the present study is to quantify the influence of cultivation parameters on gene expression and specifically focus on the influence of combinatorial (or context-specific) effects of the cultivation parameters. To this end, we have compiled a microarray compendium of well-defined chemostat cultivations of yeast and employed a computational framework to analyze the effect of the cultivation parameters on gene expression. The compendium of chemostat-based transcriptome datasets is comprised of 170 microarray measurements, which have been performed over the past years in the Kluyver Centre's yeast research program. These measurements, the majority (111 out of 170) of which have been previously published separately, encompass 55 unique growth conditions with (mostly three) independent biological replicates for each condition. Across the 55 different conditions, there are ten varying cultivation parameters, such as growth-limiting substrate, specific growth rate, aeration, pH and temperature. A forward step-wise regression model was designed and applied to quantify the (combinatorial) effect of individual environmental parameters on transcriptional regulation. This strategy is based on the assumption that the observed difference in the transcript level of a gene between two microarrays can be fully attributed to the difference in environmental parameters (and measurement noise) between these arrays. The results show that mainly due to the accurate control and measurement of the growth parameters enabled by steady-state chemostat cultivation, this assumption holds to a large degree. By employing these results from the regression analysis, we explore the significance of

context dependency throughout the compendium. Its applicability for functional analysis of (uncharacterized) genes and pathways is demonstrated using the inferred causal relationship between environmental parameters and gene expression.

## 6.3 Results and Discussion

This section starts by describing the steady-state chemostat microarray compendium and the regression analysis to assess the influence of cultivation parameters on gene expression. Then, the combinatorial effects of cultivation parameters on the transcriptome are investigated using enrichment tests and through biological interpretation of these effects on genes of functional categories and biochemical pathways. To demonstrate the usefulness of the compendium, this section concludes by presenting two case studies concerned with, firstly, the functional analysis of uncharacterized and dubious genes, and secondly, the interpretation of shake-flask-based transcriptome studies using the compendium.

### 6.3.1 Inferring the influence of cultivation parameters on gene expression

The *Saccharomyces cerevisiae* laboratory reference strain CEN.PK 113-7D (*MAT*a) was grown at steady state in chemostat cultures under 55 different conditions. A condition can be characterized by a specific configuration of the settings of ten different cultivation parameters. One of these cultivation parameters is the available carbon source. Throughout the compendium five different carbon sources were used, i.e. acetate, ethanol, galactose, glucose and maltose. Thus, these five compounds form the settings that the cultivation parameter carbon source can assume. Table 6.1 provides an overview of the settings for all cultivation parameters. Figure 6.1 depicts the expression levels of the gene *UPC2* across all 55 conditions. The lower part of this figure is a schematic representation of the settings of the ten cultivation parameters over all conditions. Note that the expression levels are absolute expression levels that come from a single-channel microarray system and not relative expression levels, where a reference condition is employed. A regression model was designed to assess the influence of the cultivation parameters on gene expression. The model was applied to all differentially expressed genes individually. (A large majority (6005 of 6383) of the genes in the *S. cerevisiae* genome was found to be differentially expressed in at least one of the 55 conditions.) Using a step-wise approach, the regression model iteratively selects significant predictors in order to reconstruct the expression pattern of a gene.
Here, the cultivation parameters form the predictors. We incorporated single effects and two types of combinatorial effects. See Figure 6.2 for a schematic example of genes that are influenced by these effects. A single effect is constituted by one setting of one cultivation parameter. For example, limiting element carbon is a predictor. (This will be a significant predictor for genes, which show differential expression between carbon-limited growth and growth that is limited by the residual quantity of other substrates.) In Figure 6.2 gene g1 responds solely to a single effect. The first type of combinatorial effect is constituted by applying the logic AND function between two settings of two different cultivation parameters. For example, limiting element carbon AND aerobic growth (in short: aerobic carbon-limited growth) form such a combinatorial effect. Of course, the

**Table 6.1 – Settings within the cultivation parameters.**
This table presents the different settings within each of the ten cultivation parameters. Each of the 55 conditions in the chemostat compendium is characterized by a combination of settings of the ten cultivation parameters. The colored matrix in Figure 6.1 is a schematic representation of the settings of the cultivation parameters for each condition. Abbreviations of cultivation parameter settings used in the schematic representation are stated between parentheses in this table.

| Aeration type | C-source | N-source | S-source | Limiting element |
|---|---|---|---|---|
| Aerobic | Acetate (Ace) | Ammonium chloride (A.cl.) | Methionine (Met) | Carbon |
| Anaerobic | Ethanol (Eth) | Ammonium sulfate (A.s.) | Sulfate | Iron (Iro) |
| | Galactose (Gal) | Asparagine (Asp) | | Nitrogen |
| | Glucose | Leucine (Leu) | | Phosphorus (Pho/Phos) |
| | Maltose (Mal) | Methionine (Met/Meth) | | Sulfur (Sul/Sulf) |
| | | Phenylalanine (Phen) | | Zinc (Zin) |
| | | Proline (Pro) | | |

| Growth rate | Temperature (C) | pH | Extra compound | Protocol |
|---|---|---|---|---|
| 0.03 | 12 | 3.5 | Acetate (Ace) | B |
| 0.05 | 30 | 5 | Benzoate (Benz) | A |
| 0.1 | | 6.5 | $CO_2$ | |
| 0.2 | | | Ethanol 18.72 mM (Eth) | |
| | | | Ethanol 9.38mM (Eth) | |
| | | | Formate (For) | |
| | | | Propionate (Pro) | |
| | | | Sorbate (Sor) | |
| | | | Tween 80 (Twe) | |
| | | | none | |

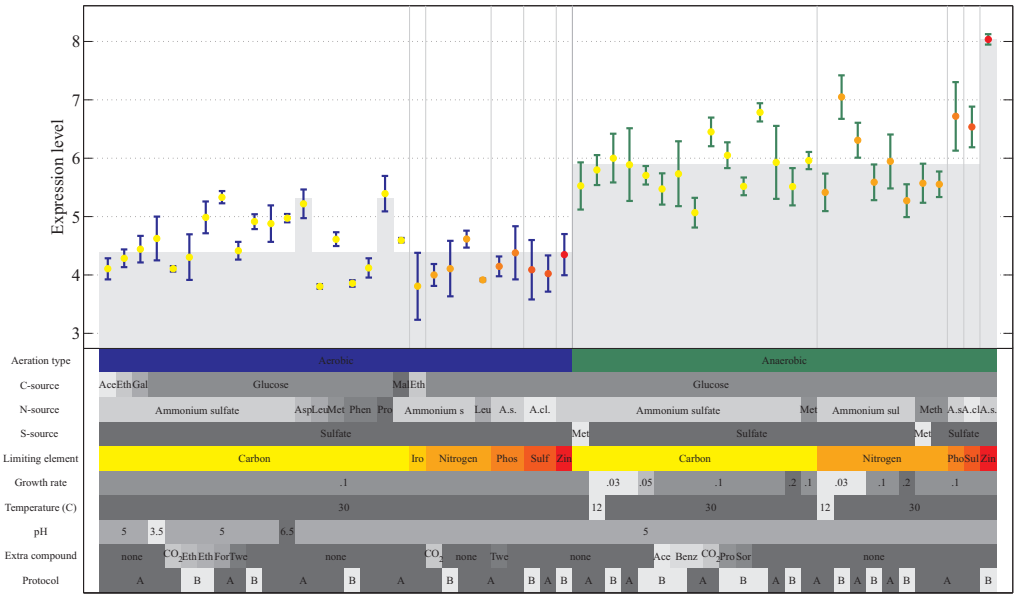



**Figure 6.1 – Expression levels of *UPC2* across the 55 cultivation conditions.**
The colored matrix is a schematic representation of the settings of the ten cultivation parameters over the 55 conditions. The colored lanes indicate the cultivation parameters that are employed to order the experiments, in this case, aeration type and limiting element. The applied regression model was able to explain 71% of the variance in the expression of this gene. The model selected one significant single effect, i.e. aeration type, and two significant combinatorial effects, i.e. aeration type anaerobic together with limiting element zinc and the usage of proline or asparagine as nitrogen source. The reconstructed expression pattern based on these three effects is indicated by the shaded area.

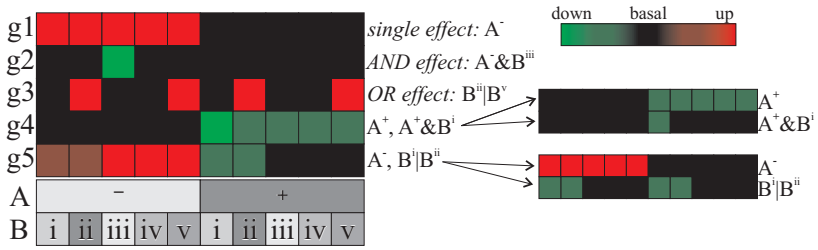



**Figure 6.2 – Schematic representation of the normalized expression patterns of genes affected by a single effect, combinatorial effect or a mixture of these.** In this example there are two cultivation parameters, A and B, which can assume two and five different values, respectively. Genes g1, g2 and g3 are affected by a single effect, AND effect and OR effect, respectively. The expression of genes g4 and g5 is constituted by the influence of both a single effect and a combinatorial effect.

cell's transcriptome and metabolome are known to respond in a combinatorial fashion to particular environmental conditions or parameters. That is, the simultaneous presence of certain environmental factors results in a transcriptional and metabolic state that is not a simple aggregation of the states reached based on the single presence of one of these factors. For example, when glucose is present, it is utilized in different ways by *S. cerevisiae*, depending on the presence of oxygen. Including these AND effects enables the systematic investigation of the influence of combinations of cultivation parameters on gene expression. Gene g2 in Figure 6.2 responds to an AND effect. The second type of combinatorial effect is constituted by applying the logic OR function on two different settings within the same cultivation parameter. Here, carbon-limited OR iron-limited growth forms an example. This effect is included, because we expect that closely related settings within a cultivation parameter, e.g. similar carbon sources, will have a similar effect on gene expression. Gene g3 in Figure 6.2 responds to an OR effect. In the case of *UPC2* (Figure 6.1), the regression model successively selected the single effect aeration type, the AND combinatorial effect anaerobic zinc-limited growth and the OR combinatorial effect nitrogen source proline or asparagine. (Note that cultivation parameter aeration type can assume only two settings, i.e. aerobic growth and anaerobic growth. Since these two predictors are mutually redundant, only one of them (aerobic growth) is included as a predictor in the regression model and labeled as aeration type. A positive regression coefficient for aeration type indicates that the gene is more highly expressed under aerobic conditions; a negative coefficient indicates the reverse scenario.) The regression model keeps on adding cultivation parameters as predictors, until no further significant improvement can be made. For example, for g4 in Figure 6.2 the single effect $A^+$ is selected first, followed by the combinatorial effect $A^+\&B^i$. See Methods section for details.

### 6.3.2 The expression of many genes responds to combinatorial effects

For most genes the regression model was able to explain 60 to 80% of the variance, which is present in their expression patterns across the 55 conditions. See Figure 6.3a. The amount of explained variance does not depend that much on the average expression level of a gene, although there is a steady increase in explained variance with increasing

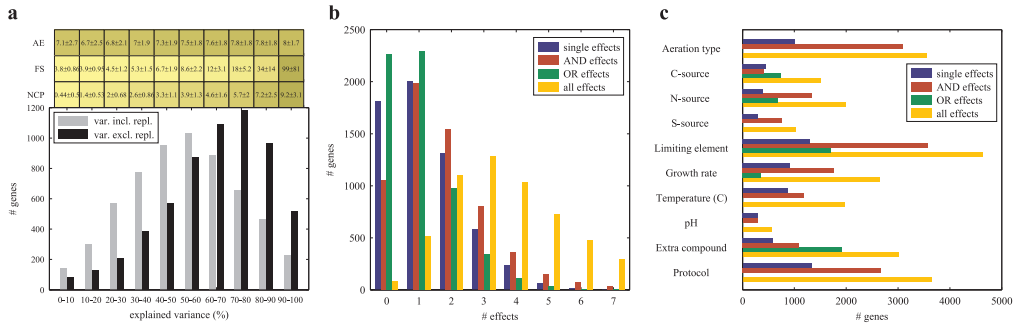



**Figure 6.3 – General statistics of the applied regression model.**
**a:** Histogram plot indicating how much variance within the gene expression patterns could be explained by the regression model for all (differentially expressed) genes. The black bars indicate the percentage of explained variance when excluding the variance present in the replicates, and which, therefore, cannot be explained by the regression model. Above the histogram are the mean and variance of the average expression level (AE), the F-statistic (FS) and the number of selected cultivation parameters (NCP) for the groups of genes with explained variance (including replicate variance) as stated on the x-axis of the histogram. **b:** Histogram plot indicating the number of single and combinatorial effects as well as the total number of effects that were selected to explain the observed gene expression patterns. **c:** Histogram plot indicating the number of genes influenced by particular cultivation parameters, either as a single effect, AND effect, OR effect or independent of the effect type ('all effects'). The 'all effects' bar is not the sum of the other three, because genes can be affected by a cultivation parameter both as a single effect and as a combinatorial effect.

average expression level. Much more important is the degree to which a gene is differentially expressed. The F-statistic, i.e. the ratio between the variance of the average expression levels across the 55 conditions and the average replicate variance across these conditions, is strongly correlated with the degree to which the gene's expression pattern can be reconstructed. The expression levels of genes with small F-statistics are obscured by measurement noise and do not differ significantly between the growth conditions. Also not surprisingly, when more significant cultivation parameters are selected by the regression model, more of the variance of the gene can be explained. Figure 6.3b,c outlines which and how many cultivation parameters were selected to reconstruct the expression patterns of all genes. On average, a gene is influenced by 1.25 (± 1.18) single effects, 1.73 (± 1.43) AND effects and 1.01 (± 1.04) OR effects. The limiting element, aeration type and protocol (which is dealt with in more detail below) are the most prominent factors that influence gene expression behavior. Here it should be noted that the setup of the cultivation parameters in the compendium is not fully combinatorial, i.e. not all possible combinations of cultivation parameters are present in the dataset. For example, across the 55 conditions, 53 have been cultivated under pH 5, while only a single condition was performed with a lower pH (3.5) and similarly for a higher pH (6.5), thereby precluding combinatorial effects between the higher or lower pH and other environmental parameters. Thus, the numbers of genes, which are influenced by a particular cultivation parameter (as visualized in Figure 6.3c), are biased by the number of different settings of the cultivation parameters and the number of combinations of cultivation parameters present in the compendium. Anyhow, the results indicate that the expression of many genes is influenced, not only independently by particular cultiva-

tion parameters, but also in a combinatorial fashion, i.e. there are many combinatorial effects between cultivation parameters that affect gene expression behavior.
The regression analysis was repeated using only the single effects as predictors. For most genes this resulted in a lower percentage of explained variance. See Figure 6.4a. Of course, this result could be expected based on the fact that many combinatorial ef-

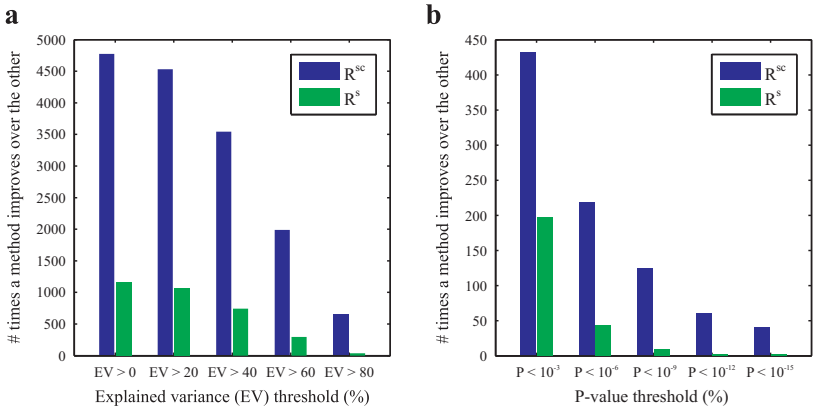



**Figure 6.4 – Comparison between the regression analysis including including both the single *and* the combinatorial effects ($R^{sc}$) and the regression analysis including *only* the single effects ($R^{s}$).**
**a:** Histogram plot indicating how many times one method ($R^{sc}$ or $R^{s}$) leads to a higher percentage of explained variance (EV) of a gene given that the EV of this gene is larger than the EV threshold (x-axis) for at least one of both methods. **b:** Histogram plot indicating how many times one method ($R^{sc}$ or $R^{s}$) leads to a higher enrichment value (lower P-value) for a functional category given that the enrichment of this category is below a P-value threshold (x-axis) for at least one of both methods.

fects were selected as significant predictors in the original regression model. Subsequent enrichment analysis provided additional evidence for combinatorial regulation. Genes, of which their expression levels are manipulated by a particular single effect or combinatorial effect, were grouped and checked for functional overrepresentation. Additional file 1 of [Knij 09] online provides an overview of all enrichment analysis results. It reveals the many cases ($> 1000$) in which a particular combination of environmental parameters leads to the up- or downregulation of a group of functionally related genes. Also, functional enrichment was compared between the regression analysis including both single and combinatorial effects and the analysis including only single effects. Genes were clustered based on their reconstructed expression patterns that were obtained for both regression models and these clusters were evaluated for enrichment in functional annotation categories. Figure 6.4b shows that the inclusion of the combinatorial effects leads to increased functional enrichment, and thus further substantiates the existence of the combinatorial influence of the presence of environmental factors and the importance of modeling them. Additional file 2 of [Knij 09] online describes the complete comparison between the regression models including and excluding the combinatorial effects.

### 6.3.3 The sample preparation protocol has a large impact on the measured gene expression levels

As indicated in Table 6.1 and Figure 6.1 the tenth cultivation parameter is termed "Protocol". Unlike the nine other parameters, "Protocol" is not directly related to the cultivation conditions under which yeast is grown, but refers to the protocol to process RNA samples. Several years ago, an improved sample preparation kit was introduced [Affy 04]. This kit obviated the need for the expensive and time-consuming poly-A mRNA purification step included in the original procedure. The decision to omit the purification step, which was also made in other yeast research groups, was supported by information indicating that samples prepared with or without this step were similar [Affy 00]. Thus, two different protocols were used to generate the chemostat compendium's samples for microarray hybridization: Protocol A and Protocol B. The main difference between these protocols is that Protocol A includes the polyA-mRNA isolation step (with cDNA synthesis being performed on purified mRNA), while Protocol B excludes the purification step (with cDNA synthesis being performed on total RNA). (The Methods section and Additional file 3 of [Knij 09] online provide the complete details on both protocols.)

As apparent from Figure 6.3c, the measured transcript levels of many genes appeared to be influenced by the protocol. Enrichment analysis revealed a significant overrepresentation of characterized genes amongst the genes that have higher apparent transcript levels under protocol B; all three GO root-categories (biological process, cellular component and molecular function) were highly enriched. On the other hand, significantly many uncharacterized genes yielded higher apparent transcript levels under protocol A. Further investigation revealed a trend between transcript level and protocol influence: Genes with higher average expression level tended to yield a higher transcript level in protocol B and genes with a lower average transcript level tended to yield lower transcript levels under protocol B (Figure 6.5). In general, uncharacterized genes have a lower expression than characterized genes, which explains the results from the enrichment analysis. Further evidence for this hypothesis is found when analyzing the genes that encode ribosomal proteins (RP genes), whose mRNA's are highly abundant. Again, significantly many RP genes exhibit higher expression when analyzed with protocol B (middle and bottom plots in Figure 6.5).

The relationship between mRNA abundance (expression level) and protocol is only weak and does not hold for each gene individually. It may, for example, be influenced by the average length of the polyA-tail of different transcripts. Indeed, analysis of mitochondrial genes lacking a poly-A tail demonstrated a large influence of the protocol. Of the 52 transcripts on the microarray representing mitochondrial genes, 27 (amongst which 16 unique mitochondrial genes) were influenced by the protocol, i.e. the regression model selected the protocol as a significant predictor of the expression pattern of these genes. All these 27 mitochondrial genes showed a higher (apparent) transcript level under protocol B. These results illustrate that not only different microarray platforms, labs, and strains, but also the hybridization preparation steps can affect the outcome of microarray analyses. This strongly underlines previous warnings on the challenges involved in comparing microarray results from different experiments.

The chemostat compendium allows us to adequately model the influence of the hybridization protocol on expression. In particular, the compendium contains 18 growth conditions (9 sets of two), where the only differing cultivation parameter is the protocol

setting: The growth conditions were identical in these nine cases, only the protocol was different. This provides extra statistical power in the regression procedure and enables us to successfully model the protocol effect. This allows us analyze the influence of the environmental cultivation parameters without interference of the protocol's confounding effect.

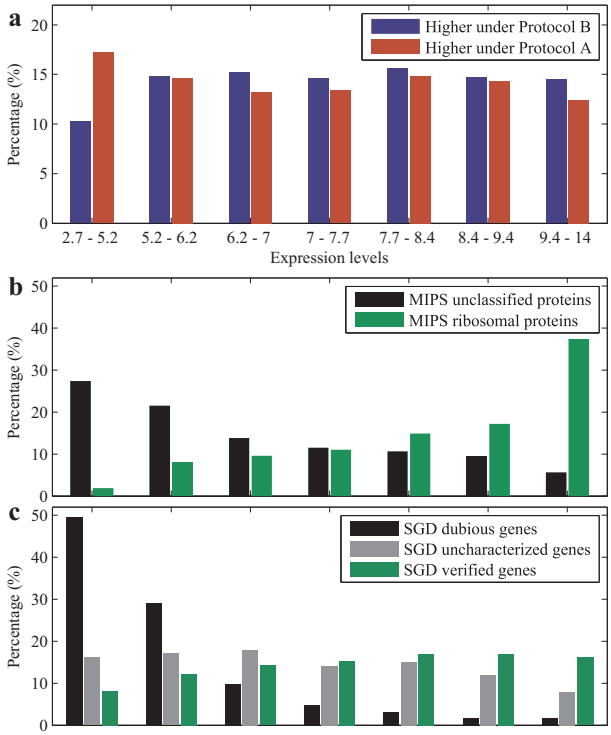



**Figure 6.5 – The influence of the protocol on gene expression.**
All genes that are affected by the modifications to the protocol, either as a single effect or as an interaction effect, are analyzed. First, the mean expression levels of these genes across all 55 conditions are computed. The genes are divided in seven groups based on their mean expression levels such that each group holds the same amount (i.e. 14,3%) of the genes. Each group is characterized by a lower and a higher bound on the expression value; these two numbers represent the range of the mean expression levels of the genes within the group. Also, we dichotomize the genes into the ones with positive regression weights (i.e. upregulation under Protocol B with respect to Protocol A) and the ones with negative regression weights. **a:** The blue bars indicate the percentage of genes with positive regression weights (higher under Protocol B) across these groups (or expression ranges). Similarly, the red bars indicate these percentages for the genes with negative coefficients (higher under Protocol A). **b,c:** For the same ranges, each bar represents the percentage of genes in the range annotated to a particular functional category over all of the genes that are annotated with this category and affected by the protocol.

### 6.3.4 Functional categories are specifically associated with combinations of environmental parameters

Many functional categories are specifically influenced by a combinatorial effect. Many genes within such a category are influenced by a combinatorial effect, whereas none or

**Table 6.2 – MIPS functional categories specifically associated with combinatorial effects.**
The combinatorial effects 'carbon source acetate OR ethanol' and 'Limiting element phosphorus OR sulfur' are specifically associated to the listed MIPS functional categories. P-values of the enrichment of genes within these categories that are affected by the combinatorial effect are given in the rightmost column. Also, enrichment P-values of genes affected by each and by both of the single effects that constitute this combinatorial effect are given.

| MIPS category | Enrichment P-values single effects | | | comb. effect |
|---|---|---|---|---|
| | Acetate | Ethanol | both | Acetate \| Ethanol |
| METABOLISM | 0.065 | 0.077 | 1 | $7.8 \cdot 10^{-18}$ |
| metabolism of glutamate | 1 | 0.048 | 1 | $1.4 \cdot 10^{-6}$ |
| C-compound and carbohydrate metabolism | 0.027 | 0.082 | 1 | $1.4 \cdot 10^{-22}$ |
| C-compound and carbohydrate utilization | 0.02 | 0.043 | 1 | $1.3 \cdot 10^{-17}$ |
| C-compound, carbohydrate catabolism | 0.2 | 1 | 1 | $8 \cdot 10^{-13}$ |
| sugar, glucoside, polyol and carboxylate catabolism | 0.44 | 1 | 1 | $9.3 \cdot 10^{-11}$ |
| ENERGY | 0.013 | 1 | 1 | $1 \cdot 10^{-17}$ |
| glycolysis and gluconeogenesis | 1 | 1 | 1 | $3.8 \cdot 10^{-9}$ |
| tricarboxylic-acid pathway | 1 | 1 | 1 | $2.2 \cdot 10^{-11}$ |
| | Phosphorus | Sulfur | both | Phosphorus \| Sulfur |
| transcriptional control | 0.13 | 0.017 | 1 | $4.3 \cdot 10^{-8}$ |
| RNA processing | 0.86 | 0.32 | 1 | $1.5 \cdot 10^{-6}$ |
| rRNA processing | 0.5 | 0.83 | 1 | $3.3 \cdot 10^{-6}$ |

only a few genes are affected by the single effects that constitute this combinatorial effect. See Methods section for these details. This analysis was performed on all MIPS categories. In total 153 significant combinatorial effect-MIPS category pairs were identified. These are depicted in Additional file 4 of [Knij 09] online. Here, we focus on the biological interpretation of two such combinatorial effects: Carbon source acetate OR ethanol, and, Limiting element phosphorus OR Sulfur. See Table 6.2.
The first example is provided by the OR effect of carbon sources ethanol and acetate on metabolism and energy household. These C2-compounds share a drastically different impact on central metabolism when compared to using the sugars glucose, maltose and galactose as carbon source. During growth on sugars, all metabolic building blocks can be derived from glycolysis, the tricarboxylic acid cycle and the pentose phosphate pathway, while during growth on C2-compounds, gluconeogenesis and the glyoxylate cycle are essential for the provision of some of these precursors. Furthermore, the higher ATP requirement for biosynthesis during growth on the C2-compounds implies that, at a fixed specific growth rate, dissimilatory fluxes have to be higher with the C2-compounds than with a sugar as the sole carbon source. This is supported by the significant shared influence of the C2 carbon sources on the genes of gluconeogenesis and the tricarboxylic acid pathway.
Besides this and other examples that can be easily explained by current knowledge, there are also many interactions that might represent as of yet unknown regulatory mechanisms. For example, we find that the limiting elements sulfur and phosphorus have a similar effect (i.e. OR effect) on transcription regulation genes. A close inspection of the genes influenced by this OR effect revealed the presence of five genes encoding subunits of Mediator (*MED3/PGD1* (complex tail), *MED7* and *MED10/NUT2* (middle), *MED11* and *MED18/SRB5* (head)), an evolutionarily conserved coregulator of RNA polymerase II [Pepp 05] and nine genes encoding chromatin remodeling enzymes (*ARP7*, *GCN5*, *HST2*, *RIF1*, *RSC6*, *RVB2*, *SFH1*, *SNF6* and *SPT8*). In eukaryotes,

gene transcriptional regulation depends on a complex interplay between signal transduction, specific and general gene regulators and complexes that modify chromatin and RNA polymerase II. Under sulfur limitation *S. cerevisiae* adapts its transcriptome in order to reduce the expression of sulfur rich genes and proteins [Fauc 02, Boer 03]. This response is mediated by Met4 the main sulfur metabolism regulator. The transcriptional changes upon phosphate limitation are mainly related to high affinity phosphate transport, phosphate assimilation and polyphosphate metabolism [Boer 03, Tai 05]. Although *S. cerevisiae* requires the transcription of different specific genes under sulfur or phosphate limitations, it is tempting to speculate that the mechanisms that govern the transcription control of these specific sets of genes are shared and depend on shared mechanisms involving specific subunits of the Mediator complex. Such high degree of specificity was demonstrated with the implication of Med2 (a Mediator tail subunit) in the regulation of the low iron response regulon [Pepp 05].

### 6.3.5 Combinatorial regulation within biochemical pathways provides further insight into sulfur metabolism and scavenging

As demonstrated above, we can assess whether groups of genes are influenced by particular (combinations of) environmental parameters using enrichment tests. This opens up the interesting possibility to correlate new and previously known patterns of regulation of individual genes with the regulation of larger families of genes connected to each other in pathways. In contrast to other gene groups, in a metabolic pathway clear connections exist between the gene products and their functions, which allows for more in-depth analysis. Here, we focus on biochemical pathways as described in SGD, which depict the series of chemical reactions converting metabolites, and the enzymes catalyzing these reactions. Enrichment analysis indicated that 5 of the 9 downloaded 'SGD superpathways' were influenced by at least one significant combinatorial effect (at $P < 10^{-3}$, $Q < 0.08$).

An illustrative example is presented by analyzing the expression profiles of the gene family involved in sulfur- and sulfur containing amino acid-metabolism in yeast (Figure 6.6). Sulfur amino acid biosynthesis involves a considerable number of enzymes required for the de novo biosynthesis of methionine and cysteine and the recycling of organic sulfur metabolites. Expression of the genes encoding the enzymes for this metabolic network is tightly controlled by the available sulfur source, through modulation of the intracellular S-adenosyl-methionine levels. Six different cultivation parameters were significantly often selected to explain the expression patterns of the genes in this pathway ($P < 10^{-3}$). Five of these are combinatorial cultivation parameters. Not surprisingly, the only single effect is sulfur limitation, which causes the upregulation of ten out of the eighteen genes [Thom 97]. See box 1 of the bars near the enzyme names in Figure 6.6. Despite large variations in expression under different combinations of conditions, many of the *MET*-, *CYS*-, *SAM*- and *HOM*-genes invariably respond to the presence of methionine in the growth medium by clearly reduced expression. See Figure 6.7, which depicts the normalized gene expression patterns of all genes of the pathway. This response is independent of the presence of oxygen or growth limitation by carbon or nitrogen sources. Only in the case where methionine is utilized both as sulfur and nitrogen source *and* methionine is the limiting element, we observe that the expression of the corresponding genes is not reduced (but even slightly induced, mimicking the (known) response under sulfur

limitation). This explains the selection of the combinatorial effects involving methionine depicted by boxes 4, 5 and 6 in Figure 6.6.

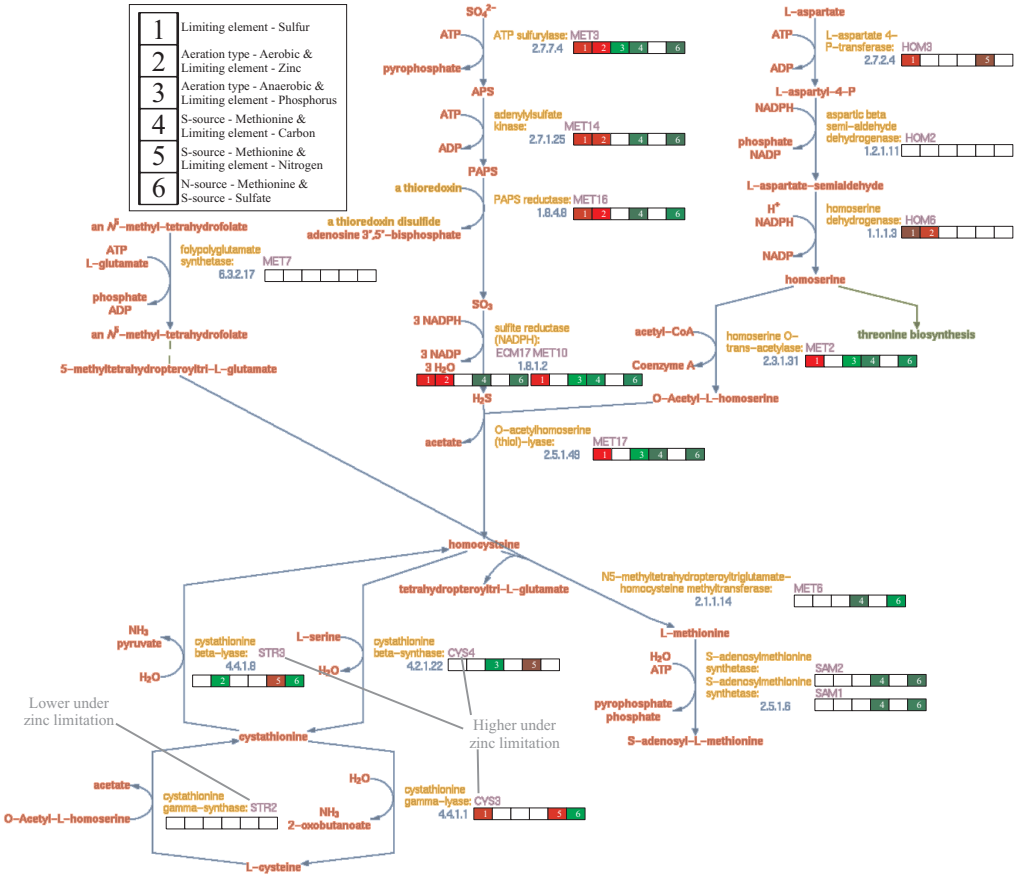



**Figure 6.6 – Superpathway of sulfur amino acid biosynthesis.**
Near each enzyme (gene product) is a bar representing the regression weights of the six significant cultivation parameters. These parameters are stated in the legend in the upper-left corner of this figure. A blank box indicates that the cultivation parameter is not selected by the regression model. Red and green boxes indicate positive (upregulation) and negative (downregulation) regression weights, respectively. Darker colors indicate larger regression weights.

Interestingly, two genes involved in this sulfur-metabolizing network in part respond differently. *HOM2*, which is involved in homoserine biosynthesis, responds reciprocally to the availability of methionine in the growth medium compared to the other *HOM* genes, especially under aerobic conditions. The same observation is made for *STR2*, which is involved in cystathionine biosynthesis. (In Figure 6.7 magenta boxes mark the conditions, where methionine is part of the growth medium.) This discrepancy is indicative of a differential regulatory mechanism operating between the *HOM2*, *HOM3* and *HOM6* genes of the homoserine pathway, and of the complex regulation of the transsulfuration pathway, involving *CYS3*, *CYS4*, *STR2* and *STR3*. Further detailed analysis would be required to elucidate the molecular mechanisms operating in these differential

combinatorial controls. Such differential controls operating within a pathway are likely to be involved in intricate flux balancing mechanisms.

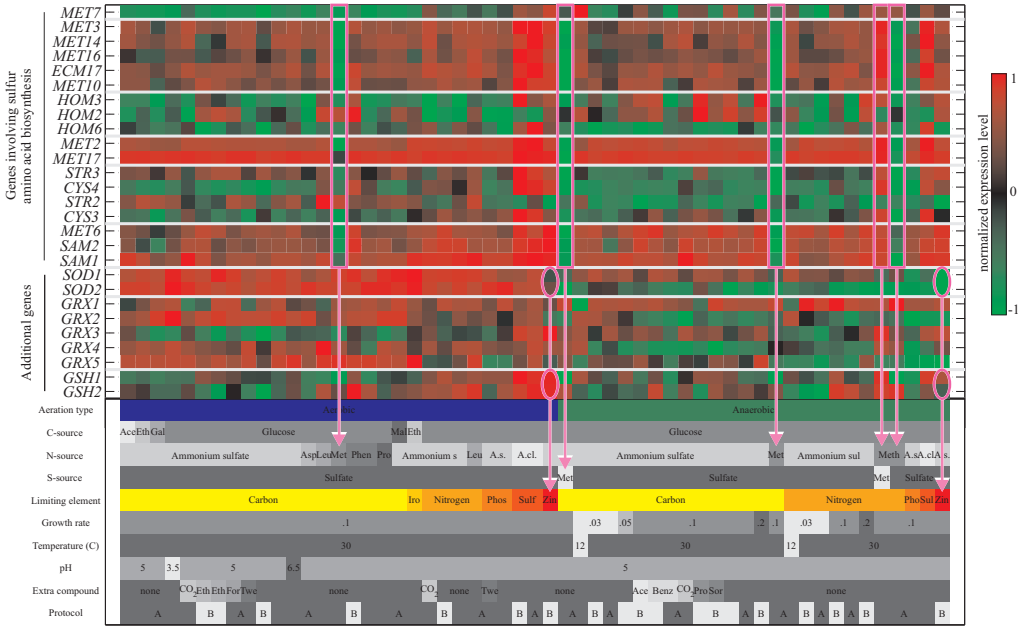



**Figure 6.7 – Normalized gene expression patterns of the genes that are part of the superpathway of sulfur amino acid biosynthesis and additional genes discussed in the text.**
The expression values of each gene are linearly scaled to range from $-1$ to 1. Here, $-1$ represents the lowest expression value and 1 indicates a gene's highest expression value. These normalized expression patterns are projected on the green-black-red colormap to derive the heatmap visualization.
Separate branches of the pathway are indicated by the grey horizontal lines. For the group denoted as "Additional genes", the grey horizontal lines split the genes in functionally related groups. The magenta boxes and arrows indicate the cultivation parameters, where methionine is used as nitrogen or sulfur source. The magenta ellipses and arrows highlight the expression levels of the *SOD* and *GSH* genes under zinc limitation.

Surprisingly, for many of the genes in the pathway under investigation expression levels under zinc limitation are almost as high as under sulfur limitation, especially under aerobic conditions. Moreover, the genes of the transsulfuration pathway are highly expressed under zinc and sulfur limitation, yet lower expressed under the other nutrient limitations. Also here, *STR2* responds reciprocally and is lower expressed under zinc limitation. Although transcript levels per se cannot be used as flux indicators, this expression behavior is consistent with an upregulation of the flux towards cysteine under zinc limitation via the increased synthesis of the corresponding enzymes. (See the graph structure of the pathway near cysteine in Figure 6.6.) The exact nature of this response is not immediately apparent. However, it provides an interesting hypothesis on the oxidative stress response of *S. cerevisiae* under zinc limitation. As previously described [Mora 00], a "first line of defense" in oxidative stress response is formed by the superoxide dismutase genes *SOD1* and *SOD2*, which are induced under aerobic conditions. See Figure 6.7. The dithiol glutaredoxin genes *GRX1* and *GRX2* [Luik 98], and the monothiol glutaredoxin genes *GRX3-GRX5* [Moli 04], which also participate in

the response against oxidative stress, exhibit highly differential transcriptional profiles. This may provide new insight into the specific roles for each of the varying combinations of glutaredoxins under different growth conditions. Surprisingly, under zinc limitation not only the Cu, Zn-dependent *SOD1* gene is lower expressed; also the *SOD2* gene, encoding the mitochondrial superoxide dismutase, which is dependent on Mn and not on Zn, is much less induced. A boost in glutathione synthesis apparently takes over the main defense, since the glutathione synthase genes *GSH1* and *GSH2* are clearly induced, especially under zinc-limited aerobic conditions. This can be seen from the magenta ellipses in Figure 6.7. This fits with the fact that significantly many genes in the sulfur scavenging pathway are upregulated under zinc-limited aerobic growth, presumably leading to an induced cysteine pool, cysteine being one of the three components of the tripeptide glutathione.

### 6.3.6 Functional characterization of uncharacterized and dubious genes using the chemostat compendium

In a recent review [Pena 07] it was pointed out that many ($> 1000$) genes in the yeast genome are still uncharacterized. Possible reasons for this include genetic redundancy, lack of strong growth phenotype and the possibility that not all of them are real genes. Additionally, genes may be involved in environmental and metabolic responses, which are normally not queried in the lab. Concerning the "characterized" genes, it can be noted that the function of many annotated genes is derived from large-scale studies, and hence, in-depth detailed analysis is lacking for these genes.
We conjecture that the visualization of the expression behavior of a gene over the conditions of the compendium, together with the identification of the significant cultivation parameters to which the gene responds, provides valuable information regarding gene function. With this information, one can design directed biological experiments or assays that probe a specific pathway or activity in order to advance towards the functional characterization of a gene. We mapped our regression results to SGD's genome snapshot, upon which the division of *Saccharomyces cerevisiae* ORF's into verified ORF's, uncharacterized ORF'S and dubious ORF's in [Pena 07] was based. For 1350 genes the regression model lead to a good reconstruction of the observed expression pattern (explained variance including replicate variance $> 70\%$). According to SGD, 1009 of these genes were verified ORF's; 286 were uncharacterized and 54 were classified as dubious genes. Amongst the uncharacterized genes, many genes were found to be expressed under conditions which have not been extensively studied before. For example, amongst the 286 uncharacterized genes, five genes are most significantly influenced by zinc limitation, i.e. zinc limitation was the first condition selected by the regression model. One of these, *YOR387C*, is only expressed under zinc limitation. These results immediately link the function of a gene to a particular cultivation parameter or a specific biological process related to this cultivation parameter. The expression pattern of these five zinc responsive genes as well as the other genes to be discussed in this section are visualized in Figure 6.8. Also, amongst the 54 dubious genes, there are many genes that are highly expressed under one or a few cultivation parameters, while having a constant expression over the remaining conditions. For example, *YJL119C* is only highly expressed under phosphorus limitation. *YBL070C* also responds to phosphorus limitation, yet particularly when the yeast is grown aerobically. The expression of *YBR292C* is influenced by aerobic sulfur-limited growth and *YBL065W* is only expressed when grown

at a low temperature (12 °C). 35 of the 54 dubious genes were affected by the aeration effect or the interaction effect between carbon limitation and aeration. These genes were screened against a recent proteomics study, where expression data of yeast grown in aerobic and anaerobic carbon-limited chemostats was measured [Groo 07]. We found that for three genes unique peptides were quantified. This establishes the existence of the proteins encoded by these "dubious" genes. See Additional file 5 of [Knij 09] online for a list of the 54 dubious genes and details on the peptide identification. Notably, 51 of the 54 dubious genes are no longer present on YG 2.0, the successor of the Affymetrix YG S98 GeneChip, after comparative genomics [Kell 03] and phylogentic footprinting [Clif 03] approaches identified these as false ORF's. However, our analysis reveals a clear-cut influence of environmental conditions on the expression levels of many of these genes, implying that these genes *do* have a functional role, at least in the *Saccharomyces cerevisiae* strain that was used in this study.

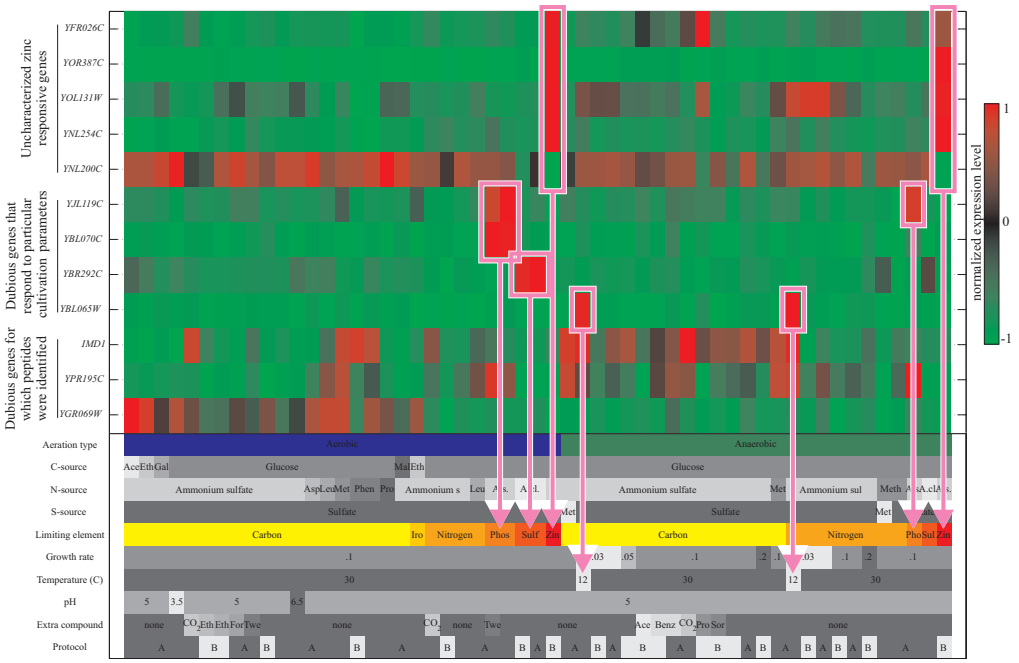



**Figure 6.8 – Normalized gene expression patterns for twelve uncharacterized or dubious genes.**
The expression values of each gene are linearly scaled to range from −1 to 1. Here, −1 represents the lowest expression value and 1 indicates a gene's highest expression value. These normalized expression patterns are projected on the green-black-red colormap to derive the heatmap visualization. The magenta boxes and lines highlight the cultivation parameters that influence the expression of the genes.

### 6.3.7 Analysis of shake-flask experiments with the chemostat compendium

Changes in the extracellular environment or perturbations on genetic level do not only affect (signaling) pathways in which the change or perturbation has direct involvement,

but can also impact the cell's viability, metabolism or other processes in the cell. For example, there are many experimental conditions and genetic perturbations that will impact the growth rate of the cell. For shake flask cultivations it is not possible to distinguish between the direct and indirect effects, since cultivation parameters like growth rate and nutrient availability cannot be controlled. This also confounds the analysis of gene expression data from shake flask experiments [Rege 06]. By screening a group of genes, which were grouped together on the basis of shake flask experiments, against the compendium, some of the confounding effects can be resolved. The group can be subdivided into clusters of genes that respond to particular environmental parameters within the compendium and thereby identify the cultivation parameters or biological processes that could have played a role in the original shake flask experiment, even when these have not been measured.

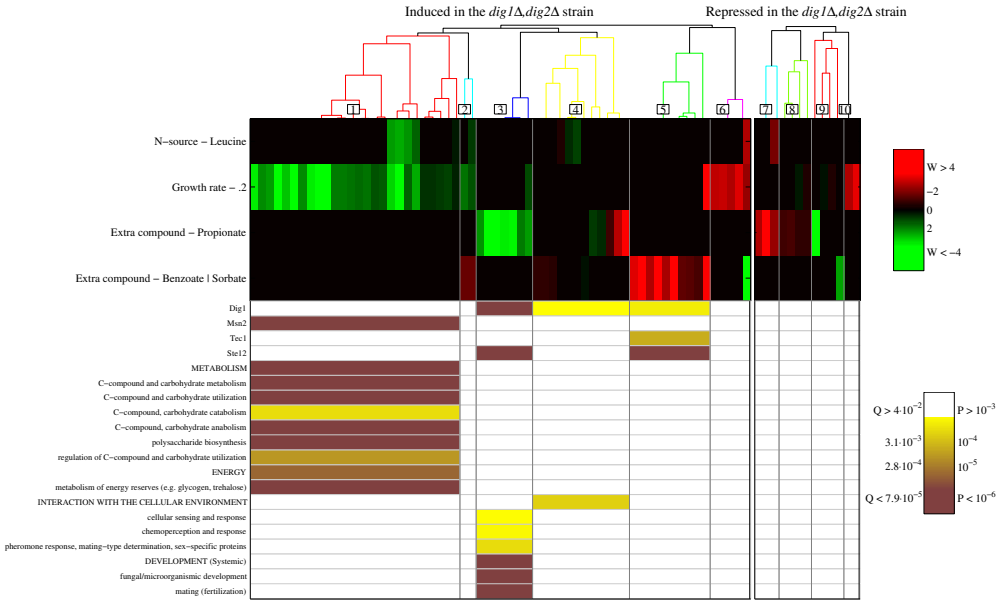



**Figure 6.9 – Analysis of two groups: The genes upregulated in a *dig1*Δ,*dig2*Δ strain and the genes downregulated in this strain.**
**Middle:** Normalized regression weights for the significant cultivation parameters across the gene groups. **Top:** The genes were clustered based on these regression weights. **Bottom:** Schematic representation of the enrichment P-values and related false discovery rates (Q-values) for each of the uncovered clusters when related to TF binding data and MIPS functional categories.

To this end, we apply the following strategy: First, we select the (combinatorial) cultivation parameters that are significant for the group under investigation. These are the cultivation parameters that are significantly often selected by the regression model to explain the expression pattern of the genes in the group when compared to the complete genome. Next, the genes are clustered based on the normalized regression coefficients under these cultivation parameters. Finally, these newly obtained clusters are consulted for enrichment of annotation categories. See Methods section for details. As an example, Figure 6.9 depicts the results of this analysis for the groups of genes, which were found to be induced or repressed in a *dig1*Δ,*dig2*Δ mutant strain grown in a shake-flask

[Hugh 00]. To make the induced and the repressed gene groups, we consulted the gene expression data of this study (i.e. the Hughes *et al.* yeast mutant microarray compendium [Hugh 00]). The induced group is formed by all genes that are upregulated by one fold-change or more in the *dig1*$\Delta$,*dig2*$\Delta$ mutant strain compared to the wild-type strain. The repressed group is formed in a similar fashion by identifying the genes that are downregulated by one fold-change or more.
The results show a clear difference between direct and indirect effects. On the one hand, the enrichment analysis on the TF binding data tells us that the genes in Clusters 3, 4 and 5 form a significantly large part of Dig1's regulon, i.e. the direct targets of TF Dig1. The known role of Dig1 and Dig2 in regulating mating-specific and pheromone-responsive genes is confirmed by the enrichment of these functional categories in Cluster 3. Also, binding sites of TFs Tec1 and Ste12, which together with Dig1 form a regulatory complex involved in mating and filamentation [Chou 06], are enriched for Cluster 5 and Clusters 3 and 5, respectively. Interestingly, the genes within Clusters 3, 4 and 5 were clustered together based on their response to the addition of organic acids propionate, benzoate and sorbate. (The clusters are characterized by the shared transcriptional response of their genes to these acids.) On the other hand, a large set of genes that is induced after the knockout of *DIG1* and functionally redundant *DIG2*, is affected by growth rate in the chemostat microarray compendium. See Clusters 1, 6 and 10. The genes of Cluster 1 show high enrichment for metabolism and energy functional categories as well as for general stress response TF Msn2. From this observation we conclude that besides the genes that are directly affected, the double knockout also has a large impact on the metabolism and energy household of the cell when grown in a shake-flask.

## 6.4 Conclusions

The compendium of chemostat-based transcriptome data is a valuable resource for yeast systems biology that can be queried online. Additional file 6 of [Knij 09] online contains the complete dataset (expression data and description of the cultivation conditions). Additional file 7 of [Knij 09] online is an interactive tool to visualize the gene expression across all conditions in the compendium; this file can be downloaded from the author's website.
Using a forward step-wise regression strategy, we were able to quantify the influence of (combinatorial) cultivation parameters on the expression of genes and (using enrichment tests) groups of functionally related genes. The regression results demonstrate the large extent to which regulation of individual genes results from the integration of multiple external signals. In fact, the analysis yielded only few "signature transcripts", i.e. transcripts whose level showed a unique up- or downregulation under a single condition in the compendium relative to all other conditions. This observation has important implications for the applicability of so-called signature transcripts to diagnose cellular status (e.g. starvation for a nutrient, stress or, in higher organisms, disease). Our results indicate that the "signature" status of a gene with respect to an individual environmental parameter can depend strongly on other ("background") environmental signals to which the cell is exposed. In this respect, it should be stressed that the current compendium of chemostat-based data represents only a minute fraction of the infinite range of combinatorial conditions to which yeast cells can be exposed in nature, in industry and in the laboratory.

The relevance of the proposed approach for functional analysis of genes and pathways is exemplified by the observed combinatorial effects of zinc and sulfur availability in the pathway of sulfur amino acid biosynthesis. Furthermore, the compendium approach has provided clear indications that 54 *S. cerevisiae* genes that had previously been labeled as 'dubious' and have even been removed from some commercial DNA microarrays, exhibited a specific and reproducible transcriptional response to some of the investigated culture conditions. These examples illustrate the potential for enabling more focused functional analysis studies through a correlation of a wide range of cultivation conditions and gene expression data. The results provide a strong incentive for further extending the range of cultivation conditions included in the compendium.

The systematic dissection of the impact of (combinations of) individual culture parameters on transcriptional regulation enabled by chemostat-based microarray analysis can be applied to interpret transcriptome data generated in less extensively controlled, but highly relevant cultivation conditions in industry and in the laboratory. This is exemplified by the additional interpretation of previously published data from shake-flask-based transcriptome analysis of a *dig1*$\Delta$,*dig2*$\Delta$ mutant (Figure 6.9).

In view of the excellent reproducibility of chemostat-based microarray analysis [Pipe 02], it should be possible to extend the compendium with data from other research groups, provided that yeast strain, cultivation procedures and procedures for microarray analysis are rigorously standardized. The effect of a change in the mRNA processing protocol, as identified in the regression strategy, provides a clear caveat on the possible impact of even small differences in experimental procedures.

One promising avenue to be explored is the use of the compendium in deriving transcriptional regulation networks. Given that changes in gene expression can be ascribed to changes in the activity of TFs and chromatin remodeling proteins, the compendium dataset provides the means to investigate how cultivation parameters influence the activity of the proteins that control transcription. Since the cultivation parameters, such as the employed carbon source, are closely linked to the actual molecular signals that are detected by the cell, it may be possible to also relate transporters and signaling cascades to the observed expression under different environmental conditions. This allows for a genome-wide analysis of the complete chain of regulatory relationships that cause changes in the extracellular environment to lead to changes in gene expression.

In the employed regression model, the (combinatorial) cultivation parameters are assumed to have an additive effect on gene expression. In previous work [Knij 07] the aeration type was modeled as a linear effect with both additive and multiplicative components. This approach was not possible for the cultivation parameters within the current framework. Furthermore, a more complex modeling or incorporation of higher-order effects results in a highly under-determined system and possible computational complexity issues. Given the high-degree of non-linearity in biological systems, the application of logic (Boolean) functions might provide a sensible alternative to the commonly used linear modeling. Irrespective of the structure of the models, incorporating combinatorial effects in models for (transcriptional) regulation is crucial. Only in this way, the goal of systems biology to investigate and understand the interactions between different components and/or levels in biological systems can be complemented by an equally integrative approach towards the complex environmental context in which cells grow and survive.

## 6.5 Methods

### 6.5.1 Chemostat cultivation and microarray data

Prototrophic *Saccharomyces cerevisiae* strain CEN.PK113-7D (*MAT*a) [Dijk 00] was grown at 30°C (or at 12°C) in 2-liter chemostats (Applikon) with a working volume of 1.0 liter as described in van den Berg *et al.* [Berg 96]. Cultures were fed with a defined mineral medium that limited growth by either carbon, nitrogen, phosphorus, sulfur, zinc or iron with all other growth requirements in excess and at a constant residual concentration. The dilution rate ranged from 0.03 to 0.2 $h^{-1}$. The pH was measured online and kept constant at 5.0 (or 3.5 and 6.5) by the automatic addition of 2 M KOH using an Applikon ADI 1030 bio controller. Stirrer speed was 800 rpm, and the airflow was 500 ml $min^{-1}$. Dissolved oxygen tension was measured online with an Ingold model 34-100-3002 probe and was above 50% of air saturation. The off-gas was cooled by a condenser connected to a cryostat set at 2°C, and oxygen and carbon dioxide were measured offline with an ADC 7000 gas analyzer. When required, anaerobic conditions were maintained by sparging the medium reservoir and the fermentor with pure nitrogen gas (500 ml $min^{-1}$). Furthermore, Norprene tubing and butyl septa were used to minimize oxygen diffusion into the anaerobic cultures [Viss 94]. Steady-state samples were taken after ∼10-14 volume changes to avoid strain adaptation due to long term cultivation [Fere 99]. Dry weight, metabolite, dissolved oxygen and gas profiles had to be constant over at least 3 volume changes before sampling for RNA extraction. The detailed culture media recipes, used in the 55 different conditions presented in this study, can be retrieved from the individual GEO [Geno] array reports. The GEO accession numbers can be found in Additional file 6 of [Knij 09] online.
In this study, two different sample preparation protocols were employed: Protocol A (for 36 of the 55 conditions) and Protocol B (for 19 of the 55). For Protocol A, sampling of the chemostat cultures, probe preparation and hybridization to the single-channel Affymetrix GeneChip YG S98 microarrays is described in Piper *et al.* [Pipe 02]. Protocol B has the following modifications with respect to Protocol A: In stead of harvesting ∼700 $\mu$g of total RNA and applying a Poly-A mRNA isolation step before cDNA synthesis (Protocol A), ∼15 $\mu$g of total RNA is harvested and the purification step is omitted (Protocol B). Thus, in Protocol B cDNA synthesis is performed on total RNA, while in Protocol A the synthesis is performed on Poly-A purified mRNA. Additional file 3 of [Knij 09] online provides the complete details on both protocols and references to the used AffyMetrix manuals.
Across the 55 conditions, ten different varying cultivation parameters can be identified. A cultivation parameter, e.g. the carbon source, is described as a categorical variable and contains two or more settings, e.g. the used carbon source can be either acetate, ethanol, galactose, glucose or maltose. Each condition is characterized by a configuration of these settings across the ten cultivation parameters. See Figure 6.1, Table 6.1 and Additional file 6 of [Knij 09] online for an overview of the relevant settings within the environmental parameters per condition. In total, 180 microarray measurements were performed. There is a variable number of independent biological replicates per condition, however for most (39) conditions three replicates were performed. Chip quality control, condensing probe intensities to gene expression levels and normalization was performed using GeneData Refiner Array [Gene]. 170 high quality chips, i.e. gradient severity $\leq 0.165$, defective area $\leq 0.5\%$ and outlier area $\leq 0.59\%$, were retained. Ten chips,

which did not meet these criteria were dismissed. The RMA algorithm was used to derive the log scale measure of the expression levels [Iriz 03]. Quantile normalization was applied to normalize between arrays [Bols 03]. The normalized expression data is given in Additional file 6 of [Knij 09] online. The raw array data used in this study can be retrieved at Genome Expression Omnibus [Geno] with series number GSE11452.

### 6.5.2 Detecting differential expression

A gene was called differentially expressed when 1) the gene was present in at least one of the arrays (present call P-value $< 0.05$) and 2) the gene showed significant differential expression in at least one condition (one-way ANOVA with 55 classes, P-value $< 0.05/9335$). 9335 is the total number of transcripts on the array.

### 6.5.3 Inferring the influence of cultivation parameters on gene expression using regression

A designmatrix was created, containing both main (or single) effects and interaction (or combinatorial) effects: Each setting within each cultivation parameter is represented by a binary indicator column with 170 entries. These columns represent the main effects, which indicate for each array whether the yeast was grown under the relevant setting of a particular cultivation parameter. Two types of combinatorial effects were included in the model, i.e. "AND" and "OR" effects. The AND interaction effect columns were obtained by applying the logical AND function to all possible pair-wise combinations of main effect columns. The OR interaction effect columns were obtaining by applying the logical OR function to all possible pair-wise combinations of main effect columns that are associated with the same cultivation parameter. Thus, only OR effects that are constituted of two settings within the same cultivation parameter were modeled. Redundant columns and columns with all zeros were removed. This resulted in the binary $[170 \times 227]$ designmatrix $\mathbf{D}$, which includes 38 single effects, 101 AND effects and 88 OR effects. A visualization of this matrix is found in Additional file 8 of [Knij 09] online.
A forward step-wise ordinary least squares regression strategy was applied to each gene individually:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{6.1}$$

Here, $\mathbf{y}_i$ denotes the measured gene expression level of a particular gene for array $i$, with $i = 1, \ldots, 170$; $\mathbf{X}$ is the predictor matrix, $\beta$ represents the regression coefficients and $\epsilon$ the error, which is assumed to be independent zero-mean normally distributed. Initially, $\mathbf{X}$ contains only the intercept, i.e. a column of 170 ones. In an iterative fashion, columns from $\mathbf{D}$ are added to $\mathbf{X}$. For this we applied a leave-one-out cross validation (loocv) scheme, where a single sample is used for testing, while the remaining (169) samples are used for training the regression model. This was repeated such that each sample is used once as the test data. The column from $\mathbf{D}$, with the smallest root-mean-squared (rms) loocv error and absolute regression coefficient larger than 0.3, was selected and added. The iterative process of adding columns is discontinued when the P-value, as output by a t-test that determines whether the regression coefficient significantly differs from zero, exceeds $0.05/227$. To prevent the inclusion of spurious AND effects, the following strategy is applied: When an AND effect column is selected, we check whether the addition in explained variance is larger than the addition is explained variance when

adding the two main effect columns that constitute the combinatorial effect. Only in this case, we add the AND effect column, otherwise the two main effect columns are added, provided that they satisfy the P-value threshold and their absolute regression coefficients are larger than 0.3.
Note that only coefficients larger than 0.3 or smaller than $-0.3$ were allowed. This was done to focus on large changes in gene expression. Inclusion of absolute weights smaller than 0.3 did not increase enrichment scores of functional categories (see next section). Although small regression coefficients might be biologically relevant, this indicates that there are also many spurious results amongst the small regression weights.
The choice for a step-wise regression approach is substantiated in Additional file 9 of [Knij 09] online.

### 6.5.4 Enrichment analysis

For each main or interaction effect, i.e. a column from **D**, we group all genes, for which that effect turned out to be a significant predictor with a positive regression coefficient (or regression weight). This procedure was also carried out to group genes with negative weights for the significant predictors, and to group genes irrespective of the sign of the weight. The latter grouping is basically a union of the genes with positive weights and the genes with the negative weights. In addition, for the cultivation parameters that can assume more than two settings, we group all the genes that respond to at least one of the settings of that cultivation parameter as a main effect. Basically, we select all main effect columns from **D** that represent a setting of one particular cultivation parameter and group the genes, for which at least one of these settings is a significant predictor. (Note that the cultivation parameters that can only assume two settings, i.e. Aeration type, S-source, Temperature and Protocol, only have one main effect column in **D**, since the two settings are mutually redundant and only one of them is included in **D**.) Also here, we make the distinction between positive and negative regression coefficients and the union of these. The hypergeometric test is employed to assess the significance of the overlap between all these groups and gene sets from GO [Ashb 00], MIPS [Mewe 97], KEGG [Kane 00] and TF binding data [MacI 06, Harb 04]. Additional file 1 of [Knij 09] online provides an overview of the significant results ($P < 10^{-6}$, $Q < 8.5{\cdot}10^{-4}$). Here, for each triplet of P-values, associated with the positive weights, the negative weights or all weights, the most significant (smallest P-value) is selected and color coded accordingly. See page 1 of Additional file 10 of [Knij 09] online for a flowchart describing the steps of this analysis.

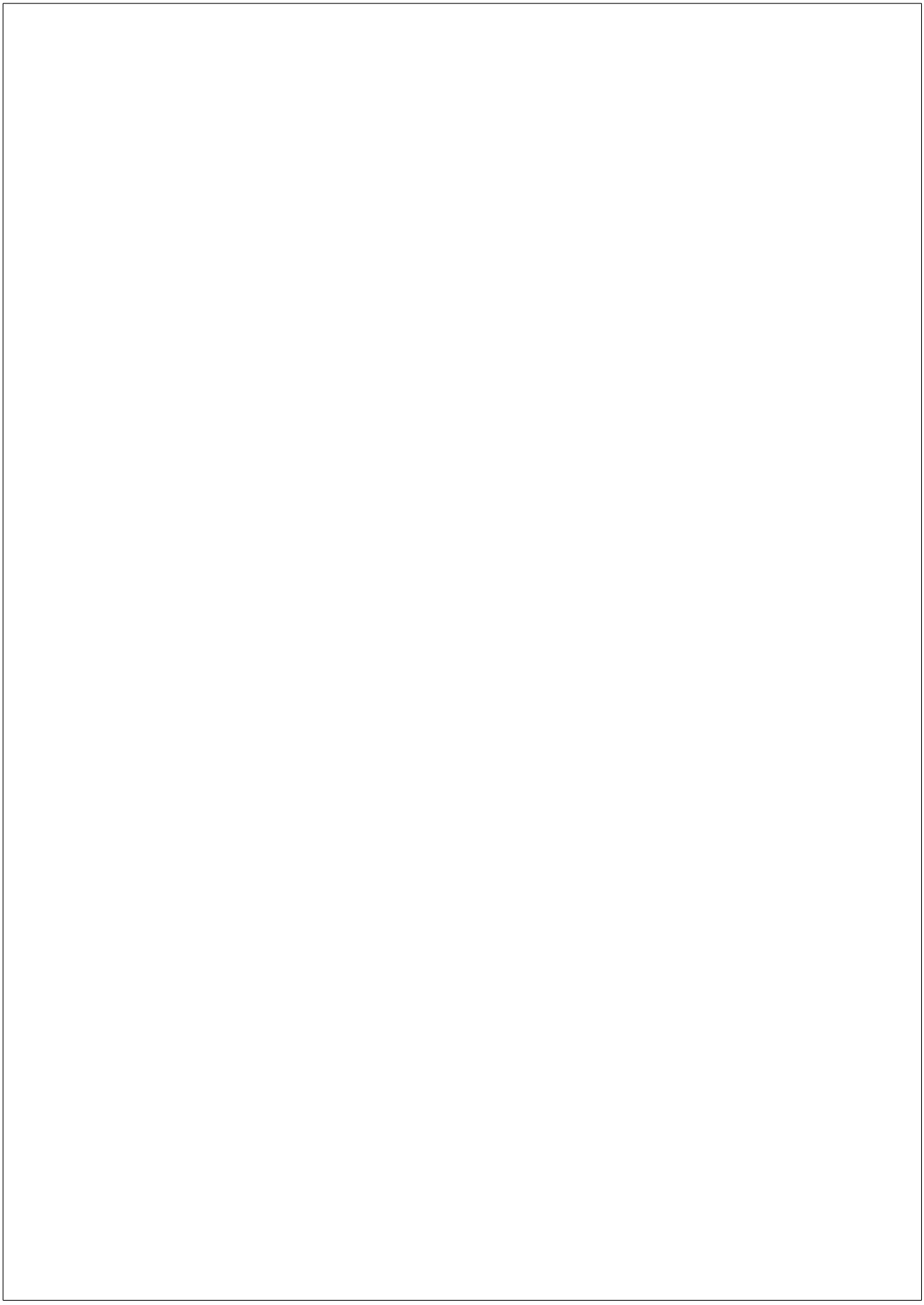
### 6.5.5 Functional categories specifically influenced by a combinatorial effect

To find a combinatorial effect that is specific for a functional category we group all the genes for which this effect was selected as a significant predictor by the regression model (irrespective of the sign of the weight). Also, for this effect, we make three other gene sets by grouping the genes which are influenced by 1) one of the single effects that constitute the combinatorial effect 2) the other single effect and 3) by both these single effects. (If a gene is influenced by both the combinatorial effect and a single effect, we only consider the effect that was selected first and then add this gene to the appropriate group.) Functional categories, which are overrepresented in the first group ($P \leq 10^{-5}$)

and not overrepresented in the three other groups ($P \geq 10^{-2}$) are called "specifically influenced by the combinatorial effect". See page 2 of Additional file 10 for a flowchart describing the steps of this analysis.

### 6.5.6 Clustering of genes based on regression coefficients

Given a group of genes, the hypergeometric test is employed to select those (interaction) effects, i.e. columns from **D**, which are significantly often selected by the regression model for the genes in this group when compared to all genes in the genome. Columns with $P < 10^{-5}$ are kept. Next, we create matrix **R**, which contains the normalized regression weights for the selected columns of all genes in the group under investigation. The normalized weights of a gene are obtained by dividing the original regression weights by the variance of the gene. A consensus clustering algorithm [Grot 06] is applied to cluster the genes based on the normalized regression weights in **R**: The data is clustered using a Bayes mixture of Gaussians EM algorithm. The number of clusters is varied from 2 to 20 (or the number of genes in the group if this is smaller than 20) and repeated 50 times for each number of clusters. The total of all clusterings is used to build a co-occurrence matrix, which indicates how many times a pair of genes was found in the same cluster amongst all clusterings. This co-occurrence matrix is transformed into a distance matrix. The distance matrix is zero, when a pair of genes was clustered together in all attempts; the matrix is one, when a pair never clustered together. We apply hierarchical clustering with complete linkage on this distance matrix and cut the dendrogram at 0.9 to create the final clusters. These clusters are consulted for enrichment of annotation categories using the hypergeometric test as explained before. See page 3 of Additional file 10 for a flowchart describing the steps of this analysis to create matrix **R**.



# CHAPTER 7

# COMBINATORIAL INFLUENCE OF TFS

In this chapter the regression results from Chapter 6 are used to construct regulatory transcription networks. Here, TF binding data is employed to 'explain' the influence of cultivation parameters on gene expression. The method described in this chapter aims to estimate under which cultivation parameters a TF becomes active as an enhancer or a repressor to (co-)regulate the expression of a gene. The interplay between activated enhancers and repressors that bind a gene promoter determine the possible up- or downregulation of the gene. The model is translated into a linear integer optimization problem and solved accordingly. This study is the first to demonstrate how environmental parameters can be employed to derive transcriptional regulation networks.

This chapter is published as:

'Combinatorial influence of environmental parameters on TF activity'

Theo A. Knijnenburg, Lodewyk F.A. Wessels and Marcel J.T. Reinders

Bioinformatics, Volume 24, Special issue: ISMB 2008 Conference proceedings, p. i172-i181, July 2008

## 7.1 Abstract

Cells receive a wide variety of environmental signals, which are often processed combinatorially to generate specific genetic responses. Changes in transcript levels, as observed across different environmental conditions, can, to a large extent, be attributed to changes in the activity of transcription factors (TFs). However, in unraveling these transcription regulation networks, the actual environmental signals are often not incorporated into the model, simply because they have not been measured. The unquantified heterogeneity of the environmental parameters across microarray experiments frustrates regulatory network inference.
We propose an inference algorithm that models the influence of environmental parameters on gene expression. The approach is based on a yeast microarray compendium of chemostat steady-state experiments. Chemostat cultivation enables the accurate control and measurement of many of the key cultivation parameters, such as nutrient concentrations, growth rate and temperature. The observed transcript levels are explained by inferring the activity of TFs in response to combinations of cultivation parameters. The interplay between activated enhancers and repressors that bind a gene promoter determine the possible up- or downregulation of the gene. The model is translated into a linear integer optimization problem. The resulting regulatory network identifies the combinatorial effects of environmental parameters on TF activity and gene expression.

## 7.2 Introduction

TFs mediate the activation or repression of gene expression by binding specific regulatory sequences (motifs) in gene promoters. The combinatorial interactions of multiple TFs play an essential role in transcriptional regulation. A classical example is *E. coli*'s lactose system, where the *lac* operon is expressed only if the concentration of TF CRP is high and that of TF LacI is low. Presently, many studies have revealed an important role for combinatorial interactions between different TFs in establishing the complex patterns of gene expression [Bala 06]. The advent of high-throughput genomic measurement techniques enabled the application of genome-wide computational approaches aimed at inferring these regulatory relations. Sequence data, microarray gene expression data and ChIP-chip TF binding data have been integrated in many different ways to derive regulatory networks. Several approaches fit expression data using linear regression models, where the predictors are the TFs, i.e. their binding potential or number of motifs in a gene promoter [Buss 01, Gao 04, Nguy 06]. The effect of multiple TFs on gene expression is modeled as the weighted sum of the contribution of individual TFs. Combinatorial regulation by TFs, i.e. synergistic or antagonistic effects of multiple TFs on gene expression, are not incorporated into these models. Most methods that do include combinatorial effects limit the scope to TF pairs, e.g. [Das 04, Yu 06, Chan 06, Bonn 06]. Bonneau *et al.* [Bonn 06] employ continuous versions of logic functions (OR, AND, and XOR) of the activities of TF pairs as additional predictors in the regression model. Although, in principle, these methods can be extended to model the combinatorial effects of more than two TFs, the model will be too complex to reliably estimate its parameters given the currently available data. Segal *et al.* [Sega 03] and Yeang and Jaakkola [Yean 06] present quite different approaches to the problem of combinatorial regulation in transcription networks. Segal *et al.* constructed regulatory networks by building

decision trees. Genes are grouped into regulatory modules, which are defined by a hierarchical decision tree, where the decisions at the nodes of the tree are based on the expression levels of TFs. In Yeang and Jaakkola, a TF is characterized as an enhancer or a repressor, being either necessary or sufficient to cause up- or downregulation of a gene. The combinatorial function of all TFs that can bind a gene promoter is modeled as the consensus prediction of the individual TFs. It should be noted that these two approaches, as well as many of the abovementioned ones, rely on the often incorrect assumption that the activity of a TF can be derived from the expression of the gene that codes for the TF.

So far, regulatory networks have been presented as graph structures showing the (combinatorial) regulatory effect of TFs on individual genes, modules of similarly expressed or otherwise related genes or on other TFs. The extracellular signals that trigger the activation or deactivation of TFs are usually not part of the generated network. Yet they could provide more direct and trustworthy evidence to infer TF activity than other signals, such as the gene expression of a TF. Three main reasons for their exclusion can be identified. First, many studies on yeast are based on shake-flask cultures, where parameters like growth rate and nutrient availability are continuously changing and cannot be controlled or accurately measured. Consequently, conditions can not be accurately defined. Second, very often research questions are approached from a single perspective, i.e. a condition of interest is compared to a reference condition. Differential gene expression is then attributed to the difference between the condition of interest and the reference condition. These approaches ignore combinatorial effects of growth parameters, the presence of which have been established by various studies, e.g. [Knij 07, Rege 06, Cast 07]. That is, if the measurements were repeated using a different medium composition or temperature, chances are that a different set of differentially expressed genes would be identified. Thus, these approaches only model the *differences* between growth conditions, and not the growth conditions *themselves*. Note that this strategy is implicitly incorporated into two-channel microarray measurements, which output the gene expression ratio between the condition of interest and the reference condition. Third, when combining different microarray experiments, differences in mRNA extraction protocols, microarray platform, and possibly normalization and summarization algorithms, add to the already large amount of unquantified heterogeneity amongst experimental conditions [Tan 03, Bamm 05].

The context dependency of regulatory networks has been identified and acknowledged in many studies. For example, in Bar-Joseph *et al.* [Bar 03] annotation data is employed to identify the biological context in which the inferred regulatory interactions are assumed to take place. In Luscombe *et al.* [Lusc 04] condition-specific regulatory networks were derived. In this case, condition-specific refers to one of five phenomena (cell cycle, sporulation, DNA damage, stress response or diauxic shift), which were investigated with five different microarray datasets. Myers and Troyanskaya [Myer 07] propose a Bayesian approach for context-sensitive integration of diverse genomic data. Note however, that in these approaches, the *precise* environmental conditions under which the microarray measurements were taken are *not* included in the model. In this work we do incorporate the actual cultivation parameters in the computational framework and use this information to infer combinatorial regulation by TFs. The work is based on a yeast transcriptome compendium, comprised of 170 microarray measurements [Knij 09]. These measurements encompass 55 unique growth conditions with a variable number of independent biological replicates per condition. All cultivations were performed in

chemostat fermentors under steady-state conditions. In a chemostat, culture broth (including biomass) is continuously replaced by fresh medium at a fixed and accurately determined dilution rate. When the dilution rate is lower than $\mu_{max}$, the maximal specific growth rate of the micro-organism, a steady-state situation will be established in which the specific growth rate equals the dilution rate. In such a steady-state chemostat culture, $\mu$ is controlled by the (low) residual concentration of a single growth-limiting nutrient. Across the 55 different conditions, there are nine varying cultivation parameter types, including limiting element, growth rate, carbon source, aeration and temperature. Each type can assume a unique set of values. For example, in a given experiment, the employed limiting element is either carbon, nitrogen, sulfur, phosphorus, zinc or iron. Thus, each condition is characterized by a configuration of settings of these nine cultivation parameter types. See Figure 7.1. In order to model the effects of the cultivation

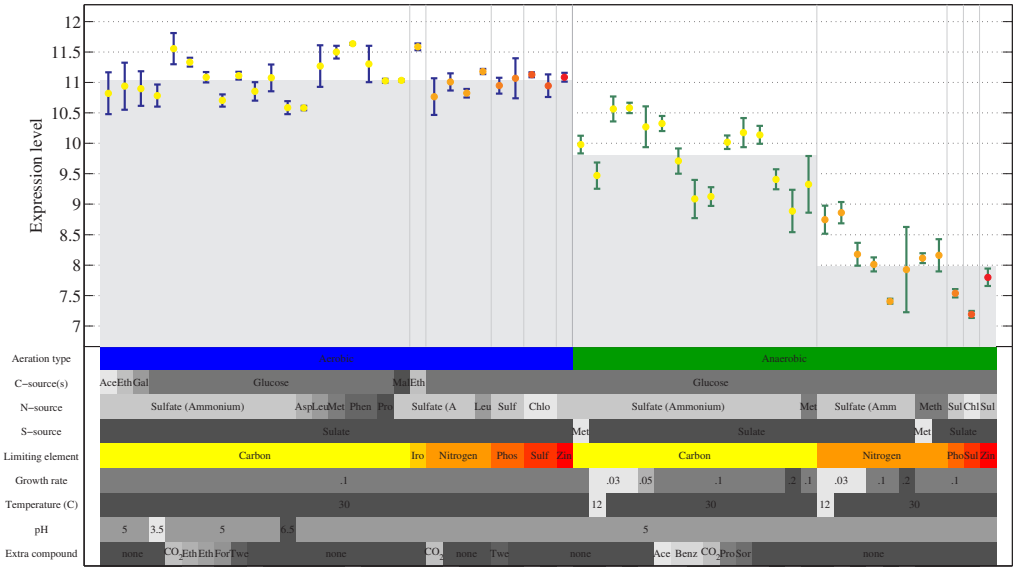



**Figure 7.1 – Expression levels of a gene (*COX5A*) across the 55 cultivation conditions.**
The colored matrix is a schematic representation of the settings of the nine cultivation parameter types across the 55 conditions. The colored lanes indicate the cultivation parameter types that are employed to order the experiments, in this case, aeration type and limiting element. The regression model which models the gene expression as a function of the cultivation parameters, selected one single effect, i.e. aeration type, and one combinatorial effect, i.e. aeration type anaerobic together with limiting element carbon. The reconstructed expression pattern based on these two effects is indicated by the shaded area.

parameters on gene expression while explicitly incorporating TFs, we follow a two-step procedure. An overview of this procedure is presented in Figure 7.2. First, we apply a forward stepwise regression strategy to quantify the (combinatorial) effect of these environmental parameters on gene expression. The regression is performed for each gene individually. Figure 7.1 depicts the results of the regression analysis for one particular gene. The influence of a cultivation parameter on the expression of a gene is represented by its regression weight. These weights are discretized by mapping non-zero elements to 1 or -1, depending on the sign of the weight. Given that changes in gene expression

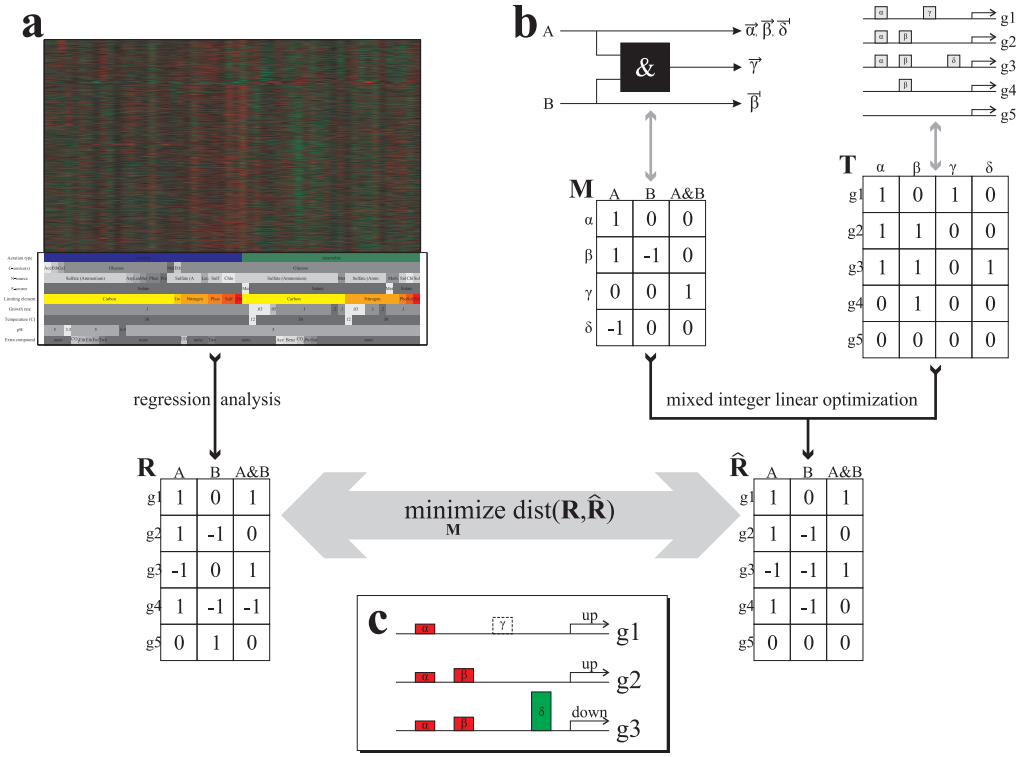



**Figure 7.2 – Schematic overview of the approach.**
The goal is to build $\widehat{\mathbf{R}}$, the optimal approximation of the discretized regression coefficients in $\mathbf{R}$. **a:** The coefficients in $\mathbf{R}$ are derived from a regression analysis, which assesses the influence of cultivation parameters on gene expression by employing these parameters as predictors in the regression model. The discretization procedure maps non-zero regression weights to 1 or -1, depending on their sign. (The schematic representation of $\mathbf{R}$ is given for five genes and three cultivation parameters.) **b:** The elements of $\widehat{\mathbf{R}}$ are determined by $\mathbf{T}$ and $\mathbf{M}$. $\mathbf{T}$ is fixed and indicates binary TF binding potential to gene promoters. The elements of $\mathbf{M}$ are estimated and indicate the activity of TFs as enhancers or repressors under the different (combinatorial) cultivation parameters. A logic circuit derived from $\mathbf{M}$ is graphically depicted above the representation of $\mathbf{M}$. **c:** Visualization of the active TFs on the gene promoters of genes g1, g2 and g3 under cultivation parameter A. Enhancers are depicted as red boxes; repressors are depicted as green boxes. (TF $\gamma$ can bind the promoter of g1, but is not active under A.) The height of a box indicates the enhancer or repressor strength. The strength of a particular enhancer or repressor is the same for all genes. A gene is upregulated when its activator strength, i.e. the sum of the heights of the red boxes, is larger than the repressor strength, which equals the sum of the heights of the green boxes. Downregulation is inferred in the opposite situation. See text for details.

levels as observed across different environmental conditions can be attributed to changes in the activity of TFs, we aim to infer the activity of TFs as a function of the cultivation parameters. This forms the second step of our approach. The goal is to estimate $\mathbf{M}$, such that $\widehat{\mathbf{R}}$ is the optimal approximation of the discretized regression coefficients in $\mathbf{R}$. The elements of $\mathbf{M}$ are -1, 0 or 1 and indicate whether a TF is activated as an enhancer (1) or a repressor (-1) under a (combinatorial) cultivation parameter. Additionally, each TF has a particular generic enhancer strength and a repressor strength. In the procedure we employ auxiliary matrix $\mathbf{T}$, which is derived from ChIP-chip experiments and literature and indicates whether a TF can bind a gene promoter. To decide whether a gene is upregulated, downregulated or not affected by a particular cultivation parameter, indicated by a 1, -1 and 0 in $\widehat{\mathbf{R}}$, respectively, we use the following rules concerning transcriptional regulation: If there is at least one active enhancer in a gene promoter, then the gene *can be* upregulated. If there are only active enhancers in a gene promoter, then the gene *is* upregulated. Similar rules apply to the repressors. If there are both active enhancers and repressors in a gene promoter, we compare total enhancer strength, which is the sum of the strengths of the activated enhancers, with its repressor counterpart. When the enhancer strength is larger than the repressor strength, the gene is upregulated. The gene is downregulated when the repressor strength exceeds the enhancer strength. Figure 7.2c visualizes the active TFs that bind the gene promoters of genes g1, g2 and g3 under cultivation parameter A. From $\mathbf{M}$ we deduce that three TFs are activated; $\alpha$ and $\beta$ are enhancers, $\delta$ is a repressor. From $\mathbf{T}$ we deduce that $\alpha$ binds all three promoters, $\beta$ binds the g2 and g3 promoters and $\delta$ only binds the promoter of g3. Gene g1 and g2 are upregulated, since only active enhancers bind the promoters. For gene g3, the repressor strength of TF $\delta$ exceeds that of the sum of the two enhancers, thereby downregulating the gene. The concept of TF strength enables the inference of hierarchical or combinatorial effects amongst TFs that bind a gene promoter. The inference algorithm is translated into a linear mixed integer optimization problem and solved accordingly. Both the elements of $\mathbf{M}$ as well as the TF strengths are estimated, such that the predicted gene regulation in $\widehat{\mathbf{R}}$ maximally corresponds with the discretized regression coefficients in $\mathbf{R}$. The abovementioned rules become constraints in the optimization problem. See the Methods section for details. The resulting model identifies the combinatorial influence of cultivation parameters on TF activity and gene expression. Furthermore, it infers the combinatorial regulatory code of multiple TFs in gene promoters.

## 7.3 Methods

### 7.3.1 Microarray data

The *Saccharomyces cerevisiae* laboratory reference strain CEN.PK 113-7D (*MAT*a) was grown in chemostat fermentors under 55 different conditions. For each condition, a variable number of independent biological replicates was performed, although mostly three, summing up to 170 microarray measurements. Across the 55 conditions, nine different cultivation parameter types can be identified. A cultivation parameter type, e.g. the carbon source, is described as a categorical variable and contains two or more categories, e.g. the used carbon source can be either maltose, glucose or ethanol. Each condition is characterized by a specific combination of these categories across the nine cultivation parameter types. Figure 7.1 presents an overview of the relevant categories assumed by

the parameter types per condition. Sampling of the chemostat cultures, probe preparation and hybridization to single-channel Affymetrix GeneChip YG S98 microarrays was performed as previously described [Pipe 02]. Chip quality control, condensing probe intensities to gene expression levels and normalization was performed using GeneData Refiner Array. The RMA algorithm was used to derive the log2 scale measure of the expression levels [Iriz 03]. Quantile normalization was applied to normalize between arrays [Bols 03].

### 7.3.2 Inferring the influence of cultivation parameters on gene expression

A design matrix was created, containing both main (or single) effects and interaction (or combinatorial) effects: Each category within each cultivation parameter type is represented by a binary indicator column with 170 entries. These columns represent the main effects, which indicate, for each array, under which category of a particular cultivation parameter type, the yeast was grown. Interaction effect columns were obtained by applying the logic AND function to all possible pair-wise combinations of main effect columns. Redundant columns and columns containing only zeros were removed, resulting in 112 columns, of which 37 represent main effects and 75 represent interaction effects. This data is stored in the binary $[A \times C]$ design matrix $\mathbf{D}$. Here, $A$ equals 170 and is the number of arrays. $C$ equals 112 and is the number of (combinatorial) cultivation parameters.
A forward stepwise ordinary least squares regression strategy was applied to each gene individually:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon \tag{7.1}$$

Here, $\mathbf{y}_i$ denotes the measured gene expression level of a particular gene for array $i$, with $i = 1 \ldots A$; $\mathbf{X}$ is the predictor matrix, $\theta$ represents the regression coefficients and $\epsilon$ the error, which is assumed to be independent zero-mean normally distributed. Initially, $\mathbf{X}$ contains only the intercept, i.e. a column of $A$ ones. In an iterative fashion, columns from $\mathbf{D}$ are added to $\mathbf{X}$. For this we apply a leave-one-out cross validation (loocv) scheme, where a single sample is used for testing, while the remaining $(A-1)$ samples are used for training the regression model. This is repeated such that each sample is used once as test data. The column from $\mathbf{D}$, with the smallest root-mean-squared (rms) loocv error and absolute regression coefficient larger than one, is selected and added. The iterative process of adding columns is discontinued when the P-value, as output by a t-test that determines whether the regression coefficient significantly differs from zero, exceeds $0.05/C$. To prevent the inclusion of spurious combinatorial effects, the following strategy is applied: When a combinatorial effect column is selected, we check whether the addition in explained variance is larger than the addition is explained variance when adding the two main effect columns that constitute the combinatorial effect. Only in the cases where this is true, we add the combinatorial effect column. Otherwise the two main effect columns are added, provided that they satisfy the P-value threshold and their absolute regression coefficients are larger than one.
Note that only coefficients larger than one or smaller than minus one are allowed. In terms of the absolute expression measure, this means we only take into account expression differences of one fold change or more. (The expression data is on log2 scale.) So, we focus on the cases where a cultivation parameter has a large influence on expression.

Finally, the regression coefficients for all yeast genes are discretized and put in $\mathbf{R}$ ($[G \times C] \in \mathbb{Z}[-1, 0, 1]$), where $G$ is the number of yeast genes. The discretization procedure maps positive coefficients to 1 and negative coefficients to -1. $\mathbf{R}$ is quite sparse since for most of the genes only two or three columns from $\mathbf{D}$ were selected as significant predictors.

### 7.3.3 TF binding data

For 111 TFs we extracted their known regulatory sites from TRANSFAC [Wing 00] and ChIP-chip data [Harb 04, MacI 06] (no conservation, binding P-value cutoff 0.001). These gene-TF pairs were put in the binary $[G \times F]$ TF binding matrix $\mathbf{T}$, where a 1 indicates that a TF can bind a gene promoter. $F$ equals 111 and is the number of TFs.

### 7.3.4 Inferring TF activity and TF strengths

The goal of our optimization problem is to infer the activity of TFs as a function of cultivation parameters, such that we can optimally explain the regression coefficients, which were distilled from the observed gene expression data. These TF activities form tertiary matrix $\mathbf{M}$ ($[F \times C] \in \mathbb{Z}[-1, 0, 1]$). A nonzero element in $\mathbf{M}$ indicates that a TF is activated under a cultivation parameter and either acts as an enhancer (1) or a repressor (-1). Other data used in the optimization problem are: TF binding matrix $\mathbf{T}$ ($[G \times F] \in \mathbb{Z}[0, 1]$), discretized regression coefficient matrix $\mathbf{R}$ ($[G \times C] \in \mathbb{Z}[-1, 0, 1]$) and its reconstructed version $\widehat{\mathbf{R}}$ ($[G \times C] \in \mathbb{Z}[-1, 0, 1]$). First, from the tertiary matrix $\widehat{\mathbf{R}}$ two binary matrices with the same dimensions, $\widehat{\mathbf{R}}^+$ and $\widehat{\mathbf{R}}^-$, are derived. $\widehat{\mathbf{R}}^+$ has non-zero entries, where $\widehat{\mathbf{R}}$ contains 1's, and thus indicates the elements, where genes are upregulated under influence of a particular cultivation parameter. $\widehat{\mathbf{R}}^-$ has non-zero entries, where $\widehat{\mathbf{R}}$ contains -1's, and thus indicates the downregulated elements. A similar procedure is undertaken for tertiary matrix $\mathbf{M}$, thereby obtaining $\mathbf{M}^+$, which contains the active enhancers and $\mathbf{M}^-$, which contains the active repressors. Now, all variables consist of binary integers (and are restricted to remain binary integers).
The objective function for the optimization problem is as follows:

$$\text{minimize} \sum_{\forall g,c \in I^+} [\mathbf{R}_{gc} - \widehat{\mathbf{R}}^+_{gc}] + \sum_{\forall g,c \in I^-} [-\mathbf{R}_{gc} - \widehat{\mathbf{R}}^-_{gc}] + \\ + \lambda \sum_{f=1}^{F} \sum_{c=1}^{C} [\mathbf{M}^+_{f,c} + \mathbf{M}^-_{f,c}] \tag{7.2}$$

where $I^+$ is the set of index pairs referring to the elements where $\mathbf{R}$ is 1, and similarly, $I^-$ refers to the negative elements of $\mathbf{R}$. Thus, we only try to explain the nonzero elements of $\mathbf{R}$, which represent the large expression changes due to the influence of the cultivation parameters. The zero elements of $\mathbf{R}$ do not only contain cases where there is no change in expression, but they contain the whole spectrum of no change in expression up to moderately large changes in gene expression. Therefore, we do not want to enforce TFs to be deactivated because of these zero elements. The last term of Eq. 7.2 restricts the model complexity by penalizing the number of activated TFs. Parameter $\lambda$ can be interpreted as the number of non-zero elements in $\mathbf{R}$ that a TF needs to help explain in order for it to be activated. Below, the constraints of the optimization problem are

stated. These constraints are linear in $\mathbf{M}^+$, $\mathbf{M}^-$, $\widehat{\mathbf{R}}^+$ and $\widehat{\mathbf{R}}^-$, which are the variables in the system. In the appendix a detailed explanation for constraints **c5**, **c8** and **c12** is given.
The first two constraints are straightforward. Constraint **c1** states that a TF cannot be an active repressor and an active enhancer at the same time. Constraint **c2** states that a gene cannot be upregulated and downregulated at the same time.

$$\begin{aligned}
&\textbf{c1:} \quad \mathbf{M}^+_{fc} + \mathbf{M}^-_{fc} \leq 1 && \forall f, c \\
&\textbf{c2:} \quad \widehat{\mathbf{R}}^+_{gc} + \widehat{\mathbf{R}}^-_{gc} \leq 1 && \forall g, c
\end{aligned}$$

Constraint **c3** states that if there is at least one active enhancer in a gene promoter, i.e. the inner product is positive then the gene *can be* upregulated, i.e. the regression coefficient can be 1. Constraint **c4** is the analogue constraint for the case of active repressors. Constraint **c5** forces a gene to be either upregulated or downregulated, when there is at least one active enhancer or one active repressor in the gene promoter.

$$\begin{aligned}
&\textbf{c3:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^+_{\cdot c} \rangle \geq \widehat{\mathbf{R}}^+_{gc} && \forall g, c \\
&\textbf{c4:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^-_{\cdot c} \rangle \geq \widehat{\mathbf{R}}^-_{gc} && \forall g, c \\
&\textbf{c5:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^+_{\cdot c} \rangle + \langle \mathbf{T}_{g\cdot}, \mathbf{M}^-_{\cdot c} \rangle \leq F \cdot (\widehat{\mathbf{R}}^-_{gc} + \widehat{\mathbf{R}}^+_{gc}) && \forall g, c
\end{aligned}$$

To decide upon upregulation or downregulation when multiple active enhancers and repressors bind a promoter, four continuous variables are introduced: $\mathbf{S}^+$ and $\mathbf{S}^-$; both ($[F \times C] \in \mathbb{R}[0, F]$) and $\widetilde{\mathbf{S}}^+$ and $\widetilde{\mathbf{S}}^-$; both ($[F \times 1] \in \mathbb{R}[1, F]$). $\mathbf{S}^+_{fc}$, represents the "strength" of TF $f$ as an enhancer under cultivation parameter $c$. $\mathbf{S}^+_{fc}$ is zero when $\mathbf{M}^+_{fc}$ is zero, i.e. when $f$ is not activated as an enhancer under $c$. This rule is stated in constraint **c6**. $\mathbf{S}^+_{fc}$ equals the generic TF strength for $f$, $\widetilde{\mathbf{S}}^+_f$, when $\mathbf{M}^+_{fc}$ is one. Thus, the strength of a TF $f$ is the same for all genes under the cultivation parameters, where the gene is activated (and zero otherwise). This rule is stated in constraints **c7** and **c8**. Analogue rules apply for $\mathbf{S}^-$ and $\widetilde{\mathbf{S}}^-$. The corresponding constraints **c9**, **c10** and **c11** are omitted for brevity.

$$\begin{aligned}
&\textbf{c6:} \quad \mathbf{S}^+_{fc} \leq F \cdot \mathbf{M}^+_{fc} && \forall f, c \\
&\textbf{c7:} \quad \mathbf{S}^+_{fc} \leq \widetilde{\mathbf{S}}^+_f && \forall f, c \\
&\textbf{c8:} \quad \mathbf{S}^+_{fc} - \widetilde{\mathbf{S}}^+_f \geq F \cdot (\mathbf{M}^+_{fc} - 1) && \forall f, c
\end{aligned}$$

Constraint **c12** states that when the sum of the strengths of active enhancers that bind a gene promoter is larger than its repressing counterpart, the gene is upregulated. Constraint **c13** encloses the reverse scenario. Note that if an identical set of enhancers and repressors is active on a promoter, this will lead to the same reconstructed regression coefficient for any gene and under any cultivation parameter.

$$\begin{aligned}
&\textbf{c12:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{S}^+_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^-_{\cdot c} \rangle \geq (F^2 + F^{-2}) \cdot \widehat{\mathbf{R}}^+_{gc} - F^2 && \forall g, c \\
&\textbf{c13:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{S}^-_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^+_{\cdot c} \rangle \geq (F^2 + F^{-2}) \cdot \widehat{\mathbf{R}}^-_{gc} - F^2 && \forall g, c
\end{aligned}$$

The optimization problem is implemented within the MATLAB environment and executed using the MOSEK optimization toolbox with standard settings for mixed integer optimization. Given constraints **c1** to **c13**, MOSEK estimates variables $\mathbf{M}^+$, $\mathbf{M}^-$, $\widehat{\mathbf{R}}^+$, $\widehat{\mathbf{R}}^-$, $\mathbf{S}^+$, $\mathbf{S}^-$, $\widetilde{\mathbf{S}}^+$ and $\widetilde{\mathbf{S}}^-$ such that the optimization function in Eq. 7.2 is minimized.

## 7.4 Results

### 7.4.1 TF activity in response to changes in oxygen and carbon presence

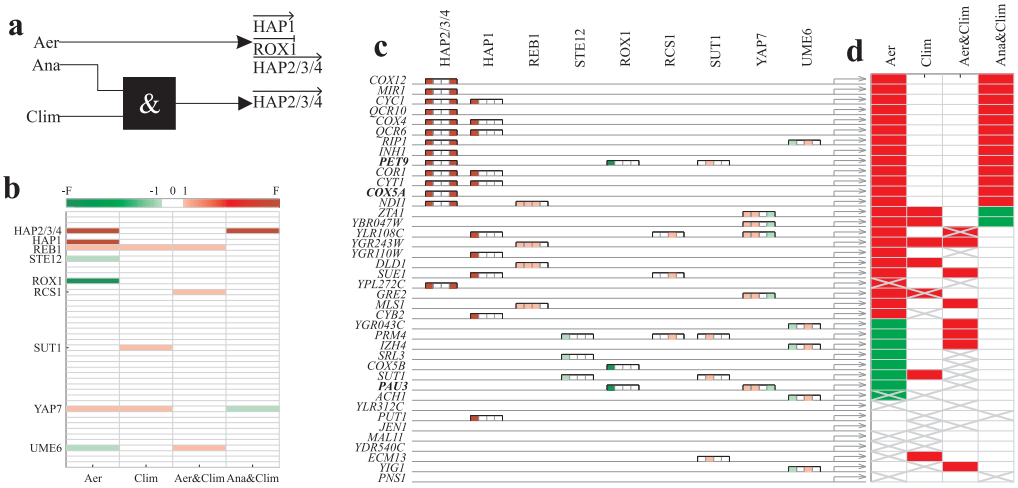



**Figure 7.3 – Overview of the results obtained for the oxygen and carbon limitation data.**
**a:** Inferred influence of cultivation parameters aerobic growth (Aer), anaerobic growth (Ana) and carbon limitation (Clim) on TF activity. Only the three dominating TFs are reported. **b:** Representation of **S**, indicating the strength of the activated TFs under each of the four cultivation parameters. Enhancers are depicted in red; repressors are depicted in green. **c:** Representation of **T**, indicating which gene promoters can be bound by the activated TFs. The enhancer or repressor strengths for the four cultivation parameters are visualized by the colored blocks inside the rectangle that represents a binding site. **d:** Representation of $\widehat{\mathbf{R}}$, indicating the inferred regression coefficients. Upregulation is indicated by red; downregulation is indicated by green. Incorrectly inferred elements are marked with a grey cross. White boxes without a cross are the zero elements of **R**. These elements are not part of the optimization scheme.

The regulatory network inference algorithm is run on a subset of the data. In particular, we focus on oxygen and carbon; two environmental factors, which have a large and well studied effect on the transcriptional program of *Saccharomyces cerevisiae*. Four cultivation parameters are selected, i.e. aeration type, carbon-limitation and the combinatorial cultivation parameters, carbon-limited aerobic growth and carbon-limited anaerobic growth. Note that aeration type is actually a cultivation parameter type that assumes two values, i.e. aerobic growth and anaerobic growth. Since these are mutually redundant, only aerobic growth was included in the regression model and subsequent optimization algorithm. (Downregulation under aerobic growth and upregulation under anaerobic growth are interchangeable.) There are 40 genes, which are influenced by at least two of these four cultivation parameters, i.e. there are 40 rows in **R** with at least two non-zero elements in the four columns of interest. These 40 genes are bound by 46 different TFs. In this experiment $\lambda$ is set to two. The algorithm correctly inferred the regression coefficients of 58 of the 84 (70%) nonzero elements in **R**. A particularly large concentration of incorrectly predicted values appears towards the bottom of $\widehat{\mathbf{R}}$, where

zeros are predicted while the true expression coefficients are non-zero. See Figure 7.3d. This stems from the fact that the promoters of these genes have almost no motifs for the activated TFs, in which case the model can not explain the up- or downregulation.

*Inferred TF activity*

In total, nine different TFs were activated across the four cultivation parameters, some under more than one cultivation parameter. Three of these TFs, Hap1, Hap2/3/4 and Rox1, have a significantly larger strength, when compared to the others. See Figure 7.3a,b. The large strength indicates their dominating effect on transcriptional regulation. If one of these TFs is active and binds the promoter, it will determine the direction of transcriptional regulation. E.g., under aerobic conditions (Aer) the promoter of gene *PAU3* (the tenth gene from the bottom in Figure 7.3c) is bound by one active enhancer, i.e. Yap7, and one active repressor, i.e. Rox1. Since the repressor strength of Rox1 is (much) larger than the enhancer strength of Yap7, the gene is (correctly) predicted to be downregulated. Interestingly, in the resulting network for this data, the TF strength of Rox1 equals 45.9995, which is very close to the maximum value of 46, the number of TFs $F$. However, this number is slightly smaller than the strength of Hap2/3/4 which has the maximal strength of 46. This difference can be attributed to gene *PET9* (the ninth gene from the top in Figure 7.3c). Both Hap2/3/4 and Rox1 can bind the *PET9* promoter. To ensure that this gene is upregulated when grown aerobically, as was deduced from the regression analysis, the active enhancers should have a larger strength than the active repressors. Therefore, the strength of Rox1 is set a bit smaller than the strength of Hap2/3/4, however, still large enough to dominate other active enhancers.

*Regulation of gene expression by oxygen*

The role of the three dominant TFs in the regulation of gene expression by oxygen is widely reported in the literature. Both Hap1 and the Hap2/3/4 complex activate genes in response to heme, which is synthesized only in the presence of oxygen [Zito 92]. TF Rox1 is needed for the repression of hypoxic or heme-repressed genes under aerobic conditions [Lowr 88]. Also, the relation between carbon source and the Hap2/3/4 complex has been investigated. The Hap2 and Hap3 proteins enable DNA binding of the complex, whereas Hap4 contains the transcriptional activation domain. The synthesis of the activator subunit Hap4 is regulated by the carbon source. More specifically, the expression of Hap4 is repressed by glucose, *Saccharomyces cerevisiae*'s preferred carbon source [Fors 89]. Tai *et al.* [Tai 05] reports that Hap4 mRNA is present in carbon-limited cultivations even under anaerobic conditions, where Hap4 has no obvious role. We can corroborate and even further substantiate these findings with the observation that the Hap4 protein is an activator under carbon-limited anaerobic conditions. Note that all genes, which are upregulated under carbon-limited anaerobic growth are also upregulated under aerobic growth. See the top 13 genes in Figure 7.3c. The expression profile of one of these genes, *COX5A*, across all conditions is depicted in Figure 7.1. This expression profile is typical for all the 13 members of this group. It shows that these genes are most highly expressed when grown aerobically. Yet, in the anaerobic case, where the expression is in general lower, these genes show different expression behavior in carbon-limited growth compared to other nutrient limitations. I.e., these genes have a higher expression level in carbon-limited cultivations, where there is hardly any glucose, compared to the situation, where glucose is abundant.

Also, for the other TFs, which are activated according to the inference algorithm, evidence is found in literature. E.g., Reb1, which acts as an enhancer under three cultivation parameters, is a RNA polymerase I enhancer binding protein as well as an activator for many genes transcribed by RNA polymerase II [Ju 90]. Ste12 is known to activate genes associated with pseudohyphal (low oxygen) growth [Norm 99]. Sut1 is reported to encode a glucose transporter [Weie 99], however Sut1 also has a putative role in the regulation of some hypoxic genes [Regn 01]. In general, the precise regulatory role of these TFs in (an)aerobiosis and response to the carbon source is not known. The results of this analysis provide hints for elucidating the regulatory mechanisms of these factors.

*Setting $\lambda$*

Parameter $\lambda$, which restricts the model complexity by penalizing the number of activated TFs, is chosen using a 5 fold cross-validation (cv) scheme. The genes are divided into five parts, where consecutively four parts are used for training and one part is used for testing. The $\mathbf{M}$ and $\mathbf{S}$ matrices, which are computed on the training set, are applied to the test set to obtain the reconstructed regression coefficients for the test set, $\widehat{\mathbf{R}}^{test}$. The error on the test set is defined as:

$$Err = \frac{1}{J} \sum_{\forall g,c \in I} \left| \mathbf{R}^{test}_{gc} - \widehat{\mathbf{R}}^{test}_{gc} \right| \tag{7.3}$$

where $I$ is the set of index pairs referring to the non-zero elements of $\mathbf{R}^{test}$ and $J$ the number of these non-zero elements. The cross-validation scheme is repeated ten times. Figure 7.4 depicts the average error over all cv runs. For small values of $\lambda$, many TFs are activated in order to approximate the regression coefficients. Clearly, this strategy is prone to overfitting, which is also illustrated by the large cv error. For large values of $\lambda$, activating a TF is severely penalized, such that only a few TFs will be activated. (For $\lambda = 20$, no TF is activated and every element of $\widehat{\mathbf{R}}^{test}$ is zero). The high cv error in this case, indicates that a lot of true regulation is missed. The optimal $\lambda$ will be found between these extremes. In this experiment, $\lambda = 2$ led to the smallest cv error and was therefore selected.

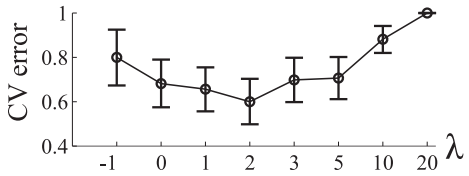



**Figure 7.4 – Cv errors for different values of $\lambda$.**

### 7.4.2 Transcriptional regulation of nitrogen metabolism

Across the conditions of the compendium, yeast was grown on six different nitrogen sources. This inspired the second experiment, where we analyzed the transcriptional regulation of the genes that comprise the nitrogen compound metabolism category of

GO biological processes [Ashb 00]. 119 of these genes are influenced by at least one cultivation parameter and bound by one of 78 different TFs. In total, there are 68 cultivation parameters that cause up- or downregulation of at least one of these 119 genes. The resulting transcription regulation network (with $\lambda_{opt} = 2$) revealed the activation of 14 different TFs under 28 different cultivation parameters, of which 11 are combinatorial. Figure 7.5 depicts the network for the cultivation parameters, which are most straightforwardly related to nitrogen metabolism, i.e. the different nitrogen sources, nitrogen as growth limiting element and combinatorial effects involving these cultivation parameters. The six different nitrogen sources can be dichotomized into preferred and non-preferred

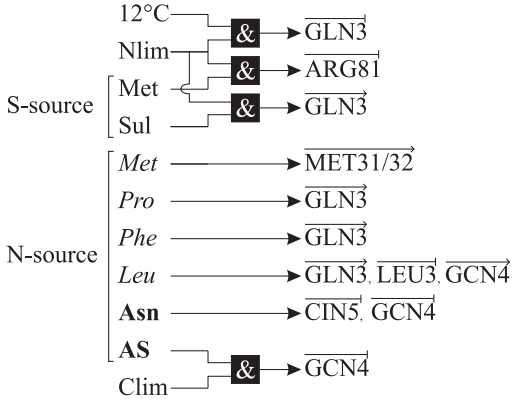



**Figure 7.5 – Inferred TF activity derived from genes, which are involved in nitrogen metabolism.**
Preferred nitrogen sources are printed in bold; non-preferred nitrogen sources are printed in italic style. Abbreviations for the nitrogen and sulfur sources are explained in the text.

nitrogen sources. The preferred nitrogen sources are asparagine (Asn) and ammonium (in ammonium sulfate (AS)). Proline (Pro), phenylalanine (Phe), methionine (Met) and leucine (Leu) are non-preferred (or poor) nitrogen sources [Maga 02, Boer 07]. In *S. cerevisiae*, the use of nitrogen sources is controlled by a transcriptional regulation mechanism known as nitrogen catabolite repression (NCR). When a good nitrogen source is present, NCR shuts down the pathways for the use of poor nitrogen sources. NCR is mediated by a four-member family of GATA-binding TFs: Gln3, Gat1, Dal80 and Gzf3 [Hofm 99]. In the absence of a good nitrogen source, Gln3 is activated and in turn activates the transcription of NCR-sensitive genes. Indeed, for three of the four non-preferred nitrogen sources, Gln3 acts as an enhancer. When methionine is the nitrogen source, the Met31/32 complex is activated. This complex controls the biosynthesis of sulfur containing amino-acids [Blai 97]. (Methionine is also used as a sulfur source.) In the case of leucine, two additional TFs are activated; Leu3 and Gcn4, the two key regulators in the regulation of branched-chain amino acid metabolism [Boer 05]. The inferred role of Gcn4 as an activator in the presence of a poor nitrogen source and as a repressor in the presence of good nitrogen sources corroborates the work of Sosa *et al.* [Sosa 03]. It further supports the fact that NCR is not solely achieved through the action of the abovementioned family of GATA factors, but conceivably also through Gcn4.

*Missing and dubious TF activity*

Remarkably, the other tree GATA factors, Gat1, Dal80 and Gzf3, are not part of the

generated network. Inspection of the TF binding data in the promoters of the 119 nitrogen metabolism genes revealed that Gat1, Dal80 and Gzf3 bind only 3, 4 and 0 genes, respectively. This could indicate that their targets are not transcriptionally regulated under the influence of the cultivation parameters. However, this observation should also be related to the ChIP-chip data. From TRANSFAC, we extracted many TF-gene pairs, which are not present in the ChIP-chip data. This indicates that not all TF targets are detected by the ChIP-chip experiments. Furthermore, Gao *et al.* [Gao 04] estimate that 40% of the ChIP-chip TF targets are non-functional. Obviously, this complicates regulatory network inference. Another dubious result was identified when analyzing the cases in which two or more TFs were active on a promoter. In this experiment, there are 72 such cases, of which 10 are unique. Amongst the most frequent cases, we found the combinatorial regulation of TFs, which have already been reported in literature, e.g. the interplay between Leu3 and Gcn4 [Boer 05] and that of Cbf1 and Gcn4 [OCon 95]. Also, Gln3 and Gcn4 were found activated together in a set of nine gene promoters.

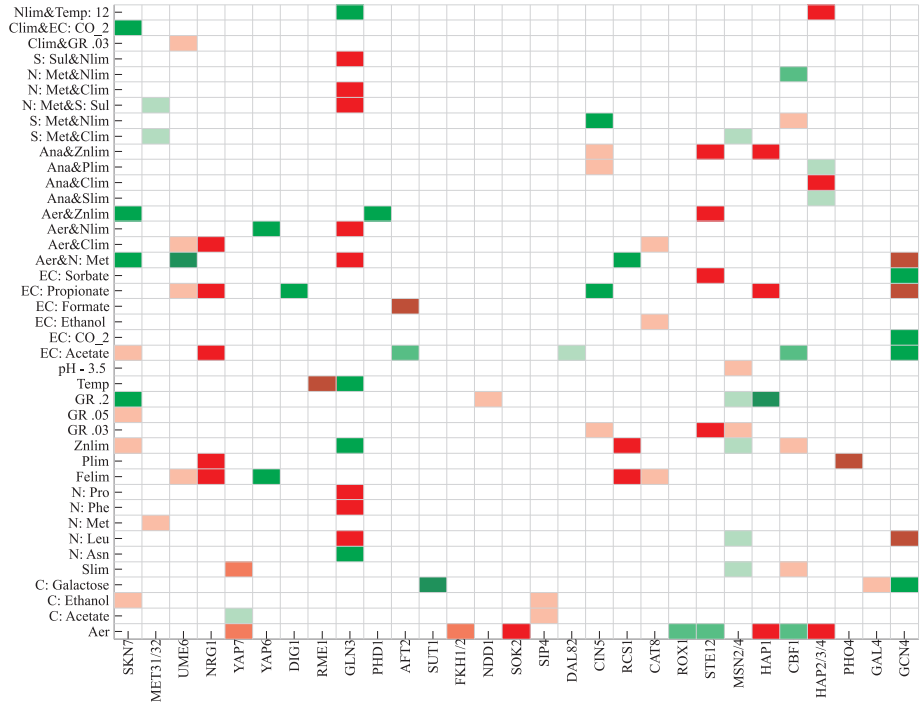



**Figure 7.6 – Representation of S for the regulatory program inferred using the compendium.**
Color coding is identical to Figure 7.3b.

These nine genes were upregulated under two cultivation parameters, i.e. sulfur limitation and zinc limitation, where both Gln3 and Gcn4 are enhancers. However, under another cultivation parameter, i.e. where leucine is used as a nitrogen source, the *same* genes were *downregulated*, where now Gln3 acts as a repressor (which is stronger than enhancer Gcn4). These results seem implausible and imply that this regulation pattern should involve another TF, which might not be present in the employed TF binding data set. Preliminary experiments with artificial datasets have demonstrated that especially

missing TFs (simulated by removing columns from **T**) can have a large negative effect on the ability to reconstruct the correct regulatory network. (Results not shown.)

### 7.4.3 Compendium analysis

The algorithm was also run on the complete compendium for all genes that are up- or downregulated under at least two cultivation parameters and for all cultivation parameters that influence the expression of at least ten genes ($G = 766, C = 67, F = 101, \lambda_{opt} = 5$). In the resulting regulatory network, 41 (61%) cultivation parameters activated at least one of the TFs, resulting in 29 (29%) different activated TFs in total. See Figure 7.6. Network inference on the complete dataset allows for a more rigorous and unbiased estimation of the regulatory program. It reveals confounding factors, with respect to the previously discussed programs, which were based on a subset of the data. For example, the regulatory program of GATA factor Gln3, as discussed before, is also depending on other (combinatorial) cultivation parameters, e.g. zinc limitation and nitrogen limitation at low temperature. These results offer interesting leads, however the combinatorial regulation of TFs, as inferred by this analysis, becomes complicated. There are up to four active TFs on gene promoters. This calls for an automated procedure that uses these inferred TF activities and accompanying strengths to derive logic rules, in which the influence of multiple TFs on transcriptional regulation is formalized.
Note that the inference algorithm was run on a selection of genes and cultivation parameters. The number of variables and constraints in optimization problem is $4FC + 2GC$ and $7FC + 6GC$, respectively, which becomes quite large for the complete dataset. It is yet unclear (due to computation time) if converge is reached for the dataset with all genes and cultivation parameters.

## 7.5 Discussion

The transcriptional program of a cell is largely determined by its extracellular environment. The accurate measurement of environmental parameters, e.g. with chemostat cultures, have inspired several approaches that analyze the (combinatorial) effect of environmental parameters on gene expression. In this study, we have, for the first time demonstrated how environmental parameters can be employed to derive transcriptional regulation networks. In these networks, the cultivation parameters form the signals that trigger the activation or deactivation of TFs. Since many TFs are regulated post-transcriptionally, this approach seems more natural than the often employed strategy of deducing the TF activity from the mRNA expression of TFs. The inference algorithm was translated into a linear optimization problem, solvable without having to rely on greedy and/or heuristic search strategies.
The combinatorial regulatory code of multiple TFs that are able to bind a promoter, is modeled using the linearly weighted sum of inferred enhancers and repressor strengths. Previous approaches have also modeled gene expression as a linearly weighted sum of TF contributions, e.g. [Gao 04]. The main improvement of our method is the fact that the activity of TFs can be explicitly turned on or off, and that the inference algorithm optimizes this choice with respect to the direction of regulation, i.e whether a gene is up- or downregulated. This strategy enables the inference of combinatorial effects between TFs. For example, a repressor, which interacts directly with the TATA binding protein,

thereby completely blocking transcription independent of the possible enhancers that bind the promoter, would acquire a strength that is larger than the sum of the strengths of all enhancers that can bind the promoter. Thus, the repressor, when active, will cause downregulation of the gene, thereby nullifying the influence of the enhancers. This is in contrast with the linear regression strategies, where these enhancers would still have influence on the gene expression level.
Additional validation experiments indicate that more pairs of TFs, which are simultaneously active according to our approach, are found to co-occur in PubMed abstracts when compared to TF pairs uncovered with Gao *et al.* [Gao 04]. This difference can be attributed to the fact that we decompose the expression in terms of cultivation parameters, and analyze these cultivation parameters separately. When using only the expression data itself, some cultivation parameters (such as aeration type) can have a much larger influence than others, thereby dominating the expression pattern and thus controlling which TFs are found to be the most significant, leading to less diversity in activated TFs (and thus fewer TF pairs). An overview of this comparison can be found Supplementary material of [Knij 08] online.
A future challenge lies in the integral interpretation of the inferred regulatory networks, which must be accompanied by a computational approach that derives logic rules, which are able to describe the interplay of multiple TFs on gene promoters.

## 7.6 Appendix

A detailed explanation for constraints **c5**, **c8** and **c12** is given.

$$\textbf{c5:} \quad \langle \mathbf{T}_{g\cdot}, \mathbf{M}^{+}_{\cdot c} \rangle + \langle \mathbf{T}_{g\cdot}, \mathbf{M}^{-}_{\cdot c} \rangle \leq F \cdot (\widehat{\mathbf{R}}^{-}_{gc} + \widehat{\mathbf{R}}^{+}_{gc}) \qquad \forall g, c$$

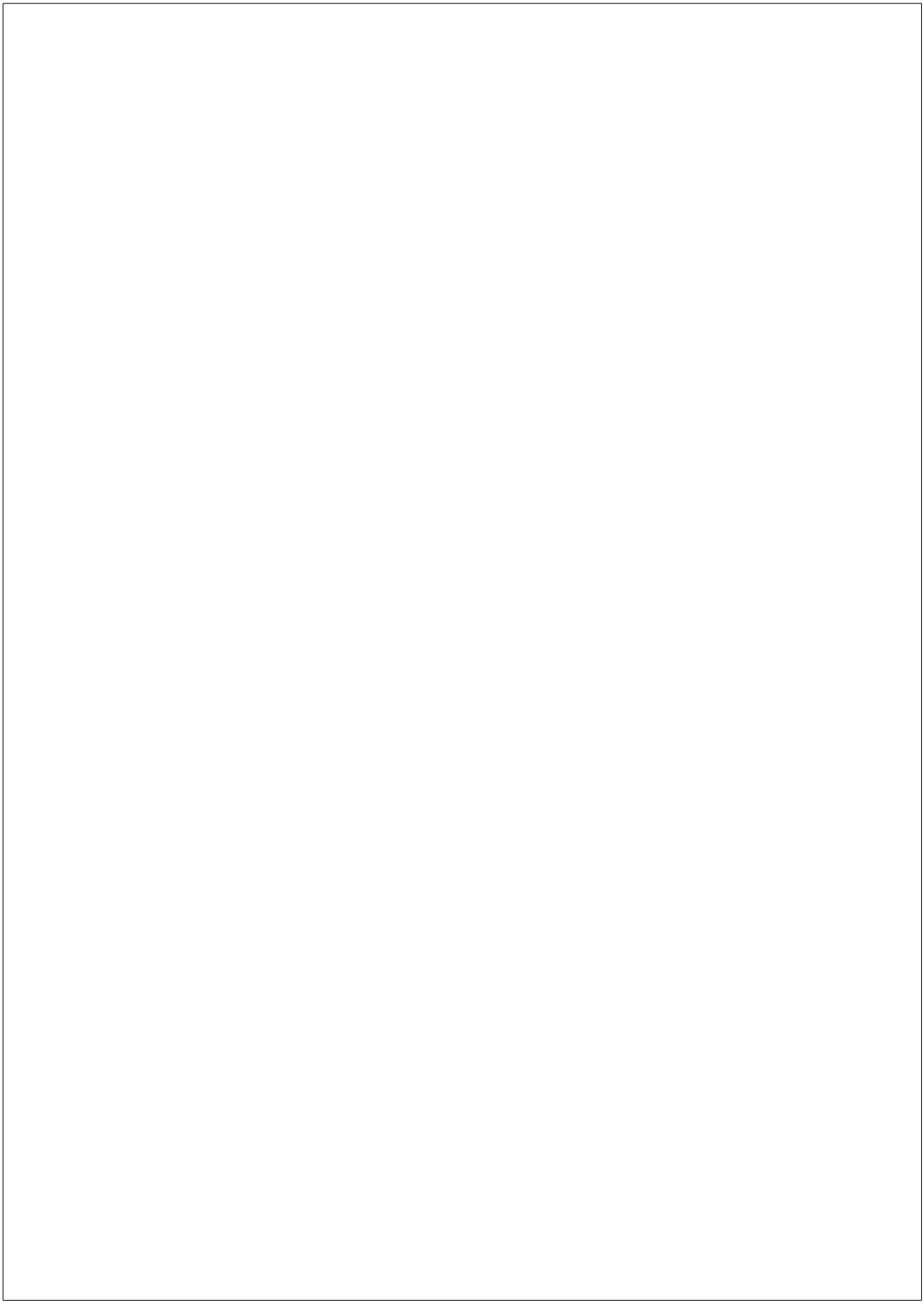
$\langle \mathbf{T}_{g\cdot}, \mathbf{M}^{+}_{\cdot c} \rangle$ is the inner product of row $g$ from binary matrix $\mathbf{T}$ and column $c$ from binary matrix $\mathbf{M}^{+}$ and indicates the number of active enhancers that binds a gene promoter. $\langle \mathbf{T}_{g\cdot}, \mathbf{M}^{-}_{\cdot c} \rangle$ indicates the number of active repressors in the promoter. The sum of these two terms is an integer between 0, when no active TFs bind the promoter, and $F$, when all TFs are activated and bind the promoter. The right side of constraint **c5** can be either 0, when both $\widehat{\mathbf{R}}^{-}_{gc}$ and $\widehat{\mathbf{R}}^{+}_{gc}$ are 0, or $F$, when one of the two equals 1. (Note that because of constraint **c2** the $\widehat{\mathbf{R}}$ coefficients cannot both be 1.) When the sum of the two inner products is zero, both $\widehat{\mathbf{R}}$ coefficients can be 0 or one of them can be 1, since $0 \leq 0$ and $0 \leq F$. However, if there is at least one active enhancer or repressor that binds the promoter, i.e. the sum of the inner products, denoted by $x$, is positive, then one of the $\widehat{\mathbf{R}}$ coefficients must be 1, since $x \not\leq 0$ and only $x \leq F$ holds. Consequently, constraint **c5** forces a gene to be either upregulated or downregulated, when there is at least one active enhancer or one active repressor in the gene promoter.

$$\textbf{c8:} \quad \mathbf{S}^{+}_{fc} - \widetilde{\mathbf{S}}^{+}_{f} \geq F \cdot (\mathbf{M}^{+}_{fc} - 1) \qquad \forall f, c$$

Constraint **c6** ensures that $\mathbf{S}^{+}_{fc}$, the strength of TF $f$ under cultivation parameter $c$, is 0, when the $\mathbf{M}^{+}_{fc}$ is 0, i.e. when $f$ is not activated under $c$. In this case constraint **c8** becomes $-\widetilde{\mathbf{S}}^{+}_{f} \geq -F$, which is always satisfied, since $\widetilde{\mathbf{S}}^{+}_{f}$, the general enhancer strength of $f$ is at most $F$. In the case that the TF is activated, i.e. $\mathbf{M}^{+}_{fc}$ is 1, constraint **c8** becomes $\mathbf{S}^{+}_{fc} \geq \widetilde{\mathbf{S}}^{+}_{f}$. Together with constraint **c7**, which states that $\mathbf{S}^{+}_{fc} \leq \widetilde{\mathbf{S}}^{+}_{f}$, it forces $\mathbf{S}^{+}_{fc}$ to be equal to $\widetilde{\mathbf{S}}^{+}_{f}$ in the case that $\mathbf{M}^{+}_{fc}$ is 1.

**c12:** $\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle \geq (F^2 + F^{-2}) \cdot \widehat{\mathbf{R}}^{+}_{gc} - F^2 \qquad \forall g, c$

$\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle$ is the inner product of row $g$ from binary matrix $\mathbf{T}$ and column $c$ from continuous matrix $\mathbf{S}^{+}$ and indicates the sum of the strengths of all active enhancers that can bind the promoter of $g$ under cultivation parameter $c$. $\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle$ indicates the total repressor strength. From constraints **c3**, **c4** and **c8** we know that if there are no active TFs that can bind the promoter, both inner products as well as $\widehat{\mathbf{R}}^{+}_{gc}$ are 0. In that case, constraint **c12** becomes $0 \geq -F^2$ and is satisfied. In the case that there is at least one active enhancer or repressor that binds the promoter, the difference between the inner products can range from $-F^2$, when all $F$ TFs are active, bind the promoter and act as repressors with the maximal strength of $F$, to $F^2$, when the enhancer strength is at its maximum. If we want to call a gene upregulated, i.e. $\widehat{\mathbf{R}}^{+}_{gc}$ is 1, than constraint **c12** becomes: $\langle \mathbf{T}_{g\cdot}, \mathbf{S}^{+}_{\cdot c} \rangle - \langle \mathbf{T}_{g\cdot}, \mathbf{S}^{-}_{\cdot c} \rangle \geq F^{-2}$. Here, $F^{-2}$ plays the role of a small positive number. Consequently, a gene can only be upregulated, i.e. $\widehat{\mathbf{R}}^{+}_{gc}$ can only be 1, when the enhancer strength is larger than the repressor strength.



# CHAPTER 8

# GENE SET ACTIVITY PROFILES

This chapter presents an alternative to the hypergeometric test procedure used to test gene groups for functional enrichment. The test described in this chapter is based on the central limit theorem. In contrast to the rest of the thesis, the method is applied to time series microarray data in order to create gene set activity profiles, which represent the enrichment of a gene set over time. Since for each gene set a unique activity profile can be derived, differences in the activity of e.g. biological processes or transcription factors in terms of the degree of enrichment and timing can be analyzed, thereby offering profound insight in (the hierarchy of) regulatory mechanisms.

This chapter is published as:

'Creating gene set activity profiles with time-series expression data'

Theo A. Knijnenburg, Lodewyk F.A. Wessels and Marcel J.T. Reinders

International Journal of Bioinformatics Research and Applications (IJBRA), Vol. 4, No.3, p. 306 - 323, 2008

## 8.1 Abstract

The use of predefined gene sets has become crucial in the interpretation of genomewide expression data. A limitation of the existing techniques that relate gene expression levels to gene sets is that they cannot readily be applied to time-course microarray data. The ability to attach statistical significance to the behavior of biological processes over time would greatly contribute to understanding the complex regulatory mechanisms in the cell.
We propose a statistical testing procedure based on the central limit theorem to assess the enrichment of a gene set. The technique is applied on time-course microarray data to generate gene-set specific 'activity profiles'.

## 8.2 Introduction

The use of predefined gene sets has become crucial in the interpretation of genomewide expression data [Curt 05]. By examining the expression of a set of genes, which are grouped on the basis of a shared biological property, one is able to establish the possible characteristic behavior of this gene set with respect to the experimental conditions under consideration. In a fast and easy way, this can provide insight into the active biological mechanisms or changes thereof.
A limitation of the existing techniques that relate gene expression levels with gene sets is that they cannot readily be applied to time-course microarray data. Measuring the transcriptional response to e.g. temperature changes, stress responses and developmental stages over time has become increasingly popular in the past few years. In December 2005 it was estimated that time-series experiments account for over 30% of all microarray studies [Erns 05]. This number is expected to grow in the coming years; not only due to the decreasing cost of such experiments, but also because of the clear advantages over static expression experiments. For example, time-course analysis enables one to derive regulatory networks [Bar 03] and investigate the regulation (in terms of reaction speed to environmental perturbations and temporal hierarchy) of the transcriptome, proteome and metabolome [Kres 06]. Thus, with the emergence of time-course expression data, also rises the need for techniques that relate this data to gene sets in order to interpret the measured transcriptional response over time. The ability to monitor the activity of a biological process over time would greatly contribute to understanding the complex regulatory mechanisms in the cell.
By far, the most common way to relate gene expression levels with gene sets is through the hypergeometric test (or Fisher's exact test) [Khat 05]. For example, to test whether the genes associated with a particular Gene Ontology category (set 1) are over-represented in the set of differentially expressed genes (set 2). Basically, two sets of genes are compared to assess the significance of their overlap under the assumption (or null hypothesis) that at least one of the groups is randomly drawn from the genome.
Such an analysis has been applied to time-course data, e.g. in Kresnowati *et al.* [Kres 06]. Here, clusters of yeast genes that exhibit similar expression over time are related to the functional categories of MIPS [Ruep 04] as well as to binding targets of known transcription factors [Harb 04]. One of the major disadvantages of this approach is that only one number ($P$-value) is produced for a particular cluster and gene set combination. This number conveys information regarding the significance of the overlap, but

does not directly relate to the different time points of the experiment. One could look at the prototype signal of the cluster and use this as an 'activity profile' of the enriched functional category or transcription factor. However, the interpretation of different degrees of enrichment for different gene sets with respect to the same cluster and, thus, the same prototype signal becomes highly ambiguous. An additional problem of this approach lies in the fact that gene expression clustering is by no means a transparent problem. Choices concerning the clustering algorithm and number of clusters are very difficult to substantiate [Dhae 05]. This is even more so when clustering time-series expression data [Bar 04].
We propose a statistical testing procedure based on the central limit theorem to assess the enrichment of a gene set. For every (interpolated) time point in a time-course experiment an enrichment $P$-value is generated, such that a gene-set specific 'activity profile' can be derived. The method employs gene scores, such as expression values or fold changes, and not a grouping based on these scores. Thus, the need for clustering in computing enrichment scores is circumvented. A gene set is enriched, when the sum of the scores of the genes in the set significantly deviates from expected sum of a randomly drawn gene set. The method can also be applied to static expression experiments. In that case, selecting a cut-off to decide which genes are differentially expressed is not necessary. The proposed method is widely applicable and its theoretical framework allows avoiding computationally expensive permutation schemes. Since for each gene set a unique activity profile can be created, it is easy to monitor the activity of a gene set over time and compare this with other gene sets. This can greatly contribute to the analysis of time-series expression data. Using several microarray datasets and comparisons with other techniques, we demonstrate the rationale and usefulness of our approach.

## 8.3 Methods

### 8.3.1 Enrichment computation

Given set $\mathcal{G}$, which contains scalar elements (in our case gene scores) $x_g$ with $g = 1 \dots G$ and set $\mathcal{S}$, which is a subset of $\mathcal{G}$ ($\mathcal{S} \in \mathcal{G}$) and contains $S$ elements, the following statistic is computed:

$$Y(\mathcal{S}) = \sum_{g \in \mathcal{S}} x_g \tag{8.1}$$

Under the assumption that each $x_g$ is an outcome (or realization) of random variable $X_g$ and that all these random variables are i.i.d., we can estimate the mean and unbiased variance of $Y(\mathcal{S})$ as follows:

$$\mu = E(Y(\mathcal{S})) = S \cdot \frac{\sum_{g=1}^{G} x_g}{G} \tag{8.2}$$

$$\sigma^2 = \mathrm{var}(Y(\mathcal{S})) = S \cdot \left[ \left( \frac{\sum_{g=1}^{G} x_g^2}{G-1} \right) - \frac{G}{G-1} \left( \frac{\sum_{g=1}^{G} x_g}{G} \right)^2 \right] \tag{8.3}$$

Next, we compute the following Z score:

$$Z = \frac{Y(\mathcal{S}) - \mu}{\sigma} \tag{8.4}$$

Note that this Z score resembles the one derived in Newton *et al.* [Newt 06], except for a term $(G-S)/G$ in the variance. This difference is explained by the fact that in their random set method, $\mathcal{S}$ is randomly sampled *without replacement* from $\mathcal{G}$, while our method is equivalent to $\mathcal{S}$ being randomly sampled *with replacement* from $\mathcal{G}$. Although drawing a gene set from all genes in the genome is a case of sampling without replacement, our approach of approximating this by sampling with replacement conforms to the i.i.d. assumptions and thus permits application of the central limit theorem (CLT). (Because of the replacement, every draw of a gene is from the same distribution, characterized by mean $\mu/S$ and variance $\sigma^2/S$, and is independent of a previous draw.) According to the CLT, $Z$ is standard normally distributed under the null hypothesis that $\mathcal{S}$ is randomly drawn with replacement from $\mathcal{G}$, which we equate to $\mathcal{S}$ not being enriched. By employing the CLT it is easy to compute a $P$-value for a particular Z-score and thus give an indication of the significance of the enrichment of a particular set $\mathcal{S}$. We will refer to this test as the CLT test. According to the CLT test, $\mathcal{S}$ is enriched when the sum of its elements significantly deviates from $\mu$. This will either be the case when all of its elements are slightly and consistently different with respect to the average of all elements in $\mathcal{G}$ or when some of its elements strongly deviate (or a combination of these extreme cases). The CLT test is implemented as a two-tailed test, which means that significant results will be reached when there is over-representation of high scores (right tail) or overrepresentation of low scores (left tail).

Warnings have been issued concerning the use of the CLT in hypothesis testing, e.g. in Yates and Goodman [Yate 99]. The approximation of the Z score (Eq. 8.4) by a normal distribution might be poor for small $S$, especially in the tails, where hypothesis testing takes place. To attain more insight in this matter, we have performed several experiments that compare the CLT approximation of $Z$, i.e. a standard normal, with the theoretical distribution of $Z$. The theoretical distribution of a sum of i.i.d. random variables is obtained by convolving the pdf of the random variable of interest [Star 86]. The results of this experiment for the continuous uniform distribution and a discrete uniform distribution with two outcomes are displayed in Figures 8.1 and 8.2, respectively.

These results indicate that for our range of interest ($P < 0.01$) the CLT approximation is always conservative, also in the case where the largest density is found in the tails of the distribution as is the case with the discrete uniform distribution in Figure 8.2. This can intuitively be expected since the normal distribution has finite probabilities for any range of numbers between $-\infty$ and $\infty$, while the random variable of interest has a finite range. Moreover, the deviation from the correct $P$-value is only large for very small $S$, and quickly dissolves with increasing $S$.

The CLT test is a 'competitive' test, which means that the scores of a gene set are compared to a standard defined by either all genes or the complement of the gene set. In contrast to this is a 'self-contained' test, which compares the gene set to a fixed standard that does not depend on the measurements of genes outside the gene set. Although by far most enrichment tests are competitive tests, it has been argued that self-contained tests are more powerful and sensible to use in gene set testing [Goem 07]. The CLT test can be transformed to a self-contained test when a proper null hypothesis or null distribution of the gene scores can be formulated. A good example is the case where the gene scores are $P$-values. $P$-values are the results of a hypothesis test and uniformly distributed between zero and one for genes for which the null hypothesis (e.g. no differential expression) is true. Then, the mean and variance of this uniform distribution can be used in the CLT test. Note that this resolves the bias introduced by the assumption

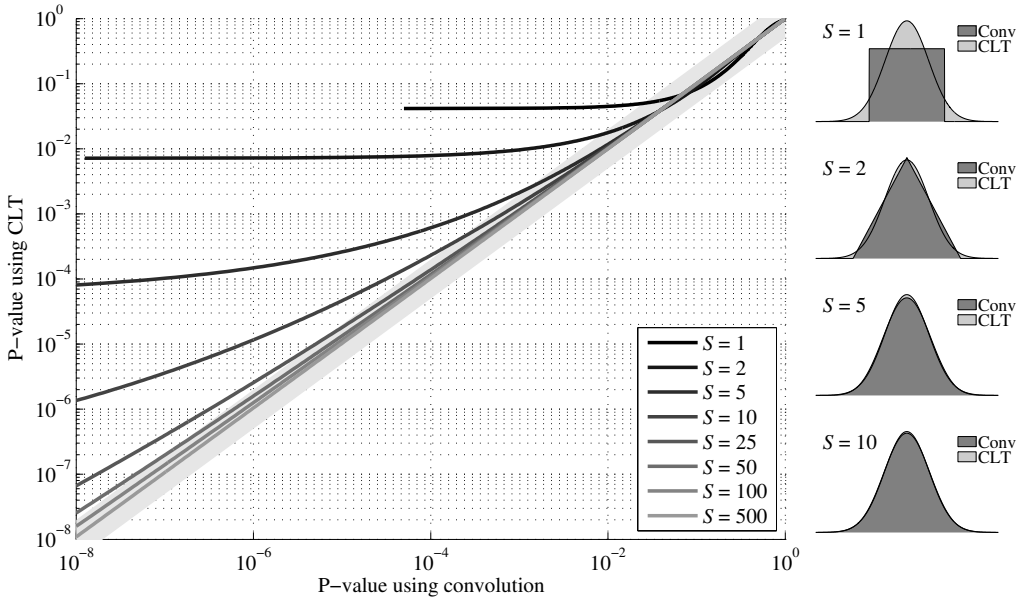



**Figure 8.1 – The theoretical distribution of a sum of i.i.d. uniform distributions, as derived via convolution, is compared with its CLT approximation.**
For the range of all possible outcomes, the $P$-values (probabilities from the cdf's of these distributions) were computed and plotted against each other. The shaded area indicates a deviation between $P$-values that is smaller than a factor two. On the right, distributions for different $S$ are visualized.

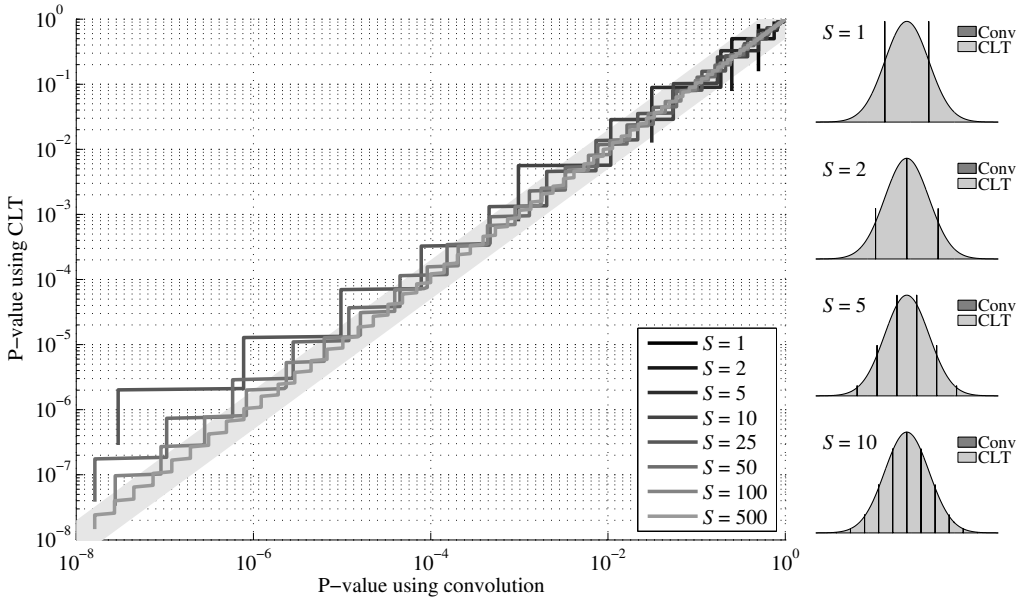



**Figure 8.2 – Identical to Figure 8.1, except now for the discrete uniform distribution with two outcomes.**
Note that the auc's in the figures on the right do not match; the normal curve is scaled up for visibility.

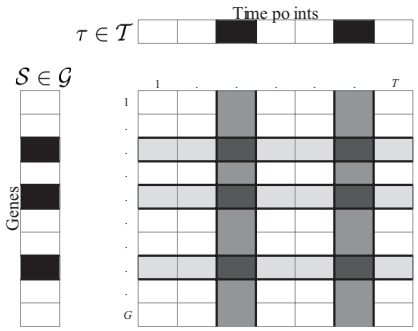
of sampling without replacement. However, for many other scores, such as ranks, or fold-changes it is not straightforward to define a null distribution.

### 8.3.2 Application to time-course expression data

Within the setting of microarray gene expression data, set $\mathcal{G}$ contains a score for each gene in the genome, while $\mathcal{S}$ is the gene set under investigation. Using the enrichment computation as explained above, we determine the significance of the deviation of the sum of scores of the gene set from the expected sum, which is obtained under the null hypothesis that this gene set is randomly drawn from the genome. Any type of gene score can be input into the algorithm, as long as the mean and variance over all scores exist and are finite. Examples include (log) $P$-values, (log) fold-changes and ranks.
In time course microarray experiments, there is a set of $G$ scores for every time point $t$ with $t = 1 \dots T$. The CLT test can simply be applied to gene scores at every time point *separately*. This will result in an enrichment score for every $t$, which we define as the activity profile of $\mathcal{S}$. Here, interpolation techniques can be employed to exploit the dependence between time points to create many more time points and, consequently, smooth activity profiles.
Also, the i.i.d. assumption (or hypothesis) can be extended to hold over time. Then, a new set, $\mathcal{H}$, is defined. $\mathcal{H}$ contains $G \times T$ elements $x_{gt}$ for all genes in set $\mathcal{G}$ and all time points in set $\mathcal{T}$. In this case, a new statistic, $Y'(\mathcal{S}, \tau)$, is computed as the sum of gene scores for the gene(s) in $\mathcal{S}$ ($\mathcal{S} \in \mathcal{G}$) at time point(s) $\tau$ ($\tau \in \mathcal{T}$):

$$Y'(\mathcal{S}, \tau) = \sum_{g \in \mathcal{S} \wedge t \in \mathcal{T}} x_{gt} \tag{8.5}$$

while $\mu$ and $\sigma$ (Eqs. 8.2 and 8.3) are determined by computing the mean and variance over all elements of $\mathcal{H}$ and multiplying this by the product of $S$ and the number of time points in $\tau$.
In general, by introducing a second factor (time), different hypotheses can be posed, because there exist multiple ways to define the set from which the parameters of the null distribution are estimated, i.e. different ways of formulating the statistical test. See Figure 8.3. For example, given the scores of gene set $\mathcal{S}$ at time points $\tau$, we can assess its possible divergent behavior with respect to either 1) the other genes at these time points (the time-specific scenario in Figure 8.3; TS_CLT), or 2) the other time points for this gene set (denoted by the gene-specific scenario in Figure 8.3; GS_CLT) or 3) all genes and all time points (the global scenario in Figure 8.3; G_CLT). Special cases arise when $\tau$ (or $\mathcal{S}$) consists of one time point (or gene) or all time points (or genes). Then for example, we can test divergent behavior (differential expression) of a single gene. In this study, we limit ourselves to the analysis of gene sets at one time point $t$, i.e. we examine the possible divergent behavior (enrichment) of a gene set at time point $t$ with respect to the scores of all genes at time point $t$ (TS_CLT) and we investigate the enrichment of a gene set at time point $t$ given the scores of all other genes at *all* possible time points (G_CLT). The latter approach allows us to detect global patterns over time, e.g. when a large portion of genes in the genome becomes upregulated during the experiment.




| Scenario | μ and σ computation | Label |
|---|---|---|
| Global | Over all elements | G_CLT |
| Time−Specific | Over the grey columns | TS_CLT |
| Gene−Specific | Over the light−grey rows | GS_CLT |

**Figure 8.3 – Visualization of matrix $\mathcal{H}$ and the different scenarios to compute the null distribution.**
The set of genes, $\mathcal{S}$, is depicted by the vector on the left, with a black element indicating that the corresponding gene is a member of the set. Similarly, the set of time points, $\tau$, is depicted by the vector at the top, with a black element indicating membership. The overlap between the genes of set $\mathcal{S}$ and the time points of $\tau$ are indicated by the dark grey squares. $Y'(\mathcal{S}, \tau)$ is the sum of these dark-grey elements, while $\mu$ and $\sigma$ are computed over either 1) all elements of $\mathcal{H}$ (denoted by G_CLT), 2) the columns defined by $\tau$ (denoted by TS_CLT) or 3) the rows defined by $\mathcal{S}$ (denoted by GS_CLT).

## 8.4 Results

### 8.4.1 Comparison to the hypergeometric test

A common first step in the analysis of gene expression data is to find the genes that exhibit differential expression between the measured cultivation conditions, developmental stages or patient/sample classes. Most algorithms for this, such as the T-test, ANOVA and SAM (Significance Analysis of Microarrays) [Tush 01] output a $P$-value and/or a Q-value (false discovery rate) that indicates the degree of differential expression of a gene. In order to compute the enrichment of a gene set using the hypergeometric test, a cut-off on the $P$- or Q-value is selected to dichotomize the genes in a group of differentially expressed and non-differentially expressed genes.
For this comparison, we employ a microarray gene expression dataset of yeast grown under four different nutrient limitations in both aerobic and anaerobic chemostat cultures [Tai 05]. More specifically, differential expression between carbon-limited aerobic growth and nitrogen-limited anaerobic growth is analyzed. A two-tailed T-test, comparing the two cultivation conditions, was performed for all (6383) genes in the yeast genome, resulting in a $P$-value for each of those genes. To compute enrichment of MIPS functional categories using the hypergeometric test, different $P$-value cut-offs to determine differential expression were selected, i.e. $P_{\text{cutoff}} = 5{\cdot}10^{-x}$ with $x = 2, 2.1, \ldots, 4.9, 5$. For the CLT test, no cut-off is selected, yet an appropriate score should be chosen as input. Here, it is sensible to use the logarithm of the $P$-values, since this transformation emphasizes small $P$-values, which is similar to setting the cut-off in the hypergeometric tests at very small $P$-values.
From Figure 8.4 it is evident that both methods (the CLT test and the hypergeometric test) give highly similar results. Notably, the CLT test, is two-tailed and produces small $P$-values for both small and large Z scores. Thus, also gene sets that are comprised of genes with significantly large scores are considered highly enriched. In the context of the hypergeometric test, these categories are under-represented in the set of differentially

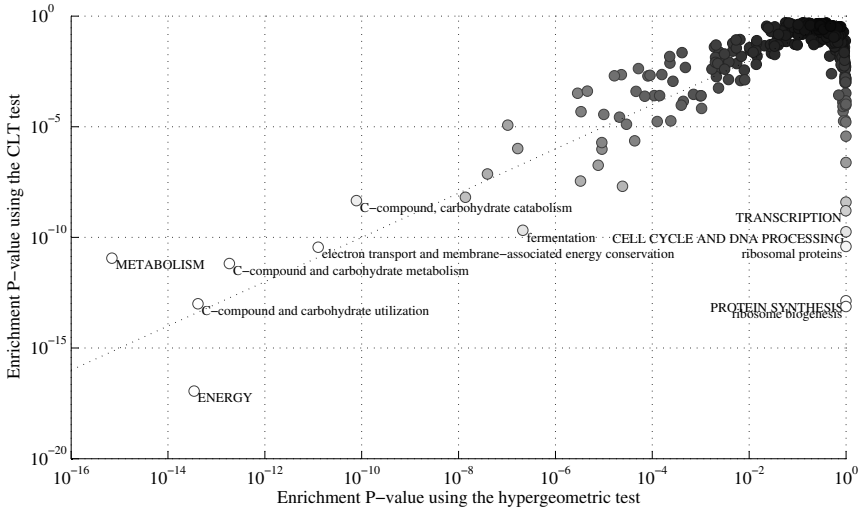



**Figure 8.4 – Comparison of the enrichment $P$-values for all MIPS categories with more than ten genes, obtained with both the hypergeometric test and the CLT test.**
In the case of the hypergeometric test, for each MIPS functional category, the minimum $P$-value over all different cut-offs was selected. The dotted diagonal line represents identical enrichment values for both methods.

expressed genes (or over-represented in the non-differentially expressed set). Although under-representation can be computed using the hypergeometric test, it is usually not done. However, the categories found here with small CLT $P$-values, and large Z scores, i.e. protein synthesis, cell cycle and transcription are very interesting in the light of the experimental setup of the microarray data set under consideration. In the chemostats the growth rate of yeast can be controlled and was kept constant and identical (0.1 $h^{-1}$) for all cultivation conditions. Therefore, it is not only interesting to find differentially expressed gene sets, but also to find the categories that exhibit much less differential expression when compared to the rest of the genome, since this can provide clues to how the cell senses the limiting nutrient and regulates itself to maintain a determined growth rate.

Additionally, we analyzed the ranking of the gene sets, since it is quite common not to look at the obtained enrichment score itself, but at the order of the most highly enriched gene sets. From Figure 8.5 we can deduce that both methods perform comparable, since the ranks of the gene sets derived with the CLT test usually fall within the variation created by the different cut-off levels used for the hypergeometric tests. However, it should be noted that gene sets can have a very different ranking (and enrichment value) based on a fundamental difference between both methods: The hypergeometric test only selects a short list of genes with extreme scores and determines over-representation of this list in a gene set, while the CLT test determines the enrichment by summing over all gene scores in the gene set. In Newton *et al.* [Newt 06] the already intuitive notion that 'summing' approaches (such as the CLT test) are more powerful when the gene set under consideration contains lots of scores that deviate only a little from the mean score, and 'selection' approaches (such as the hypergeometric test) are more powerful for gene sets, which have extreme scores for a small number of genes, was proven using an artificial

data set. In real data sets, both sorts of gene sets will be present to a more or less pronounced degree. Therefore, both approaches can have benefits in practice, and possibly complement each other. However, we conjecture that for general data interpretation the results are reasonably equivalent. Similar results were obtained using different two-class and multi-class comparisons within this dataset, as well as between different cultivation conditions in other yeast chemostat microarray data sets and between poor and good prognosis samples from the metastasis dataset, which is described in detail in the next section. (Results not shown.)

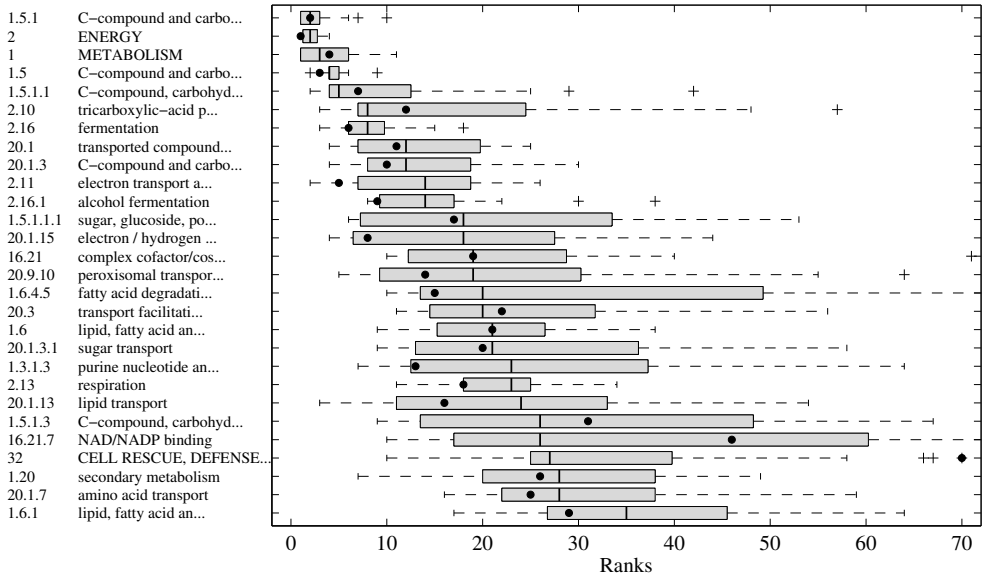



**Figure 8.5 – Boxplot of the ranks of the most highly enriched MIPS gene sets according to the hypergeometric test.**
For each $P_{\text{cutoff}}$ a ranking of the gene sets was derived on the basis of their enrichment $P$-value obtained with the hypergeometric test. These outcomes are represented by the boxplots. Also, a ranking of the MIPS gene sets based on the CLT test was derived (based on the Z score). These ranks are denoted by the filled circles.

### 8.4.2 Comparison to GSEA

Similar to the CLT test, GSEA uses all gene scores of a gene set to compute enrichment and does not place a threshold on these scores. The first implementation of GSEA [Moot 03] only uses the ranking of the genes based on their scores to compute enrichment values. The goal of GSEA is to determine whether the members of a gene set tend to occur toward the top or the bottom of the rank-ordered list of genes. For this, a Kolmogorov-Smirnov statistic (KS) is computed. In a later version [Subr 05], the scores are used as weights, resulting in a weighted KS. Permutation tests are performed to compute an empirical $P$-value by counting the number of times the KS, as computed on a permuted data set, exceeds the original KS. In GSEA this permutation takes place on the class labels of the microarray data set under investigation. This is in contrast to the CLT test, which assumes gene sets that are randomly drawn from the genome and

is equivalent to permuting the gene labels. This results in a different hypothesis being tested. (This fact is overlooked by Kim and Volsky [Kim 05], where also a Z-score for a gene set is computed, in their case using the fold change between two condition as gene scores. Here, we will not address the theoretical concerns of this issue. For that discussion, we refer to Tian *et al.* [Tian 05] and Efron and Tibshirani [Efro 07].)

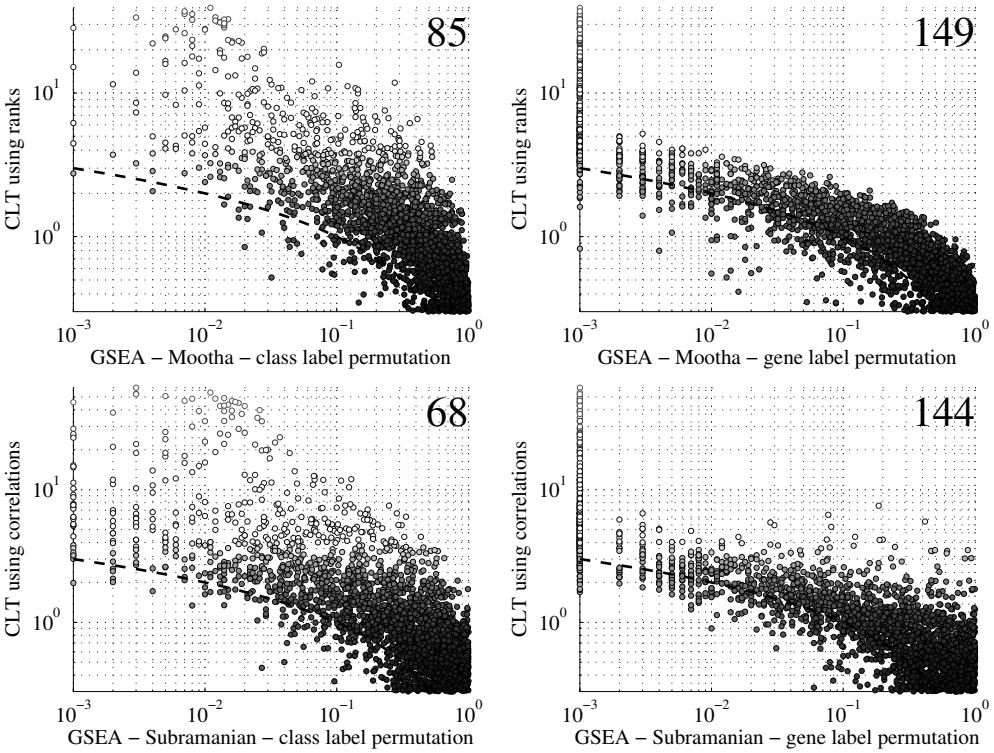



**Figure 8.6 – Comparison of the enrichment $P$-values for more than 2500 gene sets (with more than ten genes), obtained with both GSEA and the CLT test on the Van de Vijver data [Vijv 02].**
For visibility, the -$\log_{10}$($P$-value) is given for the CLT test. The gene sets are derived from MSigBD [Subr 05], Gene Ontology [Ashb 00] and others. For GSEA, $10^3$ permutations were performed. In these plots, GSEA $P$-values of zero were set to $10^{-3}$. The number in the upper right corner indicates the overlap of the 150 most enriched gene sets for both methods. The dashed line represents identical enrichment values for both methods.

For our comparison, we employed the breast cancer microarray dataset of Van de Vijver [Vijv 02], which contains genomewide expression measurements for 258 patients (65 with poor prognosis, 193 with good prognosis). For all (24481) genes, the Pearson correlation between the class labels and expression pattern of a gene was computed. We compare 1) the normalized version of GSEA of Mootha *et al.* [Moot 03] (implementation of Subramanian *et al.* [Subr 05] with $p = 0$) to the CLT test using the ranks (based on the correlations between the class labels and expression patterns) as gene scores, and 2) GSEA of Subramanian *et al.* [Subr 05] ($p = 1$) to the CLT test using the correlations as gene scores. Furthermore, we apply both class label permutation and gene label permutation to compute GSEA's enrichment $P$-value. For GSEA we performed 1000

permutations. From Figure 8.6 it is clear that the enrichment $P$-value from the CLT test and from GSEA are similar in the case of gene label permutation. There is, however, a large difference in the $P$-value range between both approaches. This is because the number of permutations determines the $P$-value resolution. The permutation scheme is computationally intensive and time consuming. (Remember that for the CLT test no permutations are necessary.) Additionally, in the case of class label permutation in combination with a dataset with relatively small classes (e.g. triplicate measurements of yeast cultivation experiments), the number of possible permutations is very limited, resulting in a low resolution $P$-value.

### 8.4.3 Activity profiles for a glucose pulse

In Kresnowati *et al.* [Kres 06], the global transcriptional response of the yeast *S. cerevisiae* to a glucose pulse was investigated. Initially, the yeast was growing in glucose-limited chemostats, where metabolism is fully respiratory, after which the glucose concentration was instantaneously increased. Triplicate samples were taken at $t = 0, 30, 60, 120, 210, 300$ and 330 seconds after glucose addition.
Here, we employ this time-course microarray dataset to create activity profiles of gene sets. First, we interpolated the expression profile of each gene using piecewise cubic spline interpolation [Bar 04], such that we have an expression level for each second after the glucose pulse. Next, we computed the $\log_2$ fold-change between the expression level at each time point and the expression level at $t = 0$. These $\log_2$ fold-changes are used as gene scores in our algorithm. The employed gene sets are the MIPS functional categories [Ruep 04] and the binding targets of known transcription factors [Harb 04]. In Kresnowati *et al.*, these gene sets were related to one of two clusters, i.e. a cluster of upregulated and a cluster of downregulated genes, through the hypergeometric test. In contrast, with our approach we create a unique activity profile for each gene set without clustering the genes beforehand. Figures 8.7 and 8.8 display activity profiles for some functional categories and transcription factors that were over-represented in one of the clusters of Kresnowati *et al.*. In general, we found that gene sets with a larger hypergeometric test $P$-value in Kresnowati *et al.*, had a larger maximum enrichment value in their activity profiles. Furthermore, since we create a unique profile for each gene set in stead of relating a gene set to the upregulated or downregulated cluster, we are able to detect differences in the transcriptional response time after the glucose pulse for different gene sets. For example, the targets of stress responsive element transcription factors, Msn2 and Msn4, which are part of the glucose-sensing pathway [Gela 03] are downregulated at an earlier stage than Nrg1 and Sko1, which are involved in glucose catabolite repression [Rep 01, Berk 04]. See Figure 8.8. (Results for Msn4 and Sko1 and not shown, but are very similar to Msn2 and Nrg1, respectively). These results provide clues towards the dynamics of the glucose-induction signaling and high osmolarity MAPK signaling pathways in *Saccharomyces cerevisiae* as recently reconstructed in Arga *et al.* [Arga 07]. Another notable observation is that genes involved in amino acid metabolism have a higher $\log_2$ fold-change w.r.t. the glucose-limited steady-state ($t = 0$) compared to the other genes in the genome, already from right after the glucose addition.
In Kresnowati *et al.*, it was established that there is no or only very little transcription response until between 120 to 210 seconds after the pulse, when major transcriptional changes start to occur. Thus, when applying the CLT test by computing the mean and

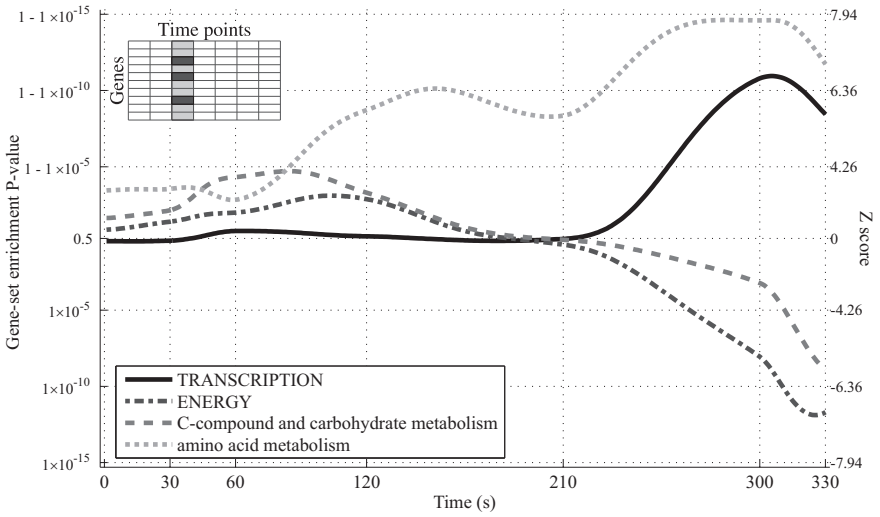



**Figure 8.7 – Activity profiles for four functional categories derived by applying the TS_CLT test at each time point.**
The small embedded figure (top-left) is similar to Figure 8.3 and visualizes the scenario used to derive the activity profile. In this TS_CLT case the score of a gene set at one time point (dark-grey squares) is compared to the scores of all genes at that time point (grey column). The left vertical axis indicates the enrichment $P$-value; the right vertical axis indicates the corresponding Z score. In Kresnowati *et al.*, Transcription and amino acid metabolism were related to the cluster of upregulated genes; Energy and C-compound and carbohydrate metabolism were related to the cluster of downregulated genes, all with $P < 10^{-14}$.

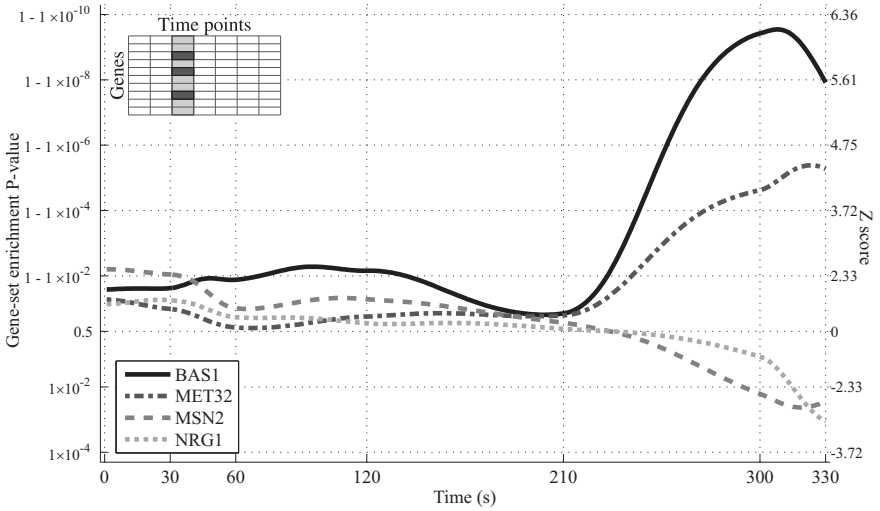



**Figure 8.8 – Identical to Figure 8.7, except now for four transcription factors.**
In Kresnowati *et al.*, Bas1 and Met32 were related to the cluster of upregulated genes with $P = 5.94 \cdot 10^{-11}$ and $P = 1.80 \cdot 10^{-3}$ respectively. Msn2 and Nrg1 were related to the cluster of downregulated genes with $P = 7.50 \cdot 10^{-6}$ and $P = 9.99 \cdot 10^{-4}$ respectively.

variance over all gene scores at all time points (G_CLT), enriched gene sets at the earlier time points are no longer found significant. See Figure 8.9. This is because the variance

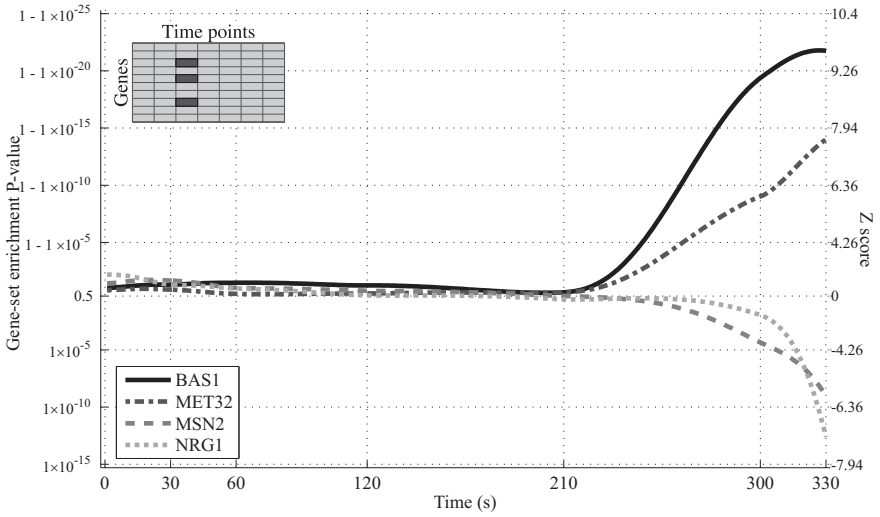



**Figure 8.9 – Identical to Figure 8.8, except now by applying the G_CLT test for each time point.**

over the $\log_2$ fold-changes at all time points is larger than the individual variances for the earlier time points, thereby shrouding these subtle variations and only uncovering the global patterns. In general, profiles derived with G_CLT and TS_CLT will exhibit significant differences, when during the time-course there are large changes in the overall activity of the transcriptome.

### 8.4.4 Activity profiles for yeast's cell cycle

The TS_CLT test was applied on the cell cycle microarray dataset of Spellman *et al.* [Spel 98]. In this work, yeast cultures were synchronized using three different methods: $\alpha$ factor arrest, elutriation and arrest of a cdc15 temperature-sensitive mutant. Also the data of Cho *et al.* [Cho 98] was included, where the cultures were synchronized using a cdc28 mutant. $\text{Log}_2$ fold-changes were obtained by comparison with gene expression measurements of asynchronous cultures of the same cells growing exponentially at the same temperature in the same medium. Again, we use these as gene scores to be input in the algorithm. In this case, no interpolation was performed. Activity profiles were derived for gene sets comprised of the binding targets of known transcription factors [Harb 04]. To relate a time point to a cell cycle phase (M/G1, G1, S or G2/M), we applied the TS_CLT test to gene sets, which are comprised of genes that are known to be regulated in a particular cell cycle phase. These gene sets, which are determined by traditional methods, are also used in the methodology of Spellman *et al.* [Spel 98]. Time points were labeled with the cell cycle phase corresponding to the most highly upregulated gene set. For the cell cycle dataset, there are no significant differences between the results of G_CLT and TS_CLT.
We compared the derived transcription factor activity profiles to the findings of Simon

*et al.* [Simo 01]. In this work, an attempt was made to identify the serial regulation of transcription factors in yeast's cell cycle. Regulators were related to a particular phase of the cell cycle using the 800 cell cycle related genes as found by Spellman *et al.*. Figures 8.10 and 8.11 display the activity profiles of Ace2, Mbp1, Fkh2 and Mcm1, which according to Simon *et al.* [Simo 01] are involved in the regulation of the (subsequent) M/G1, G1/S, G2 and G2/M cell cycle phases, respectively. Indeed, we obtain this

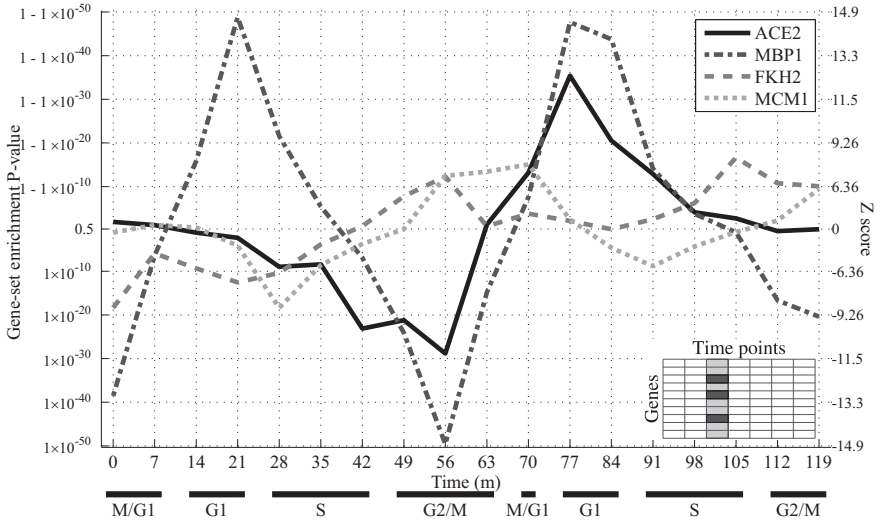



**Figure 8.10 – Activity profiles for four highly enriched TF's.**
Synchronization of the yeast cultures in this cell cycle data set, was achieved using $\alpha$ factor arrest. Each time point was linked to a cell-cycle phase. This is indicated below the x-axis.

pattern, which can be seen from the fact that the peak (and the rise) of the profile of Ace2 is followed by the peak of Mbp1, followed by the peaks for Fkh2 and Mcm1, after which the cycle begins again. Note that the profiles were generated using the $\log_2$ fold-changes of all genes in the genome, and not by using the 800 cell-cycle related genes. Moreover, the cell-cycle phases, which are assigned to the time points using the TS_CLT test, correspond to the activity of the transcription factors, as determined in Simon *et al.*, thereby providing additional justification of the truthfulness and power of our analysis.

## 8.5 Discussion

In this study, we have introduced a technique that employs time-course expression data to derive activity profiles, which represent the enrichment of a gene set over time. Since for each gene set a unique activity profile can be derived, differences in the activity of e.g. biological processes or transcription factors in terms of the degree of enrichment and timing can be analyzed, thereby offering profound insight in (the hierarchy of) regulatory mechanisms. Other approaches have been proposed to derive an activity profile of e.g. a transcription factor [Rone 06]. However, their approach heavily relies on the complex modeling of the mRNA quantity and parameter estimation based on a-priori knowledge. Our algorithm, on the other hand, is a fast and easy tool in data interpretation of time

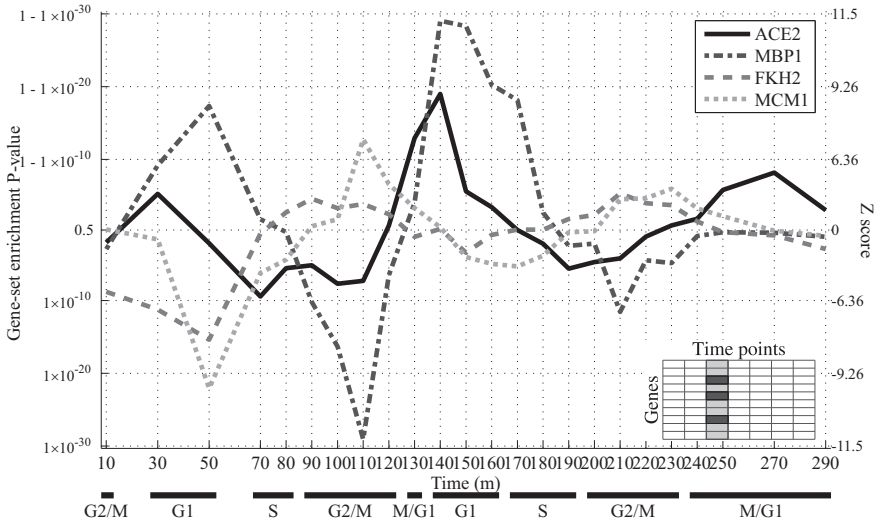
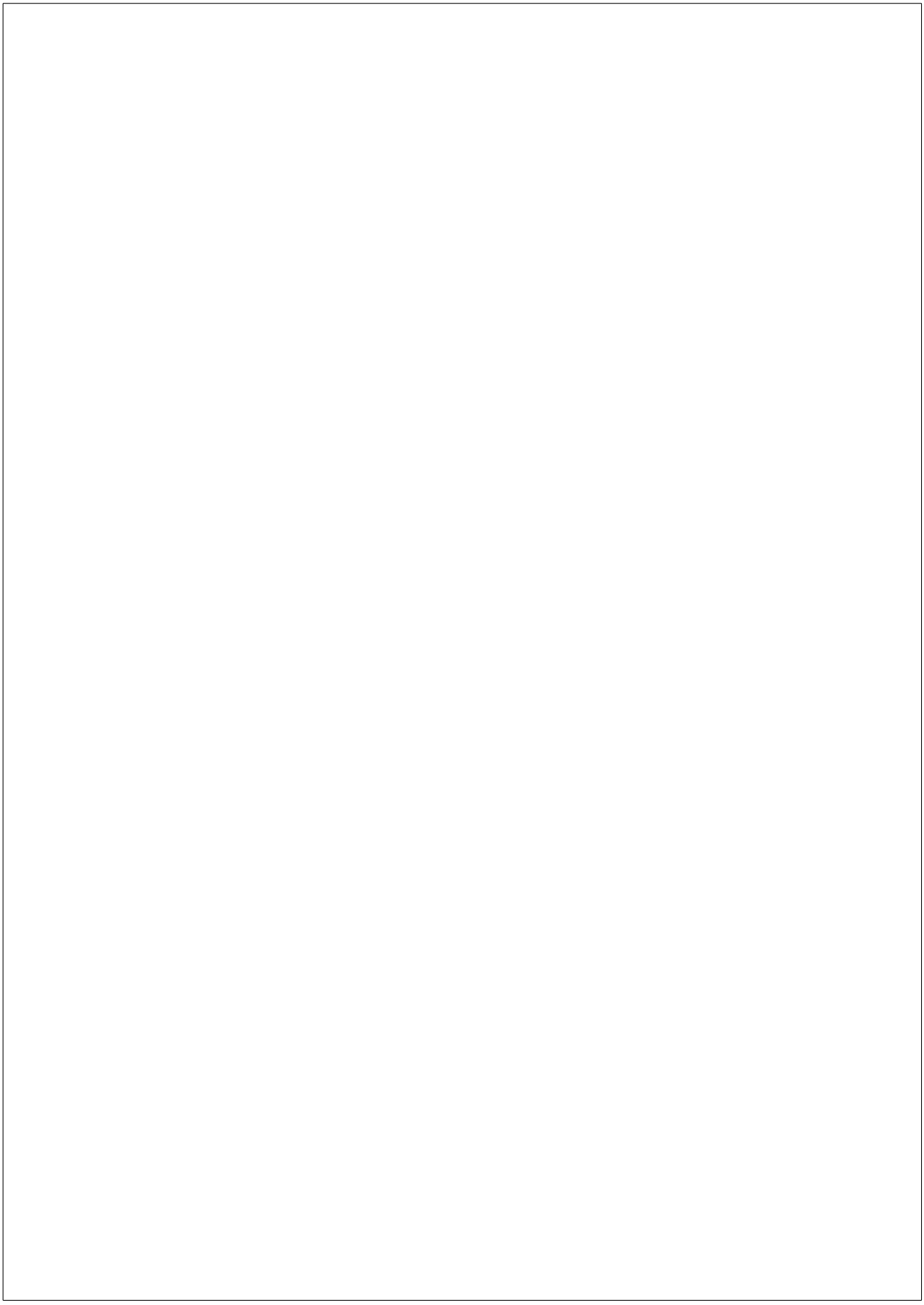



**Figure 8.11 – Identical to Figure 8.10, except here a different synchronization technique was used.**
In this case synchronization was performed using arrest of a cdc15 temperature-sensitive mutant.

course expression data.

The underlying statistical test to assess the enrichment of a gene set is based on the central limit theorem. For each time point in the time-series, the CLT test evaluates the significance of the scores of the gene set with respect to all genes in the genome. Any type of score can be used. We have demonstrated that by choosing appropriate gene scores, we can obtain similar results in comparison to two widely-used enrichment tests, i.e. the hypergeometric test and gene set enrichment analysis (GSEA). The CLT test has the following advantages: In comparison to the hypergeometric test, no cut-off needs to be selected to dichotomize genes into differentially or non-differentially expressed genes or into clusters. Especially in the range of small $P$-values, the hypergeometric test is very sensitive to cut-off selection in the sense that a small change in the chosen cut-off can lead to large differences in the enrichment score. In comparison to GSEA, the time-consuming permutation scheme can be avoided.



# DISCUSSION

All computational methods that have been presented in this thesis integrate gene expression data with the growth conditions under which the microarrays were performed. The main motivation behind the incorporation of the growth conditions into the computational model is to enable the interpretation of the results in terms of the growth conditions. Throughout this thesis different computational techniques were employed to properly address the biological questions that accompanied the gene expression datasets. In this section, we discuss some important issues concerning the employed computational techniques. Thereafter, we sketch directions for future work in this area.

## Discretization

In Chapters 2, 3 and 4, discretization of the gene expression patterns was employed to describe gene and clusters of genes in terms of their transcriptional response to particular cultivation parameters. This offers an advantage with respect to standard clustering algorithms, where interpretation of the clusters in terms of the growth conditions is often ambiguous. See Figure 1a. On the other hand, for some genes the discretized expression patterns do not do justice to their complex continuous expression patterns. For example, most genes in the five k-means clusters of Figure 1b-f do not have an obvious discretized (tertiary) expression pattern. The Davies-Bouldin index [Davi 79] or other cluster validity measures can be employed to decide upon the quality of the discretization. Genes for which no compact and well-separated clusters of conditions are found, could be omitted from analysis or placed in a special group and analyzed separately. Alternatively, the number of discretization levels could be increased such that a condition could occupy more than three discrete expression states. However, in that case it is no longer possible to call a gene upregulated, downregulated or having basal expression.

In Chapter 6 we have seen that the majority of genes has widely differing expression levels across the 55 growth conditions, even when compensating for the mRNA extraction protocol effect. In most cases, there is no expression level, which can be called the basal expression level of that gene, simply because the expression level is fluctuating across most conditions in stead of being constant across the majority of conditions. The discretization strategy to choose a basal expression level and identify conditions which are up- or downregulated with respect to this basal expression level can therefore be a serious oversimplification when applied to a large set of conditions. Therefore, the

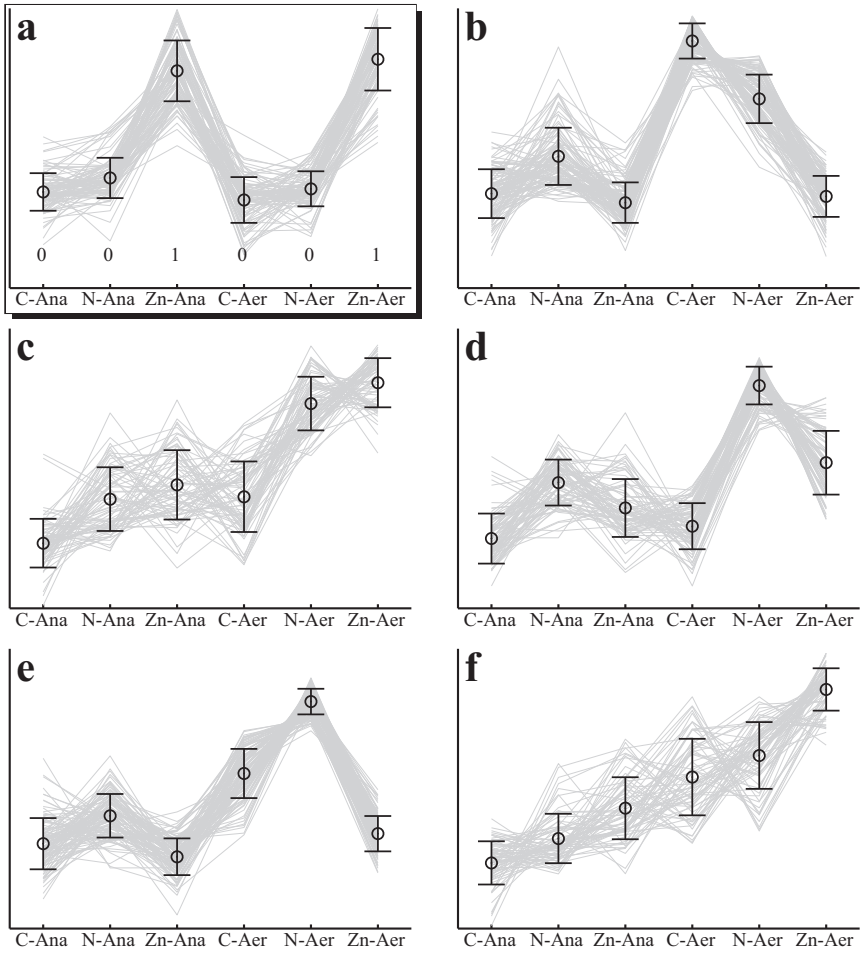



**Figure 1 – Normalized expression patterns of different clusters.**
**a:** Normalized expression patterns of genes in Cluster (or Module) 1 from Chapter 3. These genes were clustered together based on their discretized expression patterns as explained in Chapter 3. The cluster can be characterized as upregulated under zinc limitation irrespective of oxygen presence, the discretized expression pattern being [0 0 1 0 0 1]. **b-f:** A portion of clusters derived from the same data (1500 differentially expressed genes) using k-means clustering. The five clusters have a similar number of genes, are more compact (smaller inter cluster distances) compared to the first cluster and have similar enrichment in functional categories. However, the interpretation of the expression pattern of the genes in the clusters in terms of the growth conditions is ambiguous, i.e. one cannot clearly indicate under which cultivation parameters genes are up- or downregulated.

tertiary discretization strategy is only sensible to apply on expression datasets with a relatively small number of conditions ($< 10$).

## Model complexity

In modeling (building a conceptual representation of some phenomenon) there is always the trade-off between model complexity and predictive accuracy. Modeling (part of) the biological cell is a great challenge for three reasons. First, cellular mechanisms are characterized by complex, nonlinear functions. The components that comprise the model of some cellular phenomenon (such as CRPs, TFs and promoter regions to model gene transcription rates) form an intricate maze; their properties and mutual interactions give rise to complicated behavior (such as competitive binding) that can only be captured by complex models. Second, relevant components might still be unknown and, therefore, missing in the model. Third, intracellular measurement data is often lacking (when the measurement technique is unavailable) or noisy and of too small-sample size to learn complex models.

In this thesis, growth conditions were described using all cultivation parameters that differ between the experiments of the microarray dataset under investigation. Ideally, the setup of cultivation parameters is fully combinatorial (or factorial), i.e. the experimental design includes all possible combinations of settings across all cultivation parameters (factors). From a practical point of view, such approaches are only feasible with a very limited number of cultivation parameters. For example, both Chapter 4 and Brauer *et al.* [Brau 08] use a factorial design with two cultivation parameters. For factorial designs with many factors the number of necessary microarray experiments explodes. For example, the ten cultivation parameters in Chapter 6 would result in 403200 distinct growth conditions when a full combinatorial setup would be applied.
In addition, from a statistical point of view the number of (independent biological) replicates should be as large as possible. However, normally not more than three replicates are available per growth condition, which is rather small. In general, more samples means more statistical power and the ability to apply more complex models.
Here, a trade-off can be made between sample size and heterogeneity. By neglecting particular cultivation parameters (because they are thought to be of minor importance), growth conditions that only differed in terms of these cultivation parameters will be merged, leading to more samples per growth condition. Especially in the medical setting, where e.g. groups of patients with different disease development are compared, patient-specific factors like gender, age, etc. are often ignored. On the one hand, this leads to confounding effects and thus more heterogeneity within a group of patients; on the other hand, the larger number of samples within one group allows for more complex (classification) models. Obviously, such trade-offs depend on the intended purpose and should not be made beforehand. Both biological validation and statistical validation (e.g. classification error or model fit criteria) can be used to decide whether samples from different conditions should be pooled. In this work, such a trade-off was not considered, because all cultivation parameters that characterize a growth condition were considered potentially relevant and we were interested in their effect on gene expression.

This thesis employs reasonably simple models to relate cultivation parameters to gene expression and TF activity. In Chapter 6, a forward step-wise ordinary least squares regression strategy has been employed, where cultivation parameters have an additive effect on gene expression. Non-linear effects were introduced by applying logic functions (AND, OR) to the cultivation parameters, which were represented as binary predictors in the model. However, when adding these interaction effects one must guard against overfitting and loss of statistical power. For this reason, (similar to other approaches, e.g. [Bonn 06]) it was not desirable to incorporate second (or higher) order interaction effects.
The smaller, but factorial design of the microarray dataset in Chapter 4 allowed for a slightly more complex approach with regard to the modeling of one particular cultivation parameter, i.e. aeration type. The 'oxygen effect' was successfully modeled to have both an additive as well as a multiplicative effect on gene expression, hinting at the fact that the additive model of Chapter 6 is inadequate. On the other hand, the setup of cultivation parameters in Chapter 6 does not allow for such a modeling approach to be straightforwardly applied. That is, the aerobic growth conditions do not completely match with the anaerobic conditions. For example, acetate was only used as a carbon source under aerobic conditions and yeast was only grown at lower temperature (12°C) in the anaerobic case. This complicates estimation of the 'oxygen effect' when compared to the factorial design of Chapter 4.
Chapter 7 presents the most complex model of this thesis, where we assess which cultivation parameters activate which TFs and estimate the strength of TFs such that their interplay on a gene's promoter explains the gene expression level. Throughout this model we used binary (0, 1) and tertiary (-1, 0, 1) variables. (The continuous regression coefficients of Chapter 6 were also discretized.) The use of integers facilitates complex (non-linear) relations between the variables in the model. Boolean logic and if-then logic (e.g. 'If only enhancers bind the promoter, then a gene is upregulated') can be easily implemented. Furthermore, integer programming can be used to solve the complex optimization problem effectively and efficiently. On the other hand, a discrete representation of cellular events might be a poor approximation of the 'continuum' that is the cell.

In bioinformatics research, the trade-off between model complexity and prediction accuracy is not straightforward. Of course, it is important to satisfy (to some degree) the assumptions on the data imposed by the employed model. For example, in the regression approach (Chapter 6) the use of log RMA expression values in stead of the standard absolute MAS expression values was considered more suitable, because the variance of the error (or noise) component in this model is assumed to be constant. For absolute expression levels the measurement error was observed to be proportional to the expression level (multiplicative), while for log expression values the error becomes less dependent on the expression level and can more easily be modeled as a constant additive component. Furthermore, internal validation measures and strategies like cross-validation can be applied to avoid overfitting. However, the biological interpretation of the results is also very important to assess the suitability of the model. For example, the regression coefficients of significant predictors were required to have an absolute value larger than a particular threshold (0.3 in this case) to prevent regression coefficients near zero. Although small (absolute) weights can be significant according to the regression model, they are likely not to have any biological significance (i.e. effect in the cell).

The behavior of a particular model when applied to the often high-dimensional and complex cellular data is not something that can easily be predicted or thought through beforehand. Internal validation, external validation (like functional enrichment tests) and biological interpretation are crucial aspects to determine the suitability of a model in describing the data. These validation techniques should be used to deduce clues and inspiration to adjust the model in order to advance to an appropriate description of the data and thereby answer the biological question.
Therefore, we think that proper bioinformatics research should follow a cycle that is similar to the scientific cycle (or method) used to generate new scientific theories or hypotheses. See Figure 2. In the scientific method observations (and existing knowledge)

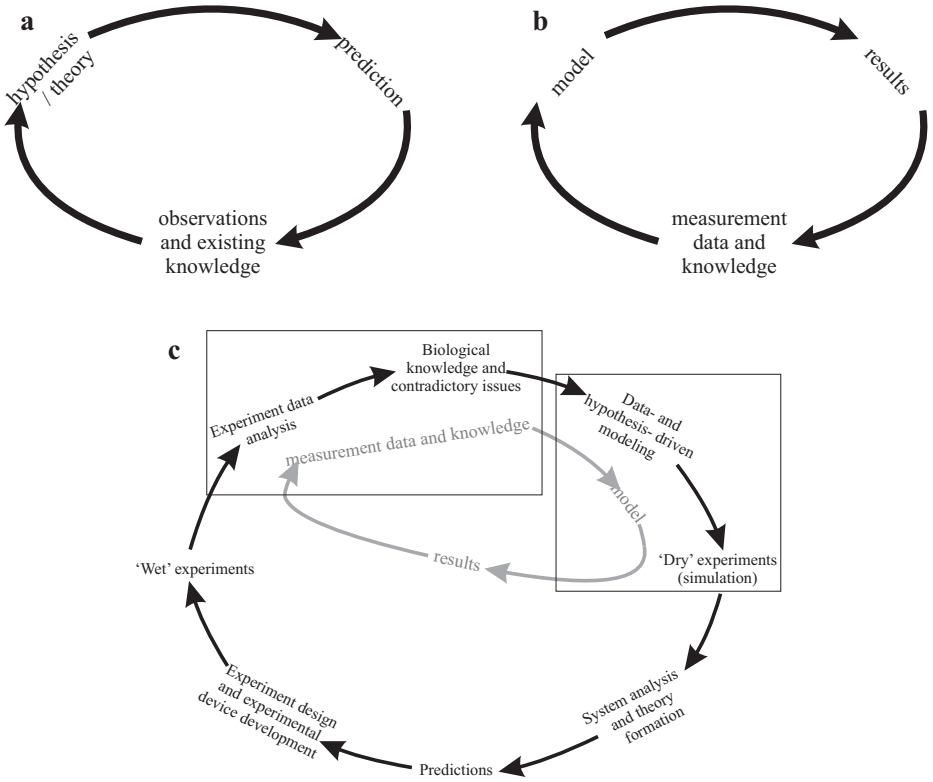



**Figure 2 – Three cycles for hypothesis generation**
**a:** The general scientific cycle. **b** The proposed bioinformatics cycle. **c** In black (font color and arrows): The hypothesis driven systems biology cycle as postulated by Kitano [Kita 02]. In grey: the proposed bioinformatics cycle as an innerloop of Kitano's cycle.

are used to generate some hypothesis or theory. See Figure 2a. Based on this hypothesis predictions can be made, which can be verified using the/new observations and provide hints towards the validity of the hypothesis and possible adjustments (improvements) to the hypothesis. Then, new predictions can be made based on the adjusted hypothesis, and so on. We envision a similar cyclic process in bioinformatics research: Based on cellular measurement data and existing knowledge (about cell biology and statistics/machine learning) a computational model is derived to answer the biological

question. The results obtained when applying this model to the data are validated using both biological interpretation and analysis of internal/external statistical scores. Again, this would allow one to adjust the model to more suitably describe the data and tackle the biological question. See Figure 2b.
Kitano [Kita 02] proposed a hypothesis driven cycle for systems biology research. In this cycle, computational ('dry') experiments are used to generate hypotheses or predictions that are tested using lab ('wet') experiments. We believe this cycle can be improved by introducing the proposed bioinformatics cycle as an innerloop on the 'dry' side. See Figure 2c. In principle, the proposed bioinformatics cycle does not use new experimental data (although this is possible), but statistical and biological validation on the original experimental data as evaluation and subsequent improvement of the model. In our opinion and experience, several iterations (cycles) of adjustments to the model are necessary to derive a suitable description of the data that can be used to generate new theories or hypotheses to be tested in the lab.

## Condition specificity

The interplay between TFs to control the transcriptional rates remains elusive. Currently, there are no models (logical, biophysical or other) that allow one to successfully predict a gene's transcription rate given the upstream binding of different TFs. Here, Chapter 7 and other related work (cited in Chapter 7) form a starting point to uncover the synergistic and antagonistic effects between TFs. Many of these approaches are hampered by the employed TF binding data. The most frequently used and by far the largest TF binding dataset [Harb 04] does not only contain many false positives [Gao 04], but is inadequate in describing complete TF binding potential. That is, only a small number of different growth conditions were employed to profile the binding locations of the many TFs in this dataset. Under other growth conditions different binding locations will be uncovered. This is due to the possible interaction with different TFs, co-factors or other proteins that could be (in)activated in new conditions, but also due to the altered accessibility of the upstream region due to chromatin remodeling [Beye 06]. Many of the false positive binding sites found with motif scanning approaches are regions inaccessible to TFs due to chromatin [Narl 07]. At this moment, no genome-wide chromatin occupancy datasets are available under a range of different growth conditions. However, the activity of chromatin remodeling proteins and nucleosome occupancy appears to be very much condition dependent [Pokh 05].
In general, the extracellular environment has a large influence on *all* ("omics") levels of the cell and, consequently, on measurements thereof. For example, synthetic-sick-and-lethal interactions are also condition dependent. The synthetic-sick-and-lethal screens have been performed on yeast growing under rich media conditions [Tong 04]. Yet, in the steady-state chemostat microarray compendium, there are certain growth conditions, where both genes of a lethal interaction pair are not expressed, while the yeast is still happily growing. This observation of context-dependency of (measurements on) cellular components is very important to take into account when integrating different data sources.

## Directions for future work

*Microarray experiments of dynamic processes*

To infer the effect of cultivation parameters on gene expression, this thesis only uses steady-state microarray data. This provides a static picture of the organism's transcriptional response to a set of cultivation parameters. It is possible to uncover the changes in gene expression in relation to the changes in cultivation parameters. However, it is very hard to find out how these changes came to be. Transcriptional adaptation to differences in the extracellular environment can be analyzed using time-series microarray experiments, where microarray measurements and (simultaneous measurements of environmental parameters) are taken over time as extracellular conditions change. In that case, the cultivation parameters can no longer be modeled as categorical variables, but assume continuous values. In principle, similar strategies, such as the described regression approach in Chapter 6, can be applied with time as an additional parameter. However, strategies that model the time aspect in a more sensible or sophisticated way, such as Markov chains or differential equations, are likely more promising. Anyhow, the modeling of dynamic experiments enables one to derive regulatory network models that incorporate temporal hierarchy and causal relations amongst and between the TFs and cellular processes that mediate the transcriptional response that enables the organism to adapt to changed conditions.

*Combinatorial regulation by TFs*

As mentioned above, the systematic analysis of combinatorial regulation by the TFs (and histones) remains underdeveloped. Bussemaker [Buss 06] draws the analogy between transcriptional regulation and Ohm's law, which applies to passive electrical circuits containing only conductors. Here, the rate of transcription is directly proportional to the activity of TFs, where the conductivity (or proportionality constant) is represented by the regulatory coupling between TFs and their target genes. This conductivity can be modeled as the binding affinity of a TF to the promoter of a gene. Although this linear model of transcription regulation suffices as a first-order approximation, many genome-wide studies and especially studies on the mechanisms behind the transcription control of individual genes (or genes in a single pathway) demonstrate a far more complex role for TFs. Transcriptional regulation should therefore be compared to electrical circuits with active components, where TFs can act upon each other and in a combinatorial (synergistic or antagonistic) way influence gene expression.

One way to model these combinatorial effects is by using logic functions that can be applied after discretization of TF activity, TF binding and/or gene expression. Logic functions (or networks) form a reasonable approximation to many known mechanisms involving multiple TFs, such as the cooperation between TFs, hierarchy amongst TFs, TFs repressing activators or even directly interacting with the basal transcription machinery (thereby inhibiting its function), etc. Furthermore, these logic functions result in interpretable TF networks, from which testable hypotheses can be deduced. Most methods that employ combinatorial effects do not explicitly model these effects as logic functions. For example, in Beer and Tavazoie [Beer 04] the presence (or absence) of motifs is used to explain cluster membership. In Chapter 7 the presence of activated enhancers and repressors on a gene promoter together with their inferred strengths determine, whether a gene is upregulated or downregulated.

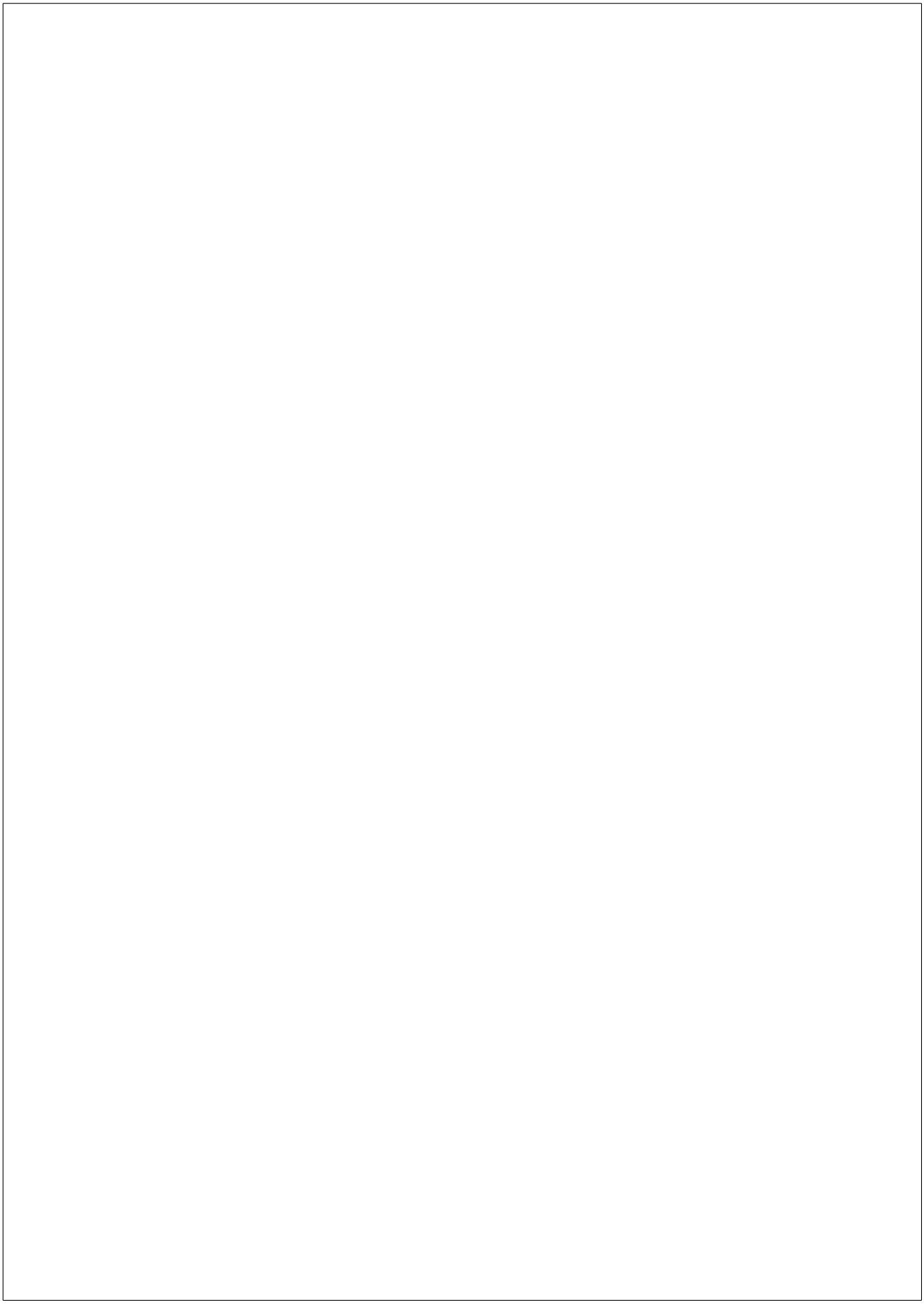
However, these models do not specify the combinatorial or logic relations (AND, OR) that exist between the motifs/TFs. Methods that do explicitly incorporate logic functions are usually limited to pairwise interactions between TFs due to model complexity issues. Promising techniques that are able to model higher order logic are inductive logic programming (ILP), logic regression and integer optimization tools. ILP uses background knowledge (e.g. TF binding data and the model that active TFs can (combinatorially) affect gene expression) and positive and negative examples (e.g. inferred TF activities and gene expression measurements) to induce hypotheses about transcription regulation. Logic regression can be used to concatenate logic gates and derive a Boolean network that explains (discretized) gene expression levels. Integer optimization tools enable the search of high dimensional discrete solution spaces. Constraints on the variables can be formulated to represent the interplay between TFs.
Also, a challenge lies in extending the biophysical model of TF binding as introduced by Foat *et al.* [Foat 06] with combinatorial effects. The biophysical model is based on the thermodynamic equilibrium that exists between the associated state of a TF and a DNA sequence (i.e. the TF is bound to the DNA) and the unassociated or unbound state. It provides a more realistic view of this cellular mechanism. In the current model a DNA binding site is fully accessible to the TF. However, chromatin structure can have a large effect on this accessibility. Other (bound) TFs can also facilitate or hinder TF binding, e.g. by steric hindrance. In the extreme case where two TFs have overlapping binding sites, we basically have two compounds (TFs) competing for the same substrate (DNA), resulting in an equilibrium state, which is obviously different from the equilibrium states of the individual TFs with the binding site.

*Single sample, multiple measurements*

A promising avenue for future work is the "single sample, multiple measurements" approach, where different biological quantities are measured on a single biological sample. In this case, this implies that for different growth conditions not only gene expression levels are measured, but also the binding (or occupancy) of TFs as well as chromatin (and its remodeling proteins). Such an approach excludes confounding (non-biological, external) effects that will arise when integrating data from different labs, different protocols, different yeast strains, etc. Further, instead of inferring the activity of TF binding and chromatin occupancy (as done in Chapter 7), these quantities are measured and can directly be related to the growth conditions and used to model the gene expression levels.
As a side note, expression data (complemented with TF/chromatin data) is not enough to understand the complete route from extracellular quantities to adaptation of transcription rates. This route includes the import and/or signaling of these extracellular quantities via diverse mechanisms that eventually manipulate the activity of TFs and chromatin remodeling proteins, which, in the end, regulate gene expression. In general, the activity of transporters, signaling proteins, enzymes, etc. is not well correlated with the expression of the genes that code for these proteins. Therefore, other types of measurements, such as protein and metabolite measurements need to be integrated in order for computational models to infer the influence of extracellular stimuli on all levels of the cell. Additionally, the nature of cultivation parameters and the way in which the cell reacts to them can be completely different. Consider, for example, the difference between parameters that are imported by the cell and metabolized, such as

carbon source glucose, and parameters that are only sensed by the cell as chemical or physical stimuli, such as temperature. This would necessitate directed and specialized computational approaches on equally specific datasets.

Conclusively, the work described in this thesis provides a starting point to explore future directions that investigate how the cell responds to its environment.



# BIBLIOGRAPHY


[Abbo 07] D. A. Abbott, T. A. Knijnenburg, L. M. I. de Poorter, M. J. T. Reinders, J. T. Pronk, and A. J. A. van Maris. "Generic and specific transcriptional responses to different weak organic acids in anaerobic chemostat cultures of Saccharomyces cerevisiae.". *FEMS Yeast Res*, Vol. 7, No. 6, pp. 819–833, Sep 2007.

[Affy 00] Affymetrix. "GeneChip Expression Analysis Technical Manual P/N 701021 rev 1". 2000.

[Affy 04] Affymetrix. "Technical Note: The New GeneChip IVT Labeling Kit: Optimized Protocol for Improved Results P/N 701466 rev 2". 2004.

[Albe 03] M. Albertsen, I. Bellahn, R. Kramer, and S. Waffenschmidt. "Localization and function of the yeast multidrug transporter Tpo1p.". *J Biol Chem*, Vol. 278, No. 15, pp. 12820–12825, Apr 2003.

[Alon 99] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.". *Proc Natl Acad Sci U S A*, Vol. 96, No. 12, pp. 6745–6750, Jun 1999.

[Arga 07] K. Y. Arga, Z. I. Onsan, B. Kirdar, K. O. Ulgen, and J. Nielsen. "Understanding signaling in yeast: Insights from network analysis.". *Biotechnol Bioeng*, Vol. 97, No. 5, pp. 1246–1258, Aug 2007.

[Ashb 00] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.". *Nat Genet*, Vol. 25, No. 1, pp. 25–29, May 2000.

[Bail 94] T. L. Bailey and C. Elkan. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers.". *Proc Int Conf Intell Syst Mol Biol*, Vol. 2, pp. 28–36, 1994.

[Bala 06] S. Balaji, M. M. Babu, L. M. Iyer, N. M. Luscombe, and L. Aravind. "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast.". *J Mol Biol*, Vol. 360, No. 1, pp. 213–227, Jun 2006.

[Bamm 05] T. Bammler, R. P. Beyer, S. Bhattacharya, G. A. Boorman, A. Boyles, B. U. Bradford, R. E. Bumgarner, P. R. Bushel, K. Chaturvedi, D. Choi, M. L. Cunningham,

S. Deng, H. K. Dressman, R. D. Fannin, F. M. Farin, J. H. Freedman, R. C. Fry, A. Harper, M. C. Humble, P. Hurban, T. J. Kavanagh, W. K. Kaufmann, K. F. Kerr, L. Jing, J. A. Lapidus, M. R. Lasarev, J. Li, Y.-J. Li, E. K. Lobenhofer, X. Lu, R. L. Malek, S. Milton, S. R. Nagalla, J. P. O'malley, V. S. Palmer, P. Pattee, R. S. Paules, C. M. Perou, K. Phillips, L.-X. Qin, Y. Qiu, S. D. Quigley, M. Rodland, I. Rusyn, L. D. Samson, D. A. Schwartz, Y. Shi, J.-L. Shin, S. O. Sieber, S. Slifer, M. C. Speer, P. S. Spencer, D. I. Sproles, J. A. Swenberg, W. A. Suk, R. C. Sullivan, R. Tian, R. W. Tennant, S. A. Todd, C. J. Tucker, B. V. Houten, B. K. Weis, S. Xuan, H. Zarbl, and M. of the Toxicogenomics Research Consortium. "Standardizing global gene expression analysis between laboratories and across platforms.". *Nat Methods*, Vol. 2, No. 5, pp. 351–356, May 2005.

[Bane 02] N. Banerjee and M. Q. Zhang. "Functional genomics as applied to mapping transcription regulatory networks.". *Curr Opin Microbiol*, Vol. 5, No. 3, pp. 313–317, Jun 2002.

[Bar 03] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. "Computational discovery of gene modules and regulatory networks.". *Nat Biotechnol*, Vol. 21, No. 11, pp. 1337–1342, Nov 2003.

[Bar 04] Z. Bar-Joseph. "Analyzing time series gene expression data.". *Bioinformatics*, Vol. 20, No. 16, pp. 2493–2503, Nov 2004.

[Bara 01] Y. Barash, G. Bejerano, and N. Friedman. "A simple hypergeometric approach for discovering putative transcription factor binding sites". *Algorithms in Bioinformatics: Proc. First InternationalWorkshop, number 2149 in LNCS*, Vol. 2149, pp. 278–293, 2001.

[Barr 04] M. H. Barros, A. Johnson, and A. Tzagoloff. "COX23, a homologue of COX17, is required for cytochrome oxidase assembly.". *J Biol Chem*, Vol. 279, No. 30, pp. 31943–31947, Jul 2004.

[Baue 03] A. Bauer and B. Kuster. "Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes.". *Eur J Biochem*, Vol. 270, No. 4, pp. 570–578, Feb 2003.

[Bede 79] G. W. Bedell and D. R. Soll. "Effects of low concentrations of zinc on the growth and dimorphism of Candida albicans: evidence for zinc-resistant and -sensitive pathways for mycelium formation.". *Infect Immun*, Vol. 26, No. 1, pp. 348–354, Oct 1979.

[Beer 04] M. A. Beer and S. Tavazoie. "Predicting gene expression from sequence.". *Cell*, Vol. 117, No. 2, pp. 185–198, Apr 2004.

[Berg 96] M. A. van den Berg, P. de Jong-Gubbels, C. J. Kortland, J. P. van Dijken, J. T. Pronk, and H. Y. Steensma. "The two acetyl-coenzyme A synthetases of Saccharomyces cerevisiae differ with respect to kinetic properties and transcriptional regulation.". *J Biol Chem*, Vol. 271, No. 46, pp. 28953–28959, Nov 1996.

[Berk 04] C. D. Berkey, V. K. Vyas, and M. Carlson. "Nrg1 and nrg2 transcriptional repressors are differently regulated in response to carbon source.". *Eukaryot Cell*, Vol. 3, No. 2, pp. 311–317, Apr 2004.

[Beye 06] A. Beyer, C. Workman, J. Hollunder, D. Radke, U. Mller, T. Wilhelm, and T. Ideker. "Integrated assessment and prediction of transcription factor binding.". *PLoS Comput Biol*, Vol. 2, No. 6, p. e70, Jun 2006.

[Bhus 03] S. Bhushan, B. Lefebvre, A. Sthl, S. J. Wright, B. D. Bruce, M. Boutry, and E. Glaser. "Dual targeting and function of a protease in mitochondria and chloroplasts.". *EMBO Rep*, Vol. 4, No. 11, pp. 1073–1078, Nov 2003.

[Bind 01] H. Binder, K. Arnold, A. S. Ulrich, and O. Zschrnig. "Interaction of Zn2+ with phospholipid membranes.". *Biophys Chem*, Vol. 90, No. 1, pp. 57–74, Mar 2001.

[Bird 06] A. J. Bird, M. Gordon, D. J. Eide, and D. R. Winge. "Repression of ADH1 and ADH3 during zinc deficiency by Zap1-induced intergenic RNA transcripts.". *EMBO J*, Vol. 25, No. 24, pp. 5726–5734, Dec 2006.

[Blai 05] A. Blais and B. D. Dynlacht. "Constructing transcriptional regulatory networks.". *Genes Dev*, Vol. 19, No. 13, pp. 1499–1511, Jul 2005.

[Blai 97] P. L. Blaiseau, A. D. Isnard, Y. Surdin-Kerjan, and D. Thomas. "Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism.". *Mol Cell Biol*, Vol. 17, No. 7, pp. 3640–3648, Jul 1997.

[Blai 98] P. Blaiseau and D. Thomas. "Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA". *EMBO J.*, Vol. 17, No. 21, pp. 6327–6336, 1998.

[Boer 03] V. M. Boer, J. H. de Winde, J. T. Pronk, and M. D. W. Piper. "The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur.". *J Biol Chem*, Vol. 278, No. 5, pp. 3265–3274, Jan 2003.

[Boer 05] V. Boer, J. Daran, M. Almering, J. de Winde, and J. Pronk. "Contribution of the Saccharomyces cerevisiae transcriptional regulator Leu3p to physiology and gene expression in nitrogen- and carbon-limited chemostat cultures". *FEMS Yeast Res.*, Vol. 5, No. 10, pp. 885–897, 2005.

[Boer 07] V. M. Boer, S. L. Tai, Z. Vuralhan, Y. Arifin, M. C. Walsh, M. D. W. Piper, J. H. de Winde, J. T. Pronk, and J.-M. Daran. "Transcriptional responses of Saccharomyces cerevisiae to preferred and nonpreferred nitrogen sources in glucose-limited chemostat cultures.". *FEMS Yeast Res*, Vol. 7, No. 4, pp. 604–620, Jun 2007.

[Bohm 97] S. Bohm, D. Frishman, and H. W. Mewes. "Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins.". *Nucleic Acids Res*, Vol. 25, No. 12, pp. 2464–2469, Jun 1997.

[Bols 03] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.". *Bioinformatics*, Vol. 19, No. 2, pp. 185–193, Jan 2003.

[Bonn 06] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson. "The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.". *Genome Biol*, Vol. 7, No. 5, p. R36, 2006.

[Brau 08] M. J. Brauer, C. Huttenhower, E. M. Airoldi, R. Rosenstein, J. C. Matese, D. Gresham, V. M. Boer, O. G. Troyanskaya, and D. Botstein. "Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast.". *Mol Biol Cell*, Vol. 19, No. 1, pp. 352–367, Jan 2008.

[Brom 97] S. K. Bromberg, P. A. Bower, G. R. Duncombe, J. Fehring, L. Gerber, V. K. Lau, and M. Tata. "Requirements for Zinc, Manganese, Calcium, and Magnesium in Wort". *J. Am. Soc. Brew. Chem.*, Vol. 55, pp. 123–128, 1997.

[Brow 93] J. L. Brown, S. North, and H. Bussey. "SKN7, a yeast multicopy suppressor of a mutation affecting cell wall beta-glucan assembly, encodes a product with domains homologous to prokaryotic two-component regulators and to heat shock transcription factors.". *J Bacteriol*, Vol. 175, No. 21, pp. 6908–6915, Nov 1993.

[Brow 94] J. L. Brown, H. Bussey, and R. C. Stewart. "Yeast Skn7p functions in a eukaryotic two-component regulatory pathway.". *EMBO J*, Vol. 13, No. 21, pp. 5186–5194, Nov 1994.

[Buss 01] H. J. Bussemaker, H. Li, and E. D. Siggia. "Regulatory element detection using correlation with expression.". *Nat Genet*, Vol. 27, No. 2, pp. 167–171, Feb 2001.

[Buss 06] H. J. Bussemaker. "Modeling gene expression control using Omes Law.". *Mol Syst Biol*, Vol. 2, p. 2006.0013, 2006.

[Carm 98] V. Carmelo, R. Santos, C. A. Viegas, and I. S-Correia. "Modification of Saccharomyces cerevisiae thermotolerance following rapid exposure to acid stress.". *Int J Food Microbiol*, Vol. 42, No. 3, pp. 225–230, Jul 1998.

[Cast 07] J. Castrillo, L. Zeef, D. Hoyle, N. Zhang, A. Hayes, D. Gardner, M. Cornell, J. Petty, L. Hakes, L. Wardleworth, B. Rash, M. Brown, W. Dunn, D. Broadhurst, K. O'donoghue, S. Hester, T. Dunkley, S. Hart, N. Swainston, P. Li, S. Gaskell, N. Paton, K. Lilley, D. Kell, and S. Oliver. "Growth control of the eukaryote cell: a systems biology study in yeast.". *J Biol*, Vol. 6, No. 2, p. 4, Apr 2007.

[Chan 06] Y.-H. Chang, Y.-C. Wang, and B.-S. Chen. "Identification of transcription factor cooperativity via stochastic system model.". *Bioinformatics*, Vol. 22, No. 18, pp. 2276–2282, Sep 2006.

[Chen 95] Y. Chen and B.K.Tye. "The yeast Mcm1 protein is regulated posttranscriptionally by the flux of glycolysis". *Mol. Cell. Biol.*, Vol. 15, No. 8, pp. 4631–4639, 1995.

[Cho 98] 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*, Vol. 2, No. 1, pp. 65–73, Jul 1998.

[Chou 06] S. Chou, S. Lane, and H. Liu. "Regulation of mating and filamentation genes by two distinct Ste12 complexes in Saccharomyces cerevisiae.". *Mol Cell Biol*, Vol. 26, No. 13, pp. 4794–4805, Jul 2006.

[Chua 04] G. Chua, M. Robinson, Q. Morris, and T. Hughes. "Transcriptional networks: reverse-engineering gene regulation on a global scale". *Curr. Opin. Microbiol.*, Vol. 7, No. 6, pp. 638–646, 2004.

[Clif 03] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. "Finding functional features in Saccharomyces genomes by phylogenetic footprinting.". *Science*, Vol. 301, No. 5629, pp. 71–76, Jul 2003.

[Cohe 99] A. Cohen, N. Perzov, H. Nelson, and N. Nelson. "A novel family of yeast chaperons involved in the distribution of V-ATPase and other membrane proteins.". *J Biol Chem*, Vol. 274, No. 38, pp. 26885–26893, Sep 1999.

[Curt 05] R. K. Curtis, M. Oresic, and A. Vidal-Puig. "Pathways to the analysis of microarray data.". *Trends Biotechnol*, Vol. 23, No. 8, pp. 429–435, Aug 2005.

[Dara 03] P. Daran-Lapujade, J. M. Daran, P. Ktter, T. Petit, M. D. W. Piper, and J. T. Pronk. "Comparative genotyping of the Saccharomyces cerevisiae laboratory strains S288C and CEN.PK113-7D using oligonucleotide microarrays.". *FEMS Yeast Res*, Vol. 4, No. 3, pp. 259–269, Dec 2003.

[Dara 04] P. Daran-Lapujade, M. L. A. Jansen, J.-M. Daran, W. van Gulik, J. H. de Winde, and J. T. Pronk. "Role of transcriptional regulation in controlling fluxes in central carbon metabolism of Saccharomyces cerevisiae. A chemostat culture study.". *J Biol Chem*, Vol. 279, No. 10, pp. 9125–9138, Mar 2004.

[Dara 09] P. Daran-Lapujade, J.-M. Daran, A. J. A. van Maris, J. H. de Winde, and J. T. Pronk. "Chemostat-based micro-array analysis in baker's yeast.". *Adv Microb Physiol*, Vol. 54, pp. 257–311, 2009.

[Das 04] D. Das, N. Banerjee, and M. Q. Zhang. "Interacting models of cooperative gene regulation.". *Proc Natl Acad Sci U S A*, Vol. 101, No. 46, pp. 16234–16239, Nov 2004.

[Davi 79] D. Davies and D. Bouldin. "A cluster separation measure". *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. PAMI-1, pp. 224–227, 1979.

[DeNi 06] R. DeNicola. *Ph.D. thesis*. PhD thesis, University of Abertay, Dundee, Scotland, 2006.

[DeRi 96] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. "Use of a cDNA microarray to analyse gene expression patterns in human cancer.". *Nat Genet*, Vol. 14, No. 4, pp. 457–460, Dec 1996.

[Deve 00] R. J. Devenish, M. Prescott, X. Roucou, and P. Nagley. "Insights into ATP synthase assembly and function through the molecular genetic manipulation of subunits of the yeast mitochondrial enzyme complex.". *Biochim Biophys Acta*, Vol. 1458, No. 2-3, pp. 428–442, May 2000.

[Dhae 05] P. D'haeseleer. "How does gene expression clustering work?". *Nat Biotechnol*, Vol. 23, No. 12, pp. 1499–1501, Dec 2005.

[Dijk 00] J. van Dijken, Bauer, Brambilla, Duboc, Francois, Gancedo, Giuseppin, Heijnen, Hoare, Lange, Madden, Niederberger, Nielsen, Parrou, Petit, Porro, Reuss, van Riel N, Rizzi, Steensma, Verrips, Vindelov, and Pronk. "An interlaboratory comparison of physiological and genetic properties of four Saccharomyces cerevisiae strains.". *Enzyme Microb Technol*, Vol. 26, No. 9-10, pp. 706–714, Jun 2000.

[Dijk 86] J. P. V. Dijken and W. A. Scheffers. "Redox balances in the metabolism of sugars by yeast". *FEMS Microbiol Rev*, Vol. 32, pp. 199–224, 1986.

[Dufo 03] J. Dufour, P. Malcorps, and P. Silcock. *Control of ester synthesis during brewery fermentation (p. 213-233 of Brewing yeast fermentation performance)*. Blackwell Science, Oxford, United Kingdom, 2003.

[Efro 07] B. Efron and R. Tibshirani. "On testing the significance of sets of genes". *Annals of Applied Statistics*, Vol. 1, pp. 107–129, 2007.

[Elbl 91] R. Elble and B. K. Tye. "Both activation and repression of a-mating-type-specific genes in yeast require transcription factor Mcm1.". *Proc Natl Acad Sci U S A*, Vol. 88, No. 23, pp. 10966–10970, Dec 1991.

[Elli 04] C. D. Ellis, F. Wang, C. W. MacDiarmid, S. Clark, T. Lyons, and D. J. Eide. "Zinc and the Msc2 zinc transporter protein are required for endoplasmic reticulum function.". *J Cell Biol*, Vol. 166, No. 3, pp. 325–335, Aug 2004.

[Elli 05] C. D. Ellis, C. W. Macdiarmid, and D. J. Eide. "Heteromeric protein complexes mediate zinc transport into the secretory pathway of eukaryotic cells.". *J Biol Chem*, Vol. 280, No. 31, pp. 28811–28818, Aug 2005.

[Enja 04] B. Enjalbert, J. L. Parrou, M. A. Teste, and J. Franois. "Combinatorial control by the protein kinases PKA, PHO85 and SNF1 of transcriptional induction of the Saccharomyces cerevisiae GSY2 gene at the diauxic shift.". *Mol Genet Genomics*, Vol. 271, No. 6, pp. 697–708, Jul 2004.

[Erns 05] J. Ernst, G. J. Nau, and Z. Bar-Joseph. "Clustering short time series gene expression data.". *Bioinformatics*, Vol. 21 Suppl 1, pp. i159–i168, Jun 2005.

[Fauc 02] M. Fauchon, G. Lagniel, J. C. Aude, L. Lombardia, P. Soularue, C. Petat, G. Marguerie, A. Sentenac, M. Werner, and J. Labarre. "Sulfur sparing in the yeast proteome in response to sulfur demand.". *Mol Cell*, Vol. 9, No. 4, pp. 713–723, Apr 2002.

[Fere 99] 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*, Vol. 96, No. 17, pp. 9721–9726, Aug 1999.

[Fern 05] A. R. Fernandes, N. P. Mira, R. C. Vargas, I. Canelhas, and I. S-Correia. "Saccharomyces cerevisiae adaptation to weak acids involves the transcription factor Haa1p and Haa1p-regulated genes.". *Biochem Biophys Res Commun*, Vol. 337, No. 1, pp. 95–103, Nov 2005.

[Foat 06] B. C. Foat, A. V. Morozov, and H. J. Bussemaker. "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.". *Bioinformatics*, Vol. 22, No. 14, pp. e141–e149, Jul 2006.

[Fors 89] S. L. Forsburg and L. Guarente. "Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer.". *Genes Dev*, Vol. 3, No. 8, pp. 1166–1178, Aug 1989.

[Fran 01] J. Francois and J. L. Parrou. "Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae.". *FEMS Microbiol Rev*, Vol. 25, No. 1, pp. 125–145, Jan 2001.

[Ganc 01] J. M. Gancedo. "Control of pseudohyphae formation in Saccharomyces cerevisiae.". *FEMS Microbiol Rev*, Vol. 25, No. 1, pp. 107–123, Jan 2001.

[Ganc 98] J. M. Gancedo. "Yeast carbon catabolite repression.". *Microbiol Mol Biol Rev*, Vol. 62, No. 2, pp. 334–361, Jun 1998.

[Gao 04] F. Gao, B. C. Foat, and H. J. Bussemaker. "Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data.". *BMC Bioinformatics*, Vol. 5, p. 31, Mar 2004.

[Gasc 00] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. "Genomic expression programs in the response of yeast cells to environmental changes.". *Mol Biol Cell*, Vol. 11, No. 12, pp. 4241–4257, Dec 2000.

[Gasc 02] A. P. Gasch and M. Werner-Washburne. "The genomics of yeast responses to environmental stress and starvation.". *Funct Integr Genomics*, Vol. 2, No. 4-5, pp. 181–192, Sep 2002.

[Ge 03] Y. Ge, S. Dudoit, and T. P. Speed. "Resampling-based multiple testing for microarray data analysis". *TEST*, Vol. 12, No. 1, pp. 1–77, 2003.

[Geis 00] A. Geissler, T. Krimmer, B. Schnfisch, M. Meijer, and J. Rassow. "Biogenesis of the yeast frataxin homolog Yfh1p. Tim44-dependent transfer to mtHsp70 facilitates folding of newly imported proteins in mitochondria.". *Eur J Biochem*, Vol. 267, No. 11, pp. 3167–3180, Jun 2000.

[Gela 03] R. Gelade, S. V. de Velde, P. V. Dijck, and J. M. Thevelein. "Multi-level response of the yeast genome to glucose.". *Genome Biol*, Vol. 4, No. 11, p. 233, 2003.

[Gene] "GeneData http://www.genedata.com".

[Geno] "Genome Expression Omnibus http://www.ncbi.nlm.nih.gov/projects/geo/".

[Geor 99] E. Georgatsou and D. Alexandraki. "Regulated expression of the Saccharomyces cerevisiae Fre1p/Fre2p Fe/Cu reductase related genes.". *Yeast*, Vol. 15, No. 7, pp. 573–584, May 1999.

[Gime 92] C. J. Gimeno, P. O. Ljungdahl, C. A. Styles, and G. R. Fink. "Unipolar cell divisions in the yeast S. cerevisiae lead to filamentous growth: regulation by starvation and RAS.". *Cell*, Vol. 68, No. 6, pp. 1077–1090, Mar 1992.

[Gita 98] R. S. Gitan, H. Luo, J. Rodgers, M. Broderius, and D. Eide. "Zinc-induced inactivation of the yeast ZRT1 zinc transporter occurs through endocytosis and vacuolar degradation.". *J Biol Chem*, Vol. 273, No. 44, pp. 28617–28624, Oct 1998.

[Gler 95] D. M. Glerum, T. J. Koerner, and A. Tzagoloff. "Cloning and characterization of COX14, whose product is required for assembly of yeast cytochrome oxidase.". *J Biol Chem*, Vol. 270, No. 26, pp. 15585–15590, Jun 1995.

[Goem 07] J. J. Goeman and P. Buhlmann. "Analyzing gene expression data in terms of gene sets: methodological issues.". *Bioinformatics*, Vol. 23, No. 8, pp. 980–987, Apr 2007.

[Golu 96] G. Golub and C. V. Loan. *Matrix Computations*. The John Hopkins University Press, Maryland, U.S.A., 1996.

[Golu 99] 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*, Vol. 286, No. 5439, pp. 531–537, Oct 1999.

[Groo 07] M. J. L. de Groot, P. Daran-Lapujade, B. van Breukelen, T. A. Knijnenburg, E. A. F. de Hulster, M. J. T. Reinders, J. T. Pronk, A. J. R. Heck, and M. Slijper. "Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals post-transcriptional regulation of key cellular processes.". *Microbiology*, Vol. 153, No. Pt 11, pp. 3864–3878, Nov 2007.

[Grot 06] T. Grotkjaer, O. Winther, B. Regenberg, J. Nielsen, and L. K. Hansen. "Robust multi-scale clustering of large DNA microarray datasets with the consensus algorithm.". *Bioinformatics*, Vol. 22, No. 1, pp. 58–67, Jan 2006.

[Gupt 03] A. Gupta and G. Rao. "A study of oxygen transfer in shake flasks using a non-invasive oxygen sensor.". *Biotechnol Bioeng*, Vol. 84, No. 3, pp. 351–358, Nov 2003.

[Harb 04] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. "Transcriptional regulatory code of a eukaryotic genome.". *Nature*, Vol. 431, No. 7004, pp. 99–104, Sep 2004.

[Hatz 03] K. Hatzixanthis, M. Mollapour, I. Seymour, B. E. Bauer, G. Krapf, C. Schller, K. Kuchler, and P. W. Piper. "Moderately lipophilic carboxylate compounds are the selective inducers of the Saccharomyces cerevisiae Pdr12p ATP-binding cassette transporter.". *Yeast*, Vol. 20, No. 7, pp. 575–585, May 2003.

[Haye 88] T. E. Hayes, P. Sengupta, and B. H. Cochran. "The human c-fos serum response factor and the yeast factors GRM/PRTF have related DNA-binding specificities.". *Genes Dev*, Vol. 2, No. 12B, pp. 1713–1722, Dec 1988.

[Haza 04] R. Hazan, A. Levine, and H. Abeliovich. "Benzoic acid, a weak organic acid food preservative, exerts specific effects on intracellular membrane trafficking pathways in Saccharomyces cerevisiae.". *Appl Environ Microbiol*, Vol. 70, No. 8, pp. 4449–4457, Aug 2004.

[Held 03] J. van Helden. "Regulatory sequence analysis tools.". *Nucleic Acids Res*, Vol. 31, No. 13, pp. 3593–3596, Jul 2003.

[Hert 99] G. Z. Hertz and G. D. Stormo. "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.". *Bioinformatics*, Vol. 15, No. 7-8, pp. 563–577, 1999.

[Higg 03] V. J. Higgins, P. J. Rogers, and I. W. Dawes. "Application of genome-wide expression analysis to identify molecular markers useful in monitoring industrial fermentations.". *Appl Environ Microbiol*, Vol. 69, No. 12, pp. 7535–7540, Dec 2003.

[Hodg 90] J. Hodgson and M. Moir. "Control of esters in brewing". *Proceedings of the 3rd Aviemore Conference on Malting, Brewing and Distilling*, pp. 266–269, 1990.

[Hofm 99] J. Hofman-Bang. "Nitrogen catabolite repression in Saccharomyces cerevisiae.". *Mol Biotechnol*, Vol. 12, No. 1, pp. 35–73, Aug 1999.

[Holy 96] C. D. Holyoak, M. Stratford, Z. McMullin, M. B. Cole, K. Crimmins, A. J. Brown, and P. J. Coote. "Activity of the plasma membrane H(+)-ATPase and optimal glycolytic flux are required for rapid adaptation and growth of Saccharomyces cerevisiae in the presence of the weak-acid preservative sorbic acid.". *Appl Environ Microbiol*, Vol. 62, No. 9, pp. 3158–3164, Sep 1996.

[Holy 99] C. D. Holyoak, D. Bracey, P. W. Piper, K. Kuchler, and P. J. Coote. "The Saccharomyces cerevisiae weak-acid-inducible ABC transporter Pdr12 transports fluorescein and preservative anions from the cytosol by an energy-dependent mechanism.". *J Bacteriol*, Vol. 181, No. 15, pp. 4644–4652, Aug 1999.

[Hora 02] C. E. Horak, N. M. Luscombe, J. Qian, P. Bertone, S. Piccirrillo, M. Gerstein, and M. Snyder. "Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae.". *Genes Dev*, Vol. 16, No. 23, pp. 3017–3033, Dec 2002.

[Hosk 05] P. A. Hoskisson and G. Hobbs. "Continuous culture–making a comeback?". *Microbiology*, Vol. 151, No. Pt 10, pp. 3153–3159, Oct 2005.

[Houg 82] J. Hough, D. E. Briggs, R. Stevens, and T. M. Young. *Malting and brewing science*. Chapman and Hall, London, United Kingdom, 1982.

[Huan 05] H.-R. Huang, C. E. Rowe, S. Mohr, Y. Jiang, A. M. Lambowitz, and P. S. Perlman. "The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function.". *Proc Natl Acad Sci U S A*, Vol. 102, No. 1, pp. 163–168, Jan 2005.

[Hugh 00] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. "Functional discovery via a compendium of expression profiles.". *Cell*, Vol. 102, No. 1, pp. 109–126, Jul 2000.

[Imai 95] T. Imai and T. Ohno. "The relationship between viability and intracellular pH in the yeast Saccharomyces cerevisiae.". *Appl Environ Microbiol*, Vol. 61, No. 10, pp. 3604–3608, Oct 1995.

[Imaz 05] H. Imazu and H. Sakurai. "Saccharomyces cerevisiae heat shock transcription factor regulates cell wall remodeling in response to heat shock.". *Eukaryot Cell*, Vol. 4, No. 6, pp. 1050–1056, Jun 2005.

[Iriz 03] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data.". *Biostatistics*, Vol. 4, No. 2, pp. 249–264, Apr 2003.

[Jaco 79] T. Jacobsen, R. Volden, S. Engan, and O. Aubert. "A chemometric study of some beer flavour components". *J. Inst. Brew.*, Vol. 89, pp. 265–270, 1979.

[Jami 98] D. Jamieson. "Oxidative stress responses of the yeast Saccharomyces cerevisiae". *Yeast*, Vol. 14, No. 16, pp. 1511–1527, 1998.

[Jans 04] M. L. A. Jansen, P. Daran-Lapujade, J. H. de Winde, M. D. W. Piper, and J. T. Pronk. "Prolonged maltose-limited cultivation of Saccharomyces cerevisiae selects for cells with improved maltose affinity and hypersensitivity.". *Appl Environ Microbiol*, Vol. 70, No. 4, pp. 1956–1963, Apr 2004.

[Jone 84] R. Jones and P. F. Greenfield. "A review of yeast ionic nutrition". *Process Biochem.*, Vol. 19, pp. 48–59, 1984.

[Ju 90] Q. D. Ju, B. E. Morrow, and J. R. Warner. "REB1, a yeast DNA-binding protein with many targets, is essential for growth and bears some resemblance to the oncogene myb.". *Mol Cell Biol*, Vol. 10, No. 10, pp. 5226–5234, Oct 1990.

[Kami 89] A. Kamizono, M. Nishizawa, Y. Teranishi, K. Murata, and A. Kimura. "Identification of a gene conferring resistance to zinc and cadmium ions in the yeast Saccharomyces cerevisiae.". *Mol Gen Genet*, Vol. 219, No. 1-2, pp. 161–167, Oct 1989.

[Kane 00] M. Kanehisa and S. Goto. "KEGG: kyoto encyclopedia of genes and genomes.". *Nucleic Acids Res*, Vol. 28, No. 1, pp. 27–30, Jan 2000.

[Kapt 01] J. C. Kapteyn, B. ter Riet, E. Vink, S. Blad, H. D. Nobel, H. V. D. Ende, and F. M. Klis. "Low external pH induces HOG1-dependent changes in the organization of the Saccharomyces cerevisiae cell wall.". *Mol Microbiol*, Vol. 39, No. 2, pp. 469–479, Jan 2001.

[Kell 03] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. "Sequencing and comparison of yeast species to identify genes and regulatory elements.". *Nature*, Vol. 423, No. 6937, pp. 241–254, May 2003.

[Kerr 00] M. K. Kerr, M. Martin, and G. A. Churchill. "Analysis of variance for gene expression microarray data.". *J Comput Biol*, Vol. 7, No. 6, pp. 819–837, 2000.

[Khat 05] P. Khatri and S. Draghici. "Ontological analysis of gene expression data: current tools, limitations, and open problems.". *Bioinformatics*, Vol. 21, No. 18, pp. 3587–3595, Sep 2005.

[Kim 05] S.-Y. Kim and D. J. Volsky. "PAGE: parametric analysis of gene set enrichment.". *BMC Bioinformatics*, Vol. 6, p. 144, 2005.

[Kita 02] H. Kitano. "Systems biology: a brief overview.". *Science*, Vol. 295, No. 5560, pp. 1662–1664, Mar 2002.

[Knij 05] T. Knijnenburg, J. Daran, P. Daran-Lapujade, M. Reinders, and L. Wessels. "Relating transcription factors, modules of genes and cultivation conditions in Saccharomyces cerevisiae". *IEEE CSBW'05*, pp. 71–72, 2005.

[Knij 07] T. A. Knijnenburg, J. H. de Winde, J.-M. Daran, P. Daran-Lapujade, J. T. Pronk, M. J. T. Reinders, and L. F. A. Wessels. "Exploiting combinatorial cultivation conditions to infer transcriptional regulation.". *BMC Genomics*, Vol. 8, No. 25, 2007.

[Knij 08] T. A. Knijnenburg, L. F. A. Wessels, and M. J. T. Reinders. "Combinatorial influence of environmental parameters on transcription factor activity.". *Bioinformatics*, Vol. 24, No. 13, pp. i172–i181, Jul 2008.

[Knij 09] T. Knijnenburg, J.-M. Daran, M. van den Broek, P. Daran-Lapujade, J. de Winde, J. Pronk, M. Reinders, and L. Wessels. "Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: A quantitative analysis of a compendium of chemostat-based transcriptome data.". *BMC Genomics*, Vol. 10, No. 53, Jan 2009.

[Koer 02] M. G. Koerkamp, M. Rep, H. J. Bussemaker, G. P. M. A. Hardy, A. Mul, K. Piekarska, C. A.-K. Szigyarto, J. M. T. D. Mattos, and H. F. Tabak. "Dissection of transient oxidative stress response in Saccharomyces cerevisiae by using DNA microarrays.". *Mol Biol Cell*, Vol. 13, No. 8, pp. 2783–2794, Aug 2002.

[Kohl 03] G. Kohlhaw. "Leucine biosynthesis in fungi: entering metabolism through the back door". *Microbiol. Mol. Biol. Rev.*, Vol. 67, No. 1, pp. 1–15, 2003.

[Kolk 06] A. Kolkman, P. Daran-Lapujade, A. Fullaondo, M. M. A. Olsthoorn, J. T. Pronk, M. Slijper, and A. J. R. Heck. "Proteome analysis of yeast response to various nutrient limitations.". *Mol Syst Biol*, Vol. 2, p. 2006.0026, 2006.

[Kreb 83] H. A. Krebs, D. Wiggins, M. Stubbs, A. Sols, and F. Bedoya. "Studies on the mechanism of the antifungal action of benzoate.". *Biochem J*, Vol. 214, No. 3, pp. 657–663, Sep 1983.

[Kred 99] G. Kreder. "Yeast assimilation of trub-bound zinc". *J. Am. Soc. Brew. Chem.*, Vol. 57, pp. 129–132, 1999.

[Kren 03] A. Kren, Y. M. Mamnun, B. E. Bauer, C. Schller, H. Wolfger, K. Hatzixanthis, M. Mollapour, C. Gregori, P. Piper, and K. Kuchler. "War1p, a novel transcription factor controlling weak acid stress response in yeast.". *Mol Cell Biol*, Vol. 23, No. 5, pp. 1775–1785, Mar 2003.

[Kres 06] M. T. A. P. Kresnowati, W. A. van Winden, M. J. H. Almering, A. ten Pierick, C. Ras, T. A. Knijnenburg, P. Daran-Lapujade, J. T. Pronk, J. J. Heijnen, and J. M. Daran. "When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation.". *Mol Syst Biol*, Vol. 2, p. 49, 2006.

[Kuhb 06] F. Kuhbeck, W. Back, and M. Krottenthaler. "Influence of lauter turbidity on wort composition, fermentation performance and beer qualitya review". *J. Inst. Brew.*, Vol. 112, pp. 215–221, 2006.

[Kuma 06] A. Kumanovics, K. E. Poruk, K. A. Osborn, D. M. Ward, and J. Kaplan. "YKE4 (YIL023C) encodes a bidirectional zinc transporter in the endoplasmic reticulum of Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 281, No. 32, pp. 22566–22574, Aug 2006.

[Kwas 02] K. E. Kwast, L.-C. Lai, N. Menda, D. T. James, S. Aref, and P. V. Burke. "Genomic analyses of anaerobically induced genes in Saccharomyces cerevisiae: functional roles of Rox1 and other factors in mediating the anoxic response.". *J Bacteriol*, Vol. 184, No. 1, pp. 250–265, Jan 2002.

[Lago 03] A. Lagorce, N. C. Hauser, D. Labourdette, C. Rodriguez, H. Martin-Yken, J. Arroyo, J. D. Hoheisel, and J. Franois. "Genome-wide analysis of the response to cell wall mutations in the yeast Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 278, No. 22, pp. 20345–20357, May 2003.

[Lamb 03] T. M. Lamb and A. P. Mitchell. "The transcription factor Rim101p governs ion tolerance and cell differentiation by direct repression of the regulatory genes NRG1 and SMP1 in Saccharomyces cerevisiae.". *Mol Cell Biol*, Vol. 23, No. 2, pp. 677–686, Jan 2003.

[Lee 02] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. "Transcriptional regulatory networks in Saccharomyces cerevisiae.". *Science*, Vol. 298, No. 5594, pp. 799–804, Oct 2002.

[Lee 95] S. Lee, K. Villa, and H. Patino. "Yeast strain development for enhanced production of desirable alcohols/esters in beer". *J. Am. Soc. Brew. Chem.*, Vol. 53, pp. 153–156, 1995.

[Lee 99] J. Lee, C. Godon, G. Lagniel, D. Spector, J. Garin, J. Labarre, and M. B. Toledano. "Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast.". *J Biol Chem*, Vol. 274, No. 23, pp. 16040–16046, Jun 1999.

[Leed 00] Y. Leedan and P. Meer. "Heteroscedastic Regression in Computer Vision: Problems with Bilinear Constraint". *Int J of Computer Vision*, Vol. 37, No. 2, pp. 127–150, 2000.

[Levi 05] D. E. Levin. "Cell wall integrity signaling in Saccharomyces cerevisiae.". *Microbiol Mol Biol Rev*, Vol. 69, No. 2, pp. 262–291, Jun 2005.

[Li 01] L. Li and J. Kaplan. "The yeast gene MSC2, a member of the cation diffusion facilitator family, affects the cellular distribution of zinc.". *J Biol Chem*, Vol. 276, No. 7, pp. 5036–5043, Feb 2001.

[Lo 98] W. S. Lo and A. M. Dranginis. "The cell surface flocculin Flo11 is required for pseudohyphae formation and invasion by Saccharomyces cerevisiae.". *Mol Biol Cell*, Vol. 9, No. 1, pp. 161–171, Jan 1998.

[Lowr 88] C. V. Lowry and R. S. Zitomer. "ROX1 encodes a heme-induced repression factor regulating ANB1 and CYC7 of Saccharomyces cerevisiae.". *Mol Cell Biol*, Vol. 8, No. 11, pp. 4651–4658, Nov 1988.

[Lu 05] H. Lu and J. Woodburn. "Zinc binding stabilizes mitochondrial Tim10 in a reduced and import-competent state kinetically.". *J Mol Biol*, Vol. 353, No. 4, pp. 897–910, Nov 2005.

[Ludo 01] P. Ludovico, M. J. Sousa, M. T. Silva, C. Leo, and M. Crte-Real. "Saccharomyces cerevisiae commits to a programmed cell death process in response to acetic acid.". *Microbiology*, Vol. 147, No. Pt 9, pp. 2409–2415, Sep 2001.

[Luik 98] S. Luikenhuis, G. Perrone, I. W. Dawes, and C. M. Grant. "The yeast Saccharomyces cerevisiae contains two glutaredoxin genes that are required for protection against reactive oxygen species.". *Mol Biol Cell*, Vol. 9, No. 5, pp. 1081–1091, May 1998.

[Lusc 04] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein. "Genomic analysis of regulatory network dynamics reveals large topological changes.". *Nature*, Vol. 431, No. 7006, pp. 308–312, Sep 2004.

[Luss 97] M. Lussier, A. M. Sdicu, E. Winnett, D. H. Vo, J. Sheraton, A. Dsterhft, R. K. Storms, and H. Bussey. "Completion of the Saccharomyces cerevisiae genome sequence allows identification of KTR5, KTR6 and KTR7 and definition of the nine-membered KRE2/MNT1 mannosyltransferase gene family in this organism.". *Yeast*, Vol. 13, No. 3, pp. 267–274, Mar 1997.

[Lyon 00] T. J. Lyons, A. P. Gasch, L. A. Gaither, D. Botstein, P. O. Brown, and D. J. Eide. "Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast.". *Proc Natl Acad Sci U S A*, Vol. 97, No. 14, pp. 7957–7962, Jul 2000.

[MacD 00] C. W. MacDiarmid, L. A. Gaither, and D. Eide. "Zinc transporters that regulate vacuolar zinc storage in Saccharomyces cerevisiae.". *EMBO J*, Vol. 19, No. 12, pp. 2845–2855, Jun 2000.

[MacD 02] C. W. MacDiarmid, M. A. Milanick, and D. J. Eide. "Biochemical properties of vacuolar zinc transport systems of Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 277, No. 42, pp. 39187–39194, Oct 2002.

[MacI 06] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. "An improved map of conserved regulatory sites for Saccharomyces cerevisiae.". *BMC Bioinformatics*, Vol. 7, p. 113, 2006.

[Maga 02] B. Magasanik and C. A. Kaiser. "Nitrogen regulation in Saccharomyces cerevisiae". *Gene*, Vol. 290, pp. 1–18, 2002.

[Mago 92] E. Magonet, P. Hayen, D. Delforge, E. Delaive, and J. Remacle. "Importance of the structural zinc atom for the stability of yeast alcohol dehydrogenase.". *Biochem J*, Vol. 287 ( Pt 2), pp. 361–365, Oct 1992.

[Mari 04] R. M. Marion, A. Regev, E. Segal, Y. Barash, D. Koller, N. Friedman, and E. K. O'Shea. "Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression.". *Proc Natl Acad Sci U S A*, Vol. 101, No. 40, pp. 14315–14322, Oct 2004.

[Mart 01] A. M. Martins, C. A. Cordeiro, and A. M. P. Freire. "In situ analysis of methylglyoxal metabolism in Saccharomyces cerevisiae.". *FEBS Lett*, Vol. 499, No. 1-2, pp. 41–44, Jun 2001.

[Mart 03] V. Martin, D. E. Quain, and K. A. Smart. *Brewing yeast oxidative stress responses: impact of brewery handling (p. 61-74 Brewing yeast fermentation performance).* Blackwell Science, Oxford, United Kingdom, 2003.

[Mash 03] M. R. Mashego, W. M. van Gulik, J. L. Vinke, and J. J. Heijnen. "Critical evaluation of sampling techniques for residual glucose determination in carbon-limited chemostat culture of Saccharomyces cerevisiae.". *Biotechnol Bioeng*, Vol. 83, No. 4, pp. 395–399, Aug 2003.

[Mats 05] T. Matsuoka, J. Wada, I. Hashimoto, Y. Zhang, J. Eguchi, N. Ogawa, K. Shikata, Y. S. Kanwar, and H. Makino. "Gene delivery of Tim44 reduces mitochondrial superoxide production and ameliorates neointimal proliferation of injured carotid artery in diabetic rats.". *Diabetes*, Vol. 54, No. 10, pp. 2882–2890, Oct 2005.

[Mats 07] S. Matsui, M. Ito, H. Nishiyama, H. Uno, H. Kotani, J. Watanabe, P. Guilford, A. Reeve, M. Fukushima, and O. Ogawa. "Genomic characterization of multiple clinical phenotypes of cancer using multivariate linear regression models.". *Bioinformatics*, Vol. 23, No. 6, pp. 732–738, Mar 2007.

[McDa 65] L. E. McDaniel, E. G. Bailey, and A. Zimmeli. "Effect of oxygen-supply rates on growth of escherichia coli. ii. comparison of results in shake flasks and 50-liter fermentor.". *Appl Microbiol*, Vol. 13, pp. 115–119, Jan 1965.

[Mend 05] D. Mendoza-Cozatl, H. Loza-Tavera, A. Hernandez-Navarro, and R. Moreno-Sanchez. "Sulfur assimilation and glutathione metabolism under cadmium stress in yeast, protists and plants". *FEMS Microbiol. Rev.*, Vol. 29, No. 4, pp. 653–671, 2005.

[Mewe 97] H. W. Mewes, K. Albermann, K. Heumann, S. Liebl, and F. Pfeiffer. "MIPS: a database for protein sequences, homology data and yeast genome information.". *Nucleic Acids Res*, Vol. 25, No. 1, pp. 28–30, Jan 1997.

[Miya 00] S. Miyabe, S. Izawa, and Y. Inoue. "Expression of ZRC1 coding for suppressor of zinc toxicity is induced by zinc-starvation stress in Zap1-dependent fashion in Saccharomyces cerevisiae.". *Biochem Biophys Res Commun*, Vol. 276, No. 3, pp. 879–884, Oct 2000.

[Miya 01] S. Miyabe, S. Izawa, and Y. Inoue. "The Zrc1 is involved in zinc transport system between vacuole and cytosol in Saccharomyces cerevisiae.". *Biochem Biophys Res Commun*, Vol. 282, No. 1, pp. 79–83, Mar 2001.

[Moch 96] F. Mochaba and E. O'Connor-Cox. "Metal ion concentration and release by a brewing yeast: characterization and implications". *J. Am. Soc. Brew. Chem.*, Vol. 54, pp. 155–163, 1996.

[Moeh 91] C. M. Moehle and A. G. Hinnebusch. "Association of RAP1 binding sites with stringent control of ribosomal protein gene transcription in Saccharomyces cerevisiae.". *Mol Cell Biol*, Vol. 11, No. 5, pp. 2723–2735, May 1991.

[Moli 04] M. M. Molina, G. Belli, M. A. de la Torre, M. T. Rodriguez-Manzaneque, and E. Herrero. "Nuclear monothiol glutaredoxins of Saccharomyces cerevisiae can function as mitochondrial glutaredoxins.". *J Biol Chem*, Vol. 279, No. 50, pp. 51923–51930, Dec 2004.

[Moll 04] M. Mollapour, D. Fong, K. Balakrishnan, N. Harris, S. Thompson, C. Schller, K. Kuchler, and P. W. Piper. "Screening the yeast deletant mutant collection for hypersensitivity and hyper-resistance to sorbate, a weak organic acid food preservative.". *Yeast*, Vol. 21, No. 11, pp. 927–946, Aug 2004.

[Moot 03] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.". *Nat Genet*, Vol. 34, No. 3, pp. 267–273, Jul 2003.

[Mora 00] P. Moradas-Ferreira and V. Costa. "Adaptive response of the yeast Saccharomyces cerevisiae to reactive oxygen species: defences, damage and death.". *Redox Rep*, Vol. 5, No. 5, pp. 277–285, 2000.

[Muta 97] T. Muta, D. Kang, S. Kitajima, T. Fujiwara, and N. Hamasaki. "p32 protein, a splicing factor 2-associated protein, is localized in mitochondrial matrix and is functionally important in maintaining oxidative phosphorylation.". *J Biol Chem*, Vol. 272, No. 39, pp. 24363–24370, Sep 1997.

[Myer 07] C. L. Myers and O. G. Troyanskaya. "Context-sensitive data integration and prediction of biological networks.". *Bioinformatics*, Vol. 23, No. 17, pp. 2322–2330, Sep 2007.

[Nare 01] N. V. Narendranath, K. C. Thomas, and W. M. Ingledew. "Effects of acetic acid and lactic acid on the growth of Saccharomyces cerevisiae in a minimal medium.". *J Ind Microbiol Biotechnol*, Vol. 26, No. 3, pp. 171–177, Mar 2001.

[Narl 07] L. Narlikar, R. Gordn, and A. J. Hartemink. "A nucleosome-guided map of transcription factor binding sites in yeast.". *PLoS Comput Biol*, Vol. 3, No. 11, p. e215, Nov 2007.

[Newc 03] L. Newcomb, J. Diderich, M. Slattery, and W. Heideman. "Glucose regulation of Saccharomyces cerevisiae cell cycle genes". *Eukaryot. Cell.*, Vol. 2, No. 1, pp. 143–149, 2003.

[Newt 06] M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist. "Random-set methods identify aspects of the enrichment signal in gene-set analysis - Technical Report 1130". Tech. Rep., UW Madison Statistics Department, 2006.

[Nguy 06] D. H. Nguyen and P. D'haeseleer. "Deciphering principles of transcription regulation in eukaryotic genomes.". *Mol Syst Biol*, Vol. 2, p. 2006.0012, 2006.

[Nico 07] R. D. Nicola, L. A. Hazelwood, E. A. F. D. Hulster, M. C. Walsh, T. A. Knijnenburg, M. J. T. Reinders, G. M. Walker, J. T. Pronk, J.-M. Daran, and P. Daran-Lapujade. "Physiological and transcriptional responses of Saccharomyces cerevisiae to zinc limitation in chemostat cultures.". *Appl Environ Microbiol*, Vol. 73, No. 23, pp. 7680–7692, Dec 2007.

[Nobe 01] H. de Nobel, L. Lawrie, S. Brul, F. Klis, M. Davis, H. Alloush, and P. Coote. "Parallel and comparative analysis of the proteome and transcriptome of sorbic acid-stressed Saccharomyces cerevisiae.". *Yeast*, Vol. 18, No. 15, pp. 1413–1428, Nov 2001.

[Nobe 90] J. G. de Nobel, F. M. Klis, J. Priem, T. Munnik, and H. van den Ende. "The glucanase-soluble mannoproteins limit cell wall porosity in Saccharomyces cerevisiae.". *Yeast*, Vol. 6, No. 6, pp. 491–499, 1990.

[Nobe 91] J. G. D. Nobel and J. A. Barnett. "Passage of molecules through yeast cell walls: a brief essay-review.". *Yeast*, Vol. 7, No. 4, pp. 313–323, 1991.

[Norm 99] T. C. Norman, D. L. Smith, P. K. Sorger, B. L. Drees, S. M. O'Rourke, T. R. Hughes, C. J. Roberts, S. H. Friend, S. Fields, and A. W. Murray. "Genetic selection of peptide inhibitors of biological pathways.". *Science*, Vol. 285, No. 5427, pp. 591–595, Jul 1999.

[OCon 92] K. F. O'Connell and R. E. Baker. "Possible cross-regulation of phosphate and sulfate metabolism in Saccharomyces cerevisiae.". *Genetics*, Vol. 132, No. 1, pp. 63–73, Sep 1992.

[OCon 95] K. F. O'Connell, Y. Surdin-Kerjan, and R. E. Baker. "Role of the Saccharomyces cerevisiae general regulatory factor CP1 in methionine biosynthetic gene transcription.". *Mol Cell Biol*, Vol. 15, No. 4, pp. 1879–1888, Apr 1995.

[Odom 07] D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. "Tissue-specific transcriptional regulation has diverged significantly between human and mouse.". *Nat Genet*, Vol. 39, No. 6, pp. 730–732, Jun 2007.

[Ogaw 00] N. Ogawa, J. DeRisi, and P. O. Brown. "New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis.". *Mol Biol Cell*, Vol. 11, No. 12, pp. 4309–4321, Dec 2000.

[Pahl 01] A. K. Pahlman, K. Granath, R. Ansell, S. Hohmann, and L. Adler. "The yeast glycerol 3-phosphatases Gpp1p and Gpp2p are required for glycerol biosynthesis and differentially involved in the cellular responses to osmotic, anaerobic, and oxidative stress.". *J Biol Chem*, Vol. 276, No. 5, pp. 3555–3563, Feb 2001.

[Palm 95] R. D. Palmiter and S. D. Findley. "Cloning and functional characterization of a mammalian zinc transporter that confers resistance to zinc.". *EMBO J*, Vol. 14, No. 4, pp. 639–649, Feb 1995.

[Pamp 89] M. E. Pampulha and M. C. Loureiro-Dias. "Combined effect of acetic acid, pH and ethanol on intracellular pH of fermenting yeast". *Applied Microbiology and Biotechnology*, Vol. 31, pp. 547–550, 1989.

[Pamp 90] M. E. Pampulha and M. C. Loureiro-Dias. "Activity of glycolytic enzymes of Saccharomyces cerevisiae in the presence of acetic acid". *Applied Microbiology and Biotechnology*, Vol. 34, pp. 375–380, 1990.

[Pass 89] S. Passmore, R. Elbe, and B.K.Tye. "A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes". *Genes Dev.*, Vol. 3, pp. 921–935, 1989.

[Pena 07] L. Pena-Castillo and T. R. Hughes. "Why are there still over 1000 uncharacterized yeast genes?". *Genetics*, Vol. 176, No. 1, pp. 7–14, May 2007.

[Pepp 05] J. van de Peppel, N. Kettelarij, H. van Bakel, T. T. J. P. Kockelkorn, D. van Leenen, and F. C. P. Holstege. "Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets.". *Mol Cell*, Vol. 19, No. 4, pp. 511–522, Aug 2005.

[Pilp 01] Y. Pilpel, P. Sudarsanam, and G. Church. "Identifying regulatory networks by combinatorial analysis of promoter elements". *Nat. Genet.*, Vol. 29, No. 2, pp. 153–159, 2001.

[Pipe 02] M. D. W. Piper, P. Daran-Lapujade, C. Bro, B. Regenberg, S. Knudsen, J. Nielsen, and J. T. Pronk. "Reproducibility of oligonucleotide microarray transcriptome analyses. An interlaboratory comparison using chemostat cultures of Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 277, No. 40, pp. 37001–37008, Oct 2002.

[Pipe 97] P. W. Piper, C. Ortiz-Calderon, C. Holyoak, P. Coote, and M. Cole. "Hsp30, the integral plasma membrane heat shock protein of Saccharomyces cerevisiae, is a stress-inducible regulator of plasma membrane H(+)-ATPase.". *Cell Stress Chaperones*, Vol. 2, No. 1, pp. 12–24, Mar 1997.

[Pipe 98] P. Piper, Y. Mah, S. Thompson, R. Pandjaitan, C. Holyoak, R. Egner, M. Mhlbauer, P. Coote, and K. Kuchler. "The pdr12 ABC transporter is required for the development of weak organic acid resistance in yeast.". *EMBO J*, Vol. 17, No. 15, pp. 4257–4265, Aug 1998.

[Pipe 99] P. W. Piper. "Yeast superoxide dismutase mutants reveal a pro-oxidant action of weak organic acid food preservatives.". *Free Radic Biol Med*, Vol. 27, No. 11-12, pp. 1219–1227, Dec 1999.

[Pitt 04] J. Pittman, E. Huang, H. Dressman, C.-F. Horng, S. H. Cheng, M.-H. Tsou, C.-M. Chen, A. Bild, E. S. Iversen, A. T. Huang, J. R. Nevins, and M. West. "Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes.". *Proc Natl Acad Sci U S A*, Vol. 101, No. 22, pp. 8431–8436, Jun 2004.

[Pokh 05] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young. "Genome-wide map of nucleosome acetylation and methylation in yeast.". *Cell*, Vol. 122, No. 4, pp. 517–527, Aug 2005.

[Post 89] E. Postma, A. Kuiper, W. F. Tomasouw, W. A. Scheffers, and J. P. van Dijken. "Competition for glucose between the yeasts Saccharomyces cerevisiae and Candida utilis.". *Appl Environ Microbiol*, Vol. 55, No. 12, pp. 3214–3220, Dec 1989.

[Qi 05] M. Qi and E. A. Elion. "MAP kinase pathways.". *J Cell Sci*, Vol. 118, No. Pt 16, pp. 3569–3572, Aug 2005.

[Rega 06] L. Regalla and T. J. Lyons. *Zinc in yeast: mechanisms involved in homeostasis (p 37-54 Molecular biology of heavy metal homeostasis and detoxification: from microbes to man)*. Springer, Berlin, Germany, 2006.

[Rege 06] B. Regenberg, T. Grotkjaer, O. Winther, A. Fausbll, M. Akesson, C. Bro, L. K. Hansen, S. Brunak, and J. Nielsen. "Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in Saccharomyces cerevisiae.". *Genome Biol*, Vol. 7, No. 11, p. R107, 2006.

[Regn 01] M. Regnacq, P. Alimardani, B. E. Moudni, and T. Bergs. "SUT1p interaction with Cyc8p(Ssn6p) relieves hypoxic genes from Cyc8p-Tup1p repression in Saccharomyces cerevisiae.". *Mol Microbiol*, Vol. 40, No. 5, pp. 1085–1096, Jun 2001.

[Rep 01] M. Rep, M. Proft, F. Remize, M. Tams, R. Serrano, J. M. Thevelein, and S. Hohmann. "The Saccharomyces cerevisiae Sko1p transcription factor mediates HOG pathway-dependent osmotic regulation of a set of genes encoding enzymes implicated in protection from oxidative damage.". *Mol Microbiol*, Vol. 40, No. 5, pp. 1067–1083, Jun 2001.

[Rep 96] M. Rep and L. A. Grivell. "MBA1 encodes a mitochondrial membrane-associated protein required for biogenesis of the respiratory chain.". *FEBS Lett*, Vol. 388, No. 2-3, pp. 185–188, Jun 1996.

[Roe 02] A. J. Roe, C. O'Byrne, D. McLaggan, and I. R. Booth. "Inhibition of Escherichia coli growth by acetic acid: a problem with methionine biosynthesis and homocysteine toxicity.". *Microbiology*, Vol. 148, No. Pt 7, pp. 2215–2222, Jul 2002.

[Rone 06] M. Ronen and D. Botstein. "Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source.". *Proc Natl Acad Sci U S A*, Vol. 103, No. 2, pp. 389–394, Jan 2006.

[Ross 06] S. Rossell, C. C. van der Weijden, A. Lindenbergh, A. van Tuijl, C. Francke, B. M. Bakker, and H. V. Westerhoff. "Unraveling the complexity of flux regulation: a new method demonstrated for nutrient starvation in Saccharomyces cerevisiae.". *Proc Natl Acad Sci U S A*, Vol. 103, No. 7, pp. 2166–2171, Feb 2006.

[Roth 98] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.". *Nat Biotechnol*, Vol. 16, No. 10, pp. 939–945, Oct 1998.

[Roui 00] A. Rouillon, R. Barbey, E. E. Patton, M. Tyers, and D. Thomas. "Feedback-regulated degradation of the transcriptional activator Met4 is triggered by the SCF(Met30 )complex.". *EMBO J*, Vol. 19, No. 2, pp. 282–294, Jan 2000.

[Ruep 04] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Gldener, G. Mannhaupt, M. Mnsterktter, and H. W. Mewes. "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.". *Nucleic Acids Res*, Vol. 32, No. 18, pp. 5539–5545, 2004.

[Russ 91] A. D. Russell. "Mechanisms of bacterial resistance to non-antibiotics: food additives and food and pharmaceutical preservatives.". *J Appl Bacteriol*, Vol. 71, No. 3, pp. 191–201, Sep 1991.

[Russ 92] J. Russell. "Another explanation for the toxicity of fermentation acids at low pH: anion accumulation versus uncoupling". *Journal of applied microbiology*, Vol. 73, pp. 363–370, 1992.

[Salm 84] C. V. Salmond, R. G. Kroll, and I. R. Booth. "The effect of food preservatives on pH homeostasis in Escherichia coli.". *J Gen Microbiol*, Vol. 130, No. 11, pp. 2845–2850, Nov 1984.

[Sava 02] T. Savard, C. Beaulieu, N. Gardner, and C. Champagne. "Characterization of spoilage yeasts isolated from fermented vegetables and Inhibition by lactic, acetic and propionic acids". *FOOD MICROBIOLOGY*, Vol. 19, pp. 363–373, 2002.

[Schu 04] C. Schueller, Y. M. Mamnun, M. Mollapour, G. Krapf, M. Schuster, B. E. Bauer, P. W. Piper, and K. Kuchler. "Global phenotypic analysis and transcriptional profiling defines the weak acid stress response regulon in Saccharomyces cerevisiae.". *Mol Biol Cell*, Vol. 15, No. 2, pp. 706–720, Feb 2004.

[Schu 64] J. S. Schultz. "Cotton closure as an aeration barrier in shaken flask fermentations.". *Appl Microbiol*, Vol. 12, pp. 305–310, Jul 1964.

[SCPD] "SCPD The Promoter Database of Saccharomyces cerevisiae http://rulai.cshl.edu/SCPD/".

[Sega 03] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.". *Nat Genet*, Vol. 34, No. 2, pp. 166–176, Jun 2003.

[Serr 83] R. Serrano. "In vivo glucose activation of the yeast plasma membrane ATPase.". *FEBS Lett*, Vol. 156, No. 1, pp. 11–14, May 1983.

[Sigg 05] E. D. Siggia. "Computational methods for transcriptional regulation.". *Curr Opin Genet Dev*, Vol. 15, No. 2, pp. 214–221, Apr 2005.

[Simo 01] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, and R. A. Young. "Serial regulation of transcriptional regulators in the yeast cell cycle.". *Cell*, Vol. 106, No. 6, pp. 697–708, Sep 2001.

[Simo 06] T. Simoes, N. P. Mira, A. R. Fernandes, and I. S-Correia. "The SPI1 gene, encoding a glycosylphosphatidylinositol-anchored cell wall protein, plays a prominent role in the development of yeast resistance to lipophilic weak-acid food preservatives.". *Appl Environ Microbiol*, Vol. 72, No. 11, pp. 7168–7175, Nov 2006.

[Sipo 02] K. Sipos, H. Lange, Z. Fekete, P. Ullmann, R. Lill, and G. Kispal. "Maturation of cytosolic iron-sulfur proteins requires glutathione.". *J Biol Chem*, Vol. 277, No. 30, pp. 26944–26949, Jul 2002.

[Smit 08] E. N. Smith and L. Kruglyak. "Gene-environment interaction in yeast gene expression.". *PLoS Biol*, Vol. 6, No. 4, p. e83, Apr 2008.

[Soll 81] D. R. Soll, G. W. Bedell, and M. Brummel. "Zinc and regulation of growth and phenotype in the infectious yeast Candida albicans.". *Infect Immun*, Vol. 32, No. 3, pp. 1139–1147, Jun 1981.

[Sosa 03] E. Sosa, C. Aranda, L. Riego, L. Valenzuela, A. DeLuna, J. M. Cant, and A. Gonzlez. "Gcn4 negatively regulates expression of genes subjected to nitrogen catabolite repression.". *Biochem Biophys Res Commun*, Vol. 310, No. 4, pp. 1175–1180, Oct 2003.

[Spel 98] 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*, Vol. 9, No. 12, pp. 3273–3297, Dec 1998.

[Star 86] H. Stark and J. W. Woods. *Probability, random processes and estimation theory for engineers.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1986.

[Subr 05] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.". *Proc Natl Acad Sci U S A*, Vol. 102, No. 43, pp. 15545–15550, Oct 2005.

[Tahe 96] M. J. Taherzadeh, G. Lidn, L. Gustafsson, and C. Niklasson. "The effects of pantothenate deficiency and acetate addition on anaerobic batch fermentation of glucose by Saccharomyces cerevisiae.". *Appl Microbiol Biotechnol*, Vol. 46, No. 2, pp. 176–182, Sep 1996.

[Tai 05] S. L. Tai, V. M. Boer, P. Daran-Lapujade, M. C. Walsh, J. H. de Winde, J.-M. Daran, and J. T. Pronk. "Two-dimensional transcriptome analysis in chemostat cultures. Combinatorial effects of oxygen availability and macronutrient limitation in Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 280, No. 1, pp. 437–447, Jan 2005.

[Tai 07] S. L. Tai, P. Daran-Lapujade, M. C. Walsh, J. T. Pronk, and J.-M. Daran. "Acclimation of Saccharomyces cerevisiae to low temperature: a chemostat-based transcriptome analysis.". *Mol Biol Cell*, Vol. 18, No. 12, pp. 5100–5112, Dec 2007.

[Taid 00] B. Taidi, B. Hoogenberg, A. I. Kennedy, and J. A. Hodgson. "Pre-treatment of pitching yeast with zinc". *MBAA Tech. Q.*, Vol. 37, pp. 431–434, 2000.

[Tan 03] P. K. Tan, T. J. Downey, E. L. Spitznagel, P. Xu, D. Fu, D. S. Dimitrov, R. A. Lempicki, B. M. Raaka, and M. C. Cam. "Evaluation of gene expression measurements from commercial microarray platforms.". *Nucleic Acids Res*, Vol. 31, No. 19, pp. 5676–5684, Oct 2003.

[Tedf 97] K. Tedford, S. Kim, D. Sa, K. Stevens, and M. Tyers. "Regulation of the mating pheromone and invasive growth responses in yeast by two MAP kinase substrates.". *Curr Biol*, Vol. 7, No. 4, pp. 228–238, Apr 1997.

[Thom 06] E. Thomsson and C. Larsson. "The effect of lactic acid on anaerobic carbon or nitrogen limited chemostat cultures of Saccharomyces cerevisiae.". *Appl Microbiol Biotechnol*, Vol. 71, No. 4, pp. 533–542, Jul 2006.

[Thom 93] D. Thomas. *Yeasts as Spoilage Organisms in Beverages. The Yeasts: Yeast Technology.* Academic Press, London, UK., 1993.

[Thom 97] D. Thomas and Y. Surdin-Kerjan. "Metabolism of sulfur amino acids in Saccharomyces cerevisiae". *Microbiol. Mol. Biol. Rev.*, Vol. 61, No. 4, pp. 503–532, 1997.

[Tian 05] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. "Discovering statistically significant pathways in expression profiling studies.". *Proc Natl Acad Sci U S A*, Vol. 102, No. 38, pp. 13544–13549, Sep 2005.

[Tong 04] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D. S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J. N. Levinson, H. Lu, P. Mnard, C. Munyana, A. B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.-M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S. L. Wong, L. V. Zhang, H. Zhu, C. G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F. P. Roth, G. W. Brown, B. Andrews, H. Bussey, and C. Boone. "Global mapping of the yeast genetic interaction network.". *Science*, Vol. 303, No. 5659, pp. 808–813, Feb 2004.

[Tran] "Transfac http://www.gene-regulation.com/".

[Tudo 93] E. Tudor and R. Board. *Food Spoilage Yeasts. The Yeasts: Yeast Technology.* Academic Press, London, UK., 1993.

[Tush 01] V. G. Tusher, R. Tibshirani, and G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response.". *Proc Natl Acad Sci U S A*, Vol. 98, No. 9, pp. 5116–5121, Apr 2001.

[Uemu 05] T. Uemura, K. Tachihara, H. Tomitori, K. Kashiwagi, and K. Igarashi. "Characteristics of the polyamine transporter TPO1 and regulation of its activity and cellular localization by phosphorylation.". *J Biol Chem*, Vol. 280, No. 10, pp. 9646–9652, Mar 2005.

[Vall 90] B. L. Vallee and D. S. Auld. "Zinc coordination, function, and structure of zinc enzymes and other proteins.". *Biochemistry*, Vol. 29, No. 24, pp. 5647–5659, Jun 1990.

[Vasc 01] M. J. Vasconcelles, Y. Jiang, K. McDaid, L. Gilooly, S. Wretzel, D. L. Porter, C. E. Martin, and M. A. Goldberg. "Identification and characterization of a low oxygen response element involved in the hypoxic induction of a family of Saccharomyces cerevisiae genes. Implications for the conservation of oxygen sensing in eukaryotes.". *J Biol Chem*, Vol. 276, No. 17, pp. 14374–14384, Apr 2001.

[Vasi 04] A. Vasiljev, U. Ahting, F. E. Nargang, N. E. Go, S. J. Habib, C. Kozany, V. Panneels, I. Sinning, H. Prokisch, W. Neupert, S. Nussberger, and D. Rapaport. "Reconstituted TOM core complex and Tim9/Tim10 complex of mitochondria are sufficient for translocation of the ADP/ATP carrier across membranes.". *Mol Biol Cell*, Vol. 15, No. 3, pp. 1445–1458, Mar 2004.

[Veen 87] M. Veenhuis, M. Mateblowski, W. H. Kunau, and W. Harder. "Proliferation of microbodies in Saccharomyces cerevisiae.". *Yeast*, Vol. 3, No. 2, pp. 77–84, Jun 1987.

[Vela 04] I. Velasco, S. Tenreiro, I. L. Calderon, and B. Andr. "Saccharomyces cerevisiae Aqr1 is an internal-membrane transporter involved in excretion of amino acids.". *Eukaryot Cell*, Vol. 3, No. 6, pp. 1492–1503, Dec 2004.

[Verd 90] C. Verduyn, E. Postma, W. A. Scheffers, and J. P. van Dijken. "Energetics of Saccharomyces cerevisiae in anaerobic glucose-limited chemostat cultures.". *J Gen Microbiol*, Vol. 136, No. 3, pp. 405–412, Mar 1990.

[Verd 92] C. Verduyn, E. Postma, W. A. Scheffers, and J. P. V. Dijken. "Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation.". *Yeast*, Vol. 8, No. 7, pp. 501–517, Jul 1992.

[Vieg 98] C. A. Viegas, P. F. Almeida, M. Cavaco, and I. S-Correia. "The H(+)-ATPase in the plasma membrane of Saccharomyces cerevisiae is activated during growth latency in octanoic acid-supplemented medium accompanying the decrease in intracellular pH and cell viability.". *Appl Environ Microbiol*, Vol. 64, No. 2, pp. 779–783, Feb 1998.

[Vijv 02] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. "A gene-expression signature as a predictor of survival in breast cancer.". *N Engl J Med*, Vol. 347, No. 25, pp. 1999–2009, Dec 2002.

[Viss 90] W. Visser, W. A. Scheffers, W. H. B. van der Vegte, and J. P. van Dijken. "Oxygen requirements of yeasts.". *Appl Environ Microbiol*, Vol. 56, No. 12, pp. 3785–3792, Dec 1990.

[Viss 94] W. Visser, A. A. van der Baan, W. B. van der Vegte, W. A. Scheffers, R. Kramer, and J. P. van Dijken. "Involvement of mitochondria in the assimilatory metabolism of anaerobic Saccharomyces cerevisiae cultures.". *Microbiology*, Vol. 140 ( Pt 11), pp. 3039–3046, Nov 1994.

[Walk 98] G. M. Walker. *Yeast metabolism*. John Wiley & Sons Ltd, 1998.

[Wang 05] W. Wang, J. Cherry, Y. Nochomovitz, E. Jolly, D. Botstein, and H.Li. "Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation". *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 102, No. 6, pp. 1998–2003, 2005.

[Wart 89] A. D. Warth. "Relationships between the resistance of yeasts to acetic, propanoic and benzoic acids and to methyl paraben and pH.". *Int J Food Microbiol*, Vol. 8, No. 4, pp. 343–349, Jul 1989.

[Wate 02] B. M. Waters and D. J. Eide. "Combinatorial control of yeast FET4 gene expression by iron, zinc, and oxygen.". *J Biol Chem*, Vol. 277, No. 37, pp. 33749–33757, Sep 2002.

[Weie 99] T. Weierstall, C. P. Hollenberg, and E. Boles. "Cloning and characterization of three genes (SUT1-3) encoding glucose transporters of the yeast Pichia stipitis.". *Mol Microbiol*, Vol. 31, No. 3, pp. 871–883, Feb 1999.

[Wing 00] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prss, I. Reuter, and F. Schacherer. "TRANSFAC: an integrated system for gene expression regulation.". *Nucleic Acids Res*, Vol. 28, No. 1, pp. 316–319, Jan 2000.

[Wing 85] D. R. Winge, K. B. Nielson, W. R. Gray, and D. H. Hamer. "Yeast metallothionein. Sequence and metal-binding properties.". *J Biol Chem*, Vol. 260, No. 27, pp. 14464–14470, Nov 1985.

[Wu 04] J. Wu, N. Zhang, A. Hayes, K. Panoutsopoulou, and S. Oliver. "Global analysis of nutrient control of gene expression in Saccharomyces cerevisiae during growth and starvation". *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 101, No. 9, pp. 3148–3153, 2004.

[Wu 07] C.-Y. Wu, A. J. Bird, D. R. Winge, and D. J. Eide. "Regulation of the yeast TSA1 peroxiredoxin by ZAP1 is an adaptive response to the oxidative stress of zinc deficiency.". *J Biol Chem*, Vol. 282, No. 4, pp. 2184–2195, Jan 2007.

[Yate 99] R. D. Yates and D. J. Goodman. *Probability and stochastic processes.* John Wiley & Sons, Inc., 1999.

[Yean 06] C.-H. Yeang and T. Jaakkola. "Modeling the combinatorial functions of multiple transcription factors.". *J Comput Biol*, Vol. 13, No. 2, pp. 463–480, Mar 2006.

[Yeas] "Yeast Protein Database http://www.proteome.com".

[Yu 06] X. Yu, J. Lin, T. Masuda, N. Esumi, D. J. Zack, and J. Qian. "Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae.". *Nucleic Acids Res*, Vol. 34, No. 3, pp. 917–927, 2006.

[Yun 01] C. W. Yun, M. Bauler, R. E. Moore, P. E. Klebba, and C. C. Philpott. "The role of the FRE family of plasma membrane reductases in the uptake of siderophore-iron in Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 276, No. 13, pp. 10218–10223, Mar 2001.

[Zakr 05] A. Zakrzewska, A. Boorsma, S. Brul, K. J. Hellingwerf, and F. M. Klis. "Transcriptional response of Saccharomyces cerevisiae to the plasma membrane-perturbing compound chitosan.". *Eukaryot Cell*, Vol. 4, No. 4, pp. 703–715, Apr 2005.

[Zara 00] O. Zaragoza and J. M. Gancedo. "Pseudohyphal growth is induced in Saccharomyces cerevisiae by a combination of stress and cAMP signalling.". *Antonie Van Leeuwenhoek*, Vol. 78, No. 2, pp. 187–194, Aug 2000.

[Zhao 96a] H. Zhao and D. Eide. "The yeast ZRT1 gene encodes the zinc transporter protein of a high-affinity uptake system induced by zinc limitation.". *Proc Natl Acad Sci U S A*, Vol. 93, No. 6, pp. 2454–2458, Mar 1996.

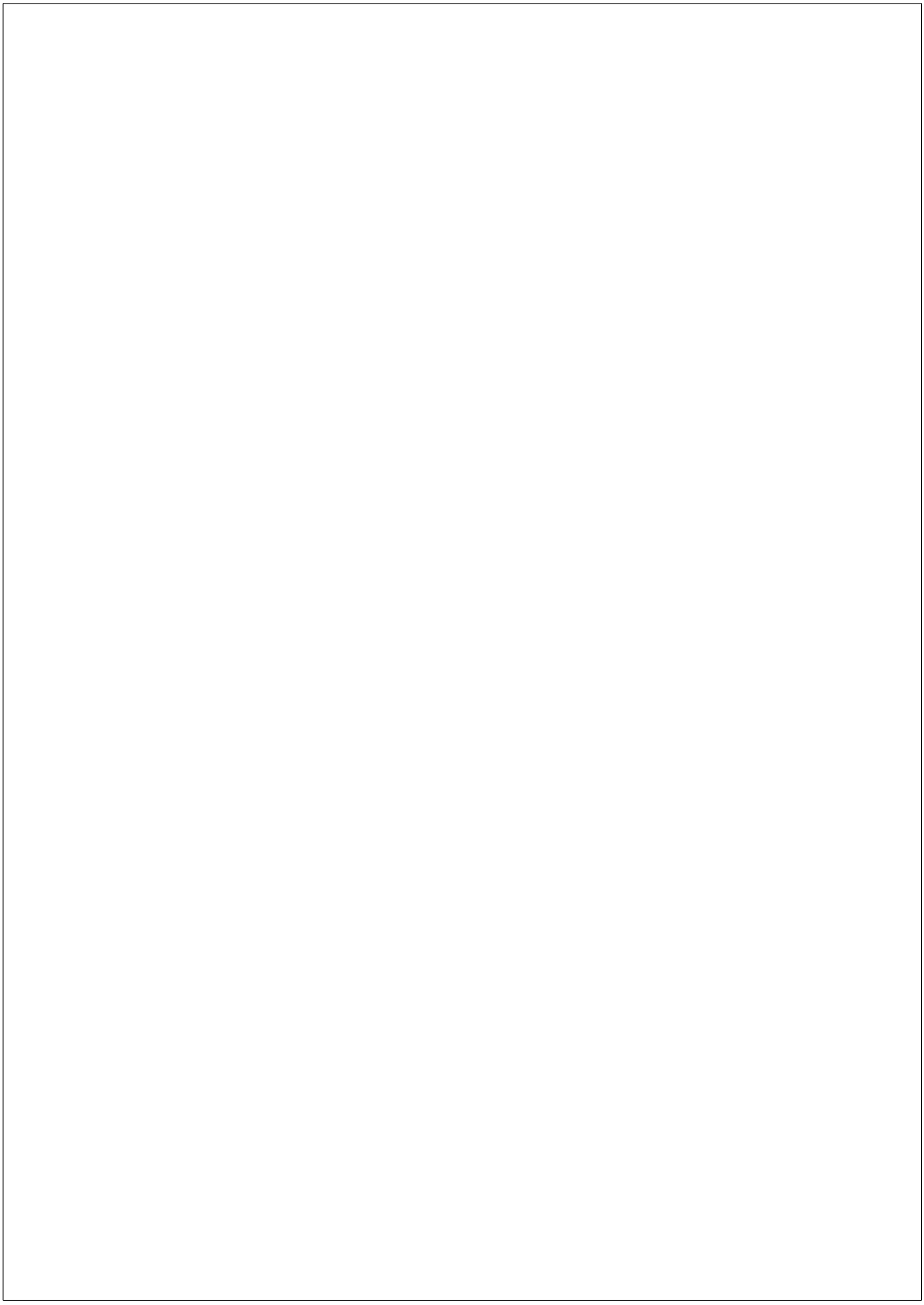
[Zhao 96b] H. Zhao and D. Eide. "The ZRT2 gene encodes the low affinity zinc transporter in Saccharomyces cerevisiae.". *J Biol Chem*, Vol. 271, No. 38, pp. 23203–23210, Sep 1996.

[Zhao 97] H. Zhao and D. J. Eide. "Zap1p, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in Saccharomyces cerevisiae.". *Mol Cell Biol*, Vol. 17, No. 9, pp. 5044–5052, Sep 1997.

[Zito 92] R. S. Zitomer and C. V. Lowry. "Regulation of gene expression by oxygen in Saccharomyces cerevisiae.". *Microbiol Rev*, Vol. 56, No. 1, pp. 1–11, Mar 1992.

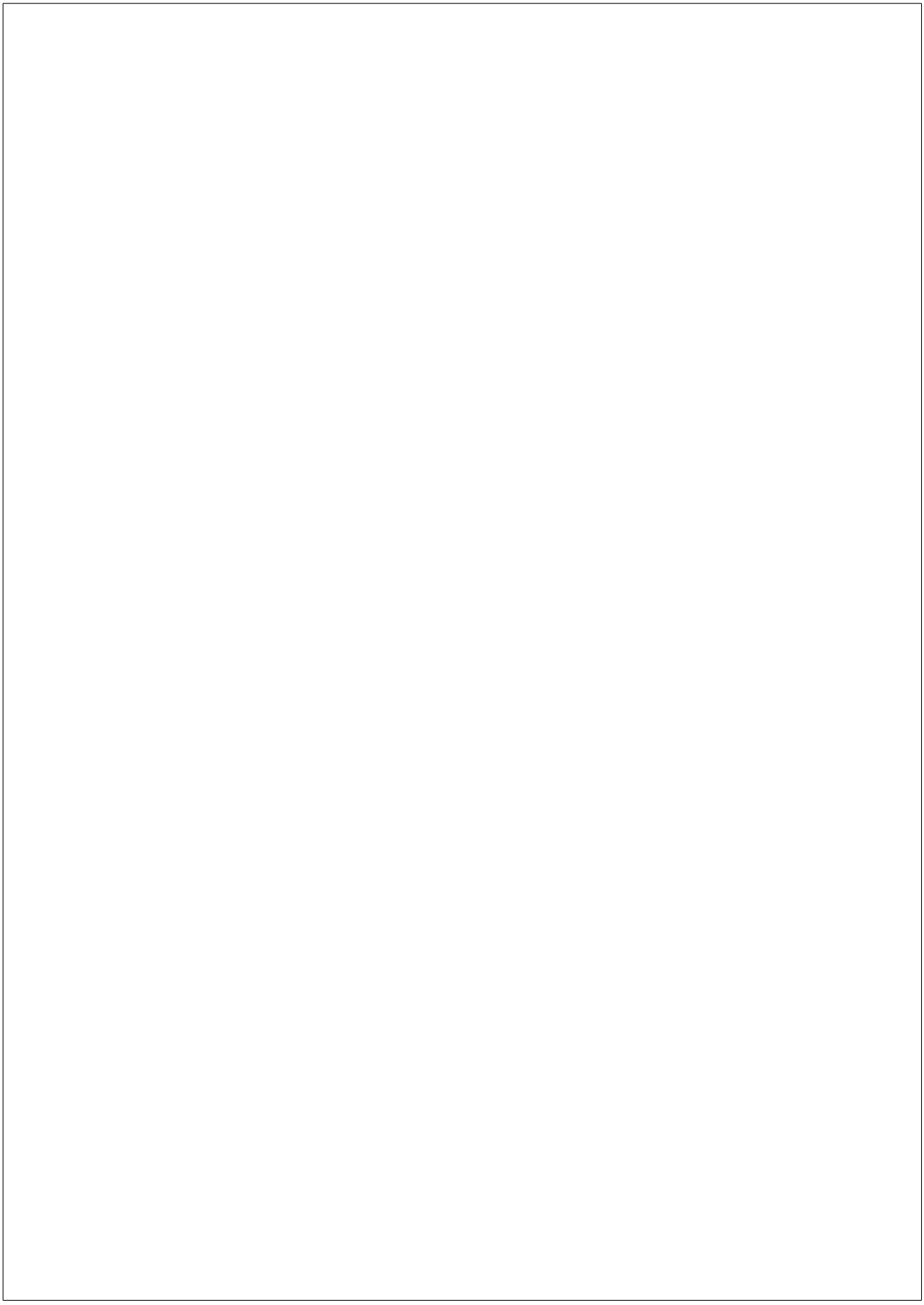
# SUMMARY

Exactly how an organism adapts its transcriptional program in response to intra- and extracellular signals remains elusive. Development of computational approaches that use the large amounts of diverse intracellular data to unravel the cell's transcriptional program is one of today's main challenges in bioinformatics research. This thesis contributes to that field by investigating the transcriptional response of the yeast *Saccharomyces cerevisiae* to multiple chemical and physical signals from its environment. In contrast to the commonly used shake-flask cultures, the gene expression data employed in this thesis originates from yeast grown in steady-state chemostat cultures. These chemostat cultures enable the accurate control, measurement and manipulation of individual cultivation parameters, such as growth rate, temperature and nutrient concentrations. A growth condition can thus be characterized by the combined settings of several cultivation parameters. Computational methods are developed that use these 'multifactorial' growth conditions to infer the effect of individual cultivation parameters and combinations of cultivation parameters on gene expression. The gene expression measurements are integrated with data about the binding potential of transcription factors (TF), the proteins that bind the DNA near a gene (promoter region) and possibly manipulate the rate at which the gene is transcribed. This integration enables us to investigate the effect of cultivation parameters on the activity of TFs. We present computational approaches that not only infer the activity of TFs as a function of the cultivation parameters, but also describe the combinatorial interplay between different TFs on gene promoters to regulate a gene's rate of transcription.



# SAMENVATTING

De precieze wijze, waarop een organisme zijn programma van gentranscriptie aanpast aan intra- en extracellulaire signalen is in het geheel nog niet duidelijk. Bioinformaticaonderzoek heeft de belangrijke rol om computertechnieken te ontwikkelen, die de grote en diverse hoeveelheden beschikbare data van metingen in de cel kunnen gebruiken om het transcriptieprogramma van een cel te ontravelen. Dit proefschrift vormt een bijdrage aan dit vakgebied door onderzoek te doen naar de transcriptionele reactie van bakkersgist op verschillende fysische en chemische signalen van buitenaf. De genexpressiemetingen, die worden gebruikt in dit proefschrift zijn afkomstig van gist groeiend in 'steady-state' chemostaat culturen. In tegenstelling tot de vaakgebruikte schudkolfen faciliteren de chemostaten de nauwkeurige controle, meting en manipulatie van individuele cultivatieparameters, zoals de groeisnelheid, temperatuur en voedingsconcentraties. Een groeiconditie kan dus worden gekenmerkt door de gecombineerde instellingen van verschillende cultivatieparameters. Computertechnieken zijn ontwikkeld om vanuit deze 'multifactoriële' groeicondities het effect van individuele cultivatieparameters, maar ook combinaties van verschillende cultivatieparameters op genexpressie af te leiden. De genexpressiemetingen zijn geïntegreerd met informatie over de binding van transcriptiefactoren (TF's); de eiwitten, die vlak bij een gen kunnen binden op het DNA om daar het proces van gentranscriptie te beïvloeden. Deze integratie stelt ons in staat om het effect van de cultivatieparameters op de activiteit van de TF's te onderzoeken. Wij introduceren computertechnieken, die niet alleen de activiteit van de TF's behandelen als functie van de cultivatieparameters, maar ook de interactie tussen verschillende TF's, die bepalend is voor de transcriptiesnelheid van een gen, beschrijven.



# PUBLICATIONS

2009

- T.A. Knijnenburg, J.M. Daran, M.A. van den Broek, P. Daran-Lapujade, J.H. de Winde, J.T. Pronk, M.J.T. Reinders and L.F.A. Wessels, 'Combinatorial effects of environmental parameters on transcriptional regulation in *Saccharomyces cerevisiae*: A quantitative analysis of a compendium of chemostat-based transcriptome data', *BMC Genomics*, vol. 10, no. 53, 2009.

2008

- T.A. Knijnenburg, L.F.A. Wessels and M.J.T. Reinders, 'Creating gene set activity profiles with time-series expression data', *International Journal of Bioinformatics Research and Applications (IJBRA)*, vol. 4, no. 3, pp. 306-323, 2008.
- T.A. Knijnenburg, L.F.A. Wessels and M.J.T. Reinders, 'Combinatorial influence of environmental parameters on transcription factor activity', *Bioinformatics*, vol. 24, no. 13, pp. i172-82, 2008.

2007

- M.J.L. de Groot, P.A.S. Daran-Lapujade, B. van Breukelen, T.A. Knijnenburg, M.J.T. Reinders, J.T. Pronk, A.J.R. Heck and M. Slijper, 'Quantitative proteomics and transcriptomics of anaerobic and aerobic yeast cultures reveals posttranscriptional regulation of key cellular processes', *Microbiology UK*, vol. 153, no. 11, pp. 3864-3878, 2007.
- T.A. Knijnenburg, J.H. de Winde, J.M. Daran, P. Daran-Lapujade, J.T. Pronk, M.J.T. Reinders and L.F.A. Wessels, 'Exploiting combinatorial cultivation conditions to infer transcriptional regulation', *BMC Genomics*, vol. 8, no. 25, 2007.
- D.A. Abbott, T.A. Knijnenburg, L.M. de Poorter, M.J.T. Reinders, J.T. Pronk and A.J.A. van Maris, 'Generic and specific transcriptional responses to different weak organic acids in anaerobic chemostat cultures of Saccharomyces cerevisiae.', *FEMS Yeast Res.*, vol. 7, issue 6, pp 819-833, 2007.
- M. Clements, E.P. van Someren, T.A. Knijnenburg and M.J.T. Reinders, 'Integration of Known Transcription Factor Binding Site Information and Gene Expression Data to Advance from Co-Expression to Co-Regulation', *Genomics, Proteomics & Bioinformatics*, vol. 5, no. 2, pp. 86 - 101, 2007.

- R. de Nicola, L.A. Hazelwood, E.A.F. de Hulster, M.C. Walsh, T.A. Knijnenburg, M.J.T. Reinders, G.M. Walker, J.T. Pronk, J.G. Daran, and P.A.S. Daran-Lapujade, 'Physiological and transcriptional responses of Saccharomyces cerevisiae to zinc limitation in chemostat cultures', *Applied and environmental microbiology*, vol. 73, no. 23, pp. 7680-7692, 2007.

2006

- T.A. Knijnenburg, M.J.T. Reinders and L.F.A. Wessels, 'Artifacts of Markov blanket filtering based on discretized features in small sample size applications', *Pattern Recognition Letters*, vol. 27, pp. 709-714, 2006.

- M. Clements (E.P. van Someren, T.A. Knijnenburg and M.J.T. Reinders thesis supervisors), 'Discovery of co-regulated gene clusters by combining known transcription factor binding motifs and gene expression profiles', *M.Sc. Thesis*, pp. 21, Delft University of Technology, April 2006.

- M.T.A.P. Kresnowati, W.A. van Winden, M.J.H. Almering, A. ten Pierick, C. Ras, T.A. Knijnenburg, P. Daran-Lapujade, J.T. Pronk, J.J. Heijnen and J.M. Daran, 'When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation', *Mol. Syst. Biol.*, vol. 2, no. 49, 2006.

- T.A. Knijnenburg, J.M. Daran, P. Daran-Lapujade, M.J.T. Reinders and L.F.A. Wessels, 'Relating transcription factors, modules of genes and cultivation conditions in Saccharomyces cerevisiae', *Twelfth annual conference of the Advanced School for Computing and Imaging*, 2006.

- T.A. Knijnenburg, L.F.A. Wessels and M.J.T. Reinders, 'Condition transition analysis reveals TF activity related to nutrient-limitation-specific effects of oxygen presence in yeast', *International Conference on Computational Methods in Systems Biology*, Trento, Italy, Proceedings p 271-284, 2006.

2005

- T.A. Knijnenburg, M.J.T. Reinders, and L.F.A. Wessels, 'The selection of relevant and non-redundant features to improve classification performance of microarray gene expression data', *Eleventh annual conference of the Advanced School for Computing and Imaging*, 2005.

2004

- T.A. Knijnenburg (M.J.T. Reinders and L.F.A. Wessels thesis supervisors), 'Selecting relevant and nonredundant features in microarray classification applications', *M.Sc. Thesis*, Delft University of Technology, Februar 2004.

- T.A. Knijnenburg (M.J.T. Reinders and L.F.A. Wessels thesis supervisors), 'Artifacts of Markov blanket filtering based on discretized features in small sample size application', *M.Sc. Thesis*, pp. 23, Delft University of Technology, Februar 2004.

# CURRICULUM VITAE

Theo Arjan Knijnenburg was born in Leidschendam, The Netherlands, on August 21, 1980. He obtained his VWO-diploma from Veurs College in Leidschendam in 1998, after which he started his study Electrical Engineering at the Delft University of Technology. During a three-month internship in 2002 he worked in British Telecom's Future Content Group in Ipswich, UK, on object detection in video for portable devices. In 2004 he obtained his M.Sc. degree in Electrical Engineering after doing his graduation work in the Information and Communication Theory group at the Delft University of Technology on feature selection in gene expression based tumor classification problems. In 2004 he started his Ph.D. study in the Information and Communication Theory Group on unraveling the transcriptional program of the yeast *Saccharomyces cerevisiae*. His Ph.D. project was performed in collaboration with the Industrial Microbiology Group of the Delft University of Technology and was part of the Kluyver Centre for Industrial Fermentation. Since October 2008 he works as a postdoctoral researcher at the Institute for Systems Biology, Seattle, US.