genomics *Saccharomyces cerevisiae* DNA fitness profiling

Chapter 1

micro-array TFIID PAC rRPE Hypergeometric CGATGAG AAAATTTT T-profiler MSN2 GSEA mRNA Slt2

General Introduction

CTA(W4)TAG Bonferonni Rlm1 ChIP-chip Environmental Stress Response Gene Ontology MIPS REDUCE Fisher’s exact TATA-box Rlm1

Genomics in *Saccharomyces cerevisiae*

In recent years the focus in molecular biology research has shifted from the study of individual genes and proteins to the relationships between the genes and proteins of an entire organism. The study of an organism's entire genome, and the use of the genes derived from sequence information is called genomics. An important driving force of genomics was the development of a technique to determine the base-pair sequence of an organism's total genome. In 1975, Frederick Sanger developed a DNA sequencing method [1] that later could be automated, and two years later he determined the first complete genomic sequence obtained from the bacteriophage phi X-174 [2]. Since then more than 1800 genomes have been sequenced [3], with an important milestone in 1996, when sequencing of the first genome of a eukaryotic organism (*Saccharomyces cerevisiae*) was completed, and later in 2001, when the human genome was finished [4]. Much effort is undertaken to understand the function of genes and proteins. For a long time it was thought that information derived from the genome sequence would explain a great deal of the functioning of organisms. However, recent technological developments have made it possible to measure simultaneously the level of every mRNA molecule in cells, in a single experiment. From these experiments we have learned that the timing of the synthesis of mRNA and mRNA levels (and subsequently protein levels), in other words the regulation of gene expression, play a very important role in the functioning of cells. Information about gene expression is also coded in the DNA. One of the most challenging tasks within biology is to understand how a DNA sequence translates into complex biological functions. This starts with the biological interpretation of such transcriptional experiments carried out on a genomic scale, which are therefore very data-rich. This thesis describes the development and application of bioinformatics tools that are designed to help interpret large and complex biological datasets. Such tools can help to elucidate the complex interactions between genes. Most of the data analysis described throughout this thesis is derived from the unicellular eukaryote *Saccharomyces cerevisiae*.

Saccharomyces cerevisiae as a model organism for molecular biology

Saccharomyces cerevisiae, or baker's yeast, is a budding yeast. Mankind has used it since ancient times for baking, brewing, and winemaking. More importantly in this context, it is also intensively studied as a eukaryotic model organism in molecular and cell biology. The knowledge of conserved basic biological processes in *S. cerevisiae* helps to understand similar processes in other organisms. As mentioned before, the genome of *Saccharomyces cerevisiae* was the first of an eukaryotic organism to be fully sequenced [5]. Its genome is composed of about 12,000,000 base pairs and contains about 6,000 functional genes. Seventy percent of the genome contains coding information. Despite many years of research, the function of approximately 25% of these genes is still not known [6, 7]. The individual deletion of about 1100 genes results in a lethal phenotype and is therefore considered essential whereas individual deletion of the remainder of the genes results in a viable phenotype, at least when the cells are cultured in rich medium under optimal conditions [8]. Importantly, more than 40% of yeast proteins share some conserved sequence with at least one known or predicted human protein, including several hundred genes implicated in human disease [9]. A classical example is the identification of the cell division-cycle (CDC) genes in *S. cerevisiae* by Leland Hartwell [10]; later it was discovered that these genes have evolutionary conserved human counterparts. In addition, *S. cerevisiae* is a food spoilage organism and is therefore also a good model to study fungal food spoilage. Besides being a model organism in cell biology, *S. cerevisiae* is also leading the way in the development of new genomic techniques, especially those used in functional genomics.

Functional genomics in *Saccharomyces cerevisiae*

One of the main goals of genomic research is to decipher, annotate and understand the biological role of every feature of a DNA sequence in the genome. Since the introduction of high-throughput techniques, the focus has been shifted from the functionality of individual genes to a more global view of how the cellular network functions and how cellular subsystems interact and function together. This systematic study of the complex interactions in biological systems has become a new field of study and is called systems biology. The availability of data produced by various genome sequencing projects is now followed up by high throughput methods that study the function and interaction of genes in a genome-wide fashion. Functional genomics is the part of molecular biology that uses high-throughput techniques like DNA micro-arrays (transcriptomics), proteomics, metabolomics, etc, to describe genome function. *S. cerevisiae* has demonstrated its value again and again and is exploited as a workhorse for the development of functional genomics techniques. In the next section a selection of the high-throughput techniques that where pioneered in baker's yeast will be discussed.

(1) **Protein-protein interaction**. The first global map of protein interactions was described in baker's yeast. High-throughput detection of protein-protein interactions is based on two methods: The yeast two-hybrid system [11-13] and protein complex purification in combination with the use of mass spectrometry to identify the proteins involved [14, 15]. These studies provide insight into how the proteome, the ensemble of expressed proteins, is organized into functional complexes.

(2) In the **Saccharomyces Genome Deletion Project** all genes of this organism have been deleted individually. This project revealed that 1105 genes (18.7 %) were essential for growth in rich medium [8]. Each deletion strain has been generated in such a way that it is labeled with a unique bar code that enables simultaneous analysis of all deletion mutants in competition experiments [16, 17]. The application of these so-called fitness profiling experiments will be discussed in more detail later.

(3) In another project, **protein localization** was studied by tagging Green Fluorescent Protein (GFP) to individual proteins [18]. Protein localization within a cell often correlates with, and indicates, function. The subcellular localization of 2,744 proteins was determined, of which 1,000 proteins with an unknown function.

(4) The same approach of proteins tagged with a reporter, was used to determine the **abundance of proteins** in the cell by using GFP fluorescence as a measure for the amount of pro-

tein [19].

(5) Saito *et al*., [20] studied the global **regulation of morphology** by automatically measuring morphological features of gene deletion mutants. Their system provides quantitative data for shapes of the daughter and mother cells, localization of the nuclear DNA and morphology of the actin patches. As with any high-throughput method, it is desirable to demonstrate the physiological relevance of such results, which is strengthened by independent and complementary methods.

(6) One of the most matured/developed high-throughput methods is measurement of the **abundance of all individual mRNA's** by the micro-array technique [21, 22]. DNA micro-arrays or transcriptomics provide information about the regulation of transcription. The analysis and interpretation of transcriptome data is a prominent part of this thesis. Therefore, in the next section the process of transcriptional regulation will be discussed.

Regulation of transcription in Eukaryotes

The distinction between man and mouse is not so much the number of different proteins that their genomes encode, but rather, the timing of the appearance of common proteins and their positions in the developing organisms is important, Ptashne *et al*. [23]. The activity of proteins and their localization can be regulated at the posttranslational level in various ways, for example, by protein (de)phosphorylation, degradation or stabilization, but one of the most important ways of regulation is at the level of initiation of transcription. Transcription is the process of transcribing particular parts of the DNA into RNA and is carried out by RNA polymerases. *Saccharomyces cerevisiae* contains three different RNA polymerases that are capable of synthesizing specific RNAs. Polymerase I is responsible for the synthesis of ribosomal RNA (rRNA), Polymerase II synthesizes messenger RNA (mRNA) and Polymerase III synthesizes transfer RNA (tRNA), 5S rRNA and other small RNA molecules. Since the DNA microarray technique mainly measures the changes of all mRNA levels, we will focus on the synthesis of protein-encoding genes, which is carried out by RNA Polymerase II (RNA Pol II).

The RNA Polymerase II Complex

RNA Pol II is a multi-subunit enzyme that requires additional proteins to recognize and reach promoter sequences in order to accurately initiate transcription. Transcription is a process that is tightly controlled on various levels. This is absolutely necessary since activation of genes under the improper conditions can be harmful or even lethal to the cell [24]. The typical RNA Pol II transcription cycle begins with the binding of gene-specific transcription factors upstream of the core promoter. The core promoter is defined as the minimal stretch of contiguous DNA sequence that is sufficient to direct accurate initiation of transcription by the RNA polymerase II machinery [25]. The binding of gene-specific transcription factors leads to the recruitment of adaptor complexes such as SAGA or Mediator, both of which in turn facilitate binding of general transcription factors (GTFs). The GTFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH) each carry out a specific function. For example, TFIIA and TFIIB are responsible for the positioning of Pol II to the promoter. TFIIH melts 11–15 bp of DNA in order to position the single-strand template in the Pol II cleft to initiate RNA synthesis. TFIID in turn is also a multi-subunit complex that is involved in promoter recognition. RNA Polymerase II together with all the above-mentioned subunits forms the Pre Iniation Complex (PIC) (**Figure 1**).

The ground state of eukaryotic transcription is restrictive

In contrast to prokaryotic gene expression, the ground state for eukaryotic transcription is restrictive [26]. This is mainly because in eukaryotic cells the DNA is packaged into nucleosomes. DNA packaged into nucleosomes is called chromatin. Chromatin is composed of repeating units of nucleosomes in which 147 base pairs of DNA are wrapped 1.7 times around the exterior of a histone complex. The chromatin structure enables the chromosomal DNA to fit into the nucleus by compaction. Besides the compaction role of chromatin, it also serves as a mechanism to control transcription by preventing RNA Pol II access to the DNA. The accessibility of DNA can further be controlled by methylation of DNA, which is called imprinting, or by nucleosome remodeling and covalent modification of histones, including methylation, phosphorylation, acetylation, sumoylation and ubiquitylation.

The TATA box-containing and TATA-less genes

The TFIID complex contains the TATA-box binding protein (TBP) and fourteen other TBP-associated factors (TAFs) [27] that are collectively thought to be the main DNA-binding proteins that regulate promoter specificity in yeast. TBP (encoded by *SPT15*) is a DNA-binding protein that binds specifically to the TATA-box, an AT-rich sequence that is located upstream of the translational start site of a gene. The TATA-box is a core promoter element with a DNA sequence of: 5’-TATAA-3’, which is usually followed by three or more adenine bases and has been highly conserved through evolution. In most textbooks on transcriptional regulation the TATA-box is presented as being present and functional in almost all genes. However, it is becoming clear that many genes lack a TATA box. Recent studies suggest that only 10– 15% of the human core promoters contain a TATA-box [28]. Promoters lacking a TATA box are called TATA-less promoters. In mammals, other core promoter elements like the initiator element and the downstream core promoter element (DPE) are used instead [29]. A protein complex named NC2 (Negative Cofactor 2) that interacts with TBP is involved in the activation of core promoters that are dependent on DPE [29]. NC2 was identified as a repressor of TATA-dependent transcription; however, more recent experiments suggest that NC2 could also positively affect gene transcription (**Figure 1**). In agreement with the notion of human core promoters, recent research suggests that 80% of the *Saccharomyces cerevisiae* genes have TATA-less promoters [30]. Interestingly, alternative core promoter elements such as the DPE have not yet been discovered in *S. cerevisiae*.

TBP-associated Factors

The TBP-associated factors (TAFs) represent another component of the RNA Polymerase II Complex. TAFs can form part of other multi-subunit regulatory factors such as the histone acetylation SAGA complex. SAGA is a large complex of over 15 different proteins. TFIID and SAGA share a common set of TAFs; they both regulate chromatin and are capable of delivering TBP to promoters. Recently, it was found that although TFIID and SAGA make an overlapping contribution to the transcription of almost all genes, TFIID is involved in the transcription of about 90% of the genes and SAGA of about 10% of the genes [31]. Interestingly, the SAGA controlled genes are expressed under stress conditions, while the TFIID-controlled genes are merely housekeeping genes. In *S.cerevisiae*, TATA box-containing promoters utilize SAGA rather than TFIID, which is used for TATA box-less promoters [30].

The Mediator complex

In addition to GTFs and TAFs a large multi-subunit co-activator complex that is named Mediator is known to be part of the transcriptional machinery. The mediator complex is conserved from yeast to humans but the number of subunits may vary from organism to organism. The Mediator complex is involved in the regulation of transcription by either forming a molecular bridge between gene specific transcription factors (see below) and the basal transcription machinery, or is recruited to modify the chromatin structure at the promoter [32].

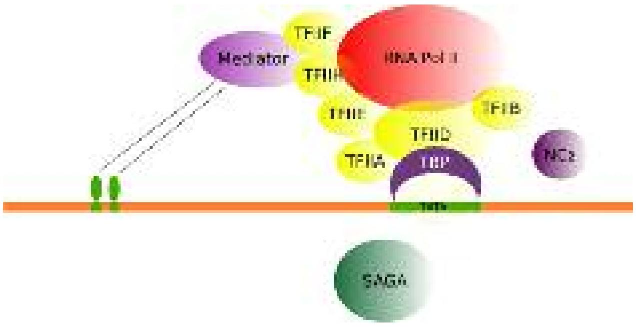


Figure 1 The Pre-initiation Complex. Figure 1 depicts the pre-initiation complex with central Polymerase II, which is surrounded by the general transcription factors (TFIIA, B, D, E, F and H). TFIIA and TFIIB are responsible for the positioning of Pol II relative to the promoter and TFIIH melts a part of the DNA for Pol II to initiate RNA synthesis. TFIID interacts with the TATA-box Binding Protein (TBP) that in turn binds the core-promoter sequence of the gene. Note that although a TATA-box is shown in the figure, only 20% of the yeast genes contain such a core-promoter element. NC2 is able to prevent the binding of TBP to the TATA-box of the gene and stimulates binding to TATA-less genes. Under stress conditions, the SAGA complex is also able to form a complex with TBP. Finally the Mediator complex is able to form a molecular bridge between transcription factors that are bound to *cis*-regulatory sequences in the promoter region of genes.

Recruitment of the Pre-initation Complex by gene-specific transcription factors

SAGA and TFIID are involved in gene regulation in a general way; the regulation of more specific gene groups is carried out by specialized transcription factors (TFs). These transcription factors, which should not be confused with the general transcription factors (GTFs), can be activators or repressors, and bind to specific *cis*-acting DNA sequences that are mostly located upstream of the core promoter elements. Initiation of transcription by RNA Pol II is controlled by these transcription factors. Currently, there are two models that describe the mechanism of transcriptional initiation by transcription factors.

Recruitment: the gene- specific transcription factor (or transcription factor complex) recruits the transcriptional machinery complex in the vicinity of the core promoter, it increases the local concentration of the transcriptional machinery complex and facilitates binding to the core promoter.

Conformational change: the transcriptional machinery complex is already bound to the core promoter in an inactive conformation. Transcription factors that bind to the same gene interact with the transcriptional machinery complex and thereby induce a conformational change. This changes the transcriptional machinery from an inactive into an active conformation.

Currently, most experimental evidence concerning transcriptional initiation supports the recruitment model. The most compelling evidence for this model is inferred from experiments in which a component of the Mediator complex (Gal11p) and the bacterial DNA binding factor LexAp are fused. This artificial protein complex was enough to drive transcription in yeast when a gene was provided with the DNA binding motif of LexAp [33]. Recently, using the more advanced Chromatin-Immuno precipitation technique three stages of recruitment could be resolved upon induction of the galactose genes. First SAGA, then Mediator, and finally RNA Pol II together with four other proteins, including TBP, were recruited to the promoter [34].

The conformational change model is not (yet) strongly supported by experimental evidence. Recently, observations from yeast cells, exiting the quiescence state, showed that RNA Pol II is already present on promoters of still repressed genes [35]. Upon exit from quiescence, these genes are activated very rapidly, suggesting that this might be triggered by conformational changes. The observation of RNA Pol II already residing on the DNA is in contradiction with the recruitment model, which assumes that the general transcription machinery is absent from promoter sequences unless an activating signal is percieved.

Regulation of transcription by gene-specific transcription factors in bakers yeast

Unicellular organisms like yeast constantly face sudden and huge fluctuations in their external conditions. Despite these fluctuations it is crucial that yeast cells adjust their internal homeostasis. Perturbation of the internal conditions may disrupt cellular functions and prevent growth. Therefore, yeast cells must rapidly adapt to new conditions by adjusting their internal milieu. One way to adapt is by reorganizing the genomic transcription program that is required for the survival in each environment. Basically, the transcription programs involved in the response to stress have aspects related to general and specific transcriptional responses. General transcriptional responses are activated in response to a broad spectrum of stresses, whereas specific transcriptional responses are triggered only when a particular stress is sensed.

Regulation of general transcriptional stress responses in bakers yeast

Two homologous and fully redundant transcription factors involved in the response to various stresses are the Cys_2His_2 zinc finger proteins Msn2p and Msn4p. Martinez-Pastor [36] showed that disruption of both *MSN2* and *MSN4* genes results in a higher sensitivity to different stresses like carbon source starvation, heat shock, and severe osmotic and oxidative stress. They also showed that Msn2p and Msn4p are required for the activation of several yeast genes such as *CTT1*, *DDR2* and *HSP12,* whose induction is mediated through stress-responsive regulatory elements (STRE). Msn2p and Msn4p bind specifically to the *cis*-regulatory motif (STRE) 5'-AGGGG-3' and its reverse complement 5'-CCCCT-3'. Msn2p and Msn4p are located in the cytoplasm, but in response to stress they are translocated to the nucleus. Msn2p and Msn4p share a conserved nuclear localization signal (NLS) that is negatively controlled by high cellular protein kinase (PK) activity [37, 38]. Interestingly, Jacquet *et al*. [39] used Msn2p and Msn4p tagged with GFP to show that both proteins repetitively shuttle into and out of the nucleus with a periodicity of a few minutes. Under non-stress conditions the proteins stay most of the time in the cytoplasm, but the balance is shifted towards the nucleus when cells are stressed.

The Environmental Stress Response

The global transcriptional response to a variety of environmental stresses has been extensively studied using the micro-array technique (see also below) [40]. These stresses include carbon- and nitrogen-source limitation, heat stress, oxidative stress and osmotic stress. Comparative analysis of the transcript profiles of the diverse environmental stresses revealed a stereotypical, transcriptional response of about 900 genes. A similar set of genes was found in a closely related study [41]. The response shown by this set of genes is named the Environmental Stress Response (ESR) and their transcription is activated by a wide variety of stresses. The ESR genes can be divided into two groups that show an opposite transcriptional behavior; a group of 600 genes that is referred to as ´repressed genes´ and a group of about 300 genes that is referred to as 'induced genes' [42]. The genes of the 'repressed' group are enriched in ribosomal biogenesis functions and their promoter sequences are overrepresented with the PAC 5'-CGATGAG-3' and rRPE 5'-AAATTTT-3' DNA motifs [43]. Recently, it was shown that the rRPE motif is associated with genes required for rapid growth [44]. This is in accordance with the observation that during most stress responses growth is temporarily halted [40]. In chapter 4 of this thesis we show that the division in a group of induced or repressed genes is an oversimplification since in time-series experiments of particular stresses (e.g. DTT-stress, hypo-osmotic stress and cold stress) the ribosomal biogenesis gene group is first up-regulated but becomes down-regulated afterwards.

Most of the genes induced in the ESR have metabolic functions and share STRE motifs in their promoter sequences. The aforementioned Msn2p and Msn4p transcription factors bind and regulate genes with such STRE motifs [36]. STRE-controlled genes show an exact opposite effect during the time series of a stress. Most stress responses studied in the experiments by Gasch *et al* [40] also show a growth-rate reduction that correlates with transcriptional activity of the ESR. During exponential growth most of the cellular free energy is used for the production of ribosomal proteins and translation. Most likely, the activation of the ESR reflects a redistribution of energy fluxes, such that the production of ribosome related RNAs and translation are temporarily down-regulated and the available energy is channeled to other processes in order to survive the stress that the cell is facing.

Regulation of specific stress responses

Specific problems ask for specific solutions; besides the general stress response, yeast cells contain a huge arsenal of transcription factors that, depending on the type of stress trigger the transcription of specific gene groups. About 200 transcription factors (or protein sequences that have a strong homology with validated transcription factors) have been identified in yeast. In the next section I will discuss three transcription factors that are used in chapters 2 to 4, respectively, and that are activated upon specific stresses. The first two (Yap1p, chapter 3; Crz1p, chapter 4) are activated as a direct consequence of the changed intracellular environment while the other (Rlm1p, chapter 2) is activated via the Slt2 MAP-kinase pathway, presumably, upon stretching of the plasma membrane.

Activation of the oxidative stress response by the gene group-specific transcription factor Yap1p

The oxidative stress response is designated as the phenomenon by which a cell responds to the imbalance between the production of reactive oxygen and a biological system's ability to readily detoxify the reactive intermediates. This can be due to the generation of reactive oxygen species (ROS) caused by the incomplete reduction of O_2 during respiration as well as to the exposure to a variety of chemicals and metal-ions. Although several transcription factors are involved in the oxidative stress response, Yap1p is the most important one. The activity of Yap1p is regulated mainly through its nuclear localization. In unstressed cells, Yap1 is freely imported into and exported from the nucleus. Upon activation by increased levels of ROS, Yap1 rapidly redistributes to the nucleus where it regulates the expression of about 70 genes. Yap1p contains a nuclear export signal (NES) sequence in the C-terminus [
]. Under non-oxidative stress conditions the exportin Crm1p recognizes the NES and exports Yap1p from the nucleus to the cytoplasm. Under oxidative-stress-inducing conditions, the interaction between Yap1 and Crm1 is inhibited and Yap1 is accumulating in the nucleus, leading to elevated expression of its target genes. Yap1p binds to the 5'-TTASTAA-3' sequence in the promoter region of its target genes. Several of the genes regulated by Yap1p are involved in glutathione metabolism, such as the genes encoding the glutathione-synthesizing enzymes, *GSH1* and *GSH2*. Furthermore, Yap1p regulates the glutathione-dependent antioxidant genes *GPX2* and *GLR1*. In general, Yap1p targets are involved in the detoxification of ROS.

Activation of the calcineurin signaling pathway by the gene-specific transcription factor Crz1p

In yeast, intracellular Ca^{2+} signaling is mediated by the Ca^{2+}/calmodulin-dependent phosphatase, calcineurin. Activation of Ca^{2+}-dependent signal transduction pathways is critical for many cellular responses. In yeast, Ca^{2+} triggers defense mechanisms under a variety of environmental stress conditions such as exposure of cells to high salt concentrations, alkaline pH, or cell wall damage. Stress conditions, that increase intracellular Ca^{2+} levels activate calcineurin, which in turn activates the transcription factor Crz1p. Dephosphorylation of Crz1p by calcineurin causes a rapid translocation from the cytosol to the nucleus, where Crz1p activates the transcription of genes whose products promote cell survival. Recent micro-array studies suggest that activation of the calcineurin pathway leads to increased expression of more than 160 genes [46]. These genes showed reduced activation in a *crz1* deletion mutant and most of them contained the Crz1p consensus binding motif (5'-GNGGC(G/T)CA-3') in their promoter regions, which suggests that they are directly activated by Crz1p. The genes that are regulated by Crz1p fall into several functional classes. Most of the genes encode products that are known or predicted to be integral membrane proteins; others are known components of the plasma membrane or cell wall. In addition, Crz1p-controlled genes code for proteins that participate in vesicle trafficking, lipid/sterol synthesis and protein degradation. Thus, in response to stress, calcineurin and Crz1p activate a specific program of gene expression that promotes remodeling of the cell surface [47].

Activation of genes involved in response to cell wall stress, via the Slt2-MAP kinase pathway

One of the sudden changes in external conditions that yeast cells may face in nature can be caused by rainfall. This changes a sugar-rich environment instantaneously into a hypo-osmotic environment. Such a change causes influx of water into the cell, which could eventually lead to lysis. Yeast cells therefore have a strong and relatively rigid cell wall. Gene regulation in response to cell wall perturbations is mediated via the transcription factor Rlm1p [48]. The activity of Rlm1p is not regulated via its nuclear localization but rather its transcriptional activity is regulated through phosphorylation by the protein kinase Slt2p (Mpk1p) [49]. The targets of Rlm1p are mostly genes that are involved in cell wall reinforcement and remodeling [48]. Rlm1p binds preferentially to the nucleotide sequence 5'-CTA(W_4)TAG-3' (W represents either an A or a T) in the upstream promoter sequences of its target genes [48, 50]. However, analysis of transcription profiles of cell wall perturbation experiments show that the motif might be refined to 5'CTA(W_4)TAGM-3' [51] (this thesis).

The activation of Rlm1p is regulated via the Cell Wall Integrity (CWI) pathway, which includes the Slt2-MAP kinase pathway. Mitogen-activated protein (MAP) kinases are serine/threonine-specific protein kinases that respond to extracellular stimuli and regulate various cellular activities such as mitosis, cell differentiation, and gene expression [52]. Extracellular stimuli that are sensed by integral membrane proteins lead to activation of a MAP kinase via a signaling cascade (**Figure 2**).

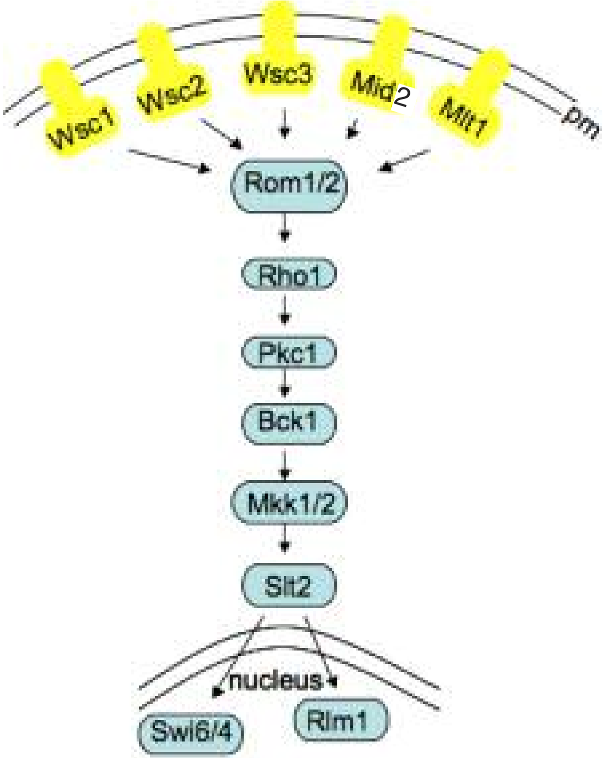


Figure 2 Schematic representation of the Cell Wall Integrity pathway. Wsc1-3, Mid2 en Mlt1 are sensors that are located in the plasma membrane. Signals are transmitted through the proteins Rom1/2, Rho1 to the MAP-kinases Pkc1, Bck1 and Mkk1/2, and finally to Slt2, which in turn activates the nuclear located transcription factor Rlm1. Besides Rlm1p, Slt2p is thought to activate Swi6/4, transcription factors that are involved in cell cycle regulation and that activates late G1-specific gene targets.

This cascade is composed of a MAP kinase kinase kinase (MAPKKK), MAP kinase kinase (MAPKK) and MAP kinase (MAPK). *Saccharomyces cerevisiae* contains 5 such MAP kinase pathways, each having a specific function that varies from response to mating, pseudohyphal/invasive growth, osmoregulation, sporulation, and to cell wall construction in case of Rlm1p [53]. The cell wall integrity or Slt2p MAP-kinase pathway contains five cell wall stress surface sensors named Wsc1-3p, Mid2p and Mlt1p that are coupled through Rom1/2 to Rho1p, which is considered to be the master regulator of the CWI pathway [54].

Among other effectors, Rho1p can activate the Map-kinase cascade via phosphorylation of Pkc1, a serine/threonine kinase. Pkc1 initiates the three-protein kinase module that starts with the MAPKKK Bck1, the redundant MAPKK Mkk1p and Mkk2p and finally the MAPK Slt2p (**Figure 2**). Besides Rlm1, phosphorylated Slt2p is thought to activate the cell cycle regulators Swi6/4p, which activates late G1-specific gene targets.

The CWI pathway can be activated by a multitude of different stimuli, like heat stress, hypo-osmotic stress, and cell wall stressing agents. Evidence is accumulating that plasma membrane stretch is the underlying physical stress that leads to activation of CWI signaling. Importantly, chlorpromazine, an amphipathic molecule that causes membrane stretch by asymmetric insertion into the plasma membrane, is a well-known activator of Slt2p [55]. Another line of evidence that supports this idea is that increasing the extracellular osmolarity can counteract the activation of the CWI pathway. This is illustrated by the phosphorylation of Slt2p by heat stress, which is prevented under hyper-osmotic conditions [56]

Functional Genomic datasets

Measuring global transcription using the micro-array technique

The classic way to measure amounts of specific mRNAs from cells has been developed by Edward Southern [57]. He developed a technique in which total cellular mRNA is isolated, run on an agarose gel for size separation, and transferred to a nitrocellulose filter for immobilization. A specific, labeled DNA probe that is complementary to the mRNA of interest is then added to the immobilized RNA. Next, the labeled probe is measured and the amount of specific mRNA can be quantified. This technique has been modified and miniaturized to what is now called a micro-array or a DNA-chip [21, 22] (**Figure 3A**). Miniaturization makes it possible to measure the abundance of the mRNA of all genes of a genome simultaneously. The DNA probe can either be a PCR product of a complete gene, or a short oligo (varying from 20 base pairs to 70 base pairs) that are complementary to a part (or parts) of a gene's coding sequence. Microarrays based on the PCR products are referred to as cDNA microarrays while those based on short oligos are referred to as oligo-based microarrays. The carrier material for cDNA microarrays is usually glass. The probes for oligo-based arrays can either be deposited on a glass slide or the probes are synthesized on the carrier material either by a photolithographic procedure (gene chips/Affymetrix) or using a modified matrix printer (Agilent).

To measure the amounts of mRNA, each mRNA is converted to cDNA, labeled with a fluorescent dye, and hybridized to a microarray slide. Specific and precise scanners have been developed that are able to detect the labeled cDNA that is hybridized to their specific DNA probe counterparts. Since the position of the specific DNA probes on the micro-array is known, the amount of labeled cDNA can directly be related to a particular gene. Micro-arrays come in two flavors. First, the two color DNA/oligo micro-array (two color DNA microarray), in which two labeled pools of mRNA are directly compared in one micro-array experiment. One part of the mRNA, for example the reference, is labeled with a red dye and the perturbed sample is labeled with a green dye. The ratio between red and green in a specific probe directly relates to the up- or down regulation of a particular gene. Affymetrix developed a system where the absolute amount of mRNA can be measured on a single micro-array (or gene chip); note that for this comparison between a reference and a perturbed sample always two chips are needed. The ratio of mRNA abundance between the reference (non-treated or healthy) and a perturbed

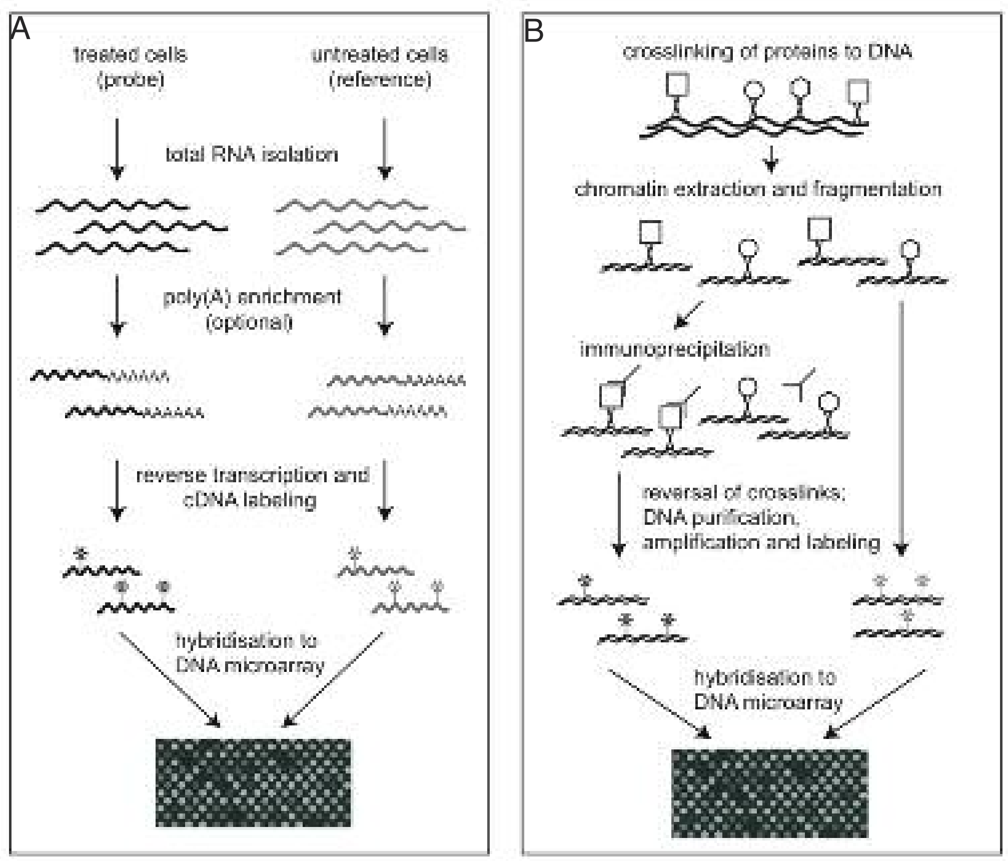


Figure 3 Global measurement of gene expression and protein DNA binding (ChIP-on-chip). (A) Global measurement of gene expression. Total RNA is isolated from treated and untreated cells. mRNA is converted into cDNA and labeled with a fluorescent label. The labeled cDNA of both treated and reference samples is hybridized to a DNA micro-array. After scanning of the micro-array the relative abundance of mRNA between treated and reference samples is obtained. (B) Global measurement of protein-DNA binding by chromatin-immune precipitation (ChIP-on-chip). Formaldehyde is used to crosslink proteins to DNA. The DNA is fragmented by shearing, and. an antibody specific for a DNA binding protein (often a transcription factor) is used to precipitate the protein and the bound DNA fragment. The enriched DNA fragments are labeled. In parallel, a reference sample is produced without the use of a specific antibody to compensate for aspecific DNA fragments. Labeled DNA from both samples is hybridized to a DNA micro-array that contains the intergenic DNA regions (figure 3, used with permission from Marijana Radonjic)

sample (gene deletion, cells that are treated with a particular compound or obtained from diseased tissue) is commonly used in this type of experiments (**Figure 3A**). Basically, micro-arrays show a snapshot of the transcriptional state of cells under a certain condition at the moment of sampling. An elegant example of the use of the microarray technique is the so-called cell-cycle experiment [58]. This experiment, using yeast cells, was set-up to measure the transcriptional behavior of genes that are regulated during the cell cycle. For this experiment, reference mRNA was isolated from exponentially growing yeast cells and was compared to mRNA isolated from cells that were synchronized (i.e. all growing in the same cell cycle phase). Every seven minutes synchronized cells were sampled to isolate mRNA. This experiment revealed that about 800 *S.cerevisiae* genes are transcriptionally regulated during the cell cycle. Furthermore, several regulatory sequence motifs controlling cell cycle-regulated genes were discovered. Another "classic" micro-array experiment is the compendium experiment conducted by Rosetta Inpharmatics [59]. In this study, the transcriptome of 280 deletion mutants, and a dozen drug treatments was profiled to create a reference database. Using this reference database, gene functions could be predicted and later verified for several uncharacterized genes. Furthermore, the method was used to characterize pharmacological perturbations, which was demonstrated by the identification of a novel target of the commonly used drug duclonine. Rosetta Inpharmatics was the first to publish a study with such a huge amount of micro-array data. Shortly after this study, two papers that studied the environmental stress response appeared almost simultaneously [40, 41]. Both studied the timing of transcriptional responses to multiple stresses like heat and oxidative stress on yeast.

Global analysis of Transcription factor binding (ChIP-on-chip)

Chromatin-Immuno Precipitation (ChIP-on-chip) experiments are used to globally determine the localization of DNA-binding proteins (mainly transcription factors) on DNA [60]. Since transcription factors mainly bind to the areas between genes, the DNA sequences of the intergenic regions were printed on chips, however, later versions of ChIP-on-chips contain both genes and intergenic regions. A ChIP-chip experiment starts by cross-linking DNA binding proteins to genomic DNA *in vivo*. The chromosomal DNA is then fragmented and the transcription factor (or another DNA binding protein) of interest is immuno-precipitated using a specific antibody (**Figure 3B**). The protein of interest can also be tagged with a myc-epitope tag, allowing antibodies against this tag to be used for immuno-precipitation. The DNA that is bound by the protein will subsequently be precipitated. In parallel, a reference sample is processed through the crosslinking and the fragmenting steps but not the immunoprecipitation step. DNA isolated from both samples is labeled and hybridized to a micro-array slide containing the intergenic DNA regions. The labeled DNA parts, bound to the immunoprecipitated transcription factor will be enriched for promoter regions of its target genes. The data from ChIP-on-chip experiments are particularly useful to elucidate binding properties of transcription factors. In 2002, Lee *et al*. [61] used the technique to elucidate the target genes of 103 transcription factors; later, Harbison *et al*. [62] expanded this, not only by the number of transcription factors but also by introducing a number of specific stress conditions under which the experiments were performed. The data from both experiments were used to create a map of the transcriptional regulatory network that describes potential pathways that yeast cells can use to regulate gene expression programs. Furthermore, their network map reveals that gene expression programs and cellular functions are highly connected through networks of transcriptional regulators that regulate other transcriptional regulators. In summary, this technique is very useful to discover *in vivo* the target genes of transcription factors or other DNA binding proteins. In this thesis we describe a method in which we use ChIP-on-chip data to analyze transcriptional profiling micro-array data.

Fitness profiling

The microarray design is also used in parallel growth experiments of a genomic library of deletion strains. These experiments examine the relative fitness of homozygous deletion strains (in case of nonessential genes) and heterozygous deletion strains (in case of essential genes) grown in the presence of an inhibitor. The homozygous set of yeast deletion mutants is a collection of all non-essential yeast genes (~4800). Each of these knockout strains is designed in such a way that the deletion leaves a ´bar code´ mark of two 20-nucleotide sequences. The bar code can be used to measure the abundance of each deletion strain [17]. To this end, the chromosomal DNA of the deletion mutants, grown in the presence of an inhibitor is collected and the DNA containing the bar code is amplified. This DNA is labeled and hybridized to a microarray that contains the complementary DNA fragments to the bar code for all individual deletion mutants. Similar experiments have also been performed by growing the individual strains on agar plates and measure their colony size [63]. Fitness profiling is used to identify pathways that buffer the cell against the effect of stresses or compounds, and thereby provide clues about their mode of action [64]. Both gene expression profiling and fitness profiling experiments are used to assess the effect of stresses and compounds. Giaever *et al* [8] measured the relative fitness of the deletion strains grown under 6 different conditions. This revealed genes that are necessary for optimal growth under these conditions. More importantly, the authors showed that only 7% of the genes that exhibit a significant increase in messenger RNA expression are also required for optimal growth in four of the tested conditions. Recently, Zakrzewska *et al*. [65] used the same approach to study the transcriptional response and global fitness of yeast cells to the plasma membrane perturbant chitosan. Again, only a partial overlap was found between the transcriptional response of the individual genes and the fitness of the individual deletion strains. However, analysis of gene expression and fitness profiles at the pathway level showed a much larger degree of similarity.

Genome sequence comparison of related species

The ChIP-on-chip method describes a sophisticated way to discover transcription factor-gene interactions. A drawback of this method is the high costs of the experiments and technical problems that make it difficult to find the true target genes of all transcription factors under all conditions. For example, the ChIP-on-chip experiments performed by Harbison *et al*. [62] do not properly detect the target genes of the Heat Shock Factor (Hsf1p) in cells confronted with heat stress. This might be caused by poor protein-DNA crosslinking at elevated temperatures. Since transcription factors bind to specific DNA sequences, the binding properties of transcription factors and the regulation of their target genes are coded in the DNA sequence. A powerful method to find regulatory sequences in genomes is to compare the promoter sequences of functionally related genes. In this way, a number of regulatory motifs have been discovered [43]. The rapidly increasing speed and rapidly decreasing cost of genome sequencing allow a new approach in the quest for new regulatory motifs. Nowadays, the genome of *Saccharomyces cerevisiae* can be compared with that of closely related species [66, 67]. Functional regions in the genome such as DNA regulatory motifs are likely to tolerate a lower mutation rate compared to non-functional regions. Besides identifying regulatory motifs, this method also allows for a more accurate identification of open reading frames. Kellis *et al*. [66] discovered that about 500 ORFs that were earlier annotated as genes, might actually be dubious genes (**Figure 4**). In addition, 72 regulatory motifs were identified, including most known *cis*-acting sequences and numerous new motifs. The same approach has also been applied to the genomes of mammals (human, mouse and rat) that contain, however, much larger intergenic DNA regions, which makes it much more difficult to find true regulatory sequences.

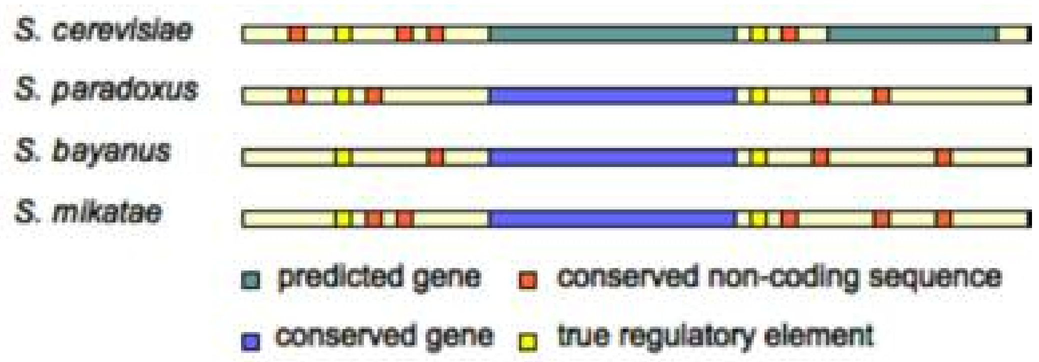


Figure 4 Sequence comparison of four related *Saccharomyces* species. Schematic representation of the sequence comparison of four related *Saccharomyces* species. Genes and non-coding sequences that are found in all four species are representing true genes and regulatory elements.

Functional classification of genes based on biological pathways or gene ontologies

Classical biochemical experiments provide functional information about genes, which is indispensable for analyzing micro-array data. These data have been stored in various databases that attempt to classify genes and proteins according to function. The KEGG (Kyoto Encyclopedia of Genes and Genomes) database is a Japanese initiative that classifies genes in metabolic pathways, basically by their homology to well characterized metabolic genes. KEGG also provides metabolic maps that are visual projections of a metabolic pathway. The Comprehensive Yeast Genome Database (CYGD; http://mips.gsf.de/genre/proj/yeast/) of the Munich information center for protein sequences aims to present information on the molecular structure and functional network of *S. cerevisiae*. This database classifies genes according to protein function, localization, presence in protein complexes, and cellular phenotype.

The most ambitious gene classification initiative is the Gene Ontology (GO;
) project that provides a controlled vocabulary to describe gene and gene product attributes in any organism. A controlled vocabulary is used in the form of a carefully selected list of words and phrases, which are used to tag units of information so that they may be more easily retrieved by a search. The Gene Ontology is split into three related ontologies: **Molecular Biology,** which describes the molecular function of a gene product; **Biological Processes,** which describes the role of gene products in multi-step biological processes, and **Cellular Components,** which describes their localization to various cellular components. Two important goals of the GO consortium are the development and maintenance of the ontologies themselves and, second, the annotation of the genes and gene products. This latter aspect can be fol-

lowed using the metadata that every GO association contains. These metadata indicate who made an annotation, the date of this annotation, and most importantly, on which type of evidence an annotation is based. Especially, the last set of metadata is important. GO uses evidence codes for this: for example, IEA means Inferred from Electronic Annotation, and is used when the output is obtained by *in silico* analysis. IDA, Inferred from Direct Assay, is used when evidence is derived from biochemical evidence like enzyme assays or binding assays. As we will see in de remainder of this thesis these data-sets are important for understanding and analyzing other kinds of high-throughput datasets such as micro-array data and fitness profiling data.

Analysis of high-throughput data sets

The need for bioinformatics

Genomics and functional genomics have transformed research in molecular biology from a relatively data-poor discipline into one that is data-rich [68]. A single yeast micro-array experiment measures the expression values of about 6000 genes. An experiment that measures the effect of a particular compound on transcription, that is performed in triplicate while using five time points will result in 90.000 data points [69]. However, this is just a modest example. Recently, an experiment was published that described more than 450 measurements using human micro-arrays, which contains each about 20.000 genes, leading to almost 10 million data points [70]. Such large datasets have driven the development of data-mining techniques to reduce the complexity of the data and to discover meaningful and useful patterns in and relationships between data [68]. Since the development and popularization of high-throughput techniques, the field of bioinformatics has grown explosively.

Besides the problem of handling large datasets, micro-arrays are complex experiments. An experiment may start with the growth of a cell culture, mRNA is extracted, converted into cDNA, labeled, and hybridized to a micro-array, and, finally, arrays are scanned and analyzed. At every stage, variation and error will be introduced. In general, the noisy nature of micro-array data is largely caused by biological variations, which correspond to real differences between the response of different biological replicates, and by experimental noise. A more or less related problem is the comparison of data derived from different laboratories and microarray platforms. Recently, two standard RNA samples were compared between seven laboratories and 12 different microarray platforms [71]. The conclusions from this study were that for most platforms within a single laboratory the reproducibility was generally good, but that the reproducibility between platforms and across laboratories was generally poor. Despite all these problems, researchers are making progress in extracting biological insights from transcriptomic data.

Gene expression data basically provide three different types of information. First, expression profiles can be used to infer information about the regulation of gene transcription. Second, a transcriptome can be regarded as a snapshot of the physiological state of the cell. Functionally related genes that respond in concert may provide information about the functional changes of a cell. Third, expression profiles may indicate the function of uncharacterized genes, for example, when their expression is strongly correlated with the expression of genes with a known function.

Data analysis techniques

Analysis of multiple microarray experiments by cluster analysis

Transcription profiles provide data about the changes in mRNA abundance for every gene that is measured. Usually, this is represented as the base-two logarithm of the fold-change between the tested condition and the control. One of the first methods that was used to analyze transcriptomic datasets is cluster analysis (**see box I**) [72]. For this type of analysis it is necessary to measure gene expression in multiple experiments. Eisen *et al*. [72] who made clustering software freely available together with a visualization tool that displayed the clusters in heat maps have popularized hierarchical clustering.

Box I – Hierarchical Cluster Analysis

Cluster analysis of microarray data starts with calculating a similarity measure of the input data. The input data is often the differential expression of genes measured under a set of conditions. A similarity measure that is widely used is Euclidian distance:

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

where the distance between vector x (for example, gene A measured under a set of conditions) and vector y is d(x,y).A dozen other similarity measures can be used like for example Pearson correlation. After the distance matrix is obtained the actual clustering starts. By using a Euclidian distance a small value in the distance matrix implies that these two clusters/objects (defined by row and column number) are more similar than clusters/objects with greater value. When performing clustering, the same matrix is scanned for the lowest value, which should be the smallest distance between two clusters. The cluster defined by the row (i) of the smallest element d(i,j), and the one defined by the column (j) are then merged. The result of such merge is a new cluster containing both of the merged elements. The two merged clusters are then removed from the distance matrix and the new cluster is added. The distance between elements in the matrix is defined by one of the three most used linkage methods:

Single linkage: The distance between two objects is defined to be the smallest distance possible between them. If both objects are clusters, the distance between the two closest members are used.

Complete linkage: This method is much like the single linkage, but instead of using the minimum of the distances, we use the maximum. So for clusters, the distance between the two farthest members are used.

Average linkage: Basically, this method takes the mean between all the members in cluster i to all the members in cluster j.

Single linkage tends to form a few big clusters early on in the hierarchy whereas complete linkage tends to form many smaller clusters. Average linkage can be a good compromise between the extremes of single and complete linkage

Box I continued:

Below a clustering example using single linkage is demonstrated:

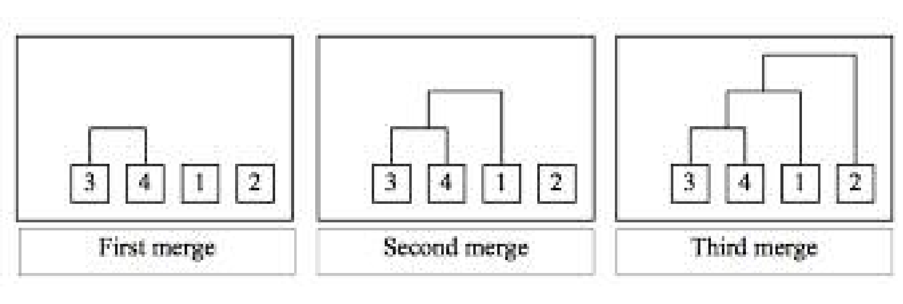
1	gene 1	gene 2	gene 3	gene 4
gene 1	0			
gene 2	6	0		
gene 3	3	5	0	
gene 4	4	4	2	0

2	gene 1	gene 2	gene (3&4)
gene 1	0		
gene 2	6	0	
gene (3&4)	3	4	0

3	gene 2	gene((3&4)&1)
gene 2	0	
gene((3&4)&1)	4	0

4	gene(((3&4)&1)&2)
gene(((3&4)&1)&2)	0

There are four elements we want to cluster (table 1), the numbers in the table are a measure of similarity (distance matrix). In the first iteration, we search the matrix for the smallest element and find that this is the combination of gene 3 and 4. These two elements are merged in table 2, and because of our selection of linkage (single) the distances that are smallest to the other elements are kept. For example, the distance between our new cluster and gene 1 is the smallest of the values 3 (gene 1- gene 3) and 4 (gene 1- gene 4), which are 3. This operation is repeated in table 3, where the merged element from the last iteration is merged with gene 1. This procedure continues until all genes are merged into one cluster (table 4). When drawing the dendrogram (see below), this final cluster will be the root of the tree. Dendogram for this example:



(The example above is adapted from “The introduction to J-Express” by Bjarte Dysvic)

Spellman *et al*. [58] used hierarchical clustering to identify clusters of genes that showed similar transcriptional regulation during the different phases of the cell cycle. Both Tavazoie *et al.* [73] and Spellman *et al.* [58] integrated genomics data with transcriptional data to identify *cis*-regulatory sequences in the promoters of tightly co-expressed gene clusters. Gene clusters also tend to be enriched for genes with similar function. Troyanskaya *et al*. [74] used this knowledge to infer a functional role for unknown genes in the same cluster. Although hierarchical clustering is widely used and has proven its usefulness, this method has a couple of shortcomings. First the classical hierarchical clustering method only allows genes to be assigned to a single cluster. Furthermore, to obtain reliable clusters most datasets require pre-selection or filtering; if all genes are used (including non-differentially expressed genes, which often form the majority of the dataset) results tend to be sub-optimal. Finally, cluster analysis requires multiple experiments as input; results of single experiments where the effect of a gene deletion or over-expression is examined cannot be analyzed in this way. For this, other techniques have been developed that apply statistical methods to single gene transcription profiles.

Analyzing gene expression profiles by Gene Class Scoring techniques

A very simple but effective way to analyze single gene expression profiles is to first identify significantly up- and down-regulated genes and subsequently characterize these sets in terms of enrichment for functional annotation [75]. When experiments have been repeated several times, statistical approaches such as the t-test and ANOVA can be applied to obtain a list of significant, differentially regulated genes [76]. Alternatively, gene lists can be produced by simply applying cut-offs based either on an assigned p-value or on a specified level of up- or down-regulation of genes. Such gene-lists can then be tested for overrepresentation of genes involved in a specific pathway or sharing a specific GO annotation. There are several statistical methods that can establish if there are more matches than expected by chance. The most commonly used are the hypergeometric distribution (**see box II**), the closely related Fisher's exact test, and the chi-square binominal distribution, the latter being more appropriate for large gene sets [77].

Box II - Scoring over-representation of predefined gene sets

Suppose that we want to know whether a specific set of genes of interest is statistically enriched for genes with a specific annotation in Gene Ontology. In this case, both features (namely, "does the gene belong to the set of genes of interest" and "is the gene associated with GO term X") are categorical, and the appropriate statistic is the overlap between both gene sets. If the two sets are chosen randomly and independently, the average overlap will be:

$$(x) - \frac{ab}{n}$$

where a is the total number of genes in set A, b the total number in set B (f.e. GO category). n is the total number of genes measured and x is the overlap between A and B. This makes sense: if a fraction b/n of all genes belongs to set B then the expected fraction of genes in set A that also belongs to set B equals x/a. In the case of over-representation, when x > (x), the exact p-value quantifying how likely it is we would get at least the same number of overlapping genes by chance, is given by:

$$P_{over}(x) - \sum_{x'=x}^{min(a,b)} H(x'|a,b,n)$$

where the hypergeometric distribution is given by:

$$H(x|a,b,n) = \frac{\binom{a}{x}\binom{n-a}{b-x}}{\binom{n}{b}} \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k!)}$$

This use of the cumulative hypergeometric distribution is also known as "Fisher's exact test." The test is by nature non-parametric because both input features are non-parametric. Under specific conditions the hypergeometric distribution may be approximated by the binomial or chi-square distribution.

(Adapted from Bussemaker et al., 2007)

Since the introduction of these methods, they have been applied in numerous bioinformatics tools. Pathway Processor [78] uses the KEGG pathways, and Onto-express [75] used Gene Ontology gene groups to score the overlap in gene content. Funspec [79] is a tool that applies the hypergeometric distribution to a great number of gene sets, varying from gene sets based on GO annotations, to gene sets that are based on phenotypic features. In principle, every gene set can be used; for example, genes that share a particular cis-regulatory motif in their promoter region can provide information about regulatory networks. As in most cases, multiple gene sets are tested simultaneously, the statistical outcome has to be corrected for this. Most tools use the Bonferonni correction or the less stringent False Discovery Rate (FDR) to correct for multiple testing (see **box III**).

Box III – Bonferroni correction

The Bonferroni correction is a safeguard against multiple tests of statistical significance on the same data. For example, when a p-value of 0.05 is used, 1 out of every 20 hypothesis-tests will appear be significant purely due to chance. The Bonferroni correction states that if an experimenter is testing n independent hypotheses on a set of data, then the statistical significance level that should be used for each hypothesis separately is 1/n times what it would be if only one hypothesis were tested. A less stringent method to correct multiple testing is the False Discovery Rate (FDR). The FDR of a set of predictions is the expected percent of false predictions. For example a method returns 100 genes with a false discovery rate of .3 then we should expect 70 of them to be correct.

(Adapted from wikipedia)

One of the disadvantages of these methods is that in order to create gene lists, (arbitrary) cut-offs often have to be applied. Furthermore, most gene lists take only those genes into account that are highly differentially regulated, and tend to ignore genes that are differential expressed with low amplitude. However, genes from some pathways can be tightly co-regulated with only minor changes in gene expression. As mentioned before, measurements of gene expression of individual is rather noisy, and thus analyzing gene expression at the level of gene sets organized in pathways might be beneficial. This can be explained by the following example; assume that we have a transcriptome with an error rate of 20% with respect to the fold-change of individual genes. In addition, let us assume that in a specific metabolic pathway all 100 gene members are up-regulated by 10%. These genes are then most probably not selected as being induced since their level of differential expression is well within the noise of individual genes. However, the error in the average expression of 100 randomly chosen genes will be on the order of 20%/√100 = 2%. The 10% change in expression at the level of the whole pathway therefore corresponds to five units of standard error and is highly statistically significant [80]. Recently, methods have been developed, that do not need a list of selected genes as input but that take into account the ratios of all genes in combination with the use of predefined gene sets. Most of these methods use a statistical test that compares the expression values from a predefined gene set with the remaining genes. One of the first examples of such a method was named Quontology [81]. This tool uses the z-value to represent the difference between the average expression in a gene set and the genome mean. This is reflected in the formula:

$$z = \frac{x_s - \mu}{\sigma_x}$$

Where X_S is the average expression of the gene set, and m is the genome-wide mean and finally $\sigma_x = \frac{\sigma}{\sqrt{n}}$ is the standard error of the mean, being the standard deviation of the genome wide distribution of log-ratios. Kim and Volskey [82] have used the same approach, which is named PAGE (Parametric Analysis for Gene Enrichment).

For larger gene sets more accurate results are obtained when the unpaired t-test for the difference between means is applied. The unpaired t-test compares the difference between the expression mean of a particular gene-set from a transcription profile with that of the remaining genes. (**Figure 5**). In this thesis this method (T-profiler) is explored to analyze gene expression profiles [83]. Just like the z-value approach, T-profiler yields a significance score that can be used for cluster analysis on the pathway level.

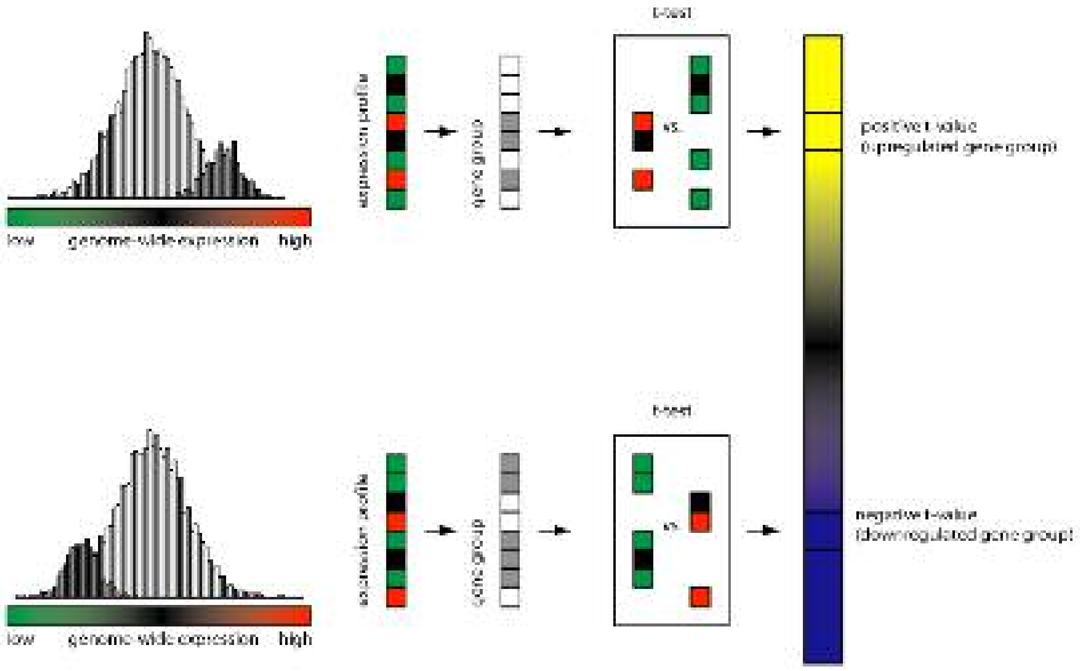


Figure 5 (page 28) Schematic representation of pathway analysis of gene expression data using T-profiler. Genes from a pre-defined gene group such as those annotated by the Gene Ontology project are combined with non-thresholded genome-wide expression data to derive a statistical measure of pathway-level activity. A pre-defined gene group (dark gray) is scored using a t-test for its expression response compared to all other genes. The obtained t-value can either be negative or positive dependent on the up- or down regulation of the corresponding gene group.

Comparison of Fisher’s exact test and T-profiler

To show the difference in performance between the Fisher’s exact test and the unpaired t-test we analyzed an expression profile where Lovastatin-treated cells were compared with untreated cells. Lovastatin is known to inhibit one of the enzymes of the ergosterol biosynthesis pathway. In response to this treatment, cells induce the ergosterol biosynthetic genes [84]. **Figure 6** shows the effect on the P-value of the Fisher’s exact test of various fold-induction thresholds on individual genes. In this example, the optimal P-value from Fisher’s exact test is slightly smaller than that obtained with the unpaired t-test. Despite this, T-profiler has the considerable advantage that no threshold needs to be optimized. Commonly, a threshold of 2-fold (equals a $\log_2$-ration of 1) induction or repression is used; at these values, T-profiler performs much better.

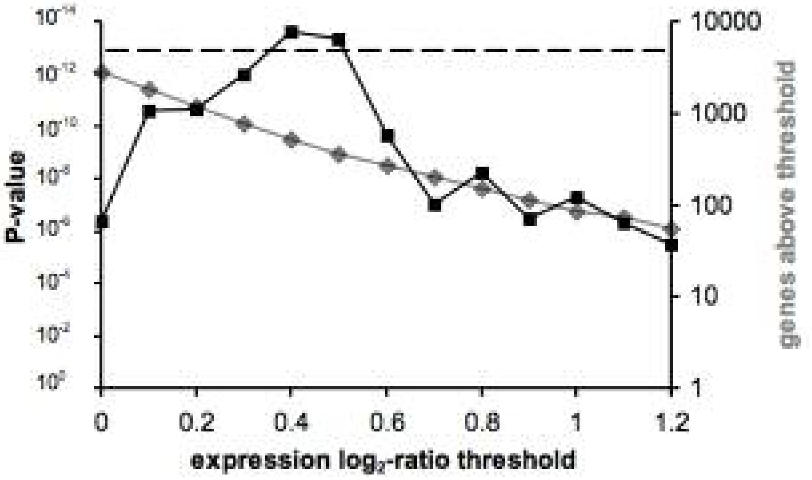


Figure 6 Scoring GO categoriesL Fisher's exact test versus T-profiler
We analyzed gene expression data for the response to the ergosterol biosynthesis inhibitor Lovastatin as measured by Hughes *et al.,* [59]. The two-sample t-test reveals that the mean expression level of genes in the GO category "ergosterol biosynthesis" is significantly higher than expected (broken line; t = 7.4; P = 1.1·10-13). Fisher's exact test can be used to score over-representation of the same GO category in the set of most induced genes. However, this requires one to first define a threshold for the expression fold-change of individual genes. The solid line shows how the P-value (black squares) from Fisher's exact test depends on this threshold. The grey line shows the number of genes after applying the threshold.

A couple of other statistical tests are used to detect differential expression of gene sets based on the distribution of expression values. Currently, Gene Set Enrichment Analysis (**GSEA see box IV**) is one of the most frequently used methods. Originally, only the non-parametric Kolmogorov-Smirnov (KS) statistic was used [85] to test whether the ranking of expression levels in a specific gene set is different from the ranking of all genes, but later the approach was modified to work more reliably [86]. In addition, a non-parametric version of the t-test, the Wilcoxon-Mann-Whitney test is also used [87].

Linear Regression Models: REDUCE

All the methods described above treat genes in a categorical way: genes are either part of a gene-set or not. This assumption might not be the most appropriate way to describe transcriptional regulatory pathways. Transcription of a gene might depend on the number of *cis*-regulatory motifs in the promoter or the precise nucleotide sequence of a motif might influence the binding affinity of a transcription factor. Also interaction with co-factors, combination with other transcription factors, the distance of the *cis*-regulatory motif from the transcriptional start site and the presence of chromatin might influence the binding. Bussemaker *et al*. [88] was one of the first who integrated information related to binding affinity in a method that is able to dis cover new *cis*-regulatory motifs from gene expression data.

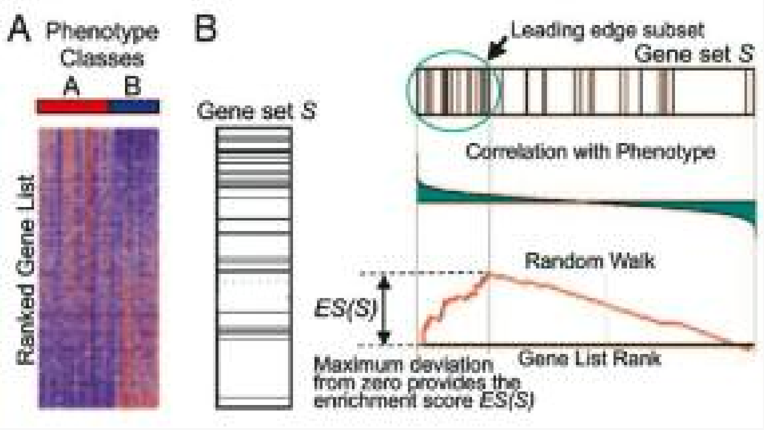
This method, named REDUCE, is based on a model in which upstream motifs contribute additively to the differential transcription of a gene. REDUCE uses the following linear regression model

$$Ag = C + F\,Ng$$

where C is the baseline expression level, which is the same for all genes, F represents the slope that estimates the change in TF activity, and the variable Ag (response) is the mRNA expression log-ratio of gene g. The independent (predictor) variable Ng represents the regulatory network connectivity between the TF and the promoter region of gene g. In REDUCE the predictor variable is the number of occurrences of a regulatory motif in the promoter region of a gene. In principle, every parameter that quantifies the strength of the binding of a transcription factor can be used. Gao *et al*. [89] used the log-ratio values of the TF binding ChIP-chip data as input parameter and Foat *et al*. [90] strongly improved the original REDUCE tool by using a binding matrix of a sequence motif as the predictor variable. Thus the better the estimate of the predictor factor, the better the global change in TF activity may be explained [91]

Box IV – Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) is a popular microarray data analysis method that uses predefined gene sets and ranks of genes to identify significant biological changes in microarray data sets (Subramanian et al., 2005) GSEA is especially useful when gene expression changes in a given microarray data set is minimal or moderate The primary result of the gene set enrichment analysis is the enrichment score (ES), which reflects the degree to which a gene set is over-represented at the top or bottom of a ranked list of genes (see figure). The enrichment score is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov–Smirnov-like statistic.



GSEA overview. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags,” i.e., location of genes from a set S within the sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.

(adapted from Subramanian 2005)

Outline of this thesis

The scope of the thesis is the development, implementation and use of transcription data analysis tools. At the start of this project only a limited number of data analysis tools were available. The most important one available was cluster analysis, which has some major shortcomings. Cluster analysis needs multiple experiments as input but more importantly it requires a (arbitrary) cut-off of the genes that are used as input. The most important problem with cluster analysis is the translation to biological interpretation. The tool we want to develop should have the following requirements: (a) no need for arbitrary cut-offs; (b) the tool should be able to analyze single transcriptomes, and (c) the interpretation of the data should be straightforward. Furthermore, since at that time many datasets of interest became available, it would be highly beneficial if the tool could be used to compare data from different platforms and laboratories. **Chapter two** describes the development and implementation of T-profiler. T-profiler scores changes in the average activity of predefined gene groups like based on categorization in Gene Ontology, TF binding (ChIP-chip) data, and genes that share a particular *cis*-regulatory motif. In **chapter three** the analysis of the global transcriptional response to the cell wall perturbants Calcofluor white and Zymolyase is described. This analysis indicates that cell wall stress results in activation of various pathways, including the cell wall integrity pathway. From the analysis of large-scale gene expression profiles using T-profiler (described in **chapter four**) a database (T-base) has been built. Predictions from T-base were validated by transcription factor localization studies. Furthermore, important predictions about the function(s) of the PAC and rRPE motifs are proposed, and from the construction of a co-modulation network of transcription factor activities, a new role for Cin5p is predicted. Finally in **chapter five**, we use T-profiler in combination with correlation analysis, to make functional predictions for uncharacterized genes based on gene expression data and on fitness profiling data.