Chapter 1

Introduction

An organism is viable or functional by expressing a certain set of proteins in a given environment. Regulation of such protein expression is, thus, critical for an organism's survival. One such regulation is executed at the level of transcription according to the central dogma of molecular biology (Alberts et al., 2002).

Traditional biology in experimental laboratories has been mostly focused on small-scale investigations of molecular mechanisms and structures. On the other hand, apart from genome sequencing projects (Hubbard et al., 2005) which allow us to identify raw material for organisms, recent technological advances have been producing a huge amount of high-throughput genome-/proteome-wide data to infer knowledge of biological processes and functions on a large scale. They include mRNA expression (Ball et al., 2005; Parkinson et al., 2005), yeast two-hybrid protein-protein interactions (Ito et al., 2001; Uetz et al., 2000) and protein-DNA interactions (Harbison et al., 2004; Lee et al., 2002).

Due to the amount of data generated from experiments, computational and statistical analyses have naturally become an appropriate and efficient way for such large-scale data analysis in the name of computational biology or bioinformatics. In particular, much effort have been driven toward understanding transcriptional regulation as key cellular responses to diverse environmental stimuli because relevant data are readily available such as gene expression and protein-DNA interactions mentioned above. Recent studies based on such large-scale data have found that transcriptional regulation itself may significantly contribute to phenotypic differences across species (Borneman et al., 2007; Kruglyak and Stern, 2007; Odom et al., 2007). The amount of data is still rapidly increasing with better quality and completeness. Accordingly appropriate bioinformatic analyses are required as an essential tool in biological research. In this thesis, we would like to contribute to such an endeavour by developing computational and statistical methods to investigate several salient issues of transcriptional regulation using heterogeneous large-scale data sources in a model organism, *Saccharomyces cerevisiae*. Below we first provide some general and essential background for our work.

1.1 Gene regulation

1.1.1 General aspects

The biological phenomenon we aim to investigate in this thesis is transcription. Transcription is the process where a DNA segment or a gene is transcribed into a messenger RNA before its becoming a final product for functionality (Alberts et al., 2002). Many regulatory proteins called transcription factors (TFs) are involved in regulating such gene expression. In simple organisms like yeast, they bind DNA segments upstream of their target genes (towards the 5’ end of the strand in question) and recruit RNA polymerases to initiate transcription of the genes. They may equally prevent the polymerases from initiating transcription. TFs are taken to be specific for genes, meaning that each of them binds particular short regions of DNA (or binding sites; about 5 to

15 base pairs) to regulate its target genes. Searching for TFs regulating genes of interest or genes regulated by TFs of interest is hence an important task. Recent DNA microarray technologies have now become a standard way to investigate such issues as we will describe in the next section.

Given a single TF, one often finds different DNA binding sites which show similar patterns. They are collectively described by representative sequences called DNA binding motifs or consensus sequences. This leads researchers to investigate DNA motifs themselves as another factor for gene regulation, what is called *cis*-regulation in contrast to *trans*-regulation by TFs. For instance, a single TF may regulate two groups of genes with a similar function each, and the two groups possess two distinct (over-represented) binding patterns or motifs respectively. This implies that the differential regulatory information lies in the distinct DNA motifs rather than the same TF. In general, collections of those distinct DNA motifs may contain different logics or codes for differential gene regulation because cis-regulation does not necessarily equal trans-regulation (Istrail and Davidson, 2005). Those patterns may be also represented by two distinct matrices, called position specific frequency matrices or PSFMs, corresponding to the two binding patterns. These PSFMs can be utilized to search for a new binding site for other genes and several databases like TRANSFAC (Matys et al., 2006) are available for those matrices. Recently, the genome-wide protein-DNA interaction study by Harbison et al. (2004) inferred a transcriptional regulatory code, if it exists, in yeast by examining binding patterns or motif organizations in diverse cellular conditions. Numerous other *in silico* studies on this motif discovery or search have been carried out by examining groups of genes which show e.g. similar functions or expression patterns. Motif analysis is not our main focus of study in this thesis and the interested reader is referred to review articles by e.g. D'haeseleer (2006) and Das and Dai (2007).

Regulatory DNA motifs are also investigated in relation to evolutionary conservation, which yields a phylogenetic footprinting method in comparative genomics (Cliften et al., 2003; Kellis et al., 2003). Phylogenetic footprints are conserved motifs in promoter regions of orthologous genes. The underlying assumption is that conserved motifs carry their functionality as regulatory elements. However, conservation alone is insufficient to detect functional signals because evolutionary divergence may have caused different TF binding events in the presence of conserved motifs across species (Borneman et al., 2007).

In higher organisms like mammals, transcription is a more complex process, so that genes may be regulated by TFs which bind DNA regions located downstream of the genes. We even do not mention here other important factors such as chromatin structures, histone modifications and methylation patterns, which we have no good knowledge of yet and are not the object of study in this thesis. Our main concern is to investigate regulatory relationships between TFs and genes given experimental data we introduce in the next section. For more comprehensive descriptions of basic biology, the reader is referred to standard textbooks such as Alberts et al. (2002).

1.1.2 Modular organization of biological systems

Modular organization of biological systems has been argued to be fundamental in understanding an organism's functionality (Hartwell et al., 1999). No single molecule in the cell can perform any function by itself. Biological molecules are constantly interacting with one another to perform function. Modules are groups of biological objects such as genes, proteins or DNA binding sites which show common properties within individual groups. Groups of genes are co-expressed, groups of proteins like TFs interact with each other, groups of DNA binding sites are clustered in intergenic regions. Hence, assuming that modularity is the key in biology, the ultimate goal is to understand what functions those autonomous modules perform and how they are related with one another in order to realize various phenotypes.

Transcriptional regulation itself has been extensively studied in the context of modular organization (Alon, 2003; Bar-Joseph et al., 2003; Ihmels et al., 2002; Lee et al., 2002; Milo et al., 2002; Segal et al., 2003; Tanay et al., 2004). Along with a large amount of genome-wide data, computational approaches have focused on identification of underlying modules. For instance, Ihmels et al. (2002) developed an algorithm which derives transcriptional modules of co-regulated genes and associated conditions using an expression dataset of more than 1000 conditions. In particular, their algorithm was able to assign genes to multiple modules. Such assignment is biologically meaningful as genes may have multiple functions but was not possible in traditional clustering algorithms. They also inferred functions of about 900 uncharacterized genes from their modules, two of which were experimentally verified. Furthermore, relationships among modules through associated conditions yielded a global view over transcriptional networks. Similarly, other computational works have generated module networks using diverse experimental data and different strategies (Bar-Joseph et al., 2003; Segal et al., 2003; Tanay et al., 2004). However, those modules or networks are usually not validated as a whole and also difficult to be compared because they are generated from noisy experimental data in the first instance. Yet, modules are very useful objects of study, even if its reality is questionable, when one tries to understand the complexity of biological systems on a global level since the reduction of complex systems becomes easier.

1.1.3 Combinatorial regulation

Combinatorial regulation is regulation of individual genes coordinated by multiple TFs, which may be viewed as another modular structure in transcription. Several TFs are all necessary to transcribe their target gene in a certain condition. This combinatorial or cooperative feature in transcriptional regulation has not been elucidated in great detail. Above all, we need to know a complete list of factors that bind DNA to regulate their common target gene, which is unknown in most cases.

Cooperative activity or DNA binding has been reported in many experimental studies (Barbaric et al., 1996; Ogata et al., 2003) and also predicted computationally using large-scale data (Banerjee and Zhang, 2003; GuhaThakurta and Stormo, 2001; Kato et al., 2004; Pilpel et al., 2001). For instance, in an experimental study by Zeitlinger et al. (2003), it was shown that a transcription factor, Ste12, regulate potential target genes in a program-specific way (mating or filamentation) dependent on a partner transcription factor, Tec1. On the other hand, their precise binding mechanism was not known. Another experimental study by Martin et al. (2004) showed that Fhl1 interacts with Ifh1 to activate ribosomal genes in wild-type cells under a rich media condition, whereas Crf1 competes with Ifh1 for binding to Fhl1 upon TOR inhibition. They also showed that Fkh2 interacts with Ndd1 cofactor to activate transcription of cyclin genes. Computational efforts have been also made in parallel using large-scale data which we will introduce in the next section. Pilpel et al. (2001) identified synergistic DNA motif pairs, without direct TF binding evidence, using gene expression data by requiring that genes with both motifs show significantly higher coherence of expression patterns than genes with either motif alone. Banerjee and Zhang (2003) applied the method of Pilpel et al. (2001) to genome-wide protein-DNA interaction data to identify synergistic or cooperative TF pairs, but without providing DNA motif informations. Kato et al. (2004) used genome-wide protein-DNA interaction data to identify over-represented motifs and then motif combinations which give rise to coherent expression patterns. In addition, they were able to assign TFs to those identified motif combinations. Despite all these small-scale experimental and large-scale computational efforts, revealing mechanisms of combinatorial regulation remains a difficult problem to resolve because of either experimental limitations or data noisiness. Therefore, we would like to explore the issue from a different perspective in this thesis.

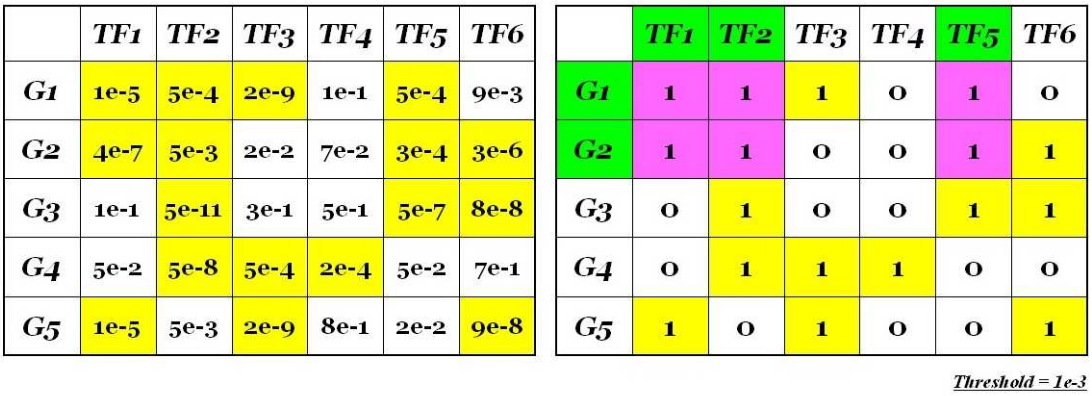
1.2 Large-scale experimental approaches

1.2.1 Protein-DNA interactions

As we mentioned above, a general picture of gene regulation is that TFs bind specific DNA regions to induce or repress transcription of downstream genes by RNA polymerases. There are both specific and general TFs and this simple picture becomes more complicated in higher organisms like humans. More diverse and complex mechanisms are thought to have evolved to control more entangled cellular networks in those higher organisms with a larger number of molecules. As a first step to decipher mechanisms of transcription, we are interested to know which TFs are bound to DNA to regulate transcription of a specific gene. Hence, this protein-DNA interaction event is an essential information for investigation of transcriptional regulation.

Experimental and computational studies of protein-DNA interactions have been mainly driven by efforts to understand the cell at an organismal level. Genome-wide experiments are now common and chromatin immunoprecipitation on microarrays, called ChIP-chip or genome-wide location analysis, have been used to study protein-DNA interactions and transcriptional networks (Harbison et al., 2004; Hawkins and Ren, 2006; Iyer et al., 2001; Lee et al., 2002; Ren et al., 2000). ChIP-chip methods purify *in vivo* DNA fragments of 200 to 1000 base pairs in length bound to a particular TF of interest by immunoprecipitating chromatin (TF-DNA-histones complex) using an antibody of that specific TF. In parallel to those enriched experimental sequences, one also purifies a sample of control DNA sequences by mock immunoprecipitation and then co-hybridizes the two samples of DNA sequences labelled by two fluorophores to a microarray of intergenic regions (DNA regions between neighbouring genes) in the genome in question. Normally the experimental sequences are dyed in red (R or Cy5) and the control sequences in green (G or Cy3). Then, the significant increase of the intensity of red light relative to that of green light on a certain spot implies that the corresponding intergenic region is bound by the TF of interest. An error model can be applied in order to deal with low-intensity spots which are a source of noise in data and to assign p-values to R/G ratios for binding confidence. In yeast, the neighbouring gene(s) located downstream of bound intergenic regions are assumed to be the target genes of the TF in question. This way one can map specific TFs to their target genes in a genome-wide way. Note that this mapping is not necessarily a true association between a TF and the downstream gene. Protein-DNA interactions are physical events, meaning that they may play no role in gene regulation other than merely binding due to pure physical interactions. Also, two divergently transcribed genes may or may not be regulated by a single TF which binds to the intergenic region between them. An additional information such as gene expression levels is needed to resolve this issue. Apart from this mapping problem, ChIP-chip assays have other several limitations such as non-specific binding of proteins to an antibody and non-availability of whole-genome microarrays in high resolution. Further details can be found in review articles such as Sikder and Kodadek (2005) and Bulyk (2006).

Data from ChIP-chip experiments are one data source for our work in this thesis and we are not concerned with those shortcomings in this thesis. Our computational analyses will be performed on publicly available data as given. ChIP-chip data we will use were



	TF1	TF2	TF3	TF4	TF5	TF6
G1	1e-5	5e-4	2e-9	1e-1	5e-4	9e-3
G2	4e-7	5e-3	2e-2	7e-2	3e-4	3e-6
G3	1e-1	5e-11	3e-1	5e-1	5e-7	8e-8
G4	5e-2	5e-8	5e-4	2e-4	5e-2	7e-1
G5	1e-5	5e-3	2e-9	8e-1	2e-2	9e-8

	TF1	TF2	TF3	TF4	TF5	TF6
G1	1	1	1	0	1	0
G2	1	1	0	0	1	1
G3	0	1	0	0	1	1
G4	0	1	1	1	0	0
G5	1	0	1	0	0	1

Threshold = 1e-3

Figure 1.1: **ChIP-chip data matrix.** Here are shown example ChIP-chip matrices we use in this thesis. The matrix on the left is of the form of original ChIP-chip data in terms of binding p-values given by experimentalists. The rows represent gene names and the columns transcription factors (TFs). The binary matrix on the right is the one we generate from the original matrix for analysis given a p-value threshold of 0.001. The yellow cells are for TF-gene pairs with p-values less than 0.001. Those three TFs and two genes in green cells in the right matrix form one transcriptional module because all the TFs bind all the genes. Note that the module is a sub-matrix with the binary value of 1 (in orchid).

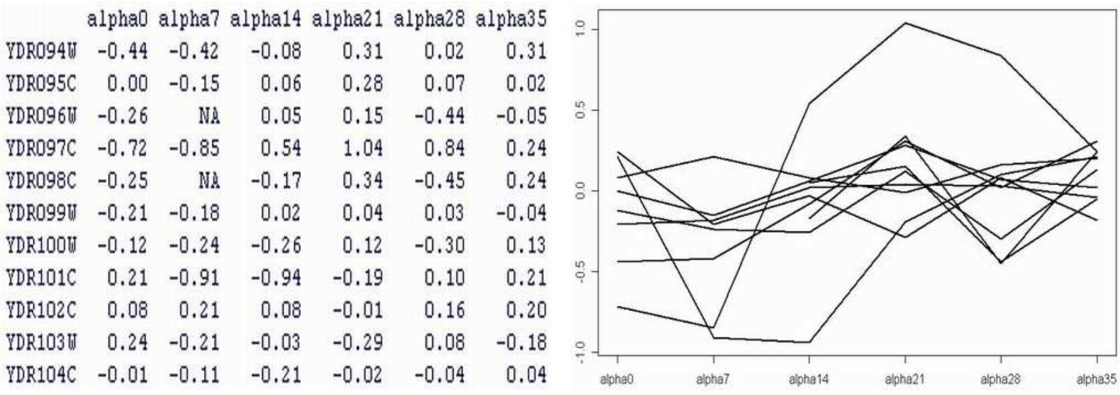
generated by the Richard Young laboratory (Harbison et al., 2004; Lee et al., 2002) and are of the form of a matrix consisting of genes in rows and TFs in columns (Figure 1.1). Each cell contains a significance value or p-value of each TF-gene interaction assigned by the Young laboratory.

1.2.2 Transcript expression profiling

Transcribed genes are expressed as functional products like proteins or RNAs. DNA microarray (or DNA chip) technologies have been used to obtain gene expression profiles of the genome of an organism (Gibson, 2003; Heller, 2002; Hoheisel, 2006). A microarray for expression profiling is an array of spots where the whole genome (the set of all gene sequences) is located to detect transcribed mRNA's simultaneously under a certain condition. In the case of ratio-based analysis of expression profiling, one uses data from two-colour systems using a single microarray which compare the amount of expressed mRNA's extracted from an experimental cells to that from control cells. Those extracted mRNA's from the two samples are reverse-transcribed into more stable cDNA's and labelled by two distinct fluorophores, normally red (R or Cy5) and green (G or Cy3) fluorescent dyes respectively. Then the two samples of cDNA's are hybridized to an array measuring intensities of the two colours. Analysis of R/G or Cy5/Cy3 ratios identifies up-regulated and down-regulated genes. Noise can be introduced because of different degradation rates and reverse-transcription rates of mRNA's, which we will not be concerned with in this thesis. There are also one-colour or single-channel microarrays to detect absolute expression levels of the genome from a single sample of mRNA's. The Affymetrix ”GeneChip” is one of the most popular one-colour systems (Lipshutz et al., 1999).

Microarray-based expression profiling is a mostly widely used way to study transcriptional responses of the whole genome simultaneously in a given condition. Diverse responses of expression profiles occur through different transcriptional regulation. Hence, gene expression analysis can be coupled to analysis of protein-DNA interactions by data integration, which is our aim in this thesis. As gene expression data from DNA microarrays were more extensively produced and analyzed than ChIP-chip data for instance, several standard public databases have been particulary developed to help researchers use those disparate data: for example, Gene Expression Omnibus (Barrett et al., 2007), ArrayExpress (Parkinson et al., 2007), and Stanford Microarray Database (Demeter et al., 2007). Those readers who are interested in analysis of expression data alone are referred to other review articles such as Sherlock (2001).

In this work, we will use gene expression datasets generated from two-colour systems under diverse cellular conditions. Each dataset we will analyze is in a format of a ma-



```
        alpha0 alpha7 alpha14 alpha21 alpha28 alpha35
YDR094W  -0.44  -0.42   -0.08    0.31    0.02    0.31
YDR095C   0.00  -0.15    0.06    0.28    0.07    0.02
YDR096W  -0.26     NA    0.05    0.15   -0.44   -0.05
YDR097C  -0.72  -0.85    0.54    1.04    0.84    0.24
YDR098C  -0.25     NA   -0.17    0.34   -0.45    0.24
YDR099W  -0.21  -0.18    0.02    0.04    0.03   -0.04
YDR100W  -0.12  -0.24   -0.26    0.12   -0.30    0.13
YDR101C   0.21  -0.91   -0.94   -0.19    0.10    0.21
YDR102C   0.08   0.21    0.08   -0.01    0.16    0.20
YDR103W   0.24  -0.21   -0.03   -0.29    0.08   -0.18
YDR104C  -0.01  -0.11   -0.21   -0.02   -0.04    0.04
```

Figure 1.2: **Gene expression data matrix.** The data matrix contains expression levels of genes (in rows) at each condition (in columns) in terms of log2(R/G) ratio values. The expression patterns are visualized on the right. Note that there are missing values (NA) as well. The data are taken from Spellman et al. (1998).

trix consisting of genes in rows and conditions in columns (Figure 1.2). Each cell is given a R/G ratio of measured expression levels. We are mostly interested in groups of genes whose expression patterns are similar. Although a recent study showed that similar expression profiles did not necessarily give rise to similar functions in mouse tissues (Yanai et al., 2006), co-expression analysis has given many insights into biological function and evolution (Stuart et al., 2003; van Noort et al., 2003).

1.2.3 Functional annotations

The most important issue in biology is to understand an organism's function or physiology. Traditional biology has contributed to a good amount of functional informations about individual proteins and genes. As an effort to use those invaluable informations easily and systematically, research communities have developed structured and controlled vocabularies for biological functions of genes and proteins and other related findings. Two main sources are the Gene Ontology (GO, Consortium, 2008) and the Munich Information Center for Protein Sequences (MIPS, Mewes et al., 2008). GO uses three ontologies to characterize genes and their products: molecular function, biological process, and cellular component. Each ontology is structured as a directed acyclic graph where nodes of terms are connected among them by their relationships, e.g. ”meiosis” $-- >$ ”cell cycle phase” $-- >$ ”cell cycle process” $-- >$ ”cellular process” $-- >$ ”biological process”. Most genes/proteins are annotated to multiple terms in each ontology as they have multiple functions in general. MIPS has a classification scheme according to the level of detail of functional annotations or categories. For example, ”meiosis” in MIPS is a functional category located under ”cell cycle” which is under the lowest level category ”cell cycle and DNA processing”. Although those manual annotations are constantly improving in quality and quantity, their precise definitions and annotations are elusive. Yet, they are a good source of large-scale information apart from those experimental data mentioned above. In this thesis, we will utilize annotation data from MIPS and further details on the data will be given where needed.

1.3 Computational approaches to data integration

Those heterogeneous high-throughput data sources mentioned above are rather sporadic and difficult to examine systematically within one framework or even under a single integrated database (Stein, 2003). In addition to the problem of data integration itself, data analysis is yet another problem we face. The tools developed so far do not produce consistent results among them when using the same data as exemplified by one recent assessment study on DNA motif discovery algorithms (Tompa et al., 2005). Nevertheless, we hope to find a suitable framework within which all related data could be modelled and described systematically. Also, large-scale data analyses can yield global views and properties of organisms, which are usually not visible from very detailed small-scale investigations. Thus, both top-down and bottom-up approaches are thought to be complementary and synthesized in the new area of systems biology (Kitano, 2002a,b; Schlitt and Brazma, 2007). It is expected that technological advances will provide us with all necessary data at all different cellular levels. Therefore, data integration is an important first step in systems biology.

A great deal of computational efforts have been under way to address transcriptional regulation by integrated analyses of heterogenous datasets in one way or another (Bar-Joseph et al., 2003; Lemmens et al., 2006; Tanay et al., 2004; Wu et al., 2006; Xu et al., 2004; Yu and Li, 2005). For example, Bar-Joseph et al. (2003) developed a 3-stage algorithm (called GRAM) to discover functional modules using ChIP-chip and gene expression data. In its first stage, the GRAM algorithm identifies target genes bound by each of all possible sets of TFs based on ChIP-chip data at a strict p-value threshold. In the second stage, it identifies a subset of those target genes which show co-expression most tightly, i.e., expression core set. This is determined by looking for a sphere in expression space to which the largest number of gene expression profiles belong. Then, GRAM revisits ChIP-chip data with a relaxed p-value threshold to expand the core set of target genes if genes with less prominent binding show the expression similarity. This yields the final module which GRAM discovers. In this way, they were able to generate a set of modules which give rise to a transcriptional network under a certain condition. Most of those modules were also found to be enriched in certain functional categories. In principle, other computational works mentioned above integrate diverse data sources using their own algorithms to provide a global view of transcriptional networks in terms of functional modules. Those different algorithms are hard to be compared objectively and often provide complementary approaches to investigate large-scale transcriptional regulation. We also note that those modules are useful concepts but their reality in themselves is questionable as we mentioned before. The members of modules (e.g. genes, proteins, and TFs) need to be examined individually as well in order to obtain detailed mechanistic understanding of how the cell works. In this thesis we will be more concerned with this issue than module networks.

1.4 Contributions of the thesis

We investigate transcriptional modules using ChIP-chip, gene expression and functional annotation data and discover dynamic aspects of modules by identifying condition-invariant and condition-specific functional modules (Chapter 2).

We develop and validate a method to prioritize gene regulatory interactions from large-scale modules and provide detailed case studies from our predictions (Chapter 3).

We develop a method to identify condition-specific co-factors of those transcription factors which change their target genes in different conditions. We use ChIP-chip data for prediction and provide evidences from gene expression, protein-protein interaction, and conserved motif data. We further show that condition-specific combinatorial regulation is statistically significant (Chapter 4).