Chapter 2

Functional analysis of transcriptional modules

2.1 Background

In order to investigate transcriptional regulation, we start with groups of genes and their regulators in which target genes regulated by common transcription factors (TFs) show similar properties such as mRNA expression levels and functional annotations. Those groups are often called transcriptional modules. We are particulary interested in global properties of transcriptional modules using three different types of genome-wide data: TF-DNA binding (ChIP-chip), mRNA expression, and functional annotation data. This chapter is based on transcriptional modules derived from ChIP-chip data. The main objective here is to identify condition-invariant and condition-specific modules which show functional signals, which are detected by coherence in both expression profiles and functional annotations among module genes.

We start with a comprehensive list of transcriptional modules derived from ChIP-chip data by Lee et al. (2002). As mentioned in Introduction, ChIP-chip data provide phys-

ical binding information rather than functional regulatory information. To overcome this problem of the ChIP-chip binding data, we integrate gene expression and functional annotation data with the binding data. By this data integration, we identify a subset of those modules which also show expression and functional coherence of target genes, i.e., functional signals. Keeping in mind the distinct regulatory mechanisms in different cellular conditions we use four different expression datasets for comparison of identified modules. Despite incompatible experimental conditions between different types of data sources, we identify condition-invariant and condition-specific modules which are claimed to be active in the respective conditions and show that identified coherent functions of genes in modules are relevant to respective conditions.

2.2 Transcriptional modules from binding data

A transcriptional module (TM) is defined to be a pair consisting of a set of transcription factors (F) and a set of their common target genes (G). Every factor in F is assumed to bind all genes in G (see Figure 1.1 for an example). We investigated a comprehensive list of 724 putative TMs generated by Manke et al. (2003) from the ChIP-chip data matrix of Lee et al. (2002) at a binding p-value threshold of 0.001. The experiments were conducted in rich media with 106 TFs and only 1671 (37%) of all analyzed 4532 intergenic regions have one or more binding sites assigned below the binding threshold. The number of corresponding genes for those intergenic regions is 2363 based on the result of Lee et al., and each module contains a group of transcription factors and target genes instead of intergenic regions bound by the transcription factors. Each intergenic region assayed is assigned to one or two downstream neighbouring genes (considering divergently transcribed genes; http://web.wi.mit.edu/young/regulatory_code/). This list of 2363 genes provides the background set for subsequent statistical analysis. The name of each module is given from the names of the TFs in each module as modules are

identified by the combinations of TFs. We removed two modules, named GAT1 and MSN2 (i.e., corresponding to the two TFs respectively), which have only one target gene because we need at least two target genes to calculate an expression correlation coefficient in subsequent analysis, so the actual number of modules we investigated is 722 without the two. The maximum number of transcription factors in a module is 10 and the maximum number of target genes in a module is 273 (module ABF1).

One does not claim that each TM completely defines the transcriptional interactions between TFs and target genes as discussed in Introduction. We normally need to filter out false ones from given data by the help of other informations. Here we are interested in utilizing expression data to identify functional modules as below.

2.3 Characterization of functional modules

Once a list of modules from binding data is given, we examine expression coherence of modules to detect a functional signal. We use mRNA expression profiles for identification of expression coherence of target genes within each module. For expression data, we use four different datasets from four publications (Hughes et al., 2000; Ihmels et al., 2002; Roberts et al., 2000; Spellman et al., 1998). The usage of different expression datasets serves to detect different functional signals in different cellular conditions. We name them after the first authors and here analyze the 722 ChIP-chip derived modules using each of the expression datasets separately as below.

(1) Ihmels Dataset

The Ihmels dataset is composed of more than 1000 conditions compiled from 34 different publications (Ihmels et al., 2002). The dataset was selected to extract strong expression signals over very diverse conditions. This dataset includes the other three

	Ihmels	Spellman	Hughes	Roberts
Conds	1011	77	300	56
Genes	2333	2162	2336	2077
SSMs	20	67	33	36
EFCs	35	87	59	62

Table 2.1: **Summary of module analysis results**. A summary of our analysis of functional modules is shown for the four expression datasets we analyzed. The four columns are for the four datasets respectively and the rows are the number of conditions ('Conds'), the number of genes ('Genes'), the number of statistically significant modules ('SSMs'), and the number of enriched functional categories in those SSMs ('EFCs'), respectively.

datasets we used (however, there were some compilation errors in the original dataset). The data matrix consists of 6206 genes as rows and 1011 conditions as columns. 42 genes in the dataset have missing values for the maximum 86 conditions (ca. 9%) and the matrix contains 40,048 missing values (ca. 0.6% of 6206 $\times$ 1011 = 6,274,266 data points). In order to estimate missing values, we used the *impute* package in R (http://cran.r-project.org/src/contrib/Descriptions/impute.html) with default parameter values, which use as estimates averages of non-missing expression values of k-nearest neighbouring genes with a Euclidean metric. Considering the intersection with the gene set from the ChIP-chip data (2363 genes; see the previous section), we finally obtained 2333 genes and 1011 conditions for our subsequent analysis (Table 2.1).

Given a transcriptional module from the previous section and an expression dataset like the Ihmels set, we examine expression coherence of module genes as follows. First, we calculate Pearson correlation coefficients, r, for all pairs of expression profiles of target genes in the module and take the average of the absolute values of the coefficients. The reason why we take the absolute value is that we consider both positive and negative correlations as the signals for possible co-regulation. We define this average value, ζ,

in general as follows,

$$\zeta \equiv \frac{1}{L}\sum_{k=1}^{L}|r_k| \ , \quad L = \binom{N}{2} \ , \tag{2.1}$$

where L is the number of all pairs of N target genes in each module and r_k is the Pearson coefficient for a pair k. Then, we take ζ as a statistic for the significance test of expression coherence. For background ζ values, we generated random modules of the same size by sampling the same number of genes from the background set of 2363 genes (see the previous section). We estimated a p-value of expression coherence, p_e, for each observed module by the fraction of the number of those random ζ's that are equal to or greater than the observed ζ with respect to the number (K) of randomly sampled groups, which is K = 10,000 in this study,

$$p_e = \frac{\| \ \{ \ \zeta_k \mid \zeta_k \geq \zeta \ , \ \ k = 1, 2, \ldots, K \ \} \ \|}{K} \ , \tag{2.2}$$

where k is an index for random modules. For the multiple testing problem, we used the *qvalue* package in R (http://cran.r-project.org/src/contrib/Descriptions/qvalue.html, with default parameter values) from the list of p-values generated. Q-values control the false discovery rate (FDR) rather than the false positive rate (Storey and Tibshirani, 2003). Transcriptional modules with q-values less than a threshold (τ_e) are considered to show co-expression or expression coherence among target genes. Those modules are termed SSMs (statistically significant modules), i.e.,

$$SSM = \{TM = (F, G) : q(G, \zeta) < \tau_e\} \ , \tag{2.3}$$

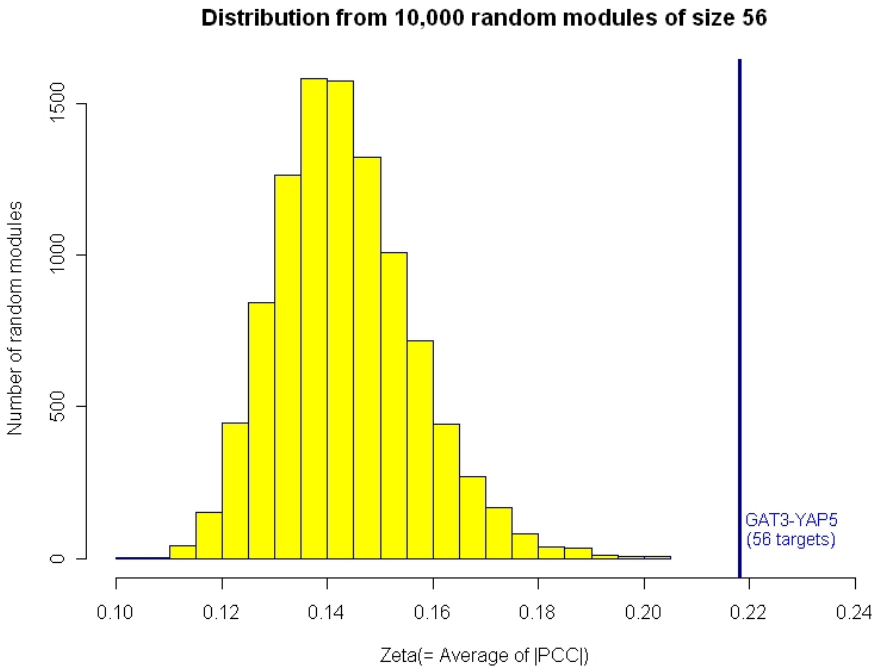

Figure 2.1: **Expression coherence test.** Distribution of 10,000 simulated ζ values from 10,000 random modules of size 56 and the ζ value (0.218) of an example GAT3-YAP5 module with 56 target genes using the Ihmels dataset. This module is deemed to possess expression coherence of target genes (q-value < 0.001).

where q is the q-value and $\tau_e = 0.001$ in this work. An example test is given in Figure 2.1.

By this method, we examined all the 722 ChIP-chip-derived modules with more than one target gene. Ranking those modules by q-values yielded 20 SSMs (Tables 2.1 and 2.2). Figure 2.2 shows the 20 SSMs and relations among them in terms of TFs. They may be seen as redundant because of subset relationships of TFs, but we consider them as independent in this study of the current chapter. In the next chapter we will take a different approach to address this issue.

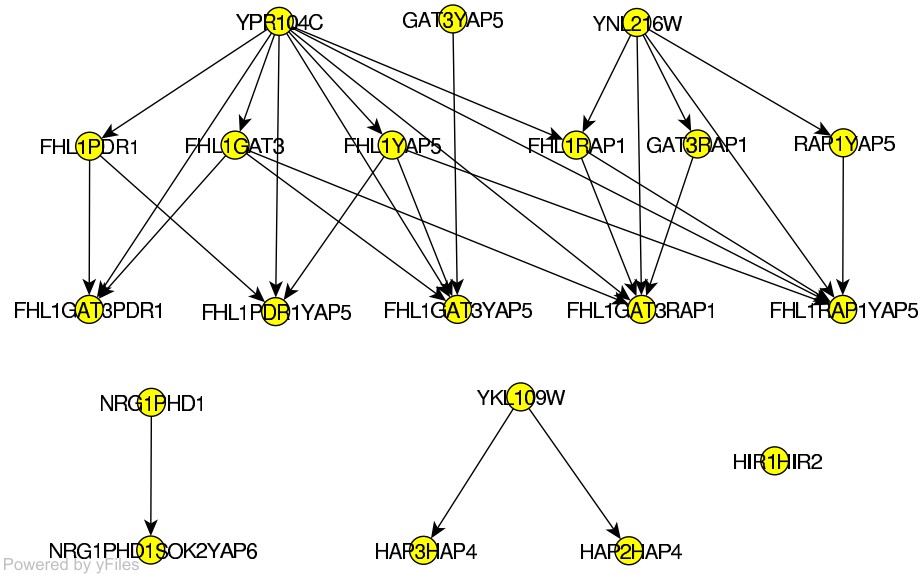

Figure 2.2: **Statistically significant modules.** The figure shows a graphical representation of the top 20 significant modules (SSMs) from the Ihmels dataset. Note that there exist subset relationships among them due to all possible TF combinations. We assumed in this work that they are all independent modules. It may also be seen as a hierarchical structure of the modules in terms of TFs.

Table 2.2: **Significant modules identified.** We show some of significant modules we identified (q-value < 0.001). 1 and 0 for each expression dataset of each column indicates that a module in a row is significant or not according to Eq. (2.3).

Significant Modules	Ihmels	Spellman	Hughes	Roberts
FHL1RAP1	1	1	1	1
GAT3RAP1	1	1	1	1
GAT3YAP5	1	1	1	1
HIR1HIR2	1	1	1	1
FKH2MBP1	0	1	0	0
FKH2MBP1SWI4	0	1	0	0
FKH2MCM1NDD1	0	1	0	0
FKH2NDD1SWI4	0	1	0	0
MBP1SWI4SWI6	0	1	0	0
CIN5NRG1YAP6	0	0	1	0
CIN5PHD1	0	0	1	0
PHD1YAP6	0	0	1	0
DIG1STE12	0	0	0	1
FKH2NDD1	0	1	0	1
MBP1SWI4	0	1	0	1
MBP1SWI6	0	1	0	1
SWI4SWI6	0	1	0	1

Table 2.3: **Significant functions identified.** We show some of enriched functions (p-value < 0.05) identified in significant modules. 1 and 0 for each expression dataset of each column indicates that a functional category in a row is enriched or not among significant modules.

MIPS categories enriched	Ihmels	Spellman	Hughes	Roberts
ribosomal proteins	1	1	1	1
DNA damage response	1	1	1	1
DNA recombination and DNA repair	1	1	1	1
DNA synthesis and replication	1	1	1	1
energy generation (e.g. ATP synthase)	1	1	1	1
mRNA synthesis	1	1	1	1
organization of chromosome structure	1	1	1	1
sugar binding	1	1	1	1
translational control	1	1	1	1
regulation of DNA processing	0	1	0	0
RNA modification	0	1	0	0
tRNA synthesis	0	1	0	0
anaerobic respiration	0	0	1	0
C-compound and carbohydrate metabolism	0	0	1	0
lipid, fatty acid and isoprenoid biosynthesis	0	0	1	0
nitrogen and sulfur utilization	0	0	1	0
phosphate metabolism	0	0	1	0
regulator of G-protein signalling	0	0	0	1
CELL CYCLE AND DNA PROCESSING	0	1	0	1
directional cell growth (morphogenesis)	0	1	0	1
nuclear and chromosomal cycle	0	1	0	1
CELL RESCUE, DEFENSE AND VIRULENCE	0	1	1	0
C-compound and carbohydrate utilization	1	1	1	0

Following the expression coherence test, functional coherence is examined for each of expression-coherent modules (i.e., SSMs). Function coherence is meant to be the same as enrichment of a functional category or annotation throughout our work. To this end, we use the functional categories provided by the Munich Information Center for Protein Sequences (MIPS, Mewes et al., 2004) and focus on those functions which are enriched in some SSMs. We test functional enrichment in each module upto the 3rd level of the category hierarchy (classification version 2.0; the most detailed category is at the 6th level, e.g. 'biosynthesis of homocysteine' in metabolism with MIPS code 01.01.06.05.01.01). More detailed annotations were pruned at the 3rd level, resulting in about 200 categories examined in total. They contain upto $\sim$750 proteins with an average of 56, excluding the category, 'unclassified proteins', which contains about 2000 proteins.

Given a module and a functional category, we assess enrichment of the functional category among target genes using the standard method (Tavazoie et al., 1999). The assessment of function coherence for each category is done by calculating a hypergeometric p-value which is defined as follows,

$$p_f = 1 - \sum_{k=0}^{K_f-1} \frac{\binom{M_f}{k}\binom{N-M_f}{S-k}}{\binom{N}{S}} , \tag{2.4}$$

where f is a functional category, N is the number of all genes which are annotated to at least one functional category, M_f is the number of all genes which are annotated to the functional category f, S is the number of all target genes in a module of interest, and K_f is the number of target genes in the concerned module which are annotated to the given functional category f. A functional category for each module is deemed coherent if p_f is less than a prescribed threshold, τ_f (0.05 in this work). Note that we

may obtain multiple coherent functions in each module. We do not correct p-values for multiple testing because for each module, p_f is defined for only those categories in which K_f is greater than 0. Also note that statistical and biological significance are not correlated in general, hence correction of p-values may be arbitrary with respect to functional significance.

For the top 20 SSMs identified above, we obtained a total of 35 enriched MIPS categories (Tables 2.1 and 2.3). Ribosomal protein genes were observed in 13 different modules, in agreement with the importance of ribosome biosynthesis in a general context and the involvement of Fhl1, Rap1 and Yap5 transcription factors (TFs) in those modules which are known to be main regulators of those genes (Lee et al., 2002). Moreover, we were interested in most frequently occurring TFs as generic TFs in the top 20 modules assuming that they would not substantially change their target genes in different conditions and contribute to the detection of strong signals by our analysis applied to the large number of conditions altogether. We found that 10 out of the top 20 modules have Fhl1 as a TF, which was previously observed in the ChIP-chip experiments by Lee et al. as the main regulator of most ribosomal genes (Lee et al., 2002). In addition, Fhl1 was classified into the condition-invariant category as well as condition-enabled in the study of Harbison et al. (2004). Another study about ribosomal protein modules showed that they are highly conserved across species together with Ifh1, Rap1 and Fhl1 binding motifs (Tanay et al., 2005). Hence, taking the previous studies as evidence, we confirm that Fhl1 is a generic TF from our analysis of the large-scale expression data. Other frequently occurring regulators in the top modules include Rap1, Yap5 and Gat3, among which Rap1 was found to be a condition-enabled TF by Harbison et al. (2004). The fact that Rap1 regulates the same set of target genes whenever it is activated in some conditions (i.e. condition-enabled) supports that Rap1 is also a generic TF in our analysis. Yap5 was found to be involved in regulation of

ribosomal protein genes forming a multi-input motif together with Fhl1 and Rap1 by Lee et al. (2002). Our result about Yap5 may signify its role as a generic TF in a large number of conditions. This TF has been also found to be active in 4 out of 5 conditions (2 endogenous and 3 exogenous conditions) by Luscombe et al. (2004). As for the remaining factor Gat3, we conclude that it is generic as well over diverse environmental conditions along with the other factors since they occur in the same module. Luscombe et al. (2004) found that Gat3 is active in 5 out of 5 distinct cellular conditions. We note again that our motivation was to find global or generic modules and corresponding TFs which function in a wide range of different conditions, which justifies our investigation of strong signals in the expression data of more than 1000 different conditions.

(2) Spellman Dataset

The Spellman dataset contains time series data about cell cycle with four different synchronization methods including one from a previous study (Spellman et al., 1998). The data matrix consists of 6178 genes and 77 conditions. We removed those 336 genes with more than 30% missing values across all the conditions and then did the imputation for the 5842 $\times$ 77 matrix using the impute package in R as above. 201 genes out of our 2363 background genes had no entries in the matrix, giving us the 2162 $\times$ 77 data matrix (Table 2.1). The Spellman dataset was selected because it gives experimental conditions which are similar to ChIP-chip ones (rich medium) in the study by Lee et al. (2002).

We retained only 715 out of the initial 722 modules because of missing genes in the data matrix. We found 67 SSMs at a q-value threshold of 0.001 (Tables 2.1 and 2.2). Those SSMs include the known cell-cycle complexes such as Swi4p-Swi6p (SBF) and Mbp1p-Swi6p (MBF), illustrating the plausibility of our statistical analysis. We also identified SSMs containing other cell-cycle related TFs: e.g. modules FKH2-MBP1-

SWI4, FKH2-MCM1-NDD1, MCM1-STE12 and MBP1-SWI6. Also, the modules for ribosome biosynthesis are among them.

The 67 SSMs contain a total of 87 coherent functional categories including 'DNA synthesis and replication', 'DNA damage response', 'organization of chromosome structure', 'DNA recombination and DNA repair' and 'mitotic cell cycle and cell cycle control' as most frequently enriched in the SSMs (Tables 2.1 and 2.3). This implies that our identified SSMs are specific for the cell cycle dataset. The category 'ribosomal proteins' is also found, as expected from the result of the Ihmels dataset above. We note that the number of SSMs is the largest among those from the four datasets we investigated. We attribute this observation to the experimental conditions similar to those of the ChIP-chip data we used. This motivates us to further develop this analysis in a biological context in the next chapter.

(3) Hughes Dataset

The Hughes dataset is a compendium of 300 experiments including various 276 deletion mutants (Hughes et al., 2000), and taken from the Ihmels combined dataset. The original data were published as log10-ratio values and have been transformed to log2-ratio values in the Ihmels dataset. The data matrix consists of 6206 genes and 300 conditions without missing values. 27 genes out of the 2363 background genes did not have entries in the dataset, giving us the 2336×300 data matrix (Table 2.1). The Hughes dataset can be considered as a synthetic condition set since it contains more than 90% deletion mutants experiments including deletions of transcription factors.

We examined all the 722 ChIP-chip modules using this dataset. We found 33 SSMs at a q-value threshold of 0.001 (Tables 2.1 and 2.2). There are 59 significant functional categories in the top 33 SSMs (Tables 2.1 and 2.3). The most frequently an-

notated category is 'carbon metabolism'. This may imply that carbon metabolism is an important or mostly affected process in responses to abrupt perturbations of cells. Some TFs belong to a number of SSMs: Nrg1 and Phd1 to 10 SSMs and Yap6 to 9 SSMs. Nrg1 has been implicated in negative repression of diverse processes, but the other two factors are not known to particulary play a role in perturbed cellular conditions (SGD). Those 19 modules associated with the three TFs are enriched in 'carbon metabolism', 'cellular import/export', 'sugar binding', 'cell-cell adhesion' and 'cell rescue/defense/virulence'.

(4) Roberts Dataset

The Roberts dataset is concerned with MAPK signaling pathways with alpha-factor pheromone treatment and deletion mutants (Roberts et al., 2000). We used its subset without missing values, having the resultant data matrix of 5627 genes and 56 conditions. A total of 286 genes out of the 2363 background genes were filtered out yielding the 2077×56 data matrix (Table 2.1). The Roberts dataset is similar to the Spellman dataset in that both sets are involved with alpha-factor treatment.

We retained 716 ChIP-chip modules and found 36 SSMs at a q-value threshold of 0.001 (Tables 2.1 and 2.2). There are 62 significant functional categories in the top 36 modules (Tables 2.1 and 2.3). The categories include 'ribosomal proteins', 'DNA damage response' and 'DNA synthesis and replication'. This list is similar to that of the Spellman dataset because the experiments were also done with alpha-factor pheromone treatment. This similarity was confirmed by detecting common SSMs in the two datasets such as those containing Fkh2, Swi4, Swi6 and Mbp1. The modules include Swi4p-Swi6p (SBF) and Mbp1p-Swi6p (MBF) complexes (Iyer et al., 2001). The interaction between the two known complexes, for certain genes, may be deduced from the module, MBP1-SWI4, connecting the two complexes. A more plausible point

from this dataset is that we also obtained the three SSMs containing Dig1 (a.k.a. Rst1), Ste12 and both respectively. Those TFs are known to be downstream targets of MAPK signalling pathways forming a complex (Roberts et al., 2000; Zeitlinger et al., 2003). Thus, we verify, from the module DIG1-STE12, the fact that Dig1 and Ste12 act together as transcriptional regulators, and either one of the two is also thought to exhibit a valid transcriptional activity provided that the experimental conditions from the ChIP-chip and gene expression are comparable.

2.4 Condition-invariance and condition-specificity

We have characterized SSMs from each of the four expression datasets by identifying coherent functional categories in them. The modules are coherent in expression patterns of target genes depending on the expression dataset used. One observes that some modules show expression coherence in all datasets while some others show coherence in a specific dataset (Table 2.2). Such condition-invariant modules include HIR1-HIR2 module as well as others we discussed above from the Ihmels dataset. The module contains the two regulators Hir1 and Hir2, six histone genes which are known targets of Hir1 and Hir2, and one additional gene, YPR195C, which is a dubious ORF (SGD). In our framework, YPR195C may be suggested to be functional in the histone module. In the case of the Roberts dataset about the MAPK signaling pathways, the Dig1 and Ste12 related modules were identified as strongly coherent among others in contrast to the other datasets (Table 2.2). This illustrates that the binding of Dig1 and Ste12 observed in rich media ChIP-chip condition may indeed account for the coherence of certain targets in some other conditions such as the pheromone response. It remains, however, an open question whether additional factors could play a role, especially because we observed reduced expression coherence of these targets in the other datasets. In fact, Zeitlinger et al. (2003) proposed a model to account for this specific mating

response due to changes in a third factor, Tec1.

Also, the observation from the comparison of enriched functions is that there are certain unique categories which are characteristic in each dataset as well as common enriched functions (Table 2.3). For example, 'ribosomal proteins' occurs as an enriched category in certain modules of each of all the expression datasets. Two categories, 'mitotic cell cycle and cell cycle control' (MIPS code 10.03.01) and 'nuclear and chromosomal cycle' (MIPS code 10.03.04), are found to be enriched in the top modules of the Spellman and Roberts datasets, both of which are relevant to cell cycle processes. On the other hand, only the Roberts dataset has an enriched category, 'regulator of G-protein signalling', which is in line with the experimental focus of these experiments on pheromone response involving signalling of G-proteins (Roberts et al., 2000). This functional comparison distinguishes this particular dataset from others and tells us about specific features of the cell's states most involved in the experimental conditions. This fact supports our approach to identify the SSMs relevant to the data in terms of over-represented functions among target genes in those modules.

However, a comprehensive analysis requires more ChIP-chip data for conditions comparable to the expression data. In this work, we used the binding data as provided by Lee et al. (2002). This dataset gives us binding information for 106 regulators assayed in rich media condition. The initial set of ChIP-chip modules was fixed regardless of expression data we used for integration. It has been pointed out that the binding of TFs to potential targets is condition-dependent. Subsequent works have highlighted this dependency (Bar-Joseph et al., 2003; Harbison et al., 2004; Luscombe et al., 2004; Zeitlinger et al., 2003). Yet, this limitation in data integration is unavoidable because of lack of ChIP-chip data compared to expression data. Therefore, our assumption is that the groups of genes and TFs derived from ChIP-chip assays do not change their

regulatory interactions in different conditions. This is likely to be unrealistic, but under that assumption we were able to examine expression changes of those modules using diverse expression datasets available. More careful utilization of condition variables in gene expression and binding data will certainly result in improved results (e.g., more comprehensive and precise condition-specific sets of SSMs and coherent functional categories). Unless we take great care of those experimental or cellular conditions, transcriptional mechanisms would be elusive in large-scale data analysis.

2.5 Summary

We have presented an approach to investigate transcriptional regulation in terms of groups of transcription factors and target genes, called transcriptional modules. To detect functional signals in modules, we applied statistical analyses of expression and functional coherence to ChIP-chip derived modules. Four different expression datasets were utilized to characterize their respective functional features. The identified functional modules include condition-invariant and condition-specific ones, showing dynamic behaviours in a condition dependent way. In order to give more support and evidence, one can think of integrating more diverse data types, e.g. protein-protein interactions, orthologous genes and subcellular localization in a fashion similar to what we used to establish coherent modules. On the other hand, one has to be careful about data integration because it is impossible to have different types of data under the exactly same cellular conditions. Since transcription phenomena are highly dynamic and condition-dependent, it is often biologically implausible to combine two types of data conducted in different laboratories, even in rich media. We may also have to use subsets of an expression dataset from the same study separately. In the next chapter, we pay more attention to such an issue of experimental conditions and attempt to prioritize TF-gene links from such identified modules by focusing on common functional

categories among modules.