Chapter 4

Condition-specific combinatorial regulation

4.1 Background

Transcriptional regulation takes place by many regulatory molecules which dynamically bind DNA by interacting with each other, i.e., combinations of regulatory molecules. Most of genes are hence regulated by multiple transcription factors (TFs) in a combinatorial way as discussed in Introduction. Although combinatorial regulation has been implied in our module analysis in the two previous chapters, we did not properly deal with the issue. For instance, two regulators in a coherent module do not necessarily act together to regulate target genes. The regulatory links between regulators and target genes in a module were derived from protein-DNA interaction data (ChIP-chip) which reflect mere physical binding events. Expression or functional coherence of module genes may result from regulation by only one of the two regulators, the other being merely bound with no regulatory effect at all. For example, Swi6 has been argued to have a regulatory function in the two heterodimers, SBF and MBF (Iyer et al., 2001). In this chapter we present novel approaches to investigate combinatorial regulation by pairs of regulators, in a condition-specific way, using protein-DNA interaction data.

The protein-DNA interaction data by Harbison et al. (2004) showed dynamic binding patterns of individual TFs in different experimental conditions. One class of such TFs is those which bind different sets of genes in different conditions, so-called "condition-altered" TFs such as Ste12 (Harbison et al., 2004). To explain these changes, they proposed that "the binding specificity of many of the transcriptional regulators might be altered through interactions with other regulators or through modifications (such as chemical) that are dependent on environment" (Harbison et al., 2004). For instance, Ste12 interacts with Mcm1 in response to pheromone to induce the expression of mating genes, whereas during filamentous growth it interacts with Tec1 to express filamentation-specific genes (Zeitlinger et al., 2003). Here we aim to formally investigate this type of scenario of selective partnership in a genome-wide way with the hypothesis that such changes of target genes in different conditions are due to changes of co-factors which interact with those condition-altered TFs.

ChIP-chip data assayed in different conditions are the main source of information for our approaches in order to identify condition-specific combinatorial TF pairs and gene expression data will be used to give support for our predictions by examining synergistic effect of combinatorial TF pairs on coherent expression of their target genes. Our two main goals of this chapter are (1) to predict condition-specific co-factors of condition-altered TFs and (2) to investigate whether the occurrence of condition-specific combinatorial regulation is statistically significant.

Condition	YPD	ACID	ALPHA	BUT14	BUT90	GAL	H2O2HI	H2O2LO	HEAT	PI	RAPA	SM
# TFs	203	2	5	8	4	5	38	28	6	2	14	34

Table 4.1: **ChIP-chip conditions and the numbers of TFs assayed.** The abbreviations of conditions come from the supplementary material of the paper by Harbison et al. (2004). For example, those conditions contain rich medium (YPD), mating inducing (ALPHA), moderately hyperoxic (H2O2LO), elevated temperature (HEAT), nutrient deprived (RAPA), and amino acid starvation (SM). These selected conditions are also used in Subsection 4.5.1.

4.2 Identification of condition-altered TFs by a hypergeometric test

As the first step of our approach, we aim to identify transcription factors (TFs) which have different sets of target genes in different conditions, i.e., condition-altered TFs. We use ChIP-chip data by Harbison et al. (2004) which were produced in 14 different conditions. As combinatorial regulation is our focus of study, we work with conditions where at least two TFs were assayed. This leaves us with 12 datasets in 12 conditions with 2 to 203 TFs and 6229 genes (Table 4.1). In addition, to reduce noise and errors in gene assignments, we use a refined list of 5714 genes obtained by Kellis et al. (2003) whose computational analyses discarded a number of genes based on whole-genome comparisons with three other yeast species. We analyze the intersection of the two gene sets by Harbison et al. and Kellis et al., which results in 5629 genes. To define target genes by each TF, we set a ChIP-chip binding p-value threshold to be 0.001. This defines a background set of 3465 genes bound by at least one TF in any condition examined.

As shown in Figure 4.1, identification of condition-altered TFs is achieved by calculating a hypergeometric p-value for the number of common genes bound in a pair of conditions by a single TF. Since we are interested in significant *changes* of target genes for a TF in a pair of conditions, we expect to have a *little overlap* between two gene sets in the condition pair in question. Hence, by taking those TFs with hypergeometric p-values > 0.1 (i.e., non-significant overlaps), we obtain a list of condition-altered TFs in all pairs of conditions. The choice of the p-value threshold of 0.1 is arbitrary. We exclude those TFs which bind no target genes in one of a pair of conditions because it may indicate the *absence* of the factors in the nucleus under that particular condition, which is not of our interest. Note that this hypergeometric test of the gene intersection being a small overlap is a practical way to identify condition-altered TFs. In principle we do not want significance of the small overlap to depend on the size of a background gene set which is one parameter in the hypergeometric probability. However, our background set of 3465 genes is fixed throughout the hypergeometric test for all TFs without greatly distorting our interpretation of significant changes of target genes. In this way, we identified 32 condition-altered TFs for 86 condition pairs from the ChIP-chip data (Table 4.2).

4.3 Systematic study of condition-specific co-factors

Given an identified condition-altered TF, TF_X, and a corresponding pair of conditions, we identify condition-specific co-factors of TF_X in each of the condition pair by a second hypergeometric test (Figure 4.1). Our focus here is the intersection of the following two gene sets : one is the set of genes bound by TF_X in one of the two conditions but not in the other (i.e., condition-specific targets) and the other the set of genes bound by another TF in that condition. We calculate a hypergeometric p-value for the number of common genes between the two gene sets. Note that we will have two such p-values given a condition pair and a TF (potential co-factor). By taking those TFs which give rise to hypergeometric p-values $< 10^{-4}$ (significant overlap) in one condition and > 0.1 (non-significant overlap) in the other, we obtain a

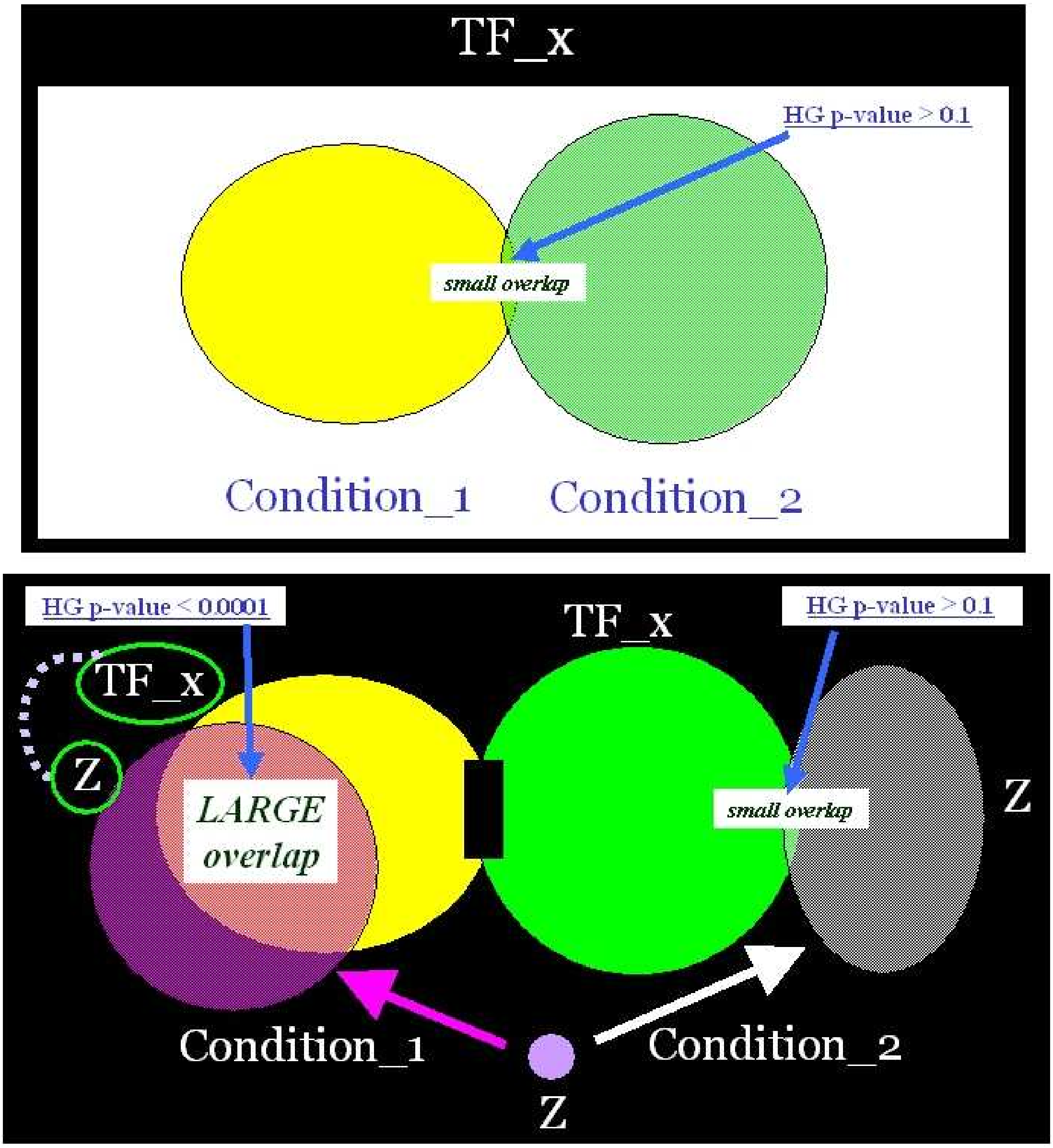

Figure 4.1: **Condition-altered TFs and condition-specific co-factors.** The upper panel shows how to identify condition-altered TFs from ChIP-chip data by a hypergeometric test. An example TF, TF_x, is a condition-altered TF which binds different sets of target genes in two different conditions (colored circles) according to the small overlap, i.e., the hypergeometric (HG) p-value cutoff of 0.1. The lower panel shows how to identify condition-specific co-factors for a given condition-altered TF by a second hypergeometric test. In Condition 1, another TF, Z, has the large overlap (HG p-value $<$ 0.0001) with TF_x, whereas the two TFs have the small overlap (HG p-value $>$ 0.1) in Condition 2. Hence, TF Z is a condition-specific co-factor of TF_x through a physical or functional interaction (dotted line).

list of condition-specific co-factors for TF_X in each of the condition pair in question. That is, those co-factors bind a significant number of common genes with TF_X in a condition-specific manner. Here the p-value thresholds are arbitrary. We note also that we imposed a constraint of at least 3 genes on the size of both condition-specific sets of target genes, given a condition pair and a condition-altered TF. This is an attempt to reduce sensitive changes of hypergeometric p-values depending on the gene set size.

As an alternative and threshold-independent way of identifying condition-specific co-factors for a condition-altered TF in a pair of conditions, we examine a difference between two distributions of ChIP-chip binding p-values given a candidate co-factor and two condition-specific gene sets of the condition-altered TF corresponding to the condition pair in question (Figure 4.2). If a particular co-factor shows a significant shift of a binding p-value distribution in one condition against the other by the Wilcoxon rank sum test with a p-value threshold of 0.001 (arbitrary choice), that co-factor is deemed a condition-specific co-factor of the condition-altered TF under consideration (Figure 4.2). As before we impose a constraint of at least 3 genes on each of the two condition-specific gene sets. We will mainly focus on the first approach above because the second alternative approach does not identify common target genes of two combinatorial TFs while our expression analysis below is applied to common target genes.

In the previous section, we identified 32 condition-altered TFs. 2 of them were not tested here because of our constraint of at least 3 genes in condition-specific gene sets. The second hypergeometric test found that 25 out of the rest of 30 condition-altered TFs have a total of 114 condition-specific co-factors in some of 5 specific conditions. The number of unique co-factors over all the 5 conditions is 51 and some of them are condition-altered TFs themselves. There are 18 such condition-altered TFs that act as condition-specific co-factors as well. This implies that 72% of those 25 condition-

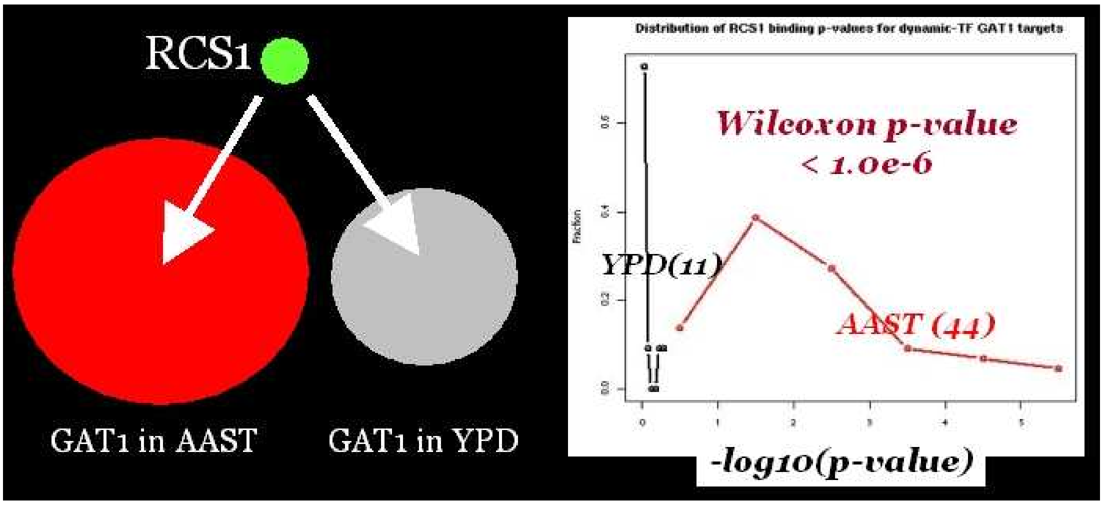

Figure 4.2: **An alternative way of identifying condition-specific co-factors.** The TF, $GAT1$, in the figure is meant to be a condition-altered TF in rich media (YPD) and amino acid starvation (AAST) conditions. The red and grey circles represent the target gene sets of $GAT1$ in the two conditions, 11 and 44 genes in the respective conditions. Then, another TF, $RCS1$, is introduced to compare two distributions of binding p-values for the two sets of $GAT1$ targets. The two distributions in the figure are significantly different in the two conditions according to the Wilcoxon test. As $RCS1$ favorably binds $GAT1$ targets in the AAST condition, $RCS1$ is a condition-specific co-factor of $GAT1$.

altered TFs (= 18/25) are involved with condition-specific combinatorial regulation. Hence, if a TF shows different binding patterns in different conditions, the regulator is likely to be subjected to combinatorial regulation rather than individual binding activity. Considering the dual role of those 18 TFs in our framework, we predict unique 104 combinatorial TF pairs under a certain condition (Table 4.2). Our predictions include novel pairs such as GAT1 and RCS1 as well as known pairs such as MSN2 and MSN4.

Table 4.2: **Predicted combinatorial TF pairs and supports.** We predicted a total of unique 104 combinatorial transcription factor pairs under a certain condition using two consecutive hypergeometric tests. Here we show 44 of those pairs which are supported by any of synergistic expression analysis, conserved motif data, and protein-protein interaction data (Section 4.5). The columns are as follows: 'TF1', condition-altered TF; 'TF2', condition-specific co-factor of TF1; 'Condition', ChIP-chip condition assayed for TF1 and TF2; 'N_genes1' and 'N_genes2', numbers of target genes of TF1 and TF2 respectively; 'N genes12', number of common target genes of TF1 and TF2; 'HG p-value', hypergeometric p-value for N_genes12; 'SEC', truth value for detection of synergistic expression coherence of the common target genes; 'Motif', fraction of those common target genes with conserved motifs for both TF1 and TF2; 'PPI', truth value for protein-protein interaction between TF1 and TF2. 'NA' values appear if there is no expression data or no conserved motif data.

TF1	TF2	Condition	N_genes1	N_genes2	N_genes12	HG p-value	SEC	Motif	PPI
GAT1	RCS1	SM	44	21	9	9.51E-13	1	0	0
MSN4	PDR1	YPD	49	72	13	6.41E-12	1	0	0
YAP5	MSN4	YPD	73	49	16	5.27E-16	1	0.06	0
YAP1	CIN5	H2O2LO	36	117	13	4.57E-11	0	0.23	0
YAP1	YAP7	H2O2LO	36	144	28	3.28E-33	0	0.32	0
GAT1	GZF3	RAPA	26	32	8	2.88E-11	0	0.12	0
YAP1	SKN7	H2O2LO	36	169	11	5.58E-07	0	0.36	1
YAP1	YAP6	H2O2LO	36	55	11	2.51E-12	0	0.09	0
DAL80	GZF3	RAPA	33	32	14	8.75E-22	0	0.14	1
GAT1	GLN3	RAPA	26	62	19	1.84E-29	0	0.21	1
MSN2	MSN4	H2O2HI	73	64	43	4.66E-63	NA	NA	1
MSN2	MSN4	H2O2LO	42	18	13	1.29E-22	0	0.23	1
MSN2	MSN4	RAPA	44	50	38	1.67E-72	0	0.24	1
MSN2	MSN4	ACID	30	6	3	1.15E-05	NA	NA	1

Continued on the next page...

RCS1	AFT2	H2O2HI	45	53	13	2.66E-14	NA	NA	1
RIM101	NRG1	H2O2LO	48	43	25	9.12E-39	0	0	1
YAP1	CAD1	YPD	65	32	13	2.68E-15	0	0.85	1
DAL80	MSN4	RAPA	33	50	8	1.11E-08	0	0.25	0
RTG3	GLN3	RAPA	48	62	10	6.13E-09	0	0.3	0
GAT1	ARG81	YPD	11	17	5	6.77E-10	0	0.4	0
HSF1	MSN2	H2O2LO	93	42	13	2.16E-11	0	0.46	0
MSN2	YAP7	H2O2LO	42	144	9	3.97E-05	0	0.56	0
MSN4	YAP7	H2O2LO	18	144	7	4.01E-06	0	0.29	0
SFP1	FHL1	SM	44	193	41	5.72E-50	0	0.68	0
YAP5	FHL1	YPD	73	180	19	2.10E-09	0	0.58	0
GAT1	GCN4	RAPA	26	150	7	8.14E-05	0	0.71	0
RTG3	GCN4	RAPA	48	150	16	4.69E-11	0	0.69	0
PHO2	DAL82	SM	21	55	5	1.41E-05	0	0.6	0
PHO2	HAP5	SM	21	36	4	5.18E-05	0	0.5	0
MSN2	CIN5	H2O2LO	42	117	9	7.32E-06	0	0.44	0
ROX1	SKN7	YPD	63	60	7	8.48E-05	0	0.43	0
MSN2	SKN7	H2O2LO	42	169	20	3.87E-16	0	0.4	0
YAP1	CIN5	YPD	65	141	11	4.64E-05	0	0.36	0
RTG3	DAL82	RAPA	48	52	6	6.45E-05	0	0.33	0
MSN2	YAP1	H2O2LO	42	36	7	1.47E-07	0	0.29	0
ROX1	YAP6	YPD	63	87	27	1.99E-28	0	0.19	0
DAL80	DAL82	RAPA	33	52	11	4.31E-13	0	0.18	0
YAP1	YAP6	YPD	65	87	12	3.97E-08	0	0.17	0
GAT1	HAP2	RAPA	26	40	7	8.84E-09	0	0.14	0
MSN4	SKN7	H2O2LO	18	169	7	1.17E-05	0	0.14	0
ROX1	CIN5	YPD	63	141	14	1.21E-07	0	0.14	0
GZF3	GLN3	RAPA	32	62	8	4.97E-08	0	0.12	0
ROX1	NRG1	YPD	63	66	21	3.65E-22	0	0.1	0
GAT1	STP1	SM	44	63	25	1.53E-34	0	0.04	0

We note that there exists the issue of selecting thresholds at various stages of our approaches such as a ChIP-chip binding threshold or hypergeometric thresholds, which can result in different sets of individual predictions. Our arbitrary choice of thresholds should be reconsidered in future investigation, which might be a challenge in terms of comparisons of results. In addition, detailed mechanisms of each combinatorial TF pair are still unknown. For instance, we need to know whether a combinatorial pair binds the same promoter region by physically interacting with each other or binds two separate binding sites and then regulates target genes via yet another factor.

4.4 Condition-specific combinatorial regulation is statistically significant

Having identified condition-specific co-factors for condition-altered TFs, we ask to what extent condition-specific combinatorial regulation takes place in the cell and how significant it is. To this end, we test statistical significance of the condition-specific combinatorial regulation by calculating two statistics and randomizing condition-specific target gene sets of condition-altered TFs in pairs of conditions.

The two statistics are (1) DCR_1, the fraction of those condition-altered TFs for which any condition-specific co-factor is identified in any condition tested and (2) DCR_2, the average number of condition-specific co-factors for each condition-altered TF in all conditions tested. From the previous section, we obtain that 25 out of 30 condition-altered TFs have any condition-specific co-factor, or DCR_1 = 83%, for both the hypergeometric and Wilcoxon tests. And on average there are 3.8 and 6.3 condition-specific co-factors for each of those 25 condition-altered TFs, or DCR_2 = 3.8 and 6.3, from the hypergeometric and Wilcoxon tests respectively.

For randomization of target genes by a condition-altered TF in a pair of conditions, we partition the union of the two condition-specific gene sets (i.e., excluding the common genes in both conditions) into two random sets of genes corresponding to the two condition-specific gene sets. We then perform the second hypergeometric or Wilcoxon test for a population of those random condition-specific gene sets to identify condition-specific co-factors. Consequently, p-values of the two statistics being less than or equal

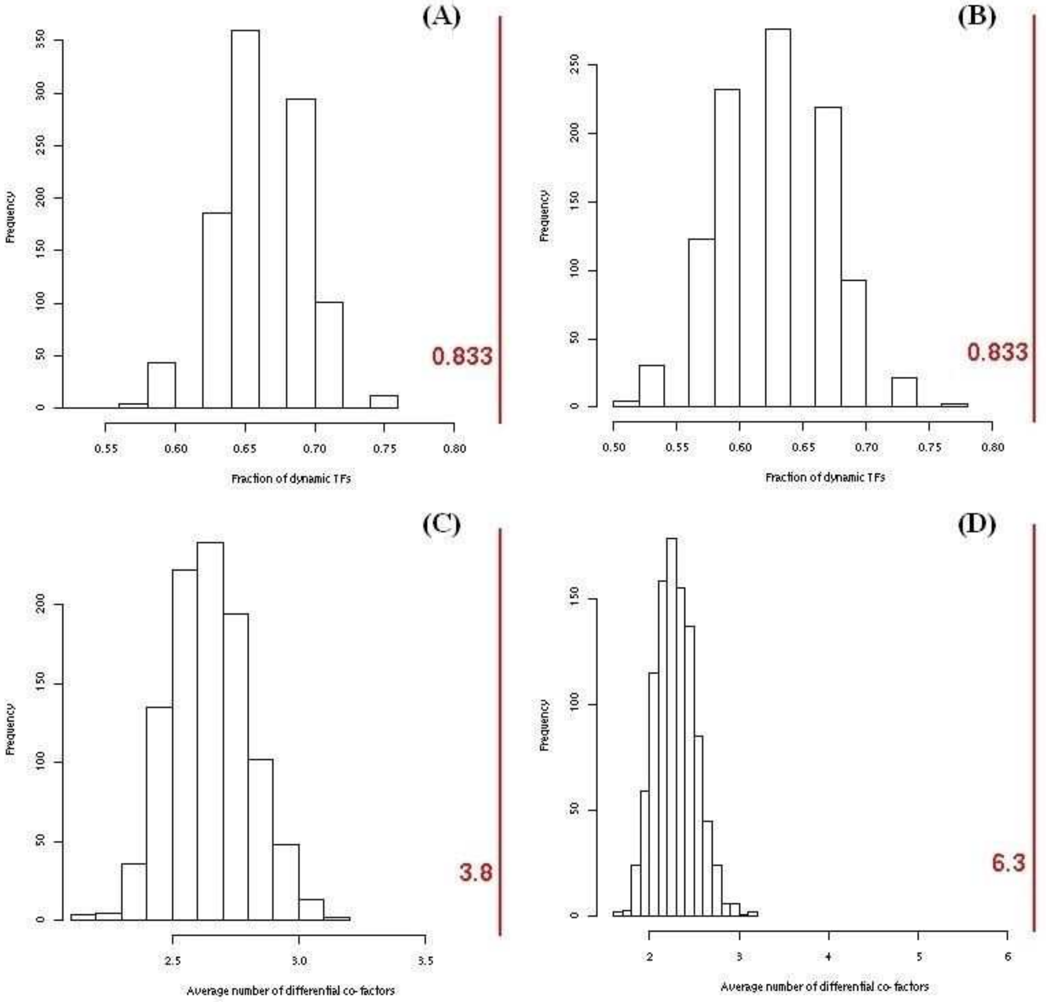

Figure 4.3: **Significance test.** The figures show results of significance tests for condition-specific combinatorial regulation. We considered two statistics, DCR_1 and DCR_2, to characterize condition-specific combinatorial regulation as described in the text. (A) and (B) are distributions of random DCR_1's by the hypergeometric and Wilcoxon tests respectively. (C) and (D) are distributions of random DCR_2's by the same two tests respectively. The red vertical lines and values are estimations from the real data we analyzed. Note that there are values which cannot be obtained in (A) and (B) because of discrete integer values in the numerators in calculating the fractions.

to random ones are obtained from corresponding empirical distributions.

As shown in Figure 4.3, all p-values are less than 10^{-3} for DCR_1 = 83% and DCR_2 = 3.8 and 6.3 from the hypergeometric and Wilcoxon tests respectively. Note that DCR_1 is not sensitive to the two methods (the hypergeometric and Wilcoxon tests), but DCR_2 is. Although we do not aim to investigate which test is superior in detecting condition-specific co-factors, the reader is reminded that the Wilcoxon test is applied to condition-specific gene sets of a condition-altered TF without a ChIP-chip binding threshold for a candidate co-factor. Another point to make is that all TFs assayed in the ChIP-chip experiments by Harbison et al. (2004) were selected on the basis of their known functions in the respective conditions except rich media. With this caveat in mind, the condition-specific combinatorial regulation we observed from the ChIP-chip data is statistically significant.

4.5 Support for condition-specific combinatorial regulation

Having established that condition-specific combinatorial regulation takes place in abundance, we now turn to inspection of individual predictions for supports from other diverse sources. Here we use expression data, conserved motifs data, and protein-protein interactions data, which provide complementary supports for our predictions based on ChIP-chip data.

4.5.1 Expression analysis

To give support for our predicted combinatorial TF pairs in Table 4.2, we first perform expression analysis of their target genes. We collected diverse expression datasets

ChIP-chip condition	Expression data sources	numbers of genes; samples compiled
Rich media (YPD)	elutriation assay (Spellman et al., 1998)	5397; 14
Moderately hyperoxic (H2O2LO)	1. H2O2 0.3mM upto 120 min (Gasch et al., 2000) 2. H2O2 0.3mM, 10 and 30 min (Carmel-Harel et al., 2001)	5194; 12
Amino acid starvation (SM)	amino acid plus adenine starvation upto 6h (Gasch et al., 2000)	3587; 5
Elevated temperature (HEAT)	heat shock data (Gasch et al., 2000)	4612; 19
Nutrient deprived (RAPA)	rapamycin treatment (Hardwick et al., 1999)	4364; 5
Mating inducing (ALPHA)	alpha-pheromone treatment (Roberts et al., 2000)	6181; 15

Table 4.3: **Expression data sources corresponding to ChIP-chip conditions.** The abbreviations of ChIP-chip conditions in the parentheses are consistent with Table 4.1.

from the Stanford Microarray Database (SMD, http://genome-www5.stanford.edu). We used R/G normalized mean values of mRNA expression and took the ratio of the respective mean values of R and G intensities if duplicate entries were observed. For consistency of experimental conditions between ChIP-chip and gene expression assays as discussed in the previous chapter, we attempted to retrieve expression data in similar conditions to ChIP-chip conditions. We were able to compile 6 expression datasets for 6 ChIP-chip conditions. For each dataset, we retain only those genes with upto 40% missing values and normalize expression ratios across all samples because it contains data from different sources or a subset of the whole data. The six datasets and the data sources are summarized in Table 4.3 together with corresponding ChIP-chip conditions.

One way to support combinatorial regulation by multiple TFs is to examine synergistic expression of their target genes. The algorithm by Pilpel et al. (2001) is one such method to address a combinatorial effect of two DNA motifs on expression coherence of genes with both motifs. Given two motifs, $m1$ and $m2$, they partitioned all genes with any of the motifs into three sets, G_{m1}, G_{m2}, and G_{m12}, which consist of genes with motif $m1$ alone, motif $m2$ alone, and both motifs $m1$ and $m2$, respectively. The two motifs are considered synergistic if G_{m12} with both motifs shows a better expression coherence score than the other two gene sets which have either motif alone. This algorithm can be equally applied to TF pairs using genome-wide ChIP-chip data, which were not yet available at the time of their work.

However, the partitioning strategy by Pilpel et al. does not necessarily address such synergistic or combinatorial effect by multiple motifs. First of all, G_{m1}, G_{m2}, and G_{m12} are disjoint, i.e., each partition has a *distinct* gene set. It is not clear that the two distinct gene sets with either motif have no functional role at all. If they do, they are likely to have a certain degree of expression coherence as well. In addition, their functions may be different from the function which is associated with the gene set with *both* motifs. Therefore, if those functions are different for those disjoint gene sets, comparisons of expression coherence among them are biologically implausible in principle. In the line of our arguments, the classification of promoter architecture by Harbison et al. (2004) provided a similar distinction. They identified a class of genes with a DNA binding site for a *single* regulator, termed ‘single regulator architecture’. Those genes were found to be involved in functions like carbon metabolism regulated by Gal4, amino acid metabolism by Gcn4, and glyoxylate cycle by Sut1. This indicates that those genes may have expression coherence too. In addition, they also identified another class of genes with multiple binding sites for multiple regulators subject to combinatorial regulation, termed ‘multiple regulator architecture’. Those genes were found to be involved in multiple metabolic pathways. Note that those two classes with a single TF and multiple TFs respectively are *independent* with *different* functions, hence cannot be compared for *better* expression coherence for instance. More TFs do not necessarily regulate target genes with better synergistic expression.

Hence, we suggest here an alternative approach to detect synergistic effects by multiple TFs on expression coherence of module genes based on our method of expression coherence assessment presented in Chapter 2. Given two TFs and two respective sets of target genes, we assess expression coherence of the common genes with respect to random sets of common genes sampled from each gene set for each TF. As in Chapter 2 we calculate the average of absolute Pearson correlation coefficients (ζ's) for all gene pairs in each set as our expression coherence score. Each TF gives a background distribution of expression coherence scores for those random sets of common genes. We then estimate two p-values of the expression coherence score of the real common gene set for two background distributions from the two TFs in question. Two TFs are deemed synergistic for the expression coherence of common target genes if those two p-values are significant (less than 0.05). This synergistic effect on gene expression is taken to be a support for combinatorial TF pairs.

Among our 104 predicted TF pairs in Table 4.2, three pairs are supported by our synergistic expression analysis (the column 'SEC' in the table). As an illustration, we predicted that Gat1 and Rcs1 are combinatorial regulators for 9 target genes in amino acid starvation condition from the ChIP-chip data. This prediction is supported by the synergistic expression analysis as shown in Figure 4.4. No reports have been made in literature about this TF pair. On the other hand, the other two pairs have been previously predicted computationally (Tsai et al., 2005). Using appropriate expression data in accordance with ChIP-chip experimental conditions as well as improved expression data with more samples or time points might further detect synergistic expression from our predicted TF pairs.

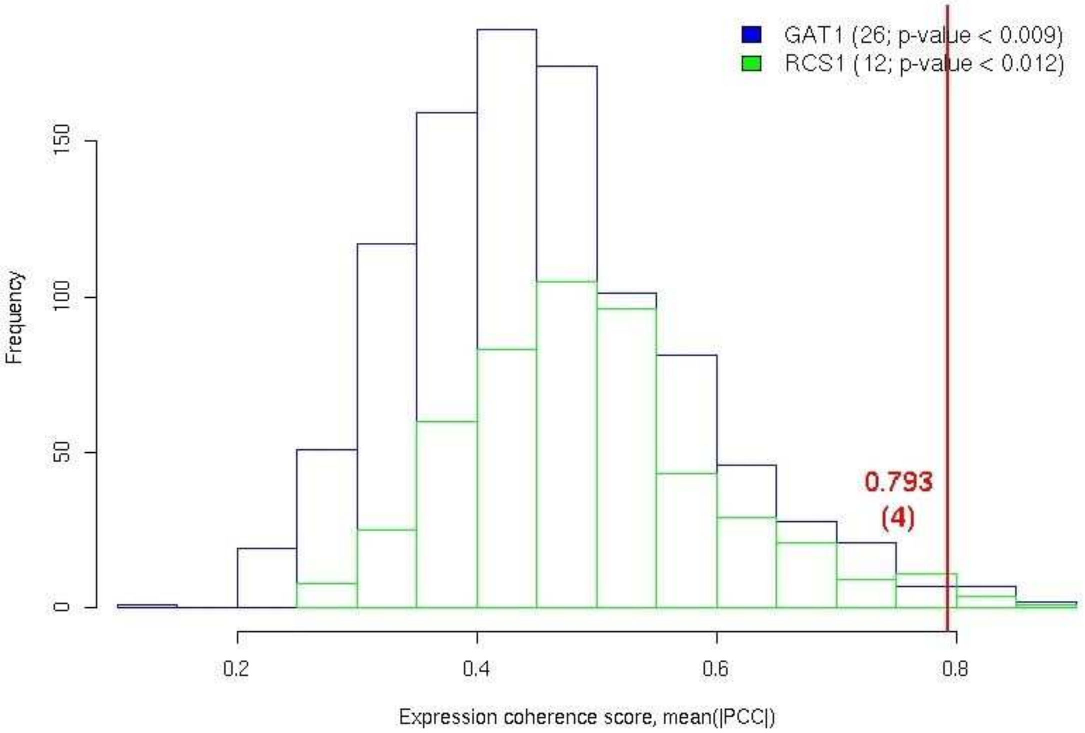

Figure 4.4: **Synergistic expression analysis.** The figure shows two distributions of expression coherence (EC) scores for random sets of 4 common genes of Gat1 and Rcs1, respectively. The red line and EC value are from real data. The two significant p-values for the two regulators (see the inset legend) suggest that their common genes are synergistically expressed by the two factors. Hence, this supports combinatorial activity of the two factors we predicted from the ChIP-chip data. Note that there are many genes missing in the expression dataset we used, so that the numbers of target genes by the regulators are less than those found in the ChIP-chip data. This resulted in 4 common genes with expression data instead of 9 from the ChIP-chip data, and in the smaller number of samples for Rcs1 ($\binom{12}{4} = 495$) than 1,000 random samples for Gat1.

4.5.2 Conserved motif

To check if there is any indication of conserved binding sites upstream of common target genes for both of each predicted TF pair, we used conserved motif data in S. cerevisiae by MacIsaac et al. (2006). They produced a refined version of regulatory interactions of Harbison et al. (2004) using two conservation-based motif discovery algorithms. We used their published results of 5201 TF binding sites of 116 TFs for 2343 target genes which are conserved in at least one other yeast species. Using these data, we calculated a fraction of common target genes of each combinatorial TF pair which contain conserved binding sites for both regulators. There are a total of 38 predicted TF pairs for which at least one target gene contains both conserved binding sites in their upstream regions (the column 'Motif' in Table 4.2). The number of predicted TF pairs for which more than 50% target genes contain both conserved binding sites is 7. We note that all target genes of the three TF pairs with synergistic expression in the previous subsection are not found to possess both conserved motifs except for one gene for the YAP5-MSN4 pair (1 / 16 $\sim$ 0.06 in Table 4.2). One may also perform independent motif analysis such as MacIsaac et al. (2006) for those predicted TF pairs and target genes, but it is beyond the scope of our work.

4.5.3 Protein-protein interaction

We also checked protein-protein interaction (PPI) data for predicted TF pairs from 5 published data sources (Gavin et al., 2006; Han et al., 2004; Krogan et al., 2006; Patil and Nakamura, 2005; Reguly et al., 2006), which resulted in 97821 unique PPI pairs in total. Those 5 studies include data from yeast two-hybrid experiments, tandem-affinity-purification coupled to mass spectrometry (TAP-MS), and literature curation. 13 predicted TF pairs are supported by those interaction data (the column 'PPI' in Table 4.2), but it should be noted that those pairs are not necessarily physically interacting partners because experimental conditions are not the same. We simply note that one can utilize protein-protein interaction data to give further support for combinatorial regulation of multiple TFs.

4.6 Summary

Transcription is often regulated by multiple transcription factors (TFs) concurrently. Such combinatorial regulation is a key mechanism in transcription. Genome-wide ChIP-chip data have shown that each transcription factor has distinct DNA-binding patterns in different environments. Combinatorial regulation by multiple regulators has been suggested as one possible mechanism for such changes of DNA-binding (and hence target genes) through interactions with each other. We investigated this scenario and focused on such condition-altered TFs and their interactions with co-factors in a condition-specific way. The hypothesis tested was that changes of target genes of each condition-altered TF are due to condition-specific co-factors with which it interacts in the respective conditions.

We employed two successive hypergeometric tests to identify condition-altered TFs and their condition-specific co-factors from ChIP-chip data. An alternative method was also proposed to identify condition-specific co-factors without using thresholds. We showed that such condition-specific combinatorial regulation is more predominant than expected by chance given the data. Our predicted combinatorial TF pairs were further inspected to obtain supports from gene expression, conserved motifs and protein-protein interactions data. Although supports from those data sources were weak, our approach was able to provide novel testable hypotheses about specific combinatorial TF pairs under a certain condition.