Chapter 3

Prioritization of gene regulatory interactions

3.1 Background

We extend our module analysis in the previous chapter to identification of refined regulatory interactions in modules. While computational methods have been developed to derive numerous modules from heterogeneous genome-wide data sources (Bar-Joseph et al., 2003; Ihmels et al., 2002; Lemmens et al., 2006; Segal et al., 2003; Tanay et al., 2004; Wu et al., 2006; Xu et al., 2004; Yu and Li, 2005) as discussed in Chapter 1, individual links between regulatory proteins and target genes still need experimental verification. Those studies mainly focused on identification of modules as independent or inter-connected functional units in regulatory networks (Chapter 1). However, experimentalists face the challenge to verify predicted modules in their functional contexts at the level of all individual links. This is currently impossible as the number of regulatory links in modules predicted from large-scale data analyses is in the order of thousands. In this chapter, we aim to provide a simple way to prioritize individual regulatory interactions in transcriptional modules as an attempt to overcome the exper-

imental issue.

The previous chapter investigated transcriptional modules by integrating three types of data sources: chromatin immunoprecipitation on microarray (ChIP-chip), gene expression and functional annotations. The data integration aims to compensate limitations of a single type of data source alone. ChIP-chip data alone do not possess functional regulatory information and gene expression data alone do not contain physical binding information. We have achieved such compensation by identifying modules using different data sources in the previous chapter. Such identified modules by data integration are likely to contain functional or regulatory (we use these two terms interchangeably) interactions between transcription factors (TFs) and target genes. Here we further develop a method to utilize those functional modules with a goal of prioritizing individual TF-gene regulatory interactions. This chapter is based on our published work in Lee et al. (2008).

3.2 Overview of our approach

This section provides an overview of our approach with a toy example as shown in Figure 3.1. Our approach starts with putative transcriptional modules (PTMs) derived from genome-wide ChIP-chip data (Step 1 in Figure 3.1). In the figure we show a toy example of ChIP-chip binary data matrix and corresponding 4 PTMs, M1 to M4. Each module contains a set of transcription factors (triangles) and a set of target genes (circles) connected by links between all of them. The genes in M2 and M3 are numbered for an illustration purpose below.

In Step 2 we identify a subset of PTMs which are (1) coherent in expression profiles of target genes and at the same time (2) enriched in functional categories. This subset of

PTMs are "coherent modules". That is, both gene expression and functional annotation data are used to extract functional signals after binding signals are retrieved from ChIP-chip data. In the given example in the figure, M1 and M4 are meant to be non-coherent and hence discarded altogether. Colours of genes in both coherent modules, M2 and M3, symbolize different functions. 'Blue' and 'red' functions are meant to be coherent (enriched) in the respective modules. Notice that the 'red' function is coherent in both modules. The fictitious 'M'-shaped expression profiles are also shown to be coherent as well in both modules. The red expression profile belongs to gene 5 which is annotated to the coherent 'red' function.

All links between TFs and target genes in those identified coherent modules are considered candidate functional links. The goal is then to narrow down those candidate functional links to core functional links in Step 3. Our key strategy is to focus on the intersections of gene sets of coherent modules for all enriched functional categories. This is illustrated by the gene 5 in Step 2 which belong to both coherent modules. The gene is annotated to the common coherent 'red' function in both modules. The union of regulators in M2 and M3 is predicted to functionally regulate the gene in this illustration. We term such genes "coherent linker genes". Notice that gene 6 belonging to both modules is not a coherent linker gene because its annotated 'yellow' function is not coherent in the modules. This short list of TF-gene pairs is our final list of predicted functional pairs and consequently has priority over the others in coherent modules for further mechanistic analysis or experimental validation.

Below we detail our approach in a formal way, show how to evaluate it, and provide a few case studies.

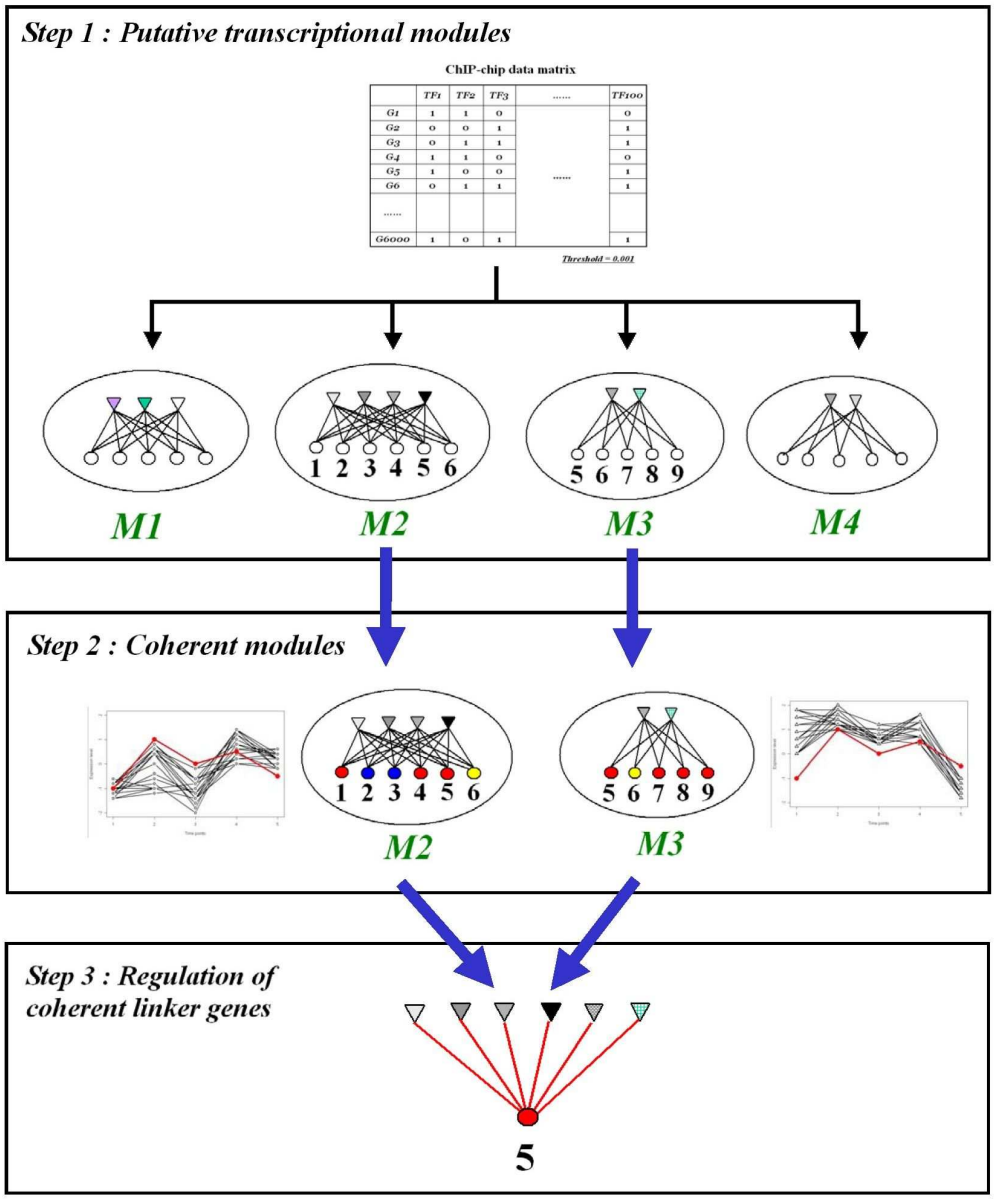

Figure 3.1: **Overview of our method.** This figure shows our 3-step approach to prioritize gene regulatory links from coherent modules. See the text for details.

3.3 Coherent modules

Our prioritization is based on functional modules which are defined by coherence of module genes in terms of their expression profiles and functional annotations as was done in Chapter 2. This section formalizes a method to derive putative transcriptional modules from ChIP-chip data and the way we define coherent modules, highlighting some differences from Chapter 2.

3.3.1 Putative transcriptional modules from binding data

In the previous chapter we utilized transcriptional modules derived by Manke et al. (2003) and here present a formal way of deriving them from a ChIP-chip data matrix with a slightly different prescription. In contrast to Chapter 2, we have a different constraint on modules and attempt to resolve the issue of redundant relationships among modules as we will explain below. We apply this method to the ChIP-chip data in rich media of Lee et al. (2002). Although this original dataset has been supplemented by new data with more TFs and conditions (Harbison et al., 2004), we mainly apply our method to the older data in order to compare our results with other methods (Subsection 3.5.3) which also used the same data of Lee et al. (2002). A brief comparison with another method using the updated binding data by Harbison et al. (2004) will be given in Subsection 3.5.4. We now explain our 2-step program to generate a set of putative transcriptional modules from binding data.

(1) Enumeration of large bicliques

Regulatory interactions between transcription factors (TFs) and target genes can be represented as a bipartite graph, with edges going from a set of TFs to a set of target genes. A biclique K is a bipartite graph such that an edge is realized from every vertex

of a TF set (F) to every vertex of a gene set (G), i.e.,

$$K = (F + G, E), \tag{3.1}$$

where E is a set of all possible edges from F to G. (i.e., $\|E\| = \|F\| \times \|G\|$). Input to our method is a set of bicliques. Generally, a bipartite graph will contain a large number of bicliques. Binding data typically are quite sparse, i.e., the number of edges in a bipartite graph is much smaller than the size of the entire TF set multiplied by the size of the entire gene set. For example, in the case of the ChIP-chip data introduced above, a p-value threshold of 0.001 results in a total of 4611 regulatory interactions and 584 bicliques generated by our program described below. Our program takes all TF-gene interactions in the ChIP-chip data at a binding threshold as true positives. Generally, a bipartite graph will contain a large number of bicliques. We have implemented a simple enumeration algorithm for large bicliques with the constraint that $\|G\| \geq 5$. This constraint is chosen to perform reasonable statistical assessment in our subsequent analysis of coherent modules in Subsection 3.3.2.

Let the set F of all factors be ordered. In the first pass of our program, each factor is inspected whether it is connected to 5 or more genes. These constitute the first set of (trivial) bicliques. The idea is then to extend those bicliques to find the bicliques with 2 factors, then with 3 factors, etc. Now assume that a set of all bicliques with m factors has been determined. The algorithm then runs iteratively through all the bicliques with m factors and adds an additional factor from the ordered list of factors to each biclique, if that factor targets 5 or more genes from the set of genes in the biclique in question. Thereby we obtain a new biclique with $m+1$ factors. The gene set of this new biclique is the intersection of the gene set in the old biclique and the set of target genes of the newly introduced factor. Since this procedure observes the order of factors, bicliques are not discovered repeatedly. However, at each step the algorithm may generate a new

biclique with an identical set of genes already contained in the old biclique, in which case we discard the old one. Notice that this prescription may still result in bicliques with the same set of genes after the whole iteration has finished.

(2) Putative transcriptional modules
Since our subsequent analysis will deal only with the gene sets induced by the bicliques derived above, we first merge those redundant bicliques which contain identical sets of genes, so as to avoid any computational overhead. The merged biclique is designed to have those transcription factors which belong to two or more of the redundant bicliques (i.e., TFs with multiplicity $>= 2$). This merging procedure was not made in the previous chapter, where we observed subset relationships of TFs among modules (see Figure 2.2) and treated those modules as independent. In this way, we generated 584 non-redundant bicliques from the ChIP-chip data by Lee et al. (2002), the maximum number of TFs in a biclique being 7. We also call them putative transcriptional modules (PTMs), and they are the input to our subsequent analysis of coherent modules (see Step 1 in Figure 3.1).

3.3.2 Defining coherent modules

Having a set of PTMs from binding data, we use gene expression and functional annotations to identify a subset of functional PTMs which define coherent modules in a similar manner to Chapter 2. It is biologically important to have independent experimental datasets in which cellular conditions are comparable if one tries to integrate them for analysis. Therefore, we selected gene expression data in view of the experimental conditions of the ChIP-chip data we used in this study. Gene expression experiments were extensively done in many diverse conditions on a genome-wide scale when compared to ChIP-chip experiments. ChIP-chip assays by Lee et al. (2002) were conducted in rich media and we focused on elutriation conditions (size-based synchronization of cell

cycle) in expression data by Spellman et al. (1998). Two other methods they used for the synchronization of cell cycle were involved with alpha-factor pheromone treatment and temperature-sensitive cdc15 mutation, which introduced characteristic artifacts of mating and heat shock respectively (Spellman et al., 1998). Those artefacts are not expected in the conditions of the ChIP-chip assays. Therefore, we used the elutriation data as the experiment was not involved with such artifacts. The data consist of 14 time points taken every 30 minutes for 6.5 hours. As for functional annotations, we used the MIPS categories (Mewes et al., 2004) as in Chapter 2.

Given a PTM and an expression dataset, we use a p-value of Eq. (2.2) in Chapter 2 to assess expression coherence of module genes. Here we attempt to use p-values, p_e, without bothering the problem of multiple testing because there is no good principle of p-value correction in a biological context. And the number of random modules, K, is 1,000 in this chapter. Functional coherence in a PTM is assessed by the p_f in Eq. (2.4) the same way as in Chapter 2.

A PTM is then called a coherent module if both p-values, p_e and p_f, are less than two thresholds, τ_e and τ_f, for expression and function coherence test respectively (Step 2 in Figure 3.1), i.e.,

$$CM = \{\, TM = (F, G) \mid p_e < \tau_e \, and \, p_f < \tau_f \,\} \,. \tag{3.2}$$

We now have coherent modules (CMs) which are derived by integrating three different data sources. Note that our module analysis in the previous chapter was concerned with SSMs (Eq. 2.3) which were not required to have functional coherence. Our goal of prioritization in this chapter is to utilize those CMs by examining functional relationships among CMs as presented below.

3.4 Prioritization of gene regulatory links

For a given list of coherent modules (CMs) at two parameter thresholds in Eq. (3.2), we now focus on those functional categories which are detected as coherent in *multiple* CMs. For a particular coherent function, we identify all CMs which *share* that function. Then, we identify *common* target genes in those CMs which are annotated to that function. We refer to this identification step as "functional intersection". Those filtered genes are called "coherent linker genes" as they link CMs. It should be noted that we require those coherent linker genes to appear in *all* those CMs. In other words, they are claimed to possess the strongest functional signal among other genes in CMs. Regulation of coherent linker genes by associated TFs in corresponding CMs constitutes our prediction of functional TF-gene pairs (Step 3 in Figure 3.1).

In contrast to earlier works on identification of functional modules themselves (or CMs in our case), our prioritization aims to identify highly reliable TF-gene functional links by way of functional intersection of CMs. The criterion of the most reliable functional links is based on the following measure of prediction accuracy, called positive predictive value (PPV), given a reference dataset. A PPV is defined to be the number of true positives (i.e., predicted TF-gene individual pairs that are found in a reference set) divided by the number of predicted TF-gene pairs. As a reference dataset, we used a set of 3962 TF-gene pairs which are combined results of literature collection and conserved motifs analysis. Further details on performance measures and reference data are given in the next section.

The prediction accuracy of PPV is dependent on the two p-value threshold parameters to define coherent modules in Eq. (3.2): one for expression coherence (τ_e) and the other for function coherence (τ_f). Given two threshold parameters, we perform functional intersection to identify coherent linker genes to predict regulatory TF-gene pairs.

		τ_f			
		0.001	0.005	0.01	0.05
τ_e	0.001	36	38	39.9	44.8
	0.005	41.5	44.1	45.8	53.1
	0.01	40.2	43.1	45.1	51.1
	0.05	40.2	43.5	45.8	48.5

Table 3.1: Positive predictive values (PPVs) from 16 combinations of the two parameters, τ_e and τ_f. The parameter combination of τ_e = 0.005 and τ_f = 0.05 corresponding to the highest PPV is the one we used for subsequent analysis.

As an attempt to optimize our PPV measure, we varied the two parameters by taking all combinations of four significant thresholds : 0.001, 0.005, 0.01, and 0.05. Then, PPVs were calculated with respect to the reference dataset of 3962 TF-gene pairs. In this work, we report all results based on τ_e = 0.005 and τ_f = 0.05, which gives the highest PPV among the 16 combinations (Table 3.1).

With this combination of p-value thresholds, we obtained 89 coherent modules with a total of 47 coherent functional categories (out of the total 557 modules tested). 20 out of the 47 enriched functions are shared by at least two of 42 coherent modules with common target genes, i.e., coherent linker genes. This functional intersection resulted in 66 coherent linker genes and 18 associated TFs, yielding 177 TF-gene functional pairs (Figure 3.2). Notice again that coherent modules themselves are not the focus of our analysis.

3.5 Evaluation of the method

We evaluated our method in two ways. First, as validation of the method, we checked if our method increases prediction accuracy in comparison to ChIP-chip data alone. Second, we compared results of our method with those of other algorithms by calculating performance measures. As performance measures, we calculated (1) positive predic-

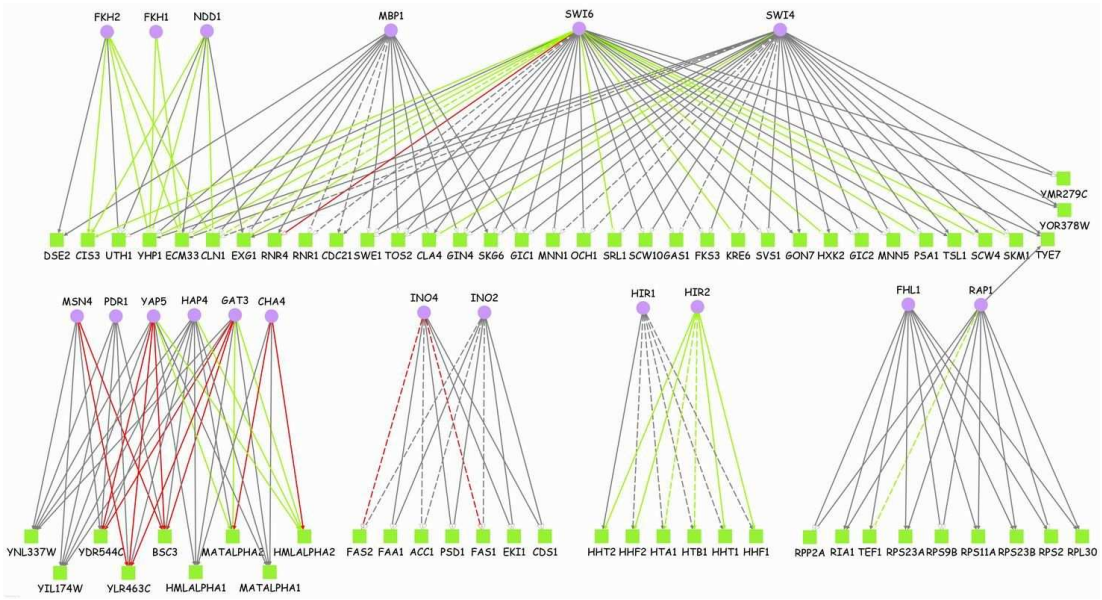

Figure 3.2: **Summary of our final predictions.** This network diagram is a graphical representation of our final predicted 177 TF-gene functional pairs between 18 TFs and 66 target genes (coherent linker genes). Dashed links show 24 literature-verified pairs. Links with white arrow heads represent 85 pairs with conserved motifs. Green and red links show additional information about expression correlation between TF-encoding genes and their target genes: high positive (Pearson coefficients > 0.661) and negative correlation (Pearson coefficients $<$ -0.628) respectively. For example, the predicted functional pair of Rap1 and TEF1 (on the bottom right) is confirmed with respect to literature and conserved motifs. In addition, the pair shows high expression correlation between them. We use this additional information for detailed case studies by considering those pairs as high confident among all our predictions (see Section 3.5). Generated by Cytoscape (http://www.cytoscape.org/) and yED graph editor (http://www.yWorks.com)

tive value (PPV) and (2) sensitivity (SNST) which are defined as the number of true positives (predicted TF-gene individual pairs that are found in a reference set) divided by (1) the number of predicted TF-gene pairs (2) the number of all reference TF-gene pairs. Notice that true negatives cannot be defined because there is no reference for the *absence* of regulatory relationships between TFs and genes. Two reference datasets we used for evaluation of our method are first presented below.

3.5.1 Reference datasets

(1) Literature collection

The first reference set we used is 1207 TF-gene pairs compiled from three literature-curation sources : (1) Lee et al.'s curation of 1049 pairs excluding computational regulatory motif results (Lee et al., 2002) (2) TRANSFAC database for 342 pairs (Matys et al., 2003) (version 10.4) (3) Siddharthan et al.'s curation of 72 pairs (Siddharthan et al., 2005). Notice that the reference data may contain TF-gene pairs where TFs act as mere DNA-binding factors rather than functional regulators.

(2) Conserved motifs

The laboratory of Richard Young recently advanced their ChIP-chip technology and applied it to yeast with 203 TFs (Harbison et al., 2004) (compare with 106 TFs in Lee et al. (2002)). Based on their binding data and sequence data from four yeast species, they identified conserved binding motifs for 102 TFs using a variety of motif detection algorithms (this was not done in the work of Lee et al. (2002)). It is widely believed that conserved motifs across species indicate their functional roles (Cliften et al., 2003; Doniger et al., 2005; Kellis et al., 2003). While the 'phylogenetic footprinting' approach will introduce errors, it provides a more comprehensive picture of regulatory links than manual curation of literature. Hence, we take the dataset of conserved motifs as a second reference set independently of the literature-based reference. We compiled

2922 TF-gene pairs from the motif analysis results of Harbison et al. (2004) maintained in the Saccharomyces Genome Database (SGD; ftp://ftp.yeastgenome.org/yeast/). The list of 2922 pairs is derived from predicted binding sites which are conserved in at least two *Saccharomyces* species, other than *S. cerevisiae*. Note that this set contains more than twice as many predicted interactions as the literature reference set.

3.5.2 Validation

For the purpose of validation of our method, we compared the performance measures from our predicted TF-gene pairs and the original ChIP-chip data we used at a binding p-value threshold of 0.001 (4611 TF-gene pairs between 96 TFs and 2326 genes). We removed all uncharacterized genes from the ChIP-chip results for the purpose of validation to avoid a possible bias of our method towards annotated genes resulting from MIPS functional data we integrated. This leaves us with 3598 TF-gene pairs between 95 TFs and 1837 genes. As shown in Table 3.2, we obtained higher PPVs at the expense of lower SNSTs, which is to be expected as we aimed at prioritization of regulatory links.

In addition, we investigated whether coherent modules themselves or functional intersection alone could have given us better performance than our combined strategy. We validated each of the two steps separately, (1) identification of coherent modules (CMs) and (2) functional intersection among CMs. The two performance measures were calculated and compared with our predictions from the combined strategy by taking (1) *all* TF-gene pairs from CMs *themselves* and (2) TF-gene pairs from functional intersection among *PTMs*, respectively. First, taking *all* pairs in CMs *without* functional intersection does not yield higher PPVs at the expense of SNSTs for both reference sets (column 'CM' in Table 3.2), indicating that functional intersection is an important step. Second, we took TF-gene pairs from functional intersection of the initial

PTMs (584 modules derived from the ChIP-chip data) *without* applying the expression coherence test. This functional intersection from the initial PTMs yields higher PPV than our predictions (18.2% vs. 13.6%) for the literature reference but lower PPV than ours (24.5% vs. 48%) for the conserved motif reference (column 'FI_TM' in Table 3.2). This suggests that functional intersection is the key to good performance with respect to literature which consists of experimentally verified interactions. However, using conserved motifs as a reference, the PPV (24.5% after functional intersection) is lower than the PPVs from either the ChIP-chip results alone (32.7%) or CMs above (35.8%) (see Table 3.2). On the other hand, the SNSTs after this functional intersection are lower than our predictions for both reference sets (Table 3.2). Therefore, the combination of both prescriptions is important for detecting regulatory signals from ChIP-chip data.

3.5.3 Comparison with other methods

As a second evaluation of our performance, we compared our predicted TF-gene pairs with those of the previous two algorithms : GRAM (Bar-Joseph et al., 2003) and MA-Networker (Gao et al., 2004), using the results provided in their original papers. Bar-Joseph et al. used the same ChIP-chip data along with a compiled expression dataset (over 500 conditions) to produce clusters of genes and regulators. We took TF-gene pairs in their final 106 clusters in rich media conditions. Gao et al. also used the same ChIP-chip data along with a compiled expression dataset (over 700 conditions). Their algorithm aimed to identify functional and non-functional target genes based on TF activity profiles they inferred using a multivariate regression model. We used the results of functional target genes and their TFs for comparison.

The GRAM algorithm predicted 1518 TF-gene pairs (in rich media condition) and the MA-Networker 1272 pairs. 66 pairs from GRAM and 67 pairs from MA-Networker

	ChIP-chip	CM	FI_TM	Our_final	GRAM	MA-Networker
PPV (%) (lit; con_mot)	4.6; 32.7	6.0; 35.8	18.2; 24.5	13.6; 48	6.3; 24.6	6.5; 38.6
SNST (%) (lit; con_mot)	13.7; 40.2	4.2; 10.5	1.7; 0.9	2.0; 2.9	7.9; 12.8	6.9; 16.8
N_pairs	3598	857	110	177	1518	1272
N_genes	1837	393	44	66	655	989
N_TFs	95	24	30	18	69	36

Table 3.2: **Comparison of performance measures.** For evaluation of our method, we compared our predicted TF-gene functional pairs (fourth column, 'Our_final') with ChIP-chip results with annotated genes only (first column, 'ChIP-chip'), two prescription steps of our method (second and third columns, 'CM' and 'FI_TM'), and two other previous algorithms (fifth and sixth columns, 'GRAM' and 'MA-Networker'). Two performance measures were calculated, PPV and SNST for two reference datasets. CM = all TF-gene pairs from coherent modules; FI_TM = TF-gene pairs from functional intersection among the initial putative transcriptional modules from ChIP-chip; lit = literature reference; con_mot = conserved motif reference; N_pairs = number of TF-gene pairs; N_genes = number of genes in the pairs; N_TFs = number of TFs in the pairs. See the text for details.

overlap with our 177 pairs and 39 pairs were predicted by all the three algorithms (469 pairs overlap in GRAM and MA-Networker). We observe that our method has higher PPV than the two methods and lower in SNSTs for both reference sets (columns 'GRAM' and 'MA-Networker' in Table 3.2). On the other hand, the two overlaps of 66 pairs and 67 pairs with the two algorithms give rise to yet higher PPVs with respect to the literature reference : 27% and 18% respectively. For the conserved motif reference case, the overlaps yield 50% and 45% PPVs respectively, which are similar to our performance of 48%. Of the 39 pairs predicted by all three algorithms, 10 pairs are found in the literature reference and 16 pairs in the conserved motif reference (25% and 41% PPVs respectively).

To illustrate the generic applicability of our approach, which does not depend on our definition of modules, we applied the functional intersection to the 106 final modules of the GRAM algorithm. This may be considered as analogous to our expression coherent modules in the absence of incorporation of functional annotation data. Then, PPVs were calculated and compared with those of their final modules for the two reference sets. The functional intersection yielded 23 pairs between 13 TFs and 9 genes (i.e., coherent linker genes) with higher PPVs than their own modules; 43.5% and 30.4% for the literature and conserved motif reference sets respectively (as compared with 6.3% and 24.6% in row 'PPV' and column 'GRAM' in Table 3.2). This illustrates that our approach of functional intersection may be applied to any set of modules identified in other works to yield more reliable regulatory links.

So far our analyses were based on the ChIP-chip data by Lee et al. (2002) to compare with the two algorithms, GRAM and MA-Networker. In order to check our performance using different datasets and compare with a more recent study, we applied our method to the updated and larger ChIP-chip dataset by Harbison et al. (2004) and com-

pared with the recent algorithm, ReMoDiscovery (Lemmens et al., 2006) which used that dataset. Lemmens et al. (2006) developed a module discovery algorithm (similar in spirit to the GRAM algorithm) which integrates ChIP-chip, gene expression and (in contrast to our method, GRAM and MA-Networker) conserved motif datasets in a concurrent way. By applying our method to the same ChIP-chip and gene expression data (Harbison et al., 2004; Spellman et al., 1998) as in their study, we predicted 108 regulatory interactions and yielded 14.8% PPV with respect to the literature reference (we did not consider the conserved motifs from Harbison et al. as a reference as this would be circular). For a comparison, we used their "seed modules" which contain 134 TF-gene interactions, a comparable number of predictions to ours. Their 134 predictions yielded 12.9% PPV with respect to the literature reference. Although the prediction accuracies are similar, there is only little overlap between the predicted sets of regulatory interactions (9 interactions in common, 3 of them are found in the literature reference), indicating the complementarity of these two methods.

3.5.4 Difficulty of comparisons

In general, it is difficult to directly compare the performance of different algorithms which are designed for different purposes. Our comparison of published results highlights the fact that different approaches have so far been used with different aims and yield different trade-offs between specificities and sensitivities. A more comprehensive evaluation study would require re-running different algorithms in different regions of parameter space. Notice though that in this work we did not vary p-value thresholds of ChIP-chip results to adjust PPV or sensitivity as was done by Bar-Joseph et al. (2003), for instance.

Therefore, we caution the reader about the interpretation of our comparison results. The increase of our PPV at the expense of sensitivity against GRAM and MA-Networker

should be expected considering the fact that we integrated one additional data source of functional annotation with the two data sources of ChIP-chip and gene expression which the other two algorithms used for their predictions. One point to make, however, is that while we utilized functional annotation data for the purpose of prediction, they used annotation data for validation of their prediction. Note also that because the validation using annotation data involves over-representation or enrichment of genes in sets of genes, it cannot serve for validation of all predicted functional target genes. In addition, while those works utilized gene expression data to derive coherent modules from ChIP-chip binding data, their published work did not focus on individual regulatory interactions. Their predicted interactions are simply all members of statistically predicted modules themselves. In contrast, our predictions do not exclusively aim at modules, but individual regulatory interactions, which we obtained by means of functional intersection. By this prioritization approach we purposefully predicted less functional associations (less sensitivity), but doubled PPV with respect to the literature reference (Table 3.2). Although this validates our approach, it may illustrate a limitation of the literature reference which covers only a fraction of all experimentally verified genes to date. Because of this limitation we also compared the different methods with respect to a more comprehensive reference set of predicted regulatory interactions. These predicted interactions are based on updated ChIP-chip data and sequence conservation across other yeast species. We took them as an indication for functional interactions. Using this reference set, we achieved 48% as compared to 25-39% from the two other works (Table 3.2). We stress that all methods compared here have their own specific aims and merits although they share the overall goal to derive functional interactions from physical interactions (as provided by ChIP-chip).

3.6 Biological examples

We now continue with detailed inspections of some of our systematic results shown in Figure 3.2. It is well known that activity profiles of TF proteins are not necessarily reflected in expression profiles of the corresponding genes because of post-transcriptional and post-translational regulations of TFs (Greenbaum et al., 2002). We took, however, any such correlation as an additional indicator of a functional relationship among our predictions and aimed at identifying all TF-target pairs with high correlation for detailed analysis. To this end, we calculated Pearson coefficients for our predicted TF-gene pairs and compared them with a background distribution of Pearson coefficients for all pairs between $\sim$200 TFs of Harbison et al. (2004) and all other genes. By taking those observed pairs whose coefficients fall within 5% of both tails from the distribution of all the coefficients (the two thresholds being 0.661 and -0.628), we obtained a list of 46 highly correlated pairs between 13 TFs and 27 target genes: 33 positively and 13 negatively correlated pairs (green and red links in Figure 3.2 respectively). In the following we restricted ourselves to some of these more specific TF-gene pairs.

3.6.1 Functionally interacting proteins

As an application from our functional TF-gene predictions, the 46 pairs with high expression correlation can provide a basis for identifying functional interactions of proteins. We hypothesize that those target genes regulated by the same TF(s) with high expression correlation have related roles in more specific biological processes than those encapsulated by the 3rd level MIPS category. In Figure 3.2, we observe that some groups of genes are highly correlated with their common TFs. They include known examples such as the associations between Hir2 and the six histone genes (Prochasson et al., 2005), and the known role of Ino4 in the regulation of FAS1 and

FAS2 (Schweizer and Hofmann, 2004).

As another such group of genes, our method yielded a group of 5 genes, KRE6, EXG1, SCW4, PSA1 and HXK2, which are highly correlated with their common regulator Swi6 (Figure 3.2). All these genes share a high-level annotation of 'C-compound and carbohydrate metabolism'. There is no literature evidence for the transcriptional regulation by Swi6, but all genes were found to have binding sites of Swi6 conserved in at least one other yeast species (Harbison et al., 2004). Previous experimental studies show that 4 out of the 5 gene products, Kre6, Exg1, Scw4 and Psa1, are related to the cell wall synthesis and that cell wall genes are controlled by cell cycle progression where Swi6 has a regulatory role (Lesage and Bussey, 2006; Zhang et al., 1999). The 4 proteins are specifically implicated in synthesis of either glucose chains (glucans) or mannose-bound proteins (mannoproteins) which are two main inter-connected components of the cell wall.

The remaining protein, Hxk2 (hexokinase 2), is known to be a major upstream regulator of the glucose signalling pathway, which also impedes on cell wall genes. Specifically, a glucan synthase subunit, Gsc2, is regulated by Hxk2 via Snf1 and Mig1 (Lesage and Bussey, 2006; Rolland et al., 2002). Hence, it is possible that Hxk2 is functionally related to the 4 other gene products through glucose regulation and utilization for glucan synthesis. Glucose signalling is also known to act downstream on the cell-cycle, although the precise mechanisms are not yet fully understood (Newcomb et al., 2003). Our result may suggest a possible feedback onto glucose regulation through the regulatory interaction of Swi6 with HXK2.

3.6.2 Conserved binding sites for three regulators of CIS3

We predicted two target genes CIS3 and UTH1 regulated by three TFs, Swi6, Fkh2 and Ndd1. The expression profile of CIS3 (glycoprotein-encoding gene in cell wall) is highly correlated with all those three TFs (Pearson coefficients are 0.856, 0.801 and 0.765, respectively), which additionally supports functional regulation of the gene by the three TFs. On the other hand, UTH1 is not well correlated with the TFs (Pearson coefficients are between -0.1 and 0.3), hence we do not postulate a functional interaction between CIS3 and UTH1, in contrast to the analysis in the previous subsection. While conserved binding sites for all the three TFs were found upstream of UTH1, Harbison et al. (2004) did not identify any conserved binding sites upstream of CIS3.

As we predicted that the three TFs functionally regulate CIS3, we searched for any putative binding sites of those TFs and their conservation across species in the upstream region of the gene. To this end, we used the matrices for Swi6, Ndd1 and Fkh2 provided by Harbison et al. (2004) and scanned the 1kb upstream region of CIS3 for matrix hits above the balanced thresholds introduced by Rahmann et al. (2003). We set the GC content of the background model to 50%. All putative binding sites detected are located within 34 base pairs (Figure 3.3). For the investigation of conservation of the putative TFBS region, we used the fungal sequence alignment tool in SGD (http://yeastgenome.org/) and found a high degree of conservation for 4 orthologous upstream regions (Figure 3.3).

It is worth noting that Ndd1-Fkh2 interactions have been suggested to be important in regulating G2/M-specific genes in cell cycle together with the MADS box protein, Mcm1, forming a permanent protein-DNA complex (Koranda et al., 2000). In fact, the position specific frequency matrix of Ndd1 from the study of Harbison et al. (2004) is very similar to that of Mcm1, so we were able to detect a binding site overlapping

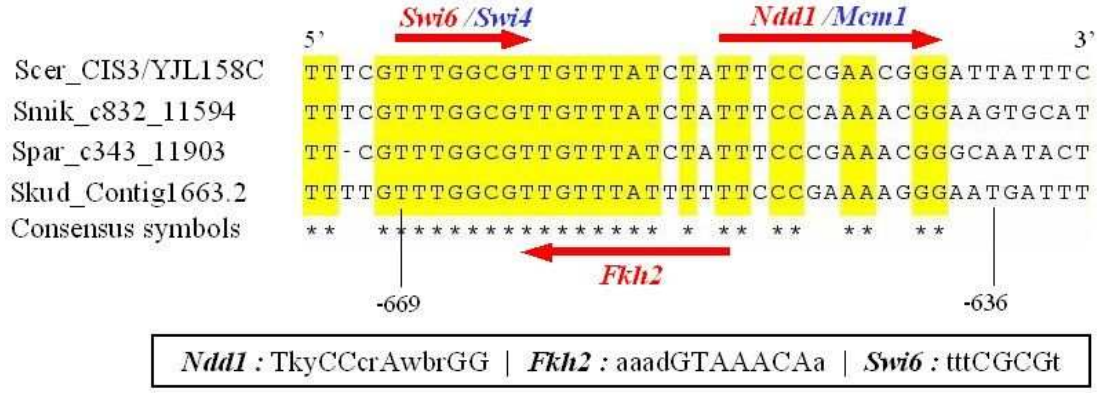

Figure 3.3: **Conservation of TF binding sites of predicted regulatory interactions** The figure shows alignment of 4 orthologous 1kb promoter regions for CIS3 and TF binding sites. We show only the region (-636 to -669 upstream of the TSS) which contains conserved binding sites for all predicted regulators (Swi6, Fkh2 and Ndd1). We used the six-species alignment from SGD (including S. mikatae, S. paradoxus and S. kudriavzevii), but removed from the SGD output S. castellii which is more distant, and S. bayanus which has a small intergenic region of only 30 nucleotides. Consensus motifs of the TFs from Harbison et al. (2004) are shown in the box using the IUPAC code. We denote degenerate binding sites of Swi4 and Mcm1 in blue. See the text for more details.

with that of Mcm1 (Figure 3.3). This indicates that Ndd1 could act as a functional co-factor, which does not necessarily bind DNA and cannot be distinguished from Mcm1 in ChIP-chip assays and motif scans. Hence, our method using gene expression correlations between TFs and their regulated genes is useful to detect a functional regulator (Ndd1) rather than a DNA-binding factor (Mcm1). Similarly, Swi6 is known to have a regulatory function forming SBF or MBF complexes with Swi4 or Mbp1 respectively (Iyer et al., 2001). We found a binding site of Swi6 overlapping with the binding site of Swi4 which is known to be a DNA-binding factor (Figure 3.3). As before, it may not be possible to differentiate between the binding properties of these two factors from binding data alone. These inspections show that our method correctly predicted TFs which have a regulatory function among the components of the TF complexes, even though the regulatory relationship may be indirect. Furthermore, the two complexes, Swi6-Swi4 (SBF) and Ndd1-Fkh2-Mcm1, may interact with each other through Fkh2 on the basis of the identified binding sites. On the other hand, a previous study on cell cycle by the Young laboratory identified CIS3 as a target of two cell cycle activators, the SBF complex and Fkh2, but not as a target of the Ndd1-Fkh2-Mcm1 complex (see Table 1 in Simon et al., 2001). Hence, our results suggest a new regulatory link between the Ndd1-Fkh2-Mcm1 complex and CIS3. It might also be the case that Fkh2 recruits either the SBF complex or the two other components of the Ndd1-Fkh2-Mcm1 complex according to distinct cell-cycle phases. Taken together, this detailed investigation highly supports our prediction of the functional regulatory links between CIS3 and the three TFs.

3.7 Summary

In this chapter we proposed a simple method to obtain reliable individual TF-gene regulatory interactions from functional modules. Starting with putative transcriptional

modules from ChIP-chip data, we first derived modules in which target genes show both expression and function coherence. The most reliable regulatory links between transcription factors and target genes were then established by identifying intersection of target genes (coherent linker genes) among coherent modules for each enriched functional category. We demonstrated that our method increased the fraction of functional interactions with respect to two different reference datasets of literature and conserved motifs, at the expense of sensitivity. Finally, we investigate our predictions in more detail and focus on those predicted TF-gene pairs whose expression profiles are highly correlated with each other. This further enables us to suggest functional interaction among gene products and novel conserved binding sites for those pairs.

By the design of functional intersection, our predictions suggest multiple transcription factors for each gene. This could be taken as a sign of combinatorial regulation where each factor is not sufficient to regulate their common target gene. The inference of combinatorial regulation requires further analyses, such as the vicinity of binding sites or a comparison of expression coherence of target genes by a set of multiple factors with that by each of the factors individually. Our predicted list of multiple transcription factors did not result from such analyses since we did not pursue the issue explicitly. Combinatorial regulation by multiple transcription factors will be investigated in the next chapter.