Bioinformatics and Biosystems 2007, Vol. 2, No. 2, pp. 62-65

Inference of Genetic Regulatory Modules Using ChIP-on-chip and mRNA Expression Data

Hyeyoung Cho, Doheon Lee

Department of Bio and Brain Engineering, KAIST

Abstract

We present here the strategy of data integration for inference of genetic regulatory modules. First, we construct all possible combinations of regulators of genes using chromatin-immunoprecipitation(ChIP)-chip data. Second, hierarchical clustering method is employed to analyze mRNA expression profiles. Third, integration method is applied to both of the data. Finally, we construct a genetic regulatory module which is involved in the function of ribosomal protein synthesis.

Keyword: Genetic Regulatory Modules, ChIP-chip, mRNA expression profiles

Introduction

Inference of genetic regulatory modules can help to reduce genetic network complexity without significant loss of explanatory power. Gene modules can be defined in the sense that they are co-bound by the same set of transcription factors and are co-expressed with the same expression pattern at the same time. This can be viewed as that the genes in the module are co-regulated, and hence likely to have a common biological function.

Expression profile reflects functional changes in mRNA levels in different conditions. On the other hand, genome-wide binding data suggests other point of view, since this data provides direct evidence of physical interactions. These two data sources can offer complementary information.

To determine binding events in genomic location data, researchers have previously used a statistical model and chosen a relatively stringent P-value threshold with the intention of reducing false positives at the expense of false negatives. However the P-values form a continuum and a strict threshold is unlikely to produce good results. In the case of gene expression profiles, the number of clusters is determined quite arbitrarily due to the inherent nature of clustering algorithm.

In the work of Bar-Joseph et al., in 2003, they introduced an algorithm which integrated genome-wide binding and expression data and finally showed an improvement than using either data source alone. This is a sort of following for the Bar-Joseph's work. In order to construct genetic regulatory modules biologically more relevant, we try to utilize genome-wide binding data and expression profiles together without determining parameters of those explicitly.

Corresponding author : Doheon Lee
(Email: doheon@kaist.ac.kr)
This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program (No. 2005-01450), and the Korean Systems Biology Research Grant (2005-00343).

Method

1. Data sets

We utilize ChIP-on-Chip data (Lee et al., 2002) which contains genome-wide binding information of 113 regulators of S. cerevisiase. We choose expression data (Spellman et al., 1998) which contains 6316 whole genomic profiles with 7 time points of S. cerevisiase. As a preprocessing, we filter out genes with missing values, small variance over time, very low absolute expression values and those with low entropy of profiles giving rise to remain 683 genes for further analysis.

2. Algorithm overview

The overall architecture of our strategy is shown in Figure 1. We describe the detail for each step as in the following:

Step1. Genome-wide DNA binding data.
For every 113 regulators, all the genes assumed to be bound by regulators are marked with one with the threshold of P-value less than 0.001. With these initial sets of binding data, we construct all possible combinations of regulator. (Fig.1. (a))

(a)

	R1	R2	R3	...	Rn
G1	1	0	1	...	0
G2	0	1	1	...	0
G3	1	0	0	...	1
...	...	...	...	...	...
Gm	1	1	0	...	1

(b)

C1	genes...
C2	genes...
C3	genes...
...	...
Cq	genes...

(c)

IM1	G11,G35,...	C11
...	...	...
IM51	G2,G32,...	C11
IM52	G42,G72,...	C11
...	...	...
IMp	G5,G16,...	C2,Cq

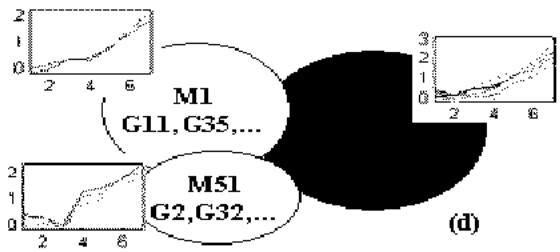

Fig. 1 The overall architecture of construction of regulatory modules. See text for details.
(a) Genome-wide DNA binding data. G and R denote a gene and a regulator respectively.
(b) Clustering results of gene expression profiles. C stands for a cluster.
(c) Intermediate modules constructed according to (a) and (b). IM represents a intermediate module.
(d) Results of constructing regulatory modules. M represents a regulatory module.

Step2. Clustering results of gene expression profiles.
Hierarchical clustering technique is employed for separating genes as many as possible, 100 clusters in this case, with the intention for the genes to be fully distinguished so as to represent functional diversity. (Fig.1. (b))

Step3. Intermediate modules constructed according to Step 2 and Step 3.
Each intermediate module consists of several regulators which are regarded as to govern the regulation of genes in the same cluster according to the result of DNA binding data. At the same time, a cluster number is assigned to each intermediate module corresponding to the genes which the module involves according to the result of expression profiles (in the red box). And the genes which appear in the same cluster of expression profiles are added even though their P-values are still higher than 0.001 in terms of the binding data. Every intermediate modules assigned to the same cluster are merged together giving rise to a big cluster. (Fig.1. (c))

Step4. Results of constructing regulatory modules.
Finally, re-clustering of the output of Step 3 is performed restricting the number of final clusters to be the number of intermediate modules which it contains. Consequently, each module is regarded to be co-regulated and co-expressed and hence likely to have a common biological function. (Fig.1. (d))

Result

In the genomic binding data, theoretically the number of possible sets of regulators is the summation of the combination of choosing i from N, which N and i denote the number of all regulators and those of chosen, respectively. However, it finally generates 564 sets which we call intermediate modules. Then, we select 148 modules which have at least 3 genes for each module in order to incorporate with expression profiles. After combining binding data and expression profiles together, 84 intermediate modules are merged so that as a result 28

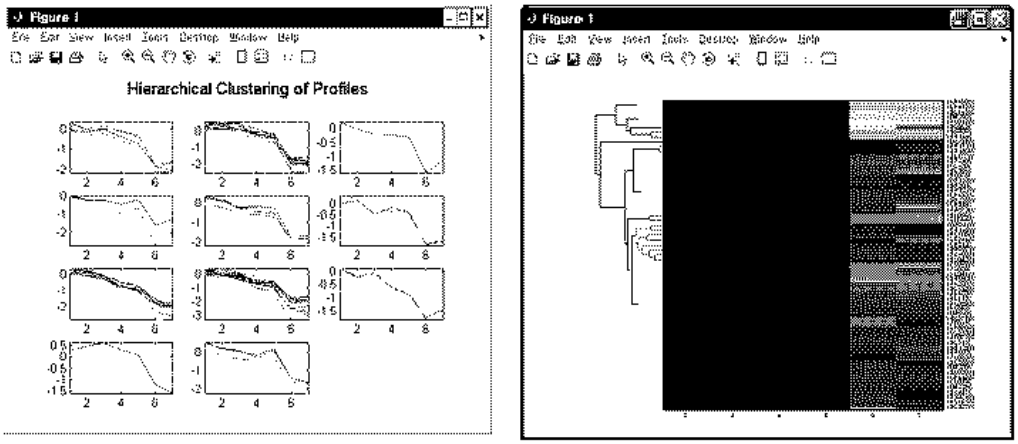

Fig. 2 An example of results after re-clustering by 11 intermediate modules.

Table 1. A ribosomal genes' regulatory module which consists of YDR450W, YLR344W, YDR471W and YNL096C. All genes are involved in ribosomal protein genes which are possibly regulated by the module including FHL1, PDR1 and RAP1 even though some P-values do not appear statistically significant (in bold text).

	FHL1	PDR1	RAP1
YDR450W	0.00000017	0.094	0.00055
YLR344W	0.0000033	0.88	0.00042
YDR471W	0.000000013	0.00044	0.00041
YNL096C	0.000000097	0.027	0.00041

clusters are uniquely determined. However, 64 intermediate modules could not be uniquely determined, because they are assigned by more than two clusters.

Among 28 regulatory modules constructed, one module that caught our attention is involving ribosomal protein genes. Ribosomes are important protein biosynthetic machines. We found that the 4 genes, YDR450W, YLR344W, YDR471W and YNL096C, are regulated by FHL1, PDR1 and RAP1 and at the same time represented almost same expression profile pattern (Table 1). One of the regulators, FHL1 is known to appear almost all ribosomal protein genes, but little else is well understood. According to the information from SGD, FHL1 appears to YDR450W and YNL096W, additionally RAP1 as well. There are no known regulators of YDR471W and YLR344W (Fig. 3). Through this module constructed, we might conclude that the four genes are regulated by the module including FHL1, PDR1 and RAP1. Even though PDR1 does not appear statistically significant, there might be some possibilities for the PDR1 to be involved in the regulation of ribosomal protein genes.

Discussion

It is evident that many genes are controlled by multiple transcription regulators. DNA binding data show that many genes are bound by multiple regulators on their promoters. In addition, the number of transcription factors is far less than the total number of genes in a genome, but most genes can be activated or repressed under multiple conditions. A possible mechanism for a small number of transcription factors to regulate a large number of genes under a variety of responses is through the combinatorial effects of these factors.

In this work, we demonstrated a sort of beginning for the

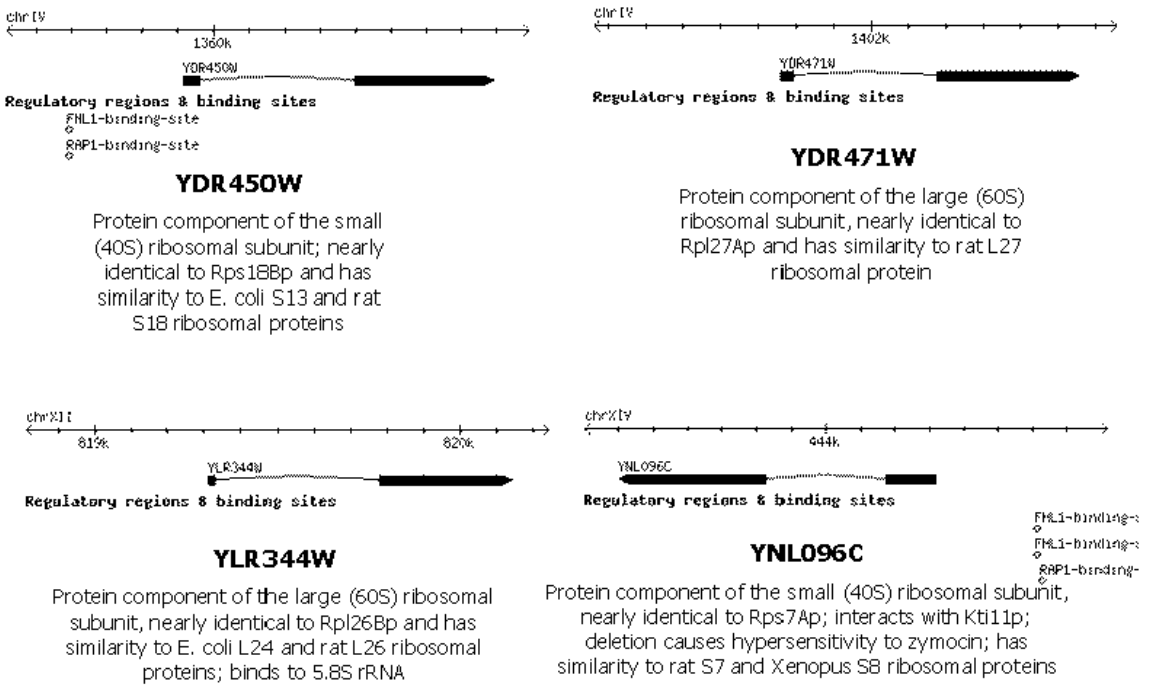

Fig. 3 Information of genes according to SGD : YDR450W, YDR471W, YLR344W and YNL096C, all genes are involved in ribosomal protein genes.

integration of high-throughput data. And there are lots of problems to be addressed. In the step for assigning expression cluster numbers to the intermediate modules, we can expect two kinds of results. In the first case, there are modules which are assigned by one cluster, we regarded this as a right case. On the other hand, there are modules which are assigned by more than two clusters, we considered these are not fully separated because there still remain boundary effects to determine to which clusters the genes should be assigned.

There are many possible directions to extend the current work of genetic regulatory modules. These fall into two categories. The first is to improve the integration strategy and the second is to broaden the types of data which integration method should deal with.

References

[1] Z. Bar-Joseph et al, 2003, "Computational discovery of gene modules and regulatory networks", Nat. Biotechnol., 21(11):1337-42.

[2] Spellman et al., 1998, "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.", Molecular Biology of the Cell, 9, 3273-3297.

[3] Lee, T.I. et al, 2002, "Transcriptional regulatory networks in S. cerevisiae", Science, 298, 799 - 804