The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Discovering Transcriptional Regulation by Integrating Protein-Protein Interaction, Gene Expression and Transcriptional Interaction Data

Fei Luo[1,*] Jinyan Li[1,†]

[1]School of Computer Engineering, Nanyang Technological University, Singapore

Abstract Understanding transcriptional regulation (TR) principles and mechanisms is a hot topic in the genomics and transcriptome research. Many experimental and computational methods have been proposed to investigate this problem. However, perhaps partially because the high complexity in the transcriptional regulation mechanism, or partially because the noise in the high-throughput detection experiments, integrative approaches are demanded to discover underlying TRs which could be missed when only single source of information is considered. In this paper, we explore the biological ideas behind co-regulated protein complexes to study this problem. Co-regulated protein complex has three remarkable characteristics: coding genes of member proteins share the same transcription factor, coding genes usually express coordinately and member proteins intensively interact and form a complex to implement a common biological function. It implies close relationships among co-regulation, co-expression and intensive interaction at the levels of transcriptional regulation, gene expression and protein-protein interaction. Based on these ideas, we integrate protein-protein interaction (PPI), gene expression (GE) and transcriptional interaction (TI) data and use them to form a framework to discover new TR relationships. Experiments on the yeast in three conditions of cell cycle, diauxic shift and DNA damaging were conducted, and 20 novel TRs were predicted and explained.

Keywords Transcription Regulation; Gene Expression; Protein-Protein Interaction

1 Introduction

Prediction of transcription regulations is one of the key tasks in the genomics and transcriptome research. However, some problems still hamper current experimental and computational methods to discover all important TRs for a species with high accuracy. For example, experimental approaches are usually influenced by the noise inherent in experimental and biological systems. Most computational approaches are based on only DNA-binding motifs. They also often suffer from over-prediction problems owing to short length of the motifs. Furthermore, presence of transcription factor (TF) at motifs just only indicates the binding by the TF to the target genes. It does not necessarily mean

*Email Address: FLuo@ntu.edu.sg
†Email Address: JYLi@ntu.edu.sg

a transcriptional function. To activate a TR process, additional environment stimuli and co-regulators are usually required. Therefore, only considering single information such as DNA-binding motifs to predict the TRs is not sufficient. In order to overcome these weaknesses, system biology approaches have drawn increasingly more attention, as they can provide a new way for TRs discovery by combining more comprehensive information to make the inference more reliable.

A case in point is combining gene expression data [1] [2], as it is widely believed that co-regulated genes would have similar expression profiles. Through extracting conserved motifs from co-expressed gene cluster, new TRs may be found by seeking the motifs in the promoter of candidate target genes. Moreover, some recent works have studied the relationship between gene expression and protein-protein interaction. Jansen [3] found that subunits of the same protein complex showed significant co-expression, both in terms of similarities of absolute mRNA levels and expression profiles. Nitin [4] studied the correlation between gene expression profiles and protein-protein interaction on four evolutionarily diverse species: human, mouse, yeast and E Coli. They found that the gene expression profiles of protein-protein interacting pairs were highly correlated in E.Coli and the likelihood of predicting protein interactions from highly correlated expression data was increased by using additional protocol for other three species. Zhang [5] observed an outstanding phenomenon that co-regulated coding genes with similar profiles often lead to intensive interaction between their protein products and forming a protein complex. Tan [6] proposed the innovative concept of co-regulated protein complex where proteins were encoded by genes that are regulated by the same TFs. Coding genes of proteins in the co-regulated complex usually exhibited coherent expression. These results imply that there is a tight linkage between transcription regulation, gene expression and protein-protein interaction. Some newest works [7] [8] have tried to integrate these multidata source for transcriptional network research.

In this paper we propose a framework to mine new TRs by integrating TI, GE and PPI data. Our proposed framework first identifies active TFs at a given condition. Then co-regulated protein groups are found to be used as seeds to do extensive search, during which a scoring function is used to measure the coherent and significant degree of the seed. Finally, we identify new TRs by comparing the seeds which have high score with their extensive search candidates.

The remaining of paper is organized as following: part two is the description of the method, which includes prediction model and working flow. Part three is the experiments on the yeast in three conditions: cell cycle, diauxic shift and DNA damaging. Finally the discussion and conclusion are made in the part four.

2 Method

Transcription factors do not all the time activate to participate a TR function. Only when transcription factors are in the state of activeness, they are able to initiate the corresponding biological reaction. Identifying the active TFs under certain conditions is challenging. Recent research work [9] adopted the assumption that the regulators are themselves transcriptionally regulated. Therefore, their expression profiles can provide informative clues to indicate their activity level. In this work, TFs are identified as being 'active' at some certain condition if they reach sufficiently high expression levels. Our

work is different from others, as we take into account the condition-oriented transcription regulation.

2.1 Model and working flow

The left-hand panel of Figure 1 is an illustration of our model. Let TF represent the active transcription factor in a condition. From the existing TI data, gene symbols a, b, c, d are known to be target genes of TF, and A, B, C, D denote their proteins. From the PPI network and gene expression profiles, if we observe the proteins A, B, C, D and an additional protein E intensively interact one another and their coding genes a, b, c, d and e express coordinately, we can predict the TF also regulate e in this condition. The reason is because E is so similar with the co-regulated protein group of A, B, C, D at the levels of gene expression and protein-protein interaction that it could infer that e also has the TI association with TF like a, b, c and d, although the known TR dataset does not indicate it.

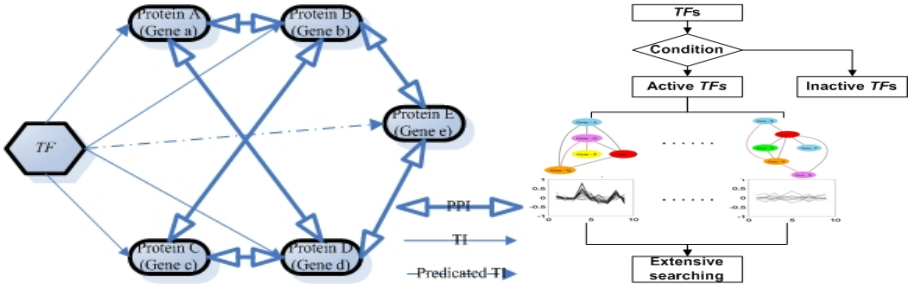

Figure 1: Model and working flow

The right-hand panel of Figure 1 demonstrates the working flow of our method. Given a condition, we first identify the active TFs and find out all target genes for each active TF. The proteins, whose target genes are regulated by the same TF, and their protein-protein interaction form a sub-network in the global PPI network. Then this sub-network is divided into several maximal connected sub-graphs (MCSG). Because MCSGs disjoint each other, one TF may correspond to multiple MCSGs. Since not all MCSGs have to co-express, we calculate significant and coherent score for each MCSG. The higher the score is for a MCSG, the more likely the MCSG is to be a co-regulated complex. Finally, The MCSGs with high significant and coherent scores and their nodes exceeding a threshold are used as seeds to search additional TRs.

2.2 Coherence and significance measurement

We evaluate and compare the coherence and significance for all seeds. Let a seed denoted by $L = (V,E)$. For any v_i, $v_j \in V$, if e=(v_i,v_j) $\in E$, we calculate a score of the coherence and significance between v_i,v_j by

$$Score(v_i,v_j) = Corr(v_i,v_j)std(v_i)std(v_j) \tag{1}$$

where $Corr(v_i,v_j)$ is the *Pearson* correlation coefficient between the coding genes of protein i and protein j to reflect their coherence, and $std()$ is the standard deviation which measures the coding gene's activity.

The coherence and significance score of L, denoted by $T(L)$ is the sum over the scores for all edges in L:

$$T(L) = \sum_{e \in E} Score(e) \tag{2}$$

We note that the coherence and significance score can be influenced by the number of edges in L. In order to compare the coherence and significance between seeds with different number of edges, for L with K edges, we randomly choose 10 000 graphs with K edges from PPI network and compute their score with formula (2), then calculate the average and standard deviation value of these 10 000 graphs and use formula (3) to standardize final coherence and significance score for seed L with K edges.After standardization, seeds with different number of edges can be compared with their coherence and significance.

$$Score(L) = \frac{T(L) - avg_k}{std_k} \tag{3}$$

2.3 Extensive search

As mentioned in Section 2.1, we discover TRs by seeking extra proteins which intensively interact and co-express with the given seed with high coherence and significance score. A protein which expresses differently with the seed will make the score decrease, while a protein which expresses consistently with most parts of the seed will increase the score. Therefore, the search process can be converted to optimize the score by adjusting the structure of the graph starting from the seed.The pseudo codes are shown in Table 1.

Table 1: Pseudo codes of our search method

input: $L_{initial}$, T_{start}, N
output: L_{rs}
step1: $L_{rs} = L_{intial}$, calculate $Score(L_{rs})$
step2: for $i = 1$ to N
step 2.1 : calculate $T_i = T_{start} \times \left(\frac{T_{end}}{T_{start}}\right)^{\frac{i}{N}}$
step 2.2 : $L_{try} = L_{rs}$
step 2.3 : randomly choose an edge $e = (v_i, v_j)$ from PPI
and at least one of v_i, v_j should belong to $V(L_{try})$
step 2.4 : if($e \in E(L_{try})$ and $e \notin E(L_{initial})$)
if(L_{try} is still a connected graph after deleting e)
delete e from L_{try}
else
add e into L_{try}
step 2.5 : calculate $Score(L_{try})$
step 2.6 : δ = $Score(L_{rs}) - Score(L_{try})$
step 2.7 : if($\Delta \succ 0$)
$L_{rs} = L_{try}$
else
$L_{rs} = L_{try}$ with the probability $p = \mathbf{e}^{\frac{\Delta}{T_i}}$
step3 :end

Because the topological structures and sizes of different complexes may vary greatly, extensive searching implements a simulated annealing procedure for every seed. Although it is a kind of heuristic searching method, it could get global optimization solution. By adding or deleting an edge operation from current solution, an opimal result with high coherence and significance score can be found. After conducting extensive search for all input seeds, we rank the seeds based on their scores. Top ones will be picked out to seek potential new TRs by comparing final result and initial seed as described in Section 2.1.

3 Experiment

Our experiments were conducted on a yeast dataset which involves three biological conditions: Cell Cycle [10], Diauxic Shift [11], DNA Damaging [12]. The Cell Cycle wet-lab experiment included expression measurements of 6 178 genes measured at 77 time points. The DNA Damaging experiment had 6 129 genes' expression values with 52 sampling points. The Diauxic Shift dataset consisted of 6 068 gene expression profiles with 7 time points. All data are shown in the Table 2.

Table 2: Data source description

Type	**Source**	**Description**
PPI	DIPs (2007.8)	4 928 proteins, 17 491 PPIs
GE	Cell Cycle, Diauxic Shift, DNA Damaging	6 178, 6 129, 6 068 genes
TR	Luscombe's [13]	142 TFs , 7 074 TRs

We used the result in Luscombe's work [13] to identify the conditional active TFs. They determined 88, 76 and 75 active TFs in Cell Cycle, Diauxic Shift and DNA Damaging conditions respectively. In the data pre-processing, we substituted zero to all missing value in GE. Because there were three different molecular types of data in our work, we unified data symbols by mapping all symbols into gene ID as standard reference. If the corresponding coding genes of the proteins in the PPI cannot be found in the GE dataset, we excluded those proteins from the PPI data. During the extensive search, the parameters were set as follows: $T_{start} = 1$, $T_{end} = 0$, $N = 300$. For each experiment, we used the MCSGs which had at least 3 nodes as the seeds. These seeds were ranked according to their coherent and significant score. Because there's no absolute threshold for the coherent and significant score to judge which of seeds could be co-regulated complex to infer new TRs, we only took the top ones in the ranking list into account to guarantee the prediction accurate. In addition, we were limited to the predicted target genes whose protein should interact with at least two members in the seed. Table 3 shows the results of new TRs predicted in the three conditions.

The seed corresponding to the TF Hsf1 had the highest score (1.883782) in the cell cycle condition. We use it as an example to illustrate how new TRs are discovered based on Hsf1. The left-hand panel in Figure 2 is the topology of the seed and the final result of the extensive search. 'Hexagon' is an added node in final result and 'Circle' is the node in the seed. The right-hand panel in Figure 2 shows the coding genes' expression profiles of proteins in seed and added proteins in final result respectively. We can note that the two sets of expression profiles exhibit a highly coherent and significant similarity, which also validates the scoring function. In order to improve the prediction accuracy, we are

Table 3: Predicted TRs in three conditions

Condition	TF	Predicted Target Genes
Cell Cycle	Hsf1	YMR186W YKL117W YOR027W
	Mbp1	YJL173C YOL090W
Diauxic Shift	Baf1	YMR049C YHR052W YHR066W YPL043W YNL110C YNL175C YER126C YJR063W YPR190C YOR224C YNR003C YBR154C
	Msn4	YMR261C YMR139W YDR247W YML100W
DNA Damaging	Hap4	YDL004W YPL078C

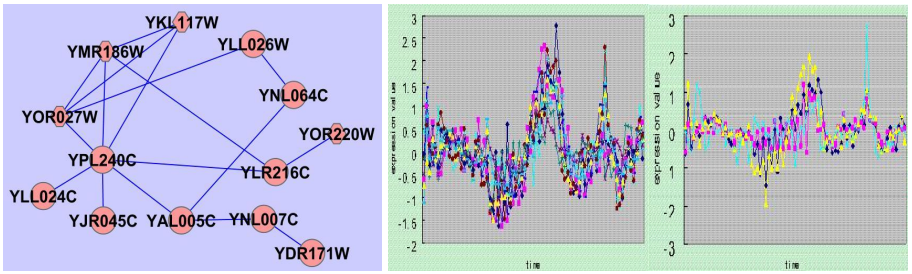

Figure 2: The topology of seed and final search result of Hsf1. Gene expression profiles of coding genes in the seed and in the result after extensive search

limited to the newly predicted target genes whose proteins should interact with at least two proteins in the seed. Then, we can infer that Hsf1 also transcriptionally regulate the target genes YMR186W, YKL117W, and YOR027W but not YOR220W.

We validated our prediction results from three aspects: (1) we retrieved and compared with literature works which predicted the same TRs as well; (2) we detected the conserved binding motifs from the target genes in the seed and examined whether there were matches in the promoter of the predicted target; (3) we examined whether the function of predicted target genes was consistent with those of target genes in the seed. Of course, the final validation for the prediction result should depend on the biology experiment in the cell cycle condition. We found that the results by [15] [16] and [17] [18] supported our newly discovered TRs: Hsf1 regulated YMR186W and Hsf1 regulated YOR027W. However, we have not found direct evidence to support that Hsf1 regulates YKL117W. Maybe, we could find evidence from binding motif to support this. There're two significant binding motifs induced by the tool MEME from the upstream 600bp of coding genes in the seed, which are shown in Figure 3. The first motif is consistent with a known Consensus Motif (GAAXXTTCXXGAA) for Hsf1. We found that there's a match to the first motif in the 600bp upstream of YKL117W, YOR027W, and there's a match to the second motif in the upstream of YMR186W. Finally, we compared the function of Hsf1, coding genes in the seed and the predicted target genes. SGD has an annotation for Hsf1 as follows: 'Hsf1 regulates the transcription of hundreds of targets, including genes involved in **protein folding**, detoxification, energy generation, carbohydrate metabolism, and cell wall organization. Deletion of Hsf1 is lethal and mutants are defective in several pro-

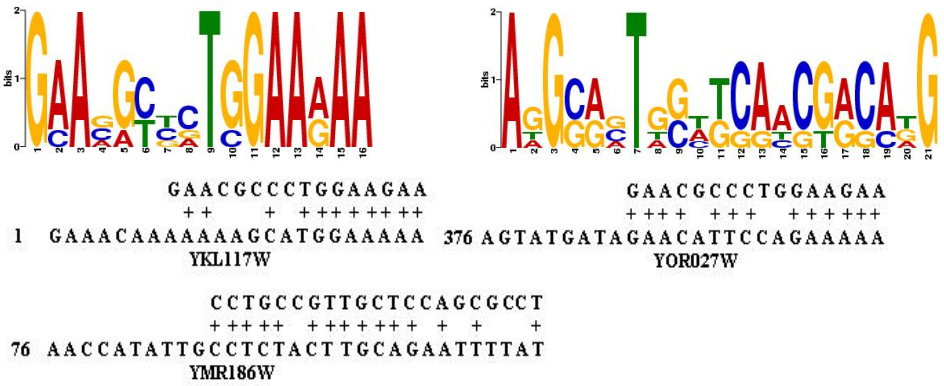

Figure 3: Two motifs predicted by MEME [14] from the upstream 600p of five coding genes in the seed. The first motif is consistent with a known Consensus Motif (GAAXXTTCXXGAA) for Hsf1 in TRANSFAC

cesses including maintenance of cell wall integrity, spindle pole body duplication, protein transport, and cell cycle progression'. Meanwhile, we conducted a function enrichment analysis for the eight genes YLR216C, YLL026W, YPL240C, YAL005C, YNL064C, **YMR186W**, **YKL117W**, **YOR027W**. Our findings is that these genes have a common function of '**protein folding**', which accords with that of Hsf1.

4 Conclusion

In this paper, we proposed a framework to discover new TRs by integrating TR, GE and PPI data. Although it cannot detect all TRs for a species one time, it provided an approach to exploit TRs from the complex mechanism. To make this method widely applicable, two real-life difficulties should be taken with caution. These include: (1) Time-course GE datasets with time points exceeding 10 for species except for yeast are not too many. In fact, most of them are knock-out experiments, which usually re-sample no more than 3 times. It's hard to measure the genes' correlation with such few time points. (2) Difference in the topology of PPI networks for different species produces different probabilities to predict the same TRs from the co-regulated complexes. For example, E. Coli PPI network trends to be a tree structure, where the phenomenon that one protein as a central node interact many other proteins (one-to-many) is more notable than that proteins in a small group interact mutually (many-to-many). When the data become abundant and available, we believe our proposed method would discover more TRs for more species.

Acknowledgement

This research work was funded by a Singapore MOE ARC Tier-2 grant (T208B2203).

References

[1] I. B. Jeffery, S. F. Madden, *et al*, Integrating transcription factor binding site information with gene expression datasets, *Bioinformatics*, 23, 2007, pp.298-305.

[2] H Li, M Zhan, Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data, *Bioinformatics*, 24, 2008, pp.1874-1880.

[3] R. Jansen, *et al*, Relating whole-genome expression data with protein-protein interactions, *Genome Research*, 12, 2002, pp.37-46.

[4] B. Nitin, H. Lu, Correlation between gene expression profiles and protein-protein interactions within and across genomes, *Bioinformatics*, 21(11), 2005, pp.2730-2738.

[5] X. Zhang, *et al*, Motifs themes and thematics maps of an integrated sacharomyces cerevisiae interaction network, *Journal of Biology*, 2005, 4(6).

[6] K. Tan, *et al*, Transcriptional regulation of protein complexes with and across species, *PNAS*, 104, 2007, pp.1283-1288.

[7] T. T. Vu, J vohradsky, Inference of active transcriptional networks by integration of gene expression kinetics modeling and multisource data, *Genomics*, 93(5), 2009, pp.426-433.

[8] M.Hecker, *et al*, Gene regulatory network inference: data integration in dynamic models-a review, *Biosystems*, 96(1), 2009, pp.86-103.

[9] D. Pe'er, A. Regev, A. Tanay, Minreg: inferring an active regulator set, *Bioinformatics*, 18 Suppl 1, 2002, pp.258-267.

[10] R. J. Cho, A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle, *Molecular Cell*, 1998, 2, pp.65-73.

[11] L. Joseph, Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale, *Science*, 278, 1997, pp.680-686.

[12] A. P. Gasch, Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p, *Molecular Biology of the Cell*, 12, 2001, pp.2987-3003.

[13] N. M. Luscombe, *et al*, Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature*, 431, 2004, pp.308-312.

[14] T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 1994, pp. 28-36.

[15] C. T. Harbison, Transcriptional regulatory code of a eukaryotic genome, *Nature*, 413, 2004, pp. 99-104.

[16] E. Boy-Marcotte, The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons, *Mol Microbiol*, 3(2), 1999, pp.274-283.

[17] C. T. Workman, *et al*, A systems approach to mapping DNA damage response pathways, *Science*, 312(5776), 2006, pp.1054-1059.

[18] D. L. Eastmond, H.C. Nelson, Genome-wide analysis reveals new roles for the activation domains of the Saccharomyces cerevisiae heat shock transcription factor (Hsf1) during the transient heat shock response, *J Biol Chem*, 281(43), 2006, pp.32909-32921.