

# Inference of Gene Regulatory Network based on Module Network Model with Gene Functional Classifications

Kohei Taki<sup>1</sup> Reiji Teramoto<sup>2</sup> Yoichi Takenaka<sup>1</sup> Hideo Matsuda<sup>1</sup>

<sup>1</sup> Graduate School of Information Science and Technology, Osaka University

<sup>2</sup> Research Division, Sumitomo Pharmaceuticals Co., Ltd.

<sup>1</sup> {k-taki, takenaka, matsuda}@ist.osaka-u.ac.jp, <sup>2</sup> teramoto@sumitomopharm.co.jp

## Abstract

We propose a novel method for an exhaustive inference of gene regulatory networks from genome-wide expression data and biological knowledge. Our method performs the inferences based on module network model. In the model a module is a set of genes with similar features, and a network represents regulatory relationships among the modules. Our method makes modules using gene functional classification together with expression data. We apply our method to inferences of the networks of yeast cell-cycle. Modules inferred by our method show consistency with experimentally-determined results on yeast cell-cycle, especially on G1 phase. Robust modules built by our method permit us to infer informative regulatory relationships.

## 1 Module Network model with biological knowledge

A module network[1] is composed of a set of modules and connections with edges among the modules. In Fig. 1 an inference process of a module network is composed of following two steps; 1) inference of modules by grouping together genes with similar features and 2) inference of regulator genes of every module. These steps update modules and connections with edges so that a score of correspondent module network is maximized. They are repeated alternatively until convergence of the score.

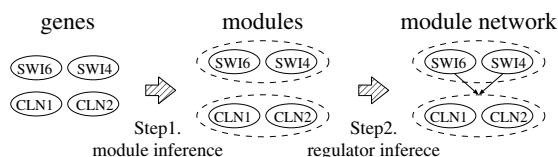


Figure 1. Inference of Module Network.

The conventional module inferences are significantly affected with measurement errors in expression data, because they are according to only expression patterns in the data. Inferred modules can include artificial interactions. In order to alleviate influence of the errors, we propose a novel method considering biological knowledge. In our method a module inference is performed using gene functional classification as the knowledge in addition to expression data. A module is made of genes classified into similar functional categories in the classification. Because genes with similar functions show similar expression, our module inference can be robust against the errors.

## 2 Module Inference using Gene Functional Classification

We use gene functional classification classified into functional categories of *Gene Ontology*. Gene Ontology provides a set of the categories defined as GO terms and a hierarchical structure that defines ‘is-a’ or ‘part-of’ relationships between the terms. We consider a similarity between two genes according to their classified categories by introducing *semantic similarity*[2]. Semantic similarity is evaluated based on the notion of information content of a category. Information content  $I(c)$  of a category  $c$  becomes higher, when fewer genes are classified into the category or its sub-categories. A definition of semantic similarity  $sim(c1, c2)$  between categories  $c1, c2$  is provided as below.

$$sim(c1, c2) = - \max_{c \in S(c1, c2)} I(c)$$

$S(c1, c2)$  is a set of super categories shared by  $c1$  and  $c2$ .

A module network is scored by log of posterior probability  $p(B_M, M|D)$ , given expression data  $D$ , where  $B_M$  and  $M$  are connections with edges and a set of modules, respectively. The probability is decomposed by Bayes’ rule.

$$score(B_M, M : D) = \log p(M) + \log p(B_M|M) + \log p(D|B_M, M)$$

A normalization constant is ignored. A definition of  $p(D|B_M, M)$  is given by Segal[1]. We assume  $p(B_M|M)$  distributes uniformly.  $p(M)$  is prior probability that a set of modules  $M$  is inferred.

We define a formula of  $p(M)$  according to the classification. The classification contributes  $p(M)$ , because it is prior knowledge about a module inference. We assume that if each module in a module inference consists of more genes classified into similar categories, the inference is more likely. We define  $p(M)$  as proportional to a sum of semantic similarities of pairs of genes in the module.

$$\begin{aligned} \log p(M|c) &= c \cdot \sum_{\mathbf{m} \in M} L(\mathbf{m}) - \log Z(c) \\ L(\mathbf{m}) &= \frac{1}{|C(\mathbf{m})|} \sum_{g1 \in \mathbf{m}} \sum_{g2 \in \mathbf{m}} l(g1, g2) \\ l(g1, g2) &= \sum_{c1 \in C(g1)} \sum_{c2 \in C(g2)} sim(c1, c2) \end{aligned}$$

where  $c$  and  $Z(c)$  are a parameter and a normalization function of the  $p(M)$ , and  $C(g)$  and  $C(\mathbf{m})$  denote sets of categories of a gene  $g$  and genes in a module  $\mathbf{m}$ , respectively.

### 3 Result of Computational Experiment

We apply our method to an inference of gene regulatory network of 800 budding yeast genes related to the cell cycle process. The network is inferred from expression data published by Spellman [3]. Our method uses the classification accumulated in Saccharomyces Genome Database.

Table 1 shows a performance comparison between our method and the conventional method using only expression data. The comparison is demonstrated with selectivity and sensitivity for inferred modules. These accuracies are evaluated for 95 genes known to be regulated in five phases named  $G_1$ ,  $S$ ,  $S/G_2$ ,  $G_2/M$  and  $M/G_1$  in the cell cycle.

**Table 1. Performance comparison between our method and the conventional method.**

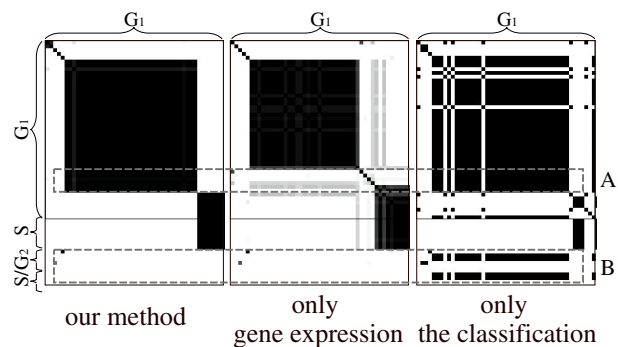
method	selectivity(%)	sensitivity(%)
only expression data	77	39
<b>our method</b>	<b>82</b>	<b>52</b>

### 4 Discussion and Conclusion

Table 1 demonstrates that sensitivity of our module inference is improved. In order to confirm that the improvement is achieved by using the classification, we perform a module inference using only the classification. Selectivity and sensitivity of the inferred modules are 52 and 49, respectively.

The sensitivity is higher than one from only expression data. Both accuracies of our method are higher than both of the method using either data. This result implies that in our method both data can complement defects of each other.

In order to confirm this assumption we show  $G_1$  phase parts of three symmetric matrices in Fig. 2. Rows and columns of the matrix correspond to a sequence of 95 known genes in the order of occurrence of regulation in the cell cycle. A black element represents that its row-gene and column-gene are grouped into the same module. Elements in broken line frames A and B contribute higher sensitivity and lower selectivity, respectively. Since modules from only the classification have more elements in A than ones from only expression data, former modules show the higher sensitivity. In the same manner, the smaller number of elements of later modules in B contributes the higher selectivity than former's one. Figure 2 shows that modules from both data have more elements in A and fewer elements in B. In this result our module inference is improved especially about  $G_1$  phase genes PCL1, RAD54, SPC110, UNG1 and so on. They are grouped into  $S/G_1$  phase module in the results using either data. This result supports the assumption stated above, and we insist that it is our advantage of using the classification together with expression data.



**Figure 2. Inferred modules by each method.**

### References

- [1] E.Segal, M.Shapira et al., "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data", *Nat. Genet.*, vol.34, no.2, Jun.2003, pp.166–176.
- [2] P.W.Lord, R.D.Stevens, and C.A.Goble, "Investigating semantic similarity measures across the Gene Ontology : the relationship between sequence and annotation", *Bioinformatics*, vol.19, no.10, Jul.2003, pp.1275–1283.
- [3] P.T.Spellman, G.Sherlock et al., "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization", *Mol. Biol. Cell*, vol.9, no.12, Dec.1998, pp.3273–3297.