연관 지식을 이용한 유전자 발현 데이터 분석

: 퍼지 클러스링과 조절 네트웍 모델링에의 응용

In-silico inferences for expression data using IGAM

: Applied to Fuzzy-Clustering & Regulatory Network Modeling

Philhyoun Lee, Hojeong Nam, Doheon Lee*, Kwang H. Lee
Department of BioSystems, KAIST, 373-1, Daejeon, 305-701, Korea

ABSTRACT

Genome-scale expression data provides us with valuable insights about organisms, but the biological validation of *in-silico* analysis is difficult and often controversial. Here we present a new approach for integrating previously established knowledge with computational analysis. Based on the known biological evidences, IGAM (Integrated Gene Association Matrix) automatically estimates the relatedness between a pair of genes. We combined this association knowledge to the regulatory network modeling and fuzzy clustering in yeast *S.Cerevisiae*. The result was found to be more effective for extracting biological meanings from *in-silico* inferences for gene expression data.

Keywords: genetic regulatory network, fuzzy-clustering, gene association score

I. INTRODUCTION

Inferring principles of regulation from genome-scale expression data is a huge challenge. Furthermore, due to the absence of standard validation method, researchers have found it difficult to demonstrate the validity of their inferences. Most researchers have used cross-validation methods or simulation techniques to justify their approaches [E. Segal et al., 2003., V. Anne Smith et al. 2002]. Verification methods devised for one specific experiment are ill-suited for other problems, and even when applied to the same problem, different methods sometimes lead to contradicting results. Besides, exhaustive search of literatures and annotation databases to corroborate their own results is time consuming and tedious. In this paper, we present a universal and consistent approach to measure the association degree of two genes. The proposed metric is devised under the assumption that related genes share their intrinsic features more often than non-related ones. We calculate the genetic

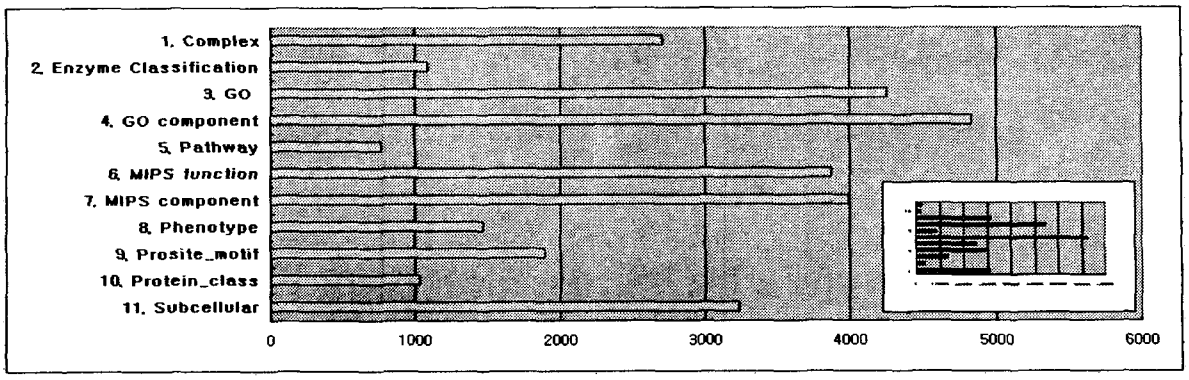

Fig. 1. the number of distinct *S.Cerevisiae* genes with annotation in each biological domain.

(inset: the number of distinct gene pairs of *S.Cerevisiae* related with each other)

interaction or regulation by examining the tendency of coexistence of their established known properties. To overcome the limited coverage of a specific knowledge domain (Fig.1), we coherently use multiple biological evidences. Measuring the association score based on the statistical framework enables us to use one metric for heterogeneous data.

II. METHODS

Data

We downloaded the *SGD GO* annotations from http://www.yeastgenome.org. *Enzyme Classification and Pathway* information was downloaded from http:// www.genome.ad.jp and *Motif* data from http://www.expasy.ch/prosite. *MIPS functional category, complex, phenotype, subcellular* annotations are available at http://mips.gsf.de/proj/yeast/catalogues/phenotype. *S.Cerevisiae* cell-cyle expression data was downloaded from http://cellcycle-www.stanford.edu/ for fuzzy C-means clustering.

Integrated Gene Association Matrix

Let $G_p = \{g_1, g_2, \ldots, g_M\}$ and $F_p = \{f_1, f_2, \ldots, f_N\}$ be a set of genes and a set of features in the biological knowledge domain p respectively. Let $N(f_k)$ be the number of genes with a feature f_k. Let $F_{pi} = \{f_{i1}, \ldots, f_{iS}\}$ be a set of features that a gene g_i may have. The association score of every gene pair is defined as follows:

$$Score(g_i, g_j) = S_{positive}(g_i, g_j) + S_{negative}(g_i, g_j)$$

$$S_{positive}(g_i, g_j) = -\left(\sum_{f_m \in F_{pi} \cap F_{pj}} \log s(g_i, g_j, f_m) \right)$$

$$S_{negative}(g_i, g_j) = \log\left(1 + \sum_{f_m \in F_{pi} - F_{pj}} \frac{{}_{t-2}C_{N(f_m)-1}}{{}_{t}C_{N(f_m)}} \right)$$

$$\text{where } s(g_i, g_j, f_m) = \frac{{}_{t-2}C_{N(f_m)-2}}{{}_{t}C_{N(f_m)}}$$

The association score satisfies the following two criteria:

i. If $N(f_k) > N(f_l)$,
$(f_k \in F_{pi} \cap F_{pj}, f_l \in F_{pk} \cap F_{pl},)$
then $s(g_i, g_j, f_k) > s(g_k, g_l, f_l)$.

ii. If $\| F_{pi} \cap F_{pj} \| = \| F_{qk} \cap F_{ql} \|$,
$(f_k \in F_{pi} \cap F_{pj}, \; f_l \in F_{qk} \cap F_{ql})$ and

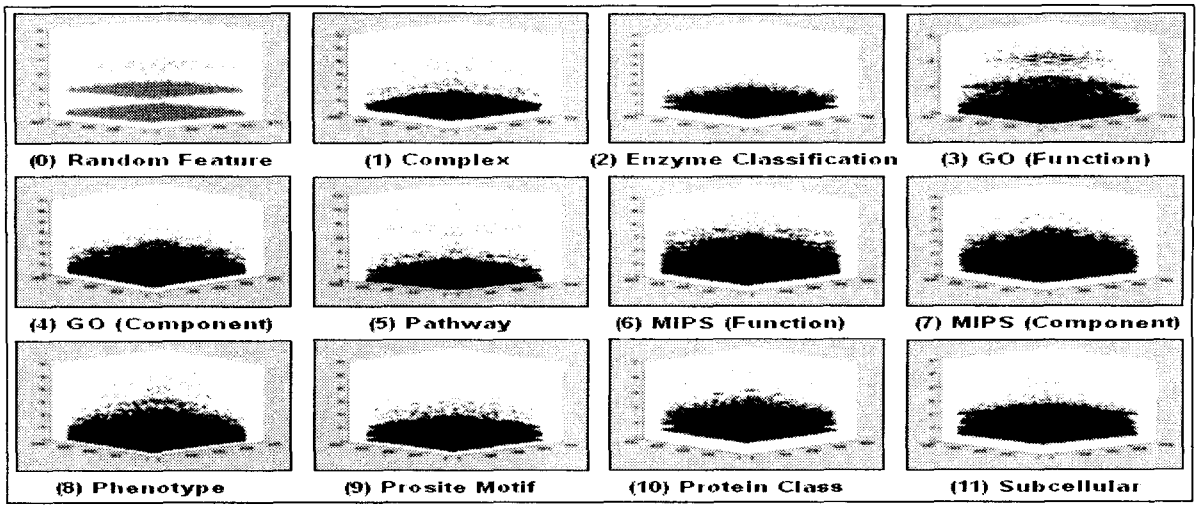

Fig. 2: Eleven IGAM score components based on each biological domain. x, y, z axes represent the first gene and the second of related gene pairs, and their association score respectively.

$N(f_k) = N(f_l)$ and
$\| G_p \| > \| G_q \|$,
then $Score(g_i, g_j) > Score(g_k, g_l)$.

The test of homogeneity with statistic χ^2 ((r-1)(c-1)) (r=20, c=20, α=0.01) was performed to confirm that each biological evidence categorizes the population independently. To test the validity of our association assessment, we created random feature data sets. Distributions obtained from random sets have shapes and ranges of score distribution distinct from those of genuine sets, as illustrated in Fig. 2.

III. RESULTS

We applied IGAM for the analysis of genetic regulation network modeling [Satoru Miyano et al., 2002]. Total 82 edges were examined for their degree of association. Directly linked genes are marked with different shades of grey indicating 4 different ranges of association scores (Fig. 3). 68 edges were revealed to be supported at least one knowledge domain. Gene pairs connected only in one knowledge domain include EST1 and ZDS2, PCL1 and PDS5, and SKN1 and BUB1. This shows that IGAM can recover subtle and hard-to-find biological associations. Gene pairs indirectly connected by mediators were also recovered as well as direct relationships. The indirectly-linked are represented by blue and subjected to further studies for hidden elements between them. Remaining 14 edges are putative candidates of literature search and wet-experiments to determine whether they represent novel interactions. In fig. 3, we can easily recognize that predictions like HSL1 and SWE1 are highly reliable even without referring to the detailed annotation; HSL1 is a negative regulator of SWE1 kinase.

Fuzzy C-means clustering with IGAM is performed with fuzzy parameters m=1.17, p = 0.45 and Pearson's correlation as a distance metric [Doulaye Dembele et al. 2003]. The coverage and performance of functional significance analysis was much

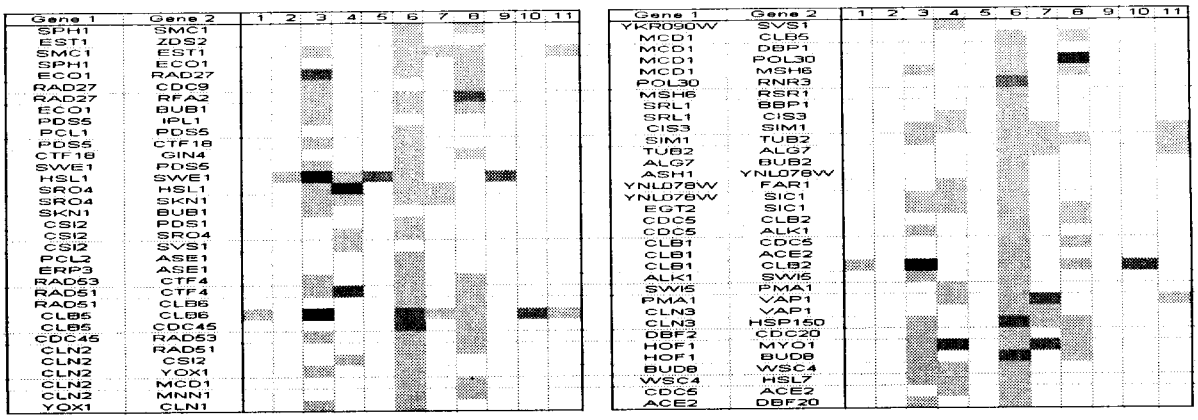

Fig 3. IGAM (Integrated Gene Association Metric) analysis result for genetic regulatory network

higher than that of evaluators based on a single evidence. Supplementary data is available at http://bioif.kaist.ac.kr/IGAM/.

IV. CONCLUSIONS

We aim to devise a generic and efficient methodology for integrating previous biological knowledge with gene expression data analysis. The novel feature of the proposed method is that the true positive inference can be easily differentiated from candidates for further investigation based on heterogeneous evidences. In addition, the sensitivity and specificity of inference can be intuitively visualized with a color map. The evaluation of regulatory network modeling and clustering results clearly demonstrates these advantages.

An extension to the evaluation of context-specific association beyond a pair-wise one is under consideration. Other domains of biological knowledge such as protein-interaction data, phylogenetic profiles, and sequence similarity will be integrated for IGAM.

V. ACKNOWLEDGEMENT

The authors are grateful to CHUNG Moon Soul Center for BioInformation and BioElectronics for supporting this work. This work was supported by the Korean Systems Biology Research Grant (M10309020000-03B5002-00000) from the KMST.

VI. REFERENCES

[1] Doulaye Dembele et al. (2003), *Bioinformatics*, Vol. 19, pp. 973-980.

[2] Satoru Miyano et al. (2002), *2002 SRCCS Statistical Workshop*.

[3] E. Segal et al. (2003), *Bioinformatics*, Vol. 19., Suppl., pp. 264-272.

[4] V. Anne Smith et al. (2002), *Bioinformatics*, Vol. 18, Suppl., pp. 216-224.

[5] Paul T. Spellman et al. (1998), *Molecular Biology of the Cell*, Vol. 9, pp. 3273-3297.