

# Automatic Extraction of Expression-Related Features Shared by a Given Group of Genes

**Takuya Oyama**<sup>1,5</sup>

oyama@isl.intec.co.jp

**Mikio Yoshida**<sup>1,5</sup>

yoshida@gic.intec.co.jp

**Satoshi Kamegai**<sup>1,5</sup>

kamegai@isl.intec.co.jp

**Kagehiko Kitano**<sup>1,5</sup>

kage@isl.intec.co.jp

**Fumihito Miura**<sup>2</sup>

fumihito@genome.c.kanazawa-u.ac.jp

**Noriko Kawaguchi**<sup>3,5</sup>

kawanori@genome.c.kanazawa-u.ac.jp

**Miyuki Onda**<sup>3</sup>

miyuki-o@ims.u-tokyo.ac.jp

**Kenji Satou**<sup>4,5</sup>

ken@jaist.ac.jp

**Takashi Ito**<sup>2,3,5</sup>

ito@k.u-tokyo.ac.jp

<sup>1</sup> INTEC Web and Genome Informatics Corp., 1-3-3, Shinsuna, Koto-ku, Tokyo 136-0075, Japan

<sup>2</sup> Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-8562, Japan

<sup>3</sup> Cancer Research Institute, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-0934, Japan

<sup>4</sup> School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan

<sup>5</sup> Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Corporation (JST), Science Plaza, 5-3, Yonban-cho, Chiyoda-ku, Tokyo 102-8666, Japan

**Keywords:** gene expression, specific feature, information extraction

## 1 Introduction

Numerous studies have been performed to examine genome-wide expression using various methods such as DNA microarray and SAGE. Consequently, a huge number of gene expression profiles under various conditions or situations have been accumulated. Many of these profiles obtained are open to the public and shared on Internet, and people can browse them through the web browsers. However, it remains a hard task for researchers to extract information of their own interests from such a vast amount of profiles. Several biological databases are constructed to provide summarized information for each gene or protein describing how it is regulated under a particular condition in natural language. Nevertheless it takes a lot of time to look over such gene expression-related summaries on dozens of genes. Thus, accumulated gene expression profiles are not always utilized effectively.

In this article, we introduce a simple approach to extract specific information or features concerned with gene expression common to dozens of given target genes that are selected according to researcher's interest or his/her own experiments.

## 2 Method

Figure 1 illustrates the outline of our approach. First of all, published profiles and source papers of them are collected and summarized to a gene-to-feature table. The summarized table provides the information of the relation between genes and features. Namely, it provides features assigned to each gene or, conversely, genes assigned to each feature. Needless to say, a "feature" is assigned to a gene if the gene has the characteristics represented by the feature. Here are some examples of the "feature", where  $X$  means a gene that has the feature.

- The  $X$  is induced by the treatment with rapamycin.
- The  $X$  is negatively regulated by RTG3.

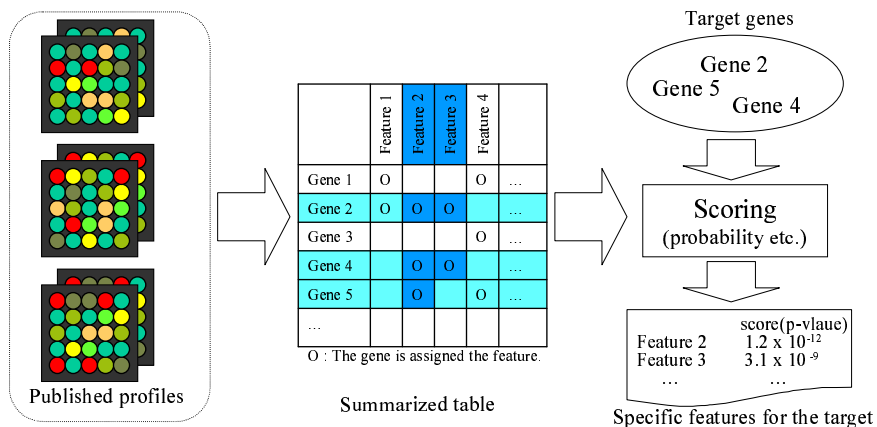


Figure 1: Outline to extract specific features for the target genes.

Next, when the target genes are given, a score for each feature is computed using the summarized table. Then, the highly ranked features are listed with the scores. The score is calculated so that it may represent the specificity of the feature for the target genes as follows:

(1) P-value: the score based on the significance

The score, p-value, is calculated as probability of  $m$  or more out of  $n$  genes being assigned to a particular feature under an assumption that the target genes are randomly selected, where  $m$  is an actual number of target genes assigned to the same feature and  $n$  is the total number of target genes. The p-value is often used in the field of statistics and represents the significance of the event [2]. If the p-value is extremely small, the feature can be regarded as specific for the target genes.

(2) Weighted score by gene expression levels

If the expression levels are submitted together with the target genes, those levels are used as weights for score calculation. The higher the level is, the greater the corresponding gene influences the score for features assigned to the gene.

### 3 Results

It takes much effort to collect profiles and to construct a summarized table from scratch. Fortunately, some databases provide profile information for individual genes. Thus, instead of construction from scratch, we extracted descriptions concerned with gene expression from Yeast Proteome Database (YPD) [1, 3], one of the popular databases containing profile information, and used them to create a summarized table.

We selected 11 genes remarkably induced by rapamycin treatment in our experiment as target genes for the analysis. Application of our method to the 11 genes indicated their relation to nitrogen metabolism, which is consistent with well-established knowledge, thereby demonstrating the effectiveness of our method.

### References

- [1] Costanzo, M.C., *et al.*, YPD, PombePD, and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information, *Nucleic Acids Res.*, 29(1):75–79, 2001.
- [2] Ewens, W.J. and Grant, G.R., *Statistical Methods in Bioinformatics - An Introduction*, Springer, 2001.
- [3] Hodges, P.E., *et al.*, Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data, *Nucleic Acids Res.*, 27:69–73, 1999.