

READ: Practical Analysis System of Mouse Transcriptome with the FANTOM Database

Hidemasa Bono¹

bono@gsc.riken.go.jp

Takeya Kasukawa^{1,2}

kasukawa@gsc.riken.go.jp

Nikaido Itoshi^{1,3}

itoshi@gsc.riken.go.jp

Masaaki Furuno¹

furuno@gsc.riken.go.jp

Yoshihide Hayashizaki^{1,3}

yoshihide@gsc.riken.go.jp

Yasushi Okazaki¹

okazaki@gsc.riken.go.jp

¹ Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center(GSC), RIKEN Yokohama Institute 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

² Multimedia Development Center, Advanced Technology Development, NTT Software Corporation, 223-1 Yamashita-cho, Naka-ku, Yokohama, Kanagawa 231-8554, Japan

³ Division of Genomic Information Resource Exploration, Science of Biological Supramolecular Systems, Yokohama City University, Graduate School of Integrated Science 1-7-29 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

Keywords: DNA microarray, transcriptome, database, cDNA, functional annotation, XML

1 Introduction

In addition to algorithmic development of analyzing gene expression data from microarray experiments, practical efforts are going on to get biological inference from the similarity of gene expression patterns in conjunction with the sequence analyses in the genomic scale. Using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays, gene expression patterns in 49 adult and embryonic mouse tissues have been systematically characterized [2]. A web-based database named READ (RIKEN Expression Array Database) has been developed to allow total access to the gene expression profiles and to easily search for genes of interest [1]. Because these cDNA clones include quite new clones to the public, the functional annotation of cDNA clones is needed before precise, clone-by-clone analyses. To add functional annotation to 21,076 RIKEN mouse cDNA clones, we held the FANTOM (Functional Annotation of Mouse) meeting in August 2000 [3]. The experts from genomics and bioinformatics extensively and substantially worked for these functional annotations. This computational resource, which organizes the functional annotation of cDNA clones and the geneexpression profiles for mouse tissues, can help us discriminating related genes from whole transcriptome, and accelerate the process of scientific discovery.

2 Method and Results

The READ system organizes all the information that needs to be referred in the analysis phase of microarray experiment. Basically these information are stored in the flat file format, though frequently referred and queried data (log-transformed ratio of gene expression intensity) and manually annotated data are stored in RDBMS (PostgreSQL) for practical reasons. Currently it can be queried not only by the RIKEN cDNA clone identifier on the chip and by keywords in the database description, but also by the log-transformed ratio of gene expression intensity in all 49 tissues.

Before the analysis of gene expression patterns, all cDNA clones on the chip are needed to be functionally annotated. Although cDNA clones from the FANTOM clone set have the functional

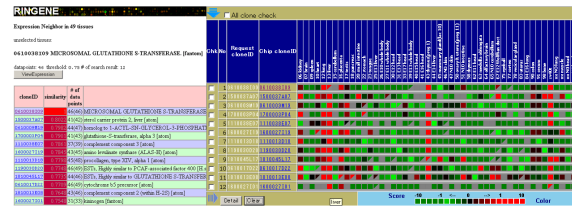


Figure 1: The result of RINGENE search in the READ system. The expression profiling of the selected set of genes appears by clicking the button of ‘ViewExpression’ in the RINGENE view.

annotation in the FANTOM database, remaining clones are required to be annotated. Thus we have developed the system named FIND (Functional Inference Descriptor) to simulate the functional annotation process that was discussed and established in a flowchart in the FANTOM meeting [3]. Rough annotations in several brief definitions can be retrieved from pre-computed results of sequence analyses with those rules in the functional annotation flowchart. These functional annotations are described in XML, called MaXML (Mouse Annotation XML), to share them in the research community for bulk genomic analysis in silico.

A new web-based tool starting from one gene, named RINGENE (READ Integrates Gene Expression Neighbor), enables us to look for genes that have similar gene expression patterns in the specified tissues 1. Values of Pearson’s correlation coefficient for correlations between the data of a specified clone identifier and those of all the other clones on the chip whose data are valid are dynamically calculated to identify pairs with high correlation. Genes that have inverse correlation can also be queried because the correlation coefficient is used to measure gene expression similarity. These inversely correlated genes could be regulated to run the program of living cells in the opposite direction. A similarity search for gene expression patterns in specified tissues enables us to predict the function of genes, which is difficult to do only from a sequence similarity search.

3 Acknowledgements

This study has been supported by Research Grant for the RIKEN Genome Exploration Research Project from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government and ACT-JST (Research and Development for Applying Advanced Computational Science and Technology) of Japan Science and Technology Corporation (JST).

References

- [1] Bono, H., Kasukawa, T., Hayashizaki, Y., and Okazaki, Y., READ: Riken Expression Array Database, *Nucleic Acids Res.*, in press.
- [2] Miki, R., Kadota, K., Bono, H., Mizuno, Y., Tomaru, Y., Carninci, P., Itoh, M., Shibata, K., Kawai, J., Konno, H., Watanabe, S., Sato, K., Tokusumi, Y., Kikuchi, N., Ishii, Y., Hamaguchi, Y., Nishizuka, I., Goto, H., Nitanda, H., Satomi, S., Yoshiki, A., Kusakabe, M., DeRisi, J.L., Eisen, M.B., Iyer, V.R., Brown, P.O., Muramatsu, M., Shimada, H., Okazaki, Y., and Hayashizaki, Y., Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays, *Proc. Natl. Acad. Sci. USA*, 98:2210–2204, 2001.
- [3] The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, Functional annotation of a full-length mouse cDNA collection, *Nature*, 490:685–690, 2001.