

Co-Expression Analysis Tool for Large Gene Expression Datasets

Hajime Harada **Natalia Polouliakh**
harada-hajime@aist.go.jp nata.polouliakh@aist.go.jp
Wataru Fujibuchi **Paul Horton**
fujibuchi-wataru@aist.go.jp horton-p@aist.go.jp

Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 153-0061, Japan

Keywords: correlation coefficient, microarray, gene expression

1 Introduction

The study of the pattern of gene expression across many experiments that survey a wide array of cellular responses were conducted. Identifying the patterns of gene expression and grouping co-expressed genes may help provide insight into biological function. Statistical analysis of gene expression levels is useful for studying gene regulation [1].

To this date, tens of thousands of microarray experiments have been conducted, and large datasets are available for public use, through repositories such as the GEO database of NCBI [2]. However this data is huge and there is a lack of suitable tools for easy search and analysis of large microarray databases. The CellMontage website adds value to publicly available microarray data (currently taken from GEO) by making it searchable, providing the results of cluster analysis and organizing the experiments based on the cell type involved. In this poster we present a program which calculates the Pearson's correlation coefficient for gene pairs from microarray data

2 Method and Results

Our program calculates Pearson's correlation coefficients of gene pairs from microarray data. The input files analyzed have two parts, the first consists of descriptions about profile such as: dataset, platform ID, types, and the other descriptions, and the second is the data lines containing (unigene ID, expression value) pairs, which are extracted from GEO data table and unigene ID number. One advantage of using the CM file format is that this format, based on the (multi)fasta file format, is portable for various applications.

For large genomes, the number of possible gene pairs is so large that the correlation matrix may not fit in memory. Our implementation solves this problem with a memory efficient multipass algorithm and can be applied to genome-scale co-expression analysis. Our program employs a command line interface which allows the user to select a file containing a list of genes or simply calculate the correlation for all genes appearing in the input data. Biases in the database may affect the results of calculation. Our program does not directly address that problem but may be used with synthetic profiles instead of actual profiles. In particular the so called "eigen cells" taken from the CellMontage website. These 112 synthetic profiles were generated by applying some normalization and eigen decomposition to the large GEO dataset and are intended to provide a representative sample of human cell expression states. As an initial application we used actual profiles and also eigen cell data to calculate

the expression correlation for MAP kinase G-protein coupling pathway-related and ribosomal genes. A small difference of correlation coefficients distribution between these two groups exists. As might be expected, the ribosomal genes, which work closely together, had higher correlation coefficients than the MAPK genes. Our results also suggest that using the “eigen cell” dataset can reduce biases in the results.

3 Discussion

We have constructed a tool which is convenient for analyzing large expression profile datasets. A test case analysis of ribosomal genes yields results which are consistent with biological expectations.

References

- [1] Quackenbush, J., Computational analysis of microarray data, *Nat Rev Genet.*, 2(6):418–427, 2001.
- [2] Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W. C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar. R., NCBI GEO: mining millions of expression profiles—database and tools, *Nucleic Acids Res.*, 33:D562–566, 2005.
- [3] CellMontage <http://cellmontage.cbrc.jp/>