A Web Tool for Expression Pattern-Based Data Retrieval and Relevant Network Discovery from vast Public Microarray Database

Chunlai Feng
feng@pharm.kyoto-u.ac.jp

Michihiro Araki
arakipharm.kyoto-u.ac.jp

Ryo Kunimoto
kunimoto@zankan.mbox.media.kyoto-u.ac.jp

Akiko Tamon
tamon@pharm.kyoto-u.ac.jp

Hiroki Makiguchi
makiguchi-hiroki@mki.co.jp

Satoshi Niijima
niijima@pharm.kyoto-u.ac.jp

Gozoh Tsujimoto
gtsuji@pharm.kyoto-u.ac.jp

Yasushi Okuno
okuno@pharm.kyoto-u.ac.jp

1 Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan
2 Kyoto University Education Unit for Global Leaders, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan
3 Bio Science Group, IT Solution Division 1, Industry Solution Business Unit, Mitsui Knowledge Industry, Osaka 531-0072, Japan
4 Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Keywords: gene expression pattern, data mining, gene co-expression network, GEO

1 Introduction

DNA microarray technology provides us with a first step toward the goal of uncovering gene functions on a genomic scale [1]. In recent years, vast amounts of gene expression data have been collected, much of which are available in public databases, such as the Gene Expression Omnibus (GEO) [2]. To date, most researchers have been manually retrieving data from databases through web browsers using accession numbers (IDs) or keywords, but gene-expression patterns are not considered when retrieving such data. The Connectivity Map was recently introduced to compare gene expression data by introducing gene-expression signatures (represented by a set of genes with up- or down-regulated labels according to their biological states) and is available as a web tool for detecting similar gene-expression signatures among their original microarray dataset, which was observed under unified experimental conditions [3]. The other tools that allow to compare users’ data to public microarray data from different experimental conditions (such as L2L and LOLA), usually use the co-occurrence of genes (the number of overlap genes) between query gene list and pre-defined lists of differentially expressed genes as the similarity metrics. Thus the existing mining tools allow users to search gene expression data from public databases, but these are also restricted by gene annotation, pre-selected gene lists, or experimental conditions. In order to support researchers to utilize the public gene expression data more effectively, we developed a web tool named GEM-TREND to retrieve similar gene expression data from a publicly available database by comparing gene-expression patterns between a query and the database entries, and generate co-expression networks for retrieved gene expression data and may provide insights into unknown functional relationships of the genes.

2 Method and Results

GEM-TREND provides both gene-expression pattern-based (accepting either gene-expression signatures or gene expression ratio data as inputs) and text-based (accepting keywords, platform IDs, or series IDs as inputs) searches to retrieve gene expression data from GEO. For the former searches, gene expression data is retrieved by comparing gene-expression pattern between the query and GEO gene expression data. The comparison methods are based on the nonparametric, rank-based pattern matching approach of Lamb et al. (Science 2006) with the additional calculation of statistical significance. For text-based searches meanwhile,
an N-gram based search engine is used, and GEO series title, series summary, platform IDs, and series IDs are considered as search criteria. Retrieved gene expression data can then be viewed as a co-expression network with gene ontology (GO) annotation [4], whereby genes and annotations are dynamically linked to external data repositories. The network is constructed based on Pearson correlation coefficients and K-means clustering, and can provide insights into unknown functional relationships of the genes (Fig.1). The web tool was tested using gene expression ratio data randomly extracted from the GEO and with in-house microarray data, respectively. We evaluated the similarity of biological annotations between the query and the retrieved microarray data, and the results validated the ability of GEM-TREND to retrieve gene expression entries biologically related to a query from GEO.

3 Discussions

GEM-TREND was developed to retrieve gene expression data by comparing query gene-expression pattern with those of GEO gene expression data and to provide network visualization. GEM-TREND could be a very useful resource for comprehensive re-analysis of the primary data by merging data from multiple studies and provide insights into unknown functional relationships of the genes through finding similar gene expression profiles and constructing its gene co-expression networks from a publicly available database. It can also be considered as an extension of the Connectivity Map and a supplementary tool of GEO. GEM-TREND was designed to be user-friendly and is expected to support knowledge discovery. GEM-TREND is freely available at http://cgs.pharm.kyoto-u.ac.jp/services/network

References


