Publime: a new tool for meta-analysis of cancer-related microarray experiments

Finocchiaro G (1,2), Mancuso F (1,2), Muller H (1,2)

(1) European Institute of Oncology, Milan, Italy
(2) IFOM, Firc Institute of Molecular Oncology, Milan, Italy

Motivation

Application of gene expression microarray technology is rapidly producing large amounts of data that represent a considerable resource to produce a general view of the transcriptional activity of several organisms, both in different phases of development and in different conditions. The analysis of a specific microarray experiment profits enormously from cross-comparing to other experiments; statistical techniques of meta-analysis have been produced to integrate and to compare raw data generated from several microarray experiments. Nevertheless the way of dataset selection for metaanalysis represents a serious limitation for the identification of new and unexpected connections between different datasets. To facilitate the solution of these problems, we developed Publime (acronym for PUBlished LIsts of Microarray Experiments), a dedicated database, where researchers can find and consult published gene lists derived from gene expression microarray analysis.

Methods

Lists of genes are manually extracted from publications concerning microarray studies by experienced curators. Each gene is annotated following a procedure similar to those we adopted to generate IFOM Affymetrix annotation tables [1]. The annotation process is necessary both to synchronize information with latest database releases and to standardize heterogeneous identifiers found in publications. Whenever available, we associated each identifier with UniGene ID, NCBI Entrez Gene ID, Official Symbol, UniGene Title, Chromosome, and GeneOntologies. Medline annotation of each publication inserted is downloaded in XML format through NCBI Entrez Utilities http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html via the perl LWP package. In particular, we extract the following information for each publication: Authors, Title, Abstract, Journal, Medical Subject Headings and Chemical that represent the controlled vocabulary of biomedical terms and chemical terms, respectively, used for indexing documents in MEDLINE. Data regarding publications and published gene lists were relationally linked in a MySQL database.

Results

Currently, 273 publications are inserted in Publime, containing 1282 gene lists. The analysis of MeSH terms allows classifying articles according to the classes of pathologies studied in the publication. Most represented categories of neoplasms are Breast, Prostatic, Ovarian and Colonic. A web interface allows biologists to retrieve information according to particular keywords. Information concerning a particular gene, a particular class of genes, or a particular class of studies can be easily queried and interpreted. Moreover a user can test for the existence of a significant overlap between a set of genes of interest and lists stored in Publime. The significance of the overlap can be estimated according to the hypergeometric distribution or Fisher's exact test. This approach helped us to identify common transcriptional targets of pRB, p16 and EWS/FLI (Ewing sarcoma breakpoint region 1) pathways having a different role in cell cycle control [2]. We annotated a total of 31243 human and 7476 mouse identifiers, representing 8187 human and 3991 unique genes, respectively. Interestingly, we observed that some genes are reported as differentially regulated much more frequently than others. Gene Ontology analysis revealed that the most frequently reported genes are predominantly involved in the regulation of cell cycle, regulation of cellular proliferation and in cell growth and mantainance.

Contact email: giacomo.finocchiaro@ifom-ieo-campus.it

References

1. Guffanti A, Finocchiaro G, Reid JF, Luzi L, Alcalay M, Confalonieri S, Lassandro L, Muller H: Automated DNA chip annotation tables at IFOM: the importance of synchronisation and crossreferencing of sequence databases. Appl Bioinformatics 2003, 2(4):245-249.

2. Finocchiaro G, Mancuso F, Muller H: Mining published lists of cancer related microarray experiments: Identification of a gene expression signature having a critical role in cell-cycle control. BMC Bioinformatics 2005, 6(Suppl 4):S14.