Genome dynamic and statistical functional annotations for biological knowledge mining from microarray data

Marco Masseroli [(1)], Dario Martucci [(2)], Francesco Pinciroli [(1,3)]

[(1)] Dipartimento di Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy
masseroli@biomed.polimi.it
[(2)] Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy
[(3)] Istituto di Ingegneria Biomedica, Consiglio Nazionale delle Ricerche, via Golgi 32, 20133 Milano, Italy

Keywords. Genomic functional annotation, Statistical analysis, Biomolecular databases, Microarray data interpretation, Biological knowledge discovery

Introduction

Statistical and clustering analyses of gene expression results from high-throughput microarray experiments produce lists of hundreds of genes candidate regulated, or with particular expression profile patterns, in the conditions under study. Independently of the microarray platforms and analysis methods used to identify and cluster differentially expressed genes, the common task any researcher faces is to translate the identified lists of genes into a better understanding of the patho-physiological phenomena involved. To this aim, many biological annotations are available within numerous heterogeneous and widely distributed databases. Although several tools have been developed for annotating lists of genes, most of them do not provide methods to evaluate the relevance of the retrieved annotations for the considered set of genes, or to estimate the functional bias introduced by the gene set present on the specific array used to identify the considered gene list. Lately, few tools have been proposed that use gene annotations provided through the Gene Ontology (GO) [1] controlled vocabularies to enrich lists of genes with biological information. Some of them (e.g. Affymetrix Data Mining Tool, DAVID, FatiGO, GoMiner, MAPPFinder) also present the GO categories more relevant for a given set of genes according to the number of genes of the considered set belonging to a given category, or in relation to their statistical evaluation performed using some basic tests. To extend these functionalities we created *GFINDer* (i.e. Genome Function INtegrated Discoverer, *http://www.medinfopoli.polimi.it/GFINDer/*), a web server able to automatically provide large-scale lists of user-classified genes with the statistically significant functional profiles that biologically characterize the different gene classes in a considered gene list.

System Architecture

The *GFINDer* web server system is implemented in a three-layer architecture based on a multi-database structure. In the first layer, the *data layer*, a MySQL DBMS server manages all different types of genomic annotations stored in three relational databases, including one containing GO terms and their semantic relationships, and another storing many different annotations and associations between genes and GO categories. These databases are kept updated by automatic procedures that automatically retrieve gene annotations and GO information from several on-line public databases [2] as soon as new releases become available.
In the second layer, the *processing layer*, a web server manages requests coming from client computers and runs all system processing and analyses. It is constituted of Active Server Page scripts and uses Microsoft ActiveX Data Object technology and Standard Query Language to communicate with the DBMS server on the data layer.
The third layer, the *user layer*, is composed of any client computer connected to the web server on the processing layer through an Internet/intranet communication network and loading in its web browser the *GFINDer* graphic user interface, implemented as web pages using Hyper Text Markup Language.
The illustrated three-layer architectural choice enhances at maximum the *GFINDer* system performances because it enables subdividing the required computational power between the two web and DBMS servers.

Statistical Analysis

When genes are selected from a predefined set or subdivided in classes, the statistical significance of their specific annotation categories, provided through controlled vocabularies in each considered classes of genes, can be evaluated. *GFINDer* considers the quantities, frequencies, and distributions of genes among the different categories, and calculate a probability of occurrence for each considered annotation category. To this aim, the binomial distribution test and the χ^2 test for equality of proportions are used [3]. When the sample size is small, the χ^2 test cannot be reliably applied and hence the Fisher's exact is automatically used instead.

Web Interface

The *GFINDer* user interface is organized in modules allowing to study the distribution of different classes of genes among GO categories, KEGG biochemical pathways, PFAM protein families, or OMIM diseases.
The *Annotation module* produces a tabular output of the uploaded gene list enriched with several annotations, including gene name, symbol, and description; identifiers in different resources; cytogenetic localization; GO categories with their evidences; EC Number; biochemical pathways, protein family names, and citations in scientific literature. Each annotation is linked to the original online source to display more information.
The *Exploration modules* perform functional categorizations of the input genes according to GO categories, KEGG biochemical pathways, PFAM protein domains, and OMIM genetic diseases and disorders the genes belong to. The results illustrate the distribution of genes among these functional characteristics. In particular, the *Gene Ontology module* exploits the GO semantic network to easily and graphically show either how many and which GO categories are related to the considered genes, or how many of those genes refer to each GO category, providing also graphical views to understand the semantic relations among the represented categories.
The *Categorization module* enables to define groups of input genes according to their membership to specific annotation categories and in relation to user-selected terms. The annotations related to user keywords are shown and the input genes with these annotations are grouped in categories that can be statistically analysed.
The *Statistical modules* allow performing statistical analyses on the GO, KEGG, PFAM, and OMIM categorizations of the input genes when these are subdivided in different classes, or a reference gene list is also provided (e.g. the list of all the genes in the microarray used to identify the input genes). This enables highlighting which biological processes, molecular functions, cellular components, biochemical pathways, protein families, and genetic diseases the input genes, or each of their classes, can be related to, and with which probability. Thus, a plain list of gene identifiers is enriched with biological meaning and statistical significances.

Conclusions

Through functional annotations and statistical evaluations of the significant membership of gene classes to different functional categories, *GFINDer* allows evaluating the functional bias of lists of candidate regulated genes identified through microarray experiments, highlighting their significant biological characteristics and discovering the relevant involvement of a set of genes in specific biological processes and functions. Moreover, it allows detecting patterns of differential expression in classes of genes with particular functional characteristics, hence facilitating a genomic approach to the understanding of the fundamental biological processes and complex cellular patho-physiological mechanisms. Thus, based on some of the publicly available genomic resources, *GFINDer* can represent an important aid in knowledge discovery from microarray experiment results.

References

[1] M. Ashburner, C. A. Ball, L. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25:25-29, 2000.

[2] G. Casella and R. L. Berger (eds.), Statistical inference, 2nd edition, Duxbury Press, Belmont, CA, 2002.

[3] R. T. Walker, D. Söll and A. S. Jones (eds.), Database issue. Nucleic Acids Res., 32, 2004.