Gene expression annotations and functional statistical analysis of human gene lists

ID - 168

Masseroli Marco[1,2], Ceresa Mario[2]

[1]Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

[2]Laboratorio di Informatica BioMedica, Dipartimento di Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Gene expression information constitutes one of the most useful resources to support better understanding of gene functions, and represents an important background for computational genome-wise analyses. To allow performing comprehensive evaluations of gene annotations sparsely available in numerous heterogeneous databanks accessible via Internet, we previously developed and constantly improved GFINDer (http://www.bioinformatics.polimi.it/GFINDer/). It is a Web system that dynamically aggregates functional and phenotypic annotations of user-uploaded lists of gene identifiers, and allows executing their statistical analysis and mining. In order to take advantage of the expression information provided by eVOC ontologies (http://www.evocontology.org/), we decided to develop and add new original modules in GFINDer. They were specifically designed and implemented to: 1) annotating numerous identifiers of user-classified human nucleotide sequences with controlled information on their anatomical system, cell type, developmental stage, and pathology expression features; 2) exploring and classifying the identifiers according to such annotation categories; and 3) statistically analyzing the obtained classifications.

Methods

GFINDer Web system is implemented in a three-tier architecture based on a multi-database structure. In the first tier, the data tier, a MySQL DBMS manages all considered genomic annotations stored in different relational databases. In one of them, we structured controlled and manually curated hierarchical gene expression annotations by eVOC ontologies. In order to efficiently exploit the eVOC hierarchies, we directly associated their terms with the annotated Gene IDs, reconstructed their hierarchical relationships, and structured them in GFINDer database tables. Then, within the GFINDer processing tier, in Javascript and Active Server Page scripts, we implemented categorical analyses of the gene expression annotations. Created analysis procedures employ hypergeometric and binomial distribution tests and the Fishers exact test to assess statistical significance of the over- and under-representation of categorical expression annotations in a group of user-classified genes. To interact with the MySQL DBMS server on the data tier, we used Microsoft ActiveX Data Object technology and Standard Query Language, whereas we employed Hyper Text Markup Language and Javascript to implement a Web graphic interface for the user tier, which is composed of any client computer connected to the GFINDer Web server on the processing tier through an Internet/intranet communication network.

Results

In GFINDer multi-database structure we imported and keep updated four of the ten ontologies available in the data release 2.7 of the eVOC human gene expression annotations. The considered four ontologies describe anatomical system, cell type, developmental stage, and pathology characteristics of human gene expression, and the hierarchical relations among such characteristics. They include 394, 161, 154, and 176 distinct terms, respectively, which represent an equal number of concepts. eVOC provides automatic annotations to each of its ten ontologies for 8,041 human clone libraries in the dbEST databank, and their dbEST IDs associated with the Entrez Gene IDs of the 23,307 genes expressed in such libraries. However, automatic annotations include also not significant terms such as 'unclassifiable', 'pending', and 'not applicable'. Only the annotations to the four ontologies we considered are also manually curated. In the used eVOC data release, out of the 8,041 clone libraries annotated, those with manually curated and significant term annotations to the considered Anatomical System, Cell Type, Developmental Stage, and Pathology ontology are 7,846 (93%), 661 (8%), 6,836 (81%), and 7,093 (84%), respectively. The new GFINDer modules developed for the utilization of eVOC structured data provide a straightforward way for the exploration and statistical analysis of gene expression annotations. The Exploration Expression module allows to easily and graphically understand how many and which location, state, and timing expression features are associated with each user-selected gene, or how many of the selected genes refer to each expression feature. When uploaded nucleotide sequence identifiers are subdivided in classes (e.g. from clustering analysis of microarray experiment results), the Statistics Expression module allows estimating relevance of eVOC controlled annotations for the uploaded genes by highlighting expression features significantly more represented within user-defined classes of genes. Thus, new GFINDer modules allow performing genomic expression annotation analyses that well complement previously provided phenotypic and functional evaluations in supporting better interpretation of gene lists (e.g. from gene expression microarray experiments), and aid unveiling new biological knowledge about the considered genes.

Availability: http://www.bioinformatics.polimi.it/GFINDer/

Email: masseroli@elet.polimi.it