Inherited disorder dynamic annotation and statistical analysis for biomedical knowledge mining from high-throughput gene lists

Masseroli M.(1), Galati O.(1), Gibert K.(2), Pinciroli F.(1)

(1) Dipartimento di Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy
(2) Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, C. Pau Gargallo 5, 08028 Barcelona, Spain

Motivation

Analysis of inherited diseases and their associated phenotypes is of great importance to gain insight into the underlying genetic interactions and could ultimately give clinically useful insights into disease processes, including complex diseases influenced by multiple genetic loci. Nevertheless, to date few computational contributions have been proposed for this purpose mainly due to lack of controlled clinical information easily accessible and structured for computational genome-wise analyses. To enable performing comprehensive evaluations of gene annotations sparsely available in numerous different databanks accessible via Internet, we previously developed GFINDer, a Web server that dynamically aggregates functional annotations of user uploaded gene lists and allows performing their statistical analysis and mining (http://www.bioinformatics.polimi.it/GFINDer/). Exploiting and structuring information present in textual form in the Online Mendelian Inheritance in Man (OMIM) databank, we developed and made available within GFINDer new original modules specifically devoted to the analysis of inherited disorder related genes. They allow annotating large numbers of user classified biomolecular sequence identifiers with morbidity and clinical information, classifying them according to genetic disease and phenotypic location categories, and 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 these, we structured genetic disease and phenotype information from OMIM. In addition to information on genetic loci, inheritance patterns, and allelic variants, many OMIM entries contain a Clinical Synopsis section, which delineates the accompanying signs and symptoms of a disease and their locations. The Clinical Synopsis section is divided into phenotype location categories, either by organ system or type of finding. To associate an inherited disorder with the involved genes or genetic loci, if any, we considered the MIM codes associated with a gene, as provided by the Entrez Gene database. Unfortunately, information in OMIM Clinical Synopsis section is not always represented in a uniform manner. Several typing errors and synonyms for the same name, and different names for overlapping concepts are often present. Thus, in GFINDer processing tier we implemented, in Javascript and Active Server Page scripts, disease and phenotype location categorical analyses based on a list of unique category terms obtained by normalizing the original Clinical Synopsis names according to a set of specifically defined synonyms. Created analysis procedures employ hypergeometric and binomial distribution tests and the Fisher's exact test to assess statistical significance of the over ad under representation of categorical biomedical and clinical 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 Web server on the processing tier through an Internet/intranet communication network.

Results

In OMIM databank we found 3,452 genetic disease entries and 995 entries containing a Clinical Synopsis section, in which we initially found 133 different phenotype location category names. After name normalization, we obtained 93 unique category terms that were structured in three hierarchical levels according to their controlled location descriptions. Each genetic disease description was divided into up to four hierarchical levels in relation to its increasing degree of detail. The GFINDer modules developed for the exploitation of such structured data provide Genetic Disorder Annotation, Exploration, and Statistics analyses. The Annotation module produces a tabular output of user-uploaded genes enriched with related genetic disease names, their inheritance mode and OMIM phenotype code, and with several other annotations automatically retrieved from many different databanks. The Exploration Genetic Disorders module enables to easily and graphically understand either how many and which diseases, phenotype locations and their specific signs and symptoms are correlated to each considered gene, or how many of the selected genes refer to each disease, location, or phenotype. When uploaded genes are subdivided in classes (e.g. from clustering analysis of microarray results), the Statistics Genetic Disorders module enables to estimate relevance of OMIM controlled annotations for the uploaded genes by highlighting genetic diseases and their phenotypic locations significantly more represented within user-defined classes of genes.

Contact email: masseroli@biomed.polimi.it

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