A WEB-BASED TOOL FOR GENOMIC FUNCTIONAL ANNOTATION, STATISTICAL ANALYSIS AND DATA MINING

D. Martucci\(^a\), F. Pinciroli\(^a,b\), M. Masseroli\(^a\)

\(^a\) Dipartimento di Bioingegneria, Politecnico di Milano, Milano, Italy
\(^b\) Istituto di Ingegneria Biomedica, Consiglio Nazionale delle Ricerche, Milano, Italy

masseroli@biomed.polimi.it

Abstract: Microarray technology is generating a massive amount of data that need to be studied with statistical approaches, clustered according to gene expression characteristics, and annotated with biological relevant information. Many web-based tools and databases have been created to annotate genes with biological information, classify genes in function-related families, and identify functional links between genes showing common regulation patterns. However, most of these tools are useful for studying one gene at a time but they are generally not designed for batch analyses of hundreds or thousands of genes at a time, as microarray experiments require. We created a web-based tool to meet the need of biological annotating a great quantity of genes through a simple user interface adequate also for users without advanced informatics knowledge. Our tool enables evaluating the functional significance of microarray experiment’s results through statistical indexes and graphical views in a well-known web browser user interface.

Introduction

The post-genomic era has led to high-throughput methodologies that generate a massive amount of experimental data at exponential rates. While in the past biologists studied single genes at a time, nowadays we have available both the complete sequences of many organisms, and the technologies that allow investigating gene expressions and mutations on a whole genomic scale. One of the most promising is the microarrays technology, which generate a huge amount of data and enable analyzing ten of thousands data points simultaneously. However, the output data points the microarrays generate are simply a vast quantity of unstructured numeric information affected by noise. For making sense of microarrays data, many efforts is been doing to statistically analyze the expression data and to construct web-accessible databases and tools to integrate the expression information of a group of genes with the available biological information \([1,2]\). These data, the gene annotations, are composed of attributes that describe the gene characteristics, its known correlations with different pathologies, and relationships with other genes. Because genes encodes proteins, the understanding of a functional relationship between genes leads consequentially to the knowledge of interactions between proteins involved in a biological process. Thus this knowledge can help biologists to create new pharmacies.

Tools such as Ensembl, LocusLink, Swiss-Prot \([3]\), SOURCE, GeneCards - using functional information including Gene Ontology, KEGG and Pfam - are some public efforts focusing on the curated annotation of gene-specific functional data. These resources provide excellent depth and coverage of the functional data available for a given gene but are not designed to effectively aggregate biological information for thousands of genes in parallel. Such aggregation and the statistical analysis of the annotations of microarrays experiment data could provide effective and useful information and hopefully improve the knowledge of related biological processes. To this aim we developed the web-tool illustrated in this paper.

Materials and Methods

Statistics - The Fisher’s Exact Test

We used some statistical analysis algorithms to investigate and to get descriptive analysis from a gene annotation set. We considered statistical indicators e.g. frequency, percentage and the Fisher’s Exact Test to evaluate the statistical significance of specific gene annotations.

The Test, based on a 2x2 table, returns a p-value that represents the significance level of a dependence hypothesis of 2 categories of variables: X and Y \([4]\). The X variable can represent a category of the Gene Ontology (GO) - and, on user demand, all its sub-classes - and the Y variable the belonging or not to a specific gene regulation class (e.g. 1, -1, 0).

The lower is the p-value returned, the stronger is the hypothesis of dependence between X and Y and the robust is the assumption that the genes of the considered class can be really involved in the considered biological process.

That is, the p-value indicates if the genes that belong to the selected class mainly belong to the considered annotation category, and the genes that do not belong to the selected class mainly do not belong to the considered annotation category. Thus, we can affirm, with the significance level indicated by the one-sided p-value, that the great part of the genes in the list that belong to the selected class are correlated with the
biological characteristic the considered annotation category represents.

To gain knowledge on gene expressions, we functionally annotate the genes and group them using the Gene Ontology (GO) controlled vocabulary classification, then we evaluate the distribution of belonging of these genes between the GO categories.

**Gene Ontology**

The Gene Ontology project is an effort to address the need for consistent descriptions of gene products in different databases [5]. The Gene Ontology Consortium is developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The ontology is composed by terms that are organized in structures called Directed Acyclic Graphs (DAGs) (Fig. 1), which differ from hierarchies in that a 'child' (more specialized term) can have many 'parents' (less specialized terms) and may have a different class of relationship with its different parents.

The link between GenBank Accession Numbers and Gene Ontology terms is established through LocusLink ID. LocusLink data are downloaded by NCBI LocusLink FTP site.

**Results and Discussion**

Our tool allows users to study experiment data through numerical tests and graphical visualization of the distribution of differentially expressed genes among functional categories using the controlled vocabulary of the Gene Ontology Consortium.

The user can load on the server his/her own gene list with regulation expressions for each gene. This list can contain hundreds of gene identifiers, obtained via microarrays experiments, which can be studied in parallel. The classification of genes (e.g. up-regulated = +1; down-regulated = -1) could be the result of a microarrays experiment but also a classification done by the user following any clustering methodology. The user can even insert a list of genes coming from two different experiments and, for example, he/she can cluster genes from the first experiment using 1 and 0 for the second.

Then the tool allows him/her to perform statistical analysis to discover if some genes can be really involved in a biological process. To do this, the tool associates gene identifiers (accession numbers, RefSeq, LocusLink ID, Affymetrix ID or Unigene ID) with annotation information available on the web (and previously retrieved from the right web-database); so a pure list of genes is increased in value by statistical significance and biological meaning.

![Figure 1: The GO Direct Acyclic Graph](image1.png)

![Figure 2: The tool system architecture](image2.png)
System Implementation

The tool core system engine is based on a relational database that keeps information about users, their uploaded gene lists, and gene-annotation associations. Another database keeps information about the Gene Ontology structure. All is implemented with MySQL server DBMS. The system is based on a 3 layer architecture (Fig. 2): a Data Base server manages the DBMS engine and contains the data, a Web server manages the HTTP query coming from the clients and runs the kernel of the application, the ASP scripts, that communicate with the Data Base server through SQL statements, on a fast LAN. The third level is composed of the clients that could be in the same LAN or connected to the Internet, scattered all over in the world. This architectural choice, enhance at maximum system performance because computational power is partitioned on the two servers and users can use this tool from every where an internet connection is available.

Microsoft ActiveX Data Object (ADO) technology was used to access the databases from the web-based user interface which was developed using Microsoft Active Server Pages (ASP) technology, with Javascript language for scripting and Hyper Text Markup Language (HTML) 4.0 for formatting the GUI implemented as web pages.

User Interface

With the web-based choice the user interface is provided inside a common web browser, without needing to install additional software on the user’s client. The user interface is intended to increase at maximum the system usage easiness and friendliness also for people without advanced informatics knowledge.

At every time the user can get the list of genes inserted by him/her that belong to a specified ontology class (with a link to Entrez-Nucleotide DataBank of NCBI), the graphical view of the ontology tree provided by EMBL-EBI Gene Ontology Browser (QuickGO) (Fig. 3) simply by clicking on an icon or a text link. The user can also click on the F icon to perform the Fisher’s Test on a specific ontology category in the list view.

The Client Side

The user can load on the server his/her own gene list with regulation expressions for each gene in pure text tab-delimited file format. The first column of the list must contain the gene identifiers, other columns (many as the user wants) can include gene regulation expressions and other user-defined gene classification expressions.

The Explore GO (Fig. 4) module performs analysis on gene-expressed ontology categories.

By choosing the level of ontology tree coverage (low levels provide high coverage but low term specificity; high levels lead to low coverage but to high term specificity), the tool shows the ontology categories represented by the loaded gene list only till the specified

Figure 3: A graph view for a selected GO category provided by EBI-EMBL Browser

Figure 4: The Explore GO module
level of the ontology tree. For each GO category, the category name and the specific ontology to which it belongs (i.e. biological process, molecular function, or cellular component), the absolute and percentage number of genes in the loaded list that represent the category and the list of these genes and links to external viewers of the ontology DAG structure from the category up to the ontology root (Fig. 3) are provided. A histogram graphical representation of the distribution of the genes in the loaded list among the represented GO category is also given. Through it is possible to navigate the represented categories. By clicking on the histogram bar representing a specific GO category, the analysis is restarted, eventually with a different selected level of ontology tree coverage, but considering only the genes belonging to that category.

The Single Category Fisher’s Exact Test module allows to perform the test on one user-inserted gene list, based on one Gene Ontology category per time. For each experiment the user can perform the same test on different regulation columns of his/her experiment file so it is easy to confront results between different regulation values of the same gene expressions.

The Multi Category Fisher’s Exact Test module allows to the user to perform the same Fisher’s Test, but automatically applied to each Gene Ontology class represented by the genes provided by the user-inserted list. The user has to specify the regulation classes (e.g. +1 vs. –1 or +1 & 0 vs. –1) and he/she gets a table that shows for each line, associated to a Gene Ontology class, the number of genes involved and the p-value of the Fisher’s Test applied to that class, accordingly to the regulation class specified before.

Other modules are provided to perform functional classifications among KEGG, PFAM and OMIM annotation to study protein families codified by genes, their biochemical pathways and diseases in which they are involved.

Testing

The tool has been tested with the well-known leukemia-A dataset from the work of Golub et al. (1999). It contains gene expression measurements corresponding to Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) samples from bone marrow and peripheral blood. The dataset involves 72 leukemia samples (47 ALL and 25 AML) of 7129 genes each. Among these genes, Golub and colleagues identified 50 genes (25 highly expressed in ALL and 25 more highly expressed in AML) most greatly correlated with the ALL-AML class distinction.

Results of the Fisher Exact Test for the cellular component analysis (Fig. 5) show that, out of the 50 considered ALL-AML class distinction genes, only 38 (21 highly expressed in the ALL class and 17 highly expressed in the AML class) have GO cellular component annotations. Out of these 38 genes, those highly expressed in the ALL class and with protein products located in the cellular nucleus are statistically significant more than those with the same nuclear localization of their protein products but highly expressed in the AML class (11 highly expressed genes in ALL vs. 1 in AML, p-value = 0.00096). Whereas, the genes highly expressed in the AML class with protein products located in the extracellular space are nearly statistically significant more than those with the same extracellular space localization of their protein products but highly expressed in the ALL class (4 highly expressed genes in AML vs. 0 in ALL, p-value = 0.05493).

Conclusions

Through the functional annotation and statistical evaluation of gene lists, our tool can stand out the most relevant biological information of a given gene set. Thus, we think our tool, based on some of the available genomic web sources, can represent an important aid in biological knowledge discovery not only from results of microarray experiments but also from biological research in general and it can represent an important aid in biological knowledge mining and discovery.

References

MEDICON and HEALTH TELEMATICS 2004
"Health in the Information Society"
Xth Mediterranean Conference on Medical and Biological Engineering

Under the High Patronage of
the President of the Republic of Italy
With the Patronages of:
Italian Ministry of Education and Scientific Research
Italian Ministry of Health
Regione Campania
Polo delle Scienze e delle Tecnologie - University of Naples Federico II
Recognized by
European Federation of Organizations for Medical Physics – EFOMP
In cooperation with
International Federation for Medical and Biological Engineering (I.F.M.B.E.)
and with
the Member Societies of Croatia, Cyprus, France, Greece, Israel, Serbia and Montenegro, Slovenia, Spain

July 31 - August 5, 2004
Ischia, Naples - Italy
www.medicon2004.unina.it