Exploration and Statistical Analysis of Human Gene Expression Annotations

Mario Ceresa¹, Marco Masseroli Ph.D. ¹,²

¹ Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan – Italy
² BioMedical Informatics Laboratory, Dipartimento di Bioingegneria, Politecnico di Milano, Milan - Italy

Abstract
Gene expression information is a relevant resource useful to better understand gene functions. To evaluate genomic annotations sparsely available in numerous databanks accessible via Internet, we previously developed GFINDer, a Web server that performs statistical analysis of functional and phenotypic annotations of gene lists. To exploit expression information provided by eVOC ontologies, in GFINDer we integrated new modules that allow annotating and statistically analyzing user-classified human nucleotide sequences with controlled information on their expression features.

Introduction
GFINDer (http://www.bioinformatics.polimi.it/GFINDer/) Web server dynamically manages functional and phenotypic genomic annotations and allows performing their statistical analysis and mining¹. To use the valuable gene expression information provided by eVOC ontologies², we implemented the two new Exploration and Statistics Expression modules in GFINDer. These modules annotate numerous user-uploaded classified human nucleotide sequence identifiers with eVOC controlled gene expression annotations, classify the uploaded identifiers according to such annotation categories, and statistically analyze the obtained classifications.

Material and Methods
GFINDer system consists of a multi-database and three-layer architecture. In its first layer, the “data layer”, we stored all genomic annotations in MySQL relational databases, including eVOC gene expression hierarchical controlled annotations, whose release 2.7 is described in Table 1. eVOC provides such annotations 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. 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 a GFINDer database table.

In GFINDer “processing layer”, we used Javascript and Active Server Page scripting technology to implement categorical analysis of the gene expression annotations using statistical hypergeometric and binomial distribution tests and Fisher’s exact test. Finally, we implemented a Web interface as “user layer”, which interacts with the system and displays results of analysis performed.

Table 1. eVOC ontologies, their distinct terms / concepts, and their manually curated annotations.

<table>
<thead>
<tr>
<th>Ontology Name</th>
<th>Terms / Concepts</th>
<th>Annotated Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomical System</td>
<td>394</td>
<td>7,846 (93%)</td>
</tr>
<tr>
<td>Cell Type</td>
<td>161</td>
<td>661 (8%)</td>
</tr>
<tr>
<td>Developmental Stage</td>
<td>154</td>
<td>6,836 (81%)</td>
</tr>
<tr>
<td>Pathology</td>
<td>176</td>
<td>7,093 (84%)</td>
</tr>
</tbody>
</table>

Results and Discussion
Developed GFINDer Exploration Expression module displays relationships between user-selected genes and location, state and timing of their expression, and how many of the selected genes refer to each expression feature. GFINDer Statistics Expression module allows estimating relevance of eVOC controlled annotations for the user-selected genes by highlighting expression features significantly related to user-defined classes of genes. Thus, the new easy to use GFINDer modules complement previously provided phenotypic and functional evaluations in supporting better interpretation of gene lists (e.g. from gene expression microarray experiments), and help unveiling new biological and biomedical knowledge about the considered genes. During the first three months of their public availability, the new modules have been used by several users worldwide.

References