Babelomics:

Introduction to Functional Annotation Databases

Javier Santoyo
jsantoyo@cipf.es
Bioinformatics and Genomics Department, CIPF
Valencia, Spain

Valencia, Mayo 2009
Introduction

- Last years has been an exponential increase in the number of annotation databases and in their content.

- *Nucleic Acids Research* online Molecular Biology Database Collection is a public repository that lists almost all *biological databases*

- Updated every year. The 2009 update includes 1170 databases!!

http://www3.oup.co.uk/nar/database/c/
Some of the biological databases contain Functional Information of the genes and sequences.
The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism.

- Last version has more than 22,000 terms
- The controlled vocabularies of terms are structured

http://www.geneontology.org/
The three categories of GO

Molecular Function
the tasks performed by individual gene products; examples are transcription factor and DNA helicase

Biological Process
broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions

Cellular Component
subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex
Functional Annotation DDBB
Gene Ontology (GO terms)

GO is a DAG (Directed Acyclic Graph)

- **biological process**: 78,842 genes
- **physiological process**: 55,602 genes
- **cellular process**: 29,557 genes
- **cell growth and/or maintenance**: 21,215 genes
- **transport**: 11,722 genes
  - **secretory pathway**: 4,505 genes
  - **vesicle-mediated transport**: 1,525 genes
  - **intracellular transport**: 2,255 genes
  - **Golgi vesicle transport**: 442 genes
  - **ER to Golgi transport**: 190 genes

Levels

More general information

More detailed information

terms are structured
Functional Annotation DDBB
Gene Ontology (GO terms)

- AmiGO provides a web interface to search and browse the ontology and annotation data
  
  http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

- QuickGO (EBI) provides also a web interface
  
  http://www.ebi.ac.uk/ego
### Functional Annotation DDBB

**Gene Ontology (GO terms)**

Example of GO annotation of **BRCA2** in **Ensembl48**

<table>
<thead>
<tr>
<th>GO</th>
<th>The following GO terms have been mapped to this entry via UniProt and/or RefSeq:</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000724</td>
<td>[from Mus_musculus ENSMUSP00000038576] [double-strand break repair via homologous recombination]</td>
</tr>
<tr>
<td>GO:0001556</td>
<td>[from Mus_musculus ENSMUSP00000038576] [oocyte maturation]</td>
</tr>
<tr>
<td>GO:0001833</td>
<td>[from Mus_musculus ENSMUSP00000038576] [inner cell mass cell proliferation]</td>
</tr>
<tr>
<td>GO:0003697</td>
<td>[from Rattus_norvegicus ENSRN0P0000001475] [single-stranded DNA binding]</td>
</tr>
<tr>
<td>GO:0005515</td>
<td>[protein binding]</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>[nucleus]</td>
</tr>
<tr>
<td>GO:0005737</td>
<td>[cytoplasm]</td>
</tr>
<tr>
<td>GO:0005813</td>
<td>[centrosome]</td>
</tr>
<tr>
<td>GO:0006291</td>
<td>[DNA repair]</td>
</tr>
<tr>
<td>GO:0006289</td>
<td>[nucleotide-excision repair]</td>
</tr>
<tr>
<td>GO:0006302</td>
<td>[double-strand break repair]</td>
</tr>
<tr>
<td>GO:0006310</td>
<td>[from Rattus_norvegicus ENSRN0P00000001475] [DNA recombination]</td>
</tr>
<tr>
<td>GO:0006355</td>
<td>[regulation of transcription, DNA-dependent]</td>
</tr>
<tr>
<td>GO:0006976</td>
<td>[from Mus_musculus ENSMUSP00000038576] [DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator]</td>
</tr>
<tr>
<td>GO:0007001</td>
<td>[from Mus_musculus ENSMUSP00000038576] [chromosome organization and biogenesis (sensu Eukaryota)]</td>
</tr>
<tr>
<td>GO:0007090</td>
<td>[regulation of S phase of mitotic cell cycle]</td>
</tr>
</tbody>
</table>
1. Metabolism

1.1 Carbohydrate Metabolism
   - Glycolysis / Gluconeogenesis
   - Citrate cycle (TCA cycle)
   - Pentose phosphate pathway
   - Pentose and glucuronate interconversions
   - Fructose and mannose metabolism
   - Galactose metabolism
   - Ascorbate and aldarate metabolism
   - Starch and sucrose metabolism
   - Aminosugars metabolism
   - Nucleotide sugars metabolism
   - Pyruvate metabolism
   - Glyoxylate and dicarboxylate metabolism
   - Propanoate metabolism
   - Butanoate metabolism
   - C5-Branch chain dicarboxylic acid metabolism
   - Inositol metabolism
   - Inositol phosphate metabolism

1.2 Energy Metabolism
   - Oxidative phosphorylation
   - Photosynthesis
   - Photosynthesis - antenna proteins
   - Carbon fixation
   - Reductive pentose cycle (CO2 fixation)
   - Methane metabolism
   - Nitrogen metabolism
   - Sulfur metabolism

1.3 Lipid Metabolism
   - Fatty acid biosynthesis
   - Fatty acid elongation in mitochondria
   - Fatty acid metabolism
   - Synthesis and degradation of ketone bodies
   - Biosynthesis of steroids
   - Bile acid biosynthesis
   - C21-Steroid hormone metabolism
   - Androgen and estrogen metabolism
   - Glycerolipid metabolism
   - Glycerophospholipid metabolism
   - Ether lipid metabolism
   - Sphingolipid metabolism

http://www.genome.jp/kegg/
Functional Annotation DDBB
MicroRNA

- Involved in gene regulation
- More than 8600 miRNAs
- The *target database* contains computationally predicted targets for microRNAs across many species

[http://microrna.sanger.ac.uk/](http://microrna.sanger.ac.uk/)
A **centralized database** of protein families, domains, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences.

[http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)

### Contents of InterPro 18.0

- Active site: 35
- Binding site: 24
- Conserved site: 166
- Domain: 4941
- Family: 11128
- PTM: 18
- Repeat: 237
Functional Annotation DDBB
CisRed

- Holds *conserved sequence motifs* identified by genome scale motif discovery, similarity, clustering, co-occurrence and coexpression calculations

http://www.cisred.org/

Databases of genome-wide regulatory module and element predictions

<table>
<thead>
<tr>
<th>Database</th>
<th>Assembly</th>
<th>Search regions</th>
<th>Search region type</th>
<th>Nbr. of input species</th>
<th>Conserved motifs</th>
<th>Discovery p-value threshold</th>
<th>Ensembl compatibility</th>
<th>Release date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 9</td>
<td>NCBI v36b</td>
<td>18.7k</td>
<td>promoter</td>
<td>41</td>
<td>236k</td>
<td>0.01</td>
<td>Build 38-49</td>
<td>26 Jul. 2007</td>
</tr>
<tr>
<td>Mouse 4</td>
<td>NCBI m37</td>
<td>17.5k</td>
<td>promoter</td>
<td>38</td>
<td>223k</td>
<td>0.1</td>
<td>Build 47-49</td>
<td>26 Sep. 2007</td>
</tr>
<tr>
<td>Mouse 3.1</td>
<td>NCBI m35</td>
<td>17.5k</td>
<td>promoter</td>
<td>38</td>
<td>223k</td>
<td>0.1</td>
<td>Build 38</td>
<td>18 Apr. 2007</td>
</tr>
<tr>
<td>Rat 1.1</td>
<td>RGSC v3.1</td>
<td>6.7k</td>
<td>promoter</td>
<td>28</td>
<td>116k</td>
<td>0.25</td>
<td>n/a</td>
<td>12 Feb. 2006</td>
</tr>
<tr>
<td>C. elegans 4</td>
<td>WormBase W5170</td>
<td>3.8k</td>
<td>promoter</td>
<td>8</td>
<td>158k</td>
<td>1.0</td>
<td>Build 44-46</td>
<td>18 Jul. 2008</td>
</tr>
<tr>
<td>Human Stat1 ChIP-seq peaks 1</td>
<td>NCBI v35</td>
<td>226</td>
<td>ChIP-seq</td>
<td>23</td>
<td>−6k</td>
<td>1.0</td>
<td>n/a</td>
<td>03 Apr. 2007</td>
</tr>
</tbody>
</table>
A curated protein sequence database with a high level of annotation, such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.

- Last release consist of 466739 entries.
- Now part of the Uniprot Knowledgebase

<table>
<thead>
<tr>
<th>UniProtKB/Swiss-Prot:BRCA2_HUMAN</th>
<th>Accession</th>
<th>UniSave</th>
<th>Description</th>
<th>GeneName</th>
<th>Species</th>
<th>Keywords</th>
<th>SeqLength</th>
</tr>
</thead>
<tbody>
<tr>
<td>P51587</td>
<td>P51587</td>
<td></td>
<td>Breast cancer type 2 susceptibility protein (Fanconi anemia group D1 protein).</td>
<td>BRCA2</td>
<td>Homo sapiens (Human).</td>
<td>3D-structure Disease mutation DNA damage DNA repair Fanconi anemia Phosphoprotein Polymorphism Repeat</td>
<td>3418</td>
</tr>
</tbody>
</table>

http://www.expasy.ch/sprot/
Database for known and predicted protein-protein interactions (direct and indirect associations)
Cover four sources of annotations: Genomic association (prokaryotes), high-throughput experiments (e.g. y2h), conserved co-expression, previous knowledge (text-mining).

A combined score is calculated for every association based on benchmarks of the different types of associations against a common reference set.
Functional Annotation DDBB
Practical exercise

- About BCL2, BRCA1, ATM and P53(TP53):
  - try to find the biological process and cellular components (GO terms), do they share some GO terms? Is that significant?
  - Are they target of the same microRNA (mirbase)?
  - What about protein functional domains (interpro)?
  - Are they regulated by the same conserved motifs (cisred)?
  - Are they involved in a common disease or pathway (kegg, biocarta)?
  - ...

...
Functional Annotation DDBB
From GEPAS to Babelomics

Microarray Data

Preprocessing (normalization, scaling, ...)
Tab matrix file

Next generation of High throughput Sequencing

GEPAS Analysis

Differential expression
Genes differentially expressed

Predictors
Predicting genes

Clustering
Genes with same expression patterns

CGH array
Genes from a deleted or duplicated region

List of genes (ie: 120 genes)

Babelomics
Functional Annotation DDBB

Babelomics try to answer these questions

- Is there any significant functional enrichment in my gene list?
- Are these genes involved in the same pathways?
- Are they sharing a specific microRNA regulation?
- Are they involved in the same disease?
- ...