Babelomics

Introduction to Functional Annotation Databases

Buenos Aires, October 2011

Javier Santoyo-Lopez
jsantoyo@cipf.es
http://bioinfo.cipf.es
Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)
Introduction

- In the last years it 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 principal *biological databases*

- Updated every year. The 2011 update includes 1330 databases!!

http://www3.oup.co.uk/nar/directory/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 33613 terms (Feb, 2011)
- The controlled vocabularies of terms are structured

http://www.geneontology.org/
Functional Annotation DDBB
Gene Ontology (GO terms)

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)

Annotations are given to the most specific (low) level.

True path rule: Annotation at a term implies annotation to all its parent terms.

Annotation is given with an Evidence Code:
- EXP: inferred from Experiment
- IDA: inferred by direct assay
- TAS: traceable author statement
- ISS: inferred by sequence similarity
- IEA: electronic annotation
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
GO slims are cut-down versions of the GO ontologies containing a subset of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms

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/

### Signature Database Information

<table>
<thead>
<tr>
<th>Signature Database</th>
<th>Version</th>
<th>Signatures*</th>
<th>Integrated Signatures**</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENE3D</td>
<td>3.0.0</td>
<td>2147</td>
<td>1021</td>
</tr>
<tr>
<td>HAMPAP</td>
<td>280509</td>
<td>1633</td>
<td>621</td>
</tr>
<tr>
<td>PANTHER</td>
<td>6.1</td>
<td>30128</td>
<td>2233</td>
</tr>
<tr>
<td>PIRSF</td>
<td>2.72</td>
<td>3222</td>
<td>2751</td>
</tr>
<tr>
<td>PRINTS</td>
<td>40.0</td>
<td>2000</td>
<td>1954</td>
</tr>
<tr>
<td>PROSITE patterns</td>
<td>20.52</td>
<td>1308</td>
<td>1290</td>
</tr>
<tr>
<td>PROSITE profiles</td>
<td>20.52</td>
<td>860</td>
<td>836</td>
</tr>
<tr>
<td>Pfam</td>
<td>24.0</td>
<td>11912</td>
<td>11285</td>
</tr>
<tr>
<td>PfamB</td>
<td>24.0</td>
<td>142303</td>
<td>0</td>
</tr>
<tr>
<td>ProDom</td>
<td>2006.1</td>
<td>1894</td>
<td>1011</td>
</tr>
<tr>
<td>SMART</td>
<td>6.0</td>
<td>809</td>
<td>801</td>
</tr>
<tr>
<td>SUPERFAMILY</td>
<td>1.69</td>
<td>1538</td>
<td>1093</td>
</tr>
<tr>
<td>TIGRFAMs</td>
<td>9.0</td>
<td>3808</td>
<td>3795</td>
</tr>
</tbody>
</table>

### Contents of InterPro 27.0 (May 2010)

- Active site: 89
- Binding site: 62
- Conserved site: 575
- Domain: 5515
- Family: 12718
- PTM: 16
- Region: 1441
- Repeat: 255
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
- Amino sugar and nucleotide sugar metabolism
- Pyruvate metabolism
- Glutamate and glutamine metabolism
- Glycolysis and dicarboxylic acid metabolism
- Propanoate metabolism
- Butanoate metabolism
- CS Branched dicarboxylic acid metabolism
- Inositol phosphate metabolism

1.2 Energy Metabolism
- Oxidative phosphorylation
- Photosynthesis
  - Photosynthesis - antenna proteins
  - Carbon fixation in photosynthetic organisms
- Reductive citric acid cycle in photosynthetic bacteria
- 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
- Steroid biosynthesis
- Primary bile acid biosynthesis
- Secondary bile acid biosynthesis
- Steroid hormone biosynthesis
- Glycerolipid metabolism
- Glycerophospholipid metabolism
- Ether lipid metabolism
- Sphingolipid metabolism
- Arachidonate metabolism
- Linoleic acid metabolism
- alpha-Linolenic acid metabolism
- Biosynthesis of unsaturated fatty acids

1.4 Nucleotide Metabolism
- Purine metabolism
- Pyrimidine metabolism

1.5 Amino Acid Metabolism
- Alanine, aspartate and glutamate metabolism
- Glycine, serine and threonine metabolism
- Cysteine and methionine metabolism
- Valine, leucine and isoleucine metabolism
- Valine, leucine and isoleucine biosynthesis
- Lysine biosynthesis

KEGG pathways

Current Statistics

<table>
<thead>
<tr>
<th>KEGG Databases</th>
<th>as of 2010/6/2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG PATHWAY</td>
<td>Pathway maps, reference (total)</td>
<td>361 (111,218)</td>
</tr>
<tr>
<td>KEGG BRITE</td>
<td>Functional hierarchies, reference (total)</td>
<td>90 (29,174)</td>
</tr>
<tr>
<td>KEGG MODULE</td>
<td>Pathway modules</td>
<td>537</td>
</tr>
<tr>
<td>KEGG DISEASE</td>
<td>Human diseases</td>
<td>374</td>
</tr>
<tr>
<td>KEGG DRUG</td>
<td>Drugs</td>
<td>9,454</td>
</tr>
<tr>
<td>KEGG ORTHOLOGY</td>
<td>KEGG Orthology (KO) groups</td>
<td>13,576</td>
</tr>
<tr>
<td>KEGG GENOME</td>
<td>KEGG Organisms</td>
<td>1,341</td>
</tr>
<tr>
<td>KEGG GENES</td>
<td>Genes in high-quality genomes</td>
<td>5,593,828</td>
</tr>
</tbody>
</table>

(131 eukaryotes + 1029 bacteria + 81 archaea)

http://www.genome.jp/kegg/
**Functional Annotation DDBB Reactome**

- It is a free, online, open-source, curated pathway database encompassing many areas of human biology. Information is authored by expert biological researchers.

http://www.reactome.org/

### Reactome - a curated knowledgebase of biological pathways

<table>
<thead>
<tr>
<th>Apoptosis</th>
<th>Axon guidance</th>
<th>Biological oxidations</th>
<th>Botulinum neurotoxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell junction organization</td>
<td>Cell Cycle Checkpoints</td>
<td>Cell Cycle, Mitotic</td>
<td>DNA Repair</td>
</tr>
<tr>
<td>DNA Replication</td>
<td>Diabetes pathways</td>
<td>Electron Transport Chain</td>
<td>Gap junction trafficking and regulation</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>Hemostasis</td>
<td>HIV infection</td>
<td>Influenza infection</td>
</tr>
<tr>
<td>Integration of energy metabolism</td>
<td>Integrin cell surface interactions</td>
<td>Metabolism of lipids and lipoproteins</td>
<td>Membrane Trafficking</td>
</tr>
<tr>
<td>Metabolism of amino acids and derivatives</td>
<td>Metabolism of carbohydrates</td>
<td>Metabolism of nitric oxide</td>
<td>Metabolism of nucleotides</td>
</tr>
<tr>
<td>Metabolism of polyamines</td>
<td>Metabolism of porphyrins</td>
<td>Metabolism of proteins</td>
<td>Metabolism of RNA</td>
</tr>
<tr>
<td>Metabolism of vitamins and cofactors</td>
<td>Muscle contraction</td>
<td>mRNA Processing</td>
<td>Myogenesis</td>
</tr>
<tr>
<td>Pyruvate metabolism and Citric Acid (TCA) cycle</td>
<td>Regulation of beta-cell development</td>
<td>Regulatory RNA pathways</td>
<td>Signaling by BMP</td>
</tr>
<tr>
<td>Signaling by EGFR</td>
<td>Signaling by FGFR</td>
<td>Signaling by GPCR</td>
<td>Signaling by PDGF</td>
</tr>
<tr>
<td>Signaling in Immune system</td>
<td>Signaling by Insulin receptor</td>
<td>Signaling by NGF</td>
<td>Signaling by Notch</td>
</tr>
<tr>
<td>Opioid Signalling</td>
<td>Signaling by Rho GTPases</td>
<td>Signaling by TGF beta</td>
<td>Signaling by VEGF</td>
</tr>
<tr>
<td>Signaling by Wnt</td>
<td>Synaptic Transmission</td>
<td>Telomere Maintenance</td>
<td>Transcription</td>
</tr>
<tr>
<td>Transmembrane transport of small molecules</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The data displayed is for *Homo sapiens*. Use the menu to change the species. Check ☑ for cross-species comparison.
Functional Annotation DDBB
MicroRNA

- Involved in gene regulation
- Last versions has 14197 entries (Release 15)
- The target database contains computationally predicted targets for microRNAs across many species

http://microrna.sanger.ac.uk/
The JASPAR database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes.

The prime difference to similar resources (TRANSFAC, etc) consist of the open data access, non-redundancy and quality.

http://jaspar.genereg.net/
Functional Annotation DDBB
ORegAnno

- It's an open database for the curation of known regulatory elements from scientific literature (TFBS)
- Annotation is collected from users worldwide for various biological assays

http://www.oreganno.org/oregano/

REGULATORY HAPLOTYPE: 7 entries.
REGULATORY REGION: 37520 entries.
TRANSCRIPTION FACTOR BINDING SITE: 14608 entries.
REGULATORY POLYMORPHISM: 175 entries.
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.

http://string.embl.de/
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 (ORegAnno)?
  - Are they involved in a common disease or pathway (kegg, reactome)?
  - ...

Functional Annotation DDBB
From Data Analysis to Functional Profiling

Data Analysis
- Preprocessing (normalization, scaling, ...)
- Tab matrix file

Differential expression
- Genes differentially expressed

Predictors
- Predicting genes

Clustering
- Genes with same expression patterns

CGH array
- Genes from a deleted or duplicated region

Next generation of High throughput Sequencing

Functional Profiling
- List of genes (i.e.: 120 genes)
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?
- ...