Babelomics

Functional interpretation of genome-scale experiments

Cambridge, Sept 2009

Ignacio Medina (Nacho)
imedina@cipf.es
http://bioinfo.cipf.es/imedina/wiki
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)
Index

- Babelomics Suite
  - FatiGO
  - Fatiscan
  - Other tools
- Next step: towards the GEPAS and Babelomics integration
Index

- **Babelomics Suite**
  - FatiGO
  - Fatiscan
  - Other tools
- Next step: towards the *GEPAS* and *Babelomics* integration
Babelomics
WEB tools suite

- A complete suite of web tools for the functional analysis of groups of genes in high-throughput experiments.
Babelomics

Key features

- Functional profiling by functional enrichment method (FatiGO) and gene set method (Fatiscan)
- Many functional and regulatory definitions (GO, KEGG, Biocarta, text-mining derived bioentities, Transcription factors, CisRed, miRNAs, InterPro, etc.)
- Wide coverage of model organisms (more than 10 species)
- Tools for automatic functional annotation of unknown sequences (Blast2GO) integrated
**Babelomics Tools**

**FatiGO**: Finds differential distributions of functional terms between two groups of genes, these terms can be: Gene Ontology, InterPro motifs, SwissProt KW, transcription factors (TF), gene expression in tissues, bioentities from scientific literature, cis-regulatory elements CisRed.

**Tissues Mining Tool**: compares reference values of gene expression in tissues to your results.

**MARMITE**: Finds differential distributions of bioentities extracted from PubMed between two groups of genes.

**FatiScan**: detect significant functions with Gene Ontology, InterPro motifs, Swissprot KW and KEGG pathways in lists of genes ordered according to different characteristics.

**MarmiteScan**: Use chemical and disease-related information to detect related blocks of genes in a gene list with associated values.
Babelomics Databases

Integrated Biological DB of Functional Annotation for more than 10 species (codename: "infrared")
- **INFRARED** database is freely distributed, to access data:
  - A web page to access to data is already integrated in Babelomics
  - *Java API* to access to data is being developed (local installation of database required)
  - *Web services* available
  - *DAS server* with the annotation is also being developed
- Next releases of *Infrared* will add more regulatory elements and pathways like *epigenetic* changes and *Reactome* database
Babelomics
Schema

Babelomics, a suite of web tools for: statistical test, multiple test corrections, blast, ...

List of genes or ids, ie: differentially expressed genes, ...

Integrated Biological DB of Functional Annotation (codename: infrared)
Index

- Babelomics Suite
  - FatiGO
  - Fatiscan
  - Other tools
- Next direction: towards the GEPAS and Babelomics integration
Babelomics
Two-steps functional interpretation

Genes are selected based on their experimental values and...
Enrichment in functional terms is tested (FatiGO, GoMiner, etc.)
Babelomics
FatiGO

- It allow us to compare functional annotation of:
  - Two list of genes
  - One list against the rest of genome
  - A genomic region against the rest of genome
- One statistical test for each Functional Block of annotation
  - Fisher's exact test
  - Multiple testing context (hundreds of annotation)
  - Filtering of annotation is convenient (the less tests the best correction)
Babelomics
FatiGO test

One Gene List (A)

Biosynthesis 60%
Sporulation 20%

Genes in group A have significantly to do with biosynthesis, but not with sporulation.

The other list (B)

Biosynthesis 20%
Sporulation 20%

Are these two groups of genes carrying out different biological roles?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biosynthesis</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>No biosynthesis</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

We do this for each GO, miRNA, Interpro, ... !!!
FatiGO
Functional enrichment

FatiGO performs functional enrichment analysis by comparing two lists of genes by means of a Fisher’s exact test. Gene modules used in the test are defined in different ways which include functional criteria (GO, KEGG, Biocarta, etc.) regulatory criteria (transcription factor targets, miRNA, etc.) and chromosomal location. Also user-defined gene modules can be imported and used for functional enrichment.

Select the organism

Choose an organism:
- Homo sapiens

List of genes

A list of genes

or a gene file
- file from your computer
- or from your projects

List of genes #2

A list of genes

or a gene file
- file from your computer
- or from your projects

Or use the rest of genome
Babelomics
FatiGO Compare options

Databases available

Select Fisher's exact test
Babelomics
FatiGO Your Annotations

Useful when you work with your own annotations or with an specie that is not in Babelomics

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>GO Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>38969_at</td>
<td>GO:0003677</td>
</tr>
<tr>
<td>37639_at</td>
<td>GO:0006306</td>
</tr>
<tr>
<td>37149_s_at</td>
<td>GO:0004674</td>
</tr>
<tr>
<td>37149_s_at</td>
<td>GO:0005525</td>
</tr>
<tr>
<td>37639_at</td>
<td>GO:0006306</td>
</tr>
<tr>
<td>38052_at</td>
<td>GO:0017111</td>
</tr>
<tr>
<td>38052_at</td>
<td>GO:0016021</td>
</tr>
<tr>
<td>38840_s_at</td>
<td>GO:0016021</td>
</tr>
</tbody>
</table>

Compare | Your annotations | Genomics | Search

List of Genes #1
- A list of genes
- or a gene file

List of Genes #2
- A list of genes
- or a gene file

Your annotations
- A list of annotations
- or an annotations file
Babelomics

FatiGO Genomic

Choose the region

Useful when you have CGH arrays or CNV data.
# Babelomics

## FatiGO results

<table>
<thead>
<tr>
<th>KEGG pathways</th>
<th>percentages</th>
<th>p-values</th>
<th>corrected p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ribosome</td>
<td>25.71%</td>
<td>0.00012</td>
<td>0.00273</td>
</tr>
<tr>
<td>Pyruvate metabolism</td>
<td>11.32%</td>
<td>0.07711</td>
<td>1.00000</td>
</tr>
<tr>
<td>Arginine and proline metabolism</td>
<td>7.75%</td>
<td>0.17804</td>
<td>1.00000</td>
</tr>
<tr>
<td>Alanine and aspartate metabolism</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>GnRH signaling pathway</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>Tight junction</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>Glycolysis / Gluconeogenesis</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>Valine, leucine and isoleucine degradation</td>
<td>7.55%</td>
<td>0.14003</td>
<td>1.00000</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle)</td>
<td>9.43%</td>
<td>0.15237</td>
<td>1.00000</td>
</tr>
<tr>
<td>Complement and coagulation cascades</td>
<td>5.71%</td>
<td>0.15843</td>
<td>1.00000</td>
</tr>
<tr>
<td>Biosynthesis of steroids</td>
<td>5.71%</td>
<td>0.15843</td>
<td>1.00000</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>11.32%</td>
<td>0.23593</td>
<td>1.00000</td>
</tr>
<tr>
<td>Carbon fixation</td>
<td>5.66%</td>
<td>0.27312</td>
<td>1.00000</td>
</tr>
</tbody>
</table>
Babelomics
FatiGO exercises

- Tool demo
- Go to the tutorial:
- and try to reproduce the examples
Babelomics
FatiGO approach may not be very powerful

If a threshold based on the experimental values is applied, and the resulting selection of genes compared for enrichment of a functional term, this might not be found significantly over-expressed in A and B. Very few genes selected to arrive to a significant conclusion on GO1 and GO2.

Functional Classes expressed as blocks in A and B.
A previous step of gene selection causes loss of information and makes the test insensitive.

The main problem is that the two-steps approach cannot distinguish between these two different cases.

We put both sides of the partition into two bags and destroy the structure of the data.

<table>
<thead>
<tr>
<th>GO</th>
<th>up</th>
<th>down</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>no GO</td>
<td>0</td>
<td>25</td>
</tr>
</tbody>
</table>

Same contingency table for GO\textsubscript{1} and GO\textsubscript{2}!!
Index

- Babelomics Suite
  - FatiGO
  - *Fatiscan*
  - Other tools
- Next step: towards the *GEPAS* and *Babelomics* integration
Babelomics
Gene-set enrichment methods

FatiScan

A B

statistic

significance

Gene set

background

Independent of the experimental design
Babelomics
Fatiscan features

- Interpret a ranked list of genes
- There is not need for choosing a cut-off (all information is included)
- One statistical test for each Functional Block of annotation
  - Multiple testing context (hundreds of annotation)
  - Filtering of annotation is convenient (the less tests the best correction)
Babelomics
Fatiscan, testing along an ordered list

- Index ranking genes according to some biological aspect under study.

- Database that stores gene class membership information.

- **FatiScan** searches over the whole ordered list, trying to find runs of functionally related genes.

![Diagram](image)

- Block of genes enriched in the annotation **A**
- Annotation **C** is homogeneously distributed along the list
- Block of genes enriched in the annotation **B**
Babelomics
Fatiscan results

Gene Ontology: biological process
Babelomics
Fatiscan results

Functional Blocks over-represented among genes over-expressed in A

Gene ranking index

Functional Blocks over-represented among genes over-expressed in B
Babelomics
Fatiscan Web tool

List of genes

Functional databases
Babelomics
Fatiscan exercises

- Tool demo
- Go to the tutorial:
- and try to reproduce the examples
Index

- Babelomics Suite
  - FatiGO
  - Fatiscan
  - Other tools
- Next step: towards the GEPAS and Babelomics integration
Babelomics

Blast2GO

Blast2GO is a web resource for high-throughput functional annotation of (novel) DNA or protein sequences. Blast2GO has two steps: blast searches to find homologous sequences and annotation to select an optimized functional assignment. The specificity and coverage of the annotation process is adjustable through the application parameters.

<table>
<thead>
<tr>
<th>Blast</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequence File in FASTA format (1-10000 seqs)</strong></td>
<td></td>
</tr>
<tr>
<td>file from your computer</td>
<td></td>
</tr>
<tr>
<td>or from your projects</td>
<td></td>
</tr>
<tr>
<td><strong>Sequence type</strong></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td></td>
</tr>
<tr>
<td>Nucleotide</td>
<td></td>
</tr>
<tr>
<td><strong>Blast DB</strong></td>
<td></td>
</tr>
<tr>
<td>nr</td>
<td></td>
</tr>
<tr>
<td><strong>Number of blast hits</strong></td>
<td></td>
</tr>
<tr>
<td>20 (1-50)</td>
<td></td>
</tr>
<tr>
<td><strong>Blast Expect Value</strong></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td><strong>Blast Program</strong></td>
<td></td>
</tr>
<tr>
<td>blastx</td>
<td></td>
</tr>
</tbody>
</table>
MARMITE allows performing functional enrichment tests using text-mining derived gene modules. Three types of definitions are used: generic functional terms (word roots), genes associated to diseases or symptoms and genes associated to chemical compounds.

<table>
<thead>
<tr>
<th>Bio-Entity</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease associated words</td>
<td>☐</td>
</tr>
<tr>
<td>Chemical products</td>
<td>☐</td>
</tr>
<tr>
<td>Word roots</td>
<td>☐</td>
</tr>
</tbody>
</table>

**Filter bio-entities**

<table>
<thead>
<tr>
<th>Minimum number of genes with a score</th>
<th>5 (0-10000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bio-entities in results</td>
<td>50 (0-10000)</td>
</tr>
</tbody>
</table>

**List in Gene names?**

- click if HUGO, HGNC, gene names [☐]

**List of Genes #1**

- or genes list file #1

**List of Genes #2**

- or genes list file #2
Babelomics and others...

- **Tissues Mining Tool (TMT)**: compares expression of two lists of genes in a set of tissues
- **ID Converter**: 10 species and almost of the existing Ids
- **GOGraphViewer**: a DAG viewer tool generates joined gene ontology graphs (DAGs)
- ...
Index

Babelomics Suite

- FatiGO
- Fatiscan
- Other tools

Next step: towards the GEPAS and Babelomics integration
Next step
GEPAS y Babelomics integration

- due in October
- New web interface based in new technologies of the WEB 2.0
- DAS server and WEB services for almost all tools
- Local installation and command line execution available
- Many bugfixes and performance improvements
- A Java framework to run new tools
- ...
Next step
New tools

- New tools:
  - Transcriptome analysis
  - Logistic Regression method to gene set analysis
  - Genome and annotation browser (DAS based)
  - Genome browser
  - SNPs and methylation analysis*
  - Next generation sequence methods*
  - Epigenome and ChIP-on-Chip analysis*
  - ...

* indicates new tools with a focus on epigenetic analysis.