Functional annotation with Blast2GO
Index

✓ Why Blast2GO? Concepts on functional annotation
  ○ Function assignment
  ○ Vocabularies
  ○ GO and GOA
  ○ The GO Direct Acyclic Graph
  ○ The problem of function transfer
✓ Blast2GO Java application + Practicals
✓ Blast2GO @ Babelomics + Practicals
Why Blast2GO?

Experiment → Data-Analysis → Gene-List

Functional interpretation

Data-Analysis

Gene-List

MNAT1
CTNNBL1
ENOX2
GTPBP1
RALY
TAGLN2
RAB3A
PPP2R5A
MAPRE1

Functional Profiling

Functional Profiling +

Functional Annotation

Functional Annotation +

Bioinformatics
What does Blast2GO do?

Generates annotations      Visualization of functional annotations
Concepts of functional annotation

✓ Gene/Protein function
  • Referees to the molecular function of a gene or a protein:
    Tyrosine kinase

✓ Functional annotation
  • More general, referees to the characterization of functional aspect of the protein.
    Stress-related, cytoplasm, ABC transporter
  • Also referees to the process of assingment of a function label
  • Habitually, standard vocabularies are used to assign function
Functional Vocabularies

The Gene Ontology

Molecular Function
Biological Process
Cellular Component

Functional motifs

Example proteins

KEGG orthologues

Metabolic pathways
The GO has a DAG structure
KEGG records pathways

Metabolism of xenobiotics by cytochrome P450 - Danio rerio (zebrafish)
InterPro indicate functional motifs

Example proteins

- **O09158** Cytochrome P450 3A25

- **O17624** Putative cytochrome P450 cyp-13B1

- **O46051** Probable cytochrome P450 4d14

- **P05177** Cytochrome P450 1A2

- **P10614** Cytochrome P450 51
Functional assignment

- Annotation
  - Empirical
    - Literature reference
    - Molecular interactions
    - Gene/protein expression
    - Biochemical assay
  - Transference
    - Structure Comparison
      - Identification of folds
    - Sequence analysis
      - Motif identification
  - Filogeny
    - Sequence homology
Automatic annotation

- GO annotations can be created by comparison to annotated sequences
- To achieve enough coverage, high-throughput, automatic annotation is required
- The most effective (also error prone) automatic annotation method is transfer from sequence similarity
Concerns in functional transfer by similarity

- Level of homology (~ from 40-60% is possible)
- The overlap query and hit sequences
- The domain or structure function association
- The paralog problem: genes with similar sequences might have different functional specifications
- The evidence for the original annotation
- Balance between quality and quantity: depends on the use
Blast2GO

✓ Suite for functional annotation and data mining on functional data
  ○ Considerations for annotation
    • Similarity
    • Length of the overlap
    • Percentage of hit sequence spanned by the overlap
    • Evidence original annotation
    • Blast hits and motif hits
    • Refinement by additional methods
  ○ Visualization:
    • Annotation charts
    • Knowledge discovery on the DAG

✓ Desktop Java application
✓ web interface @ Babelomics: Babelomics for non-model
Blast2GO Annotation strategy

**Blast**

<table>
<thead>
<tr>
<th>Sq1</th>
<th>Sq2</th>
<th>Sq3</th>
<th>Sq4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit1</td>
<td>Hit2</td>
<td>Hit1</td>
<td>Hit1</td>
</tr>
<tr>
<td>Hit3</td>
<td>Hit4</td>
<td>Hit2</td>
<td>Hit2</td>
</tr>
<tr>
<td>Hit4</td>
<td>Hit1</td>
<td>Hit3</td>
<td>Hit3</td>
</tr>
<tr>
<td>Hit1</td>
<td>Hit2</td>
<td>Hit4</td>
<td>Hit4</td>
</tr>
</tbody>
</table>

**Mapping**

<table>
<thead>
<tr>
<th>Sq1</th>
<th>Sq2</th>
<th>Sq3</th>
<th>Sq4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit1</td>
<td>go1</td>
<td>go2</td>
<td>go1</td>
</tr>
<tr>
<td>go2</td>
<td>go3</td>
<td>go4</td>
<td>go3</td>
</tr>
<tr>
<td>go3</td>
<td>go5</td>
<td>go6</td>
<td>go1</td>
</tr>
<tr>
<td>go6</td>
<td>go8</td>
<td>go1</td>
<td>go8</td>
</tr>
<tr>
<td>go1</td>
<td>go4</td>
<td>go6</td>
<td>go1</td>
</tr>
<tr>
<td>go2</td>
<td>go4</td>
<td>go2</td>
<td>go2</td>
</tr>
<tr>
<td>go2</td>
<td>go5</td>
<td>go6</td>
<td>go2</td>
</tr>
<tr>
<td>go2</td>
<td>go4</td>
<td>go2</td>
<td>go2</td>
</tr>
</tbody>
</table>

**Annotation**

<table>
<thead>
<tr>
<th>Sq1</th>
<th>Sq2</th>
<th>Sq3</th>
<th>Sq4</th>
</tr>
</thead>
<tbody>
<tr>
<td>go1</td>
<td>go2</td>
<td>go8</td>
<td>go8</td>
</tr>
<tr>
<td>go1</td>
<td>go3</td>
<td>go8</td>
<td>go8</td>
</tr>
<tr>
<td>go3</td>
<td>go1</td>
<td>go8</td>
<td>go8</td>
</tr>
<tr>
<td>go2</td>
<td>go2</td>
<td>go2</td>
<td>go2</td>
</tr>
<tr>
<td>go2</td>
<td>go2</td>
<td>go2</td>
<td>go2</td>
</tr>
<tr>
<td>go2</td>
<td>go4</td>
<td>go2</td>
<td>go4</td>
</tr>
</tbody>
</table>
Blast2GO Annotation Strategy

Sq1: go1, go2, go3, go4
Sq2: go8
Sq3: go2, go2, go2, go4
Sq4: go1, go3, go4, go6, go8

Refinement
InterPro
Annex
GOSlim
Manual

Sq1: go1, go2, go3, GO11
Sq2: go8, GO12, GO13
Sq3: go2, go4
Sq4: GO15
Blast2GO annotation rule

Lowest term above threshold

Similarity requirement

Quality of annotation source

Evidence codes

Possibility of abstraction

Recall vs. Precision

Lowest.node [(max.sim x ECw) + (#GO-1 x GOw) >= threshold]
Blast2GO annotation rule

Lowest.node \[(\text{max.sim} \times \text{ECw}) + (\#\text{GO}-1 \times \text{GOw}) \geq \text{threshold}\]

- When I have a GO with ECw = 1 and I do not allow abstraction (GOw = 0), then the Annotation Score = %similarity

- If the ECw < 1 my similarity requirement is higher to obtain the same Annotation Score

- If I allow abstraction GOw > 0, then with less similarity I can obtain the required Annotation Score at a parent node
Usage

Blast2GO citations per year

- Current year
- Previous

- mamalians: 5
- insect/artropods: 18
- fungi: 10
- microorganisms: 13
- plant: 26
- tools: 24
- fish: 22
Hands-on
Welcome to Blast2GO

Blast2GO is an ALL in ONE tool for functional annotation of (novel) sequences and the analysis of annotation data.

Main Application Features

- Easy start-up and low maintenance. Make sure you have JAVA, download Blast2GO from this site and start using the application. Updates are automatic.
- User-friendly: Blast2GO is designed for experimentalists. An intuitive interface, the many graphical parameters and the detailed user manual makes the use of the tool possible from the first try.
- High-throughput and interactive: Blast2GO can annotate THOUSANDS of sequences in one session. Users can follow and modify the annotation process at any stage.
- Highly configurable: Blast2GO is a functional annotation repositioner. You can design your custom annotation style through the many configurable parameters. Statistical charts are available to guide users in the annotation process.
- Data mining: Blast2GO does not only generate functional annotations. You can interrogate the biological meaning of your data with different graphical and statistical functions.
- Vocabularies: Gene Ontology, KEGG maps, InterPro and Enzyme Codes are supported by Blast2GO.
- Functional annotation through three steps: For Gene Ontology functional annotation follow the 3 application steps: Blast against public or private databases, Mapping against GO resources to fetch functional data. Annotation to generate useful functional assignments.
- Blast2GO is used by more than hundred labs worldwide. It has contributed to the annotation efforts in EST projects, genomic studies and microarray experiments, covering taxa from microorganisms to fungi, plants, fish, animals and humans. In the last two years the application has been cited in more than 40 scientific papers.

LAUNCH B2G
Input data (in fasta format)

>my_favorite_species_seq1 | still unknown
gtgatggaagaagaagttttgctatctgtagcatcacatggttttttcgctcggtatattgctgcctggaattaaagccctcatacataatctttctgttagag
gttatagttttttataaaacatactttttgaaacatctcttctttcagatgcagattatttagttgtgtagttatgatcgggtgtggagatttttaacgtctgtgatggactatacctctgtatatataacatcagttgtatttttagttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
Blast2GO Application

(1) Blast
(2) Mapping
(3) Annotation

Table with all the sequence information

Application statistics
Blast results
Application messages
Graph visualisation
Finding the homologues (BLAST)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blast Server URL</td>
<td>Where to run Blast</td>
</tr>
<tr>
<td>Blast DB</td>
<td>Choose Database</td>
</tr>
<tr>
<td>Number of Blast Hits</td>
<td>Number of blast hits</td>
</tr>
<tr>
<td>Blast ExpectValue</td>
<td>e-Value cut-off</td>
</tr>
<tr>
<td>Blast Program</td>
<td>Blast algorithm</td>
</tr>
<tr>
<td>Blast Mode</td>
<td>Blast mode</td>
</tr>
<tr>
<td>HSP length cutoff</td>
<td>HSP length cut-off</td>
</tr>
<tr>
<td>Save result as</td>
<td>save your results apart</td>
</tr>
<tr>
<td>Alignments from Simap first</td>
<td>(only for proteins)</td>
</tr>
</tbody>
</table>

![BLAST Configuration](image_url)

**Parameter Descriptions**

- **Blast Server URL**: Where to run Blast.
- **Blast DB**: Choose Database.
- **Number of Blast Hits**: Number of blast hits.
- **Blast ExpectValue**: e-Value cut-off.
- **Blast Program**: Blast algorithm.
- **Blast Mode**: Blast mode.
- **HSP length cutoff**: HSP length cut-off.
- **Save result as...**: save your results apart.
Mapping to existing func. information

<table>
<thead>
<tr>
<th>nr</th>
<th>Sequence name</th>
<th>seq description</th>
<th>length</th>
<th>max e-value</th>
<th>sim mean</th>
<th>#GOs</th>
<th>GO IDs</th>
<th>Enzyme</th>
<th>Interpro</th>
</tr>
</thead>
<tbody>
<tr>
<td>177</td>
<td>C0100B10</td>
<td>--NA--</td>
<td>319</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>178</td>
<td>C0100B11</td>
<td>dynamin light ch...</td>
<td>406</td>
<td>29</td>
<td>1e-47</td>
<td>75%</td>
<td>4</td>
<td>F:GO:0003777 P:GO:0005515</td>
<td>-</td>
</tr>
<tr>
<td>179</td>
<td>C0100B03</td>
<td>--NA--</td>
<td>359</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>C0100BC4</td>
<td>afl367317.1at...</td>
<td>407</td>
<td>7</td>
<td>1e-23</td>
<td>76%</td>
<td>2</td>
<td>F:GO:003674 P:GO:0008150</td>
<td>-</td>
</tr>
<tr>
<td>181</td>
<td>C0100BC5</td>
<td>sebolisin-like ...</td>
<td>396</td>
<td>20</td>
<td>1e-23</td>
<td>68%</td>
<td>7</td>
<td>C:GO:0009536 F:GO:0004289</td>
<td>-</td>
</tr>
<tr>
<td>182</td>
<td>C0100BC6</td>
<td>transformer-sr</td>
<td>418</td>
<td>20</td>
<td>1e-33</td>
<td>87%</td>
<td>1</td>
<td>F:GO:003723</td>
<td>-</td>
</tr>
<tr>
<td>183</td>
<td>C0100BC7</td>
<td>gmpk kinasesa</td>
<td>424</td>
<td>20</td>
<td>1e-47</td>
<td>66%</td>
<td>8</td>
<td>F:GO:0015901 P:GO:0005524</td>
<td>-</td>
</tr>
<tr>
<td>184</td>
<td>C0100BD2</td>
<td>--NA--</td>
<td>74</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>185</td>
<td>C0100BD6</td>
<td>--NA--</td>
<td>416</td>
<td>20</td>
<td>1e-19</td>
<td>70%</td>
<td>9</td>
<td>F:GO:0015031 C:GO:0016021</td>
<td>-</td>
</tr>
<tr>
<td>186</td>
<td>C0100BD7</td>
<td>trm7-like prot</td>
<td>382</td>
<td>14</td>
<td>1e-19</td>
<td>71%</td>
<td>9</td>
<td>C:GO:0008739 P:GO:0005741</td>
<td>-</td>
</tr>
</tbody>
</table>

**GO Graphs**

- Evidence code (EC) distribution for blast hits
- Database sources of mapping
Assigning function

... setting the annotation parameters ...
## Import and Export results

<table>
<thead>
<tr>
<th>SeqID</th>
<th>GO ID</th>
<th>SeqDesc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01001C12</td>
<td>GO:0009507</td>
<td>at1g79420 t8k14_16</td>
</tr>
<tr>
<td>C01001F11</td>
<td>GO:0003677</td>
<td>translin-associated factor x</td>
</tr>
<tr>
<td>C01001G01</td>
<td>GO:0012505</td>
<td>---NA---</td>
</tr>
<tr>
<td>C01001H01</td>
<td>GO:0009987</td>
<td>pseudo-response regulator</td>
</tr>
<tr>
<td>C01001H03</td>
<td>GO:0030001</td>
<td>proline-rich protein</td>
</tr>
<tr>
<td>C01001H03</td>
<td>GO:0046872</td>
<td></td>
</tr>
</tbody>
</table>
Add functions from protein domains

Your e-mail address

different methods offered by InterProScan from EBI
Additional Annotations: InterPro Motifs and Kegg Maps

• Ordered Kegg Map visualization
• Highlighting of involved Sequences

• Batch-InterPro-Motif searches
• Add/merge GOs from domains to annotation
Modulation of annotation: GO Second Layer

Extend annotation by the GO “Second Layer”

Ref: Myhre, Tveit, Mollestad, Lægreid: Additional Gene Ontology structure for improved biological reasoning
Bioinformatics, 2006
Modulation of annotation

Chance annotation manually.

Summarize annotation by "GoSlim"

- GO-Term ACC
- EC-Codes
- Seq. Description
- OBO GO-Slim File
Visualisation

- GO Graph Visualization as tool to explore and discover
- Interactive and zoomable graphs
- Colored graphs highlighting areas of interest

set filters for more handy graphs
Node Score

\[ \text{NodeScore} = \sum (\text{Number of genes} \times \alpha^{\text{distance to origin}}) \]

e.g. \( \alpha = 0.6 \)

Abundance analysis (Multilevel Pie)

- **Seq Distribution: molecular function (Filtered by #Seqs: cutoff=10.0)**
  - Transferase activity = 114
  - Hydrolase activity = 125
  - Oxidoenolyase activity = 110
  - Nucleoside binding = 102
  - Molecular function unknown = 107
  - Protein nucleotide binding = 87

Khatri et al. Ontological analysis of gene expression data: current tools, limitations and open problems, Bioinformatics
Blast2GO in Babelomics

- Annotation
- Visualization
Annotation generation within Babelomics
Graph visualisation in Babelomics
Several Examples/Exercises

- Blast, map and annotate several few sequences in Blast2GO by loading the 10 test sequences (within the file menu). Generate some singel-Seq GO graphs to review annotation. (right mouse click on sequence table) 
  (http://www.blast2go.org → Start → 1024MB)

- Annotated 500 Citrus-Unigenes (nt) with Blast2GO as stand-alone and in Babelomics. Analyse the annotation results. (For Babelomics: Leave default parameters; takes some 3 to 5 minutes)
  Please find the corresponding instructions here: http://bioinfo.cipf.es/babelomicswiki/tool:annotation

- Visualizes some gene annotations (breast cancer) with Babelomics GO-Graph-Viewer (using the ID-Converter Tools) (http://bioinfo.cipf.es/babelomicswiki/tool:gographviz)

- Import the gene annotations (breast cancer) as .annot into Blast2GO and try to export a handy graph as PDF.