My internship at Computational Diagnostics Group


13. Oktober 2003 - Martin Held
Outline

1. Task
2. Approach
3. Gene Ontology
4. GoGrouper
   - Functionality
   - Example
   - Perspectives
5. Summary
Development of a R tool, that supports the analysis of gene expression experiments in the following way:

- reasonable annotation of significant Affymetrix® Probe IDs (GeneName, Description, GO-Terms etc.)
- provide links to DBs with further information
- generate groups of Affy. IDs, according to their „biological information“
Approach

- Use Gene Ontology annotations to create a distance measure between probes.
- Create a probe distance matrix (#probes x #probes).
- Cluster these matrix to get groups of probes.
GO? – What’s that?

- project to provide controlled vocabularies
- vocabularies to describe molecular function, biological process, and cellular location of gene products
GO? – What’s that?

Gene ontology
- biological_process ontology
  - GO:049448
devlopment
  - GO:005783
pattern specification
  - GO:000364
  - adaxial/abaxial pattern formation
    - GO:000005
axis specification
  - GO:000177
  - ataxia/adaxial axis specification
    - GO:000003
  - polarity of the adaxial/abaxial axis
    - GO:000002

INNER NC OUTER gene product
- nucleus
  - GO:007653
- intracellular
  - GO:019611
- cell
  - GO:027473
- cellular component
  - GO:037123
- gene ontology

Transcription factor activity
  - GO:002118
DNA binding activity
  - GO:004636
nucleic acid binding activity
  - GO:007157
binding activity
  - GO:017525
molecular_function ontology
  - GO:057721
- gene ontology

Key
- A single gene, annotated to one node on each of the three ontologies.
- All the nodes leading to the gene INNER NC OUTER.
**GoGrouper - Functionality**

### Probe IDs

- 206083_at, 204455_at, 212253_x_at, 212254_s_at,
- 215016_x_at, 216918_s_at, 220154_at,
- 201605_x_at, 204636_at, 201533_at, 202514_at,
- 202515_at, 202516_s_at, 215988_s_at,
- 217208_s_at

### GoGrouper

- **ion transport**
  - 201533_at, 202514_at, 202515_at, 202516_s_at,
  - 215988_s_at, 217208_s_at

- **cell growth**
  - 206083_at, 204455_at, 212253_x_at, 212254_s_at,
  - 215016_x_at, 216918_s_at

- **immune response**
  - 220154_at, 201605_x_at, 204636_at
### GoGrouper - Functionality

#### Example: Cellular Component Distances

GoGrouper

37573 at
- GO:0005871
- GO:0005615

37908 at
- GO:0005834

Node distance = number of edges between to GO-nodes

Probe distance = minimum of all probe node distances (e.g. 3)

### Probe Distance Matrix (CC)

<table>
<thead>
<tr>
<th></th>
<th>37573_at</th>
<th>37908_at</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>37573_at</td>
<td>0</td>
<td>?</td>
<td>3</td>
</tr>
<tr>
<td>37908_at</td>
<td>?</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>...</td>
<td>?</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>
GoGrouper - Functionality

- calculating “optimal” number of clusters
- clustering of probe distance matrix → probe groups
- calculating centroid for each cluster
- get probe annotations from R packages
- generating HTML page
GoGrouper - Example
GoGrouper - Perspectives

- add statistical values to result pages
- improvement of „needed number of clusters“ calculation
- implementation of a dynamic probe dependend GO-Distance Matrix
- implementation of a more efficient boolean matrix multiplication
- providing web access to GoGrouper
- usage of a weighted GO-DAG (weight decreases with depth)
Summary

- R script
- an approach to assist gene expression analysis
- makes use of the graph structure of gene ontology to group probes
- creates a clear HTML output
Questions
Thanks for your attention!

Thanks to Claudio and Jörn, for R support!

Thanks to Jochen, for help with APSP approach!

Thanks to all other group members, for helping me!

Thanks to Rainer, for giving me the opportunity!
My internship at Computational Diagnostics Group – Martin Held