Expression Profiling

measurements over multiple tissues (or time point)
give individual gene **expression profiles** (expression vectors)
Multidimensional experiments

6000 genes in 10 patients:
- 6000 points in 10-dimensional space (gene view)
- 10 points in 6000-dimensional space (chip view)

Reduction of dimensionality:
- Principal Component Analysis
- Clustering
- Correspondence analysis
Principal component analysis

Going from 3 dimensions:

To 2 dimensions:

Second principal component

First principal component
Hierarchical clustering

Gene Expression in Patient A1

Gene expression in Patient N1

Euclidean Distance
K-means clustering
Bladder Cancer

- Superficial progression stages:
  - Ta grI
  - Ta grII
  - Ta grIII
  - T1
  - T2+grIII
  - T2+grIV

- Invasive progression stages:
  - T2+grIII
  - T2+grIV

Numbers associated with each stage:
Bladder Cancer Classifier

- Superficial:
  - Ta
- Invasive:
  - T1
  - T2+

Number of genes

Patient number

Correct classifications (%)

- 400 (correlated)
- 4000 (unbiased)
Promoter Analysis

Genes that pass the statistical significance test are clustered and their corresponding promoter regions extracted. In each cluster, transcription factor binding sites are searched using three different approaches:

1. saco-patterns. Searches for exact matches to words that are overrepresented in cluster relative to background set.
2. Gibbs sampler. Searches for matches to weight matrix description of binding site overrepresented in cluster relative to background set.
3. Transfac. Searches for overrepresentation of known transcription factor binding sites in cluster relative to background set.

The results are output in table format.
Inference of Regulatory Networks

When we delete gene $a$, we find that the expression of gene $b$ and $d$ decreases. We conclude that gene $a$ has a stimulatory effect, directly or indirectly, on genes $b$ and $d$. We can represent this information in an interaction matrix:

<table>
<thead>
<tr>
<th></th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
</tr>
<tr>
<td>$a$</td>
<td>+</td>
</tr>
<tr>
<td>$b$</td>
<td>+</td>
</tr>
<tr>
<td>$c$</td>
<td>-</td>
</tr>
<tr>
<td>$d$</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Interaction matrix between four genes.

![Redundant network](image1.png)

![Parsimonious network](image2.png)

Figure 1: Regulatory networks deduced from experimentally produced interaction matrix. Arrows mean positive regulation; bars mean negative regulation.
Regulation of nitrogen metabolism in *Bacillus subtilis*

Regulatory network as it is known from biology

Regulatory network as it was predicted by computer
Curse of dimensionality and multiple testing:

- Limit the number of conditions you are comparing and the number of genes you are testing.

- Focus on replication.

- Avoid pooling, except where it is necessary in order to get sufficient material. Measure the variation instead.

For spotted arrays of high quality, you can compare between slides, making array design much simpler (compare within slide variance to between slide variance)
Read files
Affymetrix CEL or genetable

Normalize
(Qspline)

Array type
CEL

Calculate expression
(Li-Wong)

genetable

PCA, clustering and classification of chips

LaTeX

report
(PDF)

List of significantly regulated genes

Gene clustering,
Corr. anal.,
promoter anal.

Link to GO,
LocusLink,
Protfun, KEGG Transpath

ANOVA
(Benjamini-Hochberg)

t-test
(Benjamini-Hochberg)

Number of categories

>2

2

Analysis Flowchart
GenePublisher 1.0 Server

This server accepts gene tables or Affymetrix CEL files as input, performs numerical and statistical analysis, links the results to various databases, and returns a report of the results. (Previously known as GeneMachine).

This is version 1.00 4/2/2003.

Sample output report

Input restrictions:

- The input CEL files must be compressed with gzip (available from www.gzip.org) and cannot be more than 30 Mbytes in total size. That corresponds to 12 HuGeneFL chips, 10 HG_U95A chips, 22 Focus chips, 8 HG-U133A chips, and so on. The maximum number of chips accepted on this web server is currently 12.
- Avoid using characters _, & , %, # in the input as they may interfere with analysis.

Title of report: ____________________________
Author name on report: ____________________________

<table>
<thead>
<tr>
<th>Specify CEL files (minimum 4, maximum 12):</th>
</tr>
</thead>
<tbody>
<tr>
<td>File name</td>
</tr>
<tr>
<td>Ctrl1</td>
</tr>
<tr>
<td>Ctrl2</td>
</tr>
<tr>
<td>Exp1</td>
</tr>
<tr>
<td>Exp2</td>
</tr>
<tr>
<td>Ctrl3</td>
</tr>
<tr>
<td>Ctrl4</td>
</tr>
<tr>
<td>Ctrl5</td>
</tr>
<tr>
<td>Ctrl6</td>
</tr>
<tr>
<td>Ctrl7</td>
</tr>
<tr>
<td>Ctrl8</td>
</tr>
<tr>
<td>Ctrl9</td>
</tr>
<tr>
<td>Ctrl10</td>
</tr>
<tr>
<td>Ctrl11</td>
</tr>
<tr>
<td>Ctrl12</td>
</tr>
</tbody>
</table>

OR specify:

Gene table of raw intensities for each gene in each experiment.
The intensities need to be tab-separated in a text file and use decimal notation (e.g. 712.5). You can have an annotation field in column 2. This file must not be compressed. Download (shift-click) a sample file of 3 control replicates and 3 experiment replicates from a spotted array of Bacillus subtilis as an example (no header).

Description of file contents (ID=identifier, AN=annotation, A=category A, B=category B, all space separated):

Name experiment columns (space separated):

Is there a header in the first line of the file?

Bonferroni correction for multiple testing:
Max number of genes to plot: 100  Max number of false positives: 10
Select Chip Type HU6800
Select Organism Homo sapiens

Important notes:

- The analysis can take from 20 minutes to 2 hours, depending on the number of chips and the statistical analysis performed. The status of your job (either 'queued' or 'running') will be displayed and constantly updated until it terminates and the server output appears in the browser window.
- At any time during the wait you may enter your e-mail address and simply leave the window. Your job will continue; you will be notified by e-mail when it has terminated. The e-mail message will contain the URL under which the results are stored; they will remain on the server for 24 hours for you to collect them.
- In case of problems with this server, please study any error message carefully, as it will often tell you what the problem is. Otherwise, please check the input and make sure that it conforms with the requirements.

CITATIONS

For publication of results, please cite:

**GenePublisher: Automated Analysis of DNA Microarray Data.**
Knudsen, S., Workman, C., Sicheritz-Ponten, T., and Friis, C.
Submitted 2003.

GETTING HELP

Scientific problems: Steen Knudsen  Technical problems: Kristoffer Rapacki

This file is last modified Thu Apr 3 08:41:10 2003