Gene expression analysis
Last lecture

What are **microarrays**? - Biomolecular devices measuring the transcriptome of a cell of interest.

Workflow of a **microarray experiment** - RNA extraction, cDNA rewriting, labeling, hybridization to microarray, scanning, spot detection, spot intensity to numeric values, normalization, *analysis* (today)

**Normalization** – Assumption, that the vast majority of genes is not differentially expressed between the two classes. Remove technical bias to detect the biological differences.
This lecture

- Differential expression
- Clustering
- Standards in the gene expression data management
- Databases
Differential Expression - Motivation

Why find genes that behave differently in two classes (e.g., normal and tumor)?

Better understanding of the genetic circumstances that cause the difference (disease) hopefully leads to better therapy.

Detection of marker-genes enables the early recognition of diseases as well as the recognition of subtypes of diseases.

Once a cause is identified therapy can become more specific, more effective and reduce side-effects.
We **have**: 

\[ N_1, \ldots, N_m : \text{normale samples} \]
\[ T_1, \ldots, T_n : \text{tumor samples} \]

We **look for**: genes with significant differences between N and T

Compare values of gene X from group N with those of group T

\[ N = \{n_1, \ldots, n_m\} \]
\[ T = \{t_1, \ldots, t_n\} \]

many methods, here:

- Fold change
- t-test
Visualization - Scatterplot

one point = one gene

totally identical distribution

distribution of intensity differences

outlier: interesting genes

A ... Gene expression Patient A
B ... Gene expression Patient B
Fold Change

Definition **Fold Change (FC):**

\[ FC = \log_2 \left( \frac{\text{avg}(T)}{\text{avg}(N)} \right) \]

**Significance** of result is determined by **threshold** fc:

\[ |fc| < 1 \text{ not interesting} \]
\[ 1 < |fc| < 2 \text{ interesting} \]
\[ |fc| > 2 \text{ very interesting} \]

**Why log2 ?**

<table>
<thead>
<tr>
<th></th>
<th>mean(tumor)</th>
<th>mean(normal)</th>
<th>mean(t) / mean(n)</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene x</td>
<td>16</td>
<td>1</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>gene y</td>
<td>0.0624</td>
<td>1</td>
<td>1/16</td>
<td>-4</td>
</tr>
</tbody>
</table>
Fold Change – Advantages / Disadvantages

+ intuitive measure
- independent of scatter

- independent of absolut values

→ score based only on the mean of the groups not optimal, include variance!
Hypothesis testing

Hypothesis

H0 Null hypothesis (the one we want to reject)
H1 Alternative hypothesis (logical opposite of H0)

Test statistic

Function of the sample that summarizes the data set into one value that can be used for hypothesis testing.

Significance level ($\alpha$)

Probability for a false positive outcome of the test, the error of rejecting a null hypothesis when it is actually true

p-Value

Probability of obtaining the observed test-statistic under the assumption, that the null hypothesis holds.
Hypothesis testing example

Perceived Age Change After Aesthetic Facial Surgical Procedures
Quantifying Outcomes of Aging Face Surgery

Nitin Chauhan, MD, FRCSC; Jeremy P. Warner, MD; Peter A. Adamson, MD, FRCSC


ABSTRACT

Objective To quantify the degree of perceived age change after aesthetic facial surgical procedures to provide an objective measure of surgical success.

Methods Sixty patients undergoing various aging face surgical procedures were randomly chosen for analysis. Preoperative and postoperative photographs were evaluated. Raters were presented with photographs in a random assortment and were asked to estimate the age of the patient. Perceived age difference was defined as the difference between the chronological age and the estimated age, and the change in this value after surgery was the chief outcome of interest. Statistical models were designed to account for any effects of interrater differences, preoperative chronological age, rater group, photograph order, or surgical procedure performed.
Hypothesis testing example

- N = 60 patients with estimated mean $\bar{x} = 7.177$; stdev = 2.948
- Unknown: True mean $\mu$; normal distribution is assumed
- H0 : $\mu \leq 0$; H1: $\mu > 0$
- We search probability of observing $\bar{x}$ (or a more extreme $\bar{x}$), given H0 is true

$$P\left(\bar{x} - \mu \geq \frac{\bar{x} - \mu}{\sqrt{sd^2/N}}\right)$$

$$P\left(\bar{x} - \mu \geq \frac{7.177 - 0}{\sqrt{2.948^2/60}}\right) \approx P\left(t_{59} \geq 18.86\right)$$

Furthermore we could estimate the confidence interval for the true mean.
T-test (Welch-test)

**Assumption:** The values are normally distributed (note that for the normal t-test equal variances are assumed)

**Teststatistik:**

\[
t = \frac{\text{avg}(S) - \text{avg}(T)}{\sqrt{\frac{sd(S)^2}{m} + \frac{sd(T)^2}{n}}}
\]

the greater $|t|$, the greater the differential expression of a gene.

From t statistic to p value: t-value and significance level determine the p value (look-up tables)
### Example (Welch-test)

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Significance level</th>
<th>Test statistic</th>
<th>P-Value</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0: \mu_x - \mu_T = 0$</td>
<td>$\alpha = 0.05$</td>
<td>$t = \frac{\text{mean}(N) - \text{mean}(T)}{\sqrt{\frac{s,d,(N)^2}{m} + \frac{s,d,(T)^2}{n}}}$</td>
<td>$p\text{-value} = 0.0126$</td>
<td>7.056</td>
</tr>
<tr>
<td>$H_1: \mu_x - \mu_T \neq 0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$N = \{5, 7, 6, 9, 5\}$

$T = \{2, 4, 3, 5, 3\}$

$$H_0: \mu_x - \mu_T = 0 \quad H_1: \mu_x - \mu_T \neq 0$$

$$\alpha = 0.05$$

$$t = 3.3129$$

$$p\text{-value} = 0.0126$$

$$df = 7.056$$
Example

\[ t = -55.53 \]

\[ t = -7.5 \]

\[ t = 2.37 \]

\[ t = 0.16 \]
Vulcano Plot

- Scatterplot significance versus fold change
  - Y-axis: Negative $\log_{10}$ of the p-value
  - X-axis: Foldchange
Further Methods

ANOVÀ – comparing more than one group as well as different factors.

SAM – Significance analysis of Microarrays. An 'improvement' of the t-test, as small variances can lead to very significant results without a considerable fold change.

Rank Product – sort genes by expression and determine geometric mean of rank.
Multiple Testing Correction

**Problem:** Microarrays contain up to 20,000 genes, thus an $\alpha = 0.05$ leads to approximately $20,000 \times 0.05 = 1000$ FPs.

**Solution:** Multiple testing correction. Two basic approaches:

1. **Family wise error rate (FWER)**, the probability of having at least one false positive in the set of results considered as significant.
2. **False discovery rate (FDR)**, the expected proportion of true null hypotheses rejected in the total number of rejections. (FDR measures the expected proportion of incorrectly rejected null hypotheses, i.e. type I errors).
Bonferoni (FWER)

Let $N$ be the number of genes tested and $p$ the p-value of a given probe, one computes an adjusted p-value using:

$$p_{\text{adjusted}} = p \times N$$

Only if the adjusted p-value is smaller than the pre-chosen significance value, the probe is considered differentially expressed.

**Very conservative** (many wrong rejections of $H_1$), rarely used

Bonferoni assumes independence between the tests (usually wrong)

Appropriate when a **single false positive** in a set of tests would be a problem (e.g., drug development)
Benjamini – Hochberg (FDR)

1. choose a specific $\alpha$ (e.g. $\alpha=0.05$)

2. rank all $m$ p-values from smallest to largest

3. correct all p-values: $\text{BH}(p_i)_{i=1,\ldots,m} = p_i \times \frac{m}{i}$

4. $\text{BH}(p) = \text{significant if } \text{BH}(p) \leq \alpha$

<table>
<thead>
<tr>
<th>Genes</th>
<th>p-value</th>
<th>rank</th>
<th>BH(p)</th>
<th>Significant? ((\alpha=0.05))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>0.00001</td>
<td>1</td>
<td>$1000/1 \times 0.00001=0.01$</td>
<td>yes</td>
</tr>
<tr>
<td>Gene B</td>
<td>0.0004</td>
<td>2</td>
<td>$1000/2 \times 0.0004=0.20$</td>
<td>no</td>
</tr>
<tr>
<td>Gene C</td>
<td>0.01</td>
<td>3</td>
<td>$1000/3 \times 0.01=3.3 \rightarrow 1.0$</td>
<td>no</td>
</tr>
</tbody>
</table>
Clustering - Motivation

High dimensional data possibly containing all kinds of patterns and behavior of subgroups which might represent biomedical phenomena. (explorative)

Clustering for quality control.

Expression patterns similar in spacial and temporal behavior → co-regulated / expressed genes (e.g. genes controlled by the same transcriptionfactor).

Discover new disease subtypes by clustering samples.
Clustering

**UNSUPERVISED LEARNING**
- Dataset
- Unknown Classes
- Cluster Samples
- Assign Class Labels
- Class Discovery

**SUPERVISED LEARNING**
- Known Classes
  - Class A
  - Class B
- Training Set
- Train Model
- Independent Test Set ("Unknowns")
- Apply Model
- Class Prediction

Ramaswamy & Golub 2002
Clustering

• Goal:
  • Partition observations into groups (clusters)
  • Pairwise similarities between elements assigned to the same cluster tend to be higher than to elements in different cluster

• Requires:
  • Similarity measure

• Advantages:
  • Intuitive and well known
  • Simple procedure
Clustering - Example
Hierarchical Clustering - Algorithm

1. choose a distance measure (e.g., Euclidean, Pearson, etc.)
2. compute similarity matrix S
3. while |S|>1:
   4. determine pair (X,Y) with minimal distance
   5. compute new value $Z = \text{avg} (X,Y)$, (single, average, or complete linkage)
   6. delete X and Y in S, insert Z in S
   7. compute new distances of Z to all elements in S
   8. visualize X and Y as pair
Hierarchical Clustering - graphical

\begin{align*}
(A, b) &\rightarrow c \\
(C, G) &\rightarrow d \\
(d, c) &\rightarrow e \\
(a, e) &\rightarrow f
\end{align*}
Hierarchical Clustering – real data
Hierarchical Clustering

Result: binary tree, clusters have to be determined by the user.

Cutting the dendrogram at a particular height partitions the data into disjoint clusters

For a easier determination of clusters: length of branch is set in relation to the difference of the leafs.
K means

K-means partitions the n observations into k clusters and tries to minimize the distance of the n data points from their respective cluster centres.

1. choose $k$ random cluster centers $\mu_1, \ldots, \mu_k$
2. Assign for each point $x$ in dataset $S$ the closest cluster center
3. compute a new center $\mu_i$ for every cluster $C_i$
4. repeat 2-3. until cluster centers do not change
K means

http://www.itee.uq.edu.au/~comp4702/lectures/k-means_bis_1.jpg
K means

Convergence is not assured.

Cluster quality can be computed by determining the mean distance of a gene to its cluster center.

Number of clusters has to be chosen in advance.

The initialization of the cluster centers has a great impact on the clustering quality, compute more than one initial constellation.
Standards

To determine the comparability of different experiments detailed information on the different steps is necessary.

RNA extraction,
cDNA rewriting,
labeling,
hybridization to microarray,
scanning,
spot detection,
spot intensity to numeric values,
normalization
MIAME describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.

MIAME does **not** specify a particular **format** (→ use **MAGE-TAB** or **MAGE-ML**)

MIAME does **not** specify any particular **terminology** (use **MGED-ontology**)

MIAME Specification

1. **raw data** (.CEL, .gpr)

2. final **processed** (normalized) **data**

3. **sample annotation** (incl. Experimental factors and their values, scan protocol, groove protocol)

4. **experimental design** including sample data relationships (e.g., overall design; technical or biological replicates)

5. **annotation** of the **array** (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences)

6. **laboratory** and **data processing protocols** (e.g., what normalisation method)
Standards - Overview

![Diagram showing relationships between standards]

- MIAME (formal version)
- MAGE-OM (derives from)
- MAGE-ML (simpler, tab-delimited)
- MAGE-TAB
- MGED-Ontology

---

Bioinformatics, Summer Semester 2013
## Standards - Overview

<table>
<thead>
<tr>
<th>Minimum Information Specification</th>
<th>DNA Microarray Data</th>
<th>High-throughput Sequencing Data</th>
<th>In Situ Hybridization and Immunohistochemistry Data</th>
<th>Tissue Microarray Data</th>
<th>Proteomics Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIA M E</td>
<td>MIA M E</td>
<td>MIA M E</td>
<td>MIA M E</td>
<td>MIA M E</td>
<td>MIA M E</td>
</tr>
<tr>
<td>Data Model</td>
<td>MAGE-OM</td>
<td>?</td>
<td>?</td>
<td>TMA-OM</td>
<td>PSI-OM</td>
</tr>
<tr>
<td>XML Format</td>
<td>MAGE-ML</td>
<td>?</td>
<td>?</td>
<td>TMA-DES</td>
<td>PSI-ML</td>
</tr>
<tr>
<td>TAB-del. format</td>
<td>MAGE-TAB</td>
<td>?</td>
<td>?</td>
<td>TMA-TAB</td>
<td>?</td>
</tr>
<tr>
<td>Controlled vocabulary</td>
<td>MGED-ontology</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Databases

GEO (Gene Expression Omnibus)
Array Express
GEO – Gene Expression Omnibus

NCBI public repository
RDBMS schema

**GPL**
(GEO platform)
platform description

**GSM**
(GEO sample)
raw-processed intensities from a
single or chip

**GSE**
(GEO series)
grouping of chip data,
a single experiment

**GDS**
(GEO dataset)
grouping of experiments

submitted by
manufacturer

submitted by
experimentalist

curated by
NCBI
GEO

**Gene Expression Omnibus**: a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles. More information →

---

**GEO navigation**

- **DataSets**
- **Gene profiles**
- **GEO accession**
- **GEO BLAST**

**Query**

- **DataSets**
- **Gene profiles**
- **GEO accession**
- **GEO BLAST**

**Browse**

- **DataSets**
- **Platforms**
- **Samples**
- **Series**

**Submitter login**

- **Login**
- **New account**
- **Recover account**

**Site contents**

- **Public data**
  - Platforms: 11,608
  - Samples: 938,021
  - Series: 38,704
  - DataSets: 3,341

- **Documentation**
  - Overview
  - FAQ
  - Find
  - Submission guide
  - Linking & citing
  - Journal citations
  - Construct a Query
  - Programmatic access
  - DataSet clusters
  - GEO announce list
  - Data disclaimer
  - GEO staff

- **Query & Browse**
  - Repository browser
  - GEO2R
  - FTP site
  - GEO Profiles
  - GEO DataSets

- **Submit**
  - New account
ArrayExpress (EMBL-EBI)

The ArrayExpress Archive is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. Gene Expression Atlas contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive
16266 experiments, 458692 assays
Experiment, citation, sample and factor annotations

Gene Expression Atlas
5661 experiments, 138577 assays, 18419 conditions
Genes Conditions

News
- 15 Nov 2010 - New citation for ArrayExpress
- 20 Oct 2010 - Internship for a student project in human gene expression - Filled now
  This student project is now taken.

Links
- ArrayExpress User Survey
- Old ArrayExpress Interface
- Help | Training | FAQ | Citing
- Submit Data (array based and re-sequencing)
- Programmatic Access | FTP Access
- Software Downloads and Statistics
- EFO | Bioconductor Package | Quality Metrics
- ArrayExpress Scientific Advisory Board
- Functional Genomics Group
GEO vs. ArrayExpress

- both encompass MIAME compliance
- both provide a good possibility for making data public as often requested by journals
- GEO contains more data
- ArrayExpress provides analysis tools
Diffuse large B-cell lymphoma – Subtypes

germinal center B-cell-like (GCB), activated B-cell-like (ABC) with 5-year survival rates of 59% and 30%

Wright 2003
DLBCL Subtypes

Wright 2003
DLBCL Subtypes

40 Exon arrays of DLBCL patients, subtype unknown. Do we see the division in subgroups with a different technology and different probes?
DLBCL Subtypes

Schematic representation of how gene expression results can be compared across microarray platforms.

Statistical model based on LPS values from tissue specific cDNA arrays

\[ LPS = \sum_{j=1}^{14} a_j X_j = a_1 X_1 + a_2 X_2 + \cdots + a_{14} X_{14} \]

Select genes most capable of discriminating between ABC and GCB DLBCL

Ferl G Z et al. PNAS 2003;100:10585-10587
DLBCL Subtypes
Summary

Combine t-test and fold change for optimal detection of differential expression.

More explorative analysis like clustering can detect patterns inherent in the expression data like co-regulated genes or new disease subtypes.

Public repositories like GEO and ArrayExpress offer a rich fundus of data.