Biostatistics

Grundlagen der Bioinformatik SS2019

$t = \frac{\text{variance between groups}}{\text{variance within groups}}$

- A big $t$-value = different groups
- A small $t$-value = similar groups

https://www.youtube.com/watch?v=0Pd3dc1GcHc
Agenda

- Normalization
- Differential expression
  - Fold Change
  - P-value
  - t-test
- Clustering
Experimental Design

\[ N_1, \ldots, N_m : \text{control samples} \]

\[ T_1, \ldots, T_n : \text{case samples} \]

We look for:

Genes with significant differences between \( N \) and \( T \)

Compare gene \( X \) from group \( N \) with gene \( X \) of group \( T \)

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

Many methods exist, here: Fold change t-test
Motivation normalization

- Interested in: true biological difference of mRNA expression
- What we measure: Mixture of (unwanted) technical and biological noise
- Correct undesired noise!
Quantile normalization

1. Matrix $X$
   a. Columns = samples
   b. Row = transcripts

2. Sort each column of $X \rightarrow X_{\text{sort}}$

3. Calculate row-means and store in $X'_{\text{sort}}$

4. Obtain $X_n$ by rearranging columns of $X'_{\text{sort}}$ to have the same ordering as the corresponding input vector

- Differences between the separate values retained
- Identical distribution for each array
- Information lost especially in the lower signals
What we want: comparability

Important: normalization between samples, not within one sample

Figure 7A. Ratio Intensity Plot of all probes for four pairs of chips from GeneLogic spike-in experiment

Figure 7B. As in A, after normalization by matching quantiles. Both figures courtesy of Terry Speed

### Example quantile normalization

<table>
<thead>
<tr>
<th>Indexes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1 11 13 29 26</td>
</tr>
<tr>
<td>V2</td>
<td>15 17 5 8 14</td>
</tr>
<tr>
<td>V3</td>
<td>21 2 12 20 25</td>
</tr>
<tr>
<td>V4</td>
<td>10 19 16 24 4</td>
</tr>
<tr>
<td>V5</td>
<td>18 28 3 22 27</td>
</tr>
</tbody>
</table>

#### Sort

<table>
<thead>
<tr>
<th>Indexes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1 11 13 29 26</td>
</tr>
<tr>
<td>V2</td>
<td>15 17 5 8 14</td>
</tr>
<tr>
<td>V3</td>
<td>21 2 12 20 25</td>
</tr>
<tr>
<td>V4</td>
<td>10 19 16 24 4</td>
</tr>
<tr>
<td>V5</td>
<td>18 28 3 22 27</td>
</tr>
</tbody>
</table>

#### Replace

<table>
<thead>
<tr>
<th>Indexes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>3 8 19 28 23</td>
</tr>
<tr>
<td>V2</td>
<td>19 14 8 8 14</td>
</tr>
<tr>
<td>V3</td>
<td>28 3 14 14 19</td>
</tr>
<tr>
<td>V4</td>
<td>14 19 23 23 3</td>
</tr>
<tr>
<td>V5</td>
<td>23 28 3 19 28</td>
</tr>
</tbody>
</table>

#### Reorder

<table>
<thead>
<tr>
<th>Indexes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>3 5 6 1 5</td>
</tr>
<tr>
<td>V2</td>
<td>5 6 4 4 1</td>
</tr>
<tr>
<td>V3</td>
<td>2 4 1 5 3</td>
</tr>
<tr>
<td>V4</td>
<td>4 2 3 3 2</td>
</tr>
<tr>
<td>V5</td>
<td>6 1 2 2 6</td>
</tr>
</tbody>
</table>

### Explanation
- **Sort**: Values are sorted in ascending order.
- **Replace**: Values are replaced with their corresponding sorted values.
- **Reorder**: Rows are reordered based on the sorted values.
Differential expression
### Fold Change

*Fold Change*:

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

**Thresholds (examples)**

- \(|FC| < 1\) not interesting
- \(|FC| > 2\) interesting

<table>
<thead>
<tr>
<th>Genes</th>
<th>Mean Case</th>
<th>Mean Control</th>
<th>Mean Case / Control</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>16</td>
<td>1</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>0.0625</td>
<td>1</td>
<td>0.0625</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>200</td>
<td>1</td>
<td>200</td>
<td>7.65</td>
</tr>
</tbody>
</table>
Z-score normalization

- Correct for different amount of mRNA per sample
- Z-score = scaling of counts
  - $0 = \text{average}$
- Examples: 2, -1, 0.1

$$Z = \frac{X_i - \text{mean}_\text{est}}{\text{sd}_\text{est}}$$

$X_i =$ expression gene $i$

$\text{Mean}_\text{est} : \text{(estimated) expr. average over all genes}$

$\text{Sd} : \text{(estimated) expr. standard deviation of all genes}$
Fold Change - Advantages / Disadvantages

✓ intuitive measure

✗ Independent of scatter

✗ Independent of absolute values
  ○ Score only based on mean of groups
  ○ Spread of data points essential
Variance essential

- High abs(FC) for Gene A and Gene B
- But: variance very high in the tumor samples of Gene B
- Find test for FC and variance

\[
Var(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2
\]
Hypothesis Testing 1

- Same Mean
  - Different variance
- Measure 'uncertainty' with standard deviation $sd$
- Combine both to likelihood for 'correctness'
- Assumption
  - Log-Normal distributions
  - Symmetric
  - Independent

\[
\sigma_X := \sqrt{\text{Var}(X)}
\]

\[
\text{Var}(X) = E((X - E(X))^2) = E(X^2) - (E(X))^2
\]
Hypothesis Testing 2

- **T-test (unpaired two-sample)**
  - Compares the mean of two unpaired samples
- **Assumption**
  - Values normally distributed
  - Equal variances
- **Hypothesis**
  - $H_0$ (Null hypothesis): $m_1 = m_2$ vs. $m_1 \neq m_2$ (means are not equal)
- **Test statistic**
  - Function of the sample that summarizes the data set into one value that can be used for hypothesis testing
Hypothesis Testing 3

From T-statistic to p-value

- T-value, α and number of samples determine the p-value (look-up tables)

P-value

- Probability of observing your data under the assumption that $H_0$ is true
- Probability that you will be in error if rejecting $H_0$

Significance level (α)

- Probability of a false positive outcome of the test, the error of rejecting $H_0$ when it is actually true

If $|t| > |T|$ we reject $H_0$

→ p-value is significant
  (p-value $< α$)
Hypothesis Testing - Workflow

1. Determine null and alternative hypothesis
2. Select a significance level (alpha)
3. Take a random sample from the population of interest
4. Calculate a test statistic from the sample that provides information about the null hypothesis
5. Decision
Examples

Degrees of freedom: |Samples| - 2, Here 16 - 2 = 14
Example

Hypothesis: $H_0: m_N - m_T = 0$ vs $H_1: m_N - m_T \neq 0$

Significance level: 0.05

Test statistic

\[
t = \frac{X_1 - X_2}{S_p \cdot \sqrt{\frac{1}{n_1} \cdot \frac{1}{n_2}}}
\]

P-value: 0.06

$->$ Not significant

Data from slide 9

$N = \{3.58, 4.14, 3.49, 3.37, 5.29, 5.06, 3.6\}$

$T = \{3.7, 10.9, 10.3, 3.57, 10.5, 8.18, 3.27\}$

Critical value = 2.45

$\frac{-2.27}{-2.45}$
Volcano plot

Combine P-value and Log-FC

- Y-axis: Negative log10 of the p-value
- X-axis: Fold-change

Interested in

- Upper left
- Upper right corner

http://www.hindawi.com/journals/crp/2011/532915.fig.001.jpg
Multiple Testing Correction

Problem
Microarrays has 22k genes, thus an $\alpha=0.05$ leads to approximately $22\,000 \times 0.05 \sim 1100$ 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)
Bonferroni correction

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

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

- Iff the adjusted $p$-value is smaller than the $\alpha$, the probe is considered differentially expressed.
- Bonferroni 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)

E.g. case of two $p$-values (multiply by 2)

1. 0.001 $\rightarrow$ 0.002
2. 0.03 $\rightarrow$ 0.06
Benjamini - Hochberg correction

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: \( BH(p)_{i=1,\ldots,m} = p_i \times \frac{m}{i} \)

4. \( BH(p) \) = significant if \( BH(p) \leq \alpha \)

<table>
<thead>
<tr>
<th>Genes</th>
<th>p-value</th>
<th>rank</th>
<th>BH(p)</th>
<th>Significant 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00001</td>
<td>1</td>
<td>0.00001\times1000/1 = 0.01</td>
<td>yes</td>
</tr>
<tr>
<td>B</td>
<td>0.0004</td>
<td>2</td>
<td>0.0004\times1000/2 = 0.20</td>
<td>no</td>
</tr>
<tr>
<td>C</td>
<td>0.01</td>
<td>3</td>
<td>0.01\times1000/3 = 3.3 -&gt; 1.0</td>
<td>no</td>
</tr>
</tbody>
</table>
Clustering - Motivation

Subgroups detection

Quality control

Similarity-detection in spatial and temporal behavior

- Co-regulated / expressed genes
  - E.g. genes controlled by the same transcription-factor

Discovery of new disease subtypes
Overview unsupervised clustering
Clustering

Ramaswamy & Golub 2002
Clustering

- **Goal**
  - Partitioning Biological interpretation of subtypes (clusters)
- **Requires**
  - (Useful) similarity measure
- **Advantages**
  - Intuitive Simple (you would think)

![cetuximab response in different subtypes of HNSCC](chart.png)
Hierarchical Clustering - algorithm

1. Distance measure
   a. Euclidean
   b. Pearson, etc.

2. Compute similarity matrix S

3. While |S|>1:
   a. Determine pair (X,Y) with minimal distance
   b. Compute new value Z = avg (X,Y), (single, average, or complete linkage)
   c. Delete X and Y in S, insert Z in S
   d. Compute new distances of Z to all elements in S
   e. Visualize X and Y as pair
Hierarchical Clustering

- Binary tree
- Cutting the dendrogram at a particular height partitions the data into disjoint clusters
- For an easier determination of clusters
  - Length of branch is set in relation to the difference of the leafs.

Linkage Rule essential
Hierarchical Clustering – Linkage

- Methods produce similar results for data with strong clustering tendency
  - (each cluster is compact and separated)

- **Single Linkage**
  - Single smallest distance
    \[ D(X,Y) = \min_{x \in X, y \in Y} d_{xy} \]
  - Violates the compactness property (i.e., observations inside the same cluster should tend to be similar)

- **Complete Linkage**
  - Most distant elements
    \[ D(X,Y) = \max_{x \in X, y \in Y} d_{xy} \]

- **Average Linkage**
  - Compromise
    \[ D(X,Y) = \frac{1}{N_X N_Y} \sum_{x \in X} \sum_{y \in Y} d_{xy} \]
Hierarchical Clustering