Gene expression 3

ROC curves
Multiple testing
Gene expression networks
Hierarchical clustering results

A

coronary artery 2a
coronary artery 2b
pulmonary artery 1
pulmonary artery 2
coronary artery 1
coronary artery 3
coronary artery 4
Aorta #3
Aorta #1
Aorta #2
Umbilical artery #4
Umbilical artery #3
Umbilical artery #5
Umbilical artery #2
Umbilical artery #7
Iliac artery #3
Iliac artery #2
Iliac artery #5
Iliac artery #6
Iliac artery #1
Nasal, Micro #2
Nasal, Micro #1
Nasal, Micro #3
Bladder, Micro #1
Bladder, Micro #2
Myocardium #1
Myocardium #2
Umbilical Vein #1
Umbilical Vein #5
Umbilical Vein #2
Umbilical Vein #6
Umbilical Vein #3
Umbilical Vein #4
Saphenous Vein #2
Saphenous Vein #1
Uterus, Micro #2
Uterus, Micro #3
Uterus, Micro #1
Skin, Micro #2 (Abdomen)
Skin, Micro #3 (Breast)
Skin, Micro #6 (Abdomen)
Intestinal, Micro #1
Intestinal, Micro #2
Skin, Micro #5 (Abdomen)
Skin, Micro #6 (Abdomen)
Lung, Micro, #4
Skin, Micro #3 (Abdomen)
Skin, Micro #1 (Breast)
Skin, Micro #7 (Abdomen)
Skin, Neontal (Foreskin)
Lung, Micro #3
Lung, Micro #2
Lung, Micro #1

B

Artery
Tissue (II)
Vein
Tissue (I)

Large Vessel
Microvascular

Chi et al., PNAS | September 16, 2003 | vol. 100 | no. 19 | 10623-10628

“Endothelial cell diversity revealed by global expression profiling”
Receiver operating characteristic

• A framework to compare the performance of binary classifiers
• Plot of false positive rate (sensitivity) vs true positive rate (1-specificity)
Gütemaße

• Sensitivität/Recall
  TPR = TP/P = TP/(TP+FN)

• Spezifizität
  FPR = FP/N = FP/(FP+TN)

• Precision (positive predictive value)
  PPV = TP/(TP+FP)

• False discovery rate
  FDR = FP/(TP+FP)
ROC curves with data set 'falMP'
AUC is 0.653 for 'pnc' with rank p-value 3.40e-016
Fisher's exact test p-value for 1st 250 genes is 4.41e-016
Hypothesis Testing

• H0 : Null hypothesis vs. H1 : Alternative Hypothesis

• T : test statistics C : critical value

• If |T|>C, H0 is rejected. Otherwise H0 is retained

• Example
  H0 : \( \mu_1 = \mu_2 \) vs. H1 : \( \mu_1 \neq \mu_2 \)
  \( T = (x_1 - x_2) \) / pooled standard error (se)
  If \(|T| > z(1 - \alpha/2)\), H0 is rejected at the significance level \( \alpha \)

• \( C_\alpha \)
## Hypothesis Testing

<table>
<thead>
<tr>
<th>Truth</th>
<th>Hypothesis Result</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Retained</td>
</tr>
<tr>
<td>H0</td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

- **Type I error rate** = false positives \((\alpha : \text{significance level})\)
- **Type II error rate** = false negatives
- **Power** : 1–Type II error rate

- **P-values** : \(p=\inf\{\alpha \mid \text{H0 is rejected at the significance level } \alpha \} \)
Issues in Multiple Comparison

• Given n treatments, which two treatments are significantly different? (simultaneous testing)
• Is treatment A different from treatment B?

• m treatment means: $\mu_1, \ldots, \mu_n$
  
  \[ H_j : \mu_i = \mu_j \text{ where } i \neq j \quad T_j = (x_i - x_j) / \text{pooled SE} \]

• Type I error when testing each at 0.05 significance level one by one: $1 - (0.95)^n$
• Inflated Type I error, ex) $\alpha = 1 - (0.95)^{10} = 0.401263$

• Remedies: Bonferroni Method
  
  Type I error rate = $\alpha / \# \text{ of comparison}$
**Type I Error Rates**

<table>
<thead>
<tr>
<th>Truth</th>
<th>H0</th>
<th>#retained</th>
<th>#rejected</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U</td>
<td>V</td>
<td></td>
<td>m0</td>
</tr>
<tr>
<td>H1</td>
<td>T</td>
<td>S</td>
<td></td>
<td>m1</td>
</tr>
<tr>
<td>Total</td>
<td>m-R</td>
<td>R</td>
<td></td>
<td>m</td>
</tr>
</tbody>
</table>

- Per-comparison error rate (PCER) = \( E(V) / m \)
- Per-family error rate (PFER) = \( E(V) \)
- Family-wise error rate = \( \text{pr} ( V \geq 1 ) \)
- False discovery rate (FDR) = \( E(Q) \), \( Q = V/R \), if \( R > 0 \)
  \( 0, \) if \( R = 0 \)
Type I Error Rates

Under the complete null hypothesis, each $H_j$ has Type I error rate $\alpha_j$.

- **PCER** = $E(V) / m = (\alpha_1 + ... + \alpha_m) / m$
- **PFER** = $E(V) = \alpha_1 + ... + \alpha_m$
- **FWER** = $P (V \geq 1 ) = 1 - Pr (H_j, j=1, ..., m, not rejected )$
- **FDR** = $E(V / R) = FWER$

PCER = $(\alpha_1 + ... + \alpha_m) / m \leq \max (\alpha_1 + ... + \alpha_m)$

$\alpha_1 + ... + \alpha_m \leq \text{PWER} = \text{FDR} \leq \text{PFER} = \alpha_1 + ... + \alpha_m$
Types of comparisons

• Assume $H_j, j=1, \ldots, m$, with their test statistics $T_j, j=1,\ldots, m$, which has a MN with mean $\mu=(\mu_1,\ldots,\mu_m)$ and identity covariance vector.

• Let $R_j = 1$ (Hj is rejected) and $r_j$ is observed value of $R_j$.

• Let $\gamma_j = \Pr (\text{Hj rejected under Hj})$.

• $\text{PFER} = \sum_{j=1}^{m} \gamma_j$ (Per family error rate)
• $\text{PCER} = \sum_{j=1}^{m} \gamma_j / m$ (Per comparison error rate)
• $\text{FWER} = 1 - \prod_{j=1}^{m} (1 - \gamma_j)$ (Family wise error rate)
• $\text{FDR} = \sum_{\sum r_1=01 \ldots \sum r_1=01} (\sum_{j=1}^{m} 0 r_j / \sum_{j=1}^{m} r_j) \prod \gamma_j r_j (1 - \gamma_j) 1 - r_j$ (False discovery rate)
Networks
Considerations for the analysis

- Directed vs undirected graphs
- Analysis of confounding factors
- How to assign weights?
  - Repetitions in screen
  - Outgoing and incoming edges
  - External data
- Hubs
Topological analysis

• Small worlds
  – Shortest path lengths are small
  – Degrees of separation

• Modular
  – Clustering co-efficient

\[
C_i = \frac{2E_i}{k_i(k_i - 1)}
\]

• Degree distribution
  – Random model
  \[
P(\text{deg}(v) = k) = \binom{n-1}{k}p^k(1-p)^{n-1-k}.
\]
  – Poisson with max = \(P(<k>)\)
  – Scale free
    • \(P(k) \sim k^{-r}\)
    • \(-1.5 > r > -3\)
Different networks
From Barabási (2004), Nature Reviews Genetics
Connections between hubs

Maslov and Sneppen (2002) Science

Hubs are connected to proteins of low degree, not between each other
Motifs and subgraphs
Motifs in real networks
Biological types of interactions

A proposed ontology for interactions (Lu et al.)
• Unweighted graphs
  – Hamming distance

• Weighted graphs
  – Euclidean distance
  – Correlation
    • Pearson
    • Spearman

• Boolean networks
• Probabilistic networks
  – Markov Random Fields
  – Bayesian networks