Microarray (DNA) data analysis

PRESENTED BY IVAN SLOBOZHAN

10/04/2017
Quick reminder

**Idea:**

1) Put one cell in conditions A
2) Put another cell in conditions B
3) Create DNA microarray
4) Analyze it!
Steps of image processing:

- Identification of the spots and distinguishing them from spurious signals
- Determination of the spot area to be surveyed
- Reporting summary statistics and assigning spot intensity after subtracting for background intensity.
Identification of the spots and distinguishing them from spurious signals

Idea:

1) Divide microarray into sub-arrays.
Determination of the spot area to be surveyed

Two methods:

1) Use area of a fixed size that is centered of the center of the mass of spot.

2) Precisely define the boundary for a spot and include pixels within this boundary.

Average spot ~ 314 pixels.
Reporting summary statistics and assigning spot intensity after subtracting for background intensity.

Spot median value, with the background median value subtracted from it, as the metric to represent spot intensity.

Get mean, median and other statistics

Luminance- \((0.2126*R + 0.7152*G + 0.0722*B)\)

Expression ratios

\[ T_k = \frac{R_k}{G_k} \] - relative expression level

where \( k \) – \( k^{th} \) gene of the array. \( R_k \) - spot intensity metric for the test sample. \( G_k \) - spot intensity metric for the reference sample.

\[ T_{\text{median}} = \frac{R_k^{\text{spot}} - R_k^{\text{background}}}{G_k^{\text{spot}} - G_k^{\text{background}}} \]
Problem: if genes that should not change in the two conditions, often have an average expression ratio which deviates from 1.

Why: differential labelling efficiency of the two fluorescent dyes or different amounts of starting mRNA material in the two samples.
Data normalization

- Choose a gene-set.

(a set which consists of genes for which expression levels should not change under the conditions studied, that is the expression ratio for all genes in the gene-set is expected to be 1)

- Calculate normalization factor:

\[
N_{total} = \frac{\sum_{k=1}^{N_{gene-set}} R_k}{\sum_{k=1}^{N_{gene-set}} G_k}
\]

- Normalize ratio:

\[
T'_k = \frac{R'_k}{G'_k} = \frac{R_k}{G_k \times N_{total}} = \frac{T_k}{N_{total}}
\]
Data normalization

Gene expression data before and after the normalization procedure
Analysis of gene expression data

- **Aim:** to monitor the expression level of genes and get patterns.

### Gene expression matrices

**Absolute measurement**

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>10</td>
<td>80</td>
<td>40</td>
<td>20</td>
</tr>
<tr>
<td>Gene B</td>
<td>100</td>
<td>200</td>
<td>400</td>
<td>200</td>
</tr>
<tr>
<td>Gene C</td>
<td>30</td>
<td>240</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Gene D</td>
<td>20</td>
<td>160</td>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>

**Relative measurement**

<table>
<thead>
<tr>
<th></th>
<th>C1/C4</th>
<th>C2/C4</th>
<th>C3/C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene A</td>
<td>0.50</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Gene B</td>
<td>0.50</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Gene C</td>
<td>0.50</td>
<td>4.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Gene D</td>
<td>0.25</td>
<td>2.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Distance measures

- Analysis of gene expression data is primarily based on comparison of gene expression profiles or sample expression profiles.
- In order to compare expression profiles, we need a measure to quantify how similar or dissimilar are the objects that are being considered.

Frequently used measures:
- Euclidian
- Pearson correlation coefficient
- Rank correlation coefficient
- Etc.
Aim: cluster genes or samples with similar expression profiles together, to make meaningful biological inference about the set of genes or samples.
Applications

- **Predicting binding sites.**

(DNA binding sites are a type of binding site found in DNA where other molecules may bind. In a binding site is a region on a protein or piece of DNA or RNA to which (specific molecules and/or ions) may form a chemical bond. A chemical bond is a lasting attraction between atoms that enables the formation of chemical compounds.)

- **Predicting protein interactions and protein functions**

- **Predicting functionally conserved modules**

(Genes that have similar expression profiles often have related functions)
List of existing software

<table>
<thead>
<tr>
<th>1</th>
<th>Express Yourself - An automated, online microarray data processing platform, where you can upload image files and carry out data processing and data analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Expression Profiler - A set of tools for clustering, analysis and visualization of gene expression and other genomic data. Tools in the Expression Profiler allow to perform cluster analysis, pattern discovery, pattern visualization, study and search Gene Ontology categories, generate sequence logos, extract regulatory sequences, study protein interactions, as well as to link analysis results to external databases.</td>
</tr>
<tr>
<td>3</td>
<td>Cluster &amp; Treeview - Cluster performs a variety of types of cluster analysis and other types of processing on large microarray datasets. Currently includes hierarchical clustering, self-organizing maps (SOMs). K-means clustering, principal component analysis Treeview can be used to graphically browse results of clustering and other analyses from Cluster.</td>
</tr>
<tr>
<td>4</td>
<td>Xcluster - cross platform software for analysing microarray data.</td>
</tr>
<tr>
<td>5</td>
<td>J-Express - A Java implementation of hierarchical clustering, self organized maps, and principal component analysis, with several different viewing options and output formats.</td>
</tr>
<tr>
<td>6</td>
<td>TM4 - A package of Open Source software programs for microarray analysis.</td>
</tr>
<tr>
<td>7</td>
<td>GeneXPress - A visualization and analysis tool for gene expression data, integrating clustering, gene annotation, and sequence information.</td>
</tr>
<tr>
<td>8</td>
<td>GEPAS - Gene Expression Pattern Analysis Suite.</td>
</tr>
<tr>
<td>9</td>
<td>GenMAPP - A computer application designed to visualize gene expression data on maps representing biological pathways, and other biologically meaningful groups of genes.</td>
</tr>
<tr>
<td>10</td>
<td>OligoArray - An application which computes gene specific oligonucleotides for genome-scale oligonucleotide microarray construction.</td>
</tr>
</tbody>
</table>
References

- https://en.wikipedia.org/wiki/Binding_site
Thank you for attention!