Interactive Analysis of Micro-array Data Using Spectral Maps

Rudi Verbeeck
Johnson & Johnson
Pharmaceutical Research & Development
Overview

• Micro-array analysis in an industrial environment: analysis pipeline
• Spectral map analysis (SMA)
  – Method
  – Interpretation
• Example: corticotropin releasing factor in AtT-20 cells
• Conclusions
Business case

- Application area
  end user application
- User audience
  Scientists analyzing results from micro-array experiments, bioinformaticians
- Business driver
  Provide the scientist with a single view on all the data in his/her micro-array experiment to
  - do a QC check of the samples
  - explore genes specific to a (set of) samples
- ROI
  - Reduce time by pre-calculating the data
  - Improve quality by reducing error risk (by automated analysis)
  - Reduce cost by using a standard analysis environment
Vision for data analysis in drug discovery

• The researcher is the subject matter expert. They know their data best.
• Therefore bring the analysis to the researcher instead of the research data to the analyst.
Micro-array technology

DNA Micro-arrays allow us to quantify the activity of thousands of genes.
Micro-array analysis pipeline

LIMS, lab workflow

Retrieve raw data

QC: technical report
- Present/absent calls
- Background intensity
- Housekeeping controls
- and + controls

QC: filter all present

QC: normalize data

QC: correlation

QC: QC: technical report

AN: spectral map

AN: factor loadings

AN: interactive expl.
Spectral map analysis (SMA)

• SMA is a projection method for exploratory data analysis
• SMA may reveal groups or outliers in the samples and identify genes correlated to these samples
• SMA is defined as a biplot which results from PCA of a log transformed, double centered data table (Lewi 1976)
Algorithm

Original expression levels
- $i = 1 \ldots n$ genes
- $j = 1 \ldots p$ samples

Log transformation
- Appropriate for data measured on ratio scales
- Multiplicatative space: distances become ratios
- Corrects for positive skewness

$y_{ij} = \log(x_{ij})$

Double centering
- Geometrically results in a projection on a hyperplane orthogonal to the line of identity
- Number of dimensions reduced by one
- Removes size component from the data

$z_{ij} = y_{ij} - \sum_{i=1}^{n} w_i^n y_{ij} - \sum_{j=1}^{p} w_j^p y_{ij} + \sum_{i=1}^{n} \sum_{j=1}^{p} w_i^n w_j^p y_{ij}$
Projection

\[ \sqrt{w_i^n w_j^p} z_{ij} = \sum_{k=1}^{r} u_{ik} v_{jk} \lambda_k \]

- r: rank of Z
- \( \alpha, \beta \): factor scaling coeff.
- \( \alpha, \beta = 1 \): row / column norms are preserved between Z and factor space
- \( \alpha, \beta = 0 \): factor scores / loadings are scaled to unit variance
- \( \alpha + \beta = 1 \): perpendicular projection of rows upon unipolar axis through columns (or vice versa) reproduces the values of Z

\[ s_{ik} = \frac{u_{ik} \lambda_k^\alpha}{\sqrt{w_i^n}} \quad l_{jk} = \frac{v_{jk} \lambda_k^\beta}{\sqrt{w_j^p}} \]
Spectral map interpretation

- A gene is attracted by a sample for which it is highly specific
- Genes that possess little contrast are displayed near the center of the map
- Distances are proportional to RMS contrast (small distances indicate similarly shaped profiles)
- Orthogonal projection of the genes upon an axis through two samples determines their specificity for these samples
- “Genes” and “samples” can be interchanged

Figure courtesy Luc Wouters
Additional calculations

Cartesian coordinates \((x,y,z)\)

Spherical coordinates \((r, \theta, \phi)\)

Orthogonal projection on a line through 2 points

\[
g \cdot \left( \frac{\mathbf{r}_1}{s_1} - \frac{\mathbf{r}_2}{s_2} \right)
\]
Example: corticotropin-releasing factor (CRF) in AtT-20 cells

Background

• CRF plays a central role in the regulation of the hypothalamic-pituitary-adrenal axis, mediating endocrine and behavioral responses to various stressors

• Micro-array experiments were done to elucidate the transcriptional response to CRF exposure in mouse pituitary derived AtT-20 cells

Experimental setup

1. DMSO (0.1%)
2. 1 µM CRF (in 0.1% DMSO)
3. 1 µM CRF + 1 µM antagonist
4. 1 µM antagonist

0 1 2 4 8 24
Time (h)

RNA extraction

AtT-20 cells
Example: time & CRF effects

Figure courtesy Pieter Peeters
Conclusions

• Experimental data in drug discovery should be analyzed by the subject matter experts
• Standardize QC methods for micro-array experiments and automatically generate a report
• Pre-calculate SMA data for interactive analysis by the expert
  – reveal groups or outliers in the samples
  – identify genes correlated to these samples
• SMA helps in exploring the mechanism of action of compounds by identifying gene expression signatures across different samples
• SMA reveals the presence of both time and CRF related effects in transcriptional responses to prolonged CRF exposure of AtT-20 cells
Special thanks to

- Biology: Pieter Peeters
- Micro-array experiments: Hinrich Göhlmann
- Statistics:
  - Luc wouters, Luc Bijnens
  - Paul Lewi
  - Transnational University Limburg at Hasselt, Belgium
- IT support: Steven Osselaer

References: