

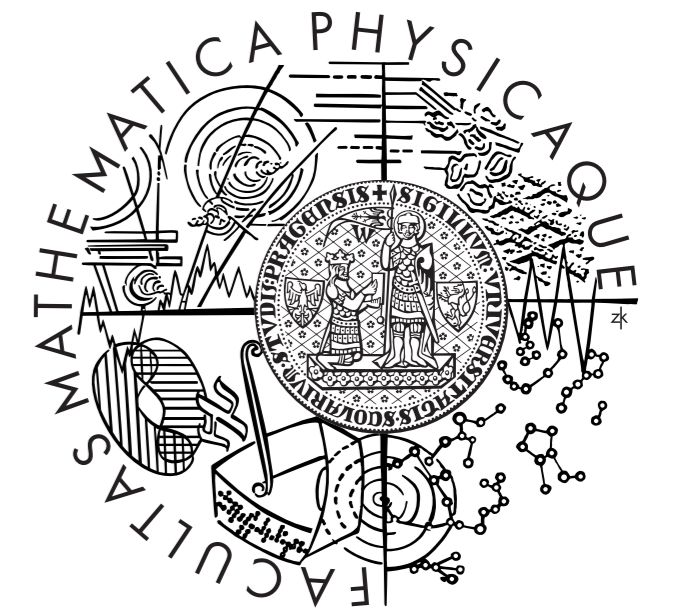
Gene Expression Data Analysis for In Vitro Toxicology



PETR ŠIMEČEK

simecek@karlin.mff.cuni.cz

Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, Prague

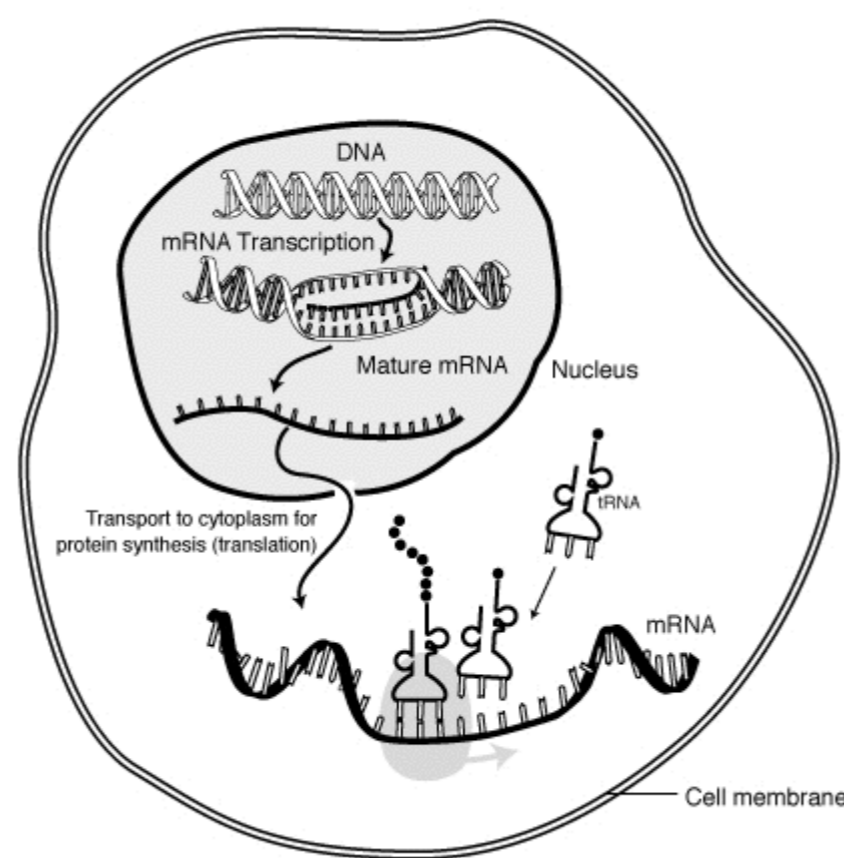


SUMMARY

The poster introduces an analysis of microarrays including preprocessing, identification of outliers and statistical tests. The methods are demonstrated on a problem of identification of genes whose expression is affected by exposure to the allergens but not by the irritants. The inference is based on a dataset containing 72 microarrays. Each microarray comprises CD34–DC sample that has been in contact with one of the 6 chemical compounds (4 allergens + 2 irritants).

BIOLOGICAL BACKGROUND

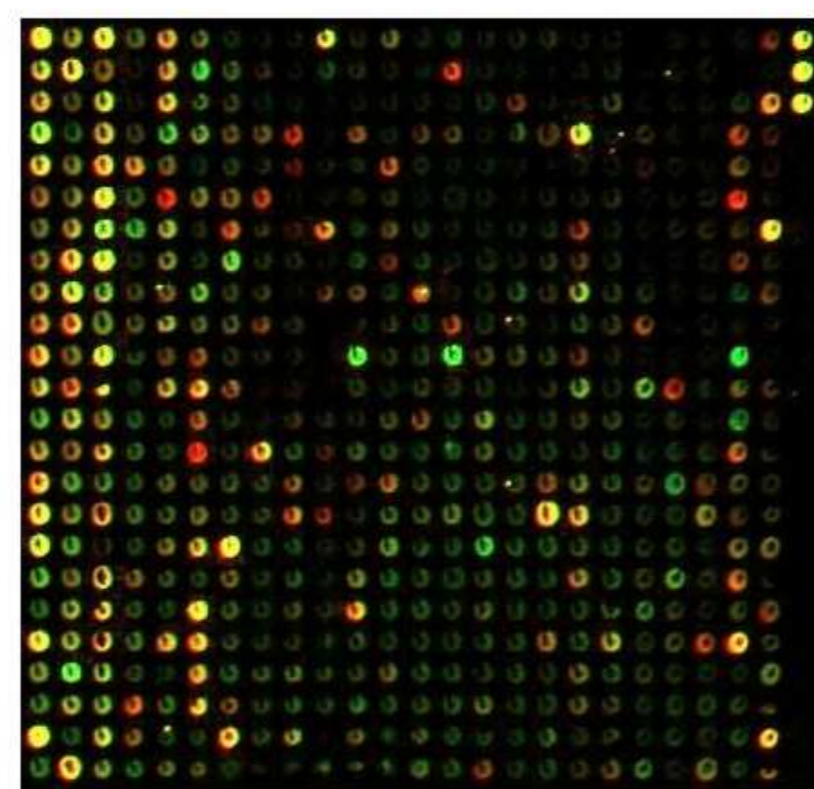
A **cDNA microarray** consists of a large number of single stranded DNA spots arranged in a grid. Microarrays are used to measure the expression levels of large numbers of different genes (encoding different proteins) simultaneously. From inspected cells **mRNA** is extracted, purified, amplified, reverse transcribed and indirectly labeled with fluorescent dyes Cy5 (red) and Cy3 (green). During hybridisation the labeled cDNA sequences present in the pooled mixture bind to their complementary sequences on the microarray. Unhybridized cDNA is washed off and the microarray is scanned in a laser scanner.



PREPROCESSING

Due to technology imperfections, substantial differences in intensity occur even among microarrays that are generated under exactly the same conditions. The purpose of **preprocessing** is to avoid such errors (cf. [1] and [2]).

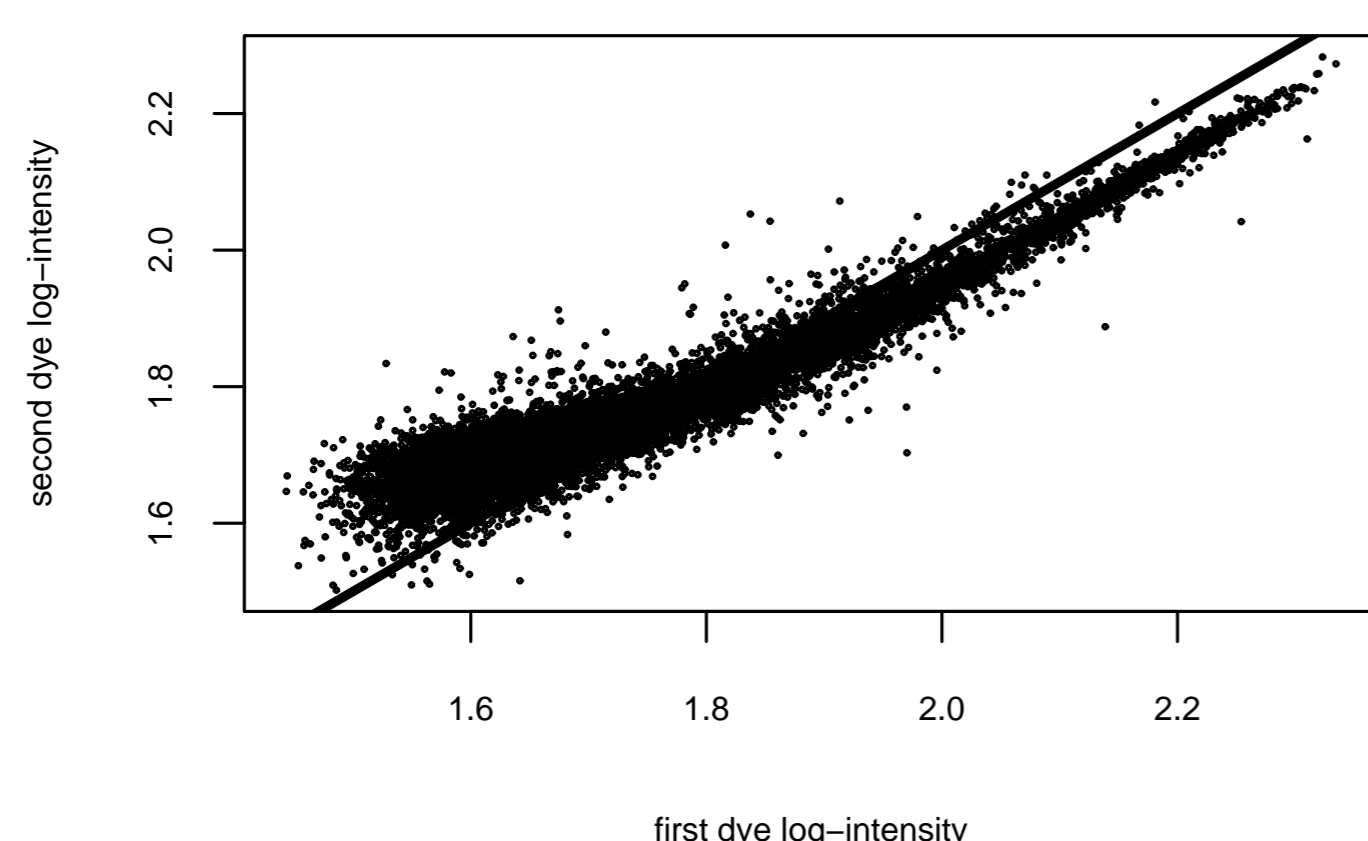
- A signal and a background must be separated. The raw intensity of the spots is strongly associated with the background intensity. That calls for a **background adjustment**. Let us denote RI_i the raw spot intensity of the i^{th} gene, and BI_i the mean background intensity for the i^{th} gene.



$$NI_i = \max(T, RI_i - BI_i)$$

- Variance of signals must be stabilized (e.g. by **log-transformation**).
- Some **normalization** technique (e.g. linear, loess, lts or quantile regression) must be used to transform arrays to the same scale. An usual assumption is that only a small number of genes is differently expressed.

Array 32



QUALITY ASSESSEMENT

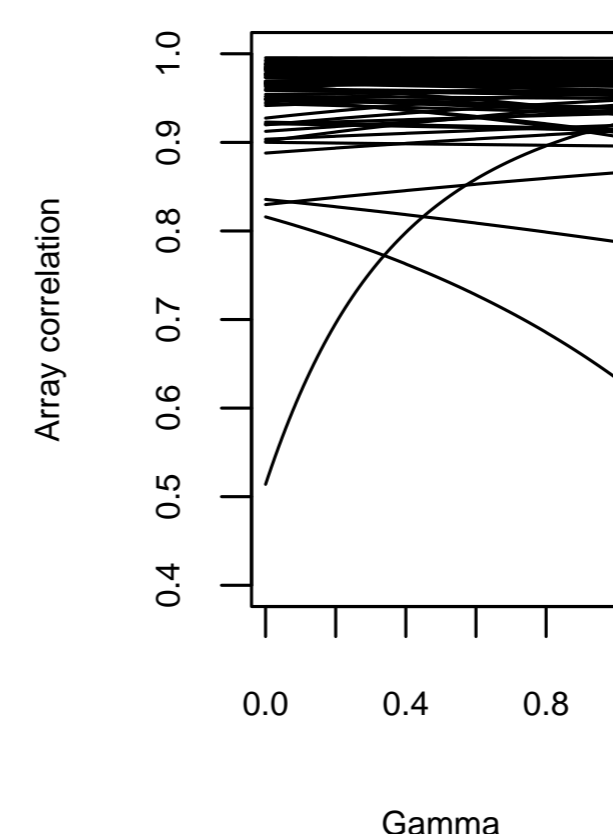
Quality of slides is usually graphically examined using scatter plots. This becomes difficult when the number of arrays is large. A helpful solution is proposed in [3].

For a given array, let $X_i^{(1)}$ and $X_i^{(2)}$ denote the log-intensities of i^{th} gene for the first and the second dye, respectively, and let A_i be a correction term computed by loess regression.

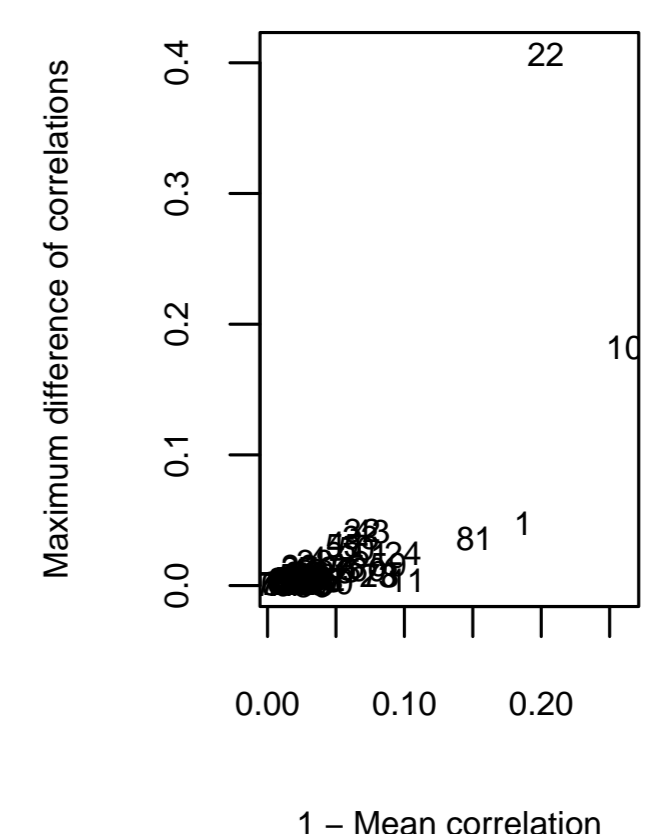
The idea is to divide A_i into two parts:

$$\begin{aligned}\tilde{X}_i^{(1)}(\lambda) &= X_i^{(1)} - \lambda \cdot A_i, \\ \tilde{X}_i^{(2)}(\lambda) &= X_i^{(2)} + (1 - \lambda) \cdot A_i.\end{aligned}$$

Diagnostic plot 1

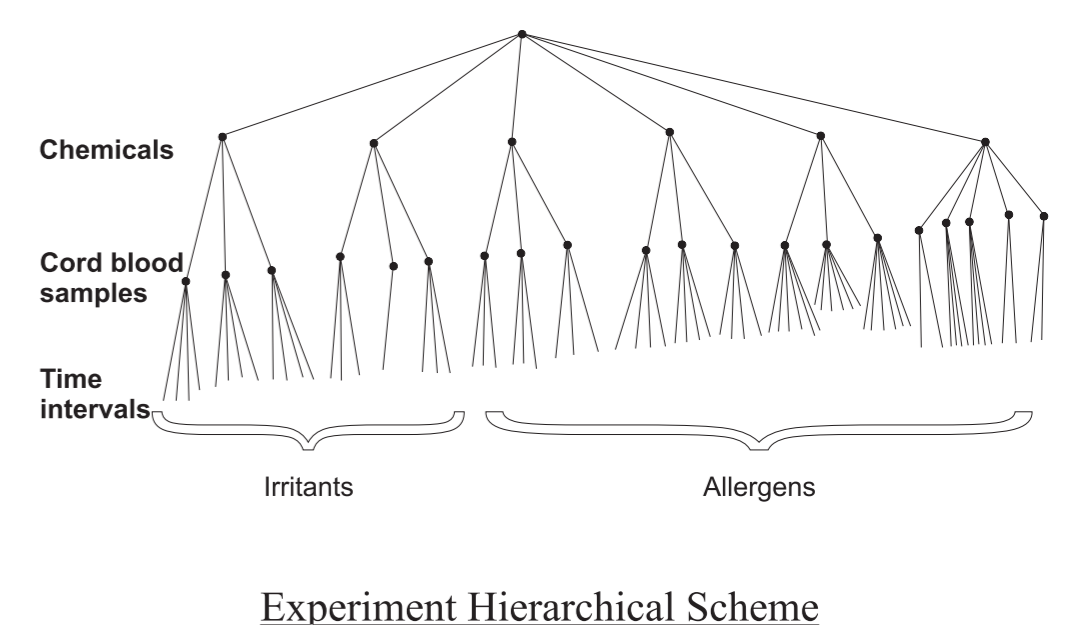


Diagnostic plot 2



STATISTICAL TESTS

When thousands of genes are tested for a change in expression due to an exposure, it can easily happen that a gene is marked as significant just by a chance. Extra attention must be therefore paid to **multiplicity adjustment** of the test level.



Several statistical tests have been performed (e.g. paired test and **ANOVA** + their nonparametric equivalents) and 68 (of 11395) genes have been found significantly differently expressed after exposure to allergens compared to irritants.

Acknowledgement: The poster was supported by the grant GAČR 201/05/H007.

References:

- [1] Amaratunga D. and Cabrera J. (2003). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley & Sons.
- [2] Draghici S. (2003). *Data Analysis Tools for DNA Microarrays*. Chapman & Hall.
- [3] Park T. et al. (2005). *Diagnostic Plots for Detecting Outlying Slides in a cDNA Microarray Experiment*. *BioTechniques* **38**.