CHARITÉ / HUMBOLDT UNIVERSITÄT ZU BERLIN
INSTITUTE FOR THEORETICAL BIOLOGY

Prof. Hanspeter Herzel
Dr. Karsten Jürchott
Institute for Theoretical Biology
D-10115 Berlin

email: h.herzel@biologie.hu-berlin.de
email: karsten.juerchott@charite.de
Tel: +49 30 2903 9106 / 6044
Fax: +49 30 2903 8801

# MODULE IV - BIOINFORMATICS: ASSIGNMENT 3

**Please email (recommended) your solutions to** karsten.juerchott@charite.de **until Thu, May 19, 24:00 or return as hard copy at the beginning of the lecture.**
Please double-check that your solutions are readable when printed on A4 paper.

## *1. Statistical data analysis*

The raw array data of 16 expression level measurements for a given probe set are:

$$\{x_i\} = (128, 64, 16, 32, 1024, 4096, 128, 256, 64, 1024, 512, 2048, 8192, 64, 256, 512).$$

- Calculate the mean value, the geometric mean value and the median.
- Calculate for the log-transformed data

  $$y_i = \log_2 x_i$$

  the mean, the variance, and plot a histogram.
- Give the formula how $y_i$ values can be transformed to a centered and standardized variable $z$ with zero mean and standard deviation equal to 1.
- Measurements after stimulation are

  $$\{u_i\} = (4096, 1024, 8192, 4096).$$

  Perform a t-test to check whether or not these levels are higher than the background values $x_i$.

## *2. Exponential decay*

The isotope $^{35}$S decays exponentially with a half-time of about 90 days.

- Sketch the time course of the $^{35}$S decay.

- Formulate a differential equation for the concentration of $^{35}$S .
- Calculate the rate constant of the decay.
- How the rate constant is related to the half-time?
- After how many months will only 0.1% of the isotope remain?

***3. Microarray data analysis***

Download a set of CEL-files from the GEO data base. We suggest GSE2639, but you can also take your own choice.

- Make a quality control (image, RNA degradation) and normalize the data. Control the normalization using scatterplots. (Skip bad arrays if necessary. Note that you have than to normalize the data set once again without the bad arrays.)
- Define two interesting groups of samples (e.g. control<->treated) and determine differentially expressed genes. Adjust the p-values by determining the false discovery rate (fdr) and draw a heatmap of significantly differentially expressed genes.
- Make a table of the top 10 differentially expressed genes including the probeset id, the gene symbol (if available), the means of the groups, the raw p-values, the fdr-adjusted p-values and a short information about the function of the gene (key words).
- Describe the analysis in a short protocol including the following information:
  - Name of the set used for the analysis,
  - Design of the set (e.g. cell line xxx, control vs. treated with xxx)
  - Biological background of the set (key words)
  - Are there bad arrays in the set?
  - How many differentially expressed genes did you have found?
  - Table of top 10 differentially expressed genes

Attach the heatmap and the R-script used for the analysis.