## Modeling Cell Fate

Prof. Dr. Volkhard Helms
Maryam Nazarieh, Thorsten Will
Summer Semester 2015

Saarland University
Chair for Computational Biology

# Exercise Sheet 1

**Due: May 22, 2015 10:00 am**

### Submission

- You are advised to work in groups of two people. If necessary, we will suggest teammates.
- Submit your solutions on paper at the beginning of the lecture in the lecture hall or in Room 3.02, both E2 1. Alternatively you may send an email with a single PDF attachment to maryam.nazarieh@bioinformatik.uni-saarland.de. Late submissions will not be considered.
- If appropriate, include source code listings into the submitted document, we will not merge and layout your source code. If relevant sources are missing on the exercise sheet, they will not be graded.
- Do not forget to mention your names/matriculation numbers.
- Discussion of this exercise will be on Tuesday, May 26th at 12:45 in the lecture room (E2 1 007).

### Exercise 1.1: Circadian rhythms (20 points)

- What are circadian rythms and essential elements of biological clocks?
- Describe the parameters of circadian clocks.
- Explain microarray and RNA-seq methods briefly and compare them by mentioning their pros and cons regarding the measuring of gene expression.
- Explain the difference between discrete time-series and continuous time-series data.

### Exercise 1.2: Minimization Problem (25 points)

Let X and Y be two random variables with $E(Y) = \mu$ and $EY^2 < \infty$.

- Show that the constant c minimizes $E(Y - c)^2$ is $c = \mu$.
- Deduce that the random variable $f(X)$ that minimizes $E[(Y - f(X))^2|X]$ is $f(X) = E[Y|X]$.
- Deduce that the random variable $f(X)$ that minimizes $E(Y - f(X))^2$ is also $f(X) = E[Y|X]$.

### Exercise 1.3: Microarray Expression Analysis of cell cycle data (55 points)

The yeast Saccharomyces cerevisiae (SC) is a single-celled organism with well-studied genetics. Thus it is used as a model organism. (The cell cycle is the series of events that takes place in a cell leading to its division and replication.) Microarray analysis of gene expression during the SC cell cycle helps to infer distinct subset of circadian genes.

Download the gene expression dataset (sine_waved.csv) which contains 10 genes with 21 samples and use the following:

- Write a small (e.g Python) script to read in the time series expression data.
- Determine by using the method of least square for each gene, the optimal phase of a sine wave that fits the data of this gene. Note that you may need to shift the sine wave to the base-level expression and you may need to adjust amplitude of the sine.
- Randomly shuffle the data points for this gene 100 times and compute optimal phase and sum of squared errors again.
- Determine using False Discovery Rate ($FDR < 0.05$) whether this gene is circadian.

Download a preprocessed (i.e. normalized) set of cell cycle expression dataset (cell_cycle.csv). It consists of 25 time points with $\sim$ 5000 genes which has been measured with interval of 5 minutes. **(25 bonus points)**

- Read the data into a data matrix where the rows correspond to the set of genes in each sample and columns correspond to the different samples.
- Use one of the cycling detection algorithms, like JTK_CYCLE, COSOPT or Fisher's G test (available in GeneCycle package in R library) to efficiently identify and characterize cycling variables in large data sets. How many genes are significantly periodically expressed?

Have fun!