

Requirements

The tasks are to be solved in groups of maximally two students. The solutions must be documented and every student should be able to explain them. They must be handed in not later than two weeks after the date of the exercise.

1 Explorative Analysis and Normalisation of microarrays

The following practical part should be done in R with `Biobase` installed. The latter provides access to many useful functions.

1. Explorative analysis. Use the `lymphoma` data set from the `vsn` package:

```
library(vsn);data(lymphoma);e <- exprs(lymphoma)
```

Now `e` contains the expression matrix. To reduce the computational cost for this exercise, select 4 chips for explorative data analysis and normalization. Remove invalid data entries. Use the following tools:

- distribution of expression values for each chip (boxplot, qq-plot)
- scatterplots: unnormalized and logarithmic scale.
- Matrix of Absolute Deviation (MAD)
- plot the standard deviation against mean (sdm plots with `meanSdPlot`), for logarithmically normalized and un-normalized data.

2. Normalisierung und Vergleich

- (a) Which transformation function is used by the VSN routine? Plot this function and the logarithmic function in the same coordinate system.
- (b) What are the advantages of VSN ?
- (c) What is the most important assumption underlying the normalization with `vsn`?
- (d) Use `vsn` to normalize the above data (**no** background correction, and **without** median polish summation).
- (e) Compare the transformed with the original data (boxplots, scatterplots)

2 Sequence-based Analysis of Regulatory Networks

In a perturbation experiment of the serum response factor (SRF) a number of human genes were found to be affected, most notably those with the following Hugo IDs: `actc1,acta1, acta2,actb,egr1,egr2,egr3, myl3,fos,jun,myc`

1. BioMart (www.ensembl.org/biomart): Extract the ± 250 bp sequences around the transcription start site (5'UTR) of the above genes and download them to a fasta file. Why could there be redundant entries? Double-check that the final sequences obtained are unique and do not overlap. Can you observed any particular sequence features ?
2. MEME (meme.nbcrl.net): Determine whether there are any sequence patterns which occur more frequently than expected. In how many sequences do they occur ?
3. JASPAR (<http://jaspar.genereg.net>): Can you associate any of the identified motifs with a transcription factor binding site? Search the Jaspas database for a transcription factor with a similar consensus sequence? Given an interpretation of your findings. What about the other sequence patterns ?
4. Genome Browser (<http://promotion.molgen.mpg.de>): Search for the *actin- α* gene on chromosome 1 in the above genome browser and locate its approximal transcription start site. Which information can be used ? Zoom out to visualize the 10kb region around *actin- α* . Assess the degree of evolutionary sequence conservation. Using the analysis tools provided on that web-page, identify predicted binding sites for the SRF motif. Are there any conserved predicted binding sites ?

3 Simple Graphs and Graph Observables

1. write down the **adjacency matrix** of the undirected graph in Fig. 1
2. determine the **degree distribution**
3. calculate the **graph spectrum**: $\{\alpha_i\}$
4. determine the **Perron-Frobenius eigenvalue** and the corresponding eigenvector. Does the Perron-Frobenius theorem hold ?
5. introduce an undirected link between nodes X1 and X4 and calculate successive powers of the new adjacency matrix: $A^k, k = 1 \dots 5$.

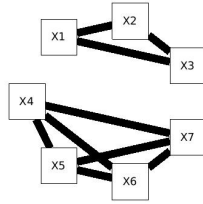


Figure 1:

6. What is the **distance** between X1 and X5? Explain how this could be obtained from appropriate matrix products of the adjacency matrix A .
7. What is the **diameter** of A ?
8. Calculate the new graph spectrum of A and the Perron-Frobenius eigenvalue.
9. calculate $\sum_i \alpha_i$ and given an interpretation
10. calculate $\sum_i \alpha_i^2$ and given an interpretation
11. calculate $\sum_i \alpha_i^3$ and given an interpretation

4 Real Networks: Protein-Protein Interactions

Protein-protein interaction networks are **undirected** networks which represent observed physical interactions as links between proteins (nodes).

1. Obtain the **protein-protein interaction network (PPI)** for budding yeast from <ftp://ftpmips.gsf.de/yeast/>
2. determine and plot the **degree distribution** $P(k)$
3. determine the maximal degree, average degree and average clustering coefficient.
4. assume as background an **Erdős-Renyi graph model** with the same number of nodes and the same average degree. What is to be expected for the diameter and clustering coefficient? How does this compare to

the observed values? Given the ER-model, what is the probability for observing the same maximal degree?

5. fit the degree distribution to (a) log-normal and (b) power-low functional form. What are the coefficients?
6. Solve one of the following problem sets
 - obtain the **Markov clustering algorithm** from <http://micans.org/mcl> and determine clusters for a range of different inflation parameters. Record the number of clusters obtained
 - Bonus: Compare the cluster solution from your favorite parameter with the set of known protein complexes (from <ftp://ftpmips.gsf.de/yeast/catalogues/complexcat>). How would you estimate the quality of the cluster solution?

or

- obtain **functional data** on protein essentiality (gene disruption) from the catalogues at <ftp://ftpmips.gsf.de/yeast/>.
- Define two classes of proteins, whose inhibition is (I) lethal or (II) viable. Compare the degrees in both groups and test (Wilcoxon or Kolmogorov-Smirnov) whether the two distributions differ.

5 Real Networks: Regulatory Networks

The nodes in regulatory networks correspond to transcription factors (TFs) and genes. They are connected by **directed links** to represent the regulation of a gene by a transcription factor. Since transcription factors are encoded by genes these networks permit more complex subgraph structures, such as feedback and feedforward loops.

1. Obtain the **regulatory network** for *E.coli* from the group of Uri Alon (<http://www.weizmann.ac.il/mcb/UriAlon>, file: coliInterfullVec.txt)
2. plot the distribution of **in-degrees** and **out-degrees** for genes and transcription factors, respectively. Attempt to fit them to exponential, log-normal and power-law.
3. How many **auto-regulatory interactions** (self-loops) are in the network? Remove them for further analysis.

4. Using **powers of the adjacency matrix**, determine all TF-gene pairs which contain one intermediate regulator (all pairs at distance 2). How many are there ? How many of those pairs also have a direct interaction ? Can you think of a simple matrix operation to determine the number of **feedforward loops** ?
5. Download the **Motif-finder** (mfinder version 1.2) from the above mentioned web-page and determine the enrichment of feedforward loops and all other 3-node subgraphs. Briefly describe the random graph model which was assumed as background model.

6 Evolving Networks

1. Generate a growing network according to the **Barabasi-Albert model**: starting with $m_0 = 2$ isolated nodes, and add one node with $m = 2$ edges at each time step during the evolution. Assume a linear **preferential attachment**.
2. determine the degree distribution for various times and estimate the **scaling coefficient** for large times.
3. repeat the simulation for a simpler model with **random attachment**. Which form has the degree distribution.

7 Graph Visualization with Cytoscape

Cytoscape is the most important tool for visualizing interacting genes and proteins.

1. If cytoscape is not yet installed on your computer, obtain and install it from <http://www.cytoscape.org/>
2. load the E.coli regulatory network from section 5. You might want to try several graphical layouts and select your favorite visualization.
3. configure the edges so that they distinguish between “activators”, “repressors” or “dual” regulation.
4. obtain the dictionary `coliInterFullNames.txt` from Uri Alon’s website (see above) and import this list as an attribute table. With the help of this dictionary it will be possible to replace the uninformative

integer IDs with the common name of the transcription factor/operon (Hint: Use VizMapper → NodeLabel).

5. Color-code the nodes according to their out-degree. This requires to calculate and load this property as an attribute table, but there might also be plugins to do such a calculations.

To document your solution of these tasks, export a single snapshot of your figure which illustrates the items above.