Bioinformatics III

Prof. Dr. Volkhard Helms
Trang Do
Summer Semester 2021

Saarland University
Chair for Computational Biology

Exercise Sheet 6

Due: June 3, 2021 12:00

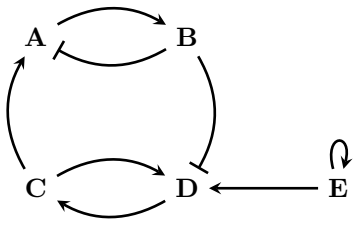
Submit your solutions to trangdht.bioinfo@gmail.com with two attachments: (1) A ZIP file containing all your source code files, potential result files, figures and whatever else is needed to generate your solution, (2) a PDF file containing your answers. Subject of the email should be in the following format: A6_LastName1_LastName2.

Please always follow the rules for submission introduced in Assignment 1!
Most tasks can be implemented using Python standard libraries and fundamental libraries (scipy, matplotlib, pandas, etc.) unless stated otherwise.

Boolean Networks and GRNs Reconstruction

Exercise 6.1: Boolean Network (50 points)

Consider the following network, which describes the mutual regulation of the hypothetical genes **A** to **E**. A line with an arrowhead denotes an activation while a flat end denotes an inhibition, i.e., if **A** is high, **B** is activated, whereas high levels of **B** inhibit the expression of **A**.



To investigate the behavior of this network use a dynamic simulation as introduced in the lecture with a synchronous update scheme.

Assume that an activation has a weight of 1, while an inhibition is always 3 times stronger than an activation. Set all thresholds to 0.

(a) **Weighted Interactions**
Set up the propagation matrix (also called condition tables in the lecture) that relates the states of the genes **A** to **E** in the next iteration to the current state.

(b) **Implementation**
Write a program to simulate the Boolean Network.

To enumerate the initial states, convert the binary levels of the genes into an integer where **A** determines the least significant bit and **E** the most significant one. An initial state where, e.g., only **A**, **C**, and **D** are on would translate into $1 + 4 + 8 = 13$.

(1) When does it make sense to stop the propagation and why?

(2) Which sequences of states (also called trajectories) do you get when you start from states 7, 13, 17, and 23?

(c) **Periodic Orbits**
To determine the attractors and the corresponding basins of attraction, go through all possible initial states and save at which state the Boolean network returns to a state that it has visited before.

(1) List these orbits with their respective lengths and basins of attraction.

(2) Give the relative coverages of the state space by the basins of attraction.

(d) **Interpretation**

(1) Give the attractors in terms of active genes and characterize them with a few words.

(2) Which ones are the special genes and what are their respective effects on the behavior of the network? For this, explain what is determining the period of the orbits.

Exercise 6.2: GRNs Reconstruction - DREAM challenge (50 points)

Apply the Noise model introduced in the lecture ($V13$ *slide* 19) to predict the directed unsigned GRN topology of *E.coli* from steady state gene expression data. The target network is of size 10 genes without self-regulatory interactions.

(a) Download the gene expression dataset containing the following data:

(1) **heterozygous.tsv** contains the steady state levels for the wild-type and the heterozygous knock-down strains for each gene. Thus, for a network of size 10 there are 11 experiments (wild-type plus knock-down of every gene).

(2) **null-mutants.tsv** contains the steady state levels for the wild-type and the null-mutant strains for each gene. Thus, for a network of size 10 there are 11 experiments (wild-type plus knock- down of every gene).

(b) Construct the Noise model and apply it on each dataset. The expected output to be submitted should be a ranked list of regulatory link predictions ordered according to the significance of each prediction (sample output file is in the supplementary).
For example: **G1 G2 score**
Where **G1** and **G2** are two different genes (no self-interactions). Note that links between them are directed: the gene in the first column regulates the gene in the second column. (If both **G1** regulates **G2** and **G2** regulates **G1**, then both lines should be included). **Score** is between 0 and 1 and indicates the confidence level you set to this link prediction.

(c) Compare the outcomes from the heterozygous knock-down strains and the null-mutant strains.

(d) Refine the prediction score by re-estimating the wild-type expression level and the Gaussian noise iteratively ($V13$ *slide* 23). Again, for each dataset, report the gene interactions and scores after performing t iterations, where t is a value of your preference applied to both datasets. Explain why you choose this t value and include your comments on the generated GRNs.