CS-E5880 Modeling biological networks (spring 2019)

Assignment project #2: Identification of gene regulatory network from gene expression time-course data

This assignment project can be done in pairs (two students).

This assignment gives you a hands-on experience on inferring the structure of biological networks.

Cantone et al. (2009) have reported a small synthetically constructed transcriptional network in yeast (see Fig. 1). The transcriptional network consists of 5 transcription factors (TF), which regulate each other in a specific manner as depicted in Fig. 1. In addition, endogenous yeast genes have a negligible effect on the operation and dynamics of this 5-gene network, i.e., we can assume that this 5-gene network operates in isolation from all other genes in yeast. Thus, you can ignore all other yeast genes during your analysis. This small yeast network is an excellent system for testing and demonstrating the power (or lack of power) of various network inference methods. We will use this 5-gene network and apply computational methods to learn its structure using experimental data only. For further details, see the original publication (Cantone et al., 2009).

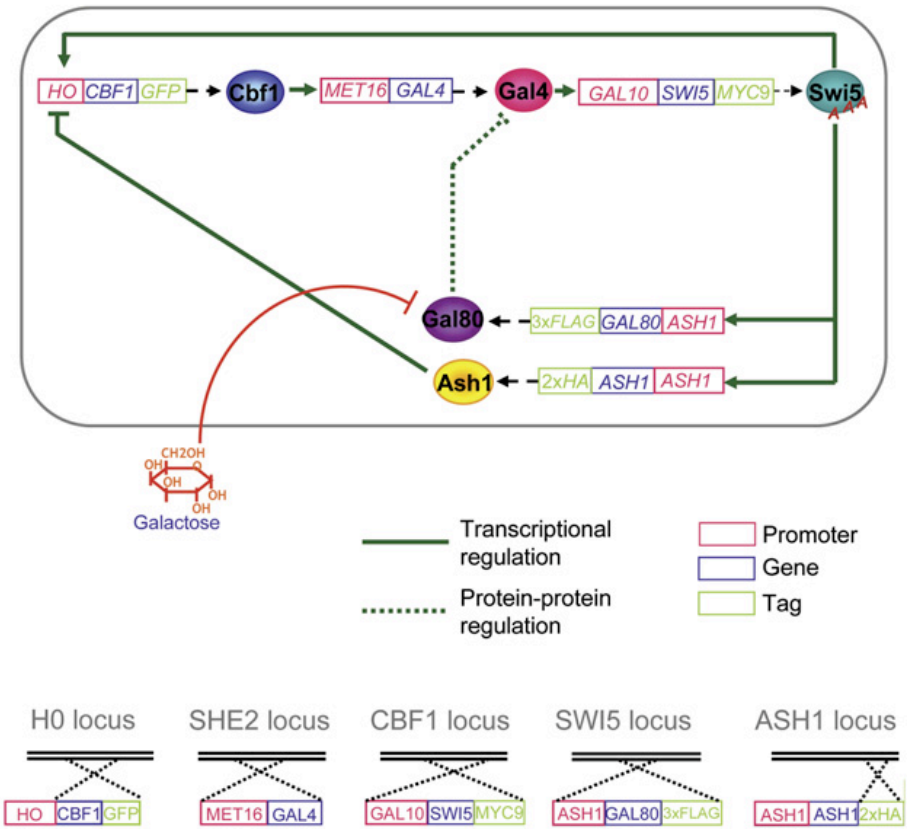

Fig. 1. A 5-gene network from (Cantone et al., 2009).

The goal of this assignment is to use computational/statistical method(s) to infer the *structure* of this 5-gene network using gene expression time-course data only. The gene expression has been measured at 10 min time intervals from 0 min to 190 min. The data for the 5 genes can be found in a tabular format at the end of this document.

You can use any (or many) of the methods that were introduced during the lectures, including ODEs, SDEs, linear/non-linear regression models, Bayesian networks, relevance networks (correlation, mutual information), dependency networks, etc., or you may use other methods from the literature as long as they are meaningful for this problem. Please justify your choice of method(s). If you choose to use a method from the literature that is not covered in the lectures, you will need to provide a more comprehensive description of the computational methods in your report as well as full reference to the publication/book/webpage/other material where the model is introduced.

If you find this assignment problem difficult, you may simplify your work by considering only a small number of candidate network structures among which you choose the "best" one. This will likely be the case e.g. if you try to learn the network structure with ODE systems without any approximations. Please explain your approach in the written report.

Performance validation[1]

To measure how well you are able to reconstruct the transcriptional network structure from the gene expression data, you can use the known network shown in Fig. 1 as a reference. You should not expect exact network structure reconstruction.

Vary the detection threshold for calling an interaction between a TF and a target gene (probability, p-value, FDR-value, cross-validation error, or other score; whatever you decide to use) for your chosen computational method and report:

- the fraction of real TF-gene interactions you detect
- the fraction of false positive TF-gene interactions you detect

and reported these numbers for different values of the threshold. What threshold value is required to detect all 6 TF-gene interactions present in the synthetically constructed network in Fig. 1? Also, how many false interactions you will detect with that threshold?

By varying the detection threshold, you can also generate a so-called receiver operating characteristics (ROC) plot of the performance. If you are not familiar with ROC before, a brief description of ROCs can be found e.g. here:

[1] This section applies only if you try to do global network analysis, i.e., try to find the best network structure among all possible structures. If you decide to choose the best network among a limited number of network structures, then you can assess accuracy of your inference among the limited set of network structures.

- Murphy KP, *Machine Learning: A Probabilistic Perspective*, MIT press, 2012 (Section 5.7.2.1; this is one of the course books and is available as a paper copy and as an e-book via our library)
- http://en.wikipedia.org/wiki/Receiver_operating_characteristic

Report

Complete the above analysis steps and write about 3 (or more) page report briefly describing your computational method(s) and summarizing your results, findings and other observations. Include a separate cover page containing your **full names** and **student numbers**.

Deadline

The deadline for the report is March 4, 2019 at 23:59 (Finnish time zone). Return your report via the course webpage.

References

Cantone I et al. (2009) A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell*, 137, 172–181.

DATA:

time\gene	SWI5	CBF1	GAL4	GAL80	ASH1
0	0.076	0.0419	0.0207	0.0225	0.1033
10	0.0186	0.0365	0.0122	0.0175	0.0462
20	0.009	0.0514	0.0073	0.0165	0.0439
30	0.0117	0.0473	0.0079	0.0147	0.0371
40	0.0088	0.0482	0.0084	0.0145	0.0475
50	0.0095	0.0546	0.01	0.0144	0.0468
60	0.0075	0.0648	0.0096	0.0106	0.0347
70	0.007	0.0552	0.0107	0.0119	0.0247
80	0.0081	0.0497	0.0113	0.0104	0.0269
90	0.0057	0.0352	0.0116	0.0142	0.019
100	0.0052	0.0358	0.0073	0.0084	0.0134
110	0.0093	0.0338	0.0075	0.0097	0.0148
120	0.0055	0.0309	0.0082	0.0088	0.0101
130	0.006	0.0232	0.0078	0.0087	0.0088
140	0.0069	0.0191	0.0089	0.0086	0.008
150	0.0093	0.019	0.0104	0.011	0.009
160	0.009	0.0176	0.0114	0.0124	0.0113
170	0.0129	0.0105	0.01	0.0093	0.0154
180	0.0022	0.0081	0.0086	0.0079	0.003
190	0.0018	0.0072	0.0078	0.0103	0.0012

Here is what the data should look like: