CS-E5885 Modeling biological networks (autumn 2024)

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

This assignment project can be done in pairs (two students). A group of two students must return a joint project report.

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

Cantone et al. (2009) have reported a small synthetically constructed transcriptional network in yeast (see Fig. 1 below). The transcriptional network consists of 5 transcription factor (TF) genes, which regulate each other in a specific manner as shown in Fig. 1. Endogenous yeast genes (i.e., any other yeast genes) have a negligible effect on the operation and dynamics of this 5-gene network. In other words, we can assume that this 5-gene network operates in isolation from all other genes in yeast and you can ignore all other yeast genes during your analysis. For further details, see the original publication (Cantone et al., 2009).

This small yeast network is an excellent system for testing and demonstrating the power, or lack of power, of various biological network inference methods. We will use this 5-gene network and apply computational methods to learn its structure using experimental data only.

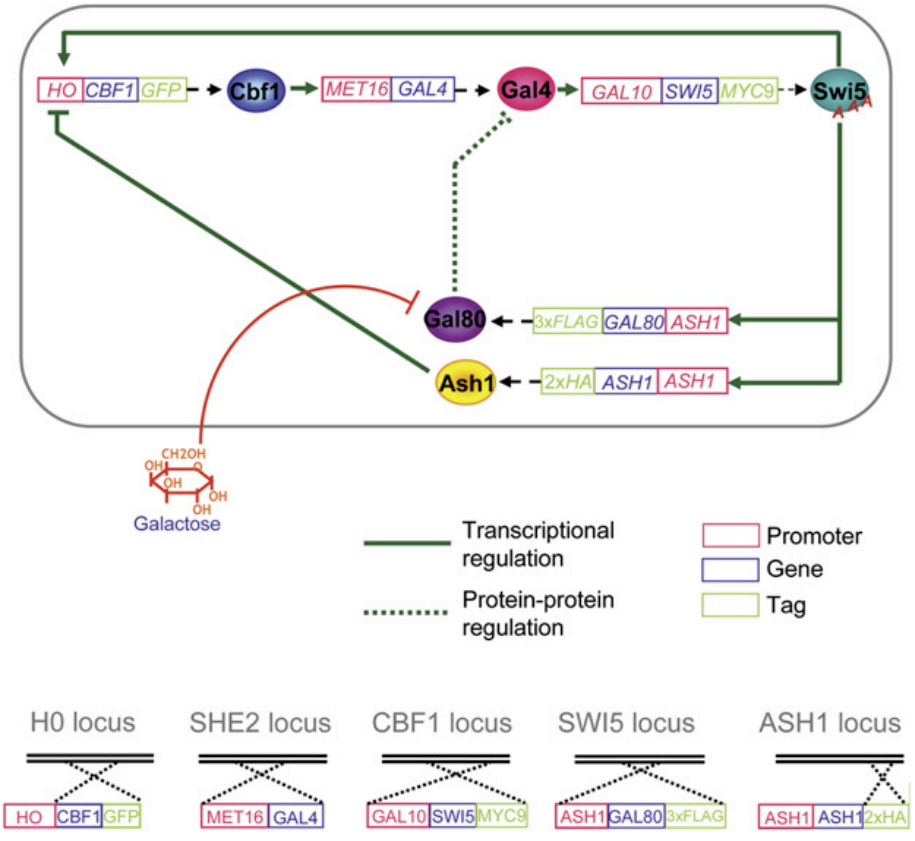

Fig. 1. A 5-gene network from (Cantone et al., 2009).

The goal of this project assignment is to use computational, statistical and/or machine learning method(s) to infer the *structure* of this 5-gene network using gene expression time-course data only. The original paper by Cantone et al. (2009) contains data from two time-series experiments. Here we will use only one them, called switch-off experiment, where the activity/abundance of the galactose is decreased at the beginning of the experiment. After switching off the galactose, 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 one or several biological network models that were introduced during the lectures, including e.g. ODEs, SDEs, linear/non-linear regression models, Bayesian networks, relevance networks (correlation, mutual information), dependency networks, etc. as well as various methods for learning their model structure using e.g. cross-validation, Bayesian methods, their approximations, etc. You may also use other methods that you have learned in other courses or can find 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 full 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.

This assignment problem can be challenging. You may simplify your work by considering only a small number of candidate network structures, and try to find the "best" from this small set of candidate networks. This simplification will likely be a necessary especially if you aim to model the gene regulatory network using more detailed mechanistic models, such as ODEs or SDEs without any simplifying approximations. Please explain your modeling approach in the written report.

Note that this is an open project assignment in a sense that you are not told to use a specific computational method but rather you need to make that decision yourself.

You do not need to implement all analysis methods yourself from scratch. You can use existing functions, software implementations and software packages. However, you need to be sure that you understand that what each software package does and you need to give citations (or web link) to methods that you use. Additionally, you still need to explain mathematically in your report that what these external software packages/implementations do (see Report section below).

Performance evaluation

To measure how well you are able to reconstruct the transcriptional network structure from the gene expression data, you can use the known network structure shown in Fig. 1 as a reference. Vary the detection threshold (probability, p-value, FDR-value, cross-validation error, or other score; whatever you decide to use) for your chosen computational method to decide whether an interaction between a regulator transcription factor gene and a target gene is true or not, and report:

- the fraction of true 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 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:

- 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 an e-book via our library, and the book can also be found from the internet)
- http://en.wikipedia.org/wiki/Receiver_operating_characteristic

Instructions for performance evaluation in this section apply directly only if you try to do global network analysis, i.e., try to find the best network structure among all (or at least a large number of) 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.

Note again that the problem is rather challenging and therefore you should not expect perfect network structure reconstruction.

Report

Complete the above analysis steps and write a report that describes your computational method(s) and summarizes your results, findings and other observations.

The second goal of this project assignment is to prepare a well-written report. An important part of the project report is a **technical description** of your method, described using mathematical and statistical formulations (i.e., using equations). If you choose to use method(s) that were covered in the lectures, then the technical description can naturally have similarities with the description in the lecture materials and the course books. However, explain the method(s) in your own words. Try to prepare a method description that is mathematically "sufficiently detailed", for example, in sense that others could re-implement your method and analysis based on your description. **Importantly, include also your motivations and explanations why you chose your method(s).**

Results section should contain performance measures as well as some additional result figures (e.g., an illustration of the model fit to the data).

Your report (including result figures) can be e.g. 3-5 pages long, but you can write more if you consider that to benefit your report overall (see also grading below). Include a separate cover page containing your **full name(s)** and **student number(s)**. Submit your report in PDF format. In addition to the 3-5 (or more) pages, it is recommended that you include your code in the report (appendix) or submit code as a separate file.

Deadline

The deadline for the report is December 17, 2024 at 23:59 (Finnish time zone). Return your report via the course webpage.

Grading / Additional instructions

Grading is **NOT** based on

- the performance measures (unless your performance is as good as random guessing; in that case you should think about tweaking your method or perhaps choosing a different method)
- the length of the report

Higher grades will be given to reports that

- are mathematically and technically correct and precise
- contain results for method(s) that are at least a bit more sophisticated than pure correlation measures
- summarize the results such that a reader gets an understanding of the performance
- explain why a specific method was chosen
- include explanations, or at least some speculations, why the chosen method(s) perform well/poorly
- contain some ideas, or at least some speculations, of how the computational analysis could be improved
- are carefully and fluently written
- are compact, yet sufficiently comprehensive

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 (on the next page):

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

Below is a figure of what the data should look like: