Mathematics of Cellular Networks

Prof. Dr. Volkhard Helms
Thorsten Will
Sommer Semester 2014

Saarland University
Chair for Computational Biology

Exercise Sheet 6

Due: Juli 15, 2014 10:15

Send a single PDF attachment with your name(s), your solution(s). Hand in all source code via mail to thorsten.will@bioinformatik.uni-saarland.de.
Please read the exercises carefully to not miss any questions.

1 Basic information theory (15 points)

(a) You learned in the lecture that maximum likelihood scoring in structure learning is related to the mutual information, a measure of dependency between two random variables with its origin in information theory. It is defined as:

$$I(X,Y) = \sum_{x \in X, y \in Y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right).$$

However, Pearson's correlation coefficient is a more common measure of dependency. Read up on both measures to have a better understanding of their meaning and implications.
What are the major differences between the two measures? Name applications where one of the two measures is better suited than the other (and vice versa). Justify your reasoning. (10 points)

(b) Another concept from information theory that can be frequently found in computational biology is the (Shannon) entropy. The Shannon entropy H is a measure of uncertainty (often also termed "surprisal") that can be used to quantify the minimal and maximal information content of a message over an alphabet $S = \{x_1, x_2, ..., x_n\}$ given the distribution of the individual x_i.
The bitwise entropy is defined as:

$$H(S) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i).$$

The four DNA bases are a common alphabet in the context of sequence analysis, thus $S_{\mathrm{DNA}} = \{A, C, G, T\}$. What is the possible range of the entropy $H(S_{\mathrm{DNA}})$ and which base distributions yield the extreme values? (5 points)

2 Learning trees and forests from data (50 points)

In this exercise you will infer the structure of a Bayesian network from data using simple methods that use measures from information theory. For convenience we assume that each random variable is influenced by at most one parent.

(a) The file "net_samples.txt" contains samples (rows) from 5 binary random variables (columns) that were generated with a Bayesian network of unknown structure.
Compute and report the mutual information for all possible pairs of random variables as well as the entropy of all individual random variables. Use the definitions given in the previous exercise. (15 points)

(b) As a first approach, construct a spanning tree following Kruskal's algorithm:
- Start with the empty graph G and the set of possible edges E (all pairwise edges).
- Sort all edges by their mutual information I.
- In order of decreasing weight, take each edge and add it to G unless it introduces a cycle.
Return this step until all vertices (the random variables) are connected.

Briefly argue that a thus constructed spanning tree G has the optimal score

$$S(G) = \sum_{(x,y)\in G} I(x,y).$$

Your graph G so far is undirected. To transform your graph to a directed one, choose for each edge the path of increased entropy (parent has lower entropy). Reason why this should be the direction of choice in the framework of Bayesian networks.
Draw the network you inferred and point out why such a method will likely overfit the data. (15 points)

(c) We now want to find the maximal spanning forest that maximizes the regularized score

$$S_r(G) = \alpha^{|G|} \sum_{(x,y)\in G} I(x,y),$$

where $|G|$ is the number of edges in G.
Apply Kruskal's algorithm again (use $\alpha = 0.9$) but stop if the addition of an edge does not increase S_r. Why is the output of the approach a forest and not a tree?
Report all intermediate S_r and your final result with directed edges.
Furthermore, comment on the differences to your result in (b) and discuss why regularization may be worthwhile. How does an increase/decrease of α affect the outcome? Which range of α would you recommend in practice (approximately)? (15 points)

(d) Unrelated to the previous parts of the exercise, you have also learned about local optimization to infer network structures that additionally allow for cycles.
Given an arbitrary starting or candidate structure, the graph is iteratively optimized (an edge added, removed or reversed) until a certain quality measure (like our scores) converges.
What is the major disadvantage of such an approach and how would you tackle it? (5 points)

3 Gene regulatory networks and data integration (35 points)

Gene regulatory networks (GRN) are directed graphs that depict the regulatory interplay between gene products. Classically they associate transcription factors (TFs) with their target genes by binding to the promotor and (most likely) regulate.

In this exercise you are provided with heavily preprocessed data for yeast:

- "tf_targets.csv" contains in each line a TF and all proteins that the protein binds to. The data consists of experimentally (e.g. ChIP-seq) and computationally derived (e.g. motif searches) binding events.
- "expression.csv" comprises normalized expression data for a cell-cycle time-series.

After a quick glance into the files their layout should be self-explanatory.

(a) Build an initial GRN using the TF/target association given in "tf_targets.csv". Report the number of proteins in the network and the number of interactions.
Discuss on this example why the inclusion of heterogeneous data sources is mostly worthwhile. (10 points)

(b) Refine your initial GRN by integrating the supplied expression data. Only allow an interaction between a TF and a target if the absolute correlation coefficient of their genes' expression is above 0.9. Skip the pair if there is no expression data.
Again, report the number of proteins in the network and the number of interactions.
Additionally, report the TF with the most outgoing interactions and argue why this seems reasonable. Why does it make sense to use the absolute value of the correlation in this context? Lastly, it is always necessary to preprocess expression data (correction of batch-effects etc.), but is it also necessary to normalize the data (transform to mean zero and variance one) if the correlation is the only measure used on the dataset? (15 points)

(c) Crucial regulatory drivers are often referred to as 'master regulators'. While there is no fixed definition of the term, it commonly means the TFs on the highest level of the regulatory hierarchy. Use the notion of topological sorting (a node ordering introduced in the fourth assignment) to determine putative master regulators in the refined network.
How many parents do the TFs in the highest level of the hierarchy have and how many equivalent TFs are on this level? (5 points)

(d) Discuss why the inclusion of expression data might introduce a strong bias in our case given that it is from the cell-cycle (vs other cellular states) and a time-series (vs steady-state) data. (5 points)