Prof. Jörg Stelling
D-BSSE
ETH Center UNG H14
Universitätstrasse 19
CH-8092 Zürich
https://csb.ethz.ch

Computational Systems Biology
Autumn 2023

Assignment 2

(Issue: 03-Oct-2025)

Probabilistic Graphical Models

Inductive Inference in Bayesian Networks

We are interested in the analysis of the small gene network shown below. We start from a set of observed gene expression profiles and from the structure of a Bayesian network to model the dependencies between the five variables. Both are given below, and a text file for the numerical expression values can be downloaded from the course website.

	C1	C2	C3	C4	C5	C6	C7
A	0.1	0.7	0.8	0.1	1	0.6	0.7
B	0.2	0.9	0	0.4	1	0	0.1
C	0.1	0.8	0.1	0.7	1	0.9	0.7
D	0	1	0.8	0.9	0.9	0	0.9
E	0	0.9	0	1	0.7	0.7	0.8

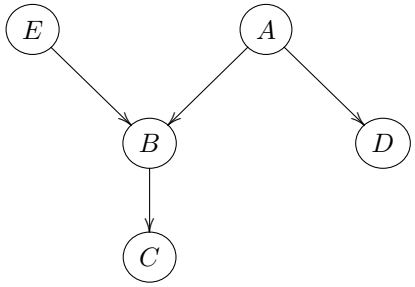

Figure 1: Expression profiles for five-gene example (conditions C1-C7, left) and model structure (right).

a) Specify the adjacency matrix for the graph that encodes the conditional dependencies in the Bayesian network model.

b) To determine the conditional probability tables for each node of the graph, first discretize the gene expression values in the table above by assuming that a gene is active (1) when its expression level is greater or equal to 0.5, and inactive (0) otherwise. Then, use conditional counting to determine the conditional probabilities, where $N(X_i^j, \mathbf{Pa}_i)$ is the number of joint observations of a state $i, X_i^j = x_j$ of a node, and of the state vector of its parents $\mathbf{Pa}_i = \mathbf{pa}_i$:

$$P(X_i^j|\mathbf{Pa}_i) \approx \frac{N(X_i^j, \mathbf{Pa}_i)}{\sum_k N(X_i^k, \mathbf{Pa}_i)}$$

Hint: You can count by hand or set up a small program that, starting from the adjacency matrix in part (a), determines the parents and the respective counts for all nodes.

c) To better understand what we can learn by observing the gene C, using Bayes' theorem, compute the posterior probabilities $P(A|C = 1)$ and $P(E|C = 1)$. Here, assume that the probabilities $P(A)$ and $P(E)$ derived from the expression data are suitable prior probabilities. Is A or E more likely to be active given that C is active?

d) Now assume that we do another set of experiments that results in the second set of data below. How do the results change when we incorporate all the available data?

	C8	C9	C10	C11	C12	C13	C14
A	0.2	0.9	0.9	1	0	0.9	0.8
B	0.1	1	0.7	0.8	0.8	0.7	0.6
C	0.3	0.7	0.7	0.6	0.8	0.8	0.8
D	0.4	0.7	1	0.9	1	0.8	0.7
E	1	0	0	0	0	0.2	0.1

How do $P(E|C = 1)$ change as a function of the cutoff for active genes (e.g., between 0.2 and 0.7)?

Notes & submission:

This exercise can be also be solved by hand, you do not necessarily need to use programming here, but it is helpful especially for part (d). Please, address questions to

alix.moawad@bsse.ethz.ch