LEARNING BAYESIAN NETWORKS FROM GENE EXPRESSION DATA

Jan Malinowski
Facculty of Physics
372202

INTRODUCTION

Bayesian network is one of the probabilistic models widely used in machine learning domain. Its operation is based on the hidden parameters, which are added to the network represented by directed acyclic graph. In this work, method implemented in the R 'Deal' package, was used. Using it, the learning process consists of three steps:

1. generating a prior structure of the network,
2. generating a prior distribution of the parameters of the local probability distribution and,
3. finding the optimal network structure using heuristic search.

Purpose of this work is to assess the uncertainty associated with network inference, which could be obtained by comparing the outcomes of the Bayesian network learning procedure based on the randomly perturbed data. Dataset used for this experiment consists of expression levels of 800 yeast's genes measured at different time points during the cell cycle.

DATA PREPARATION

Firstly, data was preprocessed in order to filter out the 10 most variant genes. Missing data was filled in replaced by the median of measurements for the specific gene. Then, the inter quartile range was calculated for every gene, and 10 genes, which range was greater than 1.6 were selected for further work.

LEARNING NETWORK FROM OBSERVED DATA

Prior structure for the network was manually created (Fig. 1), keeping the below conditions:

- YOL007C has an effect on YBR088C and,
- YNL327W has an effect on YER124C, YHR143W and YNR067C.

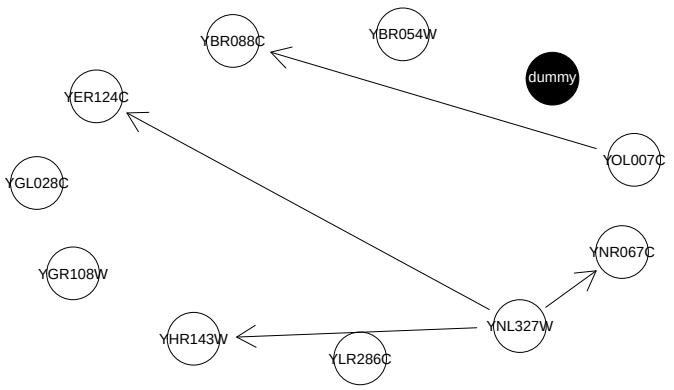

Figure 1: Prior structure of the network, with dummy node added due to the implementation dependencies.

Local probability distributions of gene YBR088C were shown in the table below.

Table 1. Local probability distributions of gene YBR088C

0.228734538	0.008498828	0.826678327

Using the previously created prior structure and generated distribution for the parameters of the joint distribution, initial Bayesian network was learned (its score was **-1101.865**). Subsequently, local search was performed in order to find the optimal network. Its structure was plotted in Fig. 2.

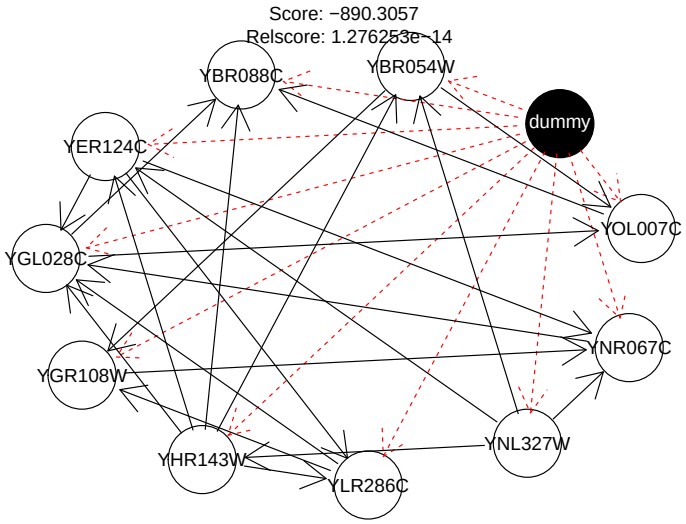

Figure 2: Structure of the BN*.

Score of the network improved to **890.3057**.

LEARNING NETWORK FROM PERTURBED DATA

Variances of selected 10 genes among the experiments were calculated and presented in the table below.

Table 2. Experiment results variances for each gene.

YBR054W	**YBR088C**	**YER124C**	**YGL028C**	**YGR108W**
1.4410599	0.9897524	1.4331767	1.4743219	0.9366332
YHR143W	**YLR286C**	**YNL327W**	**YNR067C**	**YOL007C**
1.0552875	2.0208231	1.4356611	1.6242167	1.1091768

Experimental values of each gene i was perturberd by adding a noise term distributed as $\mathrm{N}\left(0, \frac{\sigma_i^2}{10}\right)$ to each entry in the column corresponding to gene i. This procedure was repeated 30 times in order to generate 30 datasets. Plot of the experiment values distributions was presented in Fig. 3.

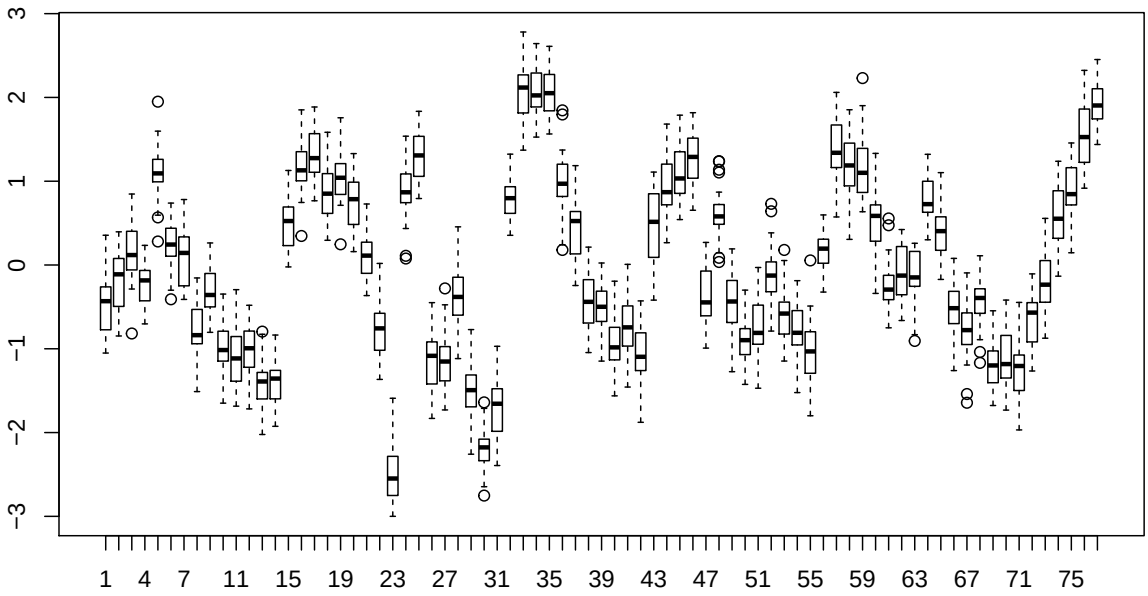

Figure 3: Box plot of the experiment values distribution of gene YHR143W. The horizontal axis displays the number of experiment and the vertical axis displays experiment value.

Optimal Bayesian networks were then found using such prepared data. Plot corresponding to the PBN5 network was shown in Fig. 4.

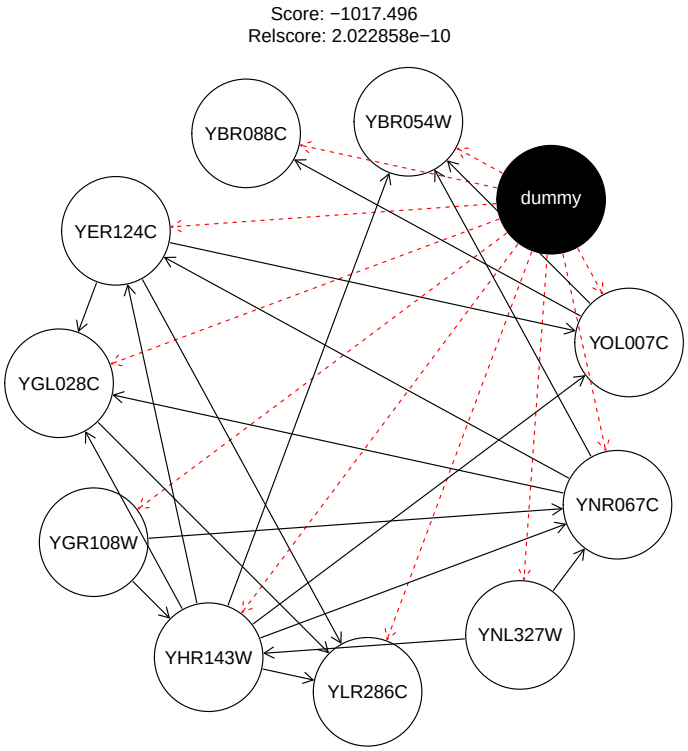

Figure 4: Structure of the PBN5.

ANALYSIS OF THE LEARNED NETWORKS EDGES

For each edge contained in BN*, its relative frequency among the PBNs was calculated. Plot of these frequencies was shown in Fig. 5. Edges, which relative frequency among all the

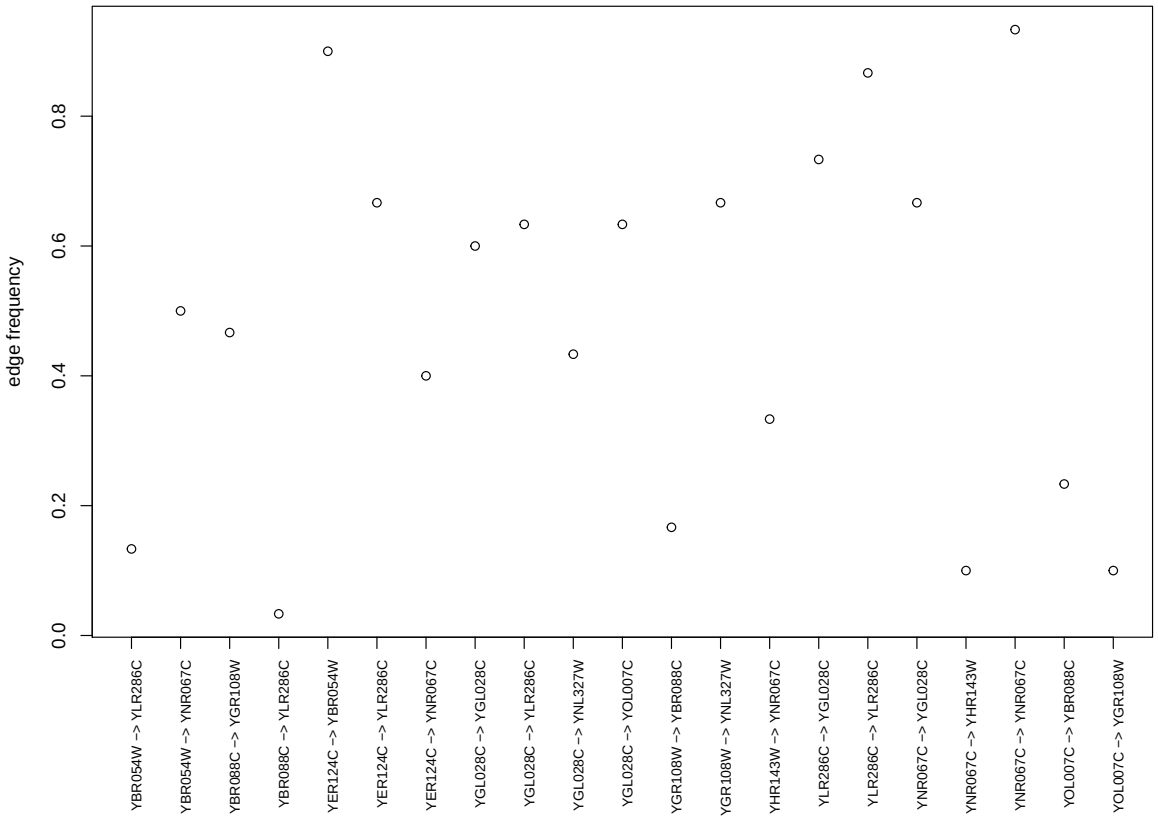

Figure 5: Plot of edge, present in BN*, relative frequencies among PBNs. The horizontal axis displays the edge frequency and the vertical axis displays.

PBNs were lower than the value corresponding to the 0.25 quantile, were considered spurious. This set consists of the subsequent edges:

- YBR054W → YLR286C,
- YBR088C → YLR286C,
- YGR108W → YBR088C,
- YNR067C → YHR143W,
- YOL007C → YGR108W.

Similar plot, for the edges, which were absent in the BN*, was shown in Fig. 6.

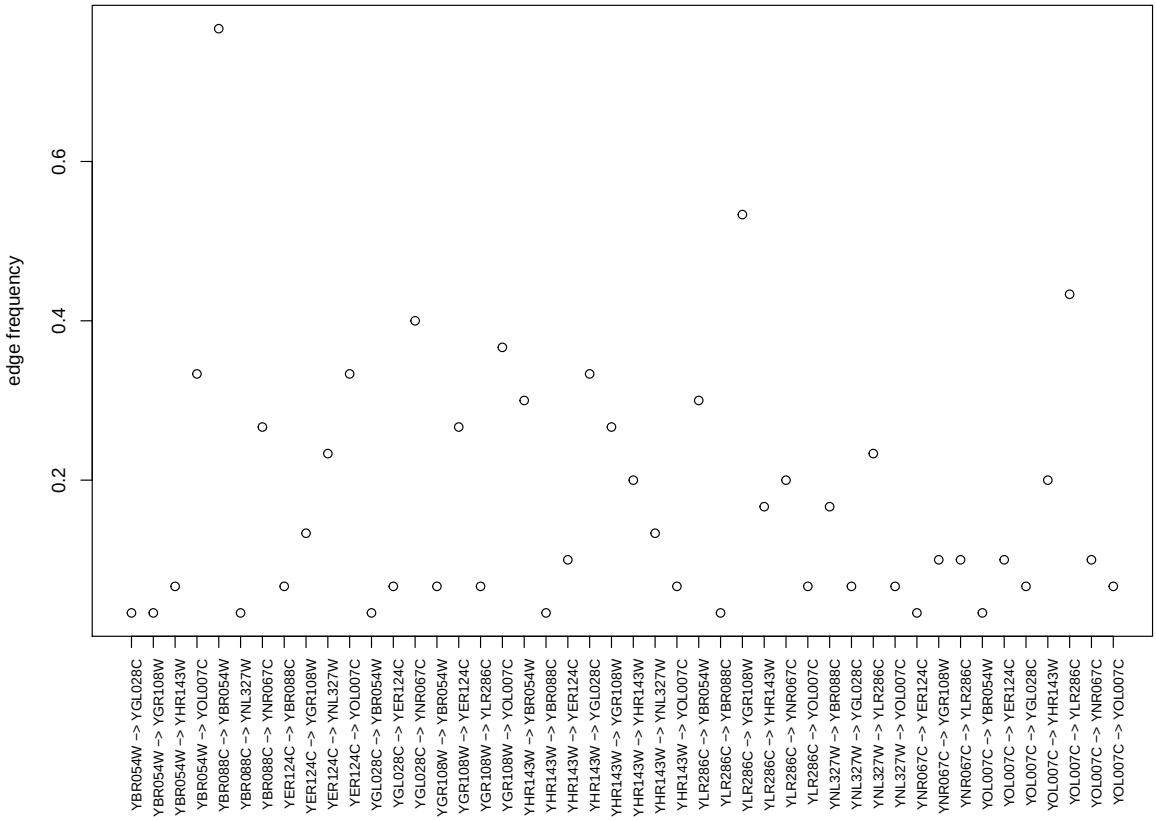

Figure 6: Plot of edge, absent in BN*, relative frequencies among PBNs. The horizontal axis displays the edge frequency and the vertical axis displays.

Edges, which relative frequency among all the PBNs were higher than the value corresponding to the 0.75 quantile, were considered missing in the optimal network based on the observed data. These edges were:

- YBR054W → YOL007C,
- YBR088C → YBR054W,
- YER124C → YOL007C,
- YGL028C → YNR067C,
- YGR108W → YOL007C,
- YHR143W → YBR054W,
- YHR143W → YGL028C,
- YLR286C → YBR054W,
- YLR286C → YGR108W,
- YOL007C → YLR286C.

This report shows the uncertainty associated with network inference from gene expression data.