The Use of Prior Distributions to Learn Genetic Networks

Olivier Gevaert[1], Steven Van Vooren[1], Bart De Moor[1]

Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Abstract. We have used simulated data to show that the use of a structure prior for reverse-engineering genetic networks with Bayesian network models can improve model selection. When using such a prior the number of errors between the selected model and the true model is lower and significantly different when using an uninformative prior. Therefore we introduced automatically generated priors based on pubmed abstracts and publicly available taxonomies and ontologies to be used in combination with real data. In the future these and similar priors can be used to construct more reliable models of genetic networks.

1 Introduction

Reverse engineering genetic networks has been a hot topic in bioinformatics for several years. An important issue within this area is that mostly the data is limited or restricted to model organisms. Therefore, in our opinion, the integration of other sources of information is very important to find reliable models that can explain the data. Probabilistic models provide a natural solution to this problem since information can be incorporated in the prior distribution over the model space. This prior is then combined with the data to form a posterior distribution over the model space which is a balance between the information incorporated in the prior and the data. We investigated the use of prior information on the structure of a genetic network in combination with Bayesian network learning on simulated data and we suggest possible priors for real data. Bayesian networks and extensions of Bayesian networks are popular models for reverse engineering genetic networks ([5, 15, 9, 10]). However in most cases the data is limited and the integration of other sources of information can improve results. For other types of data we have already proven that structure prior information improves model selection especially when few data is available ([6, 1]).

2 Bayesian networks

A Bayesian network is a probabilistic model that consists of two parts: a dependency structure and local probability models ([12, 11]). The dependency structure specifies how the variables are related to each other by drawing directed

edges between the variables without creating directed cycles. Each variable depends on a possibly empty set of other variables which are called the parents:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i | Pa(x_i)) \tag{1}$$

where $Pa(x_i)$ are the parents of x_i. Usually the number of parents for each variable is small therefore a Bayesian network is a sparse way of writing down a joint probability distribution. The second part of this model, the local probability models, specifies how the variables depend on their parents. We used discrete-valued Bayesian networks which means that these local probability models can be represented with Conditional Probability Tables (CPTs).

2.1 Structure learning

The most important step when learning a Bayesian network is finding the dependency structure that best explains the data. This is done using a scoring metric combined with a search strategy. The scoring metric describes the probability of the structure S given the data, D. When there are n variables $x_1, ..., x_i, ..., x_n$ with r_i the number of values of each variable and q_i the number of instantiations of the parents of each variable than the scoring metric is defined as:

$$p(S|D) \propto p(S) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \left[\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})} \right], \tag{2}$$

for details see [3, 8, 7]. Equation 2 allows to score structures combined with a search strategy to find a good model. An exhaustive search is infeasible since the number of structures becomes intractably large when there are much variables. Therefore we used the greedy search algorithm K2 [3] to build a Bayesian network structure. The next step consists of estimating the parameters of the local probability models for the selected model. This amounts to filling in a CPT for every variable and every possible value of its parents using the data.

2.2 Structure prior

$p(S)$ in equation 2 is the prior probability of the structure and is calculated by iterating over all the variables and each time multiplying the probability that there is an edge between the parents of a variable x_i and, the probability there is no edge between the other variables and x_i:

$$p(S) = \prod_{i=1}^{n} \prod_{l=1}^{p_i} p(Pa_l(x_i) \rightarrow x_i) \prod_{m=1}^{o_i} p(NonPa_m(x_i) \nrightarrow x_i) \tag{3}$$

with n the number of variables, Pa_l the l-th parent of x_i and p_i the number of parents of variable x_i. $NonPa_.(x_i)$ is then the set of variables which are not a parent of x_i with $NonPa_m(x_i)$ the m-th variable in this list and o_i the number

of variables that are not a parent of x_i. Next, $p(a \rightarrow b)$ is the probability that there is an edge from a to b while $p(a \nrightarrow b)$ is the inverse, i.e. the probability that there is no edge from a to b. This means that we have to specify a probability for each directed edge between all combinations of two variables in the data set. Suppose there are three variables then there are six possible edges and therefore six probabilities that have to be specified (between two variables there are two possible edges, one in each direction). After assessing these prior probabilities, this results in a matrix that specifies the probability that a directed edge occurs between any combination of two variables. The prior probability of a structure, $p(S)$ seems a good candidate to capture prior information and improve model selection.

3 Simulated data

We used simulated data by generating random discrete-valued Bayesian networks. This was done by creating networks consisting of ten variables with three states. Data sampled from these networks was then used to re-discover the original network with or without a structure prior. We investigated if a noisy structure prior, as defined earlier, could improve the model selection by comparing the selected model with the real model. Then we counted the sum of the number of missing edges, superfluous edges and reversed edges in the selected model compared to the real model. Figure 3 shows that using a structure prior reduces the number of errors made in the small sample range for ten variable-networks.

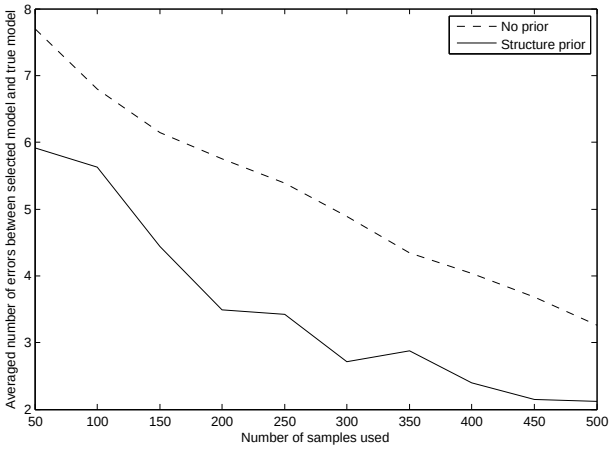

Fig. 1. This figure shows that a structure prior reduces the number of errors made in the small sample range.

4 Prior information

Since microarray data consists of thousands of genes, it is infeasible to construct a structure prior manually. Therefore methods based on automatic elicitation of a relationship between genes have to be used. Therefore we propose the use of text based gene-by-gene-similarity matrices as priors. Genes are represented in the Vector Space Model ([14]), where each position of a gene vector corresponds to a term or phrase in a fixed vocabulary. Vocabulary are based on publicly available taxonomies and ontologies such as MeSH, OMIM and Gene Ontology ([2]). For each gene, manually curated literature references are extracted from Entrez Gene. All PUBMED abstracts linked to genes are indexed along the vocabularies mentioned above. This procedure involves stemming ([4]) and stop word removal. As a result, all PUBMED abstracts are represented in a high dimensional vector space using IDF weights for non-zero vector positions. All vectors are normalised. Gene vectors are then constructed by averaging the vectors of all the abstracts referenced to that gene in Entrez Gene. The cosine measure is used to obtain gene-to-gene distances using normalized gene vectors.

5 Discussion

We showed that a noisy structure prior improves the model selection when using simulated data in a simple example. This example was used to illustrate that even when there is noise in the structure prior, it can improve model selection. Le Phillip et al performed a similar study by using prior edges and also found a beneficial effect ([13]). The way that our structure prior is defined however allows that it can be constructed automatically using publicly available data sources. Therefore we proposed priors based on co-ocurrence of genes in PUBMED abstracts. Moreover other sources of information are currently being investigated (e.g. pathway information) and can be combined in a single prior. We have applied these priors on real data and are currenlty developing methods to evaluate the results. Furthermore we will also combine structure priors with our method to integrate clinical and microarray data ([7]).

References

1. P Antal, G Fannes, D Timmerman, Y Moreau, and B De Moor. Using literature and data to learn bayesian networks as clinical models of ovarian tumours. *Artif Intell Med*, 30:257–81, 2004.
2. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34(Database issue):322–326, Jan 2006.
3. G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309 –347, 1992.
4. MF Porter et al. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
5. N Friedman, M Linial, I Nachman, and D Pe'er. Using bayesian networks to analyze expression data. *J Comput Biol*, 7:601–20, 2000.

6. O Gevaert, F De Smet, , E Kirk, B Van Calster, T Bourne, S Van Huffel, Y Moreau, D Timmerman, B De Moor, and G Condous. Predicting the outcome of pregnancies of unknown location: Bayesian networks with expert prior information compared to logistic regression. *Human reproduction*, 2006.
7. O Gevaert, F De Smet, D Timmerman, Y Moreau, and B De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Accepted at the 14th Annual International conference on Intelligent Systems for Molecular Biology (ISMB 2006)*, 2006.
8. D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.
9. D Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic bayesian networks. *Bioinformatics*, 19:2271–2282, 2003.
10. S Imoto, S Kim, T Goto, S Miyano, S Aburatani, K Tashiro, and S Kuhara. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J Bioinform Comput Biol*, 1:231–52, 2003.
11. R.E. Neapolitan. *Learning Bayesian networks*. Prentice Hall, Upper Saddle River, NJ, 2004.
12. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Matteo, California, 1988.
13. P Le Phillip, A Bahl, and LH Ungar. Using prior knowledge to improve genetic network reconstruction from microarray data. *In silico biology*, 4:335–53, 2004.
14. G Salton, A Wong, and CS Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.
15. E Segal, M Shapira, A Regev, D Pe'er, D Botstein, D Koller, and N Friedman. Module networks: indentifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34:166–76, 2003.