A framework for elucidating regulatory networks from prior information and expression data

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

Katholieke Universiteit Leuven, Dept. Electrical Engineering, ESAT-SCD,
Kasteelpark Arenberg 10, 3001 Leuven, Belgium

Abstract. Elucidating regulatory networks is an intensively studied topic in bioinformatics. Integration of different sources of information could facilitate this task. We propose to incorporate these information sources in the structure prior of a Bayesian network. We are currently investigating two complimentary sources of information: pubmed abstracts combined with publicly available taxonomies or ontologies, and known protein-DNA interactions. These priors, either separately or combined, have the potential of reducing the complexity of reverse engineering regulatory networks while creating more robust and reliable models. Moreover this approach can easily be extended with other data sources. In such a way Bayesian networks provide a powerful framework for data integration and regulatory network modeling.

1 Introduction

Reverse-engineering genetic networks from expression data has been studied extensively for several years now. The biggest challenge is that the limited data set size leads to an underdetermined problem. Therefore, in our opinion, the integration of other sources of information could provide the needed extra information which alleviates this issue. Publicly available data sources are potentially complimentary to expression data and they possibly offer views from other angles at the same problem. Moreover by using this information, one can avoid estimating features that are already known and concentrate on new interactions. 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. In this contribution we would like to present a framework based on Bayesian networks together with two sources of prior information that we are currently investigating. Pubmed abstracts combined with publicly available taxonomies or ontologies and known protein-DNA interactions. For both sources of information we propose a method to incorporate them in the structure prior of a Bayesian network. Bayesian networks and extensions of Bayesian networks are already popular models for reverse engineering regulatory networks ([1–4]). Therefore the presented work can easily be incorporated in these models. Our approach was motivated by the succesfull integration of information for other types of data where we have proven that structure prior information improves model selection especially when few data is available ([5, 6]).

2 Bayesian networks

A Bayesian network is a probabilistic model that consists of two parts: a dependency structure and local probability models. 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 Model Building

Previously we mentioned that a discrete valued Bayesian network consists of two parts: the structure and the local probability models. Consequently, there are two steps to be performed during model building: structure learning and learning the parameters of the CPTs. First the structure is learned using a search strategy. Since the number of possible structures increases super-exponentially with the number of variables, we used the well-known greedy search algorithm K2[7] in combination with the Bayesian Dirichlet (BD) scoring metric for details see[7–9]:

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

The next step consists of estimating the parameters of the local probability models for the structure with the highest BD score. This amounts to filling in a CPT for every variable and every possible value of its parents using the data. This Bayesian network can then be used to predict future data.

2.2 Structure prior

At this moment we have not yet defined $p(S)$ in Equation 2, $p(S)$ is the prior probability of the structure. When assuming that being a parent of some node is independent of any other parental relationship, the prior can be decomposed according to the variables:

$$p(S) = \prod_{i=1}^{n} p(Pa(x_i) \rightarrow (x_i)) \tag{3}$$

The probability of a local structure $p(Pa(x_i) \rightarrow (x_i))$ is then calculated by multiplying the probability that there is an edge between the parents of x_i and, the probability there is no edge between the other variables and x_i:

$$p(Pa(x_i) \rightarrow x_i) = \prod_{p \in Pa_(x_i)} p(p \rightarrow x_i) \prod_{y \notin Pa_(x_i)} p(y \nrightarrow x_i) \tag{4}$$

3 Sources of Prior information

Since the number of genes we would like to model is quite large, it is infeasible to construct a structure prior manually. Therefore methods based on automatic elicitation of relationships between genes have to be used. Currently we propose two complementary sources of prior information that can be combined with expression data to reverse-engineer regulatory networks: known protein-DNA interactions and gene-by-gene similarities based on pubmed abstracts.

3.1 Protein-DNA interaction prior

A first source of information is the incorporation of known protein-DNA interactions in the structure prior. Such interactions can be mined from publicly available databases such as the Biomolecular Interaction Network Database[10], Transfac[11] and IntAct[12]. These interactions can be incorporated in the structure prior by setting the $p(R \rightarrow G) = 1$ and $p(G \rightarrow R) = 0$ with R the regulator (i.e. protein) and G the regulated gene. All other entries in the structure prior are then given a constant probability. Currently we have parsed the Biomolecular Interaction Network Database for known protein-DNA interactions in human and incorporated this information into a structure prior. Subsequently we combined this structure prior with Bayesian network learning for modelling (in)active regulatory mechanisms in different types of cancer. We could show that when using the prior with known protein-DNA interactions, simpler and more robust models where built as compared to not using the prior[13].

3.2 Text based prior

Next, we propose the use of text based gene-by-gene-similarity matrices as priors. Genes are represented in the Vector Space Model, where each position of a gene vector corresponds to a term or phrase in a fixed vocabulary. The vocabulary is based on the publicly available National Cancer Institute Thesaurus ([14]). For each gene, manually curated literature references are extracted from Entrez Gene. All PUBMED abstracts linked to genes are indexed along the vocabulary mentioned above. This procedure involves stemming 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.

4 Discussion

We presented a framework based on Bayesian networks in combination with structure prior information for reverse-engineering regulatory networks and provided two complimentary sources of information that can be integrated in this structure prior. Firstly, we proposed the use of known protein-DNA interactions. This prior information ensures that the model building effort is invested in learning new features instead known relations. In addition, the quality of this prior is high because it is based on experimentally verified interactions but, since there are only a limited number of interactions known, this prior is very sparse. Secondly, we defined a method of mining gene-by-gene similarities from pubmed abstracts. This approach provides information for a larger number of genes but in the process also contains more noise. Hence both priors offer possibly complimentary information. We already succesfully applied the first source of prior information and could show that simpler, more robust models where built when using known protein-DNA interactions ([13]). A mayor challenge however is the validation of regulatory network models. There is still an urgent need for methods that can objectively evaluate regulatory network models and their algorithms. Finally we will also apply the presented method in a classification setting in combination with our method to integrate clinical and microarray data ([8]).

References

1. N. Friedman, M. Linial, I. Nachman and D. Pe'er, *J Comput Biol* **7**, 601 (2000).
2. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, *Nature Genetics* **34**, 166 (2003).
3. D. Husmeier, *Bioinformatics* **19**, 2271 (2003).
4. S. Imoto, S. Kim, T. Goto, S. Miyano, S. Aburatani, K. Tashiro and S. Kuhara, *J Bioinform Comput Biol* **1**, 231 (2003).
5. O. Gevaert, F. D. Smet, E. Kirk, B. V. Calster, T. Bourne, S. V. Huffel, Y. Moreau, D. Timmerman, B. D. Moor and G. Condous, *Human reproduction* **21** (2006).
6. P. Antal, G. Fannes, D. Timmerman, Y. Moreau and B. D. Moor, *Artif Intell Med* **30**, 257 (2004).
7. G. Cooper and E. Herskovits, *Machine Learning* **9**, 309 (1992).
8. O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau and B. D. Moor, *Proceedings of the 14th Annual ISMB, Bioinformatics special issue* (2006).
9. D. Heckerman, D. Geiger and D. Chickering, *Machine Learning* **20**, 197 (1995).
10. C. Alfarano *et al.*, *Nucleic Acids Research* **1**, D418 (2005).
11. V. Mathys *et al.*, *Nucleic Acids Research* **1**, D108 (2006).
12. H. Hermjakob *et al.*, *Nucl. Acids Res.* **32**, D452 (2004).
13. O. Gevaert, F. D. Smet, K. Marchal and B. D. Moor, *Submitted to PSB* (2007).
14. N. C. I. Thesaurus, http://www.mindswap.org/2003/CancerOntology/.