Reconstruction of gene regulation networks from microarray data by Bayesian networks

Hoai-Tuong NGUYEN[1], Gérard RAMSTEIN[1], Philippe LERAY[1], Yannick JACQUES[2]

[1] LINA - Laboratoire d'informatique de Nantes Atlantique, Nantes, France
La Chantrerie - rue Christian Pauc, BP 50609, 44306, Nantes Cedex 3, France
hoai-tuong.nguyen;gerard.ramstein;philippe.leray@univ-nantes.fr
[2] CRCNA, Centre de Recherche en Cancérologie Nantes/Angers, UMR 829 INSERM
9 quai Moncousu 44093 Nantes Cedex 01, France
yjacques@nantes.inserm.fr

Abstract: *In this work, we reconstruct the gene regulation networks from the microarray experiments data by Bayesian networks approach. We use the evolutionary algorithm for the search-and-score based structure learning methods. The learned network is tested by the hypothesis testing with two populations of patient data, one with treatment (drugs), other without treatment. The answer of question "How does the treatment influence to gene regulation?" is expected.*

Keywords: Gene expression, microarray, gene regulation networks, Bayesian networks, genetic algorithm, estimation of distribution algorithm.

1 Introduction

The inference of gene regulatory networks from high-throughput microarray data is a central problem of biological research. There are various machine learning and statistical methods have been proposed to reconstruct more effectively this kind of networks, such as, clustering [4], Bayesian Networks (BNs) [7], [3], [15], Graphical Gaussian Models [13]. Compared to others, BNs can solve more effectively almost principle problems of this reconstruction: (1) *The complex interactions* involving many genes usually have to be inferred from sparse and noisy data; (2) There are *a massive number of variables* (over 30.000 genes), but *a small number of samples* (dozens experiments); (3) *Computational complexity of structures and statistical significance* betweens variables in learned networks. In the other words, to have a "better" gene regulation network, we have to construct a "better" BN from microarray data. This is a question of BNs learning (consist of parameter and structure learning) from high-throughput microarray experiments data. It's the main goal of this work.

2 Methods

In the present work, we investigate an implementation of two BNs structure learning methods by the genetic algorithm and by the estimation of distribution algorithm to ProBT®, a general-purpose development toolkit for Bayesian modelling, inference, and learning, developed by ProBayes. Then, we use a hypothesis testing with two populations of patient data, one with treatment (drugs), other without treatment, to test the result of the best networks produced by the learning methods in the first step. The answer of question "How does the treatment influence to gene regulation?" is expected.

BNs are directed graphical models for representing probabilistic independence relations between multiple interacting entities. Formally, BNs are directed acyclic graphs (DAGs - a network without any directed cycles) modelling probabilistic dependencies among variables [2]. The graphical structure G of a BN consists of a set of nodes and a set of directed edges.

In the study of reconstruction of gene regulation networks, we use a gene to represent a node and direct influence/interaction between genes to represent an edge. If there is an edge from node A to another node B, then variable B depends directly on variable A (gene A regulates gene B), and A is called a parent of B. In a BN every variable is conditionally independent of its non-descendants given its parents (Markov condition). In the other words, the conditional distribution of a variable A given its parents pa_A in the graph G is $P = P(A|pa_A)$ (parameter of BN, Figure 1). With this simple condition, can infer how well a particular network explains the observed data. For example, in the BN below, the joint distribution decomposes nicely:

$$P(G_1, G_2, G_3, G_4, G_5, G_6) = P(G_1).P(G_3).P(G_2|G_1).P(G_4|G_2).P(G_5|G_2, G_3)$$

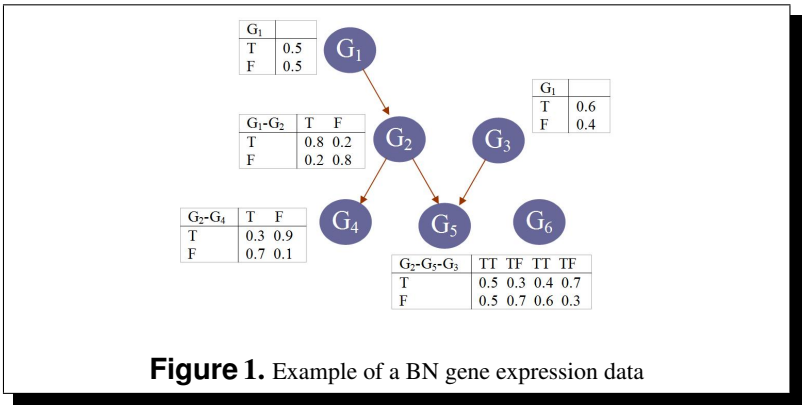

Figure 1. Example of a BN gene expression data

In the simplest case, a BN is specified by an expert and then, it is used to perform inference. However, the task of defining the network is too complex for humans. So, the network structure and the parameters of the local distributions must be learned from data. We call this task is BNs learning.

Learning a BN from data requires both identifying the model structure G (*structure learning*) and identifying the corresponding set of model parameter values (*parameter learning*). More simply, given a fixed structure, however, it is straightforward to estimate the parameter values.

To learn the BNs parameter, the common approach is to introduce a statistically motivated scoring function that evaluates each network with respect to the training data, then search for the optimal network according to this score. The most used score is BIC (Bayesian Information Criterion).

To learn the BNs structure, there are two types of methods: (1) *Constraint-based methods* search a database for conditional independence relations and then, construct graphical structures called ”patterns” which represent a class of statistically indistinguishable directed DAGs; (2) *Search-and-score methods* perform a search in the space of legal structures. Search-Scoring methods have the advantage of being able to flexibly incorporate prior knowledge and dealing with incomplete data [6]. GA, EDA are the Evolutionary Algorithms that are used as a effect heuristic search engine in the BNs structure learning problem [14], [1]. After the best structure, in order to know its real biology performance, we propose to use a hypothesis testing with two populations of patient data, one with treatment (drugs), and other without treatment. Depending on the difference of the result of this test, we can conclude the influence of the treatment on the regulation of the genes.

3 Discussion

Which BNs learning algorithms for inferring gene regulatory networks? Although in 2007, Nancy Cartwright [2] devotes half of her new book, "Hunting Causes and Using Them", to critcizing "Bayes Net Methods"–as she calls them–and what she takes to be their assumptions. All of her critical claims are false or at best fractionally true. And in a recent work, [12] applying the various search methods to real microarray data from an independently known gene expression regulatory network confirms their failure. But, various researches still concentrate motivationally on this problem [8], [10], [9], [11], [1], [5]. For each work, the authors propose their own effective methods to improve the accuracy of the inference of gene regulation networks for a specific type of microarray experiments data. Especially, we are interested in the work of C.Auliac [1] thesis that described perspectively an interesting advantage of BNs structure learning by the evolutionary algorithm. The reconstruction of the gene regulation networks by Bayesian networks will be continuously developed with the bioinformatics research.

Acknowledgements

I would like to express my thanks to the organization committee of ModGraph satellite day of JOBIM 2009 for permission to refer to this submission. I am very grateful for the cooperation and the instruction of my tutors **Mr. Gérard Ramstein**, Associate Professor of University of Nantes, **Mr. Philippe Leray**, Professor of University of Nantes and **Mr. Yannick Jacques**, Director of Research at INSERM 829, who helped me develop the ideas put forward here. I would like to acknowledge the helpful comments of my colleagues at COD-LINA. This research is supported by the BIL project, a regional project of the region Pays de Loire, France.

References

[1] C. Auliac, V. Frouin, and F.D. Alché-Buc. Approches évolutionnaires pour la reconstruction de réseaux de régulation génétique par apprentissage de réseaux bayésiens. *PhD Thesis*, 2008. 2, 3

[2] N. Cartwright. Hunting causes and using them : approaches in philosophy and economics. *ISBN 0521860814*, 2007. 3

[3] M. Dejori. Analyzing gene expression data with bayesian networks. *PhD thesis*, 2002. 1

[4] Z. Dongxiao, O. H. Alfred, C. Hong, K. Ritu, and Anand S. Network constrained clustering for gene microarray data. *Bioinformatics*, 2005. 1

[5] S.F. Emmert and M. Dehmer. Analysis of microarray data: A network-based approach. 2008. 3

[6] O. François and P. Leray. Réseaux bayésiens: de l'identification de structure à la reconnaissance des formes à partir d'informations complètes ou incomplètes. *PhD thesis*, 2006. 2

[7] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Computer Biology 7(3-4)*, pages 601–620, 2000. 1

[8] F. Geier, T. Jens, and F. Christian. Reconstructing gene-regulatory networks from time series knock-out data, and prior knowledge. *BMC Systems Biology, 1(1):11*, 2007. 3

[9] Y. Huang, J. Wang, Zhang J., Sanchez M., and Y. Wang. Bayesian inference of genetic regulatorynetworks from time series microarray data using dynamic bayesian networks. *Bioinformatics, 2:46-56*, 2007. 3

[10] P. Li, Z. Chaoyang, P. Edward, G. Ping, and Youping D. Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatory networks. *BMC Bioinformatics, 8(Suppl 7):S13*, 2007. 3

[11] F. Markowetz. A bibliography on learning causal networks of gene interactions. pages 349–356, 2007. 3

[12] J. Niemi. Accuracy of the bayesian network algorithms for inferring gene regulatory networks. *http://www.sal.tkk.fi/Opinnot/Mat-2.108/pdf-files/enie07.pdf*, 2007. 3

[13] J. Schferand and K. Strimmer. Learning large-scale graphical gaussian models from genomic data. *J. F. Mendes. (Ed.). Proceedings of CNET*, 2005. 1

[14] G. Thibault, S. Bonnevay, and A. Aussem. Learning bayesian network structures by estimation of distribution algorithms: An experimental analysis. *IEEE International Conference on Digital Information Management (ICDIM 07), Lyon, France*, 2007. 2

[15] L. Tiefei. Learning gene network using bayesian network framework. *PhD thesis*, 2005. 1