Gene regulatory network reconstruction with prior knowledge over mRNA data for COPD patients and controls

Bernal Arzola, Victor; Brandsma, C.; Faiz, Alen; Guryev, Victor; Timens, Willem; van den Berge, Maarten; Bischoff, Rainer; Grzegorczyk, Marco; Horvatovich, Peter

Published in:
Proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

Gene regulatory network reconstruction with prior knowledge over mRNA data for COPD patients and controls

Victor Bernal [12], Corry A. Brandsma[3], Alen Faiz[5], Victor Guryev[4], Wim Timens[3], Maarten van den Berge[5], Rainer Bischoff[2], Marco Grzegorczyk [1], Peter Horvatovich[2]

[1] Johann Bernoulli Institute (JBI), Groningen Rijksuniversiteit, Groningen, NL.
[2] Department of Pharmacy, Analytical Biochemistry, Rijksuniversiteit Groningen, NL.
[3] Universitair Medisch Centrum Groningen (UMCG), Department of Pathology and Medical Biology, Rijksuniversiteit Groningen, NL.
[4] Universitair Medisch Centrum Groningen (UMCG), ERIBA, Rijksuniversiteit Groningen, NL.
[5] Universitair Medisch Centrum Groningen (UMCG), Department of Pulmonary Diseases, Rijksuniversiteit Groningen, NL.

E-mail for correspondence: v.a.bernal.arzola@rug.nl

Abstract: Chronic obstructive pulmonary disease (COPD) is a type of lung disease characterized by persistent bronchitis and emphysema. Current therapy is restricted to alleviate lung tissue inflammation, but is not able to stabilize or improve lung function of patients making necessary to understand the underlying molecular mechanisms of COPD. Genome-wide gene expression of lung tissue provides a powerful tool to elucidate molecular mechanism of COPD patients. In particular, Bayesian Networks (BNs) have been applied to infer genetic regulatory interactions from microarray gene expression data. In this study we aim obtain a clearer understanding of the genes interaction in COPD patients by learning a BN over microarray expression data. A subset of genes was selected for the study fulfilling that i) the genes were significantly expressed in COPD stage 4 and ii) there is reported gene-gene experimental association. The reported associations are introduced as prior biological knowledge in the reconstruction.

Keywords: COPD; Bayesian Networks; Microarray Gene Expression; Introduction of prior knowledge.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

1 Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by distinct phenotypes as, emphysema, chronic bronchitis and fibrosis. The exact nature of the abnormal lung tissue repair processes in COPD lungs is unknown, hence a better understanding the underlying molecular mechanisms is thus crucial. Bayesian Networks (BNs) have been applied to infer genetic regulatory interactions from microarray gene expression data. This inference problem is particularly hard because of the relatively small size of the data sets. Moreover the asessment of the inference for COPD results is unfeasible as there is a lack of known gold standards from the biological literature. The aim of the present study is the reconstruction of a regulatory network from a microarray gene expression data set for COPD (stage 4) patients using BNs with prior biological knowledge coming from STRING (database of known and predicted protein-protein interactions). Inference is done using with Markov Chain Monte Carlo (MCMC) sampling following the approach of Werhli et al. (2007). Two networks are obtained and compared, one for the controls and one for the COPD data.

2 Bayesian Networks with prior knowledge

Bayesian Networks (BNs) represent probabilistic relationships between interacting agents (e.g. genes) in the form of conditional independence. For BNs the model is a Directed Acyclic Graph (DAG), a set of nodes that are connected by (directed) edges without cycles. We denote by G the network structure, which is the set of nodes and edges, and by q the interaction strength. Our objective is to learn the network structure G directly from the scratch microarray data D. Nodes are associated with genes and the edges indicate interactions between the genes. A Bayesian learning approach uses the posterior $P(G|D)$

$$P(G|D) = P(G|D)P(G)/P(D)$$

where $P(G)$ is called the prior over graphs, $P(D|G)$ the marginal likelihood, $P(D)$ is a scaling constant and $P(G|D)$ is called the posterior probability. The term $P(D|G)$ is the result of integrating the likelihood $P(D, q|G)$ over all the possible values of the parameters q. Calculating $P(G|D)$ for all possible network structure G is computationally impossible due to $P(D)$ is intractable. Gene expression data implies a large number of genes with only a relative small number of measurements. As a consequence the posterior $P(G|D)$ will not have a clear maximum G^* . A solution is to use a MCMC sampling based on Metropolis Hasting algorithm (MH). This produces a Markov Chain that reaches $P(G|D)$ as its stationary distribution (under some conditions). Generally different DAGs can represent the same probabilistic relationships (i.e they can be equivalent). The set of equivalent DAGs is called the equivalence class and is represented by a CPDAG.

Introducing prior knowledge for learning BNs is done following Werhli et al. (2007). The prior knowledge is encoded in a matrix B whose entries $B_{ij} \in [0, 1]$ represent the belief about the presence/absence of an edge. The elements B_{ij} greater than/less than 0.5 reflects evidence in favor/against the presence of the edge G_{ij}, and $B_{ij} = 0.5$ as uninformative. The prior over graphs takes the form of a Gibbs distribution,

$$P(G|\beta) = e^{-\beta E(G)}/Z(\beta)$$

The function $E(G)$ measures the agreement between a sampled network G and B (e.g. using the Hamming distance). The parameter β indicates the weight of the respective source of knowledge to the data, and $Z(\beta)$ is a normalization constant.

3 Evaluation of the COPD dataset and Implementation

The data consist of genome-wide gene expression profiling in lung tissue samples of 72 subjects, where 48 exhibited COPD (stage 4). The others 24 are non-COPD tissues (Controls). The 60 most significant genes based on p-value and $log_2(FC) > 2$ were checked in STRING data base, obtaining that 7 genes (FGG, RORC, FGA, OSMR, NR1D1, CSF3, THBS1) had reported experimental and/or data base associations. This set of genes is augmented as follows: i) the Pearson's correlation matrix ρ is built for Controls and COPD, and ii) a pair of genes is selected if $|\rho_{controls}| < 0.25$ and $|\rho_{COPD}| > 0.75$. In this way 8 additional genes are obtained, namely MT1M, IL1RL1, MT1P3, MT1A, SLCO4A, GPA33, TTN, ZBED2. In total 2 networks were reconstructed for the forementioned genes i) Controls with prior knowledge, and ii) COPD with prior knowledge. The coupling parameter β is defined in $[0, 30]$. The MCMC was implemented with the REV move from Grzegorczyk (2008) for number of $9x10^5$ iterations, with thining every 1000 DAGs. The first half of the sample was discarded as part of the burn in phase and Model Averaging was performed over the final set of 501 CPDAGs.

4 Results and Conclusions

After the burn in phase the 75th percentile of β (in Controls and COPD cases) is less than 1. This suggest that COPD's gene regulation has no apparent commonalities with previously reported interaction. Table 1 shows a summary of the interactions that considerably changed between Controls and COPD. In particular interactions like FGG-FGA, and MT1P3-MT1A are reasonable as they belong to the same family. On the other hand NR1D1 has no a well known association with fibrogenes, this interaction (FGG-NR1D1) is present in Controls (mediated by FGA) and COPD (directly). Our next research plan is to extend the gene set. We will provide more precise biological interpretations in the presentation.

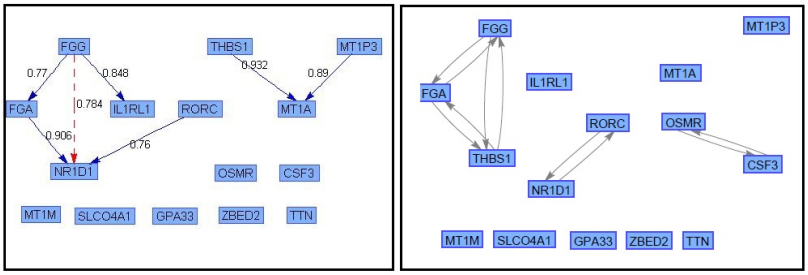

FIGURE 1. **Reconstructed regulatory network for Control and COPD.** On the right: Network from STRING with experimental and database associations. Doble arrows stand for reported associations. On the left: Reconstructed network. The blue edges stand for gene interactions in the Controls. The segmented red edges stand for gene interactions in the COPD. There are no common edges. Only edges with a frequency above 0.75 are shown.

TABLE 1. **Marginal edge frequency changes** This table presents a set of relevant gene-gene interaction whose edge frequency had a notable change between the network for the Controls and the one for COPD.

	Controls	COPD
FGG-FGA	0.770	0.314
FGA-NR1D1	0.906	0.078
MT1P3-MT1A	0.890	0.614
THBS1-MT1A	0.932	0.502
RORC-NR1D1	0.760	0.114

References

Grzegorczyk, M., Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, **71**, 265 – 305.

Werhli A., Husmeier, D. (2007). Reconstructing Gene Regulatory Networks with Bayesian Networks by Combining Expression Data with Multiple Sources of Prior Knowledge.*Stat Appl Genet Mol Biol*, **6**(1), 1 – 45.