Using Bayesian Networks to learn Gene Regulatory Networks

Nikolas Bernaola, Mario Michelis, Concha Bielza, Pedro Larranaga

Computational Intelligence Group, Departamento de Inteligencia Artificial, Universidad Politecnica de Madrid

Summary

We present a variant of the Fast Greedy Equivalence Search algorithm that can be used to learn a Bayesian network that represents the full transcriptional regulatory network of the human brain. We have fully implemented the algorithm and have some preliminary results that show that we can retrieve the Markov Blanket of individual genes of interest with good accuracy and reasonable time (2 hours in a 12 core 2.6 GHz computer). The algorithm is parallel and we are currently waiting to deploy it in a supercomputer where we expect to be able to learn the full genome network. This network could then be used as an exploratory tool by the biology community when studying the relationships between genes.

Transcriptional Regulatory Networks

We consider the regulatory network of all genetic expression in the human brain. Although this network will have some non-genetic components (i.e. hormonal regulation of gene expression or responses to chemical changes in the blood) most of this processes will be at some point regulated genetically. This suggests that we can approximate the true regulatory network of gene expression with a transcriptional one that only takes into account the concentration of mRNA in the cell for every gene in the human genome at a point in time. Using the data from the Allen Human Brain Atlas we can use Machine Learning methods to try to learn the underlying regulatory network.

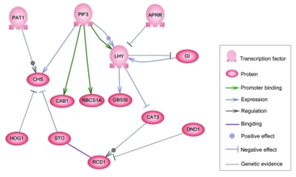

Fig. 1: Example of a small regulatory network in rice

Method

Due to the scale of the problem, with thousands of genes even in small genomes, most work on finding regulatory networks uses undirected correlation or mutual information networks. These methods, while faster, cannot be directly interpreted since due to the structure of the transcriptional network most gene expression is correlated even when two genes are completely causally independent. We decided to use Bayesian Networks since they faithfully represent the underlying independencies and can be used to predict the effects of varying one of the genes on the rest of the network. Learning the structure and parameters of a bayesian network is an NP problem and common methods scale poorly to the required number of variables.

Human Brain Project

To solve this scalability issue we implemented a variant of the Fast Greedy Equivalent Search which guarantees faithfulness of the recovered net while being paralelizable. In this method we assume the concentration of the genes is distributed normally, evaluate the BIC score difference that comes from adding a given relationship (an edge) between two of the genes and add the edge with the maximal positive score difference, greedily selecting the edges. These steps of calculating all possible edge additions have been heavily paralelized and are what allow us to have this algorithm scale to thousands of nodes. After all possible edge additions have been considered, we remove any edges that have become superfluous, that is, the score difference of deleting them is now positive and return the final network obtained.
We search in the space of equivalence classes, so some edges will be undirected. This is were we could use expert knowledge to help us direct the graph so that it is causally consistent with current knowledge of genetic regulation.
We implemented the algorithm in Python following the original by Ramsey et al. using Numba and MPI to parallelize as much as possible. To test our implementation we use a cluster of three computers in our lab with 32 cores of different capabilities, but mostly around 2.6GHz each.

Preliminary Results

We tested our method against bnlearn's database of gaussian networks and some other artificially generated networks and we can recover them with almost perfect accuracy. Our method has problems with very densely connected networks but this is not very concerning since they are not representative of transcriptional networks.
Preliminary results with real genomic data show that we can retrieve networks of up to 2000 nodes in reasonable times (up to 4 hours) but the unknown structure of the original network makes it hard to assess our accuracy. Visualization of graphs with these many nodes is extremely hard and it can hinder the interpretability of results so we have built a tool to interactively explore the graph and zoom in on nodes of interest.

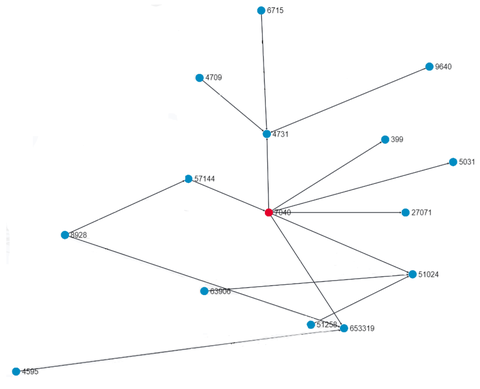

Fig. 3: Markov blanket of Transformig Growth Factor Beta 1

Conclusions

We presented an implementation of a fully scalable method that can be used to learn the structure of transcriptional regulatory networks as bayesian networks from gene expression data. The method can be used both to learn the full network of interest if considerable computing power is available or to learn a reduced neighbourhood around a node of interest. We have made both our learning algorithm and the visualization tool available from our webpage as part of the deliverables for SGA2 in HBP.

Further Work

We are currently working on deploying the method in a supercomputer to learn the full network for the human brain. Additionally, we are planning to learn the structure of the same network in different areas of the brain so that we can do differential analysis and study how gene expression changes around the brain. Finally, we are considering some improvements to the algorithm that take into account some of the properties that are characteristic of transcription networks so that we can further reduce the time needed to learn them and improve the accuracy. Some other approaches would include introducing expert knowledge to the learning step or implementing an inference method that could scale well to the size of this network. Finally, doing experiments guided by the predictions in the network would be useful to calibrate it and more gene expression data could be used to improve the accuracy of the retrieved network.

Acknowledgements

This work was done as part of Task 5.3.4 in SP5 of the Human Brain Project. We thank the Human Brain Project and the European Comision for the funding provided.

References

[1] Hidde de Jong. "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review". In: *Journal of Computational Biology* 9.1 (2002). PMID: 11911796, pp. 67–103. DOI: 10.1089/1066527025283320 8. eprint: https://doi.org/10.1089/1066527025283320 8. URL: https://doi.org/10.1089/1066527025283320 8.

[2] Joseph Ramsey et al. "A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images". In: *International Journal of Data Science and Analytics* 3.2 (Mar. 2017), pp. 121–129. ISSN: 2364-4168. DOI: 10.1007/s41060-016-0032-z. URL: https://doi.org/10.1007/s41060-016-0032-z.