Computational Discovery of Gene Regulatory Networks

Parul Jain (2012CS10240) and Sahil Loomba (2012CS10114)

August 10, 2015

1 Introduction

The interplay of interactions between DNA, RNA and proteins leads to genetic regulatory networks (GRN) and in turn controls the gene regulation. A GRN is a collection of regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins. The regulator can be DNA, RNA, protein and their complex. By identifying the interactions between genes, the discovery of GRNs allow identification of causal genes controlling the pathways of diseases and thus help in detection of disease by monitoring the gene expressions of these genes and the development of medications that target the disease by regulating the gene expressions and keeping them in desired range. The researches on GRNs of various diseases have proven to be taking further step in the direction of targeted drugs for many diseases [8, 9, 10].

2 Prior Art

The problem of discovering gene regulatory networks (GRNs) has been looked into from a variety of perspectives. The first important concern is modelling of a GRN itself. Most experimental biologists have modelled them as a set of ordinary differential equations (ODEs), with each equation describing the chemical kinetics of the “constituents” (henceforth called “interactors”). They then apply standard enzyme kinetics theories such as those of Michaelis-Menten, and then solve for the variables (concentration levels) using nonlinear dynamics (where the fixed rule is that concentration doesn’t change with time). Although biologically directly appreciable, ODE models can be quite cumbersome to realise and solve for very complex networks. Moving into a computational abstraction, GRNs are often modeled by computational biologists as boolean networks (BNs), wherein the interactors are nodes of the network and they are on or off depending on whether they are expressed or not. The boolean nature however leads to a lot of loss of information, in terms of how the interaction is exactly taking place, and thus loses out on prediction power. However, the real win of BNs has been in creating a graph abstraction for the GRN.

A BN thus essentially contains the interactors as nodes, and inhibitory/promotory directed relationships between the nodes. Before a network for a particular end can be analysed, it needs to be constructed. This construction is usually done by analysing DNA microarray data, which is essentially a time signal. We have time signals for each of the interactors in the GRN. Developing often complicated and high-order relationships between these interactomes can be difficult experimentally, which is why we need computational techniques to help guide modern biologists into realising accurate GRNs, for purposes as useful as drug target discovery for neurodegenerative diseases or even cancer.

Past literature has a variety of ways for discovering relationships between interactors. One such work was of modelling a GRN as an Artificial Neural Network without hidden layers, where the learnt weights would be indicative of the intensity of interaction between the two genes for expression [1]. Although the approach worked decently in modelling simple interactions, it was unable to model higher-order non-linear interactions due to absence of a hidden layer. This was followed by a work which tried to add hidden layers to the network, which did imporve the accuracy of the method in determining definite relationships, but made the model biologically uninterpretable [2]. (Although this model was very usable to predict gene expression levels.)

Another visualisation of the microarray data has been to treat them as signals. These signals have the potential to be related causally, in case they are related to one another. Thus, techniques of information theory can be used for such an analysis. ARCANE (Algorithm for the Reconstruction of Accurate Cellular Networks) is one such tool, which uses the idea of mutual information (MI) to find relationships [3]. MI between two signals `x` and `y` essentially measures the amount of shared information between the two signals. Although statistically significant, this so-to-speak correlation does not imply a direct relationship between the two signals, and is thus incapacitated to estimate real causation.

The idea of causation can, however, be employed through probabilistic graphical models. With an underlying bayesian network assumption, a prior work has attempted at GRN discovery using Kalman Filter [4]. The Kalman filter is an algorithm which estimates unknown variables by observing past measurements. However, this assumes a linear dynamic system, which should probably not be true for most biological networks. A good learning from this work though, is consideration of the problem prior to GRN discovery: discovery of relevant genes themselves! The authors use genetic algorithms (GAs) to first select a suitable set of candidate genes, and then apply GRN discovery.

A statistical causality measure which has become popular in systems biology literature, is that of Granger causality (GC). According to it, if a signal `x` "Granger-causes" a signal `y`, then past values of `x` should contain information that helps predict `y` above and beyond the information contained in past values of `y` alone. However, as suggested in [5] there are several shortcomings of using GC in a pairwise fashion on all interactors. The authors suggest that there can be spurious correlations in this case, especially if the order (number of timesteps looked back) of the model is not high enough and the time series data is short. But the primary issue with GC remains that it assumes in a linear regression model, which may not be the case.

As an added convenience, some work has been done to consider any naturally imposed group structure on the interactors. One such work develops graphical-GC models, wherein instead of using simple regression, they use LASSO regression (Least Absolute Shrinkage and Selection Operator) which has the added component in the objective function to minimise the intra-group variance in regression coefficients [6]. This addendum permits groups to be almost wholly present or absent from the final GRN.

3 Our Proposal

We aim to address two major problems which Computational GRN Discovery faces:

1. Incorporating non-linearity of dynamic biological systems in establishing causality relationships.

2. Allow for additional constraints to be put in determining the final GRN structure.

3.1 Incorporating Non-linearity

There are a couple of causality indicators which do not assume linearity of the system, namely transfer entropy (TE) and convergent cross mapping (CMM). The former measures the amount of directed transfer of information from one signal to another. It is the amount of uncertainty reduced in future values of `y` by knowing past values of `x` given past values of `y`. Thus, TE is essentially an extension of conditional mutual information. TE is better than GC, in that it can deal with non-linearity. However, it does require more number of time steps, which could often be lacking.

The latter is similar to GC in the sense of causality theory, however, CMM does not assume separability of the influence of variables, and is based on chaos theory to account for interplaying effects between interactors. It simply treats causal variables as belonging to the same “attractor manifold” [7]. It runs on the idea that one observation can reconstruct a shadow manifold, which can predict another variables belonging to the same true manifold. Besides CMM, other aspects of chaos theory can prove helpful in developing a more realistic picture of gene regulatory networks.

3.2 Incorporating Secondary Information

There could exist a lot of secondary information, like constraints on some network substructures, or observed trends in biological networks like obedience of the power law, which if incorporated can reveal more about the GRN. Taking away from graphical-GC, we should be able to model such constraints in our method. We suggest modelling the problem of GRN discovery, eventually, as an optimisation problem. This would make it possible to work out a globally optimum solution to the problem, and also allow for encoding network stucture constraints as constraints of the optimisation problem itself. For example, we could have an Integer Linear Program (ILP) formulation wherein boolean variables represent existence of relationships between the interactors. We can also relax and soften the ILP contraints, which intuitively might work better in finding a biologically relevant solution.

4 Datasets

The data required for this problem is the most basic form of systems biology data: DNA microarray data. There are various synthetic datasets which can be used, such as the DREAM 3 dataset used in [2]. It would make sense to use data available for bacteria, like the SOS DNA repair network in *Escherichia coli.* Alternatively, as used in [5, 6], we could use data for human cell-cycles, like the HeLa cell-cycle dataset, readily available online at `http://genomewww.stanford.edu/Human-CellCycle/Hela/`. We can essentially use any DNA microarray data, with special focus on extensive datasets, and in the field of therapeutics, such as for treatment of cancer and neurodegenerative diseases.

References

[1] Patrik D'haeseleer et al. *Linear modeling of mRNA expression levels during CNS development and injury*, Pacific Symposium on Biocomputing 4:41-52 (1999)

[2] Michael R. Smith et al. *Time Series Gene Expression Prediction using Neural Networks with Hidden Layers*

[3] Adam A. Margolin et al. *On The Reconstruction of Interaction Networks with Applications to Transcriptional Regulation*, Proc. NIPS Comp. Bio. Workshop, 2004

[4] Nikola K. Kasabov et al. Gene Regulatory Network Discovery from Time-Series Gene Expression Data – A Computational Intelligence Approach

[5] Gary Hak Fui Tam et al. *Gene regulatory network discovery using pairwise Granger causality*, IET Syst. Biol., 2013, Vol. 7, Iss. 5, pp. 195–204

[6] Aurélie C. Lozano et al. *Grouped graphical Granger modeling for gene expression regulatory networks discovery*, Vol. 25 ISMB 2009, pages i110–i118

[7] George Sugihara et al. *Detecting Causality in Complex Ecosystems*, Science 338, 496 (2012)

[8] Natascha Bushati et al. *microRNAs in neurodegeneration*, ScienceDirect

[9] Sébastien S. Hébert, et al. *Alterations of the microRNA network cause neurodegenerative disease*, Cell

[10] Sébastien S. Hébert, et al. *MicroRNA regulation of Alzheimer's Amyloid precursor protein expression*, ScienceDirect