A Project Report

on

Reconstructing GRNs with Bayesian networks using Small World Prior

Submitted by:

Mainak Mandal
Integrated BS-MS
3rd Year, IISER Kolkata

Under the Supervision of:

Dr. Sumeet Agarwal
Assistant Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
Hauz Khas, New Delhi 110016 India

August 10, 2015

1. Gene Expression Data

All cells of a multicellular organism contains the same set of genes. But their protein make-up can be drastically different both spatially and temporally due to regulation. Gene regulation is the process by which the conversion of the information stored in genes to protein end product is controlled. Gene regulation can occur at three distinct places in production of a gene product. Firstly, the transcription of the gene can be controlled this is **transcriptional regulation**. Protein products can also be regulated during transcription by **transcriptional regulation.** This switching on and off of genes at various times make the cells of a multicellular organism so diverse. With the advent of micro-arrays its now possible to get a snapshot of transcription levels of different genes of a cell. This paved the way for discovering interaction between different genes and other interactions using computational tools.

2. Bayesian Networks

Graphical models are special class of models used in statistical inference of system having multiple interacting components. Bayesian network is a type of Graphical model, Consisting of two components:

1. A Directed Acyclic Graph(G):whose vertices corresponds to random variables.
2. Parameters(θ):describe a conditional probability distribution of a vertex given its parents in G.

In Bayesian interpretation probability measures a degree of belief. Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence. Let G be a proposition(graph and parameters) and D be the evidence(data) for it.

- $P(G)$- initial degree of belief in G.
- $P(G|D)$- the posterior is the degree of belief having accounted for D.
- $P(D|G)$- represents the support D provides for G.

Bayes' Theorem states that:

$$\begin{aligned} P(G|D) &= \frac{P(G|D) * P(G)}{P(D)} \\ log(P(G|D)) &= log(P(D|G) + log(P(G)) - log(P(D)) \end{aligned}$$

Since $P(G)$ is same for all graphs it can be ignored, so the above equation can be written as:

$$log(P(G|D)) = log(P(D|G)) + log(P(G)) \tag{1}$$

The LHS is called the **posterior**. The first term on the right is called the **log-likelihood** and the 2nd term is called **log-priori**.

2.1. Mathematical Description

A Bayesian Network is a *acyclic directed graph* which describes a joint probability distribution over a set of random variables(nodes) [3]. Let $\chi = (X_1, X_2, X_3 \cdots)$ be a finite set of random variables where each variable can take values from the domain $Val(X_i)$. Then the variables $X_1, X_2, X_3 \cdots$, denote the nodes of a DAG $'G'$ and the graph denote the conditional independence statements among the random variables. They also encode the *Markov Property*, which means that any of the variables are independent of its non-descendants given its parents in G. By applying the chain rule of probability and the *Markov property* the joint probability distribution can be reduced to the form:

$$P(X_1, X_2, ..., X_{n-1}, X_n) = \prod_{i=1}^{n} P(X_i | Pa^G(X_i)) \tag{2}$$

Where $Pa^G(X_i)$ denote the parents of X_i in G. The parameter θ for a node X_i is the conditional probability of the node X_i given the state of its parents in G i.e., $Pa^G(X_i)$.

2.2. Learning Bayesian Networks

Given a training set $D = \{x^1, x^2 \cdots, x^n\}$ of independent instances of χ the problem is to find a network $G = \{G, \theta\}$ that best matches D. Algorithms for learning Bayesian network form a combination of user knowledge and statistical data. Learning algorithms have two components a scoring metric and a heuristic search algorithm.

2.2.1. Scoring Metric

The Bayesian scoring metric is a way to evaluate how good a given network explains the data. A scoring metric takes in a network, structured statistical data and user priori knowledge to return a score proportional to the posterior probability of the network given the data.

2.2.2. Search Method

Since the problem of searching for a network greater than a certain posterior score is NP Complete [4] heuristic search methods are employed to optimize the posterior score. We will be using one such method called the Genetic Algorithm, for searching networks with higher posterior scores.

2.3. Priors

Inference in Bayesian networks boils down to parameter estimation and model selection. Since we are more interested in the structure of the Gene Regulatory Network(GRN) so we will be mainly focussing on the model selection part [2]. In Bayesian setting model selection is done by sampling from the posterior using heuristic search algorithms. In equation (1) the posterior score is determined by two factors the log-likelihood and the log priori term. The function of the prior term is to influence the search towards networks which are consistent with biological prior knowledge.

2.3.1. Structure Prior:

There are many kinds of prior knowledge available about GRNs. A category of Prior knowledge is having some knowledge about the structure of the GRN which we are reconstructing. Having a knowledge about the structure of the final GRN structure, preference can be given to structures with those properties so that the structure of the final network might be influenced. Using the knowledge about the structure of the GRN as a prior is known as structure prior.

3. Small World Prior

3.1. Problem and Motivation

DNA-micro array mRNA measurement data is extremely noisy and sometimes incomplete. So it is not enough to reconstruct the GRN. With increasing sources of other types of data it is now feasible to use those sources of data as priors for the reconstruction. We will be using one such source of data, that is a structure prior.

A structure prior can be formulated by looking at some model organisms whose GRNs are already reconstructed like yeast(*Saccharomyces cerevisiae*). One scientific study reports that the yeast co-expression network has a scale-free, small-world architecture [5]. And such architecture are common in biological networks in which the nodes are connected when they are involved in the same biological process. So this architecture is nothing special for yeast and can be expected in higher order organisms also. We will be using small worldness property of GRNs as a structure-prior in our reconstruction.

3.2. Mathematical Definition

A network(graph) G, consists of a edge set $\mathbf{E}$ and a vertex set $\mathbf{V}$. For a connected graph, for any two pair of nodes v_i and v_j(both belongs to V) there may exists many paths connecting v_i and v_j. We denote the smallest path by d_{ij}. The *mean path length*($\mathbf{l}$) for a network is defined as the mean of d_{ij} for all pairs of v_i, v_j belonging to V such that $i \neq j$.

$$l^{ws} = \frac{1}{N(N-1)} \cdot \sum d_{ij} \tag{3}$$

Where N denote the cardinality of V and the sum is overall pairs of v_i, v_j belonging to V such that $i \neq j$. A small world network is a network for which the mean path length grows no faster than the logarithm of the number of vertices. That is $\mathbf{l} = O(log(N))$. Small world network are a trade-off between random networks and regular networks because they have short mean paths like random networks and high local clustering like regular graphs. Since we will be using this a s the prior clustering needs to be defined mathematically. *Local clustering coefficient* c^{ws} of node i is defined by Watts and Strogatz [6] as :

$$c^{ws} = \frac{2e_i}{k_i(k_i - 1)} \tag{4}$$

where e_i is the number of connected pairs between all neighbours of i and k_i denotes the number of neighbours of i. Actually its the ratio between the number of connections present between the neighbours of i and the maximum number of connections can possibly exist. So its a number between 0 and 1. The *global clustering coefficient* C^{ws} is defined as the the average of the local clustering coefficient over all nodes of G.

A network with n nodes and m edges is said to be small world if it has a comparable mean path length but has a higher clustering coefficient than a Erdos-Reyni network with same parameters [7]. If l^{rand} denote the mean path length of a Erdos-Reyni network and c^{rand} denote the clustering coefficient of a Erdos-Reyni network then the normalised parameters for G is given by:

$$c_G = \frac{c^{ws}}{c^{rand}} \tag{5}$$

$$l_G = \frac{l^{ws}}{l^{rand}} \tag{6}$$

From the definition of a small world network above we know that the value of $c_G \gg 1$ and $l_G \geq 1$. So if we define a metric, S_G such that:

$$S_G = \frac{c_g}{l_g} \tag{7}$$

then for small world networks S_G will be always greater than 1. To use the S_G as a small worldness prior we will use $log(S_G)$ as the *log-prior* term.

4. Network GA Sampler

The aim of the learning the Bayesian network from data and prior knowledge is to find a network(or rather a equivalence class of networks) that scores the maximum. But as stated above this problem is NP- complete. So heuristic search methods are used to maximise the score. We will be using genetic algorithm to accomplish this task.

4.1. Mathematical Description

In Genetic Algorithm a population of networks is evolved over a large number of generation to ultimately form a population which is far more fit according to the fitness metric used, than the initial population started with. We will be using the posterior score of the graph as the fitness metric in this case. We denote crossover rate as $q \in [0, 1]$ and mutation rate as $m \in [0, 1]$. In each iteration of the algorithm *1-q* fraction of the population having the highest posterior probability is promoted to the next generation. This keeps the most fit organisms of a generation in the population. Then crossover is performed between pairs from rest of the population to produce the offsprings for the next generation [8]. Then mutation is performed on m fraction of edges of each network. Mutation involves changing the state of the edge randomly. This might reduce the posterior score temporarily but it helps the algorithm to avoid local maximas and explore more areas of the search space. After a finite number of iteration the final population is formed which is then used to form the GRN structure. If an edge is present in more than a certain fraction of the networks in the final population then it is included in the final GRN or else its discarded. This way the final GRN is formed. The details of the code will be discussed in the Appendix part.

5. Appendix

Initially it was proposed that the 'ddepn'library will be modified to include the small world prior. But upon preliminary inspection the task was found to be quite elaborate. And there was the problem of modifying a R library and testing it side by side, for which no satisfactory methods could be found which could suit our needs. So we tried to extract the source code and load the source code files in R. Then it came to light that the authors of the 'ddepn'had used C functions to speed up the iterations. So we changed the code so that R equivalents of those C functions are used. But unfortunately the R versions of these functions also didn't work.

So that idea was abandoned and we started writing code from scratch and used pieces of code from 'ddepn'where ever it was found to be efficient. But the code couldn't be completed due to lack of time. The GA sampler is almost ready except the mutation part, and as for the fitness function that is the scoring metric is not complete. But the prior part is complete, although a simpler prior of clustering coefficient is implemented in the code. We were trying to use the *BDe metric* [10] as the log-likelyhood but the log-likelyhood function couldn't be constructed due to lack of time.

5.1. Code Implementation Summary

- An initial population is created using the Watts-Strogaz game and stored in an list.
- The fitness function is defined as the sum of the log-prior and log-liklyhood scores.
- The log-prior term is defined as the log of the average local clustering coefficient as defined in the Watts-Strogaz model.
- Then $1-q$ fraction of the most fit networks are promoted to the next generation. The $1-q$ fraction of individuals is roundoff to the next even integer because we need even number of individuals to perform the cross over.
- The crossover function takes in two adjacency matrix as input and returns two adjacency matrix of the child networks. It divides the matrices along the

diagonal and exchanges the two parts. The crossover parents are chosen from the remaining networks randomly.

References

[1] Steffen L. Lauritzen. Graphical Models. Oxford Statistical Science Series.O1xford University Press, New York, USA, July 1996.

[2] A Scale-Free Structure Prior for Graphical Models with Applications in Functional GenomicsP.Sheridan, T.Kamimura, H.ShimodairaDepartment of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan

[3] Using Bayesian Networks to Analyse Expression Data,JOURNAL OF COMPUTATIONAL BIOLOGY, Nir Friedman, M.Linial, I.Nachman,and D.PeerDavid Maxwell ChickeringComputer Science DepartmentUniversity of California at Los Angeles

[4] Learning Bayesian Networks is NP-Complete

[5] The yeast coexpression network has a smallworld, scalefree architecture and can be explained by a simple model, V.Noort, B.Snel, M.A.Huynen, EMBO reports

[6] Collective dynamics of small-world networks. Watts DJ, Strogatz SH, Nature 393: 440442(1998).

[7] Network Small-World-Ness: A Quantitative Method for Determining Canonical Network Equivalence, Mark D. Humphries, Kevin Gurney, Adaptive Behaviour Research Group, Department of Psychology, University of Sheffield, Sheffield, United Kingdom

[8] Inferring signalling networks from longitudinal data using sampling based approaches in the R-package ‘ddepn’ ,Bender.C, S.Heyde , F.Henjes 1 , S.Wiemann , U.Korf and T.Beibarth BMC Bioinformatics, 2011

[9] Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data, C.Bender, F.Henjes, H.Frhlich, S.Wieman, U.Korf and T.Beibarth

[10] A Bayesian Method for the Induction of Probabilistic Networks from Data , G.F.Cooper E.Herskovits