Static Gene Networks

Markus Heinonen

Helsinki October 19, 2010

UNIVERSITY OF HELSINKI
Department of Computer Science
Genetic networks indicate the genetic interactions which ultimately regulate the processes of organisms at the cellular level. A gene regulation network models indicate the activation and inhibition relations related to each gene. The principal difference amongs genetic network models is the static versus dynamic models. In static models, the direct causal relationships between genes are analysed, while dynamic models assume that the state of a gene is determined by the state of other genes at earlier time points.

A multitude of approaches have been taken to model static genetic networks. An intuitive approach is a boolean network, where genes are either “on” or “off” and regulation is mediated by logical functions. A more complex graphical models are inferred based on statistical properties of the data by looking at partial covariances of genes. Bayesian networks model the conditional independencies of genes in a compact and local manner.

In this paper an overview of learning of static gene networks is presented. In particular boolean networks, gaussian graphical models and bayesian networks are surveyed.
1 Introduction

Genes are the central elements in understanding the function of organisms at the molecular level. The expression of genes controls the functions and processes of the whole cell by RNA, proteins and small molecules of the cell. The regulation of these processes is the key to understanding life. Genetic regulation controls the activation and inhibition of genetic elements using e.g. positive and negative feedback loops and cascades. These form more complex regulatory networks consisting of regulatory structures. Understanding and inferring complex regulatory system is a challenging task where computational analyses are essential [SWBR02].

A plethora of modeling approaches have been utilized in unearthing useful gene networks. The focus has been on kinetic modeling and graphical models, perhaps due to the graphical nature of the network itself. On kinetic modeling, the network is modeled through ordinary differential equations, which operate on concentrations of mRNA’s and proteins in linear, piecewise-linear, discretized or nonlinear way [dJ02]. The problem with kinetic modeling is the size of the parameters space and difficulties in estimating good parameter values [CHC99].

On the other hand, graphical models give easily interpretable interactions between genes using statistical approach. The simplest model is a boolean network, where the binary state of each gene (“on” and “off”) is deterministically determined by a boolean function of all gene variables. Boolean networks have been extended to generalized logical networks for multiple states the gene variables can have [Tho91].

Relevance networks and graphical gaussian models infer correlation relationships between genes using Markovian assumption. Bayesian networks produce directed graphs, where causal-like relationships are explicitly constructed. The main drawback of bayesian networks is the lack of (feedback) loops.

Due to the difficulty of the problem, recent approaches often incorporate modeling features from other domains, such as adding stochasticity or interactivity against experiments in the work [SWBR02].
Figure 1: A part of the gene regulation network of Yeast containing cell cycle related genes, estimated using a non-linear Bayesian network [IGM02]. The network doesn’t contain any cycles, but has still been biologically partly confirmed.
2 Gene networks

Gene networks represent both static and dynamic aspects of the gene functions. In genetic regulation, the static approach aims at finding the mutual influences between genes and the underlying mechanisms. These causal relationships are not tied to a temporal dimension but represent the regulatory capability of certain genes. Dynamic view of the regulation focuses on the state of the system at a given time point and its relation to other time points. Most models contain elements of both models.

Gene networks are mostly learned from gene expression data. Gene expression measurements are able to quantitatively measure the amount of mRNA of every gene of the genome in various environments or at different time-steps. The inherent challenge in gene network learning arises from the fact that the amount of variables $p$ representing genes is often large, yet the number of measurements $n$ is small. This $p > n$ property makes the inferring of genetic networks inherently underdetermined problem, where a robust solutions are highly unlikely.

Several standard datasets are available. A *S. cerevisiae* cell cycle dataset contains 800 genes and 77 experiments from various stages of cell cycle [SSZ+98, FLNP00]. A breast cancer dataset contains 49 samples from 7129 probes with a 3883 genes after preprocessing [WBD+01, SS05a]. Generous amounts of data exist also in public repositories of gene expression data, such as ArrayExpress Archive$^1$ and Gene Expression Omnibus (GEO)$^2$. In Figure 1 an example inferred network is shown, where the Yeast dataset was used.

3 Boolean networks

Boolean networks were one of the first approaches used in inferring genetic interactions from gene expression data [LFS97]. A boolean network consists of gene variables $g_i \in \{0, 1\}$, where the state of genes can contain a temporal dimension with $g_i(t)$. The state vector of the system consists of the expression state of all genes at a certain time point. The next state of the system is computed either using boolean rule tables or state transition table. Typically boolean rules contain logical

\[\text{http://www.ebi.ac.uk/arrayexpress/}\]
operators such as AND, OR and NOT. The state transition table contains $2^p$ rows.

Boolean networks are dynamic models, but they also encode a static structural model of causal interactions. When simulated forward, boolean networks often end up in equilibrium states or in endless cycles.

REVEAL system [LFS97] first discretizes the gene expression data into binary states and then iteratively tries to determine a target gene state from a combination of the least possible number of confounding genes. Confounding genes are enumerated using $k = 1, \ldots, p$ input genes while trying to match input and output using all possible boolean rules inbetween.

The main drawback of boolean models is the boolean state of gene expression, where only two expression levels are allowed.

4 Graphical gaussian models

Graphical gaussian models (GGM) is a family of $p$-variate normal distributions $N(\mu, \Sigma)$ that are Markov with respect to a undirected graph $G = (V, E)$ [Dem72]. Markov property states that if a set of nodes $U$ separates sets of nodes $I$ and $J$ in $G$, the random variables $X_I$ and $X_J$ are conditionally independent given $X_U$. An intuitive interpretation is that two genes are not directly linked if their covariance is zero given the contribution of other genes. While two genes might at first seem to be connected by a non-zero covariance, there might be a third gene whose contribution in coregulating both genes cancels the common covariance between them completely.

GGM’s thus link the undirected direct effect graph $G$ and a multivariate normal distribution of the data. Central concept in GGM’s is the $q$-partial correlation coefficients $\rho_{ij,Q}$, where $Q$ with $|Q| = q$ is the set of random variables the partial correlation from $i$ to $j$ is conditioned against. Partial correlation is defined as

$$\rho_{ij,Q} = \frac{r_{ij} - \prod_{z \in Q} r_{iz} r_{jz}}{\sqrt{\prod_{z \in Q} (1 - r_{iz}^2)(1 - r_{jz}^2)}}, \quad (1)$$

where $r_{ij}$ is the standard covariance of random variables $X_i$ and $X_j$. The central Markov property then translates to having conditional independence $X_I \perp X_J | X_U$ whenever $U$ separates $I$ and $J$. This applies if and only if $\rho_{ij,U} = 0$. 
With empty $Q$ with $q = 0$, the partial correlation $\rho_{ij,Q} = r_{ij}$ is simply the covariance of two random variables. This leads to Relevance networks [BTS00], where a threshold is utilized to find the gene pairs with strongest correlation as causally dependent on each other. This model is not plausible as any two genes are most likely covariates. With $q = 1$, a triplet model emerges where all contributions caused by exactly a single confounding gene are discovered [WB06]. A full order partial correlation with $q = p - 2$ corresponds to finding the marginal correlation of two genes, which corresponds to the classical GGM approach.

For an independence relation between two genes to be established, it is enough to find a single subset of other genes which cause the partial correlation coefficient to be zero. However, there exists $\binom{p-2}{q}$ subsets against which partial correlations have to be checked. This procedure has only been applied for $q \leq 3$ [CR06]. Thus, the full partial correlation model is rendered intractable.

The full partial correlations are often not needed, as the maximum clique size of the gene regulation network indicates the maximum $q$ required. Intuitively we only need to cancel out the contribution of genes of direct effect of either of two genes to establish the independence relation.

An alternative way to determine partial correlation coefficients is to use inverted covariance matrix $K = \Sigma^{-1}$, which is called the concentration matrix. The $K_{ij}$ gives a full order partial correlation coefficient. The maximum likelihood estimate of $\hat{K} = S^{-1}$ requires that $S$ has full rank and this holds for certain if and only if $n > p$. Here, $S$ is the sample covariance matrix, where

$$S_{ij} = \frac{1}{p} \sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j),$$

(2)

where $x_i$ is a random sample vector.

The estimation of partial correlations requires in practice $n \gg p$ [YB94], while for gene expression data we have $p \ll n$. To circumvent this problem, there are several approaches. Wu et al. [WYS03] used clustering to reduce the number of variables below the number of samples. Computing only limited order partial correlations has been used by Castelo and Rovarate [CR06] and others. Third option is to use regularized GGM’s, where suitable estimates for covariance matrix and its inverse are searched using full bayesian treatment [LG06] or via variance reduction and shrinkage estimation [SS05a, SS05b].
In Castelo and Roverate [CR06] a $qp$-procedure is proposed, where partial correlations of a given $q$ are computed using sampling of subsets $\binom{p-2}{q}$, and an error model is proposed to reduce the amount of independencies established due to data noise. In simulated data of 150 genes organized into cliques of 20 genes affecting each other and each clique having 4 genes overlapping with the next, their procedure found 9751 correlating edges out of 13366, where 1206 edges are actually present. However, only 34 of the correct edges were wrong and the structure of regulation is shown. With the breast cancer dataset, a network is constructed with 7240 edges out of 11175 possible ones. This to a meaningful reduction in the search space of possible correlations.

5 Bayesian networks

Bayesian networks are a probabilistic model, where the joint distribution of a set of random variables is represented as a factorization of conditional distributions, which correspond to conditional (in)dependencies in a directed acyclic graph (DAG). The joint distribution of a set of random variables $X_1, \ldots, X_p$ is

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{n} P(X_i|Pa(X_i)),$$

where $Pa(X_i)$ is the set of parents of $X_i$ in the DAG $G$.

Due to the directed nature of the model, causalities can be explicitly modeled. However, the conditional independency requirements are the same for graphs $A \rightarrow B$ and $A \leftarrow B$. Thus, causal bayesian networks require additional semantics to be defined. Namely, in a causal bayesian network, if a state $X$ is causally affected to become $x$, then the joint probability distribution of the network changes such that all parents edges are removed from $X$ and the value of $X$ is fixed to $x$.

Bayesian networks are Markovian, and because the graph is acyclic, each variable is conditionally independent of its non-descendants given its parents. BN’s main advantage is that they offer a probabilistic interpretation of the model, and they usually produce compact models: conditional probability table of each node has to be defined only respect to its parents. With continuous variables a local probability model for each node has to be selected. A natural choice is the gaussian, which is easily representable by mean and covariance matrix, but suffers from unimodality.
and linearity.

The learning of BN’s is NP-hard, however effective heuristic approaches exist [Hec08]. Bayesian networks can be seen as a sibling method for graphical gaussian models as both extract conditional independencies by looking at partial correlations. The main difference is that bayesian networks limit themselves to be DAG’s, which allows for directed edges.

Friedman et al. [FLNP00] used discrete bayesian networks with discretized expression levels (-1, 0, +1) for variables and a linear-Gaussian model for conditional probabilities. The model was applied to a dataset of Yeast gene expression with 76 measurement samples and 6177 gene elements. A bootstrap method is used to resample the data. The confidence of the resulting model was assessed by confidence functions assessing Markov property (two genes with an edge between or a common child) and Order property (descent relation consistent). Majority of the relations were low for both models. However, the regulatory structures of the most high ranking genes were well confirmed biologically and also gave promising biological hypotheses. The prominent genes with respect to the order relation were found to be the starting points of biological pathways.

Pe’er et al [PREF01] focused on building bayesian networks with high confidence subnetworks successfully. Hartemink [HGJY01] extended the model by labeling the edges of the bayesian networks with positive and negative influences. Imoto [IGM02] introduced non-linear bayesian networks, where relationships between genes are a sum of B-spline basis functions. The conditional density $f_j(x_j|Pa(x_j))$ with parents $(p_1^{(j)}, \ldots, p_q^{(j)})$ is modeled through relationship

$$x_j = m_1(p_1^{(j)}) + \cdots + m_q^{(j)}(p_q^{(j)}) + \varepsilon_j,$$

where $\varepsilon_j$ is distributed according to $N(0,\sigma^2_j)$ and $m_k$ are smooth functions of the form

$$m_k(p_k^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk} b_{mk} p_k^{(j)},$$

where $(b_1^{(j)}, \ldots, b_{M_{jk}}^{(j)})$ are the basis functions, coefficients $\gamma$ are unknown parameters and $M_{jk}$ is the number of basis functions. Now the probability density function is
\[ f_j(x_j|Pa(x_j), \gamma_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp \left( \frac{(x_j - \sum_{k=1}^{q_j} m_k(p^{(j)}_k)^2}{2\sigma_j^2} \right) \right). \] \hspace{1cm} (6)

This leads to the standard full joint probability distribution

\[ f(x) = \prod_{j=1}^{p} f_j(x_j|Pa(x_j)). \] \hspace{1cm} (7)

The B-splines are then fitted at equidistant points along the domain \([\min(p^{(j)}_k), \max(p^{(j)}_k)\].

Non-linearity gives a surprising bonus by allowing the estimation of directionality of the edges when the effect is actually non-linear using BNRC (Bayesian Network and Nonlinear Regression) criterion [IGM02].

6 Conclusions

The problem of inferring the static genetic interactions is a difficult task for large gene sets. No single method has so far proven itself superior, but an ensemble of methods has to be used to infer useful information from the measurements. Different methods have different advantages. The small \(n\ large \(p\ problem means that all models have to make compromises.

GGM’s main advantage is the explicit construction of correlation graph with intuitive interpretation. GGM also focuses on finding independencies, which eases the job of domain experts in confirming model. The main problem with GGM’s is that partial correlations are linear and cannot model the kinetic aspects or the combined effects of multiple regulator genes. GGM is thus perhaps the best method used as the first step in order to understand the genetic interactions.

Bayesian networks seem to be the most promising model for static gene regulatory networks. They are incapable of representing feedback loops, but otherwise they provide an intuitive, compact and probabilistic representation of the gene network, while allowing for non-linear models in the choice of local probability model of the variables. Bayesian models also allow for prior knowledge to be used, of which a wealth exists.

A useful model should incorporate aspects of experimental planning. The gene
expression data and CHIP-ChIP data is still expensive and time-consuming, and thus any method should provide insight into useful perturbations to be done or indicate the most informative experimental setup to be applied based on analysis. De la Fuenta proposed an systematic perturbation scheme, where all genes are perturbated to find co-responses against a Bayesian network model [dLFBM01]. Yoo suggested a method where both interventional and non-intervenational data are used in tandem [YTC02]. Ideker has introduced a method to iteratively refine a boolean network model using a minimal number of successive experiments [ITR01]. As long as the $n < p$ problem remains, no method should overlook the experimental control.

References


FLNP00 Friedman, N., Linial, M., Nachman, I. and Pe’er, D., Using bayesian networks to analyze expression data. Proc. Fourth Int. Conf. on Computational Molecular Molecular Biology, 2000, pages 127–135.


