December 14, 2000
RT0392
Computer Science; Life Sciences 7 pages

Research Report

Inferring Genetic Networks from Gene Expression Data Using Probabilistic Boolean Network Models

Hisashi Kashima

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Inferring Genetic Networks from Gene Expression Data Using Probabilistic Boolean Network Models (preliminary version)

Hisashi Kashima

Tokyo Research Laboratory, IBM Research

Abstract

In this paper, as a model of the genetic networks, we propose the probabilistic boolean network model as the intermediate model. And then we give an efficient method to identify the probabilistic boolean network from data. As the model selection criteria, we employ the minimum desctiption length (MDL) principle. Given fixed k input variable, finding maximum likelihood probabilistic boolean concept takes $O(2^{2^k})$ by the naive solution. However, based on the computational geometric algorithm that finds region which minimizes a convex objective function introduced by Morimoto et.al. [7], we can find the (nearly) optimal model efficiently.

1 Introduction

Owing to the recent advance of the biotechnology, the sequence of the human gene has been almost determined. Though, we know few about what functions the each gene has. Therefore as the next step, here comes the problem of so-called 'functional genomics'. Recent years, DNA microarrays enable us to measure the expression level of genes simultaneously and on a massive scale. In other words, DNA microarrays can take a snapshot of the expression of all the genes at a certain time. Therefore we can expect to gain new knowledges of functions of the genes by analysing gene expression data. Since this kind of data is numeric and different from the conventional sequence data, we need the other kind of techniques for gene expression analysis.

One of the knowledge that we are expecting to find from gene expression analysis is to figure out the genetic networks that are regulatory relationships among genes. The genes have a network structure that regulates the expression of the each gene. If we can know such a regulatory network structure, we may be able to know the mechanism of gene expression and to predict the function of the genes from the dependency relations.

One of the approach to the problem of identification of the genetic network is information scientific approach, which assume the model of genetic networks such as boolean networks and differential equation and identify the structure and parameters of them from data. Some models and identification algorithms are proposed but there are few examples where they are applied to real data [4, 5] because of lack of data. Therefore, no models are known which is the most suitable for modelling the genetic networks. On the other hand, there is another strategy that infers the genetic network from vast amount of accumulated knowledge such as bibliographic data [10].

	deterministic	probabilistic
descrete	Boolean network [6, 1],weighted network[9]	Bayesian network[8, 5]
continuous	differential equations[2, 4]	

Figure 1 : genetic network models

In this paper, as a model of the genetic networks, we propose the probabilistic boolean network model as the intermediate model between the boolean network model which is deterministic and the dynamic bayesian network model which is fully probabilistic. the boolean network is a model where the next state of each variable is determined by a boolean function that has the present states of the variable subset as its inputs. In the probabilistic boolean network, the boolean function for each variable is accompanied by two parameters, one of which is the probability of the next state being TRUE if the boolean function outputs TRUE and the other is the probability of the next state being TRUE if the boolean function outputs FALSE. And then we consider efficient methods to identify the probabilistic boolean network from data. As the model selection criteria, we employ the minimum desctiption length (MDL) principle which states one should select the probabilistic model to have the shortest description length in which we can code the model and the given data. Given fixed k input variable, finding maximum likelihood probabilistic boolean concept takes $O(2^{2^k})$ in the naive solution. However, based on the computational geometric algorithm that finds region which minimizes a convex objective function introduced by Morimoto et.al. [7], we can find (nearly) optimal model efficiently.

2 The genetic network identification problem

Suppose that the time series data of experiments using DNA microarrays are represented like Figure 2. Let $v_{i,t}$ be the expression state of the i−th gene at time t. $v_{i,t} = 1$ means the $i - th$ gene expresses at time t and $v_{i,t} = 0$ means it is inhibited. Our objective is to find a regulatory mechanism that explains the data well. Here we suppose the following conditions for simplicity.

1. The expression state of a gene (i.e. expressed or inhibited) is determined by the expression state of some genes at the previous time step.
2. The expression state of each gene are determined independently.

		expression level					
time		1	2	...	j	...	M
	g_1	1	1	...			
	g_2	0	1				
	⋮	⋮	⋮		⋮		
gene	g_i	$v_{i,1}$		...	$v_{i,j}$	...	
	⋮	⋮			⋮		
	g_N	0					

Figure 2 : time series data of gene expression

The second condition enables us to determine the rules of regulation of each gene respectively. Therefore by concentrating the target gene $g_{target}(1 \leq target \leq N)$, the network identification problem reduces to the problem of finding the rule for each target gene from the table like Fugure 3. By performing the same procedure for every genes g_{target} $(target = 1, ..., N)$, we can infer the whole structure of the network.

		expression level					
time		1	2	...	j	...	M
	g_1	1	1	...			
	g_2	0	1	...			
	⋮	⋮					
gene	g_i	$v_{i,1}$	$v_{i,2}$	...	$v_{i,j}$		
	⋮	⋮					
	g_N	0					
target gene g_{target}		1	0	...	$v_{target,j+1}$	...	1

Figure 3 : time series data of gene expression for the target gene

Usually because of various factors such as descretization of the monitored expression level and the inherent stochastic behavior of gene expression and variation of the sampling interval, the gene expression data of a gene does not necessarily identical in the same condition. Furthermore the gene expression data obtained by the DNA microarray contains much noise. Therefore, by counting the number of the expression and the inhibition of the target gene for each previous expression, we can make table like below for each target gene.

		expression level					
gene	$v_{1,t}$	0	1	1	0	$\ldots$	1
	$v_{2,t}$	0	0	1	0	$\ldots$	1
	$v_{3,t}$	0	0	0	1	$\ldots$	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	$v_{i,t}$	0	0	0	0	$\ldots$	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	$v_{N,t}$	0	0	0	0	$\ldots$	1
target gene	$v_{target,t+1} = 1(\text{on})$	3	4	4	6	$\ldots$	5
	$v_{target,t+1} = 0(\text{off})$	2	5	6	1	$\ldots$	3

Figure 4 :

Now our objective is to determine the most plausible rules that explains the expression of the target gene from this table.

3 the probabilistic boolean network model and its identification algorithms

In this section, we consider a system that can present the hypothesis of the possible genetic networks from sequential gene expression data like in the previous section. Firstly we introduce the probabilistic boolean network model as a representation of the genetic network. And then we consider its identification algorithms from sequential gene expression data. Though naive solution makes this problem intractable, we propose a more efficient method based on the convexity of the likelihood function.

3.1 the probabilistic boolean network model

Since the next expression state of the target gene is not determined deterministicly given the present expession state of the genes, we cannnot describe the regulation rules as deterministic boolean function. Therefore we have to introduce probabilistic models. We model the regulation rule as a probabilistic boolean function such as

$$\Pr[v_{target,t+1} = 1] = \begin{cases} \theta & (f^k_{target}(V_t) = 1) \\ \theta' & (f^k_{target}(V_t) = 0) \end{cases} \tag{1}$$

where $f^k_{target}(\cdot)$ is a boolean function that has the expression states of k genes as its input variables and V_t is the expression states of all the genes at time t. In other words, the target gene express with probalility of θ if $f^k_{target}(V_t) = 1$ and with probalility of θ if $f^k_{target}(V_t) = 0$. Furthermore since our rule should distinguish expression and inhibition of the target gene, we introduce a new constraint to the rule as $\theta' \leq 1/2$ when $\theta \geq 1/2$ and $\theta' \geq 1/2$ when $\theta \leq 1/2$, i.e.

$$\left(\theta - \frac{1}{2}\right)\left(\theta' - \frac{1}{2}\right) < 0. \tag{2}$$

We can say our model lies between the boolean network and the dynamic bayesian network[8]. The probabilistic boolean network is a special case of the dynamic bayesian network that has constraints in its conditional probability table.

3.2 model selection using minimum description length principle

Now we are to seek for plausible probabilistic boolean function which can explain the given data well. Naturally search are performed using a criteria that reflects the statistical fitness of the model to the data such as a likelihood function. However, we must search over from simple structure to complex strucure. Moreover it is said that the number of genes that regulates the targer gene is not large in real genetic networks. Therefore as our model selection criteria, we employ the minimum desctiption length (MDL) principle which states one should select the probabilistic model to have the shortest description length in which we can code the model and the given data. The MDL score is described as

$$MDL = \min_{h \in H} L_1(D|h) + L_2(h) \tag{3}$$

where H is the set of the possible hypotheses and D is the given data. The second term represents the description lengh of the model. The first term is the description length of the data when using the model, and the shortest description length given the model is identical to the minus log likelihood of the model. Applying this to our model yields

$$MDL = \min_{\theta, \theta', f^k_{target}} - \log_2 \Pr[m_1, m_0 | \theta, \theta', f^k_{target}] + L(\theta, \theta', f^k_{target}). \tag{4}$$

The first term is minimized by the mothods that is introduced below. The second term is the length needed for description of the model. In our case, though we have to code the truth table and parameters, we do not need to consider the description length of the parameter since the number of parameters are the same for all the models. In the description length of the truth table, $\log_2 k$ bits for specifying the number of input variables k, $k \cdot \log_2 N$ bits for specifying the k input variables and 2^k bits for assigning the truth value to each entry. Usually the more the number of input variable becomes, the more likelihood becomes. Therefore we can take the second term to trade off the fitness of the model to the data and the complexity of the model. Finally, we let the MDL score for k input variables be

$$MDL(k) = \min_{\theta, \theta', f^k_{target}} - \log_2 \Pr[m_1, m_0 | \theta, \theta', f^k_{target}] + (2^k + k \cdot \log_2 N + \log_2 k) \,. \tag{5}$$

But the term 2^k seems to be overestimated, since there are cases where some variable have nothing to do with the output of the boolean function. Therefore we reconsider this from the Bayesian prespectives. The second term can be interpreted that we take the prior distribution as an uniform distributions for each of the number of input variables k and boolean concept that has k input variables respectively. 2^k is derived when we assume the number of boolean functions which has some k variable. If we only consider boolean functions that have 'strictly' k input variables i.e. no redundant input variables, the number of such boolean function is $2^{2^{k-1}}$. Therefore our MDL score is modified as

$$MDL(k) = \min_{\theta, \theta', f^k_{target}} - \log_2 \Pr[m_1, m_0 | \theta, \theta', f^k_{target}] + (2^{k-1} + k \cdot \log_2 N + \log_2 k) \,. \tag{6}$$

Our model searching strategy is the following. First, we find the maximum likelihood models for each $k = 1, ..., K$ where K denotes the maximum number of the input variables. And then finding the model that has the minimum MDL score among them. The higher level algorithm is described below.

ALGORITHM : GENETIC NETWORK IDENTIFICATION

1. *For* $k = 1, ..., K$,
   - *(a) For all the combination of* k *input variables,*
     - *Find the maximum likelihood probabilistic boolean function for the given* k *input variables.*
   - *(b) Calculate MDL(k) according to (6) for the maximum likelihood model among the models found in the previous step.*
2. *Output the model that has the minimum MDL(k).*

As mentioned before, K is said not to be so large for the real genetic networks. The central problem here is how to perform the step 1a. We explain the methods for doing this from the next subsection.

3.3 parameter estimation

Once the truth table of the boolean function f^k_{target}. The best θ and θ' that describes the data are determined by maximum likelihood estimation. Let M_1 be the number of the data where the target gene is expressed i.e. $v_{target,t+1} = 1$ and M_0 be the number of the data where the target gene is inhibited i.e. $v_{target,t+1} = 0$. And let m_1 be the number of the data that satisfies $f^k_{target}(V_t) = 1$ and $v_{target,t+1} = 1$, and m_0 be the number of the data that satisfies $f^k_{target}(V_t) = 1$ and $v_{target,t+1} = 0$. The log likelihood of this model is

$$\log \Pr[m_1, m_0|\theta, \theta', f^k_{target}] = m_1 \log \theta + m_0 \log(1-\theta) + (M_1 - m_1) \log \theta' + (M_0 - m_0) \log(1-\theta'). \quad (7)$$

and the parameters are given as

$$\theta = \frac{m_1}{m_1 + m_0} \quad (8)$$

$$\theta' = \frac{M_1 - m_1}{M - (m_1 + m_0)}. \quad (9)$$

3.4 determination of boolean rules

Once the boolean function f^k_{target} is given, maximum likelihood estimation is trivial. But the problem here is how to find the f^k_{target} itself. Given k input variables , the problem of determining a boolean function is equivalent to select the subeset of 2^k assignments of the truth value i.e. determining a truth table. Therefore the number of the possible boolean function is 2^{2^k}. Even in the case where the number of the input variables is small, we cannot see this a constant since 2^{2^k} becomes an unrealistic value ($2^{2^{10}} = 2^{1024}$) for $k = 10$. Therefore the naive search over all the probable models is intractrable solution and we seek for the more efficient solution. Here we employ the computational geometric idea of Morimoto et. al.[7] Given k input variables and the boolean function f^k_{target} fixed, the corresponding point (m_0, m_1) in the $m_0 - m_1$ plane is determined. And given the coordinate (m_0, m_1), the corresponding likelihood is determined. In other words, the likelihood function is defined over $m_0 - m_1$ plane and a fixed boolean function is a point in the plane. Using the fact that the likelihood function over $m_0 - m_1$ plane is a convex function (the nonnegativity of its second derivative is easily checked), the most likelihood point (m_0^*, m_1^*) is on the convex hull of the points in $m_0 - m_1$ plane. Therefore our problem reduces to finding the point that has the maximum likelihood on the convex hull in the region that satisfies

$$\left(\frac{m_1}{m_1 + m_0} - \frac{1}{2}\right)\left(\frac{M_1 - m_1}{M - (m_1 + m_0)} - \frac{1}{2}\right) < 0 \quad (10)$$

by substituting (8) and (9) into (2).

3.4.1 a method based on dynamic programming

Here we introduce an optimal but less efficient mothod based on the dynamic programming. Hand probing is one of the method for finding points on convex hell. The idea is to move a line that has some 傾き a to the convex hull until the line hits the convex hull. More specifically we find a point that maximizes or minimizes $b = m_1 - a \cdot m_0$. Let e_1 be the number of expression of the target gene and e_0 be the number of inhibition of the target gene. We can reduce this problem to $0 - 1$ knapsack problem [3] where items are the assignments of the truth value ,the weights are the two dimensional vectors (e_0, e_1), the values of the items are $e_1 - a \cdot e_0$ and the size of the bag is the two dimensional vector(M_0, M_1). We can solve this by dynamic programming solution for the knapsack problem and find the optimal solution in the region that satisfies (10). By continueing this for various a, we can find the maximum likelihood point.

3.4.2 a more efficient method

Though the previous method can find the optimal point in the region (10), it is not efficient. Therefore we introduce a method that finds not necessarily optimal but nearly optimal point. The idea is to forget the region constraint (10). Firstly we find the convex hull without the constraint, and then find a solution in the region (10). The difference of the two cases where with and without the constraint (10) lies around where the boundary of the region and the convex hull meet . Though this method cannot guarantee to find the optimal solution, the differrence is so that that we can neglect the error.

ALGORITHM : MAXIMUM LIKELIHOOD PROBABILISTIC BOOLEAN NETWORK

```
1. For each assignment of truth values to the input variables, calculate the ratio $s = e_1/e_0$ of the number of expression $e_1$ and the number of the expression $e_0$of the target gene.
2. Sort $s$ in descending order. Let $a_i$ be the assignment that has the $i-$th highest $s$.
3. Set the truth values for all the assignment to FALSE, and let this be the current hypothesis.
4. Until the current hypothesis does not satisfy the condition (10).
     • set the truth value of the assignment $a_i$ to TRUE and let this be the current hypothesis.
5. Output the maximum likelihood hypothesis among all the hypothesis presented so far.
```

Indeed we can combine the previous method or local search starting from the suboptimal point with this algorithm.

4 Experiments

We perform computational experments to evaluate this method. Unfortunately our method is still not realistic for real large genetic netoworks, since though this method can efficiently find the most probable probabilistic boolean function given the input variables, it search all the combination of the input variables. And Furthermore the number of data available in public is too small. Therefore as a preliminary test, we applied our method to the synthetic time series data which is generated from a randomly constructed probabilistic boolean network. We set the number of genes to 10, the number of the input variables to up to 5 and the number of data to 200. The time series data is taken 5 times starting at ramdom states and the length of the each time series is 40. In experiments, the networks were almost perfectly reconstructed when the data includes up to 5 percent noise.

5 Concluding Remarks

We introduced the probabilistic boolean network as a model of the genetic networks. It is an intermediate model between two models already been proposed, which are the boolean network model which is deterministic and the dynamic bayesian network model which is fully probabilistic. And then we considered efficient methods to identify the probabilistic boolean network from data using the minimum desctiption length (MDL) principle as our model selection criteria. Based on the computational geometric algorithm that finds region which minimizes a convex objective function introduced by Morimoto et.al. [7], we can find (nearly) optimal model efficiently. However we search all the combination of the input genes up to K genes so far. Our method is still not enough for large real genetic networks.
Our method has the following features.

- We can apply this method not only for the genetic networks, but for SNP analysis.
- This system itself is not specializes to biology and applicable for another application.

The future directions of this research can be summarized as follows.

- preprocessing method that can focus the relevant genes to the targer gene
- incorporating biological knowledge as priors
- application to the real data

References

[1] Akutsu, T., S. Miyano and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under Boolean network model. *Pacific Symposium on Biocomputing*, 4:17–28, 1999.

[2] Chen, T., H. L. He and G. M. Church. Modelling Gene Expressions with Defferential Equations. *Pacific Symposium on Biocomputing*, 4:29–40, 1999.

[3] Cormen, T. C., C. E. Leiserson and R. L. Rivest. *Introduction to Algorithm*. MIT Press, Cambridge, MA, 1990.

[4] D'haeseleer, P., X. Wen, S. Fuhrman and R. Somogyi. Linear Modelling of mRNA Expression levels During CNS Development and Injury. *Pacific Symposium on Biocomputing*, 4:42–52, 1999.

[5] Friedman N., M. Linial, I.Nachman and D. Pe'er. Using Bayesian Networks to Analyse Expression Data. *RECOMB 2000*, 2000.

[6] Liang, S, S. Fuhrman and R. Somogyi. REVEAL:A General Reverse Engineering Algorithm for Inference of Genetic Network Architectures. *Pacific Symposium on Biocomputing*, 3:18–29, 1998.

[7] Morimoto, Y. et. al. Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases. *Proceedings of the 24th VLDB conference*, pages 380–391, 1998.

[8] Murphy, K. and S.Mian. Modelling Gene Expression Data using Dynamic Bayesian Networks. *UC Berkeley Technical Report*, 1999.

[9] Noda, K. et.al. Finding Genetic Network from Experiments by Weighted Network Model. *Intl. Workshop on Genome Informatics*, 1998.

[10] Zien A., R. Kuffner, R. Zimmer and T. Lengauer. Analysis of Gene Expression Data with Pathway Scores. *ISMB 2000*, 2000.