# A POLYNOMIAL APPROACH TO STRUCTURAL GENE DYNAMICS MODELLING

**Saadia Faisal** * **Gerwald Lichtenberg** *
**Herbert Werner** *

* *Institute of Control Engineering*
*Hamburg University of Technology, Germany*
*email: {S.Faisal,Lichtenberg,H.Werner}@tu-harburg.de*

Abstract: A new approach for modelling the dynamics of gene expression from time series microarray data is presented. A modelling method based on a continuous representation of Boolean functions in the form of Zhegalkin Polynomials is proposed. Structural information known from theoretical biology like the canalizing property can be included as well as continuous measurements of gene expression levels. As an example, its applicability to yeast data is demonstrated. The complexity of the problem requires efficient methods and tools. The discrete set of all Canalizing Boolean models consistent with the measurements is large and grows exponentially as the connectivity degree of each gene increases. This set can be defined in terms of the Zhegalkin Polynomial coefficients. Moreover, this paper gives two theorems on structural properties of Canalizing Zhegalkin Polynomials. An algorithm based upon them shows how these results can be used for identifying Canalizing Boolean functions. 

Keywords: Genetic Networks, Gene Expression, Boolean Networks, Canalizing Functions, Zhegalkin Polynomials

## 1. INTRODUCTION

A current challenge in Systems Biology is to understand and unveil the gene expression mechanism in cells. Genes encode proteins which in turn regulate genes, (P.D'haeseleer, 2000). The activities of genes or gene expression levels are further controlled by a number of elements present in the cell e.g., receptor proteins etc. The cells are able to recognize and respond to molecules in the extracellular environment via so called *signalling pathways*, (C.Sniegoski and R.Somogyi, 1996). In order to model gene expression mechanism, m$RNA$ levels (which determine the level of expression of genes) have been chosen as a level of abstraction out of several levels in signalling pathways.

Microarray technology has made it possible to measure m$RNA$ levels on a large scale as well as in sequence resulting in timed expression level signals. With the availability of such measurements, understanding and modelling of gene expression dynamics is currently an aim of research as it can help to control many deadly diseases including cancer etc.

This paper discusses one new approach in this field based on Zhegalkin Polynomials, (Faisal *et al.*, 2004*a*), used to model gene dynamics using gene expression data time series. The main problem in modelling is that biological measurements - as many as they may get - only cover a small part of the complexity of the whole system.

This paper is organized as follows: In Section 2, a short review of different approaches to modelling gene dynamics is given. Moreover it also poses the gene dynamics identification problem formally. Section 3 discusses a new modelling method briefly and illustrates its applicability through an example from yeast time series data in Section 4. In Section 5, two important theorems are given that form a first step towards the development of more efficient algorithms in order to handle the complexity of the modelling problem in a better way. The applicability of these results is shown by an algorithm. The paper closes with conclusions and outlook.

## 2. MODELLING APPROACHES

It was observed by Kauffman that genetic systems share many characteristics with Boolean networks such as periodicity and have a global complex behavior, (S.A.Kauffman, 1969) and (Kauffman, 1993). The use of Boolean logic was therefore proposed for understanding the complex genetic interactions as well as to reduce the genetic system to its principle features. This Boolean idealization of genetic networks was further developed by (T.Akutsu *et al.*, 1998) and (T.Akutsu *et al.*, 1999) and (S.Fuhrman *et al.*, 1998), where a gene assumes one of two states either "on" (1) or "off" (0) and the state of a gene is determined by a Boolean function of the states of other genes. As the system evolves from one time point to another, the pattern of currently expressed genes is used as an input to a Boolean rule which determines which genes will be "on" at the next time point.

All microarray data in time series for an organism can formally be represented by a set

$$\mathcal{X} = \{(\mathbf{x}(0), \ldots, \mathbf{x}(T)) \mid \mathbf{x}(k) \in \mathbb{R}^n\} \qquad (1)$$

of sequences of $n$ measured gene expression levels $\mathbf{x}_i(k)$ for $T+1$ discrete sampling time points.

For the Boolean network approach, the continuous data sequences have to be quantized with the help of a step function $\mathbf{q}: \mathbb{R}^n \longrightarrow \{0,1\}^n$ with

$$q_i(\mathbf{x}) = \begin{cases} 0 \text{ for } x_i < \theta_i \\ 1 \text{ for } x_i \geq \theta_i \end{cases} \qquad (2)$$

and thresholds $\theta_i$ such that the set (1) can be mapped to the set

$$\mathcal{Z} = \{(\mathbf{z}(0), \ldots, \mathbf{z}(T)) \mid \mathbf{z}(k) = \mathbf{q}(\mathbf{x}(k))\} \qquad (3)$$

that only contains sequences of Boolean $n$-vectors.

In order to infer Boolean networks from the above sequence of Boolean $n$-vectors (S.Fuhrman *et al.*, 1998) and (T.Akutsu *et al.*, 1999) devised algorithms. Their algorithms *REVEAL* and *BOOL1* respectively, perform well with the assumption that each gene is effected by only a few (less than four) other genes. In contrast to this, from a biological point of view, some genes have a low connectivity degree while others are effected by much more than four other genes.

It is known that genes are governed by a special kind of Boolean functions which are called Canalizing Boolean functions (Kauffman, 1993). A constraint of this particular kind of Boolean functions along with that of data was used to reduce the search space of possible Boolean functions. By definition a *Canalizing function* is a Boolean function $f_b : \{0,1\}^n \rightarrow \{0,1\}$ having the property that at least one of its input variables has one value which alone suffices to guarantee one value of the output variable.

In general, models can be divided into classes according to their types of signals

- qualitative / discrete / Boolean, (T.Akutsu *et al.*, 1999)
- quantitative / continuous, (T.Chen *et al.*, 1999) and (J.Dehoon *et al.*, 2003)

or according to the handling of uncertainty

- stochastic / Bayesian, (S.Imoto *et al.*, 2004) and (D.K.Gifford *et al.*, 2001))
- deterministic, (S.Fuhrman *et al.*, 1998).

Taking the first criteria into account the actual situation in gene dynamics modelling can be described by Figure 1.

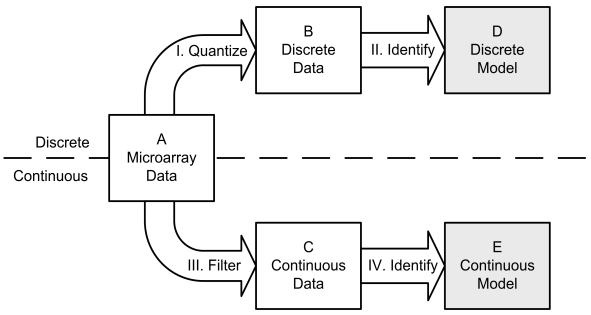


Figure 1. Known Modelling Paths

Many concurrent approaches are used for modelling that start with continuous microarray data (A). The first type of modelers starts by quantizing the data (I) and identify (II) a discrete model (D) that can be either deterministic (e.g. REVEAL) or nondeterministic (e.g. BOOL1) or stochastic (Discrete Bayesian Networks), reflecting the qualitative behavior of the measured data.

In the identification step, biological constraints (usually formulated as rules) can be included, an example is the canalizing property of Boolean functions that are used as models of the gene dynamics.

The other type of modelling techniques use quantitative approaches which do not quantize but filter and cluster the data (III). From the filtered

data, continuous models like differential equations are derived by means of identification techniques, (N.Friedman *et al.*, 2000). These efforts also include continuous Bayesian network models, (S.Imoto *et al.*, 2004), which are feasible for only small gene networks. The resulting models are able to describe the continuous levels of gene expression.

Both types of approaches have drawbacks. With the discrete approach it is not possible to model intermediate expression levels that do occur in reality as well as in microarray data. With quantitative approaches, it is not possible to define biological constraints appropriately.

With the knowledge of structural assumptions like the canalizing property on one hand and on the other hand, the availability of continuous microarray measurement data, the following identification problem for structured modelling of gene dynamics can be posed:

- **Given:** Set (1) of microarray data time series
- **Find:** Deterministic continuous model with transition vector function $\mathbf{f}$, such that

  $$\mathbf{x}(k{+}1) = \mathbf{f}(\mathbf{x}(k)) = (f_1(\mathbf{x}(k)), ..., f_n(\mathbf{x}(k)))^T$$

  which reflects not only the continuous measurements but the qualitative behavior induced by the quantization as well. This model has to fulfill structural properties, especially the canalizing property.

A modelling approach has been outlined by the authors to solve this identification problem that is different from the known standard ways and combines methods of different disciplines to reach the aim, (Faisal *et al.*, 2004*b*).

## 3. MODELLING METHOD

It can be concluded from Section 2 that any two modelling approaches that belong to two different classes (discrete/ continuous) are neither compatible nor comparable. A modelling method that nevertheless links both classes is described briefly here.

Using appropriate thresholds, the continuous valued discrete time series data of gene expression levels can be quantized to Boolean valued discrete time series data as given in Figure 1. The next task is to find all Canalizing discrete models which are consistent with the data. Therefore, contrary to Block D in Figure 1 a set of discrete models can be obtained rather than a unique model.

At this point the missing link between the qualitative and quantitative approach can be obtained by using and adapting well known methods of continuous representations of Boolean functions, in particular Zhegalkin Polynomials. They consist in general of multilinear polynomials of $n$ variables which have the form

$$\begin{aligned} f_i(\mathbf{x}) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{j=2}^{n}\sum_{i=1}^{j-1} a_{ij} x_i x_j + \\ + \sum_{k=3}^{n}\sum_{j=2}^{k-1}\sum_{i=1}^{j-1} a_{ijk} x_i x_j x_k + ... \\ + a_{123...n} x_1 x_2 x_3 ... x_n \ . \end{aligned} \tag{4}$$

These multilinear polynomials have the ability to describe both, continuous as well as discrete functions. With a restriction over the range of values that the coefficients ($a_i, a_{ij}, ...$ etc) assume, these representations coincide with Boolean functions at Boolean values of the variables $x_i$ but apparently are continuous functions, see (Franke, 1994). Arbitrary Boolean functions can be represented by Zhegalkin Polynomials. On the other hand, these representations can handle continuous values as well.

The conversion of all canalizing discrete models, that are consistent with the measurements, into their continuous representations yields a set of continuous models rather than a single continuous model as in Block E of Figure 1. In the final step, by using a suitable discrete optimization method in order to minimize the quadratic estimation error

$$J = \sum_{k=0}^{T-1} \left(f_i(\boldsymbol{x}(k)) - \boldsymbol{x}(k+1)\right)^2 \tag{5}$$

for each component $f_i$ in eqn. (4), an optimal **and** structured model can be obtained. As it is a discrete optimization problem, attention has to be paid to the search space.

Any Zhegalkin Polynomial can be defined in terms of its coefficient vector

$$\boldsymbol{a} = (a_0, a_1, a_2, ..., a_{123...n}) \in \mathcal{A}_b \subset \mathbb{Z}^{2^n} \ . \tag{6}$$

Clearly not all integer combinations of coefficients are allowed to ensure the property that each $f_i$ only takes Boolean values as outputs if the input is a Boolean vector. The subset of allowed combinations is given by $\mathcal{A}_b$.

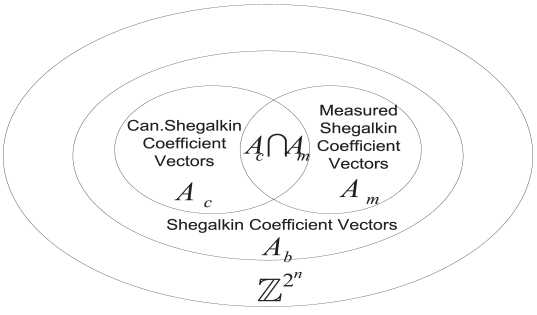


Figure 2. Coefficient sets and search space

Moreover, the Boolean functions represented by $f_i$ have to fulfill additional properties. The canalizing property reduces the allowed coefficient vectors to the set $\mathcal{A}_c$. At a final step, the Boolean functions have to be consistent with the quantized measurements. Let the set of all coefficient vectors that ensure this be denoted by $\mathcal{A}_m$. In Figure 2, the important sets and their relations are illustrated.

Thus, the search space for the optimization of the cost function (5) is given by

$$\mathcal{A} = \mathcal{A}_c \cap \mathcal{A}_m \ , \tag{7}$$

and the full structural identification problem to derive the best coefficient vector $\boldsymbol{a}^*$ (i.e. the model) given by

$$\min_{\boldsymbol{a} \in \mathcal{A}} J \ . \tag{8}$$

In order to illustrate the modelling method, a simple example using yeast gene expression data will been given in the next section.

## 4. YEAST GENE DATA MODELLING

For the sake of showing applicability of the proposed modelling cycle, the results of the whole modelling cycle for yeast time series data (http://genome-www5.stanford.edu) of four genes ($HSF1$, $SSA1$, $SSA3$, $YRO2$), which take active part in heat shock response in yeast, are presented here. Linearly normalized expression levels of the above mentioned four genes are shown in Figure 3. It is assumed that the expression level of each gene

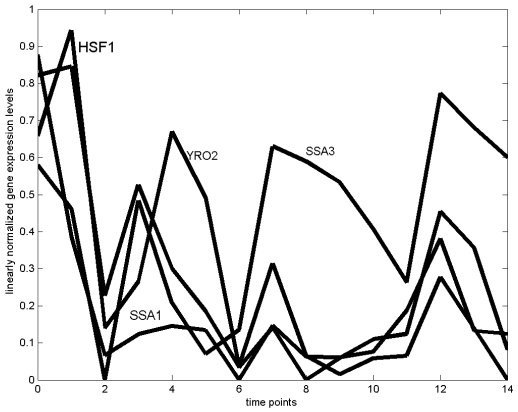


Figure 3. Normalized time series data of yeast

at time step $k+1$ depends at least upon the expression level of all four genes at the previous time step $k$. This particular example shows the result of the modelling method for a structural model corresponding to the gene $HSF1$. Following the steps of the modelling method, using mean values of the gene expression data as thresholds for quantization, 575 *Canalizing Boolean functions* were obtained out of $2^{12}$ Boolean functions consistent

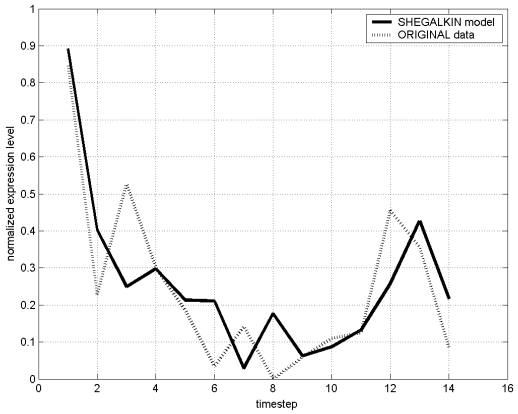


Figure 4. Time series model vs. measurements for the gene HSF1

with the measurements, corresponding to gene $HSF1$. Optimization of the Least Square error yielded an optimal Zhegalkin Polynomial that can be seen in Figure 4 in comparison to the original data.

It can be seen that there are some discrepancies between the model and the measured sequences, because the example does only in reference to four other genes causing the behavior of $HSF1$. Including additional gene expression levels profiles or increasing the number of variables of the function $f_i$ could resolve this ambiguity.

One problem to determine such models lies in the fact that the number of Zhegalkin polynomials is very large for a higher degree of connectivity. Then, separation of canalizing from non canalizing polynomials is not trivial. In the next Section, some conditions are given that distinguish coefficient vectors of canalizing form coefficient vectors of non canalizing Zhegalkin polynomials.

## 5. STRUCTURE OF CANALIZING ZHEGALKIN POLYNOMIALS

Given a Zhegalkin Polynomial representation of a Boolean function, the following theorems give necessary and sufficient conditions to test if it belongs to the canalizing or the noncanalizing class, (Faisal *et al.*, 2005).

**Theorem 1**

*A Zhegalkin Polynomial representing a Boolean function with n inputs is canalizing with the canalizing variable* $x_i, i \in \{1, \ldots, n\}$ *and the canalizing value* 0 *if and only if all following conditions hold:*

$$\begin{aligned} a_j &= 0 \quad j \neq i \ , \\ a_{jk} &= 0 \quad j,k \neq i \ , \\ a_{jkl} &= 0 \quad j,k,l \neq i \ , \\ &\vdots \end{aligned}$$

$$a_{12\ldots(i-1)(i+1)\ldots n} = 0\ . \tag{9}$$

*A Zhegalkin Polynomial representing a Boolean function with n inputs is canalizing with the canalizing variable* $x_i, i \in \{1,\ldots,n\}$ *and the canalizing value* 1 *if and only if all following conditions hold:*

$$\begin{aligned} a_j + a_{ij} &= 0 \quad j \neq i\ , \\ a_{jk} + a_{ijk} &= 0 \quad j,k \neq i\ , \\ a_{jkl} + a_{ijkl} &= 0 \quad j,k,l \neq i\ , \\ &\vdots \\ a_{12\ldots(i-1)(i+1)\ldots n} + a_{12\ldots n} &= 0\ . \end{aligned} \tag{10}$$

**Example**

The use of the above theorem can be shown by considering a Boolean function with 3 input variables and testing their canalizing property. Consider a Boolean function

$$x_1\ XOR\ (x_2\ XOR\ x_3).$$

The corresponding Zhegalkin Polynomials for this function is

$$x_1 + x_2 + x_3 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 + 4x_1x_2x_3.$$

As in this function $a_1 \neq 0$, $a_2 \neq 0$ and $a_3 \neq 0$, the function cannot be canalizing to any variable with respect to the value 0. Since $a_{12} + a_{123} \neq 0$, $a_{13} + a_{123} \neq 0$ and $a_{23} + a_{123} \neq 0$, the function cannot be canalizing to any variable with respect to the value 1. So this function is noncanalizing.

The following result describes another structural property of the Canalizing Zhegalkin Polynomials. As before only the theorem is given and the proof is omitted.

**Theorem 2**

*If* $f(\boldsymbol{x})$ *is a Zhegalkin Polynomial representation of an arbitrary Canalizing Boolean function with canalizing variable* $x_i$ *then*

$$\begin{aligned} a_{ij} &\in \{-1,0,1\} \quad \forall j = 1,...,n \\ a_{ijk} &\in \{-2,-1,0,1,2\} \quad \forall j,k = 1,...,n \\ &\vdots \\ a_{1\ldots i\ldots n} &\in \{-2^{n-2},\ldots -1,0,1,...,2^{n-2}\}. \end{aligned}$$

**Example**

Let a Zhegalkin Polynomial representation of a Boolean function of two variables be given by

$$f_2(x_1,x_2) = 1 - x_1 - x_2 + 2x_1x_2$$

The range of values that $a_{12}$ can assume to be a Zhegalkin Polynomial is given by

$$a_{12} \in \{-2,-1,0,1,2\}.$$

Since $a_{12} = 2$ in the above polynomial, and Theorem 2 doesn't allow a Canalizing Zhegalkin Polynomial to have coefficient $a_{12}$ with values in the set $\{-2,2\}$ therefore $f_2$ is noncanalizing.

Based on the Theorems, an algorithm to identify Canalizing Zhegalkin Polynomials will be presented next.

**Algorithm**

The following algorithm is formulated by using the two theorems. It checks whether a given Zhegalkin Polynomial represents a Canalizing Boolean function.

Let $f$ denotes a Zhegalkin Polynomial of $n$ variables as given in eqn. (4) and let

$S = \{-2^{n-1},...,-2^{n-2}-1, 2^{n-2}+1,...,2^{n-1}\}$ .

I. Check for non canalizing property.
   If $a_{1\ldots n} \in S$ is true
   $\Rightarrow f$ is not canalizing. STOP
   else
   NEXT.

II. Check variables $x_k$ for canalizing value 0.
   For $k = 1$ to $n$ do
   Check conditions (9) for $i = k$.
   If for any $k$ all conditions are true
   $\Rightarrow f$ is canalizing. STOP.
   end do

III. Check variables $x_k$ for canalizing value 1.
   For $k = 1$ to $n$ do
   Check conditions (10) for $i = k$.
   If for any $k$ all conditions are true
   $\Rightarrow f$ is canalizing. STOP.
   end do

IV. $f$ is not canalizing.

**Remarks**

It is clear that for an implementation of the above algorithm, redundancy should be avoided and results of tested equations have to be stored in a efficient manner for reuse.

Moreover, to solve problem (8), it is necessary to represent the elements of the search space $\mathcal{A}$ in a convenient way. Figure 5 shows that it makes sense not to use truth table representations at all, because the effort of transformation can be omitted. But because of the simplicity of the canalizing test in terms of the coefficients of the Zhegalkin polynomials, it may even be more efficient to

transform the truth table first and then check the property than to check it directly.

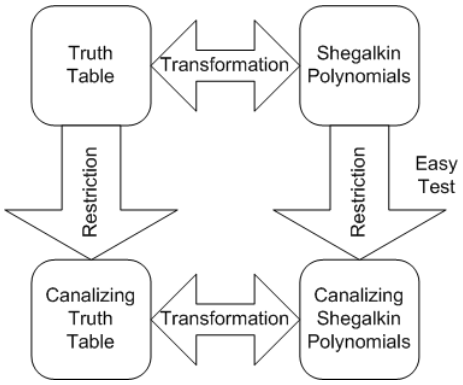


Figure 5. Canalizing Zhegalkin Polynomials

The truth table representing a Boolean function requires $n(2^n-2)$ conditions to check in the worst case, to find out if the given Boolean function is Canalizing. Whereas, by using the proposed algorithm $2^{n-1}$ functions will be identified as noncanalizing just by checking a single condition (the first step of the algorithm). However for the rest of the noncanalizing functions, it still needs up to $n(2^n-2)$ conditions to check.

## 6. CONCLUSIONS

A new modelling method for the dynamics of gene expression is given. This method inherits a polynomial model for the expression levels that can be interpreted both, qualitatively and quantitatively. The model can be identified by given timed sequences of microarray data and at the same time, it can be ensured that discrete biological rules like the canalizing property are not violated.

To do this, an underlying discrete optimization problem has to be solved that suffers from combinatorial complexity. The representation of sets of these models is possible by means of their coefficients. Two theorems are derived that give insight into the structure of allowed combinations of coefficients if the canalizing property is demanded. The next step will be to improve the algorithm for efficient implementations of identification routines.

## REFERENCES


C.Sniegoski and R.Somogyi (1996). Modeling the complexity of genetic networks, understanding multigenic and pleiotropic regulation. *Complexity* (1), 45–63.

D.K.Gifford, A.J.Hartmink, T.S.Jaakkola and R.A.Young (2001). Using graphic models and genome expression data to statistically validate models of genetic regulatory networks. In: *Pacific Symposium on Biocomputing.* number 6. pp. 422–433.

Faisal, S., G. Lichtenberg and H. Werner (2004*a*). An approach to structural modelling of gene dynamics using reed-muller-forms. In: *Proceedings of Control 2004, Bath, U.K.* University of Bath, U.K. p. 88.

Faisal, S., G. Lichtenberg and H. Werner (2004*b*). An approach using shegalkin polynomials for modelling microarray timeseries data of eucaryotes. In: *Proceedings of International Conference on Systems Biology, 2004.* DKFZ, Dchaema.

Faisal, S., G. Lichtenberg and H. Werner (2005). Structural properties of polynomial representations for canalizing boolean functions. *submitted to automatica.*

Franke, D. (1994). *Sequentielle Systeme.* Vieweg, Braunschweig.

J.Dehoon, S.Imoto, K.Kobayashi, S.Miyano and N.Ogasawara (2003). Inferring gene regulatory networks from time ordered gene expression data of bacillus subtilis using differential equations. In: *Pacific Symposium on Biocomputing.* number 8. pp. 17–28.

Kauffman, S.A. (1993). *The Origins of Order, Self Organization and Selection in Evolution.* Oxford University Press.

N.Friedman, M.Linial, I.Nachman and D.Pe'er (2000). Using bayesian networks to analyze expression data. *Journal of Computational Biology* (7), 601–620.

P.D'haeseleer (2000). Reconstructing Gene Networks from Large Scale Gene Expression Data. PhD thesis. The University of New Mexico.

S.A.Kauffman (1969). Matabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology* (22), 437–467.

S.Fuhrman, S.Liang and R.Somogyi (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In: *Pacific Symposium on Biocomputing.* number 3. pp. 18–29.

S.Imoto, S.Miyano and S.Ott (2004). Finding optimal models for small gene networks. In: *Pacific Symposium on Biocomputing.* number 9. pp. 557–567.

T.Akutsu, S.Kuhara and S.Miyano (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: *Pacific Symposium on Biocomputing.* number 4. pp. 17–28.

T.Akutsu, S.Kuhara, O.Maruyama and S.Miyano (1998). Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. In: *ACM-SIAM Symposium.* number 9. pp. 695–702.

T.Chen, G.M. Church and H.L.He (1999). Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing* (4), 29–40.