14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

# A METHOD FOR ANALYSING GENE EXPRESSION DATA TEMPORAL SEQUENCE USING PROBABALISTIC BOOLEAN NETWORKS

*Stephen Marshall, Le Yu*

Department of Electronic and Electrical Engineering, University of Strathclyde
Royal College Building, 204 George Street, G1 1XW, Glasgow, United Kingdom
phone: + (44)1415482199, fax: + (44)1415522487, email: l.yu@eee.strath.ac.uk s.marshall@eee.strath.ac.uk
web: www.eee.strath.ac.uk/ipg

## ABSTRACT

*This paper describes a new method for analysing gene expression temporal data sequences using Probabilistic Boolean Networks. Switch-like phenomena within biological systems result in difficulty in the modelling of gene regulatory networks. To tackle this problem, we propose an approach based on so called 'purity functions' to partition the data sequence into sections each corresponding to a single model with fixed parameters, and introduce a method based on reverse engineering for the identification of predictor genes and functions. Furthermore, based on the analysis of Macrophage gene regulation in the interferon pathway, we develop a new model extending the PBN concept for the inference of gene regulatory networks from gene expression time-course data under different biological conditions. In conjunction with this, a new approach based on constrained prediction and Coefficient of Determination to identify the model from real expression data is presented in the paper.*

## 1. INTRODUCTION

In recent years biological microarray technology has emerged as a high-throughput data acquisition tool in bioinformatics. It enables the measurement of the expression level of thousands of genes simultaneously in a cell at a series of time points in a specific biological process [1]. It has led to a dramatic revolution in the field of systems biology so that computational methods for modelling and simulating gene regulatory networks have been developed to study gene regulation. This has constituted a key aspect of genomic signal processing [2].

Recently, a modelling approach defined as Probabilistic Boolean Networks (PBNs) has been proposed for modelling gene regulatory networks [3]. This technology, an extension of Boolean Networks [4], is able to capture the time-varying deterministic dependencies as well as uncertainties relative to the governing model by a series of logic based predictor functions.

Cells maintain their phenotype stability until the phenotype is switched in response to an external stimulus. Under the stimulus, there are switch-like transitions observed within biological systems. Therefore, Context-Sensitive PBN modelling was proposed for inferring genetic regulatory networks in some gene expression data with different contexts for the cell [5]. The Context-Sensitive PBN model is a collection of Boolean Networks (BNs) with fixed Boolean functions set in a time-course. Our modelling approach is an extension of this concept.

For the construction of a model of gene regulatory behaviour for data under different biological conditions, it is essential to be able to partition the data into sections corresponding to different contexts of the underlying model. Therefore, we firstly proposed a method for partitioning the sample gene expression data with multi-context into different segments over a time horizon. Following partition, the individual constituent models may be fitted to each partitioned section respectively. A method is proposed by which the parameters of a PBN may be inferred directly from temporal gene expression data. In this paper we demonstrate the procedure by reverse engineering the process and recovering all the complexities of the generating model. After that, we apply modelling techniques to macrophage gene expression data under three different biological conditions. The results reveal that there exists a 'phase-changing' phenomenon on the regulator activities caused by underlying mechanisms or external inputs. To capture such phenomenon, we have extended the Context-Sensitive PBN concept to allow the network selection probability to vary with biological conditions. Finally, we have developed an approach using the Coefficient of Determination (COD) [6] for the inference of the model from the small amounts of experimental data.

Our experiments are carried out on two kinds of data. Firstly, we verify our technique by conducting experiments on a synthetic temporal gene expression data sequence generated by Genomic Signal Processing Laboratory based at Texas A&M University. According to the results of the analysis on these simulated data, we then developed a further experiment using real gene expression data supplied by the Scottish Centre for Genomic Technology and Informatics (GTI), based at University of Edinburgh. The paper is organized as follows. In Section 2, a description of an experiment on simulated data is provided. In section 3, we present the experiment on real data. Section 4 contains some concluding remarks.

## 2. VERIFICATION WITH SIMULATED DATA

The simulated data is modelled by a PBN [3] consisting of $M$ Boolean Networks (BNs) $BN^1 \ldots BN^M$. There are $n$ genes

$g_1,\ldots, g_n$ , the regulatory activities of which are controlled by a set of Boolean functions of $k$ genes, where $k<n$. The data is modelled through a sequential process whereby the state of the model at time $t$ is determined by its previous state at time $t-1$. It is a discrete-time stochastic process with the Markov property. The model parameters, that is the specific $k$ genes which control each gene and the set of functions used to determine their next states are fixed for substantial periods of time. Under external stimuli or hidden mechanisms, periodically a 'switch point' occurs where both the function and the $k$ predictor genes governing each gene change so that gene activity switches to another BN and therefore a different regulatory rule is adopted.

The first step in the process is to partition the time course data into so called 'pure' segments by finding these switch points. A pure segment corresponds to data generated from a single BN.

### 2.1 Partitioning Data

The important point to note here is that the characteristics of a BN may be determined by the pairwise transitions of the network. For a BN with fixed parameters the state of the network $S^t$ at time $t$, must always be followed by the same state, $S^{t+1}$, at time $t+1$ except when the transition is perturbed by mutation or noise. Therefore a state transition matrix compiled over a pure temporal data sequence and showing which state $S^{t+1}$, at time $t+1$ time follows $S^t$ at time $t$, will be sparse containing only one significant entry per line plus some entries due to noise. Typically the $N\times N$ $(N=2^n)$ transition matrix $Y(t1,t2)$would have the overall structure shown below,

$$\text{Current State } i=S^t \quad \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \overset{\text{Next State } j=S^{t+1}}{\begin{bmatrix} a_{11}, a_{12}, \ldots a_{1j}, \ldots a_{1N} \\ a_{21}, a_{22}, \ldots a_{2j}, \ldots a_{2N} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{i1}, a_{i2}, \ldots\ldots a_{ij}, \ldots a_{iN} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{N1}, a_{N2} \ldots a_{Nj} \ldots a_{NN} \end{bmatrix}}$$

where $a_{ij}$ indicates the number of state transitions from state $i$ to state $j$ during a time horizon i.e. between time step $t1$ and time step $t2$. State $i, j$ are in a finite state space $\{1, 2, 3,\ldots,2^n\}$. The transitions of the state in the system periodically cycle in between several attractors. Therefore, the majority of transitions will be zero and most non zero values will correspond to transitions $a_{ij}$ driven by the model, the remainder will be due to perturbation or noise.

Identification of one significant transition per state results in a one to one mapping linking pairs of successive gene states. $S^j = \psi(S^i)$ Assuming that the system consists of $n$ genes, $S=(g_1, g_2.....g_n)$ then its state at time $t+1$, may be written as a function of a subset of $k$ genes, where $k < n$.

$$S^{t+1}=(g_1, g_2.....g_n)^{t+1} \quad (1)$$

The state of each gene $g$ at time $t+1$ is a function of the genes at time $t$. More specifically it is a function of a subset of $k$ genes

$$S^{t+1}=(F_1((g_1,g_2\ldots g_n)^t),F_2((g_1, g_2\ldots g_n)^t),\ldots F_n((g_1, g_2.....g_n)^t))^{t+1} \quad (2)$$

Given sufficient data, the functions and the subset $k$ may be determined by Boolean reduction.

The objective is to identify points within the sequence where the model parameters change and between which the data is pure, i.e. derived from a single BN with fixed parameters. We now seek to define a purity measure which determines the likelihood that the data lying between two data points is 'pure'. The measure proposed is

$$\lambda(S(t_1,t_2)) = \frac{\sum_{i=1}^{N}[P_{ij} - Q_{ij}]}{\sum_i \sum_j a_{ij}} \quad (3)$$

where $P_{ij}$ is the value of the largest (major) value on each row $j$ in column $i$, $Q_{ij}$ is the value of the second peak and $a_{ij}$ is the value of the transition matrix.

Consider a data sequence with $T$ points generated from two BNs. The first section of the sequence, from $1$ to $t'$ $(t' \in [1,T])$ has been generated from the first BN and the remainder of the sequence from $t'+1$ to $T$ has been generated from the second BN. Suppose there is one switching point, the unknown switchover point is determined by splitting the sequence into two parts at point $z$. The value of $z$ is a variable which may take any value from $1$ to $T$ and it is an estimate of the value of $t'$. $z$ is varied and the data set before and after it are mapped into two different transition matrixes $W(z)$and $V(z)$. $W(z)=Y(1,z)$, $V(t)=Y(z+1,T)$. The definition of Matrix $Y$ has been shown in the left. The ideal purity factor is one which is maximised for both $W(z)$ and $V(z)$ when $z= t'$. This is shown in Figure 1. Where a data set consists of more than two pure sequences then it must be partitioned iteratively in order to determine the set of pure sequences.

Once the sequence has been successfully partitioned into pure segments then the process of identifying the functions and inputs can be carried out by Boolean reduction.

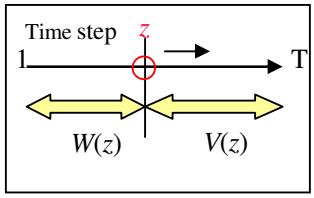


Fig 1. Data sequence divided by a sliding point $t$ and Transition matrices produced for data on each side of the partition

### 2.2 Identification of Predictor Genes and Functions

Having partitioned the gene state data into 'pure subsequences', it is now necessary to identify which subset, $k$ of the $n$ genes may best predict any given target gene. This was carried out by identifying the $k$ genes which best correlate with changes in the predicted gene. It is based on a function which is minimized when the changes in the $k$ genes specified most accurately coincide with changes in the predicted gene aggregated over each pure data subsequence.

In the pure subsequence, the next state of gene $g'_i$ is a function of $k$ genes,

$$g'_i = f_i(j_i(1), j_i(2),.., j_i(k)) \tag{4}$$

where $j_i(k)$ are selection functions determining which $k$ from $n$ genes are used as inputs to the predictor function.

According to the state transition matrix for each segmented 'pure' subsequence of data, it is possible to create a current-next state table for each pure subsequence. If the states are written in terms of their individual genes then a simple *one to one* mapping is produced from $n$ genes to $n$ genes. This was set out in Eqn (1) and (2). In the current-next state table (Table 1), each line presents states of genes corresponding to gene activity profiles in the real measurements. It is clear that the next state of any gene $g'_i$ may be written as a function of all $n$ predictor genes. The problem is to determine which subset, $k$ out of the $n$ genes may best predict any given gene.

A cost function $R(k)$ is defined which is minimized when all the output gene values are the same for the same combination of $k$ predictor genes.

$$R(k) = \sum_{g_k} r_{g_k}(k) \tag{5}$$

The cost function $R(k)$ consists of $r_{g_k}(k)$ summed over every combination of the $k$ predictor genes, where $r_{g_k}(k)$ is defined below.

$$r_{g_k}(k) = \begin{cases} \sum g'_i & if \ \sum g'_i \le n/2 \\ n - \sum g'_i & if \ \sum g'_i > n/2 \end{cases} \tag{6}$$

| $g_1$ | $g_2$ | $g_3$ | .... | $g_{n-1}$ | $g_n$ | $g'_1$ | .... | $g'_n$ |
|---|---|---|---|---|---|---|---|---|
| | | | .... | | | | .... | |
| 0 | 0 | 1 | .... | 1 | 0 | 1 | .... | 0 |
| 0 | 0 | 1 | .... | 1 | 1 | 0 | .... | 0 |
| .. | .. | .. | .. | .. | .. | .. | .... | .. |
| .. | .. | .. | .. | .. | .. | .. | .... | .. |
| 0 | 1 | 1 | .... | 1 | 0 | 0 | .... | 0 |
| 0 | 1 | 1 | .... | 1 | 1 | 1 | .... | 0 |
| | | | .... | | | | .... | |

Table 1 Current-Next state table

The value of $g'_i$ is either *0* or *1* as specified in the truth table. Hence the quantity $r_{g_k}(k)$ is minimized if the outputs for a particular combination of inputs $g_k$ are either all *0* or all *1*. There are $n!/(n-k)!k!$ ways of selecting $k$ from $n$ inputs. For small numbers of genes the $k$ inputs may be chosen by full search and the cost function evaluated for every combination. For larger sets they are chosen through genetic algorithms to minimize the cost function. In some cases the current-next state table is not fully defined. This means that the output is itself a don't care term. Tests have shown that the predictor genes may still be identified correctly even for *87.3%* of missing data [7].

Once the subset of $k$ predictor genes has been identified, the task of identification of predictor functions is a straightforward exercise in logic minimization [8]; made easier by the fact that $k$ is small.

### 2.3 Experiment Results

Experiments were conducted on synthetic time-course data, generated at Texas A&M University. The PBN consisted of *4* BNs with $n=8$ and $k \leqslant 4$. The data was processed using the method described. An example of the purity function value derived over a section of the data and clearly containing *5* switch points in is given in Figure 2. In this way the data was partitioned into pure segments. The segments were analyzed to determine the $k$ predictor genes and corresponding functions. Table 2 shows the resulting predictor genes and functions derived from the data for BN3 as compared to the actual functions. These were derived very accurately despite the fact that only *23%* of the transitions for BN3 were observed in the data.

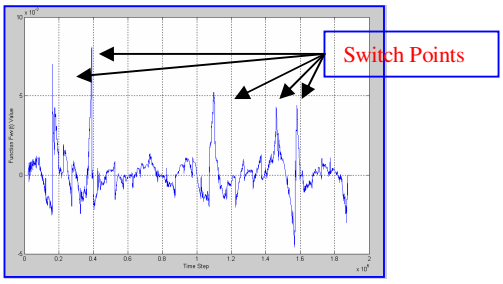


Fig.2. Purity function values for a 200,000 section of the data.

| BN3 (size 47661)<br>Observed transition state: 58 (23%) | | | | |
|---|---|---|---|---|
| Gene | Actual Predictor genes | Predicted Predictor Genes | Actual Functions | Predicted Functions |
| 1 | 1,2,6,8 | 1,2,6,8 | 0000010001101011 | 00000100011010** |
| 2 | 1,2 | 1,2 | 0100 | 0100 |
| 3 | 5 | 5 | 10 | 10 |
| 4 | 1,2,7,8 | 1,2,7,8 | 1111011010001100 | 1111011*100011** |
| 5 | 5 | 5 | 01 | 01 |
| 6 | 1,3,5,6 | 1,3,5,6 | 1000000100110000 | 100000010*11000* |
| 7 | 3,4 | 3,4 | 1000 | 1000 |
| 8 | 3,4,8 | 3,4,8 | 00110101 | 00110101 |

Table 2. The predictor genes and predictor functions in BN3

## 3. EXPERIMENTS ON REAL DATA

### 3.1 Data Analysis and Model Definition

In the experiments on real data, the gene expression data are taken from a hybridizations microarray study, in which bone marrow derived macrophages were exposed to three different biological conditions over 12 hours, with measurements made in 30 minute intervals. The three conditions were interferon treatment only (INFg), viral infection with interferon treatment (C3X_INFg) and viral infection only (C3X). All target genes and predictor genes come from a functional group containing 5 target genes in the data (Il12b, Cybb, G1p2, Itgam, Fcer2a). Figure 3 shows how the regulatory activities of the genes vary under the 3 different biological conditions. We found that most gene activities in interferon gamma treated samples were consistent with the consensus pathway (for example, Fcer2a=Irf4 $\oplus$ stat6 i.e. target gene Fcer2a is affected by the adding of Irf4 and stat6). However, this was less so for gene activities in viral infection with interferon treatment and interferon treatment without infection. Table 3 shows the proportions of which gene activities fit the known biological pathway from samples in 3 biological conditions respectively. It is clear that their proportions vary corresponding to the different biological conditions. The results on the analysis of macrophage gene expression data under three different biological conditions show that interferon treatment establishes the cognate pathway connections while infection leads to a limited engagement of the regulatory network. It revealed that there is a 'phase-changing' phenomenon on the regulator activities caused by underlying mechanisms or external inputs. To capture such phenomenon, we define a dynamic model which extends the PBN concept to allow network selection probability to vary with biological conditions.

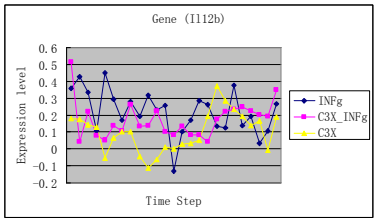


Fig.3. Normalized log-ratios (in *log2(test/control)*) of *gene (Il12b)* under 3 biological conditions.

| Target genes and logic predictor functions | % of gene activities fitting known biological pathway | | |
|---|---|---|---|
| | INFg | C3X_INFg | C3X |
| Fcer2a (Fcer2a=Irf4 $\oplus$ stat6) | 84% | 80% | 52% |
| Itgam (Itgam=Irf8) | 88% | 52% | 36% |
| Il12b (Il12b=(Irf8 $\cap$ Irf1) $\cup$ Irf2) | 84% | 64% | 52% |
| G1p2 (G1p2=(Irf8 $\cap$ Sfpi-1 $\cap$ Irf4 $\cap$ Irf2) $\cup$Irf1) | 76% | 68% | 16% |
| Cybb (Cybb=Irf8 $\cap$ Sfpi1 $\cap$ Irf1 $\cap$ Crebbp) | 92% | 84% | 64% |

Table 3. The proportions of gene activities fitting the known biological pathway from samples in 3 biological conditions

The proposed model consists of a number of PBNs and the system behaviour switches between these on a stochastic basis. The PBNs inferred from the data in 3 different biological conditions have the same structure but different parameters (the function selecting probability).

### 3.2 Model Identification

Next we present an approach based on reverse engineering for the inference of this model from the data. To simulate the real condition, a small number of synthetic data from the supposed model under 3 biological conditions will be used for the model identification. The number of target genes is set as 7. In the experiment, because of the limited amount of data, the predictor functions are constrained to fall into the class of functions known as Canalizing functions [4]. The maximum network connection is defined as $k$=3. By analyzing the data and employing logic reduction techniques from digital electronics the predictor functions and the predictor genes for each target gene may be recovered. After that, we select as the predictors, the $k$ genes which best correlate with the data for each of the 3 biological conditions. In some cases, several genes fit equally well. Table 4 shows the dominant predictor genes inferred from the data in the 3 biological conditions respectively. The numbers of the entry in the table represent the indices of the genes.

After that, the selection probabilities of predictor functions $a_{jm}^{(i)}$ ($i$ is the index of target gene, $j$ is the index of the predictor function, $m$ is the index of the biological condition) in the 3 biological conditions will be determined. Because of the limited amount of data available, a method based on the Coefficient of Determination (COD) [6] is used. The coefficient measures the degree to which the transcriptional levels of an observed gene set can be used to improve the prediction of the transcriptional state of a target gene relative to the best possible prediction in the absence of observations.

| The index of genes | Predictor genes of the constituent predictor functions | | |
|---|---|---|---|
| | Biological Condition1 | Biological Condition2 | Biological Condition3 |
| $g_1$ | 1,6,7 | 1,3,5 | 3,5,7 |
| $g_2$ | 1,3,6 | 2,3,5 | 1,2,5 |
| $g_3$ | 1,2,7 | 2,3,7 | 2,3,7 |
| | 1,2,5 | | |
| | 1,2,3 | | |
| $g_4$ | 1,4,5 | 4,6,7 | 1,4,5 |
| | | 4,5,7 | |
| | | 3,4,7 | |
| | | 2,4,7 | |
| | | 1,4,7 | |
| $g_5$ | 4,5,6 | 3,4,5 | 1,5,6 |
| $g_6$ | 4,6,7 | 1,6,7 | 2,5,6 |
| | 1,3,7 | | 1,3,6 |
| | 1,2,7 | | |
| $g_7$ | 4,6,7 | 1,4,6 | 3,4,7 |
| | 3,6,7 | | |

Table 4. The predictor genes inferred from samples in 3 biological conditions.

Define $g_i$ as the target gene, the predictor genes come from a set of genes $g_1, g_2, \ldots g_n$ . For $g_i$, the COD of a selected predictor gene set $\boldsymbol{g}_j^{(i)}$ is given as

$$\theta_j^i = (\varepsilon_i - \varepsilon_{opt})/\varepsilon_i \qquad (1)$$

where $\varepsilon_i$ is denoted as the error of best estimation for $g_i$. $\varepsilon_{opt}$ is the optimal minimum error. From the training data in all 3 biological conditions, $d_j^{(i)}$, which is the selection probability of a certain predictor function $f_j^{(i)}$ to target gene $g_i$ can be inferred from COD values with the range of *0* to *1* by the following expression [3] $d_j^{(i)} = \theta_j^i \Big/ \sum_{k=1}^{N_{(i)}} \theta_k^i$ (2)

where $N_{(i)}$ is the number of possible predictor functions for target gene $g_i$ . Similarly, $a_{jm}^{(i)}$, the selection probability of the predictor function under a certain biological condition can be inferred from the data. We define $c_m^{(i)}$ as the context selection probability for the function sets in the 3 biological conditions. $d_j^{(i)}$ can be represented as

$$d_j^{(i)} = \sum_{m=1}^{M} c_m^{(i)} \times a_{jm}^{(i)} \quad and \sum_{m=1}^{M} c_m^{(i)} = 1 \qquad (3)$$

where $M$ is the number of the biological conditions ($M=3$ in our experiment). Therefore, $c_m^{(i)}$ can be obtained from equation *(3)*. We define mean square error (MSE) $\Delta w_m^{(i)}$ as the cost function of

$$\Delta w_m^{(i)} = E\,[\,(d_j^{(i)} - d'^{(i)}_j)^2\,] \qquad (4)$$

where $d'^{(i)}_j$ is the value of equation *(3)* given a series of $c_m^{(i)}$. Once $\Delta w_m^{(i)}$ displays the minimum value, the optimal realization of $c_m^{(i)}$ will have been found.

### 3.3 Experiment Results

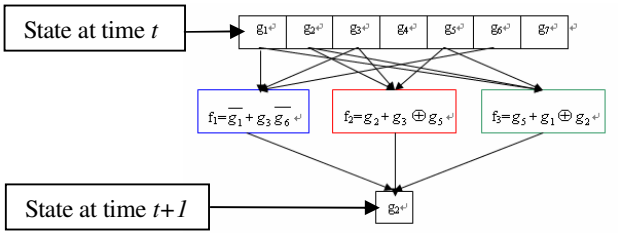


Fig.4. An example of the selected predictor functions under *3* biological conditions for $g_2$

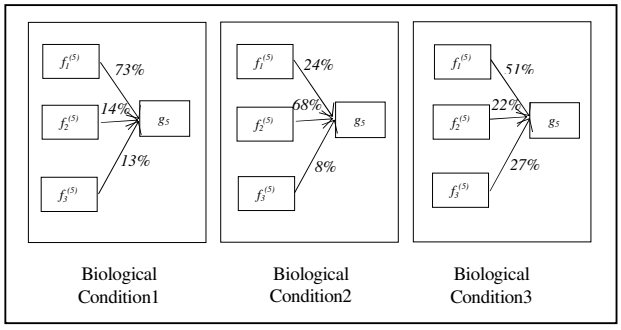


Fig. 5. An example of $a_{jm}^{(i)}$, the selection probability of predictor functions under 3 biological conditions for $g_5$ .

Figure 4 shows the constituent predictor functions which are in the class of canalizing functions for the target gene $g_2$. Figure 5 shows the selection probabilities of 3 predictor functions for the target gene $g_5$. The probabilities display quite different values for the different biological conditions. In this model, the structure of the constituent Boolean networks remains stable, whereas the network selection probabilities change in response to the different biological conditions. It implies a fixed finite discrete state space but different stationary distributions. This model keeps the rule-based structure while allowing for uncertainty. It is suitable for the inference of the gene regulatory network from gene expression data in different biological conditions.

## 4. CONCLUSION

In this paper, a method of fitting multiple Boolean Networks to time course gene expression profiles has been successfully applied. The parameters of the models used to generate the data have been accurately recovered despite the fact that only a small percentage of all possible transitions have been observed from simulated data. We also proposed a novel approach to model the gene regulator activities under different biological conditions. A method identifying the model parameters from time course gene expression profiles has been successfully applied. All these results are applied in the context of pathway biology to the analysis of an interferon gene interaction network.

## ACKNOWLEDGEMENT

We would like to acknowledge, Yufei Xiao and Edward R. Dougherty at Genomic Signal Processing Laboratory, Texas A&M University, Thorsten Forster and P. Ghazal at Scottish Centre for Genomic Technology and Informatics, University of Edinburgh for their support and help in this research.

## REFERENCES

[1] DJ Duggan, M Bitter, Y Chen, P Meltzer and JM. Trent (1999) Expression Profiling using cDNA Microarrays. Natural Genetics. 21(Sup1), 10-14

[2] ER Dougherty, A Datta, C Sima, (2005) *Research Issues in Genomic Signal Processing*. IEEE Signal Processing Magazine. Nov, 46-68.

[3] I Shmulevich, ER Dougherty, S Kim, W Zhang, (2002a) *Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks*. Bioinformatics. 18(2) 261–274.

[4] SA Kauffman, (1993) *The Origins of Order: Self-organization and Selection in Evolution*. New York: Oxford University Press.

[5] ER Dougherty, A Datta, (2005) *Genomic Signal Processing: Diagnosis and Therapy*. IEEE Signal Processing Magazine. Jan, 107-112.

[6] ER Dougherty, S Kim, Y Chen, (2000) *Coefficient of Determination in Nonlinear Signal Processing*. Signal Proc 80, 2219-2235.

[7] S Marshall, L Yu, Y Xiao, ER Dougherty (2006), *Inference of a Probabilistic Boolean Network from a Single Observed Temporal Sequence*, submitted *EURASIP Bio Journal* in July 2006.

[8] CH Roth, (1995) *Fundamentals of Logic Design*, 4th edition, Brooks Cole, 1995.