

Building Genetic Networks for Gene Expression Patterns

Wai-Ki Ching¹, Eric S. Fung², and Michael K. Ng³

The University of Hong Kong, Pokfulam Road, Hong Kong

Abstract. Building genetic regulatory networks from time series data of gene expression patterns is an important topic in bioinformatics. Probabilistic Boolean networks (PBNs) have been developed as a model of gene regulatory networks. PBNs are able to cope with uncertainty, corporate rule-based dependencies between genes and uncover the relative sensitivity of genes in their interactions with other genes. However, PBNs are unlikely used in practice because of huge number of possible predictors and their computed probabilities. In this paper, we propose a multivariate Markov chain model to govern the dynamics of a genetic network for gene expression patterns. The model preserves the strength of PBNs and reduce the complexity of the networks. Parameters of the model are quadratic with respect to the number of genes. We also develop an efficient estimation method for the model parameters. Simulation results on yeast data are given to illustrate the effectiveness of the model.

1 Introduction

An important focus of genomic research concerns understanding the mechanism in which cells execute and control the huge number of operations for normal functions, and also the way in which the cellular systems fail in disease. One of the possible reasons is the cell cycle. Cells in the body alternately divide in mitosis and interphase, and this sequence of activities exhibited by cells is called the cell cycle. Since all eukaryotic cells have important physical changes during the cell cycle, if we have better understanding of when and where the cell cycle occur, a higher prediction accuracy for cell activities can be obtained. Models based on methods such as neural network, non-linear ordinary differential equations, Petri nets have been proposed for such problem, see for instance Smolden et. al (2000), DeJong (2002) and Bower (2001).

Boolean network is commonly used for modeling genetic regulatory system because it can help discovering underlying gene regulatory mechanisms. The Boolean network is a deterministic network and it was first introduced by Kauffman (1969). In this network, each gene is regarded as a vertex of the network and is quantized into two levels only (over-express (1) or under-express (0)). However, the deterministic assumption is not practical in both biological and empirical aspect. In the biological aspect, an inherent determinism assumes an environment without uncertainty. In the empirical aspect, sample noise and relatively small amount of samples may cause incorrect results in logical rules. Later,

Shmulevich *et al.* (2002) proposed Probabilistic Boolean Networks (PBNs) that can share the appealing rule-based properties of Boolean networks and it is robust in the face of uncertainty. The dynamics of their PBNs can be studied in the context of standard Markov chain. However, the number of parameters in a PBN grows exponentially with respect to the number of genes n , and therefore heuristic methods are needed for model training, see for instance Akutsu *et al.* (2000).

The main contribution of this paper is to propose a multivariate Markov chain model (a stochastic model) which can capture both the intra- and inter-transition probabilities among the gene expression sequences. Moreover, the number of parameters in the model is only $O(n^2)$. We develop efficient model parameters estimation methods based on linear programming. We also propose a method for recovering the structure and rules for a given Boolean network.

The rest of the paper is organized as follows. In Section 2, we give a review on PBNs. In Section 3, we propose the multivariate Markov chain model. We then present estimation methods for model parameters. An example is given to illustrate the actual parameters estimation procedure in Section 4. In Section 5, we apply the proposed model and method to gene expression data of yeast, and illustrate the effectiveness of the new model. Finally, concluding remarks are given to address further research issues in Section 6.

2 Probabilistic Boolean Networks

In Shmulevich *et al.* (2000), PBNs are proposed and the model can be written as follows:

$$\mathbf{F}_i = \{f_j^{(i)}\}_{j=1, \dots, l(i)}, \quad 1 \leq i \leq n$$

where n is the number of genes, the predictor $f_j^{(i)}$ is a possible function determining the expression level (0 or 1) of the i -th gene, and $l(i)$ is the number of possible functions for the i -th gene. Let $c_j^{(i)}$ be the probability that the j th predictor $f_j^{(i)}$ is chosen to predict the i -th gene. This probability can be estimated by using the Coefficient of Determination (COD), see Dougherty *et al.* (2000).

For any given time point, the expression level of the i -th gene is determined by one of the possible predictors $f_j^{(i)}$ for $1 \leq j \leq l(i)$. The probability of a transition from $\mathbf{v}(t)$ to $\mathbf{v}(t+1)$ can be obtained as:

$$\prod_{i=1}^n \left[\sum_{k=1}^{l(i)} \left\{ c_k^{(i)} : f_k^{(i)}(\mathbf{v}(t)) = v_i(t+1) \right\} \right]$$

where $\mathbf{v}(t) = (v_1(t), \dots, v_n(t))$ and $v_l(t) = 0$ or 1. The degree of influence of the j -th gene to the i -th gene can be estimated by:

$$l_j(v_i) = \sum_{k=1}^{l(i)} \text{Prob}\{f_k^{(i)}(v_1, \dots, v_{j-1}, 0, v_{j+1}, \dots, v_n) \neq f_k^{(i)}(v_1, \dots, v_{j-1}, 1, v_{j+1}, \dots, v_n)\} \times c_k^{(i)},$$

see Shmulevich *et al.* (2000). We remark that for each set of $f^{(i)} = \{f_j^{(i)}\}$ with $1 \leq i \leq n$, the maximum number of predictors is equal to 2^{2^n} as $1 \leq l(i) \leq 2^{2^n}$, it is also true for its corresponding set of prediction probabilities $\{c_k^{(i)}\}$. It is almost not practical for using this model due to its complexity and the imprecision of parameters based on limited sample size.

3 The Multivariate Markov Chain Model

In this paper, the gene sequence in PBNs $\{v(1), \dots, v(T)\}$ is logically represented as a sequence of vectors V_1, \dots, V_T , where T is the length of the sequence, and $V_j = E_{k+1}$ (E_{k+1} is the unit column vector with the $(k+1)$ -th entry being one) if $v(j) = k$ (i.e., it is in the expression level k , $k = 0$ or 1). A first-order discrete-time Markov chain satisfies the following relationship:

$$\begin{aligned} & \text{Prob}(V_{t+1} = E_{v(t+1)} \mid V_0 = E_{v(0)}, \dots, V_t = E_{v(t)}) \\ &= \text{Prob}(V_{t+1} = E_{v(t+1)} \mid V_t = E_{v(t)}). \end{aligned}$$

These probabilities are assumed to be independent of t and can be written as

$$p_{ij} = \text{Prob}(V_{t+1} = E_{i+1} \mid V_t = E_{j+1}), \quad \forall i, j \in \{0, 1\}.$$

The matrix P , formed by placing p_{ij} in row i and column j is called the transition probability matrix.

In our proposed multivariate Markov chain model with a network of n genes, we assume the following relationship among the genes:

$$\mathbf{V}_{t+1}^{(j)} = \sum_{k=1}^n \lambda_{jk} P^{(jk)} \mathbf{V}_t^{(k)}, \quad j = 1, 2, \dots, n \quad (1)$$

where $\lambda_{jk} \geq 0$ for $1 \leq j, k \leq n$ and $\sum_{k=1}^n \lambda_{jk} = 1$, and $\mathbf{V}_t^{(i)}$ is the expression level probability distribution of the i -th gene at the time t . The expression level probability distribution of the j -th gene at the time $t+1$ depends on the weighted average of $P^{(jk)} \mathbf{V}_t^{(k)}$. Here $P^{(jk)}$ is a transition probability matrix from the expression level of the k -th gene to the expression level of the j -th gene. In matrix form, we write

$$\mathbf{V}_{t+1} \equiv \begin{pmatrix} \mathbf{V}_{t+1}^{(1)} \\ \mathbf{V}_{t+1}^{(2)} \\ \vdots \\ \mathbf{V}_{t+1}^{(n)} \end{pmatrix} = \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}P^{(12)} & \dots & \lambda_{1n}P^{(1n)} \\ \lambda_{21}P^{(21)} & \lambda_{22}P^{(22)} & \dots & \lambda_{2n}P^{(2n)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}P^{(n1)} & \lambda_{n2}P^{(n2)} & \dots & \lambda_{nn}P^{(nn)} \end{pmatrix} \begin{pmatrix} \mathbf{V}_t^{(1)} \\ \mathbf{V}_t^{(2)} \\ \vdots \\ \mathbf{V}_t^{(n)} \end{pmatrix} \equiv Q \mathbf{V}_t.$$

Although each column sum of Q is not equal to one, we still have the following two propositions, Ching *et al.* (2002).

Proposition 1. *If $\lambda_{jk} > 0$ for $1 \leq j, k \leq n$, then the matrix Q has an eigenvalue equal to 1 and the eigenvalues of Q have modulus less than or equal to 1.*

Proposition 2. *Suppose that $P^{(jk)}$ ($1 \leq j, k \leq n$) are irreducible and $\lambda_{jk} > 0$ for $1 \leq j, k \leq n$. Then there is a vector $\hat{\mathbf{V}} = [\hat{\mathbf{V}}^{(1)}, \hat{\mathbf{V}}^{(2)}, \dots, \hat{\mathbf{V}}^{(n)}]^T$ such that $\hat{\mathbf{V}} = Q\hat{\mathbf{V}}$ and $\sum_{i=1}^m [\hat{\mathbf{V}}^{(j)}]_i = 1$, $1 \leq j \leq n$.*

4 Model Parameters Estimation

In this section, we propose numerical methods to estimate $P^{(jk)}$ and λ_{jk} . For the j -th and the k -th genes, we estimate the transition probability matrix $P^{(jk)}$ by the following method by counting the transition frequency from the expression level i_k of the k -th gene at time t to the expression level i_j of the j -th gene at time $t+1$. After the usual normalization, we obtain an estimate of the transition probability matrix $\hat{P}^{(jk)}$. In the model, we require to estimate n^2 of 2-by-2 transition probability matrices in the multivariate Markov chain model. After all $P^{(jk)}$ are estimated, we can estimate the parameters λ_{jk} based on $P^{(jk)}$.

The stationary vector $\hat{\mathbf{V}}$ can be estimated from the sequences of gene expression levels by computing the proportion of the occurrence of each expression level of the gene. Let us denote it by $\hat{\mathbf{V}} = (\hat{\mathbf{V}}^{(1)}, \hat{\mathbf{V}}^{(2)}, \dots, \hat{\mathbf{V}}^{(n)})^T$. In view of Proposition 2, we therefore expect $\hat{Q}\hat{\mathbf{V}} \approx \hat{\mathbf{V}}$. From this approximation, it suggests one possible way to estimate the parameters $\lambda = \{\lambda_{jk}\}$ as follows. For each j , we solve the following optimization problem:

$$\min_{\lambda} \max_i \left| \left[\sum_{k=1}^n \lambda_{jk} \hat{P}^{(jk)} \hat{\mathbf{V}}^{(k)} - \hat{\mathbf{V}}^{(j)} \right]_i \right|$$

subject to $\sum_{k=1}^n \lambda_{jk} = 1$ and $\lambda_{jk} \geq 0$ for $1 \leq j, k \leq n$. Here $[\cdot]_i$ denotes the i th entry of the vector. This minimization problem can be formulated as a linear programming problem for each j , Ching *et al.* (2002). Since there are n independent linear programming (LP) problems of each being n constraints and n variables, the expected total computational complexity of solving such LPs is of $O(n^4)$, see for instance Fang and Puthenpura (1993).

4.1 An Example

Consider the following two binary sequences:

$$s_1 = \{0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0\} \quad \text{and} \quad s_2 = \{1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1\}.$$

We obtain the transition frequency matrix and probability matrix:

$$F^{(21)} = \begin{pmatrix} 5 & 2 \\ 3 & 1 \end{pmatrix} \quad \text{and} \quad \hat{P}^{(21)} = \begin{pmatrix} \frac{5}{8} & \frac{2}{3} \\ \frac{3}{8} & \frac{1}{3} \end{pmatrix}, \quad (\text{after normalization})$$

Similarly, the others can be computed and are given as follows:

$$\hat{P}^{(11)} = \begin{pmatrix} \frac{3}{4} & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{3} \end{pmatrix}, \quad \hat{P}^{(12)} = \begin{pmatrix} \frac{5}{7} & \frac{3}{4} \\ \frac{3}{7} & \frac{1}{4} \end{pmatrix} \quad \text{and} \quad \hat{P}^{(22)} = \begin{pmatrix} \frac{4}{7} & \frac{3}{4} \\ \frac{3}{7} & \frac{1}{4} \end{pmatrix},$$

Moreover, we get $\hat{\mathbf{V}}_1 = (\frac{3}{4}, \frac{1}{4})^T$ and $\hat{\mathbf{V}}_2 = (\frac{7}{12}, \frac{5}{12})^T$. After solving the LPs, the multivariate Markov chain model of the two binary sequences are given by

$$\begin{cases} \mathbf{V}_{t+1}^{(1)} = 0.5\hat{P}^{(11)}\mathbf{V}_t^{(1)} + 0.5\hat{P}^{(12)}\mathbf{V}_t^{(2)} \\ \mathbf{V}_{t+1}^{(2)} = 1.0\hat{P}^{(21)}\mathbf{V}_t^{(1)} + 0.0\hat{P}^{(22)}\mathbf{V}_t^{(2)}. \end{cases} \quad (2)$$

Based on the above models, we can estimate the probability $c_j^{(i)}$ of the predictor $f_j^{(i)}$ and the degree of influence of the j -th gene to the i -th gene. Let $X_{i_1, \dots, i_n}^{(d)}$ be the conditional probability vector of the d th gene when the previous expression level of the genes are i_1, i_2, \dots, i_n respectively. From (2), we obtain

$$X_{0,0}^{(1)} = (\frac{41}{56}, \frac{15}{56})^T, \quad X_{0,1}^{(1)} = (\frac{3}{4}, \frac{1}{4})^T, \quad X_{1,0}^{(1)} = (\frac{29}{42}, \frac{13}{42})^T, \quad X_{1,1}^{(1)} = (\frac{17}{24}, \frac{7}{24})^T.$$

and

$$X_{0,0}^{(2)} = X_{0,1}^{(2)} = (\frac{5}{8}, \frac{3}{8})^T, \quad X_{1,0}^{(2)} = X_{1,1}^{(2)} = (\frac{2}{3}, \frac{1}{3})^T$$

for the above two sequences. For example,

$$X_{1,0}^{(2)} = \sum_{k=1}^2 \lambda_{2k} \hat{P}^{(2k)} E_{i_k} = 1.0 \times \hat{P}^{(21)} E_{i_1} + 0.0 \times \hat{P}^{(22)} E_{i_2} = \hat{P}^{(21)}(0, 1)^T = (\frac{2}{3}, \frac{1}{3})^T.$$

With these probabilities, we obtain the probabilities $c_j^{(i)}$ as in the following table:

$v_1 \ v_2$	$f_1^{(1)}$	$f_2^{(1)}$	$f_3^{(1)}$	$f_4^{(1)}$	$f_5^{(1)}$	$f_6^{(1)}$	$f_7^{(1)}$	$f_8^{(1)}$	$f_9^{(1)}$	$f_{10}^{(1)}$	$f_{11}^{(1)}$	$f_{12}^{(1)}$	$f_{13}^{(1)}$	$f_{14}^{(1)}$	$f_{15}^{(1)}$	$f_{16}^{(1)}$
0 0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0 1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1 0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1 1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
$c_j^{(1)}$	0.27	0.11	0.12	0.05	0.08	0.04	0.04	0.02	0.1	0.04	0.04	0.02	0.03	0.01	0.02	0.01

For instance,

$$c_6^{(1)} = [X_{0,0}^{(1)}]_1 \times [X_{0,1}^{(1)}]_2 \times [X_{1,0}^{(1)}]_1 \times [X_{1,1}^{(1)}]_2 \approx 0.04.$$

Since $\lambda_{22} = 0$, the set of predictors for the second sequence can be reduced and their corresponding probabilities are given in the following table:

$v_1 \ v_2$	$f_1^{(2)}$	$f_2^{(2)}$	$f_3^{(2)}$	$f_4^{(2)}$
0 —	0	0	1	1
1 —	0	1	0	1
$c_j^{(2)}$	0.42	0.2	0.25	0.13

From the above two tables, the best predictors are $f_1^{(1)}$ and $f_1^{(2)}$ for the first and the second sequences respectively. We can also obtain the degree of influence of the i th sequence to the j th sequence ($i, j = 1, 2$) as follows:

$$l_1(v_1) = 0.4, \quad l_1(v_2) = 0.45, \quad l_2(v_1) = 0.4, \quad l_2(v_2) = 0.$$

For instance,

$$\begin{aligned} l_1(v_1) &= \sum_{k=1}^{l(1)} \text{Prob}\{f_k^{(1)}(0, v_2) \neq f_k^{(1)}(1, v_2)\} \cdot c_k^{(1)} \\ &= 0(0.27) + \frac{1}{2}(0.11) + \frac{1}{2}(0.12) + 0.05 + \frac{1}{2} = 0.4. \end{aligned}$$

According to the calculated values $l_i(v_j)$, we know that the first sequence somehow determine the second sequence. However, this phenomena is already illustrated by the fact that $\lambda_{22} = 0(\lambda_{21} = 1)$ in the multivariate Markov chain model.

4.2 Fitness of the Model

We note that the multivariate Markov chain model presented here is a stochastic model. Given all the state vectors $\mathbf{V}_t^{(k)}$ with $k = 1, \dots, n$, the state probability distribution $\mathbf{V}_{t+1}^{(k)}$ can be estimated by using (1). According to this state probability distribution, one of the prediction methods for the j -th sequence at time $t + 1$ can be taken as the state with the maximum probability, i.e., $\hat{\mathbf{V}}(t + 1) = j$ if $[\hat{\mathbf{V}}(t + 1)]_i \leq [\hat{\mathbf{V}}(t + 1)]_j$ for all $1 \leq i \leq 2$. By making use of this procedure, our multivariate Markov chain model can be used to uncover the rules (build a true table) for a PBNs. To evaluate the performance and effectiveness, the prediction accuracy of all individual sequences r

$$r = \frac{1}{nT} \times \sum_{i=1}^n \sum_{t=1}^T \delta_t^{(i)} \times 100\%, \quad \text{where} \quad \delta_t^{(i)} = \begin{cases} 1, & \text{if } \hat{\mathbf{v}}_i(t) = \mathbf{v}_i(t) \\ 0, & \text{otherwise.} \end{cases}$$

5 Gene Expression Data of Yeast

Genome transcriptional analysis is an important analysis in medicine and bioinformatics. One of the applications of genome transcriptional analysis is used for eukaryotic cell cycle in yeast. If we have better understanding of when and where the cell cycle occurs, a higher prediction accuracy for cell activity can be obtained. Several biological changes associated with the cell cycle activities, discovering this process gives important information for internal standard comparison of gene activity over time. Hartwell and Kastan (1994) showed that without appropriate cell cycle regulation leads to genomic instability, especially in etiology of both hereditary and spontaneous cancers, instances in Wang *et al.* (1994); Hall and Peters (1996). Raymond *et al.* (1998) examined the present of cell cycle-dependent periodicity in 6220 transcripts and found that cell cycle appears in about 7% of transcripts. Those transcripts are then extracted for further investigation. When the time course was divided into early G1, late G1, S, G2 and M phase based on the size of the bugs and the cellular position of the nucleus, the result showed that more than 24% of transcripts are directly

adjacent to other transcripts in the same cell cycle phase. The data set used in our study is the selected set from Yeung and Ruzzo (2001). In the discretization, if an expression level is above (below) a certain standard deviation from the average expression of the gene, it is over-expressed (under-expressed) and the corresponding state is 1 (0).

There are two problems of using PBN for such data set. The first is that the number of genes is too large and the PBN complexity is too high. The second is that the length of transcript (17 samples) is too short and therefore almost all values of $c_j^{(i)}$ are equal to 0 by using the method of COD.

The construction of the multivariate Markov chain models for such data set only requires about 0.1 second with CPU=AMD 1800+ and RAM=512Mb. This demonstrates the proposed method is quite efficient. In our study, we assume that there is not any prior knowledge about the genes. Therefore, in the construction of the multivariate Markov chain models, we consider each target gene can be related to other genes. Based on the values of λ_{ij} in our model, we can determine the occurrence of cell cycle in transcript j , i.e., the presence of inter-relationship between transcript j and the other transcripts in different phases. And we find that such cell cycle appears in 93% of the target genes in the multivariate Markov chain models. Some of the results are illustrated in the following table:

Name of target transcript	Cell cycle phase	Length of cell cycle	Related transcripts (its phase, λ_{ij} , level of influence)
YDL101c	late G1	1	YMR031c(early G1,1.00,1.00)
YPL127c	late G1	2	YDL101c (late G1,0.33,0.38) YML027w (late G1,0.33,0.39) YJL079c (M,0.33,0.38)
YLR121c	late G1	3	YPL158c (early G1,0.33,0.42) YDL101c (late G1,0.33,0.43) YKL069W (G2,0.33,0.43)
YLR015w	early G1	4	YKL113c (late G1,1.00,0.88)

In the above table, the last column displays the name of required transcripts for predicting the target transcript, the corresponding phase of required transcripts, their corresponding weightings λ_{ij} in the model, as well as an estimated value of the level of influence from related transcript to the target transcript. Although the level of influence can be estimated based on our model parameters, the computational cost increases exponentially respect to the value of n . We find in the table that the weighting λ_{ij} provides a reasonable measure for the level of influence. Finally, we present the prediction result of different lengths of cell cycle for the whole data set in the following table and the results show that the performance of the model is quite good.

Length of cell cycle phases required	No. of occurrence in this type of cell cycle (in %)	Average prediction accuracy (in %)
1	5	86
2	9	87
3	9	83
4	70	86

6 Concluding Remarks

In this paper, we proposed a multivariate Markov chain model. Efficient parameters estimation methods are presented. Experimental results on gene expression data of yeast are given to demonstrate the effectiveness of our proposed model. The model can be easily extended to the case when the gene expression data has more than two levels. The estimation method and prediction method can still be applied. Another direction for further research is to consider higher-order multivariate models, and develop estimation method for the model parameters and prediction method.

References

1. Akutsu, T., Miyano, S. and Kuhara, S. (2000). Inferring Qualitative Relations in Genetic Networks and Metabolic Pathways. *Bioinformatics*, **16**, 727-734.
2. Bower, J. (2001). Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, M.A.
3. Ching, W., Fung, E. and Ng, M. (2002). A Multivariate Markov Chain Model for Categorical Data Sequences and Its Applications in Demand Predictions. *IMA Journal of Management Mathematics*, **13**, 187-199.
4. de Jong, H. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *J. Comput. Biol.*, **9**, 69-103.
5. Dougherty, E.R., Kim, S. and Chen, Y. (2000). Coefficient of Determination in Nonlinear Signal Processing. *Signal Process*, **80**, 2219-2235.
6. Fang, S and Puthenpura, S. (1993). *Linear Optimization and Extensions*. Prentice-Hall, Englewood Cliffs, NJ.
7. Hall, M. and Peters, G. (1996). Genetic alterations of cyclins, cyclin-dependent kinases, and Cdk inhibitors in human cancer. *Adv. Cancer Res.*, **68**, 67-108.
8. Hartwell, L.H., and Kastan, M.B. (1994). Cell cycle control and cancer. *Science*, **266**, 1821-1828.
9. Kauffman, S. (1969). Metabolic Stability and Epigenesis in Randomly Constructed Gene Nets. *J. Theoret. Biol.*, **22**, 437-467.
10. Raymond J., Michael J., Elizabeth A., Lars S. (1998), A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*, **2**, 65-73.
11. Shmulevich, I., Dougherty, E., Kim S. and Zhang W. (2002). From Boolean to Probabilistic Boolean Networks as Models of Genetic Regulatory Networks. *Proceedings of the IEEE*, **90**, No.11, 1778-1792.
12. Smolen P., Baxter D. and Byrne J. (2000) Mathematical Modeling of Gene Network. *Neuron*, **26**, 567-580.
13. Wang, T.C., Cardiff, R.D., Zukerberg, L., Lees, E., Arnold, A., and Schmidt, E.V. (1994). Mammary hyperplasia and carcinoma in MMTV-cyclin D1 transgenic mice. *Nature*, **369**, 669-671.
14. Yeung, K. and Ruzzo, W. (2001). An Empirical Study on Principal Component Analysis for Clustering Gene Expression Data. *Bioinformatics*, **17**, 763-774.