G-network Modelling based Abnormal Pathway Detection in Gene Regulatory Networks

Haseong Kim[1], Rengul Atalay[2] and Erol Gelenbe[1]

[1] Department of Electrical & Electronic Engineering, Imperial College London, UK
[2] Department of Molecular Biology and Genetics, Bilkent University, Turkey

Abstract. Gene expression centered gene regulatory networks studies can provide insight into the dynamics of pathway activities that depend on changes in their environmental conditions. Thus we propose a new pathway analysis approach to detect differentially behaving pathways in abnormal conditions based on G-network theory. Using this approach gene regulatory network model parameters are estimated from normal and abnormal samples using optimization techniques with corresponding constraints. We show that in a "p53 network" application, the proposed method effectively detects anomalous activated/inactivated pathways related with MDM2, ATM/ATR and RB1 genes, which could not be observed from previous analyses of gene regulatory network normal and abnormal behaviour.

1 Introduction

One of the fundamental problems of biology is to understand complex gene regulatory networks (GRNs), and various mathematical and statistical models have been introduced for inference from GRNs [1]. Based on such networks, over-represented biological processes or pathways of a group of genes are identified by mapping them onto the gene ontology (GO) terms or regulatory structures [2]. These pathway analyses provide the annotations and functional insight of the group of genes which are usually determined by conventional statistical tests such as the t-test. However, these differentially expressed gene (DEG) derived analyses are limited in detecting defective pathways since they only observe the amount of expression of a gene itself rather than considering the flows of expression signals that communicate with neighboring genes.

Here we aim to detect the abnormal pathways of GRNs by modelling them using G-Networks [3] which is a probabilistic model of a system with special agents such as positive and negative customers, signals and triggers. In contrast to normal queuing networks, the negative customers of G-Networks describe the inhibitory effects of GRNs [4, 5]. G-networks have a product form solution which enables us to handle the dynamics of complex GRNs without heavy computation times. The parameters of the modelled GRN are inferred from normal samples with the assumed transition probabilities of gene expression signals. Then the transition probabilities of abnormal conditioned samples are estimated by minimizing the difference between the observed and predicted steady-state proba-

bilities with constraints. Finally, permutation tests are performed to determine the statistical significance of the estimated transition probabilities.

2 G-networks for gene regulatory networks

Following [4] consider the notion of a "packet" that contains the gene expression signals, and a network node that represents a gene consisting of a queue where its packets are stored and a server where the packets' fates are determined. Let λ_i^+ and λ_i^- be the positive and negative packet input rates to the ith node, respectively. μ_i is the packet firing rate (service rate) of the ith node. Furthermore we define $\mathbf{x} = \{x_1, ..., x_n\}$ a non-negative integer n-vector with $\mathbf{x}_i^+ = \{x_1, ..., x_i + 1, ..., x_n\}$, $\mathbf{x}_i^- = \{x_1, ..., x_i - 1, ..., x_n\}$, and $\mathbf{x}_{ij}^{+-} = \{x_1, ..., x_i + 1, x_j - 1, ..., x_n\}$. Let p_{ij}^+ and p_{ij}^- be the transition probabilities for packet motion from the ith node to the jth node as a positive and a negative packet, respectively. Note that a negative packet has the effect of disappearing after it destroys one packet of the target node, or it disappears also if it does not find a positive packet to destroy. Lastly, d_i denotes the probability that a packet leaves the system so that $\sum_{j=1}^{n}(p_{ij}^+ + p_{ij}^-) + d_i = 1$.

Consider now a random process $\mathbf{X}(t) = \{X_1(t), ..., X_n(t)\}$ where $X_i(t)$ is an integer-valued random variable representing the number of packets in the ith node at time $t \geq 0$. If $Pr(\mathbf{x}, t)$ is the probability that $\mathbf{X}(t)$ takes the value $\mathbf{x}$ at time t, then the G-network equations are:

$$
\begin{aligned}
Pr(\mathbf{x}, t + \Delta t) = \sum_{i=1}^{n} \Big[& (\lambda_i^+ \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^-, t) I(\mathbf{x}_i > 0) + (\lambda_i^- \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^+, t) \\
& + \sum_{j=1}^{n} \Big\{ (p_{ij}^+ \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_{ij}^{+-}, t) I(\mathbf{x}_j > 0) \\
& + (p_{ij}^- \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_{ij}^{++}, t) + (p_{ij}^- \mu_i \Delta t + o(\Delta)) Pr(\mathbf{x}_i^+, t) I(\mathbf{x}_j = 0) \\
& + \sum_{l=1}^{n} \left((p_{ijl} \mu_i \Delta t + o(\Delta t)) Pr(\mathbf{x}_{ijl}^{++-}, t) + (p_{jil} \mu_j \Delta t + o(\Delta t)) Pr(\mathbf{x}_{ijl}^{++-}, t) \right) I(\mathbf{x}_l > 0) \Big\} \\
& + (d_i \mu_i \Delta t + o(\Delta t)) Pr(\mathbf{x}_i^+, t) + (1 - (\lambda_i^+ + \lambda_i^- + \mu_i) \Delta t + o(\Delta t)) Pr(\mathbf{x}, t) \Big]
\end{aligned}
\tag{1}
$$

where $I(C)$ is 1 if C is true and 0 otherwise, and $o(\Delta t) \to 0$ as $\Delta t \to 0$. The complete equilibrium solution of (1) was given in [4]. Let q_i be the steady-state probability that the ith gene is activated:

$$
q_i = \min \left[1, \frac{\lambda_i^+ + \Lambda_i^+}{\mu_i + \lambda_i^- + \Lambda_i^-} \right] \tag{2}
$$

with

$$
\Lambda_i^+ = \sum_{j=1}^{n} q_j \mu_j p_{ji}^+ + \sum_{j,l=1, l \neq j}^{n} q_j q_l \mu_j p_{jli} \text{ and } \Lambda_i^- = \sum_{j=1}^{n} q_j \mu_j p_{ji}^- + \sum_{j,l=1, l \neq j}^{n} q_l \mu_l p_{lij}
$$

then the steady-state probability that there are x_i packets of ith node in each of the n cells is:

$$\lim_{t\to\infty} Pr(X_1 = x_1,\ ..., X_i = x_i,\ ... , X_n; t) = \Pi_{i=1}^{n} q_i^{x_i}(1 - q_i) \tag{3}$$

3 Abnormal edge detection

The packets in the G-network represent latent objects containing the gene expression signal, and we assume that the number of packets is proportional to the mRNA expression levels which are actually observable data. We also assume that the mRNA levels are observations of the steady-state. Therefore the steady-state probability that there is at least one mRNA of ith gene is $q_i = \dfrac{a_i}{a_i + 1}$ from (3) if we denote by a_i the average mRNA level (average queue length) of ith gene, also given by $a_i = q_i/(1 - q_i)$.

To determine the G-network parameters under normal conditions we use (2) where there are four sets of unknown parameters $p_{ji} = \{p_{ji}^+, p_{ji}^-, q_{jli}, q_{lij}\}$, λ_i^+, λ_i^-, and μ_i. We initially set $p_{ji} = 1/(n_j^{out} + 1)$ where n_j^{out} is the out-degree of gene j. We set the packet output rate μ_i based on the values of λ_i^+ and λ_i^- which are $\lambda_i^+ = 0.0062sec^{-1}$ and $\lambda_i^- = 0.002sec^{-1}$ ([6]), with $\mu_i = c \cdot n_i^{out}$ where c is a scaling constant. From (2) we have $q_i = f_i(\lambda_i^+, \lambda_i^-, \mu_i|\mathbf{q}_j, p_{ji})$ where $\mathbf{q} = (q_1,\ ..\ , q_n)$. Then c can be found by minimizing the following equation given the initial values of λ_i^+ and λ_i^-:

$$\tilde{c} = \arg\min_{c} \sum_{i} (q_i - f_i(c|\mathbf{q}, p_{ji}, \lambda_i^+, \lambda_i^-))^2 \tag{4}$$

Once each μ_i is determined, we can find the optimal positive input rate λ_i^+ which minimizes $(q_i - f_i(\lambda_i^+|\mathbf{q}, p_{ji}, \tilde{\mu}_i, \lambda_i^-))^2$ for each gene with the initial value λ_i^- and a constraint $0 \le \tilde{\lambda}_i^+ \le \mu_i + \lambda_i^- + \Lambda_i^- - \Lambda_i^+$. Then we determine $\tilde{\lambda}_i^-$ which produces exactly the same values of q_i.

3.1 Transition probabilities in abnormal conditions

In an abnormal condition, let q_i' be the steady-state probability that ith gene is activated and p_{ji}' be a packet transition probability from the ith gene to jth gene in the same condition. If there are k unknown p_{ji}' for ith gene, then we will denote them by a vector $\mathbf{p}_{ki}$. For the detection of the abnormally behaving pathways, $\mathbf{p}_{ki}$ needs to be estimated given the input ($\tilde{\lambda}_i^+$ and $\tilde{\lambda}_i^-$) and output ($\tilde{\mu}_i$) rates found in normal conditions. $\mathbf{p}_{ki}$ can be determined by minimizing the following sqaured error with two constraints, $0 \le \tilde{p}_{ji}'$ and $\sum_i \tilde{p}_{ji}' \le 1$:

$$\tilde{\mathbf{p}}_{ki}' = \arg\min_{\mathbf{p}_{\mathbf{ki}}^{(\mathbf{h})}} (q_i' - f_i(\mathbf{p}_{\mathbf{ki}}^{(\mathbf{h})}|\mathbf{q}_\mathbf{j}', \tilde{\lambda}_\mathbf{i}^+, \tilde{\lambda}_\mathbf{i}^-, \tilde{\mu}_\mathbf{i}))^\mathbf{2} \tag{5}$$

where $\mathbf{p}_\mathbf{k}^{(\mathbf{h})}$ is the hth hypothesis in the constrained parameter space. Our algorithm searches for the optimal solution iteratively with different initial starting values to reduce the possibility of remaining in a local minimum.

3.2 Permutation test for the estimated transition probabilities

When the estimated $\tilde{p}'_{ij}$ differs from its initially assumed value p_{ij}, it is necessary to determine if the difference is statistically significant. The null hypothesis of this test will be $\tilde{p}'_{ij} = p_{ij}$. To proceed with the test the set of samples is shuffled at random and divided into normal and abnormal groups with the same sample size of the original group. Then the proposed method is applied in the same way as the original data. Let M be the number of permutations and $\tilde{p}_{ij}^{(m)}$ be the estimated transition probability of the mth permutation. Then we can compute the emperical p-value of the $\tilde{p}'_{ij}$ as follows,

$$p\text{-value of } \tilde{p}'_{ij} = \begin{cases} \frac{1}{M}\sum_{m=1}^{M} I(p'_{ij} \leq \tilde{p}_{ij}^{(m)}) \text{ if } p'_{ij} > p_{ij} \\ \frac{1}{M}\sum_{m=1}^{M} I(p'_{ij} \geq \tilde{p}_{ij}^{(m)}) \text{ if } p'_{ij} \leq p_{ij} \end{cases}$$

where $I(C)$ is the indicator function. Thus if the p-value is less than α_2 then the null hypothesis is rejected. In our study, $\alpha_2 = 0.1$.

4 The p53 pathway

In order to evaluate our approach using experimental data, we selected the p53 pathway which is a well studied system in human cells whose most important feature is tumor suppression when DNA is damaged. The regulatory structure of the p53 pathway with 30 genes was constructed on the basis of the KEGG database, and we also downloaded two microarray mRNA expression datasets from GEO. The first dataset (GSE12941) consists of 10 non-tumor liver tissue and 10 hepatocellular carcinoma (HCC) samples. The second (GSE6222) is a dataset for the study of liver cancer progression in HCC. In this dataset, we use 2 normal and 10 HCC samples. Before applying the proposed method, the data was normalized and scaled with mean 3 so that the average number of mRNAs of a gene without its interactions in a single cell are assumed to be approximately 3, while the variance was scaled to 1. The gene input and output rates are assumed to be $0.0062sec^{-1}$ and $0.002sec^{-1}$ from [6] so that $0.0062/0.002 \approx 3$.

Figure 1 shows the average expression levels of genes in each dataset, and the corresponding p-values of t-tests which detect DEGs. The two datasets share 9 significant DEGs with 0.05 significance level while GSE12941 has 4 more DEGs. This similarity can be confirmed by observing their expression patterns in Figure 1. However interpreting the DEGs even when we know their regulatory structure is yet another challenge. Figure 2 shows the results from our proposed method. Despite the apparent lack of significance of p53 and MDM2 in the t-test, the $p53 - MDM$ feedback loop was clearly activated in cancer samples in both datasets. In [7], the $p53 - MDM2$ feedback loop appears to produce oscillatory expression patterns. Thus in terms of the system dynamics, the activation of two pathways between p53 and MDM2 in our result might be more appropriate than the activation of only one pathway from MDM2 to p53. One of the significant pathways in both datasets is $TP53 - IGFBP3 - IGF1$ [8]. Also our method

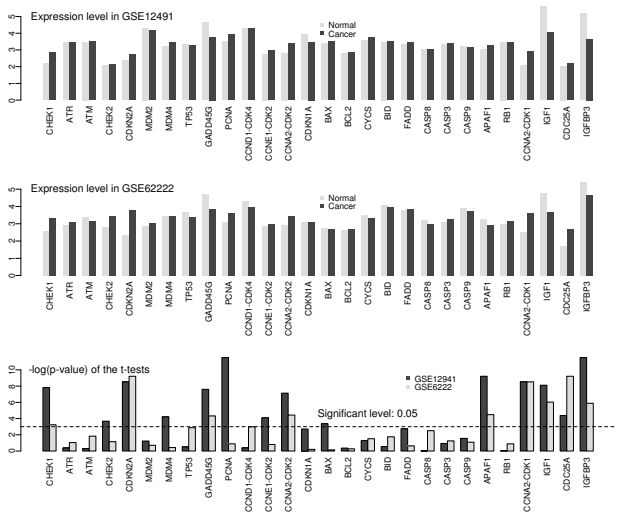

Fig. 1. Average mRNA expression levels of genes in two datasets, GSE12941 (top) and GSE6222 (middle). The p-values of the t-tests are shown in the bottom panel where the y-axis represents negative natural logarithms of the p-values. Note that multiple testing corrections are not applied in these t-tests.

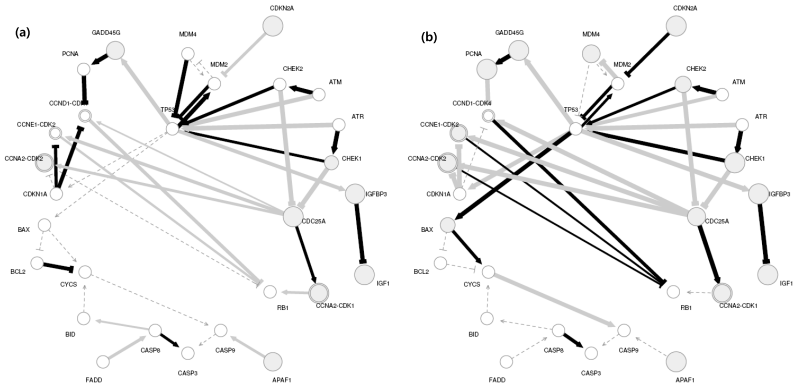

Fig. 2. The p53 network with the results of (a) GSE12941 and (b) GSE6222 dataset analysis. The solid line represents significantly activated (black) or inactivated (grey) pathways while the dashed line indicates non-significant pathways. Wider lines represent more significant pathways. The grey nodes are the selected DEGs from the t-test. The radius of a node is larger if its DEG is more significant with a 0.05 significant level. White nodes indicate non-significant genes.

properly detects two pathways, $ATM-CHEK2-TP53$ and $ATR-CHEK1-TP53$ as expected from [9] in both datasets, which cannot be detected merely by observing the p-values of the DEG test.

5 Discussion

We have proposed a new approach for detecting abnormal pathways in GRNs based on G-network modelling. This method provides an effective way to describe the flows of gene expression signals including negative or inhibitory effects on gene expression. Using some experimental data, we show that one advantage of our approach is that it can detect abnormal information flows in the dynamics of gene pathways. Thanks to existing G-network theory, the model uses a computationally tractable steady-state analysis and therefore does not require a large number of samples from time-dependent data. Moreover the analytic solution provided by G-network theory offers the possibility that our approach may be extended to very large-scale GRN systems.

In order to exploit this analytical tool, our work shows that a successful application of this method requires that the model be started with a reliable prior network structure based on real experimental data, or carefully calibrated GRN information. Though our intial experimental evaluation appears quite positive, further experimental studies will be needed to validate the proposed approach and apply it to attain biological meaningful and clinically useful results.

Acknowledgement

We would like to thank to Omer Abdelrahman and Zerrin Isik for helpful discussions.

References

1. Opgen-Rhein, R., Strimmer, K.: Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. BMC bioinformatics **8**(Suppl 2) (2007) S3
2. Beissbarth, T., Speed, T.: GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics (2004) 881
3. Gelenbe, E.: G-networks with triggered customer movement. Journal of Applied Probability (1993) 742–748
4. Gelenbe, E.: Steady-state solution of probabilistic gene regulatory networks. J. Theor. Biol Phys Rev E **76** (2007) 031903
5. Kim, H., Gelenbe, E.: Anomaly detection in gene expression via stochastic models of gene regulatory networks. BMC genomics **10**(Suppl 3) (2009) S26
6. Thattai, M., van Oudenaarden, A.: Intrinsic noise in gene regulatory networks. Proceedings of the National Academy of Sciences **98**(15) (2001) 8614
7. Wilkinson, D.J.: Stochastic modelling for quantitative description of heterogeneous biological systems. Nature Reviews Genetics **10**(2) (2009) 122–133
8. Schedlich, L., Graham, L.: Role of insulin-like growth factor binding protein-3 in breast cancer cell growth. Microscopy research and technique **59**(1) (2002) 12–22
9. Brown, C., Lain, S., Verma, C., Fersht, A., Lane, D.: Awakening guardian angels: drugging the p53 pathway. Nature Reviews Cancer **9**(12) (2009) 862–873