Computational Strategy for Discovering Druggable Gene Networks from Genome-Wide RNA Expression Profiles

Seiya Imoto, Yoshinori Tamada, Hiromitsu Araki, Kaori Yasuda, Cristin G. Print, Stephen D. Charnock-Jones, Deborah Sanders, Christopher J. Savoie, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano

Pacific Symposium on Biocomputing 11:559-571(2006)

COMPUTATIONAL STRATEGY FOR DISCOVERING DRUGGABLE GENE NETWORKS FROM GENOME-WIDE RNA EXPRESSION PROFILES

SEIYA IMOTO[1,*], YOSHINORI TAMADA[2,*], HIROMITSU ARAKI[3,*], KAORI YASUDA[3], CRISTIN G. PRINT[4,†], STEPHEN D. CHARNOCK-JONES[4], DEBORAH SANDERS[4], CHRISTOPHER J. SAVOIE[3], KOUSUKE TASHIRO[5], SATORU KUHARA[5], SATORU MIYANO[1]

[1] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan*

[2] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan*

[3] *Gene Networks International, 4-2-12, Toranomon, Minato-ku, Tokyo, 105-0001, Japan*

[4] *Department of Pathology, Cambridge University, Tennis Court Road, Cambridge, CB2 1QP, United Kingdom*

[5] *Graduate School of Genetic Resources Technology, Kyushu University, 6-10-1, Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan*

We propose a computational strategy for discovering gene networks affected by a chemical compound. Two kinds of DNA microarray data are assumed to be used: One dataset is short time-course data that measure responses of genes following an experimental treatment. The other dataset is obtained by several hundred single gene knock-downs. These two datasets provide three kinds of information; (i) A gene network is estimated from time-course data by the dynamic Bayesian network model, (ii) Relationships between the knocked-down genes and their regulatees are estimated directly from knock-down microarrays and (iii) A gene network can be estimated by gene knock-down data alone using the Bayesian network model. We propose a method that combines these three kinds of information to provide an accurate gene network that most strongly relates to the mode-of-action of the chemical compound in cells. This information plays an essential role in pharmacogenomics. We illustrate this method with an actual example where human endothelial cell gene networks were generated from a novel time course of gene expression following treatment with the drug fenofibrate, and from 270 novel gene knock-downs. Finally, we succeeded in inferring the gene network related to *PPAR*-α, which is a known target of fenofibrate.

*These authors contributed equally to this work.

†Current affiliation: Department of Molecular Medicine & Pathology, School of Medical Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

1. Introduction

The microarray technology has produced a huge amount of gene expression data under various conditions such as gene knock-down, overexpression, experimental stressors, transformation, exposure to a chemical compound, and so on. Using a large volume of microarray gene expression data, a number of algorithms together with mathematical models[1,5,7,9,12,23] for estimating gene networks has been proposed and successfully applied to the gene network estimation of *S. cerevisiae*, *E. coli* etc. As a real application of gene network estimation techniques, computational drug target discovery[19] enhanced with gene network inference[6,14,20,22] has made tremendous impacts on pharmacogenomics.

In this paper, we propose a computational strategy for discovering the druggable gene networks, which are most strongly affected by a chemical compound. For this purpose, we use two types of microarray data: One is gene expression data obtained by measuring transcript abundance responses over time following treatment with the chemical compound. The other is gene knock-down expression data, where one gene is knocked-down for each microarray. Figure 1 is the conceptual view of our strategy. First, we estimate dynamic relationships denoted by G_T between genes based on time-course data by using dynamic Bayesian networks.[17] Second, in gene knock-down expression data, since we know the information of knocked-down genes, possible regulatory relationships between knocked-down gene and its regulatees can be obtained. We denote this information by R. Finally, the gene network G_K is estimated by gene knock-down data denoted by $\boldsymbol{X}_K$ together with G_T and R by using Bayesian networks based on multi-source biological information.[13] The key idea for estimating a gene network based on multi-source biological information is to use G_T and R as the Bayesian prior probability of G_K. The prior probability of the graph proposed by Imoto *et al.*[13] only uses binary prior information, i.e. known or unknown for each gene-gene relation. In this paper, we extend the prior probability of graph[13] in order to use prior information represented as continuous values. After estimating a gene network, for extracting biologically plausible information from the estimated gene network, we have also developed a gene network analysis tool called iNET that is an extended version of G.NET.[14] The iNet tool provides a computational environment for various path searches among genes with annotated gene network visualization.

As for related works, Basso *et al.*[2] estimated a gene network of human B cells as an undirected graph by their proposed algorithm. Our aim is to

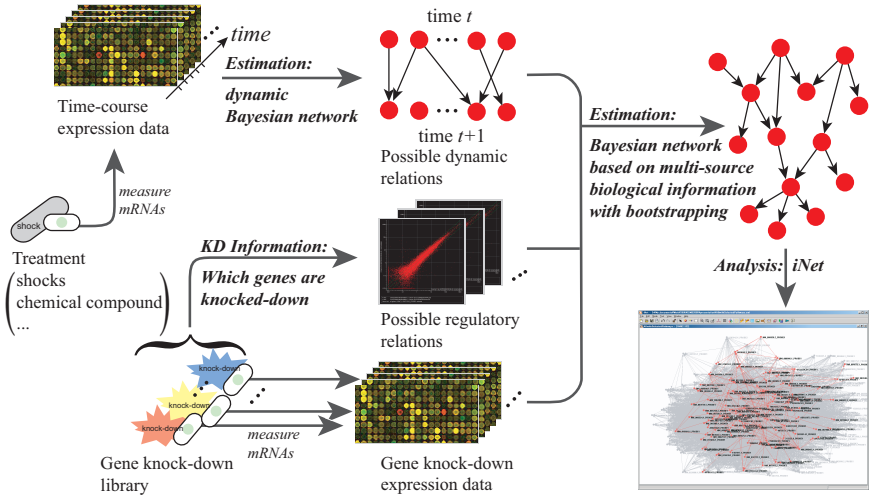

Figure 1. Conceptual view of the proposed method.

estimate druggable gene networks as directed graphs, that are sub-networks of the tissue-specific network. In this method, the edge direction is very important information and selection of compound-related genes is necessary. Therefore, the aim of this paper is clearly different from theirs. Di Bernardo *et al.*[6] proposed an interesting method for identifying mode-of-action of a chemical compound based on microarray gene expression data. Di Bernardo *et al.*[6] used statistical inference of a linear regression-based network model to find affected genes by a chemical compound. On the other hand, our interest is not only in the identification of affected genes, but also in the elucidation of their dependency as the network. In addition, since di Bernardo *et al.*[6] used examples of *S. cerevisiae* genes, more discussions might be needed in order to apply their method to human genes.

To demonstrate the whole process of the proposed method, we analyze expression data from human endothelial cells. We generate new time-course data that reveal the responses of human endothelial cell transcripts to treatment with the anti-hyperlipidaemia drug fenofibrate. We also generate new data from 270 gene knock-down experiments in human endothelial cells. The fenofibrate-related gene network is estimated based on fenofibrate time-course data and 270 gene knock-down expression data by the proposed method. The estimated gene network reveals gene regulatory relationships related to *PPAR-α*, which is known to be activated by fenofibrate. Our

computational analysis suggests that this computational strategy based on gene knock-down and drug-dosed time-course microarrays will give a new way to druggable gene discovery.

2. Methods for Reverse-Engineering Gene Networks

In the proposed method, we use Bayesian networks and dynamic Bayesian networks for estimating gene networks from gene knock-down and time-course microarray data, respectively. In this section, we briefly describe these two network models and then elucidate how we combine multi-source biological information to estimate more accurate gene networks.

2.1. *Preliminary*

Suppose that we have the observational data $\boldsymbol{X}$ of the set of p random variables $\mathcal{X} = \{X_1, ..., X_p\}$ and that the dependency among p random variables, shown as a directed graph G, is unknown and we want to estimate it from $\boldsymbol{X}$. In gene network estimation based on microarray data, a gene is regarded as a random variable representing the abundance of a specific RNA species, and $\boldsymbol{X}$ is the microarray data. From a Bayes approach, the optimal graph is selected by maximizing the posterior probability of the graph conditional on the observed data. By the Bayes' theorem, the posterior probability of the graph can be represented as

$$p(G|\boldsymbol{X}) = \frac{p(G)p(\boldsymbol{X}|G)}{p(\boldsymbol{X})} \propto p(G)p(\boldsymbol{X}|G),$$

where $p(G)$ is the prior probability of the graph, $p(\boldsymbol{X}|G)$ is the likelihood of the data $\boldsymbol{X}$ conditional on G and $p(\boldsymbol{X})$ is the normalizing constant and does not depend on the selection of G. Therefore, we need to set $p(G)$ and compute $p(\boldsymbol{X}|G)$ for the graph selection based on $p(G|\boldsymbol{X})$.

The prior probability of the graph $p(G)$ enables us to use biological data other than microarray data to estimate gene networks and the likelihood $p(\boldsymbol{X}|G)$ can be computed by Bayesian networks and dynamic Bayesian networks from gene knock-down and time-course microarray data, respectively. We elucidate how we construct $p(G|\boldsymbol{X})$ in the following sections.

2.2. *Bayesian Networks*

Bayesian networks are a graphical model that represents the causal relationship in random variables. In the Bayesian networks, we use a directed

acyclic graph encoding Markov relationship between connected nodes. Suppose that we have a set of random variables $\mathcal{X} = \{X_1, ..., X_p\}$ and that there is a causal relationship in $\mathcal{X}$ by representing a directed acyclic graph G_K. Bayesian networks then enable us to compute the joint probability by the product of conditional probabilities

$$\Pr(\mathcal{X}) = \prod_{j=1}^{p} \Pr(X_j | Pa_j), \tag{1}$$

where Pa_j is the set of random variables corresponding to the direct parents of X_j in G_K. In gene network estimation, we regard a gene as a random variable representing the abundance of a specific RNA species, shown as a node in a graph, and the interaction between genes is represented by the direct edge between nodes.

Let $\boldsymbol{X}_K$ be an $N \times p$ gene knock-down data matrix whose (i, j)-th element $x_{j|D_i}$ corresponds to the expression data of j-th gene when D_i-th gene is knocked down, where $j = 1, ..., p$ and $i = 1, ..., N$. Here we assume that i-th knock-down microarray is measured by knocking-down D_i-th gene. Since microarray data take continuous variables, we represent the decomposition (1) by using densities

$$f_{\mathrm{BN}}(\boldsymbol{X}_K | \boldsymbol{\Theta}, G_K) = \prod_{i=1}^{N} \prod_{j=1}^{p} f_j(x_{j|D_i} | \boldsymbol{pa}_{j|D_i}, \boldsymbol{\theta}_j),$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_p')'$ is a parameter vector, $\boldsymbol{pa}_{j|D_i}$ is the expression value vector of Pa_j measured by i-th knock-down microarray. Hence, the construction of the graph G_K is equivalent to model the conditional probabilities f_j $(j = 1, ..., p)$, that is essentially the same as the regression problem. For constructing $f_j(x_{j|D_i} | \boldsymbol{pa}_{j|D_i}, \boldsymbol{\theta}_j)$, we assume the nonparametric regression model with B-splines of the form

$$x_{j|D_i} = \sum_{k=1}^{|Pa_j|} m_{jk}(\boldsymbol{pa}_{j|D_i}^{(k)}) + \varepsilon_{j|D_i},$$

where $\boldsymbol{pa}_{j|D_i}^{(k)}$ is the k-th element of $\boldsymbol{pa}_{j|D_i}$, $\varepsilon_{j|D_i} \sim i.i.d.N(0, \sigma^2)$ for $i = 1, ..., N$, and m_{jk} $(k = 1, ..., |Pa_j|)$ are smooth functions constructed by B-splines as $m_{jk}(x) = \sum_{m=1}^{M_{jk}} \gamma_m^{(jk)} b_m^{(jk)}(x)$. Here $\gamma_m^{(jk)}$ and $b_m^{(jk)}(x)$ $(m = 1, ..., M_{jk})$ are parameters and B-splines, respectively.

The likelihood $p(\boldsymbol{X}_K | G_K)$ is then obtained by

$$p(\boldsymbol{X}_K | G_K) = \int f_{\mathrm{BN}}(\boldsymbol{X}_K | \boldsymbol{\Theta}, G_K) p(\boldsymbol{\Theta} | \boldsymbol{\lambda}, G_K) d\boldsymbol{\Theta}, \tag{2}$$

where $p(\boldsymbol{\Theta}|\boldsymbol{\lambda}, G_K)$ is the prior distribution on the parameter $\boldsymbol{\Theta}$ specified by the hyperparameter $\boldsymbol{\lambda}$. The high-dimensional integral can be asymptotically approximated with an analytical form by the Laplace approximation and Imoto *et al.*[12] defined a graph selection criterion, named BNRC, of the form

$$\begin{aligned}\mathrm{BNRC}(G_K) = &-2\log\{p(G_K)\} - r\log(2\pi/N) \\ &+ \log|J_\lambda(\hat{\boldsymbol{\Theta}}|\boldsymbol{X}_K)| - 2Nl_\lambda(\hat{\boldsymbol{\Theta}}|\boldsymbol{X}_K),\end{aligned}$$

where

$$\begin{aligned}l_\lambda(\boldsymbol{\Theta}|\boldsymbol{X}_K) &= \frac{1}{N}\left\{\log f_{\mathrm{BN}}(\boldsymbol{X}_K|\boldsymbol{\Theta}, G_K) + \log p(\boldsymbol{\Theta}|\boldsymbol{\lambda}, G_K)\right\}, \\ J_\lambda(\boldsymbol{\Theta}|\boldsymbol{X}_K) &= -\frac{\partial^2}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}'} l_\lambda(\boldsymbol{\Theta}|\boldsymbol{X}_K),\end{aligned}$$

r is the dimension of $\boldsymbol{\Theta}$, and $\hat{\boldsymbol{\Theta}}$ is the mode of $l_\lambda(\boldsymbol{\Theta}|\boldsymbol{X}_K)$. The network structure is learned so that $\mathrm{BNRC}(G_K)$ decreases by the greedy hill-climbing algorithm.[12] We should note that the solution obtained by the greedy hill-climbing algorithm cannot be guaranteed as the optimal. To find better solution, we repeat the greedy algorithm and choose the best one as $\hat{G}_K$. It happens quite often that the likelihood $p(\boldsymbol{X}_K|G_K)$ gives almost the same values for several network structures, construction an effective $p(G_K)$ based on various kinds of biological information is a key technique. We elucidate how we construct $p(G_K)$ in Section 2.4.

2.3. *Dynamic Bayesian Networks*

Dynamic Bayesian networks represent the dependency in random variables based on time-course data. Let $\mathcal{X}(t) = \{X_1(t), ..., X_p(t)\}$ be the set of p random variables at time t ($t = 1, ..., T$). In the dynamic Bayesian networks, a directed graph that contains p nodes is rewritten as a complete bipartite graph that allows direct edges from $\mathcal{X}(t)$ to $\mathcal{X}(t+1)$, where $t = 1, ..., T-1$. The directed graph G_T of the causal relationship among p random variables is then constructed by estimating the bipartite graph defined above. Under G_T structure, we then have the decomposition

$$\Pr(\mathcal{X}(1), ..., \mathcal{X}(T)) = \prod_{t=1}^{T}\prod_{j=1}^{p}\Pr(X_j(t)|Pa_j(t-1)), \tag{3}$$

where $Pa_j(t)$ is the set of random variables at time t corresponding to the direct parents of X_j in G_T.

Let $\boldsymbol{X}_T$ be a $T \times p$ time-course data matrix whose (t, j)-th element $x_j(t)$ corresponds to the expression data of j-th gene at time t, where $j = 1, ..., p$ and $t = 1, ..., T$. As we described in the Bayesian networks, the decomposition in (3) holds by using densities

$$f_{\mathrm{DBN}}(\boldsymbol{X}_T|\boldsymbol{\Xi}, G_T) = \prod_{t=1}^{T}\prod_{j=1}^{p} f_j(x_j(t)|\boldsymbol{pa}_j(t-1), \boldsymbol{\xi}_j, G_T),$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1', ..., \boldsymbol{\xi}_p')'$ is a parameter vector, $\boldsymbol{pa}_j(t)$ is the expression value vector of direct parents of X_j measured at time t. Here we set $\boldsymbol{pa}_j(0) = \emptyset$. We can construct f_{DBN} by using nonparametric regression with B-splines in the same way of the Bayesian networks. Therefore, by replacing f_{BN} by f_{DBN} in (2), Kim *et al.*[17] proposed a graph selection criterion for dynamic Bayesian networks, named BNRC$_{dynamic}$, with successful applications.

2.4. *Combining Multi-Source Biological Information for Gene Network Estimation*

Imoto *et al.*[13] proposed a general framework for combining biological knowledge with expression data aimed at estimating more accurate gene networks. In Imoto *et al.*[13], the biological knowledge is represented as the binary values, e.g. known or unknown, and is used for constructing $p(G)$. In reality, there are, however, various confidence in biological knowledge in practice. Bernard and Hartemink[3] constructed $p(G)$ using the binding location data[18] that is a collection of p-values (continuous information). In this paper, we construct $p(G)$ by using multi-source information including continuous and discrete prior information.

Let $\boldsymbol{Z}_k$ is the matrix representation of k-th prior information, where (i, j)-th element $z_{ij}^{(k)}$ represents the information of "gene $i \rightarrow$ gene j". For example, (1) If we use a prior network G_{prior} for $\boldsymbol{Z}_k$, $z_{ij}^{(k)}$ takes 1 if $e(i, j) \in G_{\mathrm{prior}}$ or 0 if $e(i, j) \notin G_{\mathrm{prior}}$. Here $e(i, j)$ denotes the direct edge from gene i to gene j. (2) By using the gene knock-down data for $\boldsymbol{Z}_k$, $z_{ij}^{(k)}$ represents the value that indicates how gene j changes by knocking down gene i. We can use the absolute value of the log-ratio of gene j for gene i knock-down data as $z_{ij}^{(k)}$. Using the adjacent matrix $E = (e_{ij})_{1 \leq i,j \leq p}$ of G, where $e_{ij} = 1$ for $e(i, j) \in G$ or 0 for otherwise, we assume the Bernoulli distribution on e_{ij} having probabilistic function

$$p(e_{ij}) = \pi_{ij}^{e_{ij}}(1 - \pi_{ij})^{1-e_{ij}},$$

where $\pi_{ij} = \Pr(e_{ij} = 1)$. For constructing π_{ij}, we use the logistic model with linear predictor $\eta_{ij} = \sum_{k=1}^{K} w_k(z_{ij}^{(k)} - c_k)$ as $\pi_{ij} = \{1 + \exp(-\eta_{ij})\}^{-1}$, where w_k and c_k $(k = 1, ..., K)$ are weight and baseline parameters, respectively. We then define a prior probability of the graph based on prior information $\boldsymbol{Z}_k$ $(k = 1, ..., K)$ by

$$p(G) = \prod_i \prod_j p(e_{ij}).$$

This prior probability of the graph assumes that edges $e(i, j)$ $(i, j = 1, ..., p)$ are independent of each other. In reality, there are several dependencies among e_{ij}'s such as $p(e_{ij} = 1) < p(e_{ij} = 1|e_{ki} = 1)$, and so on, we consider adding such information into $p(G)$ is premature by the quality of such information.

3. Application to Human Endothelial Cells' Gene Network

3.1. *Fenofibrate Time-Course Data*

We measure the time-responses of human endothelial cell genes to 25μM fenofibrate. The expression levels of 20,469 probes are measured by CodeLink$^{\text{TM}}$ Human Uniset I 20K at six time-points (0, 2, 4, 6, 8 and 18 hours). Here time 0 means the start point of this observation and just before exposure to the fenofibrate. In addition, we measure this time-course data as the duplicated data in order to confirm the quality of experiments.

Since our fenofibrate time-course data are duplicated data and contain six time-points, there are $2^6 = 64$ possible combinations to create a time-course dataset. We should fit the same regression function to a parent-child relationship in the 64 datasets. Under this constrain, we consider fitting nonparametric regression model to the connected data of 64 datasets. That is, if we consider gene $i \rightarrow$ gene j, we will fit the model $x_j^{(c)}(t) = m_j(x_i^{(c)}(t-1)) + \varepsilon_j(t)$, where $x_j^{(c)}(t)$ is the expression data of gene j at time t in the c-th dataset for $c = 1, ..., 64$. In the Bayesian networks, the reliability of estimated edges can be measured by using the bootstrap method. For time-course data, several modifications of the bootstrap method are proposed such as block resampling, but it is difficult to apply these methods to the small number of data points generated by short time-courses. However, by using above time-course modeling, we can define a method based on the bootstrap as follows: Let $D = \{D(1), ..., D(64)\}$ be the combinatorial time-course data of all genes. We randomly resample $D(c)$ with replacement and define a bootstrap sample $D^* = \{D^*(1), ..., D^*(64)\}$. We then re-

estimate a gene network based on D^*. We repeat 1000 times bootstrap replications and obtain $\hat{G}_T^{*1}, ..., \hat{G}_T^{*1000}$, where $\hat{G}_T^{*B}$ is the estimated graph based on the B-th bootstrap sample. The estimated reliability of edge can be used as the matrix representation of the first prior information $\boldsymbol{Z}_1$ as $z_{ij}^{(1)} = \#\{B|e(i,j) \in \hat{G}_T^{*B}, B = 1, ..., 1000\}/1000$.

3.2. *Gene Knock-Down Data by siRNA*

For estimating gene networks, we newly created 270 gene knock-down data by using siRNA. We measure 20,469 probes by CodeLink$^{\mathrm{TM}}$ Human Uniset I $20K$ for each knock-down microarray after 24 hours of siRNA transfection. The knock-down genes are mainly transcription factors and signaling molecules. Let $\tilde{\boldsymbol{x}}_{D_i} = (\tilde{x}_{1|D_i}, ..., \tilde{x}_{p|D_i})'$ be the raw intensity vector of i-th knock-down microarray. For normalizing expression values of each microarray, we compute the median expression value vector $\boldsymbol{v} = (v_1, ..., v_p)'$ as the control data, where $v_j = \mathrm{median}_i(\tilde{x}_{j|D_i})$. We apply the loess normalization method to the MA transformed data and the normalized intensity $x_{j|D_i}$ is obtained by applying the inverse transformation to the normalized $\log(\tilde{x}_{j|D_i}/v_j)$. We refer to the normalized $\log(\tilde{x}_{j|D_i}/v_j)$ as the log-ratio.

In 270 gene knock-down microarray data, we know which gene is knocked-down for each microarray. Thus, when we knock-down gene D_i, genes that significantly change their expression levels can be considered as the direct regulatees of gene D_i. We measure this information by computing corrected log-ratio as follows: The fluctuations of the log-ratios depend on their sum of sample's and control's intensities. From the normalized MA transformed data, we can obtain the conditional variance $s_j = \mathrm{Var}[\log(x_{j|D_i}/v_j)|\log(x_{j|D_i} \cdot v_j)]$ and the log-ratios can be corrected $z_{ij}^{(2)} = \log(x_{j|D_i}/v_j)/s_j$ satisfying $\mathrm{Var}(z_{ij}^{(2)}) = 1$.

3.3. *Results*

For estimating fenofibrate-related gene networks from fenofibrate time-course data and 270 gene knock-down data, we first define the set of genes that are possibly related to fenofibrate as follows: First, we extract the set of genes whose variance-corrected log-ratios, $|\log(x_{j|D_i}/v_j)/s_j|$, are greater than 1.5 from each time point. We then find significant clusters of selected genes using GO Term Finder. Table 1 shows the significant clusters of genes at 18 hours. The first column indicates how expression values are changed, i.e. "$\nearrow$" and "$\searrow$" mean "overexpressed" and "suppressed", respectively. The GO annotations of clusters with "$\searrow$" are mainly related to cell cycle,

Table 1. Significant GO annotations of selected fenofibrate-related genes from 18 hours microarray.

		GO Function	p-value	#genes
↘	GO:0007049	cell cycle	1.0E-08	35
↘	GO:0000278	mitotic cell cycle	3.7E-07	19
↘	GO:0000279	M phase	5.0E-06	17
↗	GO:0006629	lipid metabolism	1.3E-05	25
↘	GO:0007067	mitosis	1.3E-05	15
↘	GO:0000087	M phase of mitotic cell cycle	1.6E-05	15
↘	GO:0000074	regulation of cell cycle	2.7E-05	22
↗	GO:0044255	cellular lipid metabolism	4.4E-05	21
↗	GO:0016126	sterol biosynthesis	4.3E-04	6
↗	GO:0016125	sterol metabolism	4.5E-04	8
↗	GO:0008203	cholesterol metabolism	1.5E-03	7
↗	GO:0006695	cholesterol biosynthesis	2.4E-03	5
↗	GO:0008202	steroid metabolism	3.6E-03	10
↘	GO:0000375	RNA splicing, via transesterification reactions	4.1E-03	9
↘	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	4.1E-03	9
↘	GO:0000398	nuclear mRNA splicing, via spliceosome	4.1E-03	9
↗	GO:0006694	steroid biosynthesis	6.0E-03	7
↘	GO:0016071	mRNA metabolism	6.3E-03	13

the genes in these clusters are expressed ubiquitously and this is a common biological function. On the other hand, the GO annotations of clusters with "↗" are mainly related to lipid metabolism. In biology, it is reported that the fenofibrate acts around 12 hours after exposure.[8,10] Our first analysis for gene selection suggests that fenofibrate affects genes related to lipid metabolism and this is consistent with biological facts. We also focus on the genes from the 8 hour time-point microarray. Unfortunately, no cluster with specific function could be found in the selected genes from the 8 hour time-point microarray However, there also exist some genes related to lipid metabolism. Therefore we use the genes from the 8 and 18 hour time-point microarrays. Finally we add the 267 knock-down genes (three genes are not spotted on our chips) to the selected genes above, total 1192 genes are defined as possible fenofibrate-related genes and used for the next network analysis.

By converting the estimated dynamic network and knock-down gene information into the matrix representations of the first and second prior information $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, respectively, we estimate the gene network $\hat{G}_K$ based on $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and the knock-down data matrix $\boldsymbol{X}_K$. For extracting biological information from the estimated gene network, we first focus on lipid metabolism-related genes, because the clusters related this func-

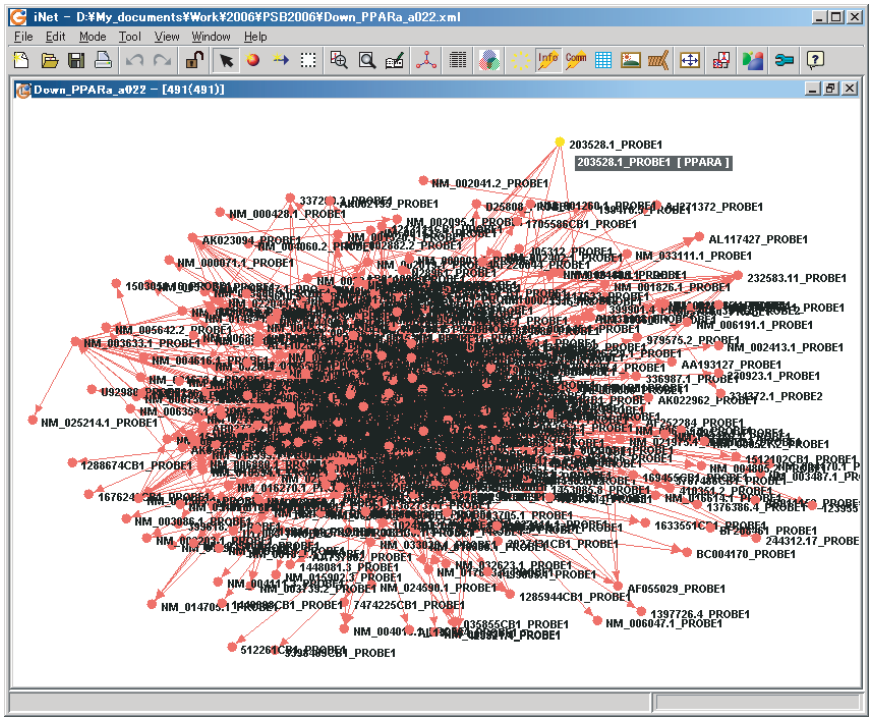

Figure 2. Down-stream of *PPAR-α*.

tion are significantly changed at 18 hours microarray. In the estimated gene network, there are 42 lipid metabolism-related genes and *PPAR-α* (*Homo sapiens* peroxisome proliferative activated receptor, alpha) is the only transcription factor among them. Actually, *PPAR-α* is a known target of fenofibrate. Therefore, we next focus on the node down-stream of *PPAR-α*. In Figure 2, the node down-stream of *PPAR-α* (491 genes). Here we consider that genes in the four steps down-stream of *PPAR-α* are candidate regulatees of *PPAR-α*. Among the candidate regulatees of *PPAR-α*, there are 21 lipid metabolism-related genes and 11 mole-

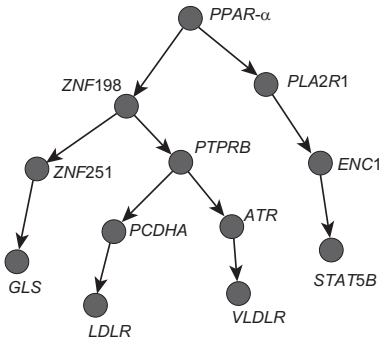

Figure 3. A sub-network related to *PPAR-α*.

cules previously identified experimentally to be related to *PPAR-α*. Actually, *PPAR-α* is known to be activated by fenofibrate. We show one sub-network having *PPAR-α* as a root node in Figure 3. One of the drug efficacies of fenofibrate whose target is *PPAR-α* is to reduce LDL cholesterol. *LDLR* and *VLDLR* mainly contribute the transporting of cholesterol and they are children of *PPAR-α*, namely candidate regulatees of *PPAR-α*, in our estimated network. As for *LDLR*, it has been reported the relationship with *PPAR-α*.[15] Moreover, several genes related to cholesterol metabolism are children of *PPAR-α* in our network. We also could extract *STAT5B* and *GLS* that are children of *PPAR-α* and have been reported their regulation-relationships with *PPAR-α*.[16,21] Therefore, it is not surprising that our network shows that many direct and indirect relationships involving known *PPAR-α* regulatees are triggered in endothelial cells by fenofibrate treatment. In the node up-stream of *PPAR-α*, *PPAR-α* and *RXR-α*, which form a heterodimer, share a parent. We could extract fenofibrate-related gene network and estimate that *PPAR-α* is the one of the key molecules of fenofibrate regulations without previous biological knowledge.

4. Discussion

From the point of view of pharmacogenomics, it is very important to know druggable gene networks. Our gene networks have the potential to predict the mode-of-action of a chemical compound, discover more effective drug target and predict side-effects. In this paper, we proposed a computational method to discover gene networks relating to a chemical compound. We use gene knock-down microarray data and time-course response microarray data for this purpose and combine multiple information obtained from observational data in order to estimate accurate gene networks under a Bayesian statistics framework. We illustrated the entire process of the proposed method using an actual example of gene network inference in human endothelial cells. Using fenofibrate time-course data and data from gene knock-downs in human endothelial cells, we successfully estimated a gene network related to the drug fenofibrate, which is a known agonist of *PPAR-α*. In the estimated gene network, *PPAR-α* has many direct and indirect regulatees including lipid metabolism related genes and this result indicates *PPAR-α* works as a trigger of the estimated fenofibrate-related network. There are many known relationships in the candidate regulatees of *PPAR-α* and we could find the relationship between *PPAR-α* and *RXR-α* in the estimated network. Peroxisome proliferator-activated receptors

(PPARs) are ligand-activated transcription factors expressed by endothelial cells and several other cell types. They are activated by ligands such as naturally occurring fatty acids and synthetic fibrates. Once activated, they heterodimerize with the retinoid-X-receptor (RXR) to activate the transcription of target genes. Many of these genes encode proteins that control carbohydrate and glucose metabolism and down-regulate inflammatory responses.[4] The further details on the relation between *PPAR-α* and *RXR-α* and their common parent will be discussed in another paper with biological evidences.

Acknowledgements

We wish to acknowledge Ben Dunmore, Sally Humphries, Muna Affara and Yuki Tomiyasu for assistance with endothelial cell culture and gene array analysis. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo.

References

1. T. Akutsu *et al.*, *Pac. Symp. Biocomput.*, **4**:17–28, 1999.
2. K. Basso *et al.*, *Nat. Genet.*, **37**:382–390, 2005.
3. A. Bernard and A.J. Hartemink, *Pac. Symp. Biocomput.*, **10**:459–470, 2005
4. A. Cabrero *et al.*, *Curr. Drug Targets Inflamm. Allergy*, **1**:243–248, 2002.
5. T. Chen *et al.*, *Pac. Symp. Biocomput.*, **4**:29–40, 1999.
6. D. di Bernardo *et al.*, *Nat. Genet.*, **37**:382–390, 2005.
7. N. Friedman *et al.*, *J. Comp. Biol.*, **7**:601–620, 2000.
8. K. Goya *et al.*, *Arterioscler. Thromb. Vasc. Biol.*, **24**:658–663, 2004.
9. A.J. Hartemink *et al.*, *Pac. Symp. Biocomput.*, **7**:437–449, 2002.
10. K. Hayashida *et al.*, *Biochem. Biophys. Res. Commun.*, **323**:1116–1123, 2004.
11. D. Heckerman *et al.*, *Machine Learning*, **20**:197–243, 1995.
12. S. Imoto *et al.*, *Pac. Symp. Biocomput.*, **7**:175–186, 2002.
13. S. Imoto *et al.*, *J. Bioinform. Comp. Biol.*, **2**:77–98, 2004.
14. S. Imoto *et al.*, *J. Bioinform. Comp. Biol.*, **1**:459–474, 2003.
15. K.K. Islam *et al.*, *Biochim. Biophys. Acta.*, **1734**:259–268, 2005.
16. S. Kersten *et al.*, *FASEB J.*, **15**:1971–1978, 2001.
17. S. Kim *et al.*, *Biosystems*, **75**:57–65, 2004.
18. T.I. Lee *et al.*, *Science*, **298**:799–804, 2002.
19. M.J. Marton *et al.*, *Nat. Med.*, **4**:1293–1301, 1998.
20. C.J. Savoie *et al.*, *DNA Res.*, **10**:19–25, 2003.
21. J.M. Shipley and D.J. Waxman, *Mol. Pharmacol.*, **64**:355–364, 2003.
22. Y. Tamada *et al.*, *Genome Informatics*, **16**:182–191, 2005.
23. E.P. van Someren *et al.*, *Pharmacogenomics*, **3**:507–525, 2002.