Journal of Bioinformatics and Computational Biology
Vol. 1, No. 3 (2003) 459–474

Imperial College Press
www.icpress.co.uk

USE OF GENE NETWORKS FOR IDENTIFYING AND VALIDATING DRUG TARGETS

SEIYA IMOTO[*,†,‡], CHRISTOPHER J. SAVOIE[*,§,¶], SACHIYO ABURATANI[*,‖,**], SUNYONG KIM[†,††], KOUSUKE TASHIRO[‖,‡‡], SATORU KUHARA[‖,§§] and SATORU MIYANO[†,¶¶]

[†]*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan*
[‡]*imoto@ims.u-tokyo.ac.jp*
[††]*sunk@ims.u-tokyo.ac.jp*
[¶¶]*miyano@ims.u-tokyo.ac.jp*
[§]*Gene Networks Inc., 560 South Winchester Blvd., Suite 500, San Jose, CA, 95128, USA*
[¶]*savoie@gene-networks.com*
[‖]*Graduate School of Genetic Resources Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan*
[**]*sachiyo@grt.kyushu-u.ac.jp*
[‡‡]*ktashiro@grt.kyushu-u.ac.jp*
[§§]*kuhara@grt.kyushu-u.ac.jp*

Received 17 October 2002
Revised 21 May 2003
Accepted 22 May 2003

We propose a new method for identifying and validating drug targets by using gene networks, which are estimated from cDNA microarray gene expression profile data. We created novel gene disruption and drug response microarray gene expression profile data libraries for the purpose of drug target elucidation. We use two types of microarray gene expression profile data for estimating gene networks and then identifying drug targets. The estimated gene networks play an essential role in understanding drug response data and this information is unattainable from clustering methods, which are the standard for gene expression analysis. In the construction of gene networks, we use the Bayesian network model. We use an actual example from analysis of the *Saccharomyces cerevisiae* gene expression profile data to express a concrete strategy for the application of gene network information to drug discovery.

Keywords: Drug targets; gene network; microarray; Bayesian network.

[*]These authors contributed equally to this work.

1. Introduction

Microarray technology has produced a large volume of genome-wide gene expression profile data under various experimental conditions such as gene disruptions, gene overexpressions, shocks, cancer cells, etc. Along with this new data production, there have been considerable attempts to infer gene networks from such gene expression profile data and several computational methods have been proposed together with gene network models such as Boolean networks,[2–5,15,16,21] differential equation models[6,7,16] and Bayesian networks.[9,10,12,13,18] While the paradigm of using microarray technology with the clustering technique has made tremendous impacts on biomedical research and practice,[8,22] the strategy enhanced with computational gene network analysis has not yet been well examined for practical applications. In this paper, we propose a novel method for identifying and validating drug targets by exploiting two computational methods for inferring gene networks from cDNA microarray data. We show the whole process of analysis by using an antifungal medicine as a drug and *Saccharomyces cerevisiae* cDNA microarrays that actually extracted candidate genes as targets of this drug.

Our strategy is summarized in Fig. 1. Gene regulatory pathway information is essential in this scheme. In order to create this information, we have prepared two kinds of cDNA microarray data for constructing gene networks. One is gene expression profile data obtained from 120 gene disruptions, where mostly transcription factors are disrupted, one for each microarray. We denote this data set as a matrix $\boldsymbol{X}$ for convenience. The other is gene expression profile data obtained from expression experiments of several dose and time responses to the drug (we denote this data set as a matrix $\boldsymbol{Y}$). Then the process of identifying drug targets consists of two steps. The first step is to identify the genes directly affected by the drug. For this purpose, the most straightforward approach might be the fold-change analysis of the data set $\boldsymbol{Y}$. However, in order to find genes directly affected by the drug, we took another method[16] which is more suited for inferring qualitative relations between genes. Based on this method, we describe a method named *virtual gene technique* where we regard the drug as a "virtual gene" and generate a multi-level directed acyclic graph with this virtual gene as the root by using both data sets $\boldsymbol{X}$ and $\boldsymbol{Y}$. From this graph, we can identify genes which may be directly affected by the drug while the fold-change analysis of the data set $\boldsymbol{Y}$ just may provide us genes directly or indirectly affected by the drug. The second step is to find "druggable genes" that regulate the drug-affected genes most strongly from the upstream of the gene network. For this purpose, we employed a method based on Bayesian network model[12,13] to construct a gene network from the data set $\boldsymbol{X}$. With this gene network, we could explore the gene network for the druggable genes related to the drug-affected genes very effectively. For this gene network exploration, we have also developed a gene network analysis tool called G.NET that provides a computational environment for various path searches among genes with annotated gene network visualization. This total system constitutes our method.

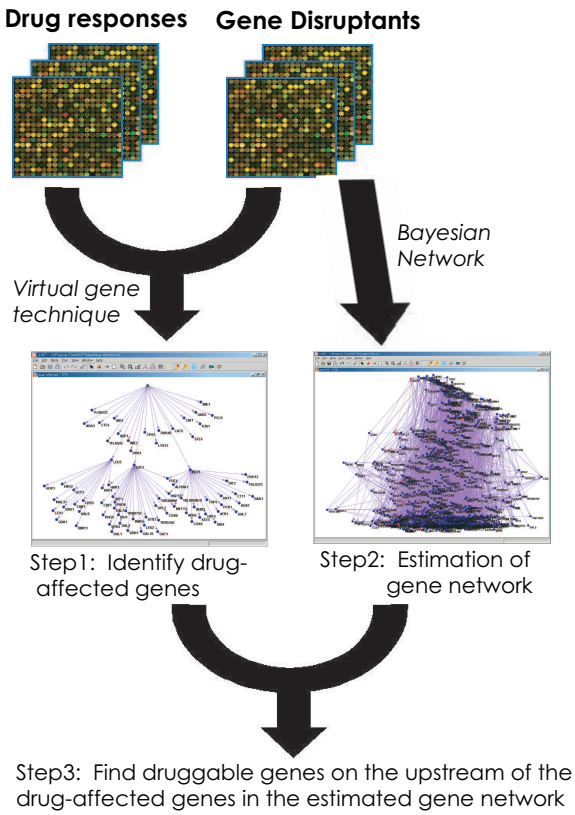

Fig. 1. Schematic view.

In this paper, we first discuss the computational methods employed for gene network analysis from microarray data. In addition, we also discuss clustering methods for our purpose since clustering[8,22] has become a standard method for analyzing microarray gene expression profile data. Then we explain how microarray data have been prepared and analyzed. By examining the computational analysis with G.NET, we demonstrate the effectiveness of our gene network strategy for identifying genes for the drug target.

2. Gene Network Models for Identifying Drug Targets

From the viewpoint of reverse engineering, several models for gene network have been proposed together with algorithms for constructing gene networks from microarray data. They may be roughly classified into three models by Boolean network model,[2–5,15,16,21] model defined as a system of ordinary differential equations,[6,7] and statistical network model.[9,10,12,13,18,23] Reverse engineering algorithms also heavily rely on characteristics of measurements, e.g. time-course data or not, disruption, overexpression, various shocks, hormone stimulus, drug response, developmental change, etc. None of them may be perfect by itself. However, a

sophisticated combination of these methods for specific microarray measurements would yield possibilities for new applications.

Two kinds of cDNA microarray measurements of *Saccharomyces cerevisiae* have been prepared for our purpose. One is the microarray data obtained by gene disruptions, and the other is a time-course data of responses to an antifungal medicine. The number of disruptants is 120 for the first data and the time-course data for one dose consists of five microarrays. The details about these data are given in Sec. 3. For these data, we employ two gene network models and algorithms for gene network estimation so that combination of two models can cover the shortcomings each other and we can obtain more reliable information by using both network methods.

2.1. *Virtual gene technique for identifying drug affected genes*

The first task is to find genes directly affected by the drug. For this purpose, we regard the drug as a "virtual gene" and we consider that the state of this virtual gene is 1 (ON) if the drug is dosed, and otherwise it is 0 (OFF).

The idealistic approach for identifying the genes directly affected by the virtual gene may be to use the Boolean network model and to apply the method developed by Akutsu *et al.*[2] for inferring Boolean network model that can suggest a series of mutants for identifying the network. However, it requires multiple disruptions and overexpressions for one mutant and the number of mutants required for identifying the network is not realistic even for a small network. Therefore this is not in our scope since we can deal with only single gene disruptants in our measurement experiment. Instead, we focus on a simpler network model whose structure is a multi-level directed acyclic graph (dag).[16] Maki *et al.*[16] have also proposed a naive algorithm for constructing a multi-level dag based on information how a single gene disruption affects other genes. We use this method for finding genes directly affected by the drug by regarding the drug as a gene and we call it the *virtual gene technique*. Here we roughly explain the method.

Let $V = \{g_1, g_2, \ldots, g_n\}$ be the set of all genes and $D = \{d_1, \ldots, d_m\} \subseteq V$ be the set of genes to be disrupted. We assume D contains the virtual genes corresponding to the drug. Our cDNA microarray compares the gene expression level of a mutant with that of a wild type for each gene. From a gene disruption experiment for gene d_i, we obtain a microarray data $E_{d_i}[g_j]$ $(1 \leq j \leq n)$. Then by setting a threshold θ, we define a relation R as follows:

$$R(a,b) = \begin{cases} 1 & \text{if } a \in D \text{ and } E_a[b] > \theta \text{ or } E_a[b] < 1/\theta \\ 0 & \text{otherwise} \end{cases}.$$

By using the transitive closure R^* of R, we define an equivalence relation $\equiv_{R^*}$ on V by $a \equiv_{R^*} b$ if and only if $a, b \in D$ and $R^*(a,b) = 1 \wedge R^*(b,a) = 1$. Then a new relation $\hat{R}$ on the equivalence classes of $\equiv_{R^*}$ is defined by $\hat{R}([a],[b]) = R^*(a,b)$ for $a, b \in V$. Note that $\hat{R}$ is well-defined by the definition of $\equiv_{R^*}$. $\hat{R}$ defines a directed

	D.gene A	D.gene B	D.gene C	D.gene D	Drug 1
gene A		Not affected	Not affected	Not affected	Affected
gene B	Affected		Not affected	Not affected	Affected
gene C	Affected	Not affected		Affected	Affected
gene D	Affected	Not affected	Affected		Affected

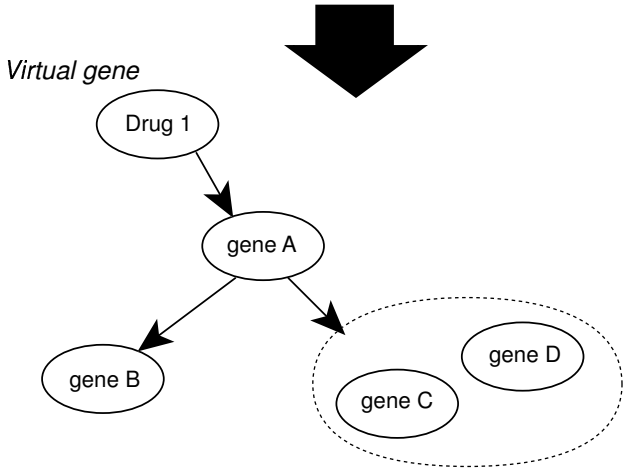

Fig. 2. Graphical view of the virtual gene technique. "D.gene A" means this microarray is observed by disrupting gene A. The dotted circle shows the equivalence class.

acyclic graph $\hat{G}$ on the set of equivalence classes, where self-loop edges are ignored. An indirect edge $([a],[b])$ of $\hat{G}$ is an edge such that there is another path from $[a]$ to $[b]$ in $\hat{G}$. By removing all indirect edges from $\hat{G}$, we obtain a multi-level dag (see Maki *et al.*[16] for more details about the construction). Figure 2 shows an example of the resulting network based on the virtual gene technique. Finally, by considering the dag whose root is the virtual gene, the children of this virtual gene would be the candidate genes directly affected by the drug.

The advantages of the use of the multi-level dag model are as follows: (a) This model is very simple and can be easily understood, (b) the model shows the parent-child relations correctly, when the data has sufficient accuracy and information. However, this method requires discretization of the expression values into two levels (0 or 1), and the quantization probably causes information loss. By choosing the threshold θ appropriately in a heuristic manner, it is reported that this naive algorithm works well in practice[1,16] although the method is not theoretically well-founded.

2.2. *Bayesian network*

After identifying the genes directly affected by the drug, the next step is to find genes upstream of the drug-affected genes. For this purpose, we employ the Bayesian network model[14] and use the method that we developed for constructing Bayesian networks from perturbated gene expression profile data.[12,13]

The Bayesian network is a graph representation of the complex relations of a large number of random variables. We consider the directed acyclic graph with Markov relations of nodes in the context of Bayesian network. We can thus describe complex phenomena through conditional probabilities instead of the joint probability of the random variables. That is, suppose that we have a gene expression value x_{ij} of ith array and jth gene for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, we have the decomposition $f(x_{i1}, \ldots, x_{ip}) = f_1(x_{i1}|\boldsymbol{p}_{i1}) \times \cdots \times f_p(x_{ip}|\boldsymbol{p}_{ip})$, where $\boldsymbol{p}_{ij} = (p_{i1}^{(j)}, \ldots, p_{iq_j}^{(j)})^T$ is a q_j-dimensional parent observation vector of x_{ij}. Here $p_{ik}^{(j)}$ is an observation of kth parent of jth gene measured by ith microarray.

Friedman *et al.*[9] proposed an interesting approach for estimating a gene network from gene expression profiles. They discretized the expression values into three values and used the multinomial distributions as the conditional distributions of the Bayesian network. However, a problem remains to be solved in choosing the threshold values for discretizing. Imoto *et al.*[12,13] recently used the nonparametric regression model $x_{ij} = m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij}$, that offers a solution that does not require quantization. Here, ε_{ij} depends independently and normally on mean 0 and variance σ_j^2 and $m_{jk}(x)$ is a smooth function, which is constructed by B-splines, of the form $m_{jk}(p_{ik}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)})$ for $k = 1, \ldots, q_j$, where $\gamma_{1k}^{(j)}, \ldots, \gamma_{M_{jk}k}^{(j)}$ are unknown coefficients and $b_{1k}^{(j)}(x), \ldots, b_{M_{jk}k}^{(j)}(x)$ are B-splines. Hence, the nonlinear Bayesian network model is defined by

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \prod_{j=1}^{p} \exp\left[-\left\{ x_{ij} - \sum_{k=1}^{q_j} \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)}) \right\}^2 \Big/ 2\sigma_j^2 \right] \Big/ \sqrt{2\pi\sigma_j^2}\,.$$

If we set the network structure, we can estimate the nonlinear Bayesian network model by a suitable procedure. However, we must choose an optimal network structure, which gives the best representation of the phenomenon described by the data. Therefore, Imoto *et al.*[12,13] derived a new criterion, named BNRC (Bayesian network and Nonparametric Regression Criterion), for selecting a network from Bayes approach. The BNRC is defined as an approximation of the posterior probability of the network by using the Laplace approximation for integrals. Imoto *et al.*[12,13] applied the proposed method to the *Saccharomyces cerevisiae* gene expression profile data and estimated a gene network. The advantages of this method are as follows: (a) We can analyze the gene expression profiles as continuous data, (b) this model can detect not only linear structures but also nonlinear dependencies between genes and (c) the proposed criterion can optimize the parameters in the model and the structure of the network automatically. Note that the Bayesian network model based on the linear regression is included in our model as a special case.

The Bayesian network has theoretical advantageous base on the mathematics, and, in this paper, we use the method proposed by Imoto *et al.*[12,13] for constructing Bayesian networks. Unfortunately, the Bayesian network cannot construct cyclic regulations and are not useful for creating multilevel directional models of regulatory effects from data created from logical joins of expression profile data from disruptants and drug response experiments. However, since our purpose is to find genes in the upstream of the drug-affected genes, this drawback is not actually serious in practice.

We applied our method to the gene expression profile data of 120 disruptants and created a Bayesian network consisiting of 735 genes. Then by mapping the drug-affected genes obtained as in Sec. 2.1 to the this network, we can analyze the genes controling them in the upstream. For this analysis, we also developed a network visualization software G.NET for various networks including Bayesian networks. Snapshots of the software are shown in Figs. 4 and 5.

3. Application

3.1. *Microarray data*

We created two libraries of cDNA microarray data from the *Saccharomyces cerevisiae* gene expression profiles. One is obtained by disrupting 120 genes, and the other is comprised of the responses to the antifungal drug. We used the BY4741 (*MATa*, *HIS3D1*, *LEU2D0*, *MET15D0*, *URA3D0*) as the wild type strain and purchased gene disruptions from Research Genetics, Inc. We selected 735 genes from the yeast genome for identifying drug targets. These genes are selected as follows: First, we have collected 314 genes which are known as transcription factors. 98 of these 314 genes have already been studied for their control mechanisms. The expression profile data for 735 genes chosen for our analysis includes the genes controlled by these 98 transcription factors from 5871 genes in addition to nuclear receptor-like genes which have a pivotal role in gene expression regulation and are popular drug targets. We have constructed the Bayesian network models of these 735 genes from 120 gene disruption conditions. As for normalizing microarray data, we first applied the total intensity notmalization that adjusts each ratio such that the mean ratio is equal to one. For local normalization, the linear regression was then performed to each pen group, which is a group of genes deposited by a spotting pen, for correcting systematic bias of each pen group. As for normalization, we referred to Quackenbush.[19]

As for the drug response microarray gene expression profile data, we incubated yeast cultures in dosages of 10, 25, 50 and 100 mg/ml of an antifungal medication in culture and took aliquots of the culture at five time points (0, 15, 30, 45 and 60 minutes) after addition of the agent. Here time 0 means the start point of this observation and just after exposure to the drug. We then extracted the total RNA from these experiments, labeled the RNA with cy5, hybridized them with cy3 labeled RNA from non-treated cells and applied them to full genome cDNA

microarrays thus creating a data set of 20 microarrays for drug response data. Actually, we measured two or three microarrays for some of each experiment. We observed that the expression patterns of our microarrays at the same condition are very close. Therefore we used one microarray against a time point or a gene disruption for the following analysis. In this paper, we use these 140 microarrays to elucidate drug targets using gene networks.

3.2. *Result*

3.2.1. *Clustering result*

For identification of the drug targets, the popular but problematic strategy is the use of clustering methods.[11,17] Clustering methods provide the gene group information via the similarity of the expression patterns. We have two types of microarray data, gene disruption and drug response, allowing us to compare drug response patterns to gene expression patterns caused by disruption. In the clustering analysis, if there would be a significant and strong similarity between the expression patterns of a single disruptant or group of disruptants and a given drug response microarray, we might conclude that the drug probably plays the same role as the disrupted gene. Moreover, if this disrupted gene would have known functional role, we could obtain more information about the response to the drug. We examined this strategy with our data.

We combine the two types of data and make the matrix $\boldsymbol{Z} = (\boldsymbol{X} : \boldsymbol{Y})$, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the gene disruption and the drug response microarray data, respectively, and implement the hierarchical clustering based on the complete-linkage method to cluster the microarrays. The similarity metric we used is the uncentered correlation.[8] Unfortunately, as is often the case with such experiments, we could not gain such a straightforward result from clustering our data. Figure 3 shows the hierarchical clustering result for the combined gene expression profile data. It is clear that the drug response microarrays make one cluster and are separated from the disruption microarrays. From this result, we cannot

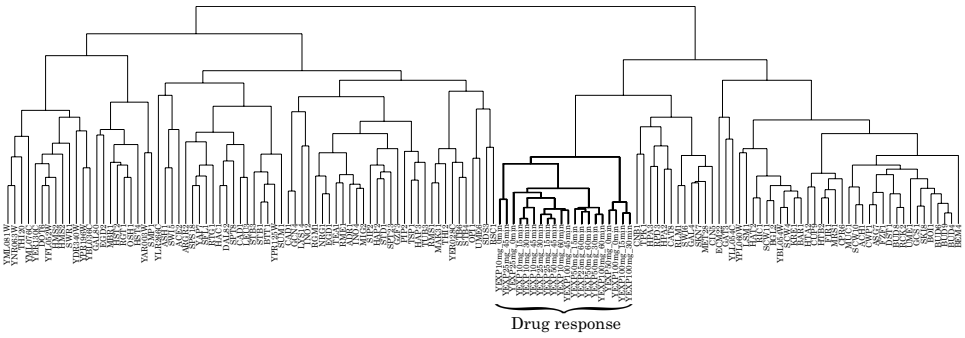

Fig. 3. The hierarchical clustering result for 140 microarrays, e.g., 20 drug responses and 120 gene disruptions.

extract any interactions between drug response data and gene disruption data. We further implemented hierarchical clustering of the gene disruption and drug response microarrays with various distances and metrics, however the results were essentially unchanged.

3.2.2. *Application of virtual gene technique*

We applied the virtual gene technique shown in Sec. 2.1 to the microarray data $\boldsymbol{Z}$ which was created by the combining gene disruption and drug response microarrays. We consider the conditions of the drug responses data as virtual genes, e.g., the condition 100 mg/ml and 30 min is given an assignment as the gene `YEXP100mg30min`. By using the network model introduced in Sec. 2.1, we can find the child genes of these virtual genes with the drug affecting these child genes in progeny generational order. That is, at first, we can find the downstream genes of the virtual gene and the information of the gene disruption microarray data can be used for understanding the regulation pathway of these downstream genes. Here, we set a different threshold θ for each microarray. That is, as for the jth microarray's threshold θ_j, we set $\theta_j = 2\hat{s}_j^2$, where $\hat{s}_j^2$ is the estimated variance of jth microarray. We then focus on genes that have five or more virtual genes as the parent genes as the putative drug-affected genes. That is, genes are under direct influence of the virtual genes. However, a gene that has only one virtual gene as its parent may be the primary drug-affected gene, depending on the mode of action for a given drug and this must be analyzed on a case by case basis. Thus the virtual gene technique highlights its advantage in the initial screening for genes under drug-induced expression influence.

In addition, fold-change analysis can provide similar information to the proposed virtual gene technique. In fact, we can obtain the differentially affected genes under certain experimental condition by fold-change analysis. However, our virtual gene technique can improve the result of the fold-change analysis. Conceptually, suppose we find that *geneA* and *geneB* are affected by the drug from fold-change analysis. The fold-change analysis cannot take into account the baseline interactions between *geneA* and *geneB*. That is, if there is a regulation pathway between *geneA* and *geneB* that *geneA* $\rightarrow$ *geneB*, the *geneB* is probably not affected by the drug directly. The virtual gene technique can take into account such interaction by using the information of the gene disruption data and thus reduces the search set to more probable target genes. In Fig. 4, we put the expression ratios of more than two-fold next to the genes. If we set two-fold as the threshold of the fold-change analysis, we conclude that the genes, which have more than two-fold expression ratios, are affected by the drug. However, by using the gene disruption microarray data, we can find the regulations of the downstream of the virtual gene and can understand whether the gene is actually affected by the drug. Indeed, some high expression ratio genes are not directly affected by the drug, they are probably the false-positives of the fold-change analysis.

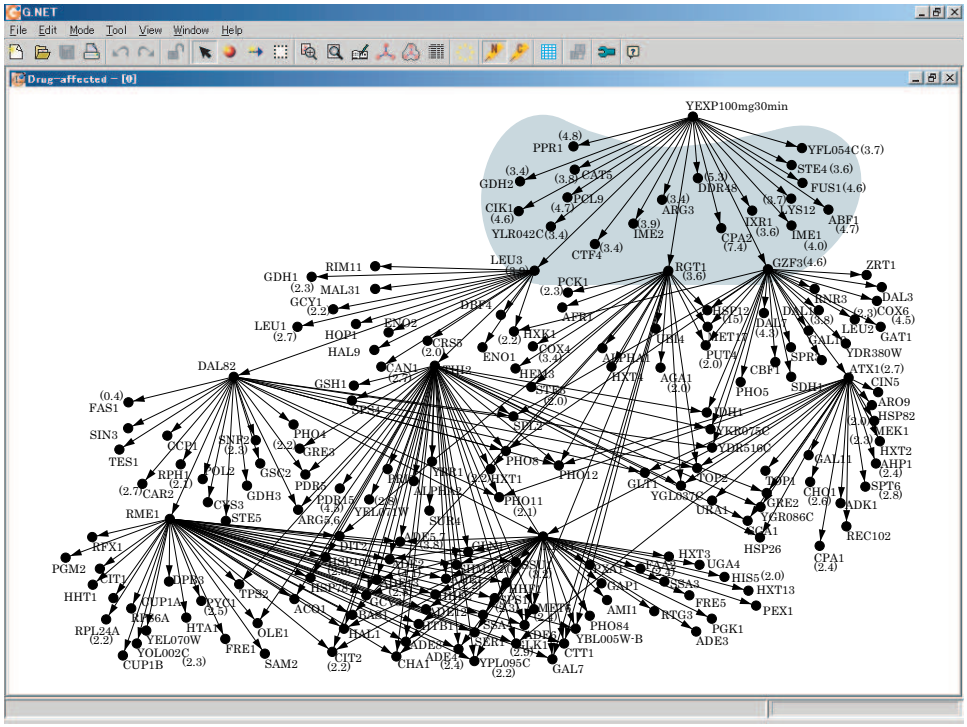

Fig. 4. The downstream pathway of the virtual gene `YEXP100mg30min`. The shadowed genes are affected by the drug.

There is no guarantee that genes that are most affected by the drug are the genes that were "drugged" by the drug agent, nor is there any guarantee that the drugged target represents the most biologically available and advantageous molecular target for intervention with new drugs. Thus, even after identifying probable molecular modes of action, we should find the most druggable genes upstream of the drug-affected genes in a regulatory network and to then screen low molecular weight compounds for drug activity on those targets. In the network model used for the virtual gene technique, the virtual genes should be placed on the top of the network. Therefore, it is difficult or sometimes impossible to find upstream information of the drug-affected genes in this network. At this stage, we can use the Bayesian network model for exploring the upstream of the drug-affected genes in an effective manner.

3.2.3. *Exploring upstream of genes with Bayesian network*

The gene network is estimated by the Bayesian network and nonparametric regression method together with BNRC optimization strategy.[12,13] We use the *Saccharomyces cerevisiae* microarray gene expression profile data obtained by

disrupting 120 genes. The druggable genes are the drug targets related to these drug-affected genes, which we want to identify for the development of novel leads. We can explore the druggable genes upstream of the drug-affected genes in the estimated gene network by the Bayesian network method. Here, we focus on the nuclear receptor-like genes as the druggable genes because: (a) In general, nuclear receptor proteins are known to be useful drug targets and together represent over 20 percent of the targets for medications presently in the market. (b) Nuclear receptors are involved in the transcription regulatory affects that are directly measured in cDNA microarray experiments.

Figure 5 shows a partial resulting network, which includes the drug-affected genes (bottom) identified by the virtual gene technique, the druggable genes (top) that are the nuclear receptor-like genes and the intermediary genes (middle). The druggable genes in the circle connect directly to the drug-affected genes and other druggable genes have one intermediary gene per one druggable gene. Of course, we can find more pathways from the druggable genes to the drug-affected genes if we admit more intermediary genes. Due to the use of the Bayesian network model, we can find the intensities of the edges and can select more reliable pathway. This is

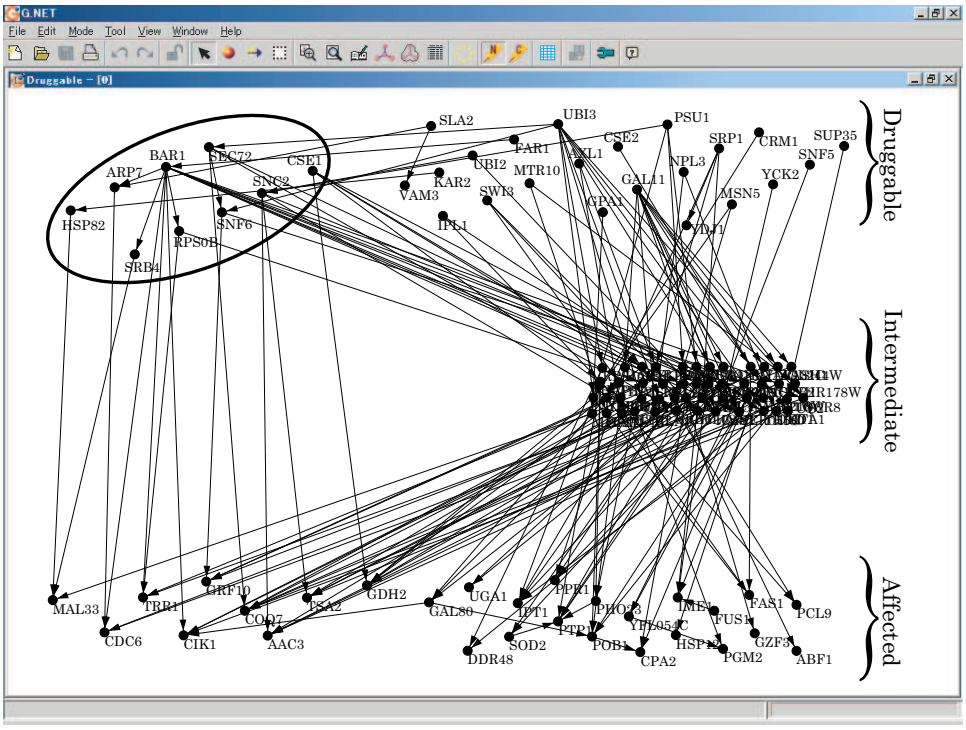

Fig. 5. A partial resulting network among the druggable (top), drug-affected (bottom) and intermediary genes (middle).

Table 1. The druggable genes of MAL33 and CDC6. "Parents" means these genes connected directly to the drug-affected genes. "Grandparents" means there is one intermediary gene between these genes and the drug-affected genes.

Drug-affected	MAL33 (YBR297W) : Maltose fermentation regulatory protein
Druggable	***Parents***
	HSP82 (YPL240C) : Heat shock protein
	SRB4 (YER022W) : DNA-directed RNA polymerase II holoenzyme and Kornberg's mediator (SRB) subcomplex subunit
	Grandparents
	BAR1 (YIL015W) : Barrierpepsin precursor
	GPA1 (YHR005C) : GTP-binding protein alpha subunit of the pheromone pathway
	KAR2 (YJL034W) : nuclear fusion protein
Drug-affected	CDC6 (YJL194W) : Cell division control protein
Druggable	***Parents***
	ARP7 (YPR034W) : Component of SWI-SNF global transcription activator complex and RSC chromatin remodeling complex
	BAR1 (YIL015W) : Barrierpepsin precursor
	Grandparents
	GAL11 (YOL051W) : DNA-directed RNA polymerase II holoenzyme and Kornberg's mediator (SRB) subcomplex subunit
	FAR1 (YJL157C) : Cyclin-dependent kinase inhibitor (CKI)
	SLA2 (YNL243W) : Cytoskeleton assembly control protein

an advantage of the Bayesian network model in searching for ideal druggable targets. From Fig. 5, we can find the druggable genes for each drug-affected gene, e.g. we can find the druggable genes for MAL33 and CDC6 shown in Table 1.

The drug-affected gene CDC6 in Table 1 is a protein that regulates the initiation of DNA replication. It binds to origins of replication at the end of mitosis, directing the assembly and disassembly of MCM proteins and the pre-replication complex. It is a member of the AAA+ family of ATPases. The genetic mechanism and effectiveness of this antifungal medication was made clear by this result. It was investigated that the localization of CDC6p is nulear. This means that any other extracellular molecule, like drug, cannot affect it directly. Thus our result indicates through the gene network analysis a mechanism that CDC6 is influenced by the extracellular molecule.

4. Discussion

In this paper, we proposed a new strategy for identifying and validating drug targets using computational models of gene networks. We showed that the clustering methods cannot provide the sufficient information and there is a need for the kind of hierarchical interaction data provided by gene network methods. We focuses on two models of gene networks for estimating gene networks from microarray gene expression profile data. Theses two methods are originally proposed from the

different points of view. However, both methods have relative strengths and weakness and we can obtain more reliable information by a harmonized use of these two methods in our analysis. We described the practical advantages of the virtual gene technique over results obtained from simple fold-change analysis for identifying drug-affected genes. On the other hand, for exploring the druggable genes upstream of the drug-affected genes, the Bayesian network approach was shown to be very effective. The Bayesian network model can provide the information of the upstream of the drug-affected genes effectively and we can thus attain a set of candidate druggable genes for each drug-affected gene. The strategy proposed by this paper is established based on the sophisticated use of a combination of two network methods. The strength of each network method can be clearly seen in this strategy and the proposed integrated method can provide a methodological foundation for the practical application of bioinformatics techniques for gene network inference in the identification and validation of drug targets. In practice, biological experiments that disrupt the target druggable genes are needed for confirming the results of the analysis.[20] We should also note that the visualized gene network analysis software G.NET played an important role in extracting subnetworks from the messy interactions of 735 genes.

Acknowledgments

The authors would like to thank two referees for their helpful comments and suggestions.

References

1. S. Aburatani, K. Tashiro, C. J. Savoie, M. Nishizawa, K. Hayashi, Y. Ito, S. Muta, K. Yamamoto, M. Ogawa, A. Enomoto, M. Masaki, S. Watanabe, Y. Maki, Y. Takahashi, Y. Eguchi, Y. Sakaki and S. Kuhara, "Discovery of novel transcription control relationships with gene regulatory networks generated from multiple-disruption full genome expression libraries," *DNA Research* **10**, 1–8 (2003).
2. T. Akutsu, S. Kuhara, O. Maruyama and S. Miyano, "A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions," *Genome Informatics* **9**, 151–160 (1998).
3. T. Akutsu, S. Miyano and S. Kuhara, "Identification of genetic networks from a small number of gene expression patterns under the Boolean network model," *Proc. Pacific Symposium on Biocomputing* **4**, 17–28 (1999).
4. T. Akutsu, S. Miyano and S. Kuhara, "Inferring qualitative relations in genetic networks and metabolic pathways," *Bioinformatics*, **16**, 727–734, 2000.
5. T. Akutsu, S. Miyano and S. Kuhara, "Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function," *J. Comp. Biol.* **7**, 331–344 (2000).
6. T. Chen, H. L. He and G. M. Church, "Modeling gene expression with differential equations," *Proc. Pac. Symposium on Biocomputing* **4**, 29–40 (1999).
7. M. J. L. de Hoon, S. Imoto, K. Kobayashi, N. Ogasawara and S. Miyano, "Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis*

using differential equations," *Proc. Pacific Symposium on Biocomputing*, **8**, in press, 2003.

8. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
9. N. Friedman, M. Linial, I. Nachman and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comp. Biol.* **7**, 601–620 (2000).
10. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young, "Combining location and expression data for principled discovery of genetic regulatory network models," *Proc. Pacific Symposium on Biocomputing* **7**, 437–449 (2002).
11. T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard and S. H. Friend, "Functional discovery via a compendium of expression profiles," *Cell* **102**, 109–126 (2000).
12. S. Imoto, T. Goto and S. Miyano "Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression," *Proc. Pacific Symposium on Biocomputing* **7**, 175–186 (2002).
13. S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara and S. Miyano, "Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network," *Journal of Bioinformatics and Computational Biology*, in press. (Preliminary version has appeared in *Proc. 1st IEEE Computer Society Bioinformatics Conference*, 219–227, 2002.)
14. F. V. Jensen, *An Introduction to Bayesian Networks*. University College London Press, 1996.
15. S. Liang, S. Fuhrman and R. Somogyi, "REVEAL, a general reverse en-gineering algorithm for inference of genetic network architectures," *Proc. Pac. Symposium on Biocomputing* **3**, 18–29 (1998).
16. Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe and Y. Eguchi, "Development of a system for the inference of large scale genetic networks," *Proc. Pacific Symposium on Biocomputing* **6**, 446–458 (2001).
17. M. J. Marton, M.J. Marton, J. L. DeRisi, H. A. Bennett, V. R. Iyer, M. R. Meyer, C. J. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, D. E. Bassett Jr., L. H. Hartwell, P. O. Brown and S. H. Friend, "Drug target validation and identification of secondary drug target effects using DNA microarrays of Expression Profiles," *Nat. Med.* **4**, 1293–1301 (1998).
18. D. Pe'er, A. Regev, G. Elidan and N. Friedman, "Inferring subnetworks from perturbed expression profiles," *Bioinformatics* **17**, Suppl.1 ISMB2001, 215–224 (2001).
19. J. Quackenbush, "Microarray data normalization and transformation," *Nature Genetics* **32**, 496–501 (2002).
20. C. J. Savoie, S. Aburatani, S. Watanabe, Y. Eguchi, S. Muta, S. Imoto, S. Miyano, S. Kuhara and K. Tashiro, "Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades," *DNA Research* **10**, 19–25 (2003).
21. I. Shmulevich, E. R. Dougherty, S. Kim and W. Zhang, "Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks," *Bioinformatics* **18**, 261–274 (2002).
22. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander and T. R. Gloub, "Interpreting patterns of gene expression with self-organizing maps:

Methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
23. H. Toh and K. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," *Bioinformatics* **18**, 287–297 (2002).

Seiya Imoto is currently a research associate of the laboratory of DNA analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. He received B.S., M.S., and Ph.D. in Mathematics from Kyushu University in 1996, 1998 and 2001, respectively. His current research interests cover analysis of high dimensional data with complex structure by using nonlinear statistical methods as well as development model selection criteria from an information theoretic or a Bayesian statistics approach, and analysis of cDNA microarray gene expression data.

Christopher Savoie is Founder, Chairman and CEO of GNI, a gene network driven pharmaceutical development company. He is also a visiting research scientist at the Institute of Medical Science at the University of Tokyo. He earned his doctor of medicine from the Faculty of Medicine of Kyushu University and is a published scholar in the fields of immunology, cancer, bioinformatics, computer science and systems biology. He previously founded Japan's first web consulting firm and was founding Chairman and CEO at Dejima, a successful Silicon Valley software venture based on his innovations in the fields of agent based software architectures and natural language processing. He is co-author on several US and international patents in both computational and biological science and was selected as a member of the MIT Technology Review TR100, a list of innovators under 35 to watch for the 21st century. His current research interests include gene and protein networks, genome-wide data analysis, and applications thereof to the fields of drug discovery and functional genomics.

Sachiyo Aburatani is currently an Assistant Professor of Laboratory of Biostatistics at Human Genome Center, Institute of Medical Science, University of Tokyo. She received her Ph.D. degree in Agriculture from Kyushu University in 2003. She has 5 peer-reviewed publications. Her current research interests cover the inference of interaction between biological molecules by combining information from various disparate resources.

Sunyong Kim is currently a master course student of the Laboratory of DNA Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. She received her B.S. in Computer Science from the University of Tokyo in 2002. Her current research interests include development of computational methods/software and establishment of statistical models for inferring gene regulatory pathways.

Kousuke Tashiro is currently an Associate Professor in Laboratory of Molecular Technics, Department of System Biology at Kyushu University. He received his Ph.D. degree in molecular biology from Kyushu University in 1981. He has more than 70 reviewed publications. His current research interests are the analysis of the mechanisms in gene regulatory networks in various biological phenomena such as cell proliferation, cell differentiation and cell adaptation against outer circumstances, using the gene expression profiles obtained by DNA-chip and computational methods.

Satoru Kuhara is currently a Professor of the Faculty of Agriculture Kyushu University. He received his Ph.D. degree in Biochemistry from the Kyushu University in 1980. He has more than 60 peer reviewed publications. His current research interests cover modeling of expression control of cells based on microarray experiments, modeling of developmental process in gene expression and generalization of relationship between structure and function of proteins.

Satoru Miyano is a Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He obtained B.S. in 1977, M.S. in 1979, and Ph.D. in Mathematics from Kyushu University. His current interests include computational gene network inference methods, modeling and simulation of biological systems, and computational knowledge discovery. He is on the Editorial Board of *Bioinformatics*, *Journal Bioinformatics and Computational Biology*, *Theoretical Computer Science* and is the Chief Editor of *Genome Informatics*.