Journal of Bioinformatics and Computational Biology
Vol. 3, No. 1 (2005) 61–77

ICP Imperial College Press
www.icpress.co.uk

BAYESIAN NETWORK LEARNING WITH FEATURE ABSTRACTION FOR GENE-DRUG DEPENDENCY ANALYSIS

JEONG-HO CHANG[†,‡], KYU-BAEK HWANG[†,‡‡],
S. JUNE OH[‖,¶] and BYOUNG-TAK ZHANG[†,§,*]

[†] *Biointelligence Laboratory, School of Computer Science and Engineering*
Seoul National University, Seoul 151-742, Korea

[‖] *Department of Pharmacology and Pharmacogenomics Research Center*
College of Medicine, Inje University, Busan 614-735, Korea

[‡] *jhchang@bi.snu.ac.kr*
[‡‡] *kbhwang@bi.snu.ac.kr*
[¶] *juno@inje.ac.kr*
[§] *btzhang@cse.snu.ac.kr*

Received 13 March 2004
Revised 23 June 2004
Accepted 2 July 2004

Combined analysis of the microarray and drug-activity datasets has the potential of revealing valuable knowledge about various relations among gene expressions and drug activities in the malignant cell. In this paper, we apply Bayesian networks, a tool for compact representation of the joint probability distribution, to such analysis. For the alleviation of data dimensionality problem, the huge datasets were condensed using a feature abstraction technique. The proposed analysis method was applied to the NCI60 dataset (http://discover.nci.nih.gov) consisting of gene expression profiles and drug activity patterns on human cancer cell lines. The Bayesian networks, learned from the condensed dataset, identified most of the salient pairwise correlations and some known relationships among several features in the original dataset, confirming the effectiveness of the proposed feature abstraction method. Also, a survey of the recent literature confirms the several relationships appearing in the learned Bayesian network to be biologically meaningful.

Keywords: Gene-drug dependency analysis; Bayesian networks; feature abstraction; microarray.

1. Introduction

The enormous amount of biological data obtained from the high-throughput experimental techniques offers a new opportunity to molecular biology and medicine. For example, combined analysis of the microarray and massive drug-activity datasets can give an insight into various correlations among gene expressions and drug activities in the malignant cell.[1,2]

*Corresponding author.

In this paper, we harness Bayesian networks to such analysis. Bayesian networks have mainly been used for the inference of gene networks from microarray data.[3–7] When adopting Bayesian networks for the analysis of huge biological data, the curse of dimensionality problem usually arises. To address this problem, we used a *feature abstraction* technique based on clustering to reduce the data dimensionality. The features, i.e. genes or drugs, are clustered according to their patterns over samples. Each resulting feature cluster maintains an *abstract feature*, called prototype, which is then used as a component (node) of the Bayesian network. It is expected that the established learning and probabilistic inference algorithms for Bayesian networks will be applicable through the feature abstraction. A similar approach was adopted for genetic network analysis using the graphical Gaussian modeling (GGM) technique in the work of Toh and Horimoto.[8] Their goal of data dimensionality reduction was to eliminate linear dependencies in the correlation coefficient matrix. Recently, Segal *et al.*[9] suggested a method for revealing gene regulation from gene expression profiles. They also adopted a clustering technique for identifying a set of co-regulated genes (modules). In contrast, the main purpose of data dimensionality reduction of our study is to make it feasible to analyze the massive biological datasets using Bayesian network learning. Moreover, we apply probabilistic inference for quantifying various relationships among genes, drugs, and cancer subtypes. Previous works did not exploit this feature of Bayesian networks for data analysis.

2. Materials and Methods

2.1. *The NCI60 dataset*

The NCI60 dataset[1] consists of 60 human cancer cell lines from 9 kinds of cancers: colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin cancers, as well as leukemias and melanomas. On each cell line, the gene expression pattern was measured by a cDNA microarray of 9703 genes including ESTs. Separate from this, 1400 chemical compounds were tested on the same 60 cell lines. The drug (chemical compound) activity on each cell line was measured by the growth inhibition activities (GI_{50}) assessed from changes in total cellular protein after 48 hours of drug treatment using sulphorhodamine B assay.

Before the analysis, we eliminated some genes and drugs for more reliable results. From 9703 genes, 1376 genes, having 4 or fewer missing values and showing strong variation among 60 cell lines (more than 3 measurements have red-green intensity ratios >2.6 or <0.38), were selected as in Scherf *et al.*'s work.[1] Furthermore, ESTs with no name were excluded from the analysis. From 1400 chemical compounds in the drug activity data, 118 anti-cancer drugs with known mechanisms were chosen for the analysis. Among them, drugs with more than 3 missing values were also thrown out. Consequently, the dataset in our analysis consists of 60 samples with 890 features: 805 gene expression levels in $\log_2$ ratios, 84 drug activities in $\log_{10}(1/GI_{50})$ values, and one additional variable for the cancer type. Figure 1 outlines the entire analysis scheme for the NCI60 dataset.

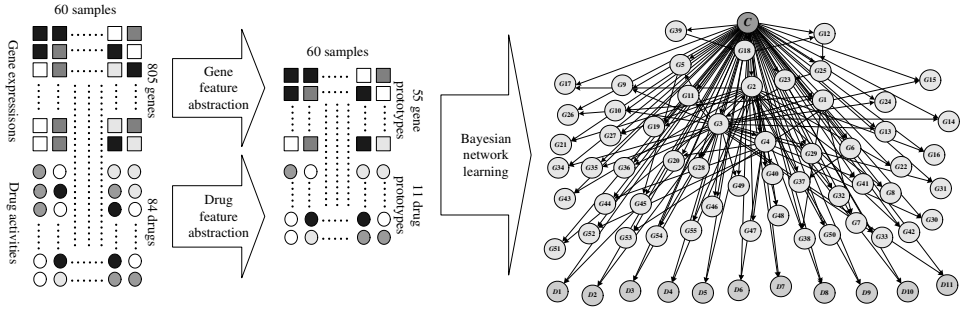


Fig. 1. The analysis procedure for the NCI60 dataset. The original dataset is condensed through the feature abstraction procedure. Then, the Bayesian network is learned from the reduced dataset.

2.2. *Bayesian networks for gene-drug dependency analysis*

A Bayesian network compactly represents the joint probability distribution over a set of M random variables, $\mathbf{X} = \{X_1, X_2, \ldots, X_M\}$:

$$P(\mathbf{X}) = \prod_{i=1}^{M} P(X_i | \mathbf{Pa}_{\mathcal{S}}(X_i)), \tag{1}$$

where $\mathbf{Pa}_{\mathcal{S}}(X_i)$ denotes the set of parents of X_i in the Bayesian network structure $\mathcal{S}$, which encodes the conditional independencies among the variables in $\mathbf{X}$, assuming a directed-acyclic graph (DAG) structure whose nodes are one-to-one correspondent to the variables. The conditional probability distribution $P(X_i|\mathbf{Pa}_{\mathcal{S}}(X_i))$ is called the local probability distribution of X_i.

Figure 2 shows an example Bayesian network for gene-drug dependency analysis. Here, the DAG shows a dependency structure among gene expressions, drug activities, and the kind of cancer. For example, we could gain an insight that the kind of cancer (*Cancer*) influences the activity level of the drug $D2$ through the expression pattern of genes $G1$ and $G2$. Such a dependency can be quantified by calculation of the conditional probability (probabilistic inference).

The Bayesian network for gene-drug dependency analysis can be learned using a greedy search algorithm.[10] The greedy search is guided by a scoring metric such as the BD (Bayesian Dirichlet) score.[11] The BD score for the structure $\mathcal{S}$ given the training data $\mathcal{D}$ is defined as

$$\begin{aligned} BD(\mathcal{S};\mathcal{D}) ::= & P(\mathcal{S},\mathcal{D}) \\ = & P(\mathcal{S})\, P(\mathcal{D}|\mathcal{S}) \\ = & P(\mathcal{S}) \int P(\mathcal{D}|\Theta,\mathcal{S})\, P(\Theta|\mathcal{S})\, d\Theta, \end{aligned} \tag{2}$$

where Θ denotes the set of the parameters for all the local probability distributions. With several assumptions such as *i.i.d.* complete data and Dirichlet prior,

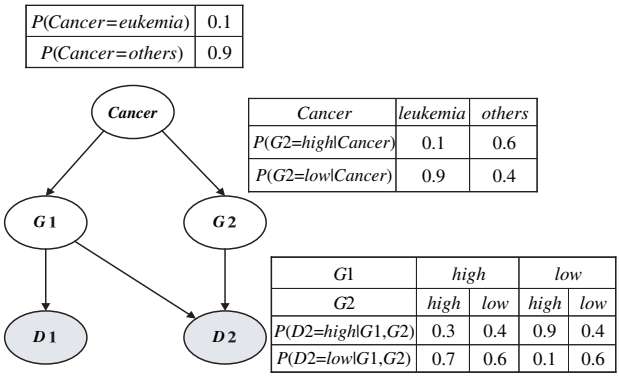


Fig. 2. A Bayesian network consisting of five variables: the kind of cancer ($Cancer$), expression levels of two genes ($G1$ and $G2$), and activity levels of two drugs ($D1$ and $D2$). Gene or drug variables have *high* and *low* as their values. The variable $Cancer$ has *leukemia* or *others* as its value. The local probability distribution of each node is represented as a table for the conditional probability distribution. The local probability distributions for $G1$ and $D1$ are not shown here.

the integral in Eq. (2) can be calculated in closed form as

$$\int P(D|\Theta,\mathcal{S})\,P(\Theta|\mathcal{S})\,d\Theta = \prod_{i=1}^{M}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{3}$$

where M is the number of variables of the Bayesian network, q_i is the number of configurations of parents of X_i, and r_i is the number of values of X_i. N_{ijk} is the number of cases in $\mathcal{D}$, where X_i assumes its kth value when its parents assume their jth configuration. α_{ijk} denotes the parameters of Dirichlet prior. $N_{ij} = \sum_k N_{ijk}$ and $\alpha_{ij} = \sum_k \alpha_{ijk}$. $\Gamma(\cdot)$ denotes the Gamma function.

2.3. *Feature abstraction for data dimensionality reduction*

The scheme described in the previous section is not appropriate for the analysis of the massive biological datasets. One reason involves the inapplicability of the greedy search algorithm. The massive biological datasets usually have more than hundreds of data features. Because the space of the possible Bayesian network structures grows super-exponentially with the number of variables, a greedy search requires an extremely large amount of time. To make matters worse, the obtained solutions are likely to be very far from the global optimum. The search space of Bayesian network structures is known to be very complicated and multimodal, having many local maxima.[12]

Another reason is related with the infeasibility of probabilistic inferences. The memory requirement or the computation time of general probabilistic inference

algorithms is usually exponential in the number of variables of the Bayesian network. Hence, the probabilistic inference to quantify the dependencies is impractical for Bayesian networks consisting of hundreds of variables. In order to mitigate these problems, we reduce the number of features in the massive biological datasets using a feature abstraction technique.

Formally, feature abstraction reduces an M-dimensional feature space $\mathbb{X} = (X_1, \ldots, X_M)$ into a K-dimensional prototype feature space $\mathbb{F} = (F_1, \ldots, F_K)$, where $K \ll M$. In this paper, the meaning of feature abstraction is confined to aggregating individual features (genes or drugs) which show similar patterns in their properties, that is, transcriptional expression (genes) or activity (drugs) across samples. A prototype feature F_k is defined by a partition subset $\mathbf{B}_k$ of the original feature set $\mathbf{X}$, where $\mathbf{B}_k \subset \mathbf{X}$ and $\mathbf{B}_k \cap \mathbf{B}_l = \emptyset$ $(l \neq k,\ 1 \leq l, k \leq K)$.

For the feature abstraction, we tried two different clustering algorithms: k-means clustering[13] with the Pearson correlation coefficient and the sequential Information Bottleneck (sIB) method.[14] The k-means algorithm is a partitioning clustering algorithm, where the total within-cluster variance is tried to be minimized. The distance between a gene (drug) x and a cluster center (average value of the cluster) c was set to

$$d(x, c) = 1 - r(x, c), \tag{4}$$

where $r(x, c)$ is the Pearson correlation coefficient between x and c. The Pearson correlation measure was chosen to group genes (drugs) which show similar relative, rather than absolute, variation pattern across patient samples. In terms of the metric between x and c, our clustering is closely related with the spherical k-means algorithm.[15]

The sIB algorithm formulates the clustering problem in an information-theoretic view and runs like the sequential k-means algorithm. Let X and Y be the random variables representing the set of genes (drugs) $x \in \mathcal{X}$ and the set of tissue samples $y \in \mathcal{Y}$, respectively. The cost of merging a gene (drug) x into a cluster c is given by

$$d(x, c) = (p(x) + p(c)) \cdot JS(p(y|x), p(y|c)), \tag{5}$$

where $JS(\cdot, \cdot)$ is the Jensen–Shannon (JS) divergence.[16]

The number of clusters (prototype features) was set by a method based on random permutation of data. Here, correlation coefficients are first calculated for all of the permutated gene (drug) pairs, and a cut-off value θ is set in such a way that only upper $\alpha\%$ of the values are higher than θ. Then, genes (drugs) in the original data are hierarchically clustered using an agglomerative method with average-linkage until all inter-cluster similarities are lower than θ. Finally, the number of clusters is set to the number of top-level subtrees in the dendrogram.[a]

[a]This method is similar to the approaches in the works of Herrero *et al.*[17] and Lukashin and Fuchs[18] in that the number of clusters is determined using a random permutation-based test, but is different in that the cut-off value is applied at the level of clusters not of individual elements.

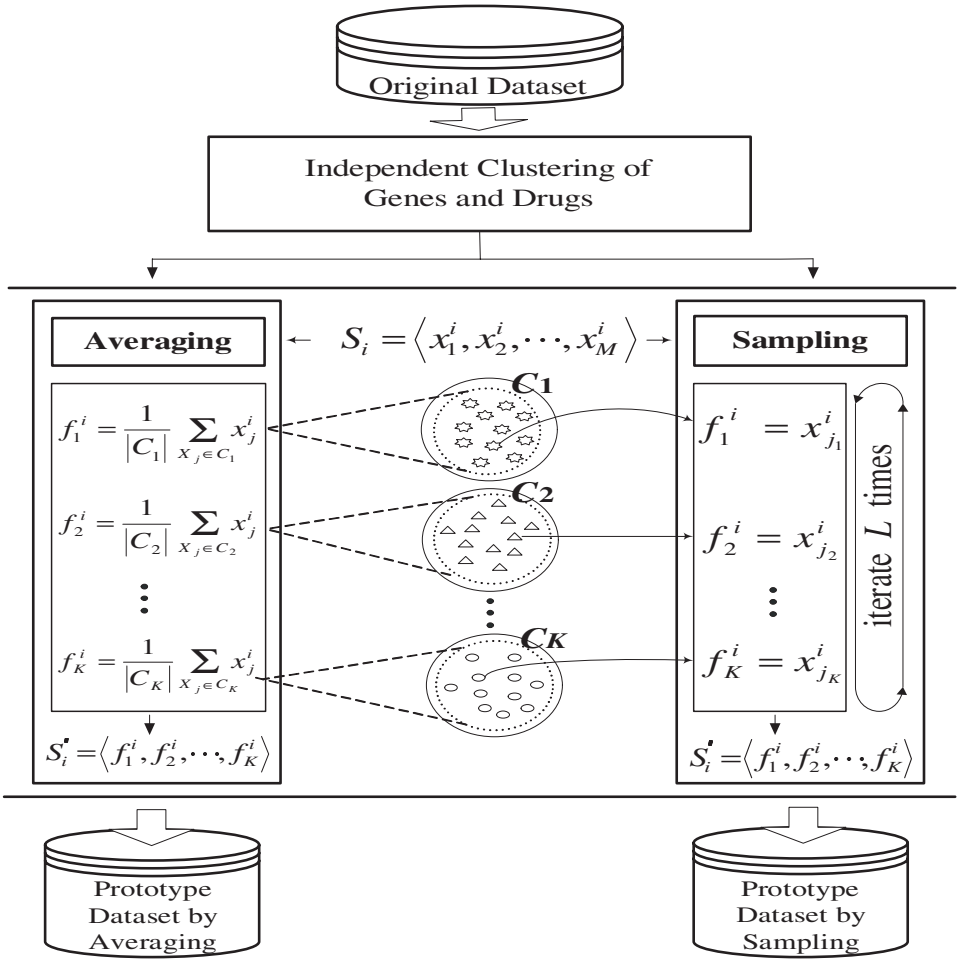


Fig. 3. Outline of the prototype dataset construction by feature abstraction. Genes and drugs are clustered independently, then prototype datasets are generated by feature averaging and feature sampling techniques, respectively. In feature averaging, the value of a prototype feature $f_k^i (1 \leq k \leq K)$ of a sample S_i is set to the average of expression (activity) values of the genes (drugs) in the cluster C_k. In feature sampling, f_k^i is set to the value of a gene (drug) randomly chosen in C_k.

2.4. *Prototype dataset construction by feature abstraction*

Let 805 genes and 84 drugs be clustered into K_1 and K_2 groups, respectively, such that the number of clusters is $K = K_1 + K_2$ in total. From the results, a condensed dataset with the reduced dimensionality ($K+1$: the number '1' is added for the cancer type) is generated by *feature averaging* or *feature sampling*, which we call the *prototype dataset* hereafter. Figure 3 summarizes the prototype dataset construction process.

In the feature averaging approach, the original sample S_i is transformed to S_i' of which the kth feature value f_k^i $(1 \leq k \leq K)$ is given by the arithmetic mean of

the expression (activity) values of genes (drugs) of the kth cluster. In the feature sampling approach, the kth feature value f_k^i of S_i' is set to the value x_j^i of an individual gene (drug) X_j randomly selected from the cluster C_k. The sampling process is repeated L times for each S_i, by which L new prototype samples are generated from S_i.

2.5. *Learning Bayesian networks from the prototype dataset*

As the local probability distribution model in learning Bayesian networks from the prototype dataset, we use the multinomial model which has mainly been adopted in the gene expression data analysis.[3–5,19b] All the feature values in the prototype dataset were discretized for the multinomial model. We divided the feature values into three levels (*low*, *normal*, and *high*) based on the mean (arithmetic average) and standard deviation of each feature. Here, *normal* just denotes the value around the mean value across all the samples. The discretization boundaries were calculated as $\mu - b \cdot \sigma$ and $\mu + b \cdot \sigma$, where μ is the mean value, σ is the standard deviation, and b is the constant which determines the breadth of the value *normal*. We tried several b values and set b as 0.60 in the experiments.

In the definition of the BD score in Eq. (2), $P(\mathcal{S})$ denotes the prior probability of the structure $\mathcal{S}$. We set $P(\mathcal{S}) \propto 2^{-\sum_i \left(\log M + \log\left(|\mathbf{Pa}_{\mathcal{S}}^{M}(X_i)|\right)\right)}$ for penalizing the complex structures. The parameters for Dirichlet prior α_{ijk} in Eq. (3) is set to the uninformative value 1.0. A greedy search algorithm[10] was adopted for the structural learning and the following constraints on the network structure were imposed to reduce the search space: 1) gene prototypes and drug prototypes can not be the parents of the cancer node, 2) drug prototypes can not be the parents of gene prototypes. The first constraint is based on the belief that the cancer type affects the gene expression patterns and the drug activities. The second is due to the nature of NCI60 dataset, where expression measurements of genes have been made on untreated cell lines, but not on those after the treatment by anti-cancer drugs, hence edges from a drug to genes are rather senseless. In addition, the maximum in-degree of the network is confined to two for compromising the extremely small sample size. Finally, to escape the local maxima that usually arise when relying on the greedy search, the greedy search with random restarts was run 20 times and the one with the highest BD score (Eq. (2)) was selected as an answer.

3. Results

3.1. *Comparison with pairwise correlation analysis*

First, we compare the results of our method with pairwise correlation analysis of the original dataset to investigate the influence of feature abstraction level (the number of prototype features) on the quality of the learned Bayesian network.

[b]Friedman *et al.*[3] experimented also with the linear Gaussian model and Imoto *et al.*[6,7] used the nonparametric regression model.

Distinct prototype datasets were generated with four confidence levels, $\alpha = 5, 10, 15, 20$, in the random permutation-based test as described in the previous section about feature abstraction by clustering. The corresponding (# of gene prototypes, # of drug prototypes) were (55, 11), (34, 9), (24, 6), and (16, 5), where the cut-off value pairs were $(\theta_{gene}, \theta_{drug}) = (0.214, 0.218)$, (0.168, 0.169), (0.136, 0.134), (0.110, 0.110), respectively. In each case, two kinds of prototype datasets were constructed by averaging and sampling. In the latter approach, 100 subsamples were generated for each cell line (that is, $L = 100$ in Fig. 3). Consequently, 16 prototype datasets, thereby 16 Bayesian networks, were constructed according to the number of clusters (4 cases), clustering method (2 cases), and the method for making prototype values (2 cases).

From the original dataset, the correlation coefficients for the following feature pairs were calculated. Among 323,610 gene-gene pairs, the top 5000 negatively-correlated ones were tested. For drug-drug pairs, the top 50 negatively-correlated pairs out of 3486 pairs were examined. For 67,620 gene-drug pairs, the top 50 negatively-correlated pairs and the top 50 positively-correlated pairs were tested.[c]

For the comparison, we defined an information-theoretic measure based on the concept of conditional probability. Assume that gene A in the gene prototype $G1$ and gene B in the gene prototype $G2$ are highly negatively-correlated in the original dataset. Then, the measure for the negative correlation between $G1$ and $G2$ in the Bayesian network, $NC(G1||G2)$ is defined as follows:

$$\begin{aligned} NC(G1||G2) := {} & 0.5 \cdot JS(P(G1 = low), P(G1|G2 = low)) \\ & + 0.5 \cdot JS(P(G1 = high), P(G1|G2 = high)) \\ & + JS(P(G1|G2 = low), P(G1|G2 = high)), \end{aligned} \tag{6}$$

where $JS(\cdot, \cdot)$ is the JS divergence. The measure for the positive correlation, $PC(G1||G2)$ is defined as the same as $NC(G1||G2)$ except that the sign of the third term, $JS(P(G1|G2 = low), P(G1|G2 = high))$ is minus. The baseline for the two measures is when $G1$ and $G2$ are independent from each other, as well as $P(G1|G2 = low)$ and $P(G1|G2 = high)$ are $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ respectively. The baseline value amounts to about 0.459 using the logarithm of base 2. All the values above (below) this baseline are considered the sign of negative (positive) correlation in the Bayesian network.

Figure 4 shows the comparison results. In the figure, it can be seen that the quality of the learned Bayesian network is affected by the level of feature abstraction. As the number of prototype features increases, the positive and negative correlations are represented well in general. In the sequel, we present the analysis results using the model built by k-means clustering of 5% confidence level, (# of

[c]In the case of the gene-gene (drug-drug) pair, positively correlated pairs are likely to be clustered into one gene (drug) prototype through the process of feature abstraction. Hence, we only consider negatively correlated gene-gene (drug-drug) pairs here.

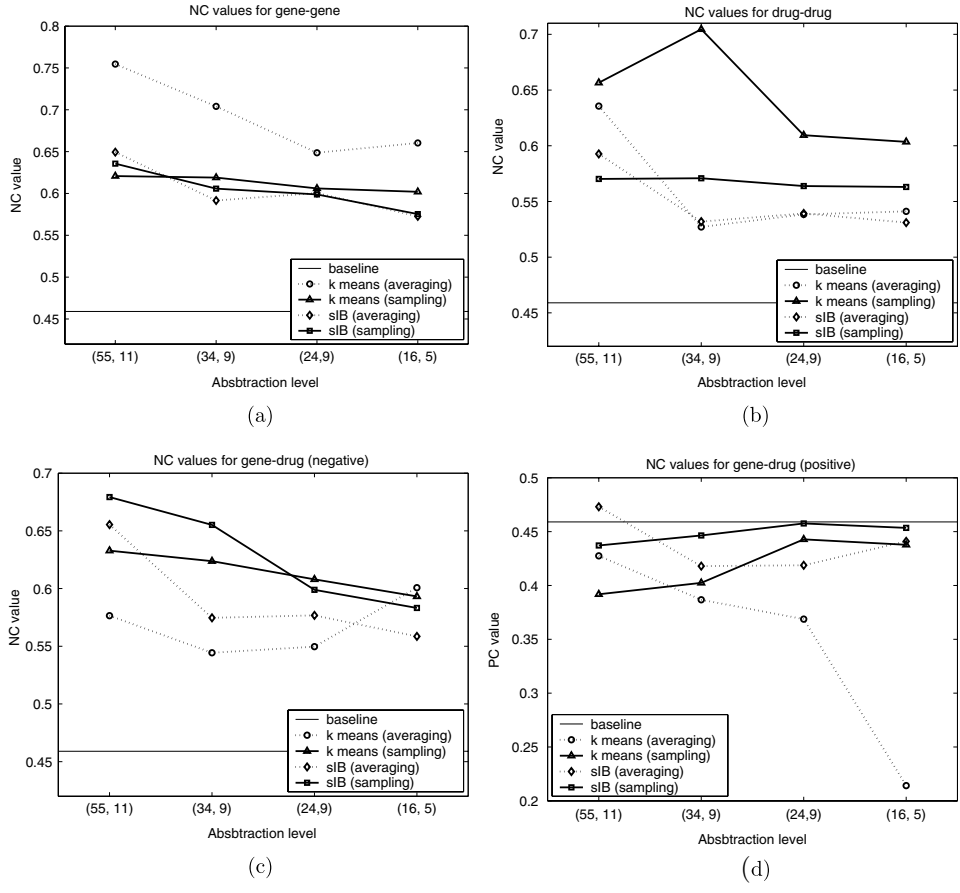


Fig. 4. Results of the comparison of the Bayesian network analysis through feature abstraction and the plain correlation analysis on the prominent pairwise correlations in the original dataset. The graphs show the mean of $NC(\cdot||\cdot)$ and $PC(\cdot||\cdot)$ for (a) gene-gene, (b) drug-drug, and (c) and (d) gene-drug pairs. It can be seen that the quality of the learned Bayesian network is affected by the level of feature abstraction. As the number of prototype features grows up, the positive and negative correlations are represented well in general.

gene prototypes, # of drug prototypes) = (55, 11), with the sampling approach for prototype data generation, since it shows acceptable agreement with four kinds of pairwise correlations. The Bayesian network is shown in Fig. 5. From this figure, we could gain insight on the dependency structure over gene prototypes, drug prototypes, and the cancer type, i.e. a condensed representation of the relationships among the features of the NCI60 dataset. Arbitrary conditional probabilities of interest can be calculated through the probabilistic inference, which is intractable on the Bayesian network consisting of hundreds of variables due to the exponential time/space complexity in the size of the network.[20]

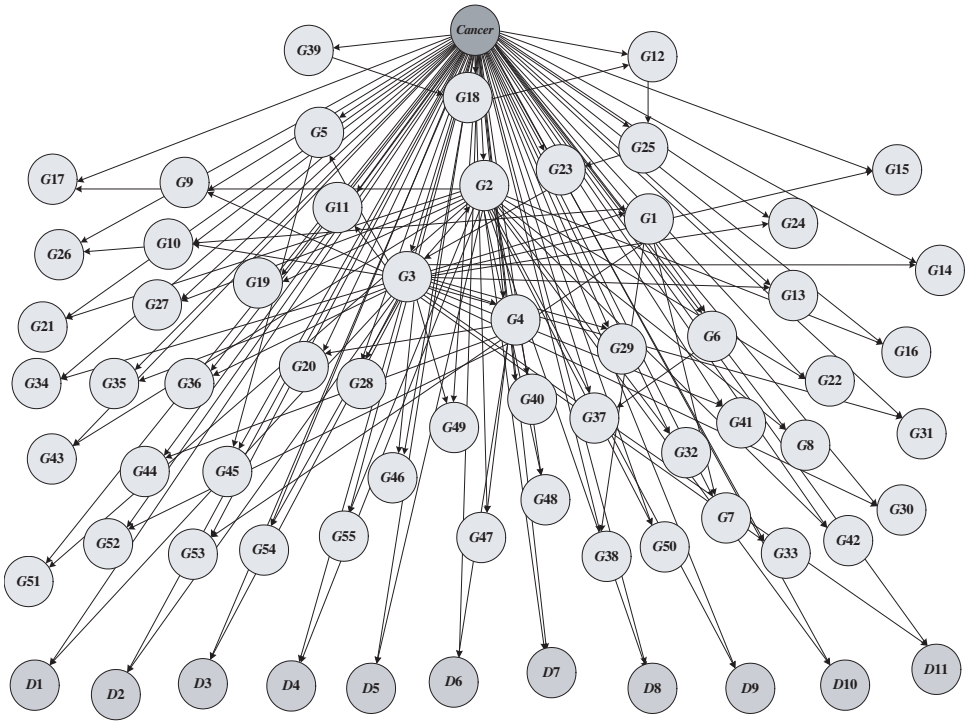
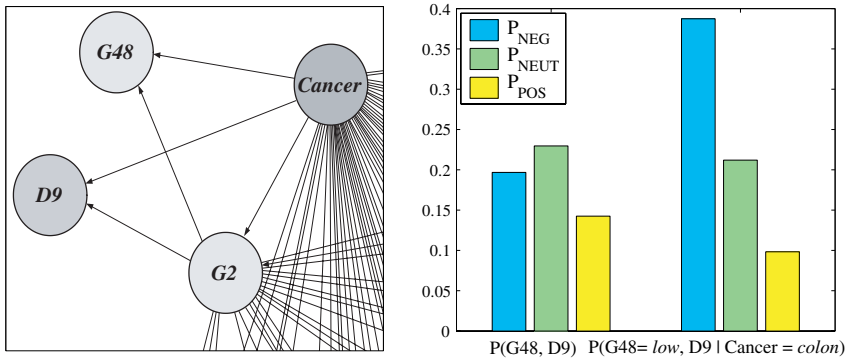


Fig. 5. The Bayesian network learned from the prototype dataset consisting of 55 gene prototypes, 11 drug prototypes, and the cancer type. This prototype dataset was built by k-means clustering with 5% confidence level and the sampling approach for making prototype feature values. The node *Cancer* influences much of the gene and drug prototypes. Several gene prototypes such as $G2$ affect many of other components. The gene prototype $G2$ includes only one member, the gene *BMI1* (murine leukemia viral (bmi-1) oncogene homolog Chr.10 [418004, (REW), 5′:W90704, 3′:W90705]).

3.2. *Beyond pairwise correlations*

Here, it is shown that more intricate relationships involving several features can be detected by our method. The target is two plausible relationships among gene expressions, drug activities, and the kind of cancer, obtained from Scherf *et al.*'s work.[1]

One is the relationship among the gene *DPYD* (SID W 278125, dihydropyrimidine dehydrogenase [5′:N94809, 3′:N63511]), the drug 5-FU (5-fluorouracil), and the kind of cancer. The drug 5-FU is clinically used to treat colorectal and breast cancers. The expression level of gene *DPYD* and the activity of drug 5-FU are highly negatively-correlated. Figure 6(a) shows the part of the Bayesian network around the gene prototype $G48$ and the drug prototype $D9$ to which *DPYD* and 5-FU belong, respectively.



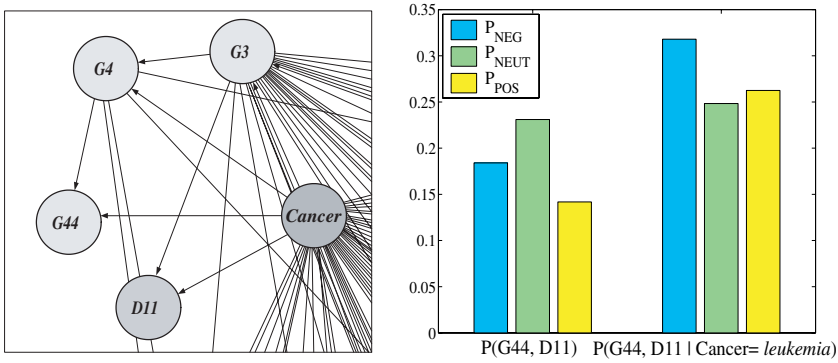
(a) (b)

Fig. 6. (a) Part of the Bayesian network around $G48$ (containing $DPYD$) — $D9$ (containing 5-FU) and (b) the probabilistic inference in the network. Each bar group in (b) depicts three kinds of probabilities: 'negative', 'neutral', and 'positive' cases from left to right. The left bar group represents $P(G48, D9)$ and the right group represents $P(G48 = low, D9|Cancer = colon)$ which is the case where the kind of cancer is colorectal and the expression level of G48 is low.

To quantify the relationship among gene (G) expressions, drug (D) activities, and the kind of cancer ($Cancer$), we infer the probabilities $P(G, D)$ and $P(G, D|Cancer)$ from the Bayesian network. In specific, the probabilities for the following three cases according to the drug activities when the gene expression level is low or $high$ are calculated: $P_{NEG}(G, D) = P(G = low, D = high) + P(G = high, D = low)$, $P_{NEUT}(G, D) = P(G = low, D = normal) + P(G = high, D = normal)$, $P_{POS}(G, D) = P(G = low, D = low) + P(G = high, D = high)$, where $P_{NEG}(G, D)$ and $P_{POS}(G, D)$ are the joint probabilities of G and D with negative and positive relations respectively, and $P_{NEUT}(G, D)$ is for the neutral case where drug activity is $normal$ irrespective of the expression level of G. $P(G, D|Cancer)$ is calculated in the same way as for $P(G, D)$, except that each probability is conditioned on the specific kind of cancer.

Figure 6(b) shows the results of probabilistic inference on $G48$ and $D9$. Here, the negative tendency between $G48$ and $D9$ is observed well.[d] Furthermore, this negative tendency becomes more definite when the kind of cancer is colorectal. For the cell lines of colorectal cancer where $G48$ is low, especially, the probability that $D9$ is $high$ ($P(G48 = low, D9 = high|Cancer = colon) = 0.387$) is prominent compared to the probability that $D9$ is low ($P(G48 = low, D9 = low|Cancer = colon) = 0.098$). In the original dataset, in practice, all of the colon-derived cell lines show low expression level of $DPYD$ and are highly susceptive to 5-FU. The

[d]Considering the bias due to the randomness in the feature sampling approach, we generated the prototype datasets ten times and built Bayesian networks respectively from these ten datasets. The variations in the estimated probabilities were very small (all standard deviations were in the range of (0.001, 0.013)) and the mean probabilities for the 10 datasets are shown here.



(a) (b)

Fig. 7. (a) Part of the Bayesian network around $G44$ (containing *ASNS*) — $D11$ (containing L-asparaginase) and (b) the probabilistic inference in the network. The left bar group represents $P(G44, D11)$ and the right group represents $P(G44, D11|Cancer = leukemia)$, which is the case where the kind of cancer is leukemia.

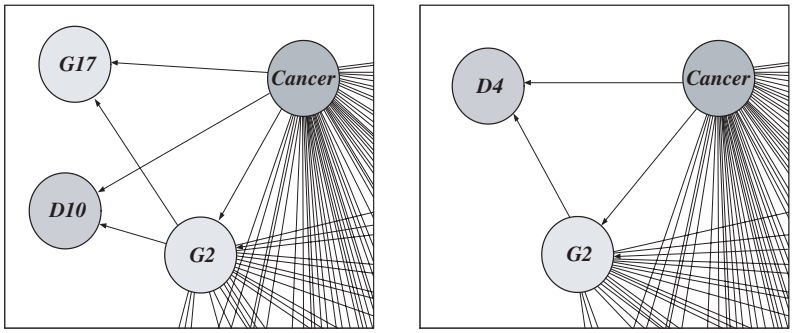
probabilities for $G48 = high$ were not included in the result, since the probability $P(G48 = high|Cancer = colon)$ inferred from the Bayesian network is so small (0.023) that it is statistically insignificant (actually, no high expression of *DPYD* are observed in colorectal lines of the original data set).

The other relationship is among the expression of gene *ASNS* (asparagine synthetase Chr.7 [510206, (IW), 5′:AA053213, 3′:AA053461]), the activity of anticancer drug L-asparaginase, and the kind of cancer. L-asparaginase depletes the amino acid asparagine external to cells. Tumor cells, especially lymphatic tumor cells, need a great amount of asparagine for their growth. There exists a moderately high negative-correlation between the expression level of *ASNS* within tumor cells and the activity of L-asparaginase. Figure 7 shows the part of the Bayesian network around the gene prototype $G44$ and the drug prototype $D11$ that contain *ASNS* and L-asparaginase, respectively, and the inference results in the network. In Fig. 7(b), it can be seen that $P_{NEG}(G44, D11)$ is higher than $P_{POS}(G44, D11)$, where $P_{NEG}(G44, D11) = 0.184$ and $P_{POS}(G44, D11) = 0.142$. For the leukemia cell lines, $P_{NEG}(G44, D11|Cancer = leukemia) = 0.318$ and $P_{POS}(G44, D11|Cancer = leukemia) = 0.263$. The results accord with the negative tendency between the expression level of *ASNS* and the activity of L-asparaginase.

3.3. *Exploratory analysis using Bayesian networks*

In this section, we illustrate several salient features discovered by the Bayesian network described in the previous section and present the results of the literature survey related to them.

Relation between *G17* and *D10*: The dependency structure for $G17$ and $D10$ is depicted in Fig. 8(a). These two components are highly negatively-correlated



(a) (b)

Fig. 8. The network structures around (a) *Cancer*, $G17$, and $D10$ and (b) *Cancer*, $G2$, and $D4$. They show prominent negative correlations in the Bayesian network measured by the $NC(\cdot||\cdot)$ metric.

($NC(G17||D10)$ is 0.720). However, the negative correlation seems to be mediated by other components, such as the kind of cancer and other genes in the Bayesian network. Eighteen genes in all belong to the prototype $G17$, including *ITGB1* (integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12) Chr.10 [486375, (IW), 5′:AA044145, 3′:AA044261]), *uPA* (SID 486215, urokinase-type plasminogen activator [5′:, 3′:AA040727]), *IL-8* (interleukin 8 Chr.4 [328692, (DW), 5′:W40283, 3′:W45324]), and so on. The drug prototype $D10$ contains six drugs, where four members (paclitaxel, colchicine, trityl-cysteine, and vinblastine-sulphate) are tubulin-active antimitotic agents and the remaining two drugs are geldanamycin and bisantrene.

Paclitaxel (taxol) is known to act as an anti-cancer agent by inducing microtubule stabilization or apoptosis of tumor cells. Two genes in $G17$, *uPA* and *ITGB1*, can affect the activation of paclitaxel in breast cancer cells, respectively. *uPA* catalyzes the conversion of plasminogen to plasmin. Plasmin induces the formation of multicellular spheroids of breast cancer cells, by which tumor cells get to acquire an increased resistance to chemotherapeutic drugs, including paclitaxel.[21] In the case of *ITGB1*, its signaling has been identified as an important survival pathway in drug-induced apoptosis in breast cancer cells.[22] Especially, the ligation of *ITGB1* by extracellular matrix ligands can considerably inhibit apoptosis induced by paclitaxel. From these, it is expected that the chemotherapeutic effect of paclitaxel might be restrained by high expression of *ITGB1* or *uPA*, by which the inferential negative relation between $G17$ and $D10$ in cells of specific cancer type (breast cancer cells, here) can be supported.

In fact, *ITGB1* and *paclitaxel* are ranked at the 1st and the 2nd in the Pearson correlation similarity with the center profiles of $G17$ and $D10$, respectively. The *uPA* — *paclitaxel* pair is within the lowest-valued top ten among all member pairs, in terms of the JS divergence between its empirical probabilistic dependency profile

and the inference result for $G17$ and $D10$ in the Bayesian network. The empirical pairwise dependency of a gene-drug pair is obtained directly from data, by calculating conditional probabilities based on their discretized patterns.

Relation between $G2$ and $D4$: Figure 8(b) shows the network structure around $G2$ and $D4$. These two nodes are directly linked in the network structure and also affected by the node of cancer type. These two prototypes are negatively correlated to each other ($NC(G2||D4)$ is 0.685). $G2$ and $D4$ contain only one member, the gene *BMI1* (murine leukemia viral (bmi-1) oncogene homolog Chr.10 [418004, (REW), 5′:W90704, 3′:W90705]) and the drug azacytidine, respectively.

The structure and functioning mechanism of the anti-cancer drug azacytidine (member of $D4$) are very similar to 5-aza-2′-deoxycytidine. And 5-aza-2′-deoxycytidine is known to promote the tumor suppressing protein p14/ARF,[23] whereas the gene *BMI1* (member of $G2$) is known to negatively regulate p14/ARF.[24] In the cancer cell line which shows high expression level of *BMI1*, it is anticipated that the functioning mechanism of azacytidine might not work because of the possibility of the interruption by *BMI1*, suppressor of p14/ARF. This hypothesis supports the observed negative correlation between the expression level of $G2$ and the activity of $D4$.

4. Discussion

We presented an analysis method using Bayesian networks that incorporates a feature abstraction technique to alleviate the curse of dimensionality problem, which thus makes it applicable to the analysis of massive biological datasets. The Bayesian network can capture the entire dependency structure immanent in the dataset. This property provides a means for encoding the valuable information contained in the large-scale datasets with noise. Further, the dependency among several features can be quantified by probabilistic inference. The probabilistic inference is used for calculating the conditional probability of interest, assisting the decision making.

It should be noted that our analysis is different from conventional regression or classification tasks, i.e. supervised learning. In our problem setting, any gene or drug prototype can be regarded as a dependent or an independent variable. In this sense, our method produces a huge number of dependency structures of similar shapes. To identify a small set of eligible ones among them, we exploited a measure for the negative correlation between gene and drug prototypes. The dependency structures with high negative-correlation ($NC(\cdot||\cdot)$) were examined by literature survey. Through this, we could confirm several dependency structures among gene expressions, drug activities, and the kind of cancer. Our findings were obtained in a purely data-driven way, and this demonstrates the potential of our method as an efficient data mining tool for the revelation of the relations between gene expression patterns and drug activities in cancer cells.

There are several directions for future work. One is to incorporate hidden nodes in Bayesian network learning so that discovery of more elaborated interactions

among genes or drugs is possible. Another direction is to combine other information sources as prior knowledge. The protein-protein or DNA-protein interaction DB can be exploited for the more precise structural learning. Additionally, other sorts of data such as polymorphisms in key enzymes or gene expression patterns from cell lines treated with some drugs (e.g., expression profile data of responses to drugs, as in the work of Imoto *et al.*[7], for cancer cell lines) could provide more biologically useful information.

Acknowledgments

This work was supported by the Korea Ministry of Science and Technology under the National Research Laboratory (NRL) Program and by the Korea Ministry of Education & Human Resources Development under the BK21-IT Program. The ICT at Seoul National University provided research facilities for this study.

References

1. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN, A gene expression database for the molecular pharmacology of cancer, *Nat Genet* **24**(3):236–244, 2000.
2. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS, Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proc Natl Acad Sci USA* **97**(22):12182–12186, 2000.
3. Friedman N, Linial M, Nachman I, Pe'er D, Using Bayesian networks to analyze expression data, *J Comput Biol* **7**(3/4):601–620, 2000.
4. Pe'er D, Regev A, Elidan G, Friedman N, Inferring subnetworks from perturbed expression profiles, *Bioinformatics* **17**(Suppl 1):S215–S224, 2001.
5. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA, Combining location and expression data for principled discovery of genetic regulatory network models, *Pacific Symposium on Biocomputing* **7**:437–449, 2002.
6. Imoto S, Kim S, Goto T, Aburatani S, Tashiro K, Kuhara S, Miyano S, Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *J Bioinform Comput Biol* **1**(2):231–252, 2003.
7. Imoto S, Savoie CJ, Aburatani S, Kim S, Tashiro K, Kuhara S, Miyano S, Use of gene networks for identifying and validating drug targets, *J Bioinform Comput Biol* **1**(3):459–474, 2003.
8. Toh H, Horimoto K, Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics* **18**(2):287–297, 2002.
9. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat Genet* **34**(2):166–176, 2003.
10. Friedman N, Goldszmidt M, Learning Bayesian networks with local structure, in Jordan MI (ed.), *Learning in Graphical Models*, pp. 421–459, MIT Press, Cambridge, 1999.
11. Heckerman D, Geiger D, Chickering DM, Learning Bayesian networks: the combination of knowledge and statistical data, *Mach Learn* **20**(3):197–243, 1995.
12. Friedman N, Koller D, Being Bayesian about network structure: a Bayesian approach to structure discovery in Bayesian networks, *Mach Learn* **50**(1):95–125, 2003.

13. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM, Systematic determination of genetic network architecture, *Nat Genet* **22**(3):281–285, 1999.
14. Slonim N, Friedman N, Tishby N, Unsupervised document classification using sequential information maximization, in *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 129–136, 2002.
15. Dhillon IS, Modha DM, Concept decompositions for large sparse text data using clustering, *Mach Learn* **42**(1): 143–175 2001.
16. Lin J, Divergence measures based on the Shannon entropy, *IEEE T Inform Theor* **37**(1):145–151, 1991.
17. Herrero J, Valencia A, Dopazo J, A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics* **17**(2):126–136, 2001.
18. Lukashin AV, Fuchs R, Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters, *Bioinformatics* **17**(5):405–414, 2001.
19. Chang J-H, Hwang K-B, Zhang B-T, Analysis of gene expression profiles and drug activity patterns by clustering and Bayesian network learning, in Lin SM, Johnson KE, (eds.), *Methods of Microarray Data Analysis II (Proceedings of CAMDA'01)*, pp. 169–184, Kluwer Academic Publishers, 2002.
20. Cooper GF, Computational complexity of probabilistic inference using Bayesian belief networks, *Artif Intell* **42**(2–3):393–405, 1990.
21. Chun MH, Plasmin induces the formation of multicellular spheroids of breast cancer cells, *Cancer Lett* **117**(1):51–56, 1997.
22. Aoudjit F, Vuori K, Integrin signaling inhibits paclitaxel-induced apoptosis in breast cancer cells, *Oncogene* **20**(36):4995–5004, 2001.
23. Magdinier F, Wolffe AP, Selective associaion of the methyl-CpG binding protein MBD2 with the silent *p14/p16* locus in human neoplasia, *Proc Natl Acad Sci USA* **98**(9):4990–4995, 2001.
24. Lindstroem MS, Klangby U, Wiman KG, p14ARF homozygous deletion or MDM2 overexpression in Burkitt lymphoma lines carrying wild type p53, *Oncogene* **20**(17): 2171–2177, 2001.

Jeong-Ho Chang is currently a Ph.D. candidate in School of Computer Science and Engineering, Seoul National University, Korea. He received his BS and MS degrees in Computer Engineering from Seoul National University in 1995 and 1997, respectively. His research interests include gene expression data analysis using probabilistic graphical models, kernel-based biological data analysis, integrated analysis of multiple genomic data, and text mining based on latent variable models.

Kyu-Baek Hwang is currently a Ph.D. candidate in School of Computer Science and Engineering, Seoul National University, Korea. He received his BS and MS degrees in Computer Engineering from Seoul National University in 1997 and 1999, respectively. His research interests include efficient (approximate) Bayesian learning for belief networks (using MCMC techniques), practical feature selection methods for sparse datasets, microarray data analysis, and gene expression pathway construction using Bayesian learning methods.

S. June Oh is currently a full-time instructor of Department of Pharmacology and Pharmacogenomics Research Center, College of Medicine, Inje University, Korea. He received Ph.D. in agricultural biotechnology from Seoul National University in 2000. Formerly, he worked as R&D section head of the Center for Bioinformation Technology (CBIT) of Institute of Computer Technology, Seoul National University. He is a member the staff of KSBI (Korean Society for Bioinformatics) and KSMCB (Korean Society for Molecular and Cellular Biology). His research interests include biological information processing, biomedical informatics, and biomolecular computing.

Byoung-Tak Zhang is currently an associate professor of School of Computer Science and Engineering at Seoul National University (SNU), Korea, and directs the Biointelligence Laboratory and the Center for Bioinformation Technology (CBIT) at SNU. He received his BS and MS degrees in Computer Engineering from SNU in 1986 and 1988, respectively, and a Ph.D. in Computer Science from University of Bonn, Germany in 1992. Prior to joining SNU, he had been a research associate at German National Research Center for Information Technology (GMD). He serves as an associate editor of IEEE Transactions on Evolutionary Computation, Advances in Natural Computation, and Genomics & Informatics. His research interests include probabilistic models of learning and evolution, DNA/biomolecular computing, molecular learning/evolvable machines.