11

ANALYSIS OF GENE EXPRESSION PROFILES AND DRUG ACTIVITY PATTERNS BY CLUSTERING AND BAYESIAN NETWORK LEARNING

Jeong-Ho Chang, Kyu-Baek Hwang, and Byoung-Tak Zhang
Biointelligence Laboratory, School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea

Abstract: High-throughput genomic analysis provides insight into a complicated biological phenomena. However, the vast amount of data produced from up-to-date biological experimental processes needs appropriate data mining techniques to extract useful information. In this paper, we propose a method based on cluster analysis and Bayesian network learning for the molecular pharmacology of cancer. Specifically, the NCI60 dataset is analysed by soft topographic vector quantization (STVQ) for cluster analysis and by Bayesian network learning for dependency analysis. Our results of the cluster analysis show that gene expression profiles are more related to the kind of cancer than to drug activity patterns. Dependency analysis using Bayesian networks reveals some biologically meaningful relationships among gene expression levels, drug activities, and cancer types, suggesting the usefulness of Bayesian network learning as a method for exploratory analysis of high-throughput genomic data.

Key words: Gene expression pattern, drug activity pattern, molecular pharmacology, soft topographic vector quantization (STVQ), Bayesian networks

1. INTRODUCTION

Recent developments in the technology for biological experiments have made it possible to produce massive biological datasets. For example, microarrays obtained from cDNA chips or oligonucleotide chips provide a parallel view of the expression pattern of tens of thousands of genes in a sample. These massive datasets provide an opportunity to broaden the knowledge of the complex biological phenomena, but also require appropriate analysis techniques different from conventional methods for the traditional one-gene-in-one-experiment paradigm. Until now, diverse methods from the statistics and machine learning fields, such as hierarchical clustering [Eisen *et al.*, 1998], principal component analysis (PCA) [Raychaudhuri *et al.*, 2000], neural networks [Khan *et al.*, 2001], and Bayesian networks [Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Hwang *et al.*, 2001], have been applied to high-throughput genomic analysis. In data analysis, it is most important to adopt the appropriate methods to the purpose of the analysis.

In this paper, the NCI60 dataset [Scherf *et al.*, 2000] is analysed for the molecular pharmacology of cancer. The NCI60 dataset consists of 60 human cancer cell lines from 9 kinds of cancers, which are colorectal, renal, ovarian, breast, prostate, lung, and central nervous system origin cancers, as well as leukaemias and melanomas. On each cell line, the gene expression pattern is measured by a cDNA microarray of 9,703 genes including ESTs. Also, 40 molecular targets other than mRNA are assessed. And 1,400 chemical compounds are tested on the 60 cell lines. These compounds include some anticancer drugs that are currently in clinical use. The drug activity on the cell line is measured by the growth inhibition assessed from changes in total cellular protein after 48 hours of drug treatment using sulphorhodamine B assay [Scherf *et al.*, 2000].

We use soft topographic vector quantization (STVQ) [Graepel, 1998] for cluster analysis and Bayesian network learning for dependency analysis. In the cluster analysis, 60 cell lines are clustered based on the gene expression patterns and drug activity patterns. Dependency analysis aims to model the probabilistic relationships among the expression level of each gene, the activity of each drug, and the kind of cancer.

The paper is organized as follows. In Section 2, the cluster analysis by STVQ is described. The dependency analysis by Bayesian network learning is described in Section 3. Finally, the conclusion and some directions for further research are given in Section 4.

2. CLUSTER ANALYSIS OF THE NCI60 DATASET

We have clustered the 60 human cancer cell lines of the NCI60 dataset based on gene expression patterns and drug activity patterns, respectively. In the experiments, we investigate if there is a common pattern in gene expression and drug activities of the cell lines from the same tissue of origin, and thus if cell lines of the same cancer type can be clustered appropriately.

2.1 Soft Topographic Vector Quantization

Soft topographic vector quantization (STVQ) [Graepel, 1998] is a clustering algorithm based on principles from statistical physics. It can provide not only a stable and good clustering solution, but also a topographic map of the clustered data.

In this algorithm, clustering is defined in terms of an optimisation problem. The cost function to be optimised is given as

$$E = \sum_{i=1}^{N} \sum_{j=1}^{M} m_{ij} e_{ij}, \qquad [1]$$

where N is the number of samples and M is the number of clusters. m_{ij} is a binary variable indicating whether the i^{th} sample belongs to the j^{th} cluster, and e_{ij} is the error occurred by assigning the i^{th} sample to the j^{th} cluster. The error term is defined as

$$e_{ij} = \frac{1}{2} \sum_{k=1}^{M} h_{jk} \left\| \mathbf{x}_i - \mathbf{z}_k \right\|^2, \quad \sum_{k=1}^{M} h_{jk} = 1 \ (\forall j), \qquad [2]$$

where $\mathbf{x}_i$ is a sample vector and $\mathbf{z}_k$ is a cluster centre whose value is determined by the average of the sample vectors assigned to it. h_{jk} is a neighbourhood function between j^{th} and k^{th} clusters,. By introducing h_{jk} for every pair of clusters, STVQ is able to visualize the cluster structure in the same way as the self-organizing map (SOM) does in the one- or two-dimensional space.

STVQ provides an efficient procedure to find a good solution to the minimization of Equation 1 based on the maximum entropy principle and the idea of deterministic annealing. It is initialised with a random configuration as a K-means algorithm and proceeds using an iterative optimisation method, the EM algorithm [Dempster *et al.*, 1977], with some annealing schedule. In the *E*-step, the expectation value of m_{ij}, that is the probability that the sample $\mathbf{x}_i$ belongs to the j^{th} cluster, is estimated for each pair of samples and clusters. Then, all the cluster centres are calculated in the *M*-step. These two steps are iteratively alternated until convergence. More details about STVQ can be found in [Graepel, 1998].

2.2 Clustering of the NCI60 Cell Lines Using STVQ

The NCI60 dataset comprises two matrices, called the **T** matrix and the **A** matrix. In the **T** matrix, each cell line is represented by 1,416 attributes that include 1,376 genes and 40 molecular characteristics. The 1,376 genes are those with strong patterns of variation among the cell lines and with less than or equal to 4 missing values [Scherf *et al.*, 2000]. Each cell line in the **A** matrix is represented by the activity values of 1,400 chemical compounds.

For each cell line in the **T** matrix, all of its attribute values were standardized (mean value is 0 and the standard deviation is 1) across 1,416 attributes, including genes and individual targets. Likewise, all the drug activity values of each cell line in the **A** matrix were standardized. Now, each cell line is represented as a vector, where the vector $\mathbf{x}_i$ corresponds to the i^{th} cell line.

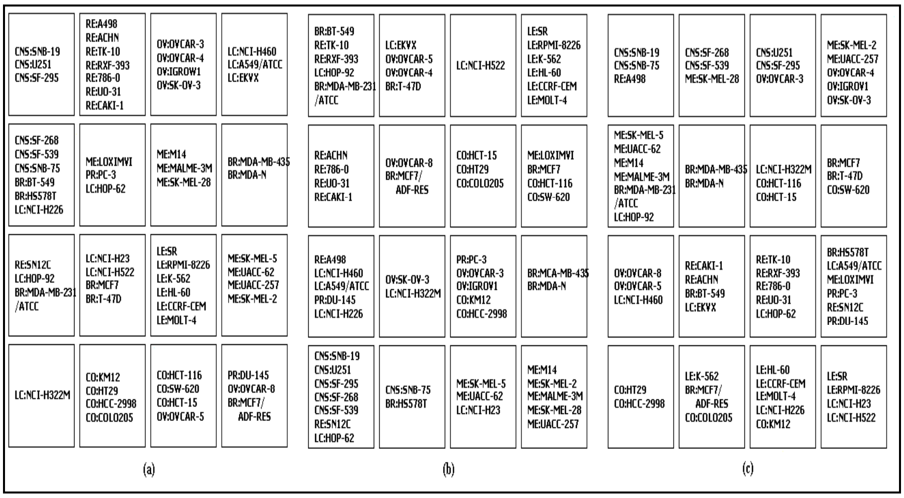
First, we have clustered the 60 cell lines based on the gene expression profiles. For each cluster centre $\mathbf{z}_k$, all of its attribute values are standardized after every update. Then the squared Euclidean distance in Equation 2 is closely related with the Pearson correlation coefficient. That is,

$$\begin{aligned}\left\|\mathbf{x}_i - \mathbf{z}_k\right\|^2 &= (\mathbf{x}_i^T\mathbf{x}_i + \mathbf{z}_k^T\mathbf{z}_k - 2\mathbf{x}_i^T\mathbf{z}_k) \\ &= \left(2D - 2D \times \frac{\mathbf{x}_i^T\mathbf{z}_k}{D}\right) = 2D(1 - r_{ik}),\end{aligned} \qquad [3]$$

where D is the number of attributes of $\mathbf{x}_i$ and $\mathbf{z}_k$, and r_{ik} is the Pearson correlation coefficient for $\mathbf{x}_i$ and $\mathbf{z}_k$. Based on this relation, we have used the squared Euclidean distance scaled by $1/D$ as the distance between $\mathbf{x}_i$ and $\mathbf{z}_k$, and the error term in Equation 2 is equivalent to

$$e_{ij} = \frac{1}{2}\sum_{k=1}^{M} h_{jk} \frac{\left\|\mathbf{x}_i - \mathbf{z}_k\right\|^2}{D} = \sum_{k=1}^{M} h_{jk}(1 - r_{ik}) \qquad [4]$$

The cell lines have been clustered with varying number of clusters, that is 9, 16, and 25. The result with 16 clusters is shown in Figure 1(a). It can be seen that each cluster or nearby clusters appropriately reflect the organ of origin of its constituent, especially for the leukaemias (LE), the colon cancer lines (CO), the CNS lines, the renal carcinoma lines (RE), and the melanoma lines (ME).



CNS:SNB-19 CNS:U251 CNS:SF-295	RE:A498 RE:ACHN RE:TK-10 RE:RXF-393 RE:786-0 RE:UO-31 RE:CAKI-1	OV:OVCAR-3 OV:OVCAR-4 OV:IGROW1 OV:SK-OV-3	LC:NCI-H460 LC:A549/ATCC LC:EKVX
CNS:SF-268 CNS:SF-539 CNS:SNB-75 BR:BT-549 BR:HS578T LC:NCI-H226	ME:LOXIMVI PR:PC-3 LC:HOP-62	ME:M14 ME:MALME-3M ME:SK-MEL-28	BR:MDA-MB-435 BR:MDA-N
RE:SN12C LC:HOP-92 BR:MDA-MB-231 /ATCC	LC:NCI-H23 LC:NCI-H522 BR:MCF7 BR:T-47D	LE:SR LE:RPMI-8226 LE:K-562 LE:HL-60 LE:CCRF-CEM LE:MOLT-4	ME:SK-MEL-5 ME:UACC-62 ME:UACC-257 ME:SK-MEL-2
LC:NCI-H322M	CO:KM12 CO:HT29 CO:HCC-2998 CO:COLO205	CO:HCT-116 CO:SW-620 CO:HCT-15 OV:OVCAR-5	PR:DU-145 OV:OVCAR-8 BR:MCF7/ ADF-RES

(a)

BR:BT-549 RE:TK-10 RE:RXF-393 LC:HOP-92 BR:MDA-MB-231 /ATCC	LC:EKVX OV:OVCAR-5 OV:OVCAR-4 BR:T-47D	LC:NCI-H522	LE:SR LE:RPMI-8226 LE:K-562 LE:HL-60 LE:CCRF-CEM LE:MOLT-4
RE:ACHN RE:786-0 RE:UO-31 RE:CAKI-1	OV:OVCAR-8 BR:MCF7/ ADF-RES	CO:HCT-15 CO:HT29 CO:COLO205	ME:LOXIMVI BR:MCF7 CO:HCT-116 CO:SW-620
RE:A498 LC:NCI-H460 LC:A549/ATCC PR:DU-145 LC:NCI-H226	OV:SK-OV-3 LC:NCI-H322M	PR:PC-3 OV:OVCAR-3 OV:IGROV1 CO:KM12 CO:HCC-2998	BR:MCA-MB-435 BR:MDA-N
CNS:SNB-19 CNS:U251 CNS:SF-295 CNS:SF-268 CNS:SF-539 RE:SN12C LC:HOP-62	CNS:SNB-75 BR:HS578T	ME:SK-MEL-5 ME:UACC-62 LC:NCI-H23	ME:M14 ME:SK-MEL-2 ME:MALME-3M ME:SK-MEL-28 ME:UACC-257

(b)

CNS:SNB-19 CNS:SNB-75 RE:A498	CNS:SF-268 CNS:SF-539 ME:SK-MEL-28	CNS:U251 CNS:SF-295 OV:OVCAR-3	ME:SK-MEL-2 ME:UACC-257 OV:OVCAR-4 OV:IGROV1 OV:SK-OV-3
ME:SK-MEL-5 ME:UACC-62 ME:M14 ME:MALME-3M BR:MDA-MB-231 /ATCC LC:HOP-92	BR:MDA-MB-435 BR:MDA-N	LC:NCI-H322M CO:HCT-116 CO:HCT-15	BR:MCF7 BR:T-47D CO:SW-620
OV:OVCAR-8 OV:OVCAR-5 LC:NCI-H460	RE:CAKI-1 RE:ACHN BR:BT-549 LC:EKVX	RE:TK-10 RE:RXF-393 RE:786-0 RE:UO-31 LC:HOP-62	BR:HS578T LC:A549/ATCC ME:LOXIMVI PR:PC-3 RE:SN12C PR:DU-145
CO:HT29 CO:HCC-2998	LE:K-562 BR:MCF7/ ADF-RES CO:COLO205	LE:HL-60 LE:CCRF-CEM LE:MOLT-4 LC:NCI-H226 CO:KM12	LE:SR LE:RPMI-8226 LC:NCI-H23 LC:NCI-H522

(c)

Figure 1. The results of cell line clustering: (a) based on gene expression profiles ($\alpha = 0.0$), (b) based on interpolated distance ($\alpha = 0.7$), and (c) based on drug activity patterns ($\alpha = 1.0$). The value of h_{jk} is inversely proportional to the Euclidean distance between j^{th} and k^{th} clusters, where each cluster is represented as a discrete position in the two-dimensional lattice. In this 4×4 lattice, the cluster in the upper-left corner is encoded as (0, 0) and that in the lower-right corner as (3, 3). Clusters at the corners and ends are not neighbouring each other in view of Euclidean distance between the coordinates in the lattice.

We then ask, will the cell lines from the same tissue of origin show similar patterns in drug activities, such that they appear in the same or nearby clusters? To investigate if this is the case, we have clustered the cell lines based on both gene expression profiles and drug activity patterns. The error occurred by assigning a cell line to a particular cluster is defined as

$$e_{ij} = \frac{1}{2}\sum_{k=1}^{M} h_{jk}\left[(1-\alpha)\left\|\mathbf{x}_i^g - \mathbf{z}_k^g\right\|^2 + \alpha\left\|\mathbf{x}_i^d - \mathbf{z}_k^d\right\|^2\right], \qquad (0 \le \alpha \le 1) \qquad [5]$$

where the cell line $\mathbf{x}_i^g$ and the cluster $\mathbf{z}_k^g$ are related with gene expression profiles, and $\mathbf{x}_i^d$ and $\mathbf{z}_k^d$ with the drug activity patterns. The constant α is used to interpolate two distances based on the gene expression profiles and drug activity patterns.

Two criteria were used to measure the quality of the clustering results: the average Pearson correlation coefficient R and the average entropy H across all the clusters. They are defined as

$$R = \sum_{j=1}^{M} \frac{N_j}{N}\left[\frac{2}{N_j(N_j-1)}\sum\nolimits_{i<k} r_{ik}^j\right], \qquad [6]$$

$$H = \sum_{j=1}^{M} \frac{N_j}{N} \left[-\sum_{k=1}^{C} \frac{N_{jk}}{N_j} \log\left(\frac{N_{jk}}{N_j} \right) \right], \qquad [7]$$

where N is the number of cell lines, M is the number of clusters, and C is the number of tissues of origin. N_j represents the number of cell lines assigned to the j^{th} cluster, and N_{jk} the number of cell lines from the k^{th} organ of origin in the j^{th} cluster. The value in the bracket in Equation 6 is the average Pearson correlation coefficient across all the pairs of cell lines in the same cluster and that in Equation 7 represents the entropy in a cluster. When the cluster size is fixed, the higher value of entropy H means that the cluster structure is less reflective of the tissue of origin of the cell lines. In the case of the Pearson correlation coefficient R, the higher value means a better quality of clustering result in terms of inner cluster similarity.

Figure 2 shows the variation of the values of R and H in clustering of the cell lines, respectively, with varying α values in Equation 5. It can be seen that, with the higher value of α, the value of R based on gene expression profiles gets lower and the value based on drug activity patterns gets higher, showing the opposite trends between the two cases. In the case of the average entropy, as the value of α increases, the entropy has a tendency of being higher (for 16 clusters, from 0.40 to 0.72), and thus the quality of clustering becomes worse.

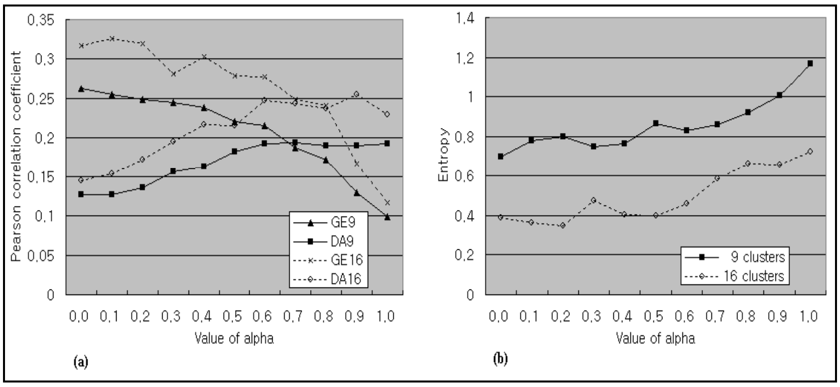


Figure 2. Values of the two measures of clustering quality over varying α. (a) Averaged Pearson correlation coefficients for 9 and 16 clusters. GE: The Pearson correlation coefficient based on gene expression profiles. DA: The coefficient based on drug activity patterns. (b) Averaged clustering entropies for 9 and 16 clusters. Only cancer types of the constituents in a cluster are considered, so just one graph suffices for each experiment.

From these two results, we can see that, in general, the similarity in gene expression profiles among a set of cell lines does not necessarily relate to a similarity in drug activity patterns among the cell lines. Also, the drug activity patterns are less related to the organ of origin, when compared with the gene expression profiles.

The cluster structure of the 60 cell lines on the basis of drug activity patterns only, that is $\alpha = 1.0$ in Equation 5, is shown in Figure 1(c). As also indicated by the value of average entropy, the cluster structure of the cell lines can be seen to be more heterogeneous than the result based on the gene expression profiles only. And Figure 1(b) shows a compromised solution with $\alpha = 0.7$. In [Scherf *et al.*, 2000], it has been proposed that this heterogeneity might be partly due to the activity of genes related to drug sensitivity and resistance, which has been supported by the fact that several cell lines with a relatively high expression level of multi-drug resistance gene *ABCB1* have been clustered in the same group. Inspired by our clustering results and the proposal, we have tried analysing the relationships among the activities of anticancer drugs and the expression levels of the genes by Bayesian network learning.

3. DEPENDENCY ANALYSIS USING BAYESIAN NETWORK LEARNING

3.1 Bayesian Networks

A Bayesian network [Heckerman, 1999] is a probabilistic graphical model that represents the joint probability distribution over a number of random variables. For an efficient representation, conditional independencies among the variables are exploited. These conditional independencies are encoded by a DAG (directed acyclic graph) structure in which a node corresponds to a random variable. The joint probability distribution over a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, given the Bayesian network for $\mathbf{X}$, is described as follows:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{Pa}(X_i)), \qquad [8]$$

where $\mathbf{Pa}(X_i)$ is the set of parents of node X_i in the Bayesian network structure. $P(X_i \mid \mathbf{Pa}(X_i))$ in the above equation is called the local probability distribution for X_i. Typically, the linear Gaussian model for continuous variables and the multinomial model for discrete variables are used for modeling the local probability distribution.

Learning Bayesian networks from data consists of two parts: learning the network structure and learning the local probability distribution for each node in the given structure. The second part corresponds to a simple calculation under some reasonable assumptions [Heckerman, 1999]. A popular approach to structural learning is the score-based search. The search space is nevertheless super-exponential in the number of variables. Because it is nearly impossible to find the best-scoring network structure even in a moderate case (7 or 8 variables), several search heuristics such as greedy search, greedy search with random restarts, and simulated annealing are used in practice [Heckerman, 1999]. In this paper, the greedy search algorithm and another search heuristic for hundreds of variables with the BD (Bayesian Dirichlet) scoring metric [Heckerman *et al.*, 1995] are used to learn Bayesian networks from the NCI60 dataset.

3.2 Applying Bayesian Networks to the Analysis of NCI60 Dataset

The NCI60 dataset contains gene expression patterns (T matrix) and drug activity patterns (A matrix) for 9 different cancer types [Scherf *et al.*, 2000]. To model the probabilistic relationships among them, we use a Bayesian network where each node corresponds to each variable. In the Bayesian network learning, the **T** matrix and **A** matrix are combined together, so that each cell line sample has gene expression levels and drug activities as its attributes.

3.2.1 Pre-Processing of the Dataset

The experiments focus on the 1,376 genes and 118 drugs as in the analysis of gene-drug correlations by Scherf *et al.* [2000]. Furthermore, genes and drugs that have more than 3 missing values across 60 samples, as well as unknown ESTs, were eliminated for robust analysis. Consequently, the analysed NCI60 dataset includes 60 samples with 890 attributes (805 gene expression levels, 84 drug activities, and one additional variable for the kind of cancer).

The number of attributes is extremely large compared to the number of samples. This might cause problems, such as a seriously slow learning speed, low confidence in learned models, and infeasibility of probabilistic inference. To cope with these problems, the number of attributes is reduced in two ways. One is to use prototypes of attributes. Genes and drugs are clustered respectively and the centre of each cluster is regarded as an attribute. The other is attribute selection. Here, all the genes and drugs are clustered together and all the members of some adjacent clusters are selected to construct the Bayesian network for the specific purpose of the analysis. The soft topographic vector quantization (STVQ) described in Section 2 is used for clustering.

All the continuous attribute values were discretized into three levels (low, normal, and high) for the multinomial local probability distribution model of the Bayesian network. The multinomial model is chosen because of its expressive power although discretization might cause some information loss. Two discretization boundary values for each attribute are calculated as $\mu + c\cdot\sigma$ and $\mu - c\cdot\sigma$. Here, μ is the mean value and σ is the standard deviation of the attribute across 60 samples. c is a constant which determines the distribution ratio of the original values in low, normal, and high.

3.2.2 A Fast Search Heuristic for Bayesian Network Learning

A general greedy search algorithm [Heckerman, 1999] is nearly inapplicable to learning Bayesian networks which consist of hundreds of nodes. Friedman *et al.* [1999] suggest a fast search heuristic for such cases and a similar approach is adopted in the experiments. The "local to global" heuristic is a kind of greedy search algorithm. Here, the search space is reduced by learning the structure around each node within small bounds before performing the greedy search procedure. The bounds are based on the concept of a Markov blanket [Pearl, 1988]. The Markov blanket of a variable satisfies the following.

$$P(X_i \mid \mathbf{X} - X_i) = P(X_i \mid \mathbf{BL}(X_i)), \qquad \mathbf{BL}(X_i) \subseteq \mathbf{X} - X_i, \qquad [9]$$

where **X** is the set of all the variables and $\mathbf{BL}(X_i)$ is the Markov blanket of X_i. Because the Markov blanket size of each node is unknown, the maximum size is pre-specified. Although the "local to global" heuristic is not guaranteed to find a good-scoring network in all cases, the learning speed is much faster than a general greedy search algorithm in the case of learning Bayesian networks with hundreds of nodes.

3.3 Experimental Results

Experimental results on the original dataset (Dataset 1), one reduced dataset with prototypes (Dataset 2), and another reduced dataset with selected attributes (Dataset 3) are given here. Table 1 shows the properties of these three datasets with applied learning methods, learning time, and the applicability of probabilistic inference. This table describes the properties of three datasets with respect to the applied learning methods, learning time, and the applicability of probabilistic inference. Samples in Dataset 2 have gene prototypes and drug prototypes as attributes. Dataset 1 is too large to apply the general greedy search algorithm. Dataset 3 is so small that the "local to global" heuristics are not required. Microsoft MSBN software - http://research.microsoft.com/research/dtg/msbn/OldMSBN.htm - was used for probabilistic inference in the analysis. The average learning time is measured on a Pentium III 1GHz machine.

Table 1. The properties of three datasets with respect to the applied learning methods, learning time, and the applicability of probabilistic inference. The numbers in the parentheses of the forth column represent the number of runs of the greedy search algorithm with random initialisations. The numbers in the parentheses of the fifth column represent the used maximum Markov blanket sizes. The rightmost column shows the applicability of probabilistic inference to the Bayesian networks learned from each dataset.

	# of genes	**# of drugs**	**Greedy search**	**"Local to global" heuristics**	**Learning time in avg. (secs)**	**Prob. inference**
Dataset 1	805	84	"—"	O (5 ~ 8)	3233.7	no
Dataset 2	40	5	O (20)	O (5 ~ 15)	123.9	yes
Dataset 3	12	4	O (100)	"—"	15.6	yes

3.3.1 Experimental Results on the Original Dataset

Three Bayesian networks were learned from the original dataset according to three different discretization boundaries (c = 0.43, 0.50, and 0.60). Probabilistic inference from the Bayesian network with 890 nodes is nearly impossible. Hence, only the number of edges connected to each node is analysed here. An edge represents direct probabilistic dependency and the node with many edges is considered to be related to many other nodes. Table 2 lists the top ten nodes that are most related to others on average in three Bayesian networks. The most related one is the cancer type node. The other nine nodes are all for genes. The results seem to be reasonable since the strong relationship between gene expression patterns and the kind of cancer is discovered from the cluster analysis in Section 2.

To investigate the influence of different discretization boundaries on the analysis, the Pearson correlation coefficient (r_{ij}) among the numbers of edges of all the nodes in two Bayesian networks was calculated as follows:

$$r_{ij} = \frac{\sum_{k=1}^{890} n_{ki}n_{kj} - \frac{1}{890}\sum_{k=1}^{890} n_{ki}\sum_{k=1}^{890} n_{kj}}{\sqrt{\sum_{k=1}^{890} n_{ki}^2 - \frac{1}{890}\left(\sum_{k=1}^{890} n_{ki}\right)^2} \cdot \sqrt{\sum_{k=1}^{890} n_{kj}^2 - \frac{1}{890}\left(\sum_{k=1}^{890} n_{kj}\right)^2}}, \qquad [10]$$

where n_{ki} is the number of edges of node k in Bayesian network i and n_{kj} is the number of edges of the same node in Bayesian network j. The average value of r_{ij} among three Bayesian networks is 0.841. The number of edges of each node does not seem to be so much influenced by different discretization boundary values.

Table 2. Top ten nodes that are closely related to the others. The first is cancer type and the other nine nodes are all for genes. The average number of edges of each node over all 890 nodes is 5.21.

Description of node	**The average number of edges**
The kind of cancer	125
SID W 487878, SPARC/osteonectin [5':AA046533, 3':AA045463]	25
Homo sapiens Cyr61 mRNA, complete cds Chr.1 [486700, (DIW), 5':AA044451, 3':AA044574]	18.3
SID W 162479, Homo sapiens epithelial-specific transcription factor ESE-1b (ESE-1) mRNA, complete cds [5':H27938, 3':H27939]	16
CDH2 Cadherin 2, N-cadherin (neuronal) Chr. [325182, (DIRW), 5':W48793, 3':W49619]	13.7
H.sapiens mitogen inducible gene mig-2, complete CDS Chr.14 [488643, (IW), 5':AA045936, 3':AA045821]	13.3
SID W 429623, Homo sapiens clone 24659 mRNA sequence [5':AA011634, 3':AA011635]	13.3
SID W 290871, Integrin alpha-3 subunit [5':N99380, 3':N71998]	13
COL4A1 Collagen, type IV, alpha 1 Chr.13 [145292, (EW), 5':R78225, 3':R78226]	12.7
COL4A1 Collagen, type IV, alpha 1 Chr.13 [489467, (IEW), 5':AA054624, 3':AA054564]	12.7

3.3.2 Experimental Results on the Reduced Dataset with Prototypes

In the Bayesian network learned from the reduced dataset with 40 gene prototypes and 5 drug prototypes, the negative correlation between *ASNS* (Asparagine synthetase Chr.7 [510206, (IW), 5':AA053213, 3':AA053461]) and L-asparaginase, as well as the negative correlation between *DPYD* (SID W 278125, Dihydropyrimidine dehydrogenase [5':N94809, 3':N63511]) and 5FU (fluorouracil) are examined [Scherf *et al.*, 2000]. Figure 3 shows two parts of the Bayesian network. In Figure 3(a), *G*4 is the gene prototype which includes *ASNS* and *D*2 is the drug prototype which includes L-asparaginase. *G*4 and *D*2 are dependent on each other directly. This suggests that these two nodes are strongly correlated with each other. In Figure 3(b), *G*8 is the gene prototype that includes *DPYD* and *D*5 is the drug prototype that includes 5FU. *G*8 and *D*5 do not directly depend on each other.

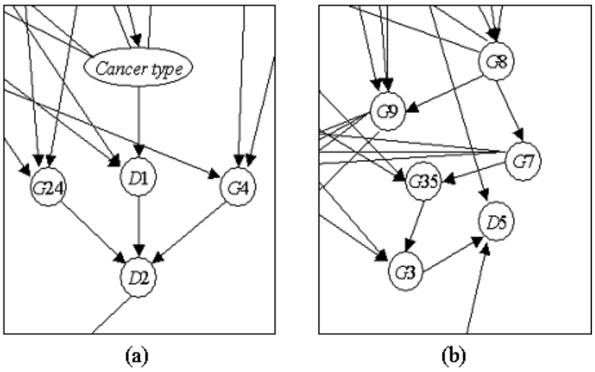


Figure 3. Two parts of the Bayesian network with 46 nodes. *G*1 ~ *G*40 are gene prototypes. *D*1 ~ *D*5 correspond to drug prototypes. In (a), *D*2 directly depends on *G*4 and vice versa. *D*5 is not directly dependent on *G*8 in (b).

Table 3 presents the results of the probabilistic inference from the Bayesian network. The inferred conditional probabilities do not show the expected negative correlation between *D*2 and *G*4 clearly. For example, *P*(*D*2 = low | *G*4 = high) should be greater than *P*(*D*2 = high | *G*4 = high). As a consequence, the Bayesian network with 46 nodes has failed to reveal some biologically known facts clearly. It might be due to the information loss induced from discretization, the use of prototypes, or both of these.

Table 3. The conditional probability table for *P*(*D*2 | *G*4) inferred from the Bayesian network in Figure 3. The negative correlation is not apparent here.

	D2 = low	**D2 = normal**	**D2 = high**
G4 = low	0.32096	0.27086	0.40818
G4 = normal	0.31387	0.41247	0.27366
G4 = high	0.32167	0.34920	0.32913

3.3.3 Experimental Results on the Reduced Dataset with Selected Attributes

To investigate the probabilistic relationships around L-asparaginase, 12 genes and 4 drugs were selected through clustering. Figure 4 shows the part of the Bayesian network with 17 nodes. In this figure, the direct probabilistic dependency is observed between the cancer type and L-asparaginase. L-asparaginase and *ASNS* are also dependent on each other directly. In addition, *ASNS* directly depends on *P5CR* (SID W 484773, PYRROLINE-5-CARBOXYLATE REDUCTASE [5':AA037688, 3':AA037689]). Tables 4 and 5 show the results of some probabilistic inferences from the Bayesian network. The conditional probabilities in Table 4 coincide with the negative correlation between *ASNS* and L-asparaginase. Moreover, when the cancer type is known to be leukaemia, the negative correlation is stronger.

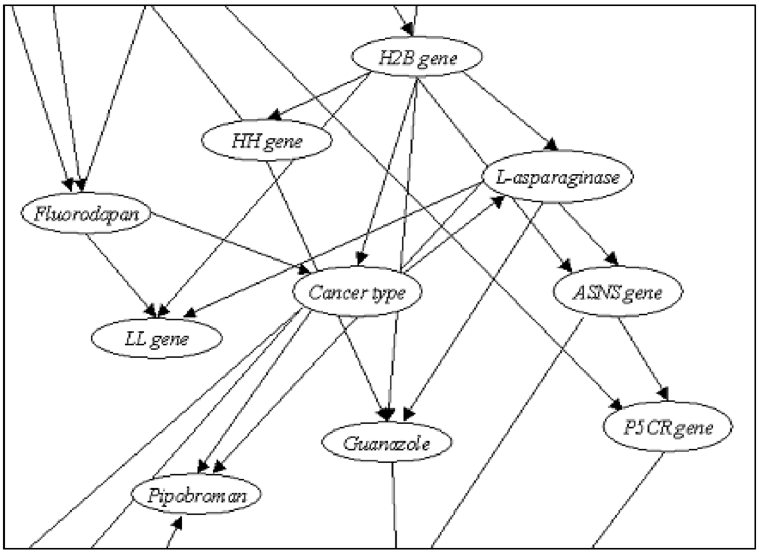


Figure 4. The Bayesian network with 17 nodes. Gene nodes are represented by acronyms. Following is the list of full names of the acronyms P5CR, ASNS, H2B, HH, and LL: SID W 484773, PYRROLINE-5-CARBOXYLATE REDUCTASE [5':AA037688, 3':AA037689] (P5CR), ASNS Asparagine synthetase Chr.7 [510206, (IW), 5':AA053213, 3':AA053461] (ASNS), SID 470936, Homo sapiens mRNA for for histone H2B, clone pjG4-5-14 [5':AA034106, 3':AA032092] (H2B), SID W 376009, HISTONE H1D [5':AA040305, 3':AA040326] (HH), SID W 430196, LACTOYLGLUTATHIONE LYASE [5':AA010331, 3':AA010332] (LL).

In addition, *P5CR* and L-asparaginase are highly negative-correlated in Table 5. *P5CR* is involved in the alanine and aspartate metabolism. *ASNS* is involved in the arginine and proline metabolism. These two metabolisms are closely located in the metabolic and regulatory pathway in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) located on the web at (http://www.genome.ad.jp/kegg). And the similarity of *P5CR* and *ASNS* in relation to the negative correlation with L-asparaginase seem to indicate a meaningful relationship.

Table 4. The conditional probability table for *P*(*L-asparaginase* | *ASNS*) and *P*(*L-asparaginase* | *ASNS*, *Cancer type* = *Leukaemia*) (the values in the parentheses). The quantified probabilistic dependency between the expression level of *ASNS* and the activity of L-asparaginase coincides with the known biological fact (the negative correlation).

	L-asparaginase = low	**L-asparaginase = normal**	**L-asparaginase = high**
ASNS = low	0.19857 (0.17536)	0.27471 (0.22838)	0.52672 (0.59626)
ASNS = normal	0.31110 (0.27128)	0.49795 (0.53790)	0.19095 (0.19081)
ASNS = high	0.42159 (0.38500)	0.36279 (0.42437)	0.21561 (0.19063)

Table 5. The conditional probability table for *P*(*L-asparaginase* | *P5CR*). The quantified probabilistic dependency between the expression level of *P5CR* and the activity of L-asparaginase is similar to that between *ASNS* and L-asparaginase.

	L-asparaginase = low	**L-asparaginase = normal**	**L-asparaginase = high**
P5CR = low	0.27510	0.35226	0.37263
P5CR = normal	0.31621	0.41072	0.27307
P5CR = high	0.33837	0.39664	0.26499

4. CONCLUSION AND FUTURE WORK

In this paper, the NCI60 dataset was analysed for the molecular pharmacology of cancer. First, the 60 cell lines were clustered using the STVQ algorithm. While the hierarchical clustering algorithm used in [Scherf *et al.*, 2000] operates in an agglomerative way and provides the tree-like cluster structure, the STVQ algorithm, starting from a coarse global structure, successively refines the cluster structure with some annealing schedule. And it finally represents the cluster structure in a two- or three-dimensional lattice.

We have performed cluster analyses based on the gene expression pattern and the drug activity pattern, respectively. The differences of the cluster structures were shown quantitatively in terms of the averaged Pearson correlation coefficient and the clustering entropy. The drug activity pattern less reflects the tissue of origin than the gene expression pattern, and it is suggested that this might be partly due to the expression of particular genes related to some drug activities. From these results, the drug activity pattern is analysed with gene expression patterns and cancer types for more detailed information, and Bayesian network learning was applied for this purpose.

In the experiments, a fast search heuristic was applied to learning the Bayesian network with hundreds of nodes. Among hundreds of attributes, only a few of them, including the cancer type and some genes, show notable relations to others. In order to perform the probabilistic inference, we reduced the dimensionality of attributes by clustering. By using prototypes, the known biological facts could not be discovered clearly. This might be due to the loss of useful information in the original data by the use of gene prototypes and drug prototypes. Hence, the dimensionality reduction by attribute selection was performed. Focusing on the discovery of relationships around L-asparaginase, we selected 12 genes and 4 drugs by clustering. The results of the analysis coincide with the known biological facts: the negative correlation between L-asparaginase and *ASNS*, as well as the influence of the kind of cancer on this negative correlation. In addition, the positive correlation between *ASNS* and *P5CR* was discovered. Biologically, *ASNS* and *P5CR* are located closely in the metabolic pathway. To summarize, the relationships among genes, drugs, and cancer types could be modelled by Bayesian network learning. This suggests that Bayesian network learning and clustering are appropriate for the exploratory analysis of high-throughput genomic data.

Directions for further research are as follows: In a complex domain such as DNA microarray analysis, the learned results are prone to be unreliable because of the small sample size compared with the number of attributes. The eMCMC (evolutionary Markov chain Monte Carlo) method [Zhang *et al.*, 2001] might be an appropriate solution. The more efficient and robust learning and inference algorithms for large Bayesian networks should also be studied. In addition, combining knowledge from biomedical literature with data analysis is a candidate for the improvement of the quality of results.

ACKNOWLEDGEMENTS

The authors would like to thank Sirk June Augh for thorough discussions on the experimental results. The authors also thank anonymous reviewers and editors for helpful comments that greatly improve the paper. This work was supported in part by BK21-IT, IMT-2000, BrainTech, and AITrc programs.

REFERENCES

Dempster, AP, Laird, NM, Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39 (1977): 1-38.

Eisen, MB, Spellman, PT, Brown, PO, Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* 95(25) (1998): 14863-14868.

Friedman, N, Nachman, I, Pe'er, D. Learning Bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI 1999)* (1999): 206-215.

Friedman, N, Linial, M, Nachman, I, Pe'er, D. Using Bayesian networks to analyze expression data. In *Proceedings of the 4th Annual International Conference on Computational Molecular Biology* (*RECOMB 2000*) (2000): 127-135.

Graepel, T. Self-organizing maps: Generalizations and new optimisation techniques, *Neurocomputing* 21 (1998): 173-190.

Hartemink, AJ, Gifford, DK, Jaakkola, TS, Young, RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing* 6 (2001): 422-433.

Heckerman, D. *A tutorial on learning with Bayesian networks*. Edited by MI Jordan. Learning in Graphical Models. MIT Press, 1999.

Heckerman, D, Geiger, D, Chickering, DM. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20(3) (1995): 197-243.

Hwang, K-B, Cho, D-Y, Park, S-W, Kim, S-D, Zhang, B-T. Applying machine learning techniques to analysis of gene expression data: cancer diagnosis. In: Lin, SM, Johnson, KF, *Methods of Microarray Data Analysis*, Norwell, MA, Kluwer Academic Publishers, (2001):167-182.

Khan, J *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 7(6) (2001): 673-679.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Edited by M Kaufmann. (1988).

Raychaudhuri, S, Stuart, JM, Altman, RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing* 5 (2000): 452-463.

Scherf, U *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* 24 (2000): 236-244.

Zhang, B-T, Cho, D-Y. System identification using evolutionary Markov chain Monte Carlo. *Journal of Systems Architecture* 47(7) (2001): 587-599.