Detection of network structure changes by graphical chain modeling: A case study of hepatitis C virus-related hepatocellular carcinoma

Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

Detection of network structure changes by graphical chain modeling: a case study of hepatitis C virus-related hepatocellular carcinoma

Shigeru Saito, Masao Honda, Shu-ichi Kaneko and Katsuhisa Horimoto

Abstract—**One of the most characteristic features of biological molecular networks is that the network structure itself changes, depending on the cellular environment. Indeed, activated molecules show a variety of responses to distinctive cell conditions, and subsequently the network structures of active molecules also change. Here we present an approach to trace the network structure changes by using the graphical chain model developed from the gene expression data. The previous procedure for applying the graphical chain model to the expression profiles of a limited number of genes has been improved to analyze the entire set of genes. Furthermore, the chain model has been rearranged according to the association strength, and was scrutinized to identify the candidates of essential gene-gene relationships for the network changes, by using the path consistency algorithm. The improved procedure was applied to the expression profiles of 8,427 genes, which were measured in two distinctive stages of liver cancer progression. As a result, the chain model of the 18 gene cluster relationships with strong associations was inferred, in which the coordination of clusters was described in the cell stage progression, and the gene-gene relationships between known cancer-related genes causing the progression were further refined. Thus, the present procedure is a useful method to model the network structure changes in the cell stage progression, and to clarify the gene candidates for the progression.**

I. Introduction

One of the remarkable relationships between molecules in living organisms is the drastic changes of network structures in response to the environment. For example, it is well known that a specific set of molecules, among all of the molecules in a cell, is activated, in response to environmental stress [1, 2]. Unfortunately, the experimental techniques for monitoring the activated molecules in a living cell still require further development. Thus, it is desirable to be able to infer the network structure of activated molecules from data measured under distinctive conditions.

Manuscript received January 28, 2009. This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (grant 20016028) and for Scientific Research (A) (grant 19201039) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

S. S. is with National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan, and also with INFOCOM Corporation, Tokyo 150-0001, Japan (e-mail: sh.saito@ infocom.co.jp).

M. H. is with Kanazawa University Graduate School of Medical Science, Kanazawa 920-8641, Japan (e-mail: mhonda@m-kanazawa.jp).

S. K. is with the Kanazawa University Graduate School of Medical Science, Kanazawa 920-8641, Japan (e-mail: skaneko@m-kanazawa.jp).

K. H. is with National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan, and also with Shanghai University, Shanghai 200444,, China (phone: +81-3-3599-8711; fax: +81-3-3599-8081; e-mail: k.horimoto@ aist.go.jp).

In human liver infected by hepatitis C virus (HCV), the infection induces the development of chronic hepatitis (CH), cirrhosis, and in some instances, hepatocellular carcinoma (HCC) [3]. The virological features of the infection were described [4, 5], and we previously reported that the gene expression profiles in chronic hepatitis C (CH-C) predominantly induced inflammatory and anti-apoptotic phenotypes. However, the network structure changes inducing these modifications in gene expression still remain to be elucidated.

Recently, we have developed a procedure [6] for tracing the network structure changes from the gene expression data by using the graphical chain model (GCM) [7-10]. In the application of GCM to the expression data measured in progressive cell stages, the block and the variables in blocks correspond the cell stage and the genes characteristically expressed in each stage, respectively. Since GCM exhibits the overlaps of the variables between the blocks, the genes responsible for distinctive stages should be selected among the set of entire genes. Indeed, in our previous application of GCM to the yeast cell-cycle [6], we adopted the gene sets of about 700 genes that characterized each cell stage, from a previous study [11]. In general, the genes that are characteristically expressed in distinctive stages are identified most effectively by discriminating between the stages. In this case, however, the genes that are continuously up-regulated or down-regulated over the stages are not selected. In the case of progressive processes, the continuously up (or down)-regulated genes may be important for identifying the molecular mechanisms underlying the stage progression.

Here, we have improved the previous procedure [6] to detect the changes of network structures more efficiently, by using the entire set of genes. The procedure was applied to the expression profiles measured in two stages of hepatocellular carcinoma, from CH to HCC [4, 5]. For each stage, all of the genes in the analyzed data were systematically classified into three groups, up-, down-, and unchanged regulated gene groups, and based on the classification, the three blocks in GCM-CH, HCC, and background-were defined. In particular, the background includes the genes that are continuously up- and down-regulated, and the influence of these genes on the progression from CH to HCC was estimated. Thus, the improved procedure allowed us to describe the network structure changes of entire sets of genes. Furthermore, the transformation of the inferred network structure helped to reveal the candidates of the gene-gene relationships causing

the cancer progression, including known cancer-related genes.

II. MATERIALS AND METHODS

A. Expression Profile Data

The expression profiles of 8,427 genes were monitored in 6 normal, 32 CH and 17 HCC samples [4]. Relative expression ratios of 8,427 genes were obtained by comparing hybridization of Cy5-labeled cDNAs from chronic hepatitis lesions and Cy3-labeled cDNA from normal liver tissue.

B. Graphical Chain Modeling (GCM)

The graphical chain model (GCM) is a probability model for multivariate random observations, in which the independence of the structure can be represented by a graph [7-10]. Here, we will briefly describe GCM.

The graph $G = (V, E)$ consists of a set of vertices V, representing the variables, and a set of edges E, representing the associations between pairs of variables. E is a set of ordered pairs (A, B), $A, B \in V$. The chain graph is based on the partitioning of V into disjointed subsets: $V = V_1 \cup V_2 \cup \ldots \cup V_T$. The subsets are called blocks or chain components. The edges within blocks are undirected, reflecting the systematic associations, and the edges between blocks are arrows pointing from blocks with lower index numbers to those with higher indices. A graphical chain model displays the independence between variables conditioned on all of the other variables in the current and previous blocks. In a graphical chain model, any direct association between two variables in the same block is assumed to be non-causal, and is represented by an undirected edge (line) in a graph. Any direct association between two variables from different blocks is assumed to be potentially causal, and is represented by a directed edge (arrow). The absence of a line or arrow between two variables in the graph indicates that there is no direct association between the variables, i.e. the variables are independent, after controlling for all of the other variables in the same and previous blocks.

The graphical chain model is fitted in a number of stages. When fitting a graphical chain model, the first step is to partition the variables into a number of ordered blocks. Then, the significant direct associations between the variables in the first block are determined. For each pair of variables, the null hypothesis when tested shows that the variables are independent, given all of the other variables in the first block, and the deviance statistics in graphical Gaussian modeling (GGM) is used [12].

Next, the significant direct associations between the variables in the second block and between the first and second blocks are determined. For each pair of variables, the null hypothesis when tested shows that the variables are independent, given all of the other variables in the first and second blocks, and again the deviance statistics is used.

The fitting continues, block by block, by determining all of the significant direct associations between the variables in the current block and between all of the variables in the current and previous blocks. The null hypothesis is now independence, given the other variables in the current and previous blocks, and again the deviance statistics in GGM is used. In other words, the procedure of GCM is the iteration of GGM. All of these tests were carried out at the 5% level, using the χ^2 distribution in deviance statistics.

In the present study, the block in the graphical chain model simply corresponds to the cell stages that are defined by biological information. By the intact correspondence to graphical chain modeling, the variable is the gene that has an expression profile with numerical values. However, since the expression profiles often show similar patterns, the genes are highly related to one another. Thus, hierarchical clustering is performed for the genes within each block, as a preprocessing step for the graphical chain modeling [7-10], and then, each gene cluster corresponds with the variable in the present procedure.

C. Improved Procedure for Applying GCM to Gene Expression Profiles

In the present study, we improved the procedure for applying GCM to gene expression measured in two cell stages. The overview of the present procedure is described in Fig. 1.

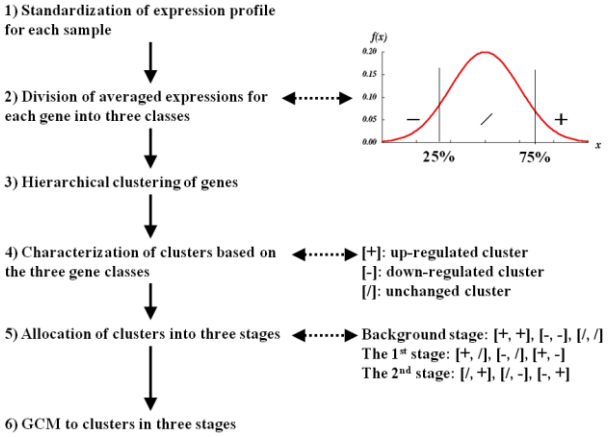

Fig. 1. Improved procedure for applying GCM to gene expression profiles in two stages.

1) Standardization

The expression profiles of all genes are standardized by the average and the standard deviation for each sample, i.e.,

$$z_{ij} = \frac{x_{ij} - AV_j}{SD_j},$$

where z_{ij} and x_{ij} are the standardized and intact expression values of the i-th gene and the j-th sample, respectively, and AV_j and SD_j are the average and the standard deviation of all genes over the j-th sample, respectively. This allows the noise of the expression profiles of each gene, due to the differences between samples, to be excluded by the transformation of the

intact expression degree into a standardized value, the z-value.

2) Division of Average Expression Values into Three Classes

The average expression values for each gene are calculated over the samples in one stage. Then, the genes are divided into three classes in terms of the average value: up-regulated gene class (‘+’, abbreviation of class), down-regulated gene class (‘-‘), and the other class (‘/’). If the average value of a gene ranges more than the 25 th percentile, less than the 75 the percentile, and between the 25 th and 75 th percentiles, then the corresponding gene is regarded as up-regulated, down-regulated, and the other class, respectively.

3) Hierarchical Clustering

All genes in the analyzed data were subjected to hierarchical clustering. In the present clustering, the metric is Pearson’s correlation coefficient of genes between the expressions of samples, and the technique is the Un-weighted Pair Group Method using the Arithmetic average (UPGMA). The number of clusters was estimated by using the variance inflation factor, defined in the previous study [13].

4) Cluster Characterization in Terms of Expression Class

Based on the gene classification into the three classes mentioned in 2), the clusters are characterized: each stage is characterized by the maximum number of gene classes. For example, if 50, 10, and 5 genes belonging one cluster in one stage are ‘+’, ‘/’, and ‘-‘, respectively, then the cluster in the stage is characterized by [+]. According to the above rule, the class pairs in the two stages were defined: the nine class pairs in the two stages are [+, +], [+, /], [+, -], [/, +], [/, /], [/, -], [-, +], [-, /], and [-, -].

5) Allocation of Three Expression Classes into Clusters

The clusters with the pairs of three expression classes for the two stages described in 4) were allocated to three groups. The rules for allocation are simple. First, if a cluster shows an up-regulated class at only one stage, then the cluster is allocated into a group that represents the corresponding stage. Second, among the remaining clusters, if a cluster also shows a down-regulated class at only one stage, then the cluster is allocated into a group that represents the corresponding stage. Finally, the remaining clusters are allocated into a hypothetical group, named the background stage. This is because the up-regulation of the gene indicates that the corresponding gene product increases, and plays a important role in the stage. Thus, for example, the clusters allocated into the first cell stage are composed as follows: according to the first rule, the up-regulated cluster class in the first cell stage and the other cluster class in the second cell stage [+, /] and the up-regulated cluster class in the first cell stage and the down-regulated cluster class in the second cell stage [+, -] are allocated. The down-regulated cluster class in the first cell stage and the other cluster class in the second cell stage [-, /] are then allocated, according to the second rule. The clusters are allocated into the second cell stage according to the same rule: [/, +], [-, +], and [/, -]. Finally, the clusters allocated into the background stage are the remaining clusters, i.e., [+, +], [-, -], and [/, /]. Here, the background stage is a hypothetical stage for considering the influence of the expression of uncharacterized genes on that of well-characterized genes, depending on the two cell stages. In other words, the variables in the hypothetical stage are viewed as purely explanatory variables, whereas the variables in the second and subsequent blocks are viewed as responses to the variables in the preceding blocks. Note that the usual approach, based on the gene selection by discrimination between the two stages, does not consider the effect of the genes showing a constant degree of gene expression.

6) GCM

Finally, we perform GCM for the three groups of clusters that were allocated in 5). In this case, the order of the above groups is set in the order of the background stage, the first stage, and the second stage. It seems natural that the background stage influences both the first and second stages. Thus, we can estimate the causal relationship between the clusters in the first stage and the background and between the clusters in the second stage and the background, as well as between the clusters in the first and second stages.

D. Securitization of Chain Model

The chain model obtained by the standard algorithm of GCM frequently has many edges, and thereby adopts a messy form with many nodes and edges. Even in a sparse form, each node obtained by cluster analysis for the entire set of genes also frequently contains many genes. To scrutinize the genes responsible for the network structure change in the chain model, therefore, we further devised two techniques: one for the former issue is the evaluation of the association strength of each edge in the model, and the other for the latter issue is the inference of the gene-gene association in the selected clusters. The details of the two techniques are described below.

1) Evaluation of Association Strength in Chain Model

When there are many edges, drawing them all on one graph produces a mess or `spaghetti' pattern, which would be difficult to read. Indeed, in some examples of the application of GGM to actual profiles, the intact networks derived by GGM still showed complicated forms with many edges [14]. The similar situation may be expected in GCM, which is the iteration of GGM. Thus, the strength of the association between clusters is evaluated in a statistical way: the intact network can be rearranged according to the partial correlation coefficient value, to interpret the associations between clusters. The strength of the association can be assigned by a standard test for the partial correlation coefficient, $r_{ij,rest}$, between the variables i and j, given the resting variables [15]. By Fisher’s Z transformation of partial correlation coefficients, i.e.,

$$Z = \frac{1}{2}\log\left(\frac{1+r_{ij\cdot rest}}{1-r_{ij\cdot rest}}\right),$$

Z distributes according to the following normal distribution:

$$N\left(\frac{1}{2}\log\left(\frac{1+r_{ij\cdot rest}}{1-r_{ij\cdot rest}}\right),\frac{1}{\{N_c-(M-2)\}-3}\right),$$

where N_c and M are the number of conditions and the number of clusters, respectively. Thus, we can statistically test the observed correlation coefficients under the null hypothesis with a significance probability.

2) Application of Path Consistency Algorithm

The application of GCM to the entire data set with the combination of cluster analysis produces a macroscopic view of the causal relationships between them. This is partly because the entire data set contains a similar pattern of gene profiles, and partly because the noise in the data due to various effects prevents us from inferring the entire relationship at a fine level, such as the causal relationship between the genes. However, if partial sets of genes are selected by GCM, then the selected sets of genes may be appropriate for applying a fine analysis to infer a causal relationship between the genes. Thus, the relationships between the constituent genes in the clusters whose relationships are inferred by GCM are also inferred by another graphical modeling method, the path consistency (PC) algorithm [16]. The PC algorithm is composed of two parts: the undirected graph inference by the partial correlation coefficient and the following directed graph, obtained by using the orientation rule based on the inferred undirected graph. A brief overview of the PC algorithm and the modifications for the present analysis is described below.

The first part of the algorithm is simple. The relationship between two variables is tested from the lower partial correlation coefficient to the higher one. For example, the relationship between the two variables is first tested by the zero-th partial correlation coefficient. If the null hypothesis is accepted, i.e., no association between the two variables, then, the relationship between the two variables is tested by the first partial correlation coefficient. If it is rejected, then the test is not performed. In general, the (m-2)-th order of the correlation coefficient is calculated between two variables, given (m-2) variables, i.e., $r_{ij,rest}$, between X_i and X_j, given the '*rest*' variables, $\{X_k\}$ for k=1, 2, …, m, and $k\neq i$, j, and after calculating the (m-2)-th order of correlation coefficient, the algorithm naturally stops. However, the algorithm does not usually request the (m-2)-th order of correlation coefficient for the natural stop. This is because no adjacent variables will be found after excluding the variables, even in the calculation of the lower order of the correlation coefficient. In the practical analysis of sample data, the zeroth-order of the correlation coefficient is calculated by Pearson's correlation coefficient, r_{ij}, expressed by

$$r_{ij}=\frac{\mathrm{cov}(X_i,X_j)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)}},$$

where cov (X_i, X_j) and var(X_i) are the covariance between X_i and X_j, and the variance of X_i, respectively. The higher order of the correlation coefficients is the partial correlation coefficient, $r_{ij,rest}$, expressed by

$$r_{ij,rest}=\frac{-r^{ij}}{\sqrt{r^{ii}\cdot r^{jj}}},$$

where (*ij*, *rest*) means $\{1,2,\ldots,p\}\backslash\{i,j\}$, and r^{ij} is the *i*-*j* element of the inverse correlation coefficient matrix. Note that the dimension of the correlation coefficient matrix corresponds to the orders of the correlation coefficients. The n th-order correlation coefficient is calculated from the (n+2) dimension of the correlation coefficient matrix. The correlation coefficient is statistically tested by using the Z-statistic [15], as in the case of the association strength evaluation in i).

Based on the inferred undirected graph, C, the direction of each graph is decided in the second part, according to the orientation rule [17]. The rules for the direction decision in C are as follows:

i) If there is an undirected relationship, X−Y−Z, and X and Z are not adjacent, then the direction is decided as being X→Y←Z.

ii) If there is a relationship, X→Y−Z, and X and Z are not adjacent, then the direction is decided as being X→Y→Z.

iii) If there is a directed path between X and Y, and there is a relationship, X−Y, then the direction is decided as being X→Y.

The key point in the present network inference is a modification of the original PC algorithm, for application to the expression profiles. The modification corrects the algorithm in the calculation of the partial correlation coefficient. Since many genes frequently show very similar patterns of expressions, the difficulty arises in the numerical calculation of correlation coefficients, due to the multi-colinearity between the variables. The original PC algorithm accidentally stops, if only one correlation between a pair of variables shows a violation of the numerical calculation, against the high similarity of the expressions. To escape the accidental stop by the highly associated gene pairs, the original PC algorithm was modified as follows: if the calculation of any order of correlation coefficient between the variables is violated, then the corresponding pair of variables is regarded as being dependent.

E. Software

All calculations of the present clustering and GGM were performed by the ASIAN web site [18](http://eureka.cbrc.jp /asian) and "Auto Net Finder", the commercialized PC version of ASIAN, from INFOCOM CORPORATION (http://www.infocom.co.jp/bio/download/).

III. RESULTS AND DISCUSSION

A. Clustering and Its Allocation into the Three Stages

According to the procedure described in the preceding section, first, all genes characterized by the three degrees of expression were subjected to the cluster analysis, and then the

TABLE I
ALLOCATION OF CLUSTERS INTO THE THREE STAGES

Cluster Number	Number of members	Cluster class pair	Number of genes in corresponding class pair	Allocated stage
1	146	[+, /]	66	CH
2	476	[+, /]	271	CH
3	654	[/, /]	394	B
4	198	[/, +]	54	HCC
5	440	[−, −]	173	B
6	229	[/, −]	127	HCC
7	51	[+, +]	47	B
8	9	[+, +]	9	B
9	288	[−, /]	85	CH
10	400	[/, +]	137	HCC
11	318	[−, /]	77	CH
12	272	[/, +]	115	HCC
13	188	[−, /]	85	CH
14	277	[−, −]	191	B
15	179	[/, /]	70	B
16	63	[+, /]	14	CH
17	310	[/, /]	86	B
18	184	[+, /]	38	CH
19	47	[−, −]	44	B
20	110	[−, −]	102	B
21	160	[−, /]	28	CH
22	950	[/, /]	565	B
23	1011	[+, /]	287	CH
24	23	[−, −]	23	B
25	392	[+, +]	281	B
26	127	[−, −]	92	B
27	274	[/, +]	126	HCC
28	40	[−, +]	6	HCC
29	316	[−, −]	293	B
30	115	[/, −]	22	HCC

clusters were allocated into the three stages. The clusters and their allocations are listed in TABLE I

As seen in TABLE I, all of the genes were grouped into 30 clusters, with various numbers of members. Indeed, the minimum and maximum numbers of genes among the 30 clusters were 9 in cluster 8 and 1,011 in cluster 23, respectively, and the average number of genes was 274.9, with a standard deviation of 243.3 genes. Based on the rule in II-C-5), the clusters were allocated to the three stages, Background, CH, and HCC. The classes of clusters in the CH and HCC stages are listed in the third column in the table, and the numbers of genes belonging to the class pairs of the two stages are listed in the fourth column. The total numbers of genes belonging to the class pairs were 3,908 among the 8,247 genes analyzed in the present study: the average fraction of the numbers of genes in the fourth column to those in the second column was 0.513, with a standard deviation of 0.275. Thus, more than half of the genes were responsible for the allocation of the clusters into the three stages. Finally, the numbers of clusters allocated to Background, CH, and HCC were 14 (2,370 genes), 9 (951 genes), and 7 (587 genes), respectively. The present allocation of clusters, with its small bias, may reveal the relationships between the clusters in the progression from CH to HCC.

B. Chain Model

The partial correlation coefficient matrix generated by

TABLE II
PARTIAL CORRELATION COEFFICIENT MATRIX FOR THE THREE STAGES BY GCM

	Background														CH									HCC						
Cluster	3	5	7	8	14	15	17	19	20	22	24	25	26	29	1	2	9	11	13	16	18	21	23	4	6	10	12	27	28	30
3	-1																													
5	0.275	-1																												
7	0	0.255	-1																											
8	0	-0.16	0	-1																										
14	0.248	0.181	0	0	-1																									
15	0	-0.28	0	0	0	-1																								
17	0	0	0	0	0.178	0.298	-1																							
19	0	0	-0.223	0	0.617	0	-0.220	-1																						
20	0.147	-0.26	0.290	0	0	0	0	-0.249	-1																					
22	-0.248	0.222	-0.350	0	-0.26	0	0	-0.255	0	-1																				
24	-0.253	0	0	-0.174	0	0	0	0.574	0.681	0.332	-1																			
25	0.421	0.165	0	0.158	-0.65	0	0	0.299	0	0	0.132	-1																		
26	-0.184	0.188	-0.226	-0.376	-0.24	0	0.448	0.215	0.315	0	-0.37	0	-1																	
29	0	0.53	-0.512	0.232	0	0	-0.132	-0.198	0.644	-0.31	0	-0.25	0	-1																
1	0	0	0.286	-0.224	0	-0.201	0	-0.440	-0.257	-0.23	0	0	0	0.394	-1															
2	0	0	0	0.186	-0.33	0	0.233	0	0	0.292	0	0	0	-0.11	0	-1														
9	0	-0.44	-0.475	0	0	-0.260	0	0	-0.322	0	0.158	0	-0.41	0	0	0	-1													
11	-0.146	0	-0.321	0.270	0	-0.289	-0.395	0	0	0	0	0	0	-0.31	0	-0.398	-0.381	-1												
13	0.443	0	0	0.277	0	-0.196	0.207	0.275	-0.224	0	0	0	-0.33	0	-0.21	-0.356	-0.209	-0.19	-1											
16	0.281	-0.21	0	0	-0.24	0.226	0.201	0.284	0	0.241	0	0	0	0	0.174	-0.377	0	0	-0.27	-1										
18	-0.203	0	0	0	-0.3	0.305	0.468	0	0	0	0.244	0	0.417	-0.22	0	-0.402	0	0	0	-0.22	-1									
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1								
23	0.329	-0.21	-0.290	0	0	0	0	0	-0.289	0.264	0.206	0.261	-0.42	0	0	0	-0.625	-0.16	-0.34	-0.17	0	0	-1							
4	0	0.263	0	0.458	0	0	0	0.212	-0.530	-0.42	0.349	-0.41	0	-0.37	0	0	-0.291	0.242	0	0	-0.25	0	-0.258	-1						
6	0.221	-0.22	0.474	-0.451	0	0.144	0.341	0.222	0.416	0.336	-0.28	-0.17	-0.38	0.746	0.743	0	0	0	-0.32	0	0.190	0.228	0.335	0.447	-1					
10	-0.238	-0.2	0.164	0	0	-0.243	-0.522	-0.442	-0.153	-0.14	0	0	0	0	-0.41	0	0.858	0.157	0.247	-0.24	0	-0.58	0	0.169	0	-1				
12	-0.237	-0.48	0.337	0	-0.35	-0.415	-0.357	-0.183	-0.484	-0.52	0.238	-0.56	-0.31	0	-0.27	-0.191	0.157	0.676	0	-0.45	-0.32	-0.5	-0.342	-0.180	0	-0.47	-1			
27	-0.263	-0.41	0.295	-0.250	-0.23	-0.475	0	-0.125	-0.366	-0.48	0	-0.32	-0.46	-0.18	-0.23	-0.512	-0.215	0	0	-0.46	-0.49	-0.4	-0.477	0	0	-0.17	-0.42	-1		
28	-0.149	-0.16	-0.232	0	0	-0.231	-0.234	-0.113	0	-0.19	0	0	0	-0.11	0	-0.123	0	-0.25	0	0	0	-0.24	0	0	0	0	0	0	-1	
30	-0.280	-0.43	0.272	-0.357	-0.21	-0.437	-0.218	-0.180	-0.166	-0.49	0	-0.46	-0.32	0	-0.13	-0.230	0	0.242	0	-0.38	-0.3	-0.42	-0.240	0	-0.19	-0.31	-0.66	-0.52	-0.18	-1

GCM is shown in TABLE II. By GCM, 195 (44.8%) of the possible 435 edges in the full model of 30 clusters were deleted; thus, there were still 242 edges in the inferred chain model, and its form was too complicated to interpret the entire association between the clusters. Within and between the three stages, the numbers of deleted edges were relatively uniform, except for those between Background and HCC: only 26 (26.5%) of 98 edges (=14x7) were deleted. In other words, the influence of the clusters in the Background to those in HCC was inferred in the model.

C. Arrangement of Chain Model

According to the procedure for evaluating the association strength, the chain model corresponding to the partial correlation coefficient matrix was arranged. To interpret the model, the strong associations with the significant (p<0.01, r>0.510) edges were selected: in the rearranged model, 22 clusters showed the strong association, and 8 clusters were isolated (not depicted in the figure).

By the above arrangement, a simple chain model emerged in Fig. 2. As seen in the figure, the number of edges within the three stages was less than those between them. In particular, there was only one edge in CH, and HCC had two edges, whereas there were 6 edges in Background. In contrast, 10 edges emerged between the three stages: there were 5 edges from Background to HCC, and in contrast, there were no edges from Background to CH. Thus, the bias of the numbers of edges between the stages to those within each stage suggests that the genes were coordinately expressed in accordance with the cancer progression.

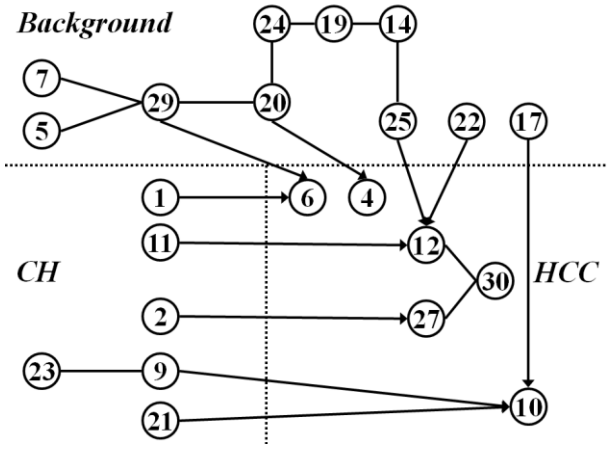

Fig. 2. Chain model for hepatitis C virus-related hepatocellular carcinoma. In the chain model, the edges showing strong associations with less than 1% significance probability are depicted, but the isolated edges are not. The numbers in the circles correspond to the cluster numbers in TABLE I. The clusters allocated to the background, CH, and HCC are located in the upper, left, and right parts of the figure, respectively, and are separated by dotted lines.

The following relationships are particularly remarkable: cluster 29 in Background, cluster 1 in CH, and cluster 6 in HCC ({29→6←1}), {(22, 25)→12←11}, and {17→10←(9, 21)}. In the above cluster sets, the clusters in Background and in CH were coordinately related to the cluster in HCC. Since conventional approaches consider only the differences in gene expression between the two stages, the above relationships between the three stages, Background, CH, and HCC, would frequently be neglected. Thus, these clusters have the possibility of containing the essential genes responsible for cancer progression, which were not detected by the conventional discrimination approaches.

D. Gene-Gene Relationships between Clusters

Here, we focus on the genes responsible for the relationships between the three cluster sets in the preceding section. However, since the clusters still contain many genes, further analysis is needed to refine the candidates of the genes responsible for the cluster-cluster relationships. For this purpose, the PC algorithm was performed for the genes belonging to the respective sets. Then, the gene-gene relationships consistent with the relationship between the clusters were selected. For example, in the first set, {29→6←1}, all of the genes belonging to clusters 29, 1 and 6 were attributed by the PC algorithm. If the inferred relationship of the genes belonging to the respective clusters is consistent with that between clusters 29, 1, and 6, then the gene sets were extracted from all of the gene-gene relationships inferred by the PC algorithm.

TABLE III
GENE-GENE RELATIONSHIPS IN COORDINATED CLUSTER-CLUSTER RELATIONSHIPS

Cluster Relationships	Gene Relationships
{29→6←1}	ARL6IP→CDH1←FGFR2, Hs.209450→BHMT2← SFRS11, SCP2→BHMT2←SFRS11, CYP4X1→NDRG1← ELYS, GPAM→SYP←KIAA0182, RRAS2→CLDN12← PXMP3, ARCN1→VDAC1←Hs.433078, SHANK2→ PDSS2←COX7B, MYO1B→KTN1←FKBP3, ARCN1→ KIAA0073←RNF146
{22→12←11}	NEO1→KRT14←FLJ12270, NEO1→KRT14←IGHMBP2, RANBP3→IL1RL1←SSRP1, Hs.79241→CD2←NFKB2, MGC4606→SLA/LP←EFNB2, PKD2L1→TK2←GNAI2, ADA→FLJ46603←HOXA9, GGA3→RAP1A←LSM14B, LMLN→RAP1A←LSM14B, ZNF638→GTPBP6← Hs.298289, DCT→GTPBP6←Hs.298289, Hs.191356→ PTMS←ATP2A3, Hs.378847→UBE2G2←TEAD3, LIM→ UBE2G2←TEAD3, TCF7L2,FLJ23556→ CTAG1B,CTAG1A←Hs.107410, SAPS3→MYO1E← ZNF175, SYNE1→SLC12A2←Hs.127657, C4.4A→PTEN ←NAPB, MGC23985→DDB1←HOXA9
{25→12←11}	T→MYCN←SLC30A4, CENTB1→CD2←NFKB2, SPINT2 →IRF3←ACVRL1, Hs.504960→TRAF2←BYSL, Hs.504960→TRAF2←SP192, MGC11266→CSNK1E← MCRS1, Hs.23871→ELAVL2←Hs.439153, GFPT1→BCL3 ←Hs.2173, KIAA1030→SEC5L1,HUS1B←TEB4
{17→10←9}	NAB2→STARD3←OPHN1, CINP→KRT8P12←MYF5

The gene-gene relationships narrowed down by the above procedure are listed in Table III. As seen in the table, 40 gene-gene relationships were narrowed down from a large number of possible gene relationships in the cluster

relationships, {29→6←1}, {22→12←11}, {25→12←11}, and {17→10←9}: in {17→10←21}, no gene relationships were detected. Interestingly, known cancer-related genes were included in the constituent genes of the relationships. Indeed, 17 genes in 17 relationships are described as the cancer-related genes in OMIM (Online Mendelian Inheritance in Man) [19]: CDH1, FGFR2, RRAS2, NEO1, NFKB2, TK2, HOXA9, ATP2A3, TCF7L2, CTAG1B, C4.4A, PTEN, MYCN, SPINT2, TRAF2, BCL3, and STARD3. Note that both known cancer-related genes and genes with other functions were included in most relationships. In these cases, genes with other functions may be involved in the cancer progression. In addition, some genes with unidentified functions were also included in the relationships above. This may suggest that one of the functions of the genes may be related to the cancer progression. At any rate, the gene-gene relationships, which were narrowed down with reference to network structure changes, reflect well the knowledge about the genes responsible for the cancer, and show the possibility of unknown relationships related to the cancer progression.

E. Merits and Pitfalls

We proposed a procedure for inferring a model for progressive stages from the entire data set, by using the graphical chain model. Furthermore, the following analyses of the evaluation of association strength by a statistical test and the selection of gene-gene relationships by the PC algorithm narrowed down the candidates of the gene sets causing the inferred cluster-cluster relationships. By using the above procedure, we analyzed 8,427 gene expression profiles in the two stages of hepatocellular carcinoma from CH to HCC. By the analyses, the chain model including the background stage was constructed, and several gene cluster connections were found to cause the progression from CH to HCC. Furthermore, 40 candidates of gene-gene relationships responsible for the progression emerged, with reference to the cluster-cluster relationships. Thus, we successfully described a framework of network structure changes for cancer progression, and based on the inferred changes, further refined the causal gene-gene relationships for the cell stage progression in a rational and systematic manner.

It is interesting in considering the present procedure in the case of more than two cell stages. For example, we can allocate 27 cluster sets of three cell stages into four groups, according to the rule adopted in the two cell stage: the first group, [+, /, /], [+, -, /], [+, -, -], [+, /, -], and [-, /, /]; the second group, [/, +, /], [-, +, /], [-, +, -], [/, +, -], and [/, -, /]; the third group, [/, /, +], [-, /, +], [-, -, +], [/, -, +], and [/, /, -]; the background group, [+, +, +], [-, -, -], [/, /, /], [+, +, -], [+, +, /], [+, -, +], [+, /, +], [-, +, +], [/, +, +], [-, -, /], [-, /, -], and [/, -, -]. In the allocation of the above clusters into the background group, some ambiguity emerged; the clusters showing up- or down-regulation at two cell stages are included. However, the present allocation rule may be reasonable, on the assumption that the network structure change is responsible for the up- and down-regulated classes characterizing each cell stage. Although the ambiguity of the allocation into the background group emerges as the number of stages increases, the present procedure may help to reveal the gene-gene relationships as well as to capture the network structure change through the distinctive cell stages in a systematic manner.

REFERENCES

[1] T. Finkel and N.J. Holbrook, "Oxidants, oxidative stress and the biology of ageing," Nature, 408, 239-247, 2000.

[2] H.C. Causton, B. Ren, S.S. Koh, C.T. Harbison, E. Kanin, E.G. Jennings, T.I. Lee, H.L. True, E.S. Lander and R.A. Young, "Remodeling of Yeast Genome Expression in Response to Environmental Changes," Mol. Biol. Cell, 12, 323-337, 2001.

[3] K. Kiyosawa, T. Sodeyama, E. Tanaka, Y. Gibo, K. Yoshizawa, Y. Nakano, S. Furuta, et al., "Interrelationship of blood transfusion, non-A, non-B hepatitis and hepatocellular carcinoma: analysis by detection of antibody to hepatitis C virus," Hepatology, 12, 671-675, 1990.

[4] M. Honda, S. Kaneko, H. Kawai, Y. Shirota, K. Kobayashi, "Differential gene expression between chronic hepatitis B and C hepatic lesion," Gastroenterology, 120, 955-966, 2001.

[5] M. Honda, T. Yamashita, T. Ueda, H. Takatori, R. Nishino, S. Kaneko, "Different signaling pathways in the livers of patients with chronic hepatitis B or chronic hepatitis C," Hepatology, 44, 1122-1138, 2006.

[6] S. Aburatani, S. Saito, H. Toh and K. Horimoto, "A Graphical Chain Model for Inferring Regulatory System Networks from Gene Expression Profiles," Statistical Methodology, 3, 17-28, 2006

[7] N. Wermuth and S. L. Lauritzen, "On substantive research hypotheses, conditional independence graphs and graphical chain models," J. R. Statist. Soc. B, 52, 21-50, 1990.

[8] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, Chichester, 1990.

[9] D. Edwards, *Introduction to Graphical Modelling*, Springer, New York, 1995.

[10] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996.

[11] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," Mol. Biol. Cell, 9, 3273–3297, 1998.

[12] P. Dempster, "Covariance selection," Biometrics, 28, 157-175, 1972.

[13] K. Horimoto and H. Toh, "Statistical Estimation of Cluster Boundaries in Gene Expression Profile Data," Bioinformatics, 17, 1143-1151, 2001.

[14] S. Aburatani, F. Sun, S. Saito, M. Honda, S. Kaneko and K. Horimoto, "Gene systems network inferred from expression profiles in hepatocellular carcinogenesis by graphical Gaussian model," EURASIP J. Bioinfo. Systems Biol., 47214, 2007.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd Edition. New York, John Wiley & Sons, 1984.

[16] P. Spirtes, C. Glymour and R. Scheines *Causation, Prediction, and Search* (Springer Lecture Notes in Statistics, 2nd edition, revised). MIT Press, Cambridge, 2001.

[17] T. Verma and J. Peral, "An algorithm for deciding if a set of observed independence has a causal explanation," Proc. 8th Conf. on Uncertainty in AI, Stanford, Morgan Kaufmann, p. 323-330, 1992.

[18] S. Aburatani, K. Goto, S. Saito, H. Toh and K. Horimoto, "ASIAN: A Web Server for Inferring a Regulatory Network Framework from Gene Expression Profiles," Nucleic Acids Res, 33, W659-W664, 2005.

[19] http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim