Hindawi Publishing Corporation
EURASIP Journal on Bioinformatics and Systems Biology
Volume 2007, Article ID 47214, 11 pages
doi:10.1155/2007/47214

Research Article

Gene Systems Network Inferred from Expression Profiles in Hepatocellular Carcinogenesis by Graphical Gaussian Model

Sachiyo Aburatani,[1] Fuyan Sun,[1] Shigeru Saito,[2] Masao Honda,[3] Shu-ichi Kaneko,[3] and Katsuhisa Horimoto[1]

[1] *Biological Network Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan*
[2] *Chemo & Bio Informatics Department, INFOCOM CORPORATION, Mitsui Sumitomo Insurance Surugadai Annex Building, 3-11, Kanda-Surugadai, Chiyoda-ku, Tokyo 101-0062, Japan*
[3] *Department of Gastroenterology, Graduate School of Medical Science, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-8641, Japan*

Received 28 June 2006; Revised 27 February 2007; Accepted 1 May 2007

Recommended by Paul Dan Cristea

Hepatocellular carcinoma (HCC) in a liver with advanced-stage chronic hepatitis C (CHC) is induced by hepatitis C virus, which chronically infects about 170 million people worldwide. To elucidate the associations between gene groups in hepatocellular carcinogenesis, we analyzed the profiles of the genes characteristically expressed in the CHC and HCC cell stages by a statistical method for inferring the network between gene systems based on the graphical Gaussian model. A systematic evaluation of the inferred network in terms of the biological knowledge revealed that the inferred network was strongly involved in the known gene-gene interactions with high significance ($P < 10^{-4}$), and that the clusters characterized by different cancer-related responses were associated with those of the gene groups related to metabolic pathways and morphological events. Although some relationships in the network remain to be interpreted, the analyses revealed a snapshot of the orchestrated expression of cancer-related groups and some pathways related with metabolisms and morphological events in hepatocellular carcinogenesis, and thus provide possible clues on the disease mechanism and insights that address the gap between molecular and clinical assessments.

1. INTRODUCTION

Hepatitis C virus (HCV) is the major etiologic agent of non-A non-B hepatitis, and chronically infects about 170 million people worldwide [1–3]. Many HCV carriers develop chronic hepatitis C (CHC), and finally are afflicted with hepatocellular carcinoma (HCC) in livers with advanced-stage CHC. Thus, the CHC and HCC cell stages are essential in hepatocellular carcinogenesis.

To elucidate the mechanism of hepatocellular carcinogenesis at a molecular level, many experiments have been performed from various approaches. In particular, recent advances in techniques to monitor simultaneously the expression levels of genes on a genomic scale have facilitated the identification of genes involved in the tumorigenesis [4]. Indeed, some relationships between the disease and the tumor-related genes were proposed from the gene expression analyses [5–7]. Apart from the relationship between tumor-related genes and the disease at the molecular level, the information about the pathogenesis and the clinical characteristics of hepatocellular carcinogenesis has accumulated steadily [8, 9]. However, there is a gap between the information about hepatocellular carcinogenesis at the molecular level and that at more macroscopic levels, such as the clinical level. Furthermore, the relationships between tumor-related genes and other genes also remain to be investigated. Thus, an approach to describe the perspective of carcinogenesis from measurements at the molecular level is desirable to bridge the gap between the information at the two different levels.

Recently, we have developed an approach to infer a regulatory network, which is based on graphical Gaussian modeling (GGM) [10, 11]. Graphical Gaussian modeling is one of the graphical models that includes the Boolean and Bayesian models [12, 13]. Among the graphical models, GGM has the simplest structure in a mathematical sense; only the inverse of the correlation coefficient between the variables is needed,

and therefore, GGM can be easily applied to a wide variety of data. However, straightforward applications of statistical theory to practical data fail in some cases, and GGM also fails frequently when applied to gene expression profiles; here the expression profile indicates a set of the expression degrees of one gene, measured under various conditions. This is because the profiles often share similar expression patterns, which indicate that the correlation coefficient matrix between the genes is not regular. Thus, we have devised a procedure, named ASIAN (automatic system for inferring a network), to apply GGM to gene expression profiles, by a combination of hierarchical clustering [14]. First, the large number of profiles is grouped into clusters, according to the standard approach of profile analysis [15]. To avoid the generation of a nonregular correlation coefficient matrix from the expression profiles, we adopted a stopping rule for hierarchical clustering [10]. Then, the relationship between the clusters is inferred by GGM. Thus, our method generates a framework of gene regulatory relationships by inferring the relationships between the clusters [11, 16], and provides clues toward estimating the global relationships between genes on a large scale.

Methods for extracting biological knowledge from large amounts of literature and arranging it in terms of gene function have been developed. Indeed, ontologies have been made available by the gene ontology (GO) consortium [17] to construct a functional categorization of genes and gene products, and by using the GO terms, the software determines whether any GO terms annotate a specified list of genes at a frequency greater than that expected by chance [18]. Furthermore, various software applications, most of which are commercial software, such as MetaCore from GeneGo http://www.genego.com/, have been developed for the navigation and analysis of biological pathways, gene regulation networks, and protein interaction maps [19]. Thus, advances in the processing of biological knowledge have enabled us to correspond to the results of gene expression analyses for a large amount of data with the biological functions.

In this study, we analyzed the gene expression profiles from the CHC and HCC cell stages, by ASIAN based on the graphical Gaussian Model, to reveal the framework of gene group associations in hepatocellular carcinogenesis. For this purpose, first, the genes characteristically expressed in hepatocellular carcinogenesis were selected, and then, the profiles of the genes thus selected were subjected to the association inference method. In addition to the association inference, which was presented by the network between the clusters, the network was further interpreted systematically by the biological knowledge of the gene interactions and by the functional categories with GO terms. The combination of the statistical network inference from the profiles with the systematic network interpretation by the biological knowledge in the literature provides a snapshot of the orchestration of gene systems in hepatocellular carcinogenesis, especially for bridging the gap between the information on the disease mechanisms at the molecular level and at more macroscopic levels.

2. MATERIALS AND METHODS

2.1. Gene selection

We selected the up- and downregulated genes characteristically expressed in the CHC and HCC stages, as a prerequisite for defining the variables in the network inference by the graphical Gaussian modeling. This involved the following steps. (1) The averages and the standard deviations in the respective conditions, AV_j and SD_j, for $j = 1, \ldots, N_c$, are calculated. (2) The expression degree of the ith gene in the jth condition, e_{ij}, is compared with $|AV_j \pm SD_j|$. (3) The gene is regarded as a characteristically expressed gene, if the number of conditions that $e_{ij} \geq |AV_j \pm SD_j|$ is more than $N_c/2$. Although the criterion for a characteristically expressed gene is usually $|AV_j \pm 2SD_j|$, the present selection procedure described above is simply designed to gather as many characteristically expressed genes as possible, and is suitable to capture a macroscopic relationship between the gene systems estimated by the following cluster analysis.

2.2. Gene systems network inference

The present analysis is composed of three parts: first, the profiles selected in the preceding section are subjected to the clustering analysis with the automatic determination of cluster number, and then the profiles of clusters are subjected to the graphical Gaussian modeling. Finally, the network inferred by GGM is rearranged according to the magnitude of partial correlation coefficients, which can be regarded as the association strength, between the clusters. The details of the analysis are as follows.

2.2.1. Clustering with automatic determination of cluster number

In clustering the gene profiles, here, the Euclidian distance between Pearson's correlation coefficients of profiles and the unweighted pair group method using arithmetic average (UPGMA or group average method) were adopted as the metric and the technique, respectively, with reference to the previous analyses by GGM [11, 16]. In particular, the present metric between the two genes is designed to reflect the similarity in the expression profile patterns between other genes as well as between the measured conditions, that is,

$$d_{ij} = \sqrt{\sum_{l=1}^{n} (r_{il} - r_{jl})^2}, \tag{1}$$

where n is the total number of the genes, and r_{ij} is the Pearson correlation coefficient between the i and j genes of the expression profiles that are measured at N_c conditions, p_{ik}, $(k = 1, 2, \ldots, N_c)$:

$$r_{ij} = \frac{\sum_{k=1}^{l} (p_{ik} - \overline{p_i}) \cdot (p_{jk} - \overline{p_j})}{\sqrt{\sum_{k=1}^{l} (p_{ik} - \overline{p_i})^2 \cdot \sum_{k=1}^{l} (p_{jk} - \overline{p_j})^2}}, \tag{2}$$

where $\overline{p_i}$ is the arithmetic average of p_{ik} over N_c conditions.

In the cluster number estimation, various stopping rules for the hierarchical clustering have been developed [20]. Recently, we have developed a method for estimating the cluster number in the hierarchical clustering, by considering the following application of the graphical model to the clusters [10]. In our approach, the variance inflation factor (VIF) is adopted as a stopping rule, and is defined by

$$\mathrm{VIF}_i = r_{ii}^{-1}, \tag{3}$$

where r_{ii}^{-1} is the ith diagonal element of the inverse of the correlation coefficient matrix between explanatory variables [21]. In the cluster number determination, the popular cutoff value of 10.0 [21] was adopted as a threshold in the present analysis, also with reference to the previous analyses.

After the cluster number determination, the average expression profiles are calculated for the members of each cluster, and then the average correlation coefficient matrix between the clusters is calculated from them. Finally, the average correlation coefficient matrix between the clusters is subjected to the graphical Gaussian modeling. Note that the average coefficient correlation matrix avoids the difficulty of the above numerical calculation, due to the distinctive patterns of the average expression profiles of clusters. This means that the GGM works well for the average coefficient correlation matrix.

2.2.2. *Graphical Gaussian modeling*

The concept of conditional independence is fundamental to graphical Gaussian modeling (GGM). The conditional independence structure of the data is characterized by a conditional independence graph. In this graph, each variable is represented by a vertex, and two vertices are connected by an edge if there is a direct association between them. In contrast, a pair of vertices that are not connected in the graph is conditionally independent.

In the procedure for applying the GGM to the profile data [11], a graph, $G = (V, E)$, is used to represent the relationship among the M clusters, where V is a finite set of nodes, each corresponding to one of the M clusters, and E is a finite set of edges between the nodes. E consists of the edges between cluster pairs that are conditionally dependent. The conditional independence is estimated by the partial correlation coefficient, expressed by

$$r_{i,j|\mathrm{rest}} = -\frac{r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}}, \tag{4}$$

where $r_{ij|\mathrm{rest}}$ is the partial correlation coefficient between variables i and j, given the rest variables, and r_{ij} is the (i, j) element in the reverse of the correlation coefficient matrix.

In order to evaluate which pair of clusters is conditionally independent, we applied the covariance selection [22], which was attained by the stepwise and iterative algorithm developed by Wermuth and Scheidt [23]. The algorithm is presented as Algorithm 1.

The graph obtained by the above procedure is an undirected graph, which is called an independence graph. The independence graph represents which pair of clusters is conditionally independent. That is, when the partial correlation coefficient for a cluster pair is equal to 0, the cluster pair is conditionally independent, and the relationship is expressed as no edge between the nodes corresponding to the clusters in the independence graph.

The genes grouped into each cluster are expected to share similar biological functions, in addition to the regulatory mechanism [24]. Thus, a network between the clusters can be approximately regarded as a network between gene systems, each with similar functions, from a macroscopic viewpoint. Note that the number of connections in one vertex is not limited, while it is only one in the cluster analysis. This feature of the network reflects the multiple relationships of a gene or a gene group in terms of the biological function.

Step 0: Prepare a complete graph of $G(0) = (V, E)$. The nodes correspond to M clusters. All of the nodes are connected. $G(0)$ is called a full model. Based on the expression profile data, construct an initial correlation coefficient matrix $C(0)$.

Step 1: Calculate the partial correlation coefficient matrix $P(\tau)$ from the correlation coefficient matrix $C(\tau)$. τ indicates the number of the iteration.

Step 2: Find an element that has the smallest absolute value among all of the nonzero elements of $P(\tau)$. Then, replace the element in $P(\tau)$ with zero.

Step 3: Reconstruct the correlation coefficient matrix, $C(\tau + 1)$, from $P(\tau)$. In $C(\tau + 1)$, the element corresponding to the element set to zero in $P(\tau)$ is revised, while all of the other elements are left to be the same as those in $C(\tau)$.

Step 4: In the Wermuth and Sheidt algorithm, the termination of the iteration is judged by the "deviance" values. Here, we used two types of deviance, dev1 and dev2, with the following:

$$\begin{aligned} \mathrm{dev1} &= N_c \log\left(\frac{|C(\tau+1)|}{|C(0)|}\right), \\ \mathrm{dev2} &= N_c \log\left(\frac{|C(\tau+1)|}{|C(\tau)|}\right). \end{aligned} \tag{5}$$

Calculate dev1 and dev2. The two deviances follow an asymptotic χ^2 distribution with a degree of freedom = n, and that with a degree of freedom = 1, respectively. n is the number of elements that are set to zero until the $(\tau + 1)$th iteration. In our approach, n is equal to $(\tau + 1)$. $|C(\tau)|$ indicates the determinant of $C(\tau)$. N_c is the number of different conditions under which the expression levels of M clusters are measured.

Step 5: If the probability value corresponding to dev1 ≤ 0.05, or the probability value corresponding to dev2 ≤ 0.05, then the model $C(\tau + 1)$ is rejected, and the iteration is stopped. Otherwise, the edge between a pair of clusters with a partial correlation coefficient set to zero in $P(\tau)$ is omitted from $G(\tau)$ to generate $G(\tau + 1)$, and τ is increased by 1. Then, go to Step 1.

Algorithm 1

2.2.3. Rearrangement of the inferred network

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 by GGM still showed complicated forms with many edges [11, 16]. Since the magnitude of the partial correlation coefficient indicates the strength of the association between clusters, the intact network can be rearranged according to the partial correlation coefficient value, to interpret the association between clusters. The strength of the association can be assigned by a standard test for the partial correlation coefficient [25]. By Fisher's Z transformation of partial correlation coefficients, that is,

$$Z = \frac{1}{2}\log\left(\frac{1 + r_{ij\cdot\text{rest}}}{1 - r_{ij\cdot\text{rest}}}\right), \tag{6}$$

Z is approximately distributed according to the following normal distribution:

$$N\left(\frac{1}{2}\log\left(\frac{1 + r_{ij\cdot\text{rest}}}{1 - r_{ij\cdot\text{rest}}}\right), \frac{1}{\{N_c - (M-2)\} - 3}\right), \tag{7}$$

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.

2.3. Statistical significance of the inferred network with the biological knowledge

The inferred network can be statistically evaluated in terms of the gene-gene interactions. The chance probability was estimated by the correspondence between the inferred cluster network and the information about gene interactions. The following steps were used. (1) The known gene pairs with interactions in the database were overlaid onto the inferred network. (2) The number of cluster pairs, upon which the gene interactions were overlaid, was counted. (3) The chance probability, in which the cluster pairs connected by the established edges in the network were found in all possible pairs, was calculated by using the following equation:

$$P = 1 - \sum_{i=0}^{f-1} \frac{\binom{g}{i}\binom{N-g}{n-i}}{\binom{N}{n}}, \tag{8}$$

where N is the number of possible cluster pairs in the network, n is the number of cluster pairs with edges in the inferred network, f is the number of cluster pairs with edges in the inferred network, including the known gene pairs with interactions, and g is the number of cluster pairs, including the known gene pairs with interactions.

2.4. Evaluation of the inferred network in terms of the biological knowledge

The inferred network can be evaluated in terms of the biological knowledge. For this purpose, we characterize the clusters by GO terms, and overlay the knowledge about the gene interactions onto the network. For this purpose, we first use GO::TermFinder [18] to characterize the clusters by GO terms with the user-defined significance probability (http://search.cpan.org/dist/GO-TermFinder). Then, Pathway Studio [19] is used to survey the biological information about the gene interactions between the selected genes.

2.5. Software

All calculations of the present clustering and GGM were performed by the ASIAN web site [26, 27] (http://www.eureka.cbrc.jp/asian) and "Auto Net Finder," the commercialized PC version of ASIAN, from INFOCOM CORPORATION, Tokyo, Japan (http://www.infocom.co.jp/bio/download).

2.6. Expression profile data

The expression profiles of 8516 genes were monitored in 27 CHC samples and 17 HCC samples [28].

3. RESULTS AND DISCUSSION

3.1. Clustering

Among the 8516 genes with expression profiles that were measured in the previous studies [28], 661 genes were selected as those characteristically expressed in the CHC and HCC stages. As a preprocessing step for the association inference, the genes thus selected were automatically divided into 18 groups by ASIAN [26, 27]. Furthermore, each cluster was characterized in terms of the GO terms, which define the macroscopic features of the cluster in terms of the biological function.

Figure 1 shows the dendrogram of clusters, together with their expression patterns. As seen in Figure 1, the genes were grouped into 18 clusters, in terms of the number of members and the expression patterns in the clusters. The average number of cluster members was 36.7 genes (SD, 14.2), and the maximum and minimum numbers of members were 69 in cluster 14 and 18 in cluster 9, respectively. As for the expression pattern, five clusters (10, 12, 14, 15, and 18) and ten clusters (1–7, 9, 16, and 17) were composed of up- and downregulated genes, respectively, and three clusters (8, 11, and 13) showed similar mixtures of up- and downregulated genes.

Table 1 shows the GO terms for the clusters (clusterGOB), which characterized them well (see details at http://www.cbrc.jp/~horimoto/suppl/HCGO.pdf). Among the 661 genes analyzed in this study, 525 genes were characterized by the GO terms, and among the 18 clusters, 11 clusters were characterized by GO terms with $P < .05$. In addition, 188 genes (28.3% of all characterized genes) corresponded to the GO terms listed in Table 1. As seen in the table, although most clusters are characterized by several GO terms, reflecting the fact that the genes function generally in multiple pathways, the clusters are not composed of a mixture of genes with distinctive functions. For example, cluster 2 is characterized by 10 terms, and most of the terms

are related to the energy metabolism. Thus, the GO terms in the respective clusters share similar features of biological functions, which cause the hierarchical structure of the GO term definitions.

In Table 1, most of the clusters characterized by GO terms with $P < .05$ are related to response function and to metabolism. Clusters 1, 6, 8, 12, and 13 are characterized by GO terms related to different responses, and clusters 2, 3, 4, and 7 are characterized by GO terms related to different aspects of metabolism. Although the genes in two clusters, 14 and 16, did not adhere to this dichotomy, the genes characteristically expressed in HCC in the above nine clusters were related to the responses and the metabolic pathways. As for the remaining clusters with lower significance, three clusters (9, 10, and 11) were also characterized by response functions, and four clusters (5, 15, 17, and 18) were related to morphological events at the cellular level. Note that none of the clusters characterized by cellular level events attained the significance level. This may be because the genes related to cellular level events represent only a small fraction of genes relative to all genes with known functions, in comparison with the genes related to molecular level events in the definition of GO terms.

It is interesting to determine the correspondence between the up- and downregulated genes and the GO terms in the clusters. In the five clusters of upregulated genes, clusters 10 and 12 were characterized by different responses, and two clusters were characterized by morphological events, which were the categories of "cell proliferation" in cluster 15 and of "development" in cluster 18. The remaining cluster, 14, was characterized by regulation, development, and metabolism. As for the clusters of downregulated genes, four of the ten clusters were characterized by GO terms related to various aspects of metabolism. In the remaining six clusters, three clusters were characterized by GO terms related to responses, two clusters were characterized by morphological events, and one cluster was characterized by mixed categories.

In summary, the present gene selection and the following automatic clustering produced a macroscopic view of gene expression in hepatocellular carcinogenesis. Although the clusters contain many genes that do not always share the same functions, the clusters were characterized by their responses, morphological events, and metabolic aspects from a macroscopic viewpoint. The clusters of upregulated genes were characterized by the former two categories, and those of the downregulated genes represented all three categories. Thus, the present clustering serves to interpret the network between the clusters in terms of the biological function and the gene expression pattern.

3.2. Known gene interactions in the inferred network

The association between the 18 clusters inferred by GGM is shown in Figure 2. In the intact network by ASIAN, 96 of 153 possible edges between 18 clusters (about 63%) were established by GGM. Since the intact network is still messy, the network was rearranged to interpret its biological meaning by extracting the relatively strong associations between the clusters, according to the procedure in Section 2.2.3. After the rearrangement, 34 edges remained by the statistical test of the partial correlation coefficients with 5% significance. In the rearranged network, all of the clusters were nested, but each cluster was connected to a few other clusters. Indeed, the average number of edges per cluster was 2.3, and the maximum and minimum numbers of edges were seven in cluster 15 and one in cluster 9, respectively. In particular, the numbers of edges are not proportional to the numbers of constituent genes in each cluster. For example, while the numbers of genes in clusters 15 and 17 are equal to each other (24 genes), the number of edges from cluster 15 (2 edges) differs from that from cluster 17 (5 edges). Thus, the number of edges does not depend on the number of genes belonging to the cluster, but rather on the gene associations between the cluster pairs.

To test the validity of the inferred network in terms of biological function, the biological knowledge about the gene interactions is overlaid onto the inferred network. For this

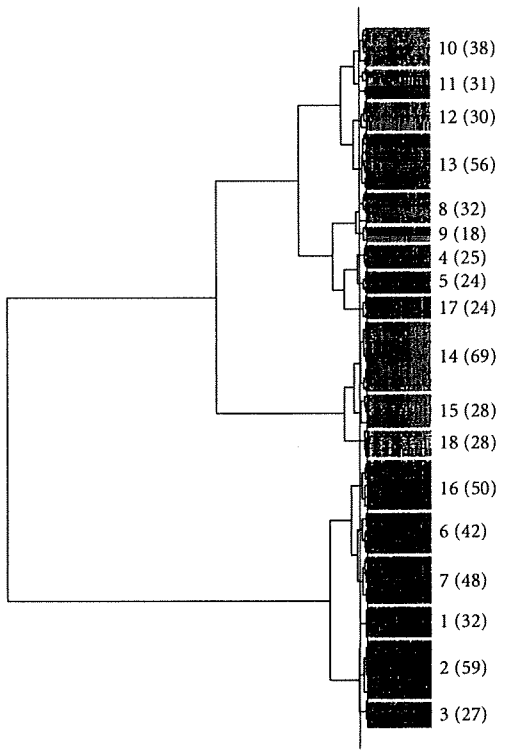

FIGURE 1: *Dendrogram of genes and profiles.* The dendrogram was constructed by hierarchical clustering with the metric of the Euclidian distances between the correlation coefficients and the UPGMA. The blue line on the dendrogram indicates the cluster boundary estimated automatically by ASIAN. The gene expression patterns of the respective clusters in the CHC and HCC stages are shown by the degree of intensity: the red and green colors indicate relatively higher and lower intensities. The cluster number and the number of member genes in each cluster (in parentheses) are denoted on the right side of the figure.

purpose, all of the gene pairs belonging to cluster pairs are surveyed by Pathway Assist, which is a database for biological knowledge about molecular interactions, compiled based on the gene ontology [17]. Among the 661 genes analyzed in this study, the interactions between 90 gene pairs were detected by Pathway Assist, and 50 of these pairs were found in Figure 2. Notice that the number of gene pairs reported in the literature does not directly reflect the importance of the gene interactions, and instead is highly dependent on the number of scientists who are studying at the corresponding genes. Thus, we counted the numbers of cluster pairs in which at least one gene pair was known, by projecting the gene pairs with known interactions onto the network. By this projection, the interactions were found in 35 (g in the equation of Section 2.3) cluster pairs among 153 (N) possible pairs (see details of the gene pair projection at http://www.cbrc.jp/~horimoto/suppl/GPPN.pdf). Then, 19 (f) of the 35 cluster pairs were overlapped with 34 (n) cluster pairs in the rearranged network. The chance probability that a known interaction was found in the connected cluster pairs in the rearranged network was calculated as $P < 10^{-4.3}$. Thus, the rearranged network faithfully captures the known interactions between the constituent genes.

Furthermore, the genes with known interactions were corresponded to the genes responsible for the GO terms of each cluster, as shown in Table 1. The genes responsible for the GO terms were distributed over all cluster pairs, including gene pairs with known interactions, except for only two pairs, clusters 15 and 17, and 15 and 18. Thus, the network can be interpreted not only by the known gene interactions but also by the GO terms characterizing the clusters.

3.3. Gene systems network characterized by GO terms

3.3.1. Coarse associations between the clusters

To elucidate the associations between the clusters, the cluster associations with 1% significance probability were further discriminated from those with 5% probability. This generated four groups of clusters, shown in Figure 3(a).

First, we will focus on the groups including the clusters that were characterized by GO terms with a significance probability, and that were definitely occupied by up- or downregulated genes (clusters depicted by triangles with bold lines in the figure). Groups I and III attained the above criteria. In group I, the clusters were a mixture of the clusters of the up- and downregulated genes. Note that three of the six clusters were composed of upregulated genes, which were characterized by responses (cluster 12), mixed categories (cluster 14), and morphological events (cluster 15). In group III, all three clusters were of downregulated genes. One cluster was characterized by responses, and two were characterized by amino-acid-related metabolism. In contrast, groups II and IV were composed of the clusters that were somewhat inadequately characterized by GO terms and expression patterns. Thus, groups I and III provide the characteristic features about the orchestration of gene expression in hepatocellular carcinogenesis.

Secondly, a coarse grinning for group associations provides another viewpoint, shown in Figure 3(b). When the groups with at least one edge between the clusters in the respective groups were presented, regardless of the number of edges, groups I, II, and IV were nested, and group III was connected with only group I. In the second view, group I, which includes three of the five clusters of upregulated genes in all clusters, was associated with all of the other groups. This suggests that group I represents a positive part of the gene expression in hepatocellular carcinogenesis, which is consistent with the interpretation by the first view, from the significant GO terms and the clear expression patterns. Interestingly, among the clusters characterized by morphological events (clusters 5, 15, 17, and 18), three of the four clusters were distributed over groups I, II, and IV, and the distribution was consistent with the nested groups. This suggests that the upregulated genes of the clusters in group I are responsible for the events at the cellular level.

Thirdly, the clusters not belonging to the four groups were clusters 1, 3, and 5. Clusters 1, 3, and 5 were directly connected with groups I, III, and IV, groups I and III, and group IV, respectively. Interestingly, cluster 1, characterized by only "*anti-inflammatory response*," was connected with five clusters belonging to three groups, in which four clusters were downregulated clusters. Although cluster 5 was not clearly characterized by the GO terms, cluster 3 was characterized by metabolic terms that were quite similar to those for cluster 2, a downregulated cluster. Thus, the three clusters may be concerned with downregulation in hepatocellular carcinogenesis.

3.3.2. Interpretations of the inferred network in terms of pathogenesis

The coarse associations between the clusters in the preceding section can be interpreted on the macroscopic level, such as the pathological level. The interpretation of the network inferred based on the information at the molecular level will be useful to bridge the gap between the information about the disease mechanisms at the molecular and more macroscopic levels.

One of the most remarkable associations is found in group I. Cluster 12, with upregulation, was associated at a 1% significance level with cluster 2, with downregulation. The former cluster is characterized by the GO terms related to the immune response, and the latter is characterized by those involved with metabolism. In general, CHC and HCC result in serious damage to hepatocytes, which are important cells for nutrient metabolism, and the damage induces different responses. Indeed, HCC is a suitable target for testing active immunotherapy [29]. Furthermore, cluster 2 was also associated at a 1% significance level with cluster 14, characterized by prostaglandin-related terms. This may reflect the fact that one mediator of inflammation, prostaglandin, shows elevated expression in human and animal HCCs [30]. Thus, the associations in group I are involved in the molecular pathogenesis of the CHC and HCC stages.

TABLE 1: Cluster characterization by GO terms#.

Cluster no.	GO no.	Category	*P*-value	Fraction
1	GO:0030236	Anti-inflammatory response	0.18%	2 of 22/6 of 26081
2	GO:0006094	Gluconeogenesis	0.06%	3 of 37/19 of 26081
2	GO:0006066	Alcohol metabolism	0.12%	6 of 37/312 of 26081
2	GO:0006091	Generation of precursor metabolites and energy	0.14%	9 of 37/961 of 26081
2	GO:0019319	Hexose biosynthesis	0.34%	3 of 37/33 of 26081
2	GO:0046165	Alcohol biosynthesis	0.34%	3 of 37/33 of 26081
2	GO:0046364	Monosaccharide biosynthesis	0.34%	3 of 37/33 of 26081
2	GO:0006067	Ethanol metabolism	0.48%	2 of 37/5 of 26081
2	GO:0006069	Ethanol oxidation	0.48%	2 of 37/5 of 26081
2	GO:0006629	Lipid metabolism	1.47%	7 of 37/722 of 26081
2	GO:0009618	Response to pathogenic bacteria	4.96%	2 of 37/15 of 26081
3	GO:0006094	Gluconeogenesis	0.61%	2 of 15/19 of 26081
3	GO:0019319	Hexose biosynthesis	1.87%	2 of 15/33 of 26081
3	GO:0046165	Alcohol biosynthesis	1.87%	2 of 15/33 of 26081
3	GO:0046364	Monosaccharide biosynthesis	1.87%	2 of 15/33 of 26081
3	GO:0009069	Serine family amino acid metabolism	4.49%	2 of 15/51 of 26081
4	GO:0006725	Aromatic compound metabolism	0.07%	4 of 20/140 of 26081
4	GO:0009308	Amine metabolism	0.38%	5 of 20/454 of 26081
4	GO:0006570	Tyrosine metabolism	0.59%	2 of 20/11 of 26081
4	GO:0050878	Regulation of body fluids	1.65%	3 of 20/113 of 26081
4	GO:0006950	Response to stress	2.70%	6 of 20/1116 of 26081
4	GO:0006519	Amino acid and derivative metabolism	4.12%	4 of 20/398 of 26081
4	GO:0007582	Physiological process	4.63%	20 of 20/17195 of 26081
5	GO:0006917	Induction of apoptosis*	16.06%	2 of 13/132 of 26081
5	GO:0012502	Induction of programmed cell death*	16.06%	2 of 13/132 of 26081
6	GO:0009613	Response to pest, pathogen, or parasite	0.00%	8 of 29/522 of 26081
6	GO:0043207	Response to external biotic stimulus	0.00%	8 of 29/557 of 26081
6	GO:0006950	Response to stress	0.00%	10 of 29/1116 of 26081
6	GO:0009605	Response to external stimulus	0.05%	10 of 29/1488 of 26081
6	GO:0006953	Acute-phase response	0.05%	3 of 29/25 of 26081
6	GO:0006955	Immune response	0.34%	8 of 29/1098 of 26081
6	GO:0006956	Complement activation	0.48%	3 of 29/52 of 26081
6	GO:0006952	Defense response	0.68%	8 of 29/1209 of 26081
6	GO:0050896	Response to stimulus	1.15%	11 of 29/2619 of 26081
6	GO:0009607	Response to biotic stimulus	1.65%	8 of 29/1372 of 26081
6	GO:0006629	Lipid metabolism	2.20%	6 of 29/722 of 26081
7	GO:0006559	L-phenylalanine catabolism	0.83%	2 of 31/9 of 26081
7	GO:0019752	Carboxylic acid metabolism	1.00%	6 of 31/590 of 26081
7	GO:0006082	Organic acid metabolism	1.02%	6 of 31/592 of 26081
7	GO:0006558	L-phenylalanine metabolism	1.26%	2 of 31/11 of 26081
7	GO:0009074	Aromatic amino acid family catabolism	1.26%	2 of 31/11 of 26081
7	GO:0006519	Amino acid and derivative metabolism	1.67%	5 of 31/398 of 26081
7	GO:0019439	Aromatic compound catabolism	1.79%	2 of 31/13 of 26081
7	GO:0006629	Lipid metabolism	3.04%	6 of 31/722 of 26081
7	GO:0009308	Amine metabolism	3.09%	5 of 31/454 of 26081
8	GO:0001570	Vasculogenesis	0.09%	2 of 21/4 of 26081
8	GO:0006950	Response to stress	0.42%	7 of 21/1116 of 26081
8	GO:0050896	Response to stimulus	2.33%	9 of 21/2619 of 26081

Table 1: Continued.

9	GO:0009611	Response to wounding*	11.19%	3 of 13/394 of 26081
10	GO:0009607	Response to biotic stimulus*	6.66%	6 of 19/1372 of 26081
11	GO:0050896	Response to stimulus*	72.68%	6 of 17/2619 of 26081
12	GO:0006955	Immune response	0.01%	8 of 18/1098 of 26081
12	GO:0006952	Defense response	0.01%	8 of 18/1209 of 26081
12	GO:0050874	Organismal physiological process	0.02%	10 of 18/2432 of 26081
12	GO:0009607	Response to biotic stimulus	0.03%	8 of 18/1372 of 26081
12	GO:0050896	Response to stimulus	0.39%	9 of 18/2619 of 26081
12	GO:0030333	Antigen processing	0.97%	3 of 18/108 of 26081
12	GO:0019882	Antigen presentation	2.62%	3 of 18/151 of 26081
12	GO:0019884	Antigen presentation, exogenous antigen	3.97%	2 of 18/32 of 26081
12	GO:0019886	Antigen processing, exogenous antigen via MHC class II	4.22%	2 of 18/33 of 26081
13	GO:0009611	Response to wounding	0.08%	6 of 30/394 of 26081
13	GO:0009613	Response to pest, pathogen, or parasite	0.38%	6 of 30/522 of 26081
13	GO:0043207	Response to external biotic stimulus	0.55%	6 of 30/557 of 26081
13	GO:0006955	Immune response	3.12%	7 of 30/1098 of 26081
13	GO:0006950	Response to stress	3.44%	7 of 30/1116 of 26081
13	GO:0050874	Organismal physiological process	3.98%	10 of 30/2432 of 26081
14	GO:0051244	Regulation of cellular physiological process	0.51%	8 of 45/665 of 26081
14	GO:0007275	Development	0.94%	13 of 45/2060 of 26081
14	GO:0001516	Prostaglandin biosynthesis	3.30%	2 of 45/9 of 26081
14	GO:0046457	Prostanoid biosynthesis	3.30%	2 of 45/9 of 26081
14	GO:0051242	Positive regulation of cellular physiological process	4.35%	5 of 45/289 of 26081
15	GO:0008283	Cell proliferation*	29.37%	4 of 26/488 of 26081
16	GO:0042221	Response to chemical substance	0.16%	5 of 31/237 of 26081
16	GO:0008152	Metabolism	1.29%	25 of 31/11891 of 26081
16	GO:0009628	Response to abiotic stimulus	1.89%	5 of 31/400 of 26081
16	GO:0006445	Regulation of translation	2.82%	3 of 31/87 of 26081
17	GO:0050817	Coagulation*	13.92%	2 of 12/118 of 26081
18	GO:0007275	Development*	11.67%	6 of 16/2060 of 26081

*The gene ontology terms in each cluster, detected with 5% significance probability by using GO::TermFinder [18], are listed. When the terms with that significance probability were not found in the cluster, the terms with the smallest probability were listed as indicated by an asterisk. In the last column, "Fraction," the numbers of genes belonging to the corresponding category in the cluster, of genes belonging to the cluster, of genes belonging to the corresponding category in all genes of the GO term data set, and of all genes are listed.

The associated clusters 4 and 7 in group III, which were characterized by GO terms related to amino acid and lipid metabolism, also show downregulation. Indeed, the products of dysregulated (aberrant regulation) metabolism are widely used to examine liver function in common clinical tests [8]. In addition, the connection between the clusters in groups III and I implies that the downregulation of the clusters in group III may be related to abnormal hepatocyte function.

In addition, cluster 15 in group I, which is characterized by the GO term "proliferation," was associated with different clusters in groups I, II, and IV. It is known that abnormal proliferation is one of the obvious features of cancer [31]. This broad association may be responsible for the cellular level events in hepatocellular carcinogenesis.

In summary, the inferred network reveals a coarse snapshot of the gene systems related to the molecular pathogenesis and clinical characteristics of hepatocellular carcinogenesis. Although the resolution of the network is still low, due to the cluster network, the present network may provide some clues for further investigations of the pathogenic relationships involved in hepatocellular carcinoma.

3.3.3. *Interpretations of the inferred network in terms of gene-gene interactions*

In addition to the macroscopic interpretations above, the gene functionality from the gene-gene interactions listed in Figure 2 is also discussed in the context of hepatocellular carcinoma. Although the consideration of gene-gene interactions is beyond the aim of the present study, some examples may provide possible clues about the disease mechanisms.

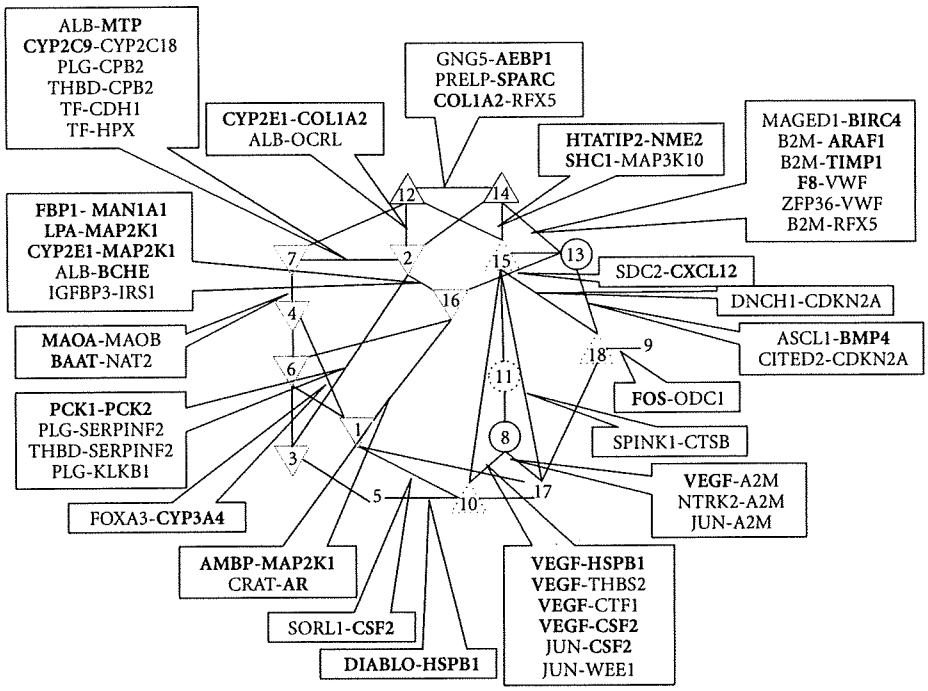

FIGURE 2: *Network between clusters, together with a projection of biological knowledge about the gene interactions.* The clusters are indicated by triangles and circles, in which the cluster numbers correspond to those in Figure 1, and the edges between the clusters are associations with 5% significance probability. The red triangles, the green upside-down triangles, and the circles indicate the clusters of up- and downregulated genes, and the mixture of them, respectively, and the dotted triangles indicate the clusters that were not characterized by GO terms with less than 5% significance probability. The known gene interactions in Pathway Assist are indicated between the clusters, in which the genes highlighted by bold letters are characterized by the GO terms in Table 1.

First, we surveyed the frequencies of GO terms (geneGOB listed in the supplemental data at http://www.cbrc.jp/~horimoto/suppl/HCGO.pdf) in the selected genes in the present analysis, to investigate the features of gene-gene interactions in the inferred network. A few general terms appeared frequently, such as "response" (122 times in the geneGOB column of the supplemental data at http://www.cbrc.jp/~horimoto/suppl/HCGO.pdf) and "metabolism" (183), as expected from the coarse associations between the clusters in the preceding section. As for more specific terms about the gene function, "lipid" (46), "apoptosis" (31), and "cell growth" (27) are remarkably found in the list. The "lipid" is expected from the relationship between groups I and III, and the "apoptosis" and the "cell growth" are also expected from the frequent appearance of GO terms (clusterGOB listed in Table 1) related to the morphological events. Since the frequent appearance of "lipid" may be a sensitive reflection of the protein-protein interactions in lipid metabolic pathways to the expression profiles, here, we focus on the gene-gene interactions characterized by the "apoptosis" and the "cell growth."

Among the gene-gene interactions listed in Figure 2, the gene-gene interactions characterized by the cell growth or death are found in the coarse associations between the clusters. Group I contains the gene-gene interactions related to apoptosis. The expression of HTAIP2 (HIV-1 Tat interactive protein 2, 30 kd) in cluster 14 induces the expression of a number of genes, including NME2 (nonmetastatic cells 2, protein) in cluster 15 as well as the apoptosis-related genes Bad and Siva [32]. MAGED1 (melanoma antigen, family D, 1) in cluster 13, and its binding partner BIRC4 (baculoviral IAP repeat-containing 4) in cluster 14 are known to play some roles in apoptosis [33]. In addition, the expression of COL1A2 (collagen, type I, alpha 2) in cluster 12, which is related to cell adhesion and skeletal development, is regulated by RFX5 (regulatory factor X, 5) in cluster 14 [29, 34]. In group IV, the expression of CSF2 (colony-stimulating factor 2) in cluster 8 is dependent on the cooperation between NFAT (nuclear factor of activated T cells) and JUN (Jun oncogene) in cluster 10 [35]. Between groups I and II, ASCL1 (achaete-scute complex-like 1) in cluster 13 and BMP4 (bone morphogenetic protein 4) in cluster 18 share the function of cell differentiation [36].

As a result, the gene-gene interactions listed above are related to the mechanisms of cell growth or death at the molecular level. On the other hand, the cluster associations reveal the relationship between the cancer-induced events and various aspects of metabolisms at the pathogenesis and clinical characteristics. Thus, the metabolic pathways might directly influence the mechanisms of cancer-induced cell growth or death at the molecular level in unknown ways.

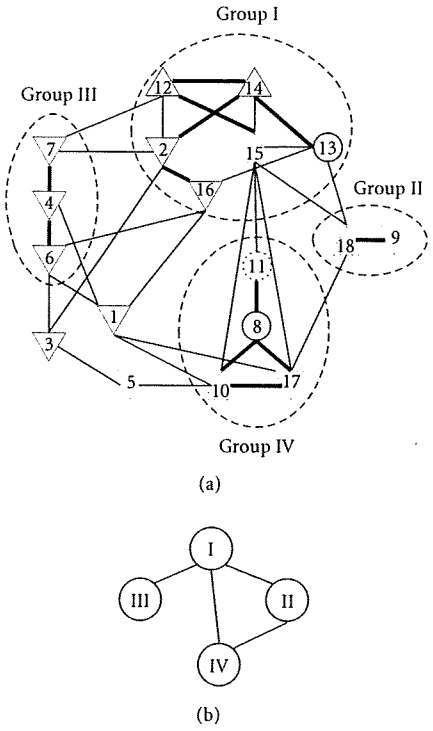

FIGURE 3: *Orchestration of gene systems.* (a) The association with 1% significance probability is indicated by a bold line, and the clusters with 1% significance association are naturally divided into four groups, which are enclosed by broken lines. (b) The connections between the groups are drawn schematically, as a coarse grinning of the cluster association.

3.4. Merits and pitfalls of the present approach

The present analysis reveals a framework of gene system associations in hepatocellular carcinogenesis. The inferred network provides a bridge between the events at the molecular level and those at macroscopic levels: the associations between clusters characterized by cancer-related responses and those characterized by metabolic and morphological events can be interpreted from pathological and clinical views. In addition, the viewpoint of the gene-gene interactions in the inferred network indicates the relationship between cancer and cell growth/death. Thus, the gene systems network may also be useful as a bridge between the gene-gene interactions and the observations at macroscopic levels, such as clinical tests.

The present method assumes linearity in the cluster associations by using a partial correlation coefficient to identify the independence between clusters. It is well known that the interactions among genes and other molecular components are often nonlinear, and the assumption of linearity misses many important relationships among genes. In the present study, our aim was not the inference of detailed gene-gene interactions, but of coarse gene system interactions. Indeed, the use of a partial correlation coefficient is employed as a feasible approach for gene association inference as a first approximation in some studies [37, 38]. Thus, the assumption of the linearity is not suitable for a fine analysis of dynamic gene behaviors, but may be useful for the approximate analysis of static gene associations.

ACKNOWLEDGMENTS

S. Aburatani was supported by a Grant-in-Aid for Scientific Research (Grant 18681031) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and K. Horimoto was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (Grant 18016008) and by a Grant-in-Aid for Scientific Research (Grant 19201039) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. This study was supported in part by the New Energy and Industrial Technology Development Organization (NEDO) of Japan and by the Ministry of Health, Labour, and Welfare of Japan.

REFERENCES

[1] M. J. Alter, H. S. Margolis, K. Krawczynski, et al., "The natural history of community-acquired hepatitis C in the United States. The sentinel counties chronic non-A, non-B hepatitis study team," *The New England Journal of Medicine*, vol. 327, no. 27, pp. 1899–1905, 1992.

[2] A. M. Di Bisceglie, "Hepatitis C," *The Lancet*, vol. 351, no. 9099, pp. 351–355, 1998.

[3] S. Zeuzem, S. V. Feinman, J. Rasenack, et al., "Peginterferon alfa-2a in patients with chronic hepatitis C," *The New England Journal of Medicine*, vol. 343, no. 23, pp. 1666–1672, 2000.

[4] S. S. Thorgeirsson, J.-S. Lee, and J. W. Grisham, "Molecular prognostication of liver cancer: end of the beginning," *Journal of Hepatology*, vol. 44, no. 4, pp. 798–805, 2006.

[5] N. Iizuka, M. Oka, H. Yamada-Okabe, et al., "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection," *The Lancet*, vol. 361, no. 9361, pp. 923–929, 2003.

[6] H. Okabe, S. Satoh, T. Kato, et al., "Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression," *Cancer Research*, vol. 61, no. 5, pp. 2129–2137, 2001.

[7] L.-H. Zhang and J.-F. Ji, "Molecular profiling of hepatocellular carcinomas by cDNA microarray," *World Journal of Gastroenterology*, vol. 11, no. 4, pp. 463–468, 2005.

[8] J. Jiang, P. Nilsson-Ehle, and N. Xu, "Influence of liver cancer on lipid and lipoprotein metabolism," *Lipids in Health and Disease*, vol. 5, p. 4, 2006.

[9] A. Zerbini, M. Pilli, C. Ferrari, and G. Missale, "Is there a role for immunotherapy in hepatocellular carcinoma?" *Digestive and Liver Disease*, vol. 38, no. 4, pp. 221–225, 2006.

[10] K. Horimoto and H. Toh, "Statistical estimation of cluster boundaries in gene expression profile data," *Bioinformatics*, vol. 17, no. 12, pp. 1143–1151, 2001.

[11] H. Toh and K. Horimoto, "Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling," *Bioinformatics*, vol. 18, no. 2, pp. 287–297, 2002.

[12] S. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, UK, 1996.

[13] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, John Wiley & Sons, New York, NY, USA, 1990.

[14] H. Toh and K. Horimoto, "System for automatically inferring a genetic network from expression profiles," *Journal of Biological Physics*, vol. 28, no. 3, pp. 449–464, 2002.

[15] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nature Genetics*, vol. 32, no. 5, pp. 502–508, 2002.

[16] S. Aburatani, S. Kuhara, H. Toh, and K. Horimoto, "Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling," *Signal Processing*, vol. 83, no. 4, pp. 777–788, 2003.

[17] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[18] E. I. Boyle, S. Weng, J. Gollub, et al., "GO::TermFinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004.

[19] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—the analysis and navigation of molecular networks," *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, 2003.

[20] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York, NY, USA, 1990.

[21] R. J. Freund and W. J. Wilson, *Regression Analysis: Statistical Modeling of a Response Variable*, Academic Press, San Diego, Calif, USA, 1998.

[22] A. P. Dempster, "Covariance selection," *Biometrics*, vol. 28, no. 1, pp. 157–175, 1972.

[23] N. Wermuth and E. Scheidt, "Algorithm AS 105: fitting a covariance selection model to a matrix," *Applied Statistics*, vol. 26, no. 1, pp. 88–92, 1977.

[24] L. F. Wu, T. R. Hughes, A. P. Davierwala, M. D. Robinson, R. Stoughton, and S. J. Altschuler, "Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters," *Nature Genetics*, vol. 31, no. 3, pp. 255–265, 2002.

[25] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1984.

[26] S. Aburatani, K. Goto, S. Saito, et al., "ASIAN: a website for network inference," *Bioinformatics*, vol. 20, no. 16, pp. 2853–2856, 2004.

[27] 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 Research*, vol. 33, pp. W659–W664, 2005.

[28] M. Honda, S. Kaneko, H. Kawai, Y. Shirota, and K. Kobayashi, "Differential gene expression between chronic hepatitis B and C hepatic lesion," *Gastroenterology*, vol. 120, no. 4, pp. 955–966, 2001.

[29] T. Wu, "Cyclooxygenase-2 in hepatocellular carcinoma," *Cancer Treatment Reviews*, vol. 32, no. 1, pp. 28–44, 2006.

[30] H. Xiao, V. Palhan, Y. Yang, and R. G. Roeder, "TIP30 has an intrinsic kinase activity required for up-regulation of a subset of apoptotic genes," *The EMBO Journal*, vol. 19, no. 5, pp. 956–963, 2000.

[31] W. B. Coleman, "Mechanisms of human hepatocarcinogenesis," *Current Molecular Medicine*, vol. 3, no. 6, pp. 573–588, 2003.

[32] Y. Xu, P. K. Sengupta, E. Seto, and B. D. Smith, "Regulatory factor for X-box family proteins differentially interact with histone deacetylases to repress collagen α2(I) gene *(COL1A2)* expression," *Journal of Biological Chemistry*, vol. 281, no. 14, pp. 9260–9270, 2006.

[33] P. A. Barker and A. Salehi, "The MAGE proteins: emerging roles in cell cycle progression, apoptosis, and neurogenetic disease," *Journal of Neuroscience Research*, vol. 67, no. 6, pp. 705–712, 2002.

[34] Y. Xu, L. Wang, G. Buttice, P. K. Sengupta, and B. D. Smith, "Interferon γ repression of collagen *(COL1A2)* transcription is mediated by the RFX5 complex," *The Journal of Biological Chemistry*, vol. 278, no. 49, pp. 49134–49144, 2003.

[35] F. Macian, C. Garcia-Rodriguez, and A. Rao, "Gene expression elicited by NFAT in the presence or absence of cooperative recruitment of Fos and Jun," *The EMBO Journal*, vol. 19, no. 17, pp. 4783–4795, 2000.

[36] J. Fu, S. S. W. Tay, E. A. Ling, and S. T. Dheen, "High glucose alters the expression of genes involved in proliferation and cell-fate specification of embryonic neural stem cells," *Diabetologia*, vol. 49, no. 5, pp. 1027–1038, 2006.

[37] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, no. 6, pp. 754–764, 2005.

[38] A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes, "Discovery of meaningful associations in genomic data using partial correlation coefficients," *Bioinformatics*, vol. 20, no. 18, pp. 3565–3574, 2004.

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
BioSystems 90 (2007) 486–495
BioSystems
www.elsevier.com/locate/biosystems

Derivation of rigorous conditions for high cell-type diversity by algebraic approach

Hiroshi Yoshida[a,*], Hirokazu Anai[b], Katsuhisa Horimoto[c]

[a] *Laboratory of Biostatistics, Institute of Medical Science, The University of Tokyo, Shirokane-dai 4-6-1, Minato-ku, Tokyo 108-8639, Japan*
[b] *IT Core Laboratories, Fujitsu Laboratories Ltd./CREST, JST., Kamikodanaka 4-1-1, Nakahara-ku, Kawasaki 211-8588, Japan*
[c] *Computational Biology Research Centre (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan*

Received 10 July 2006; received in revised form 23 November 2006; accepted 24 November 2006

Abstract

The development of a multicellular organism is a dynamic process. Starting with one or a few cells, the organism develops into different types of cells with distinct functions. We have constructed a simple model by considering the cell number increase and the cell-type order conservation, and have assessed conditions for cell-type diversity. This model is based on a stochastic Lindenmayer system with cell-to-cell interactions for three types of cells. In the present model, we have successfully derived complex but rigorous algebraic relations between the proliferation and transition rates for cell-type diversity by using a symbolic method: quantifier elimination (QE). Surprisingly, three modes for the proliferation and transition rates have emerged for large ratios of the initial cells to the developed cells. The three modes have revealed that the equality between the development rates for the highest cell-type diversity is reduced during the development process of multicellular organisms. Furthermore, we have found that the highest cell-type diversity originates from order conservation.

Keywords: Cell-type diversity; Lindenmayer system; Quantifier elimination; Algebraic computation

1. Introduction

In a multicellular organism, a single cell – an egg – or a group of cells develops into a certain pattern with a variety of cell types (Gilbert, 2003). These different cell types are created through cell differentiation, which starts with an initial type, and then cells change into several intermediate types before differentiating into the final type. The process of cell differentiation can be shown as a cell lineage. One representative of a real cell lineage is the development of blood cells, wherein a stem cell is capable of extensive proliferation, creating more stem cells as well as more differentiated cellular progeny.

The theoretical study of cell differentiation and morphogenesis was pioneered by Turing (1952), who showed that a reaction–diffusion system can produce an inhomogeneous, stable pattern. The concentrations of chemicals form a stripe or wave pattern, independently of the initial conditions, and this pattern formation process is robust against

* Corresponding author. Present address: Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan. Tel.: +81 92 642 7396; fax: +81 92 642 7396.
E-mail address: phiroshi@math.kyushu-u.ac.jp (H. Yoshida).

0303-2647/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2006.11.008

perturbations. Turing's theory provides the basis for a dynamic system for the morphogenesis and potentiality of cell differentiation. Embryogenesis with an *increase of cell numbers* was, however, not studied, and the intracellular dynamics were not sufficiently complex. In fact, resource chemicals are transported into the cell, and a complex catalytic reaction network within the cell changes the cell's state over time. Genes are expressed and repressed in response to these intracellular dynamics. Kauffman (1993) proposed that each cell-type should be regarded as an attractor of such intracellular dynamics, where each cell-type is represented as an attracting state of a genetic network. Again, morphogenetic processes with cell differentiation were not studied. By considering Turing's study and intracellular dynamics, together with the cell division process to increase the cell numbers, Kaneko and Yomo (1997, 1999) proposed *isologous diversification*. This allows spontaneous cell differentiation through cell division processes and cell-to-cell interactions. These studies have provided a basis for the cell-type diversity of a multicellular organism. However, the explicit relevance of the proliferation rates and the transition rates between cell types to cell-type diversity has not been studied.

Apart from the approach above, Lindenmayer system (abbreviated as L-system) is a parallel rewriting system that was introduced originally to model the development of multicellular organisms (Lindenmayer, 1968a,b). Indeed, an L-system is used to model the development process of various organisms (Yoshida et al., 2005c). Furthermore, stochastic aspects can be introduced into an L-system, termed a stochastic L-system (Eichhorst and Savitch, 1980; Eichhorst and Ruskey, 1981). The stochastic L-system can account for the influences of proliferation and transition rates, depending on the cell types.

The aim of this work is the derivation of rigorous algebraic relations between the proliferation and transition rates for high cell-type diversity with the conservation rule. For this purpose, we have constructed a model based on a stochastic L-system with interactions and have analysed it by using quantifier elimination (abbreviated as QE). The derivation allows us to understand the explicit algebraic relations between the cell-type order conservation rule and the high cell-type diversity of multicellular organisms.

The present paper is organized as follows. First, in Section 2, we provide a brief overview of our previous model and results (Yoshida et al., 2005b), wherein the cell-type order conservation rule appeared spontaneously. In Section 3, we introduce a model of a multicellular organism consisting of one-dimensional cells. This model postulates the cell-type order conservation rule as interaction terms. We briefly explain the QE method in Section 4. The results of the algebraic computation by using QE are given in Section 5, which describes the rigorous algebraic relations between the proliferation and transition rates. The growth matrix of the model analysed in this study is described in Section 5.1, and some features of the growth matrix are discussed in Section 5.2. In Section 5.3, the rigorous relations analysed by QE are presented, and based on the relations, the conditions for the highest cell-type diversity are scrutinized in Sections 5.4 and 5.5. Lastly, we summarize this work.

2. Background

In this section, we briefly review of our previous work (Yoshida et al., 2005b), which is the basis for the construction and analysis of the model in this work.

In a multicellular organism, a single cell – an egg – develops correctly into a prospectively determined pattern. This morphogenesis is robust against environmental perturbations, and the same pattern is always generated from an egg. In other words, recursive production is repeated. At the same time, the developmental process in a multicellular organism produces a variety of cell types. The compatibility of these two points is surprising, because 'recursive production' is the reproduction of the same pattern of an individual cell, while 'cell-type diversity' is the existence of various patterns, namely various cell types, within an individual. The question we addressed in our previous work was the selection of initial cell(s), to allow for compatibility between recursive production and cell-type diversity.

We present our previously developed model of a multicellular organism in Fig. 1. Within each cell, catalytic and autocatalytic chemical reactions maintain the cell itself and synthesize some chemicals for the cell membrane.

Our numerical results indicated that by starting with an initial object consisting of both the chaotic cell-type with diverse chemicals and the regular-dynamics cell-type with less chemical diversity, the recursive production of a multicellular organism with cell-type diversity has been realized. In addition to recursive production, a remarkable regeneration pattern, which is analogous to the intercalary regeneration in cockroach legs (see Fig. 2), and planarian and salamander limb blastema (Gilbert, 2003), was observed in our previous work (Yoshida et al., 2005b). There, starting

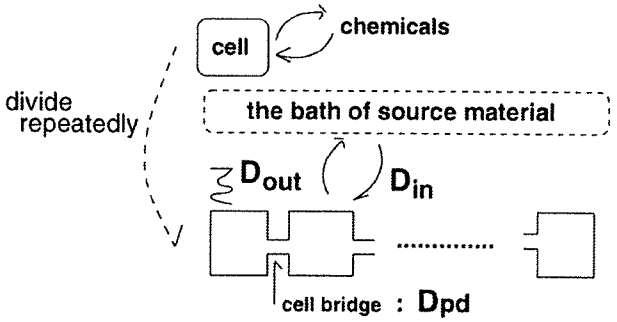

Fig. 1. Schematic representation of our previous model. The cells are surrounded by a bath of source material with a constant concentration. After a division, the cells are connected to one another by forming a cell bridge. The cells are thus connected to one another as a one-dimensional chain.

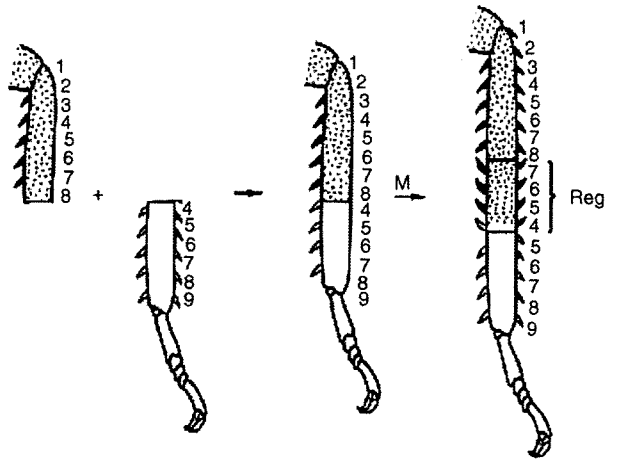

Fig. 2. Intercalary regeneration in cockroach legs (Alberts et al., 2002). When mismatched portions of the growing legs are grafted together, new tissue is intercalated to fill in the gap so that the noncontiguous positional values disappear.

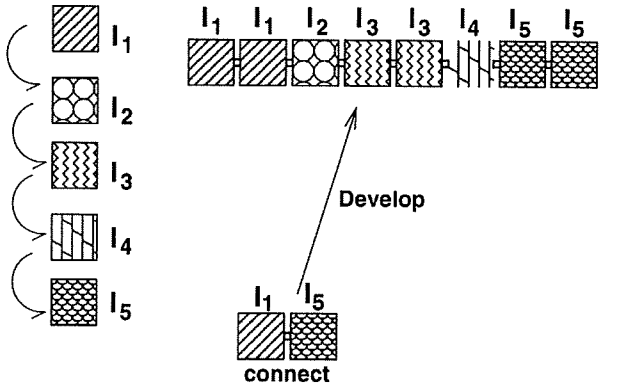

Fig. 3. Regeneration of cell-type sequence, as observed in (Yoshida et al., 2005b). The cell differentiates from I_1 to I_5 sequentially. Starting with $I_1 I_5$, patterns without noncontiguous numbers, such as $I_1 I_1 I_2 I_3 I_3 I_4 I_5 I_5$, are eventually produced. Thus, noncontiguity will disappear during the development process.

with the two cells corresponding to I_1 and I_n, the regeneration pattern corresponding to $I_1 I_2 \ldots I_n$ was eventually produced, as illustrated in Fig. 3.

3. Model

Now, we present a simple model of a multicellular organism in which the cell lineage can be represented as a line, that is, only sequential differentiation occurs. Our model is schematically illustrated in Fig. 4. We assume

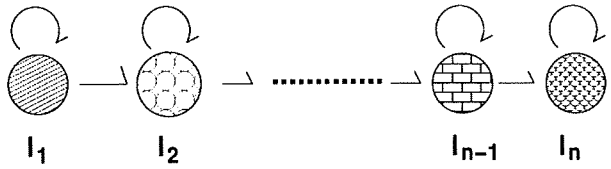

Fig. 4. Schematic representation of our model. Cell differentiation proceeds as follows: $I_1 \rightarrow I_2 \rightarrow \ldots \rightarrow I_n$.

that cell differentiation starts with an initial type, I_1, and then the cell differentiates into several intermediate types $I_2 \rightarrow I_3 \rightarrow \ldots \rightarrow I_{n-1}$ before differentiating into the final type, I_n. The regeneration phenomena mentioned in the previous section can be described as the following rewriting rule, named the *cell-type order conservation rule*:

$$I_iI_j \rightarrow I_iI_{i+1}\cdots I_{j-1}I_j,\ I_jI_i \rightarrow I_jI_{j-1}\cdots I_{i+1}I_i, \quad j > i+1. \tag{1}$$

The proliferation and transition rates of cell-type i ($1 \leq i \leq n$) are defined as follows:

$$I_i \rightarrow \begin{cases} I_iI_i & p_{i,i} \\ I_{i+1} & p_{i,i+1} \\ I_i & 1-p_{i,i}-p_{i,i+1} \end{cases}, \quad 1 \leq i < n, \qquad I_n \rightarrow \begin{cases} I_nI_n & p_{n,n} \\ I_n & 1-p_{n,n} \end{cases} \tag{2}$$

with $0 \leq p_{i,i} < 1$ ($1 \leq i \leq n$), $0 < p_{i,i+1} < 1$ ($1 \leq i < n$), $p_{i,i} + p_{i,i+1} < 1$ ($1 \leq i < n$). In addition to the above rewriting rules, we further adopt rewriting rules that appear as interaction terms and describe the cell-type order conservation rule: (1): $I_iI_j \rightarrow I_iI_{i+1}\cdots I_{j-1}I_j$, $I_jI_i \rightarrow I_jI_{j-1}\cdots I_{i+1}I_i$ ($j > i+1$), which guarantees the contiguity of cell types.

4. Analytical method

The key point in this work is the usage of QE, which is one of the main subjects in computer algebra (Caviness and Johnson, 1998). In general, QE deals with first-order formulae, which consist of polynomial equations, inequalities, quantifiers ($\exists$, $\forall$) and Boolean operators. QE computes an equivalent quantifier-free formula for a given first-order formula over the real closed field. For example, for the input ${}^{\forall}x(x^2 + bx + c > 0)$, QE outputs the equivalent quantifier-free formula $b^2 - 4c < 0$. It follows from this that we can obtain a condition for unquantified variables that makes the input formula true by QE.

We can also obtain the maximum value of an objective polynomial under certain constraints by adding one extra variable, ϵ, which is assigned to the objective polynomial. For instance, we can transform a problem: max y s.t. $x^2 + y^2 \leq 1$ and $y \leq x^2$ into the following form:

$${}^{\exists}x{}^{\exists}y(x^2 + y^2 \leq 1 \wedge y \leq x^2 \wedge y \geq \epsilon).$$

For this formula, QE outputs $\epsilon \leq (\sqrt{5}-1)/2$, which shows that the maximum value of y is $(\sqrt{5}-1)/2$. Recently, by using this ability, we performed a symbolic–numeric optimization for the biochemical kinetic model (Orii et al., 2005; Anai and Horimoto, 2006) and an algebraic computation for the multicell development model (Yoshida et al., 2005a, 2006).

5. Results and discussion

5.1. Growth matrix in a stochastic L-system

We calculate the *growth matrix* M of the two contiguous cell types I_iI_i, I_iI_{i+1}, $I_{i+1}I_i$ ($1 \leq i < n-1$), which enables us to estimate the composition of $I_\ell I_k$ ($k = \ell - 1, \ell, \ell + 1$) at step m. It should be noted that the other two contiguous cell types (e.g., I_iI_{i+3}) never appear at any step, by virtue of the cell-type order conservation rule. Although we could use a growth matrix of more than two, for simplicity we have calculated the simple growth matrix with the two contiguous cell types. If one starts with I_1I_1, then the composition at step m can be calculated generally by the following formula:

$$(1, 0, 0, \ldots)M^m. \tag{3}$$

Here, we have studied the case of $n = 3$, showing the existence of three cell types. For the sake of simplicity, let A, B and C denote I_1, I_2 and I_3, respectively, in what follows. In this case ($n = 3$), the growth matrix M is:

$$\begin{pmatrix} 2p_{1,1}+(1-p_{1,2})^2 & (1-p_{1,2})p_{1,2} & (1-p_{1,2})p_{1,2} & p_{1,2}^2 & 0 & 0 & 0 \\ p_{1,1} & 1-p_{1,2} & 0 & p_{1,2}+p_{2,2}-p_{1,2}p_{2,3} & p_{2,3} & 0 & 0 \\ p_{1,1} & 0 & 1-p_{1,2} & p_{1,2}+p_{2,2}-p_{1,2}p_{2,3} & 0 & p_{2,3} & 0 \\ 0 & 0 & 0 & 2p_{2,2}+(1-p_{2,3})^2 & (1-p_{2,3})p_{2,3} & (1-p_{2,3})p_{2,3} & p_{2,3}^2 \\ 0 & 0 & 0 & p_{2,2} & 1-p_{2,3} & 0 & p_{2,3}+p_{3,3} \\ 0 & 0 & 0 & p_{2,2} & 0 & 1-p_{2,3} & p_{2,3}+p_{3,3} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1+2p_{3,3} \end{pmatrix},$$

with its eigenvalues:

$$1-p_{1,2},\ 1+2p_{1,1}-p_{1,2},\ (1-p_{1,2})^2,\ 1-p_{2,3},\ 1+2p_{2,2}-p_{2,3},\ (1-p_{2,3})^2 \text{ and } 1+2p_{3,3}. \tag{4}$$

Note that three of the seven eigenvalues, $1+2p_{1,1}-p_{1,2}$, $1+2p_{2,2}-p_{2,3}$ and $1+2p_{3,3}$, can be greater than 1.

5.2. Features of the growth matrix analysed in this study

Let S denote the diagonal matrix: $\mathrm{Diag}(1-p_{1,2}, 1+2p_{1,1}-p_{1,2}, (1-p_{1,2})^2, 1-p_{2,3}, 1+2p_{2,2}-p_{2,3}, (1-p_{2,3})^2, 1+2p_{3,3})$. The features of the growth matrix M are as follows.

If the eigenvalues differ from one another, then M can be divided into PSP^{-1}, where P is a regular matrix. In this case, $(1,0,0,0,0,0,0)P$ is

$$(0, 2, e_3, 0, e_5, e_6, e_7),$$

where e_3, e_5, e_6 and e_7 are nonzero values. These facts lead to the composition (3): $(1,0,0,0,0,0,0)PS^mP^{-1}$ to

$$(0, 2(1+2p_{1,1}-p_{1,2})^m, e_3(1-p_{1,2})^{2m}, 0, e_5(1+2p_{2,2}-p_{2,3})^m, e_6(1-p_{2,3})^{2m}, e_7(1+2p_{3,3})^m)P^{-1}.$$

As m approaches infinity, the composition above can be described as follows:

$$(0, 2(1+2p_{1,1}-p_{1,2})^m, 0, 0, e_5(1+2p_{2,2}-p_{2,3})^m, 0, e_7(1+2p_{3,3})^m)P^{-1}.$$

The second, fifth and seventh rows of P^{-1} are $(f_1, f_2, f_2, f_3, f_4, f_4, f_5)$, $(0,0,0,g_1,g_2,g_2,g_3)$ and $(0,0,0,0,0,0,1)$, respectively, where f_i $(1 \le i \le 5)$ and g_i $(i = 1, 2, 3)$ are nonzero values. Finally, as m approaches infinity, the composition approaches

$$\begin{aligned}(AA, AB, BA, BB, BC, CB, CC) &= (1,0,0,0,0,0,0)PS^mP^{-1} \\ &= 2(1+2p_{1,1}-p_{1,2})^m(f_1, f_2, f_2, f_3, f_4, f_4, f_5) \\ &\quad + e_5(1+2p_{2,2}-p_{2,3})^m(0,0,0,g_1,g_2,g_2,g_3) \\ &\quad + e_7(1+2p_{3,3})^m(0,0,0,0,0,0,1).\end{aligned}$$

Therefore, only the second eigenvalue, $1+2p_{1,1}-p_{1,2}$, can give rise to AA, AB, BA cell types as m approaches infinity. This indicates one of the necessary conditions, that AA, AB, BA, BB, BC, CB and CC are mingled as m approaches infinity:

$$1+2p_{1,1}-p_{1,2} > 1 \wedge 1+2p_{1,1}-p_{1,2} > 1+2p_{2,2}-p_{2,3} \wedge 1+2p_{1,1}-p_{1,2} > 1+2p_{3,3}. \tag{5}$$

Under the condition (5), let m approach infinity. In other words, as the chain of cells becomes sufficiently long, the composition (3) of $n = 3$ becomes the following:

$$N(AB) = N(BA) = \frac{\gamma(p_{1,2}-p_{2,3})(1-p_{1,2}-p_{2,3})}{\gamma(p_{1,2}-p_{2,3})+p_{2,3}}, \qquad N(BC) = N(CB) = \gamma N(AB),$$

$$N(BB) = N(CC) = \gamma,$$

$$
\begin{aligned}
p_{1,1} &= \frac{p_{1,2}(1-p_{1,2})(p_{2,3}+\gamma(p_{1,2}-p_{2,3}))}{2\gamma(p_{1,2}-p_{2,3})(1-p_{1,2}-p_{2,3})},\\
p_{2,2} &= \frac{\begin{array}{l}p_{1,2}p_{2,3}(-(1-p_{1,2})p_{1,2}^2+p_{2,3}-p_{1,2}p_{2,3}^2)+(p_{1,2}^4(3-5p_{2,3})-(2-p_{2,3})(1-p_{2,3})p_{2,3}^3\\ -p_{1,2}^5(1-2p_{2,3})+p_{1,2}p_{2,3}^2(-1+2(2-p_{2,3})(1-p_{2,3})p_{2,3})+p_{1,2}^2p_{2,3}(5-9p_{2,3}+6p_{2,3}^2)\\ -p_{1,2}^3(2+p_{2,3}-7p_{2,3}^2+4p_{2,3}^3))\gamma+(p_{1,2}-p_{2,3})^2(1-p_{2,3})p_{2,3}(2-p_{1,2}-p_{2,3})\gamma^2\end{array}}{2(p_{1,2}-p_{2,3})(-1+p_{1,2}+p_{2,3})\gamma((-1+p_{1,2})p_{1,2}-p_{2,3}^2+(p_{1,2}-p_{2,3})(-2+p_{1,2}+p_{2,3})\gamma)},\\
p_{3,3} &= \frac{\begin{array}{l}p_{2,3}((1-p_{1,2})p_{1,2}p_{2,3}-(p_{1,2}-p_{2,3})(p_{1,2}^2+(1-p_{2,3})p_{2,3}^2-p_{1,2}(1+p_{2,3}+p_{2,3}^2))\gamma\\ -(p_{1,2}-p_{2,3})^2(2-p_{1,2}-p_{2,3})(1-2p_{1,2}+p_{2,3})\gamma^2)\end{array}}{2(p_{1,2}-p_{2,3})(-1+p_{1,2}+p_{2,3})\gamma((-2+p_{1,2})p_{1,2}\gamma-p_{2,3}(1-(2-p_{2,3})\gamma))},
\end{aligned}
\tag{6}
$$

where $N(XY)$ denotes the number of sequence: XY as m approaches infinity and γ denotes that the ratio of the initial cells to the developed cells is $1/\gamma$. In the equations above (6), furthermore, $N(AA)$ is normalized, i.e., $N(AA)=1$, and the following constraints are assumed:

$$\gamma N(AA)=N(BB)=N(CC)\wedge\gamma N(AB)=N(BC). \tag{7}$$

In summary, $N(XY)$, $(X, Y\in\{A, B, C\})$, $p_{1,1}$, $p_{2,2}$ and $p_{3,3}$ can explicitly be represented as functions of $p_{1,2}$ and $p_{2,3}$. Notice that $N(AB)=N(BA)$ and $N(BC)=N(CB)$ always hold true, because of the construction of the rewriting rules (2), and that the cell-type diversity becomes highest as $N(AB)(=N(BA))$ approaches 1.

5.3. Relations between proliferation and transition rates in the highest diversity of cell types by QE analysis

We now investigate the relations between the proliferation and transition rates in the highest cell-type diversity. For this purpose, we have calculated the relations that maximize $N(AB)$ (or $N(BC)$) under the constraints of (5)–(7). It may be worth noting that it seems difficult to estimate rigorous relationships between the rates under such complicated constraints by the existing numerical methods. Actually, in our previous analysis by the numerical method, we estimated a set of rates that realize high cell-type diversity by searching a huge number of points over the five-dimensional rate space, but obtained no relations between the rates. Although the rate values provide a snapshot for the system behaviour, the relation between the rates will provide more profound insights into the mechanism of the system to analyse. Here, we have utilized the QE approach to obtain rigorous rate relations.

First, we determined the maximum values of $N(AB)$, which designate the highest diversity of cell types. In the end, the determination is reduced by solving the following QE problem:

$$^{\exists}p_{1,2}{}^{\exists}p_{2,3}\,(\psi(p_{1,2}, p_{2,3}, \gamma)\wedge N(AB)\geq\epsilon), \tag{8}$$

where $\psi(p_{1,2}, p_{2,3}, \gamma)$ is a formula derived by combining all equations and inequalities appearing in (5)–(7), conjunctively. For a fixed value of γ, the QE procedure (8) outputs the following inequalities:

$$\epsilon\leq\frac{\sqrt{881}-9}{40}\sim 0.517041 \quad\text{and}\quad \epsilon\leq\frac{\sqrt{89801}-99}{400}\sim 0.50167$$

when the γ values are 10 and 100, respectively. As seen in Section 4, from these inequalities, we can determine the maximum values of $N(AB)$ as follows:

$$(AA, AB, BA, BB, BC, CB, CC)=(1, f(\gamma), f(\gamma), \gamma, \gamma f(\gamma), \gamma f(\gamma), \gamma) \tag{9}$$

with $f(10)=(\sqrt{881}-9)/40$ and $f(100)=(\sqrt{89801}-99)/400$. Thus, by the QE method, we have obtained the exact maximum value by effectively pruning a huge number of candidates for the maximum.

By using the conditions for the maximum values obtained above, the QE method has enabled us to obtain rigorous algebraic relations between the proliferation and transition rates: $\mathcal{P}$, as follows:

$$^{\exists}\epsilon\,(\psi(\mathcal{P}, \gamma)\wedge N(AB)\geq\epsilon).$$

Interestingly, three modes for the highest diversity of cell types emerge. For instance, the relation between $p_{1,2}$ and $p_{2,3}$ can be obtained as follows:

$$^{\exists}\epsilon\,(\psi(p_{1,2}, p_{2,3}, \gamma) \wedge N(AB) \geq \epsilon).$$

The three modes when γ is 10 are expressed rigorously as follows.

- Mode I:

$$p_{2,3} = \text{the minimum real root of the equation in } x,$$
$$190p_{1,2}^2 - 490p_{1,2}^3 + 200p_{1,2}^4 + (-391p_{1,2} + 681p_{1,2}^2 - 100p_{1,2}^3)x + (200 + 120p_{1,2} - 310p_{1,2}^2)x^2$$
$$+(-310 + 100p_{1,2})x^3 + 110x^4 = 0\,(0 < p_{1,2} < p_0),$$

  where p_0 is exactly the minimum real root of the equation in x,

$$399 - 3274x + 9188x^2 - 10232x^3 + 3920x^4 = 0,$$

  and is approximately 0.293122.
- Mode II:

$$p_{2,3} = \frac{1 + 18p_{1,2} - \sqrt{1 + 36p_{1,2} - 76p_{1,2}^2}}{20}, \quad p_0 \leq p_{1,2} < \frac{2}{5}$$

- Mode III:

$$p_{2,3} = \frac{20 - 9p_{1,2} - \sqrt{400 - 1960p_{1,2} + 2481p_{1,2}^2}}{40}, \quad 0 < p_{1,2} \leq \frac{2}{5}$$

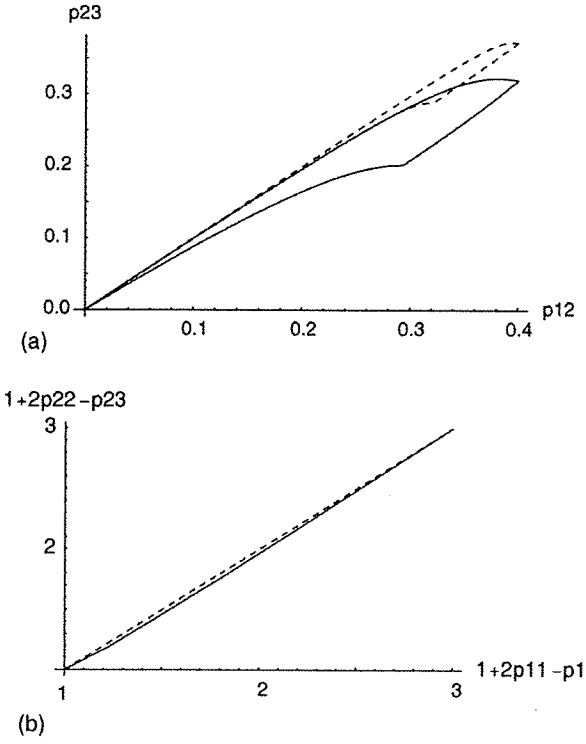

Fig. 5. Rigorous relations when the maximum values are satisfied. The solid and broken lines denote the relations when the γ values are 10 and 100, respectively, *with* the cell-type order conservation rule. (a) Relations between $p_{1,2}$ and $p_{2,3}$. Modes I–III correspond to the two curves (lines) into which the points where the curve is not smooth separate the whole region. Mode I includes the origin. (b) Relations between $1 + 2p_{1,1} - p_{1,2}$ and $1 + 2p_{2,2} - p_{2,3}$.

Notice that Modes I–III, described by the rigorous algebraic functions of $p_{1,2}$, have been derived using the QE method.

5.4. Characteristic conditions for the proliferation and transition rates

Fig. 5 shows the relations obtained when the γ values are set to 10 and 100. These values were chosen because in our previous simulation (Yoshida et al., 2005b), the constraint (7) over $N(XY)$, $(X, Y \in \{A, B, C\})$ was observed, and partly because there are few initial-type cells (A in this work) corresponding to stem cells in real biological systems (Gilbert, 2003). Remember that $1/\gamma$ was defined as the ratio of the initial cells to the developed cells in Section 5.2.

As Fig. 5 (a) shows, the three modes are contracted to approach the line of $p_{1,2} = p_{2,3}$, from a comparison between the modes with γ values of 10 and 100. Although the three modes do not disappear even with a large γ value, the relation between $p_{1,2}$ and $p_{2,3}$ can be approximated as:

$$p_{1,2} = p_{2,3}, \tag{10}$$

when γ is sufficiently large.

Another explicit relation between $(1 + 2p_{1,1} - p_{1,2})$ and $(1 + 2p_{2,2} - p_{2,3})$ is further obtained when the values of γ are set to 10 and 100, as shown in Fig. 5 (b). The $(1 + 2p_{1,1} - p_{1,2}, 1 + 2p_{2,2} - p_{2,3})$ curve when $\gamma = 100$ is closer to the line $1 + 2p_{1,1} - p_{1,2} = 1 + 2p_{2,2} - p_{2,3}$ than that when $\gamma = 10$. Thus, the following relation is observed:

$$1 + 2p_{1,1} - p_{1,2} \sim 1 + 2p_{2,2} - p_{2,3}. \tag{11}$$

By considering the relation in Eq. (10): $p_{1,2} = p_{2,3}$, the Eq. (11) indicates that $p_{1,1}$ approaches $p_{2,2}$ when γ becomes sufficiently large.

We will translate the above relations between the rates into biological terms. First, the rate relation in Eq. (10) indicates that the initial and final transition rates are almost equal at the highest cell-type diversity, on the assumption

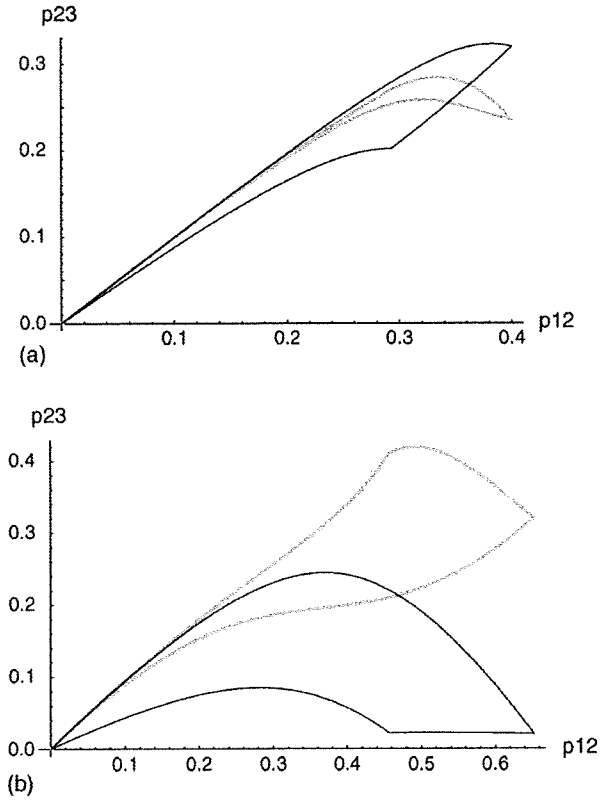

Fig. 6. Relations between the points that are lowered from the highest cell-type diversity curve by 0.001. The grey lines designate these relations. The black lines designate the original curves in which the cell-type diversity is the highest. (a) *With* the cell-type order conservation rule. (b) *Without* the conservation rule.

that the developed tissue is composed of a few initial cell types. In other words, the equality of the transition rates implies that the cell-type diversity is realized by the similar rates for transition between distinctive cell types. In addition, the assumption may be naturally accepted by the fact that the number of stem cells is very small. Secondly, the proximity of $(1 + 2p_{1,1} - p_{1,2})$ and $(1 + 2p_{2,2} - p_{2,3})$ indicates that the proliferation rates of the initial cell-type and of the next-developed type are also almost equal, $p_{1,1} = p_{2,2}$. This indicates that the distinctive cells increase with similar rates. The similar degree of cell increase in distinctive types may be responsible for the cell-type diversity.

5.5. Relation between cell-type diversity and the conservation rule

To evaluate the relation between cell-type diversity and the cell-type order conservation rule, we have calculated the difference between the rate relations of the highest diversity and lower diversity in the cases described above, with and without the conservation rule. In the evaluation, γ is set to 10.

Fig. 6 shows the rate relation differences with and without the conservation rule. In Fig. 6 (a), the rate relation with the conservation rule at the highest diversity is similar to that for the lower diversity. Indeed, although the forms of the three modes in the two cases are slightly different from each other, the calculated values of $p_{1,2}$ and $p_{2,3}$ are in similar ranges. In contrast, the forms and the values of $p_{1,2}$ and $p_{2,3}$ are quite different from each other without the conservation rule in Fig. 6 (b). In other words, the rates show distinctive relations without the conservation rule, depending on the degree of diversity. Furthermore, the form and the range of the highest diversity in Fig. 6 (a and b) are also quite different, with and without the conservation rule. In particular, the rate relation without the conservation rule is far from the relation of $p_{1,2}$ and $p_{2,3}$ that is observed with the conservation rule: in Fig. 6 (b), $p_{1,2}$ is larger than $p_{2,3}$, which indicates that the initial transition rate is faster than the final transition rate. Intuitively, it is natural to conclude that the bias of the transition rates may not be responsible for the cell-type diversity. At any rate, the conservation rule is prerequisite to the realization of cell-type diversity in this model.

6. Conclusion

In the present work, the relation between the proliferation and transition rates in the highest cell-type diversity has been investigated, based on the L-system with interactions. Remarkably, the rigorous algebraic relations have been derived with the aid of quantifier elimination. Indeed, the two rates for proliferation and transition are almost equal to each other in the highest cell-type diversity, on the assumption that the number of initial cells is very small, which implies that the similar transition rates between distinctive cell types are prerequisite for high cell-type diversity. Furthermore, cell-type order conservation is a prerequisite to realizing high cell-type diversity. Although three cell types were assumed in the analysed model, the present approach of discrete model and algebraic computation will shed some light on the mechanism of cell-type diversity within actual multicellular organisms.

Acknowledgments

We wish to express our gratitude to Professor Takashi Yokomori and Professor Kunihiko Kaneko for helpful suggestions and useful discussions. One of the authors (K.H.) was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas "Systems Genomics" (grant 18016008) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. Molecular Biology of the Cell, fourth ed. Garland Science.

Anai, H., Horimoto, K., 2006. Symbolic–numeric estimation of parameters in biochemical models by quantifier elimination. J. Bioinform. Comput. Biol. 4, 1097–1107.

Caviness, B.F., Johnson, J.R., 1998. Quantifier Elimination and Cylindrical Algebraic Decomposition. Springer-Verlag, Wien.

Eichhorst, P., Ruskey, F., 1981. On unary stochastic Lindenmayer systems. Inf. Control 48 (1), 1–10.

Eichhorst, P., Savitch, W.J., 1980. Growth functions of stochastic Lindenmayer systems. Inf. Control 45 (3), 217–228.

Gilbert, S.F., 2003. Developmental Biology, seventh ed. Sinauer Associates.

Kaneko, K., Yomo, T., 1997. Isologous diversification: a theory of cell differentiation. Bull. Math. Biol. 59 (1), 139–196. doi:10.1006/S0092-8240(96)00044-4.