Integrative Clustering of Cancer Genome Data using Infinite Relational Models

Yoichi Chikahara, Atsushi Niida, Rui Yamaguchi, Seiya Imoto and Satoru Miyano
Human Genome Center, the Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
(ychika, aniida, ruiy, imoto, miyano)@ims.u-tokyo.ac.jp

Abstract

From the late 90's until today, the advances in high-throughput measurement technologies are remarkable and producing a huge amount of cancer genomic data. Due to the complexity of data, however, we have not still got a fully integrated view of genetic and transcriptional changes that differ among individuals. To visualize the differences in genetic and transcriptional data among patient samples, we focus on grouping of three types of features, i.e., genes, patient samples, and expression modules. We propose an integrative framework based on the biclustering of multiple types of biological data, i.e., copy number, gene expression, and module activity, by extending the Infinite Relational Models (IRM), a non-parametric Bayesian model used to perform a biclustering of binary data, for continuous data. We demonstrate an utility of the model using a colorectal cancer (CRC) dataset. Our result discovers a clinical insight that the activity of modules related to an immune system is associated with CRC patients survival, which demonstrates the ability of our novel integrative approach to group not only genes and modules but also patient samples based on their genetic and transcriptional alterations.

keywords: Cancer Systems Biology, Non-parametric Bayesian Model, Infinite Relational Model (IRM), Tensor Data Clustering, Personalized Medicine

1 Introduction

Massive datasets on cancer generated by advanced high-throughput technologies have raised a research problem: how we can obtain biological insights by an integrative analysis of multiple types of biological data, especially, DNA copy number and mRNA expression. The aim of the integrative analysis can be separated mainly into two [2]. One is to reveal the target genes of particular genes, e.g., driver gene. Driver genes are defined as genes driving the tumorigenesis by their genetic alterations, e.g., changes in DNA copy number. To elucidate the gene regulatory mechanism of cancer, Akavia, *et al.* [1] have grouped genes in order to obtain an expression module associated with driver genes, i.e., targets of driver genes showing a coherent expression. The other is to discover tumor subtypes defined as subgroups of patient samples that are characterized by concordant genetic and transcriptional changes. To identify tumor subtypes, Mo, *et al.* [4] have performed the clustering of patient samples. Patient samples belonging to the same tumor subtypes may exhibit a similar prognosis or drug responses, therefore, the approach is important in terms of the personalized medicine.

In this study, we aimed to achieve the above two goals simultaneously and proposed a novel integrative approach considering not only genes and expression modules but also the diversity in gene signatures across patient samples. We aimed to group three types of features simultaneously, i.e., genes, patient samples, and expression modules by integrating three types of biological data, i.e., DNA copy number, mRNA expression, and module activity. For this goal, we developed a model-based clustering method, which is based on the IRM [3], a non-parametric Bayesian model. The IRM was proposed for the discovery of hidden relationship among objects by simultaneous clustering of multiple features in tensor data without the number of clusters given a priori. Although IRM was originally developed for binary data representing the existence of relationship between objects, we developed an extended version of the model to allow us to analyze the continuous biological data.

This paper is organized as follows. Section 2 describes the datasets of CRC patients used in this work, the mathematical model for the simultaneous clustering of tensor data, and the method to identify clusters representing the upstream driver genes and the downstream expression modules from the clustering results. Section 3 shows the performance of the proposed model and demonstrates that the obtained module cluster can possibly be the prognostic factor candidates for CRC patients survival. Finally, Section 4 discusses the

advantages of our approach comparing with the related works. Figure 1 shows the overall framework of our proposed method described in Section 2. We expect that this novel approach will be a powerful tool for the identification of further unexpected relationships between genes, samples, and modules across trans-omics data.

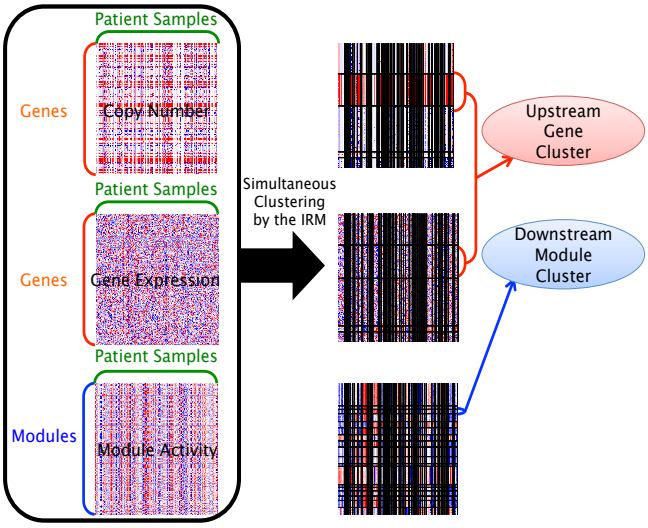

Figure 1: Outline of our proposed method

2 Materials and Methods

2.1 Data and Preprocessing

In order to prepare the three types of biological data matrices, i.e., DNA copy number, mRNA expression, and expression module activity of CRC patients, we used DNA copy number and mRNA expression data of CRC patients [1] downloaded from the TCGA data portal (`https://tcga-data.nci.nih.gov/tcga/` [6]). Note that the former two types of matrix data have elements of *gene* × *sample* and the last one have elements of *module* × *sample*.

For DNA copy number data matrix, we transformed the values to the logarithmic scale, converted the probe set IDs to gene symbols, and prepare the segmented values across chromosomes to genes on the chromosomal loci.

For mRNA expression data matrix, after converting the values to the logarithmic scale and the probe set IDs to gene symbols, we normalized the data to visualize the difference among patient samples. The normalization of the matrix data was performed so that each row has mean 0 and variance 1. Then, we selected the top 2000 genes with the highest variances of expression profiles. Therefore, we set the number of rows in copy number and gene expression data 2000.

After transforming the expression values to the logarithmic scale and the probe IDs to gene symbols as a data preprocessing, we applied the Extraction of Expression Modules (EEM) method [7] to the gene expression data to identify expression modules and to obtain the activity values of the expression modules. An expression module is defined as a group of genes, e.g., genes harboring the same cis-regulatory motif, in which the members of genes coherently express across samples. The *module activity* of an expression module for each sample is defined as the averaged expression value among the coherently expressed genes; it well represents an activity of biological process contributed by those genes. For the obtained module activity data, we performed the normalization as well as that of the mRNA expression data.

2.2 Clustering using IRM

Integrating three types of biological data matrices, we simultaneously organized three types of features, i.e., genes, samples, and modules, into clusters with similar characteristic each other using the IRM.

2.2.1 Target Distribution of Gibbs Sampling

Let $R^1, R^2 \in \mathbb{R}^{N^1 \times N^2}$, and $R^3 \in \mathbb{R}^{N^3 \times N^2}$ be data matrices of DNA copy number, mRNA expression, module activity, respectively, where N^1, N^2, and N^3 are the numbers of genes, samples, and expression modules, respectively. To represent cluster assignments of the N^1 genes, the N^2 samples, and the N^3 modules, let z^1, z^2, z^3 be vectors of length N^1, N^2, N^3, respectively, where $z^1_i \in \{1, \cdots, C^1\}$ for $i = 1, \cdots, N^1$, $z^2_j \in \{1, \cdots, C^2\}$ for $j = 1, \cdots, N^2$, and $z^3_m \in \{1, \cdots, C^3\}$ for $m = 1, \cdots, N^3$. Note that C^1, C^2, and C^3 are the number of clusters assigned to each gene, sample, and module. Here, our goal is to infer the cluster assignments that maximize the posterior distribution $P(z^1, z^2, z^3 | R^1, R^2, R^3) \propto P(R^1, R^2, R^3, z^1, z^2, z^3)$. Assuming that R^1, R^2, R^3 are conditionally independent given the cluster assignments z^1, z^2, z^3 and that z^1, z^2, z^3 are independent, this distribution can be factorized as:

$$P(R^1, R^2, R^3, z^1, z^2, z^3) = \prod_{t=1}^{3} P(R^t | z^1, z^2, z^3) \prod_{d=1}^{3} P(z^d)$$

[1] `https://tcga-data.nci.nih.gov/docs/publications/coadread_2012/`

Here, note that copy number and gene expression matrix, R^1, R^2, are dependent on cluster assignments of genes and samples, z^1, z^2 and that module activity matrix, R^3, is dependent on cluster assignments of modules and samples, z^3, z^2.

We first describe how we can obtain a cluster assignment of the i-th gene. Let R^1_i and R^2_i be i-th row vectors extracted from R^1, R^2 and Θ be a set of parameters. Fixing cluster assignments of genes other than the i-th one, z^1_{-i}, and a vector of sample cluster assignments, z^2, the conditional distribution on the i-th gene's cluster assignment $P(z^1_i = k|z^1_{-i}, z^2, R^1_i, R^2_i, \Theta)$ can be written as:

$$\begin{aligned} & P(z^1_i = k|z^1_{-i}, z^2, R^1_i, R^2_i, \Theta) \\ & \propto \quad P(z^1_i = k|z^1_{-i}) \prod_{t=1}^{2} P(R^t_i|z^1_i = k, z^1_{-i}, z^2, \Theta) \quad (1) \end{aligned}$$

where $k \in \{1, \cdots, C^1\}$ is a cluster index for genes. For the assignment of a cluster index to the i-th gene for each i, we just draw a sample from this distribution performing the Gibbs sampling. The first term represents the prior on the i-th gene's cluster assignment and the second one does the likelihood of R^1_i, R^2_i.

In the same way, the conditional distribution on the assignment of the j-th sample and the m-th module can be given as:

$$\begin{aligned} & P(z^2_j = l|z^2_{-j}, z^1, z^3 R^1_j, R^2_j, R^3_j, \Theta) \\ & \propto \quad P(z^2_j = l|z^2_{-j}) \prod_{t=1}^{3} P(R^t_i|z^2_j = l, z^2_{-j}, z^1, z^3, \Theta) \quad (2) \\ & P(z^3_m = p|z^3_{-m}, z^2, R^3_m, \Theta) \\ & \propto \quad P(z^3_m = p|z^3_{-m}) P(R^3_m|z^3_m = p, z^3_{-m}, z^2, \Theta) \quad (3) \end{aligned}$$

where $l \in \{1, \cdots, C^2\}$, $p \in \{1, \cdots, C^3\}$ are cluster indices for samples and modules.

2.2.2 Prior on Cluster Assignments

Using the IRM, we can get the optimal cluster assignments for each data without the numbers of clusters given a priori. For that purpose, we need to select the prior on cluster assignments that should encourage the model to select the number of clusters based on the data. In case of IRM, the Chinese Restaurant Process (CRP, [8]) is used as the prior on cluster assignments. Under the CRP, the distribution over cluster assignments of N objects can be given as:

$$P(z_i = k|z_{-i}, N, \alpha) = \begin{cases} \frac{N_k}{N-1+\alpha} & (N_k > 0) \\ \frac{\alpha}{N-1+\alpha} & (N_k = 0) \end{cases} \quad (4)$$

where N_k is the number of objects already assigned to the k-th cluster, and α is a parameter ($\alpha > 0$).

2.2.3 Generative Models for Continuous Data

In order to analyze continuous matrix data, referring to the previous work on nonparametric Bayesian models for continuous data [9], we introduced a generative model based on the IRM.

We can decompose the likelihood on the i-th gene over samples as:

$$\begin{aligned} & \prod_{t=1}^{2} P(R^t_i|z^1_i = k, z^1_{-i}, z^2, \Theta) \\ & = \prod_{t=1}^{2} \prod_{l=1}^{C^2} \prod_{j:\ z^2_j = l} P(R^t_{i,j}|z^1_i = k, z^2_j = l, z^1_{-i}, z^2_{-j}, \Theta) \end{aligned} \quad (5)$$

The likelihood on the j-th samples over genes and modules, $\prod_{t=1}^{3} P(R^t_j|z^2_j = l, z^2_{-j}, z^1, z^3, \Theta)$, and the likelihood on the m-th module over samples, $P(R^3_m|z^3_m = p, z^3_{-m}, z^2, \Theta)$, can be given in a similar way.

For these formulation, we can simply consider a generative model for the t-th type of matrix data, $R^t_{i,j}$ ($t \in \{1, 2\}$), using the (k, l) block parameters, $\mu^t_{k,l}$ and $s^t_{k,l}$ as:

$$P(R^t_{i,j}|z^1_i = k, z^2_j = l) \quad = \quad \mathcal{N}(R^t_{i,j}|\mu^t_{k,l}, \{s^t_{k,l}\}^{-1}) \quad (6)$$

where $\mu^t_{k,l}$ and $s^t_{k,l}$ are the means and the precisions (inverse variances) in the (k, l) block, and $\mathcal{N}$ is the probability density function of the Gaussian distribution. In case of $t = 3$, the generative model for $R^t_{m,j}$ can be given in the same way as:

$$P(R^t_{m,j}|z^3_m = p, z^2_j = l) \quad = \quad \mathcal{N}(R^t_{m,j}|\mu^t_{p,l}, \{s^t_{p,l}\}^{-1}) \quad (7)$$

For all $t \in \{1, 2, 3\}$, the means, $\mu^t_{k,l}$, are given by Gaussian priors, and the precisions, $s^t_{k,l}$, are given by Gamma priors as:

$$\begin{aligned} P(\mu^t_{k,l}|\lambda^t, r^t) \quad &= \quad \mathcal{N}(\mu^t_{k,l}|\lambda^t, \{r^t\}^{-1}) \quad (8) \\ P(s^t_{k,l}|\beta^t, w^t) \quad &= \quad \mathcal{G}(s^t_{k,l}|\beta^t, \{w^t\}^{-1}) \quad (9) \end{aligned}$$

where mean, λ^t, precision, r^t, shape, β^t, mean, $\{w^t\}^{-1}$, are the fixed hyperparameters common to all the blocks for each type of matrix data and $\mathcal{G}$ represents the probability density function of the Gamma distribution. For $t \in \{1, 2\}$, the conditional posterior distribution for the means, $\mu^t_{k,l}$, are given as the product of the likelihood in the Equation (6) conditioned by the priors Equation (8), and the conditional posterior distribution for the means, $s^t_{k,l}$, are given as the product of the likelihood in the Equation (6) conditioned by the priors

Equation (9):

$$P(\mu^t_{k,l}|z^1, z^2, R^t, s^t_{k,l}, \lambda^t, r^t)$$
$$= \mathcal{N}(\mu^t_{k,l}|\frac{\overline{R^t_{k,l}}\ n_{k,l}s^t_{k,l} + \lambda^t r^t}{n_{k,l}s^t_{k,l} + r^t}, \frac{1}{n_{k,l}s^t_{k,l} + r^t})$$
$$(\text{where } \overline{R^t_{k,l}} = \frac{1}{n_{k,l}} \sum_{(i,j)} R^t_{i,j})$$
$$P(s^t_{k,l}|z^1, z^2, R^t, \mu^t_{k,l}, \beta^t, w^t)$$
$$= \mathcal{G}(s^t_{k,l}|\beta^t + n_{k,l}, [\frac{w^t\beta^t + \sum_{(i,j)}(R^t_{i,j} - \mu^t_{k,l})^2}{\beta^t + n_{k,l}}]^{-1}) \quad (10)$$

where $n_{k,l}$ is the number of members belonging to the (k,l) block, and the summations of (i,j) in Equation (10) run among $S_{k,l}$, i.e., $(i,j) \in S_{k,l}$, where $S_{k,l}$ is a set of index pairs (i,j) belonging to the (k,l) block. For $t = 3$, we can derive these conditional posterior distributions in a similar way.

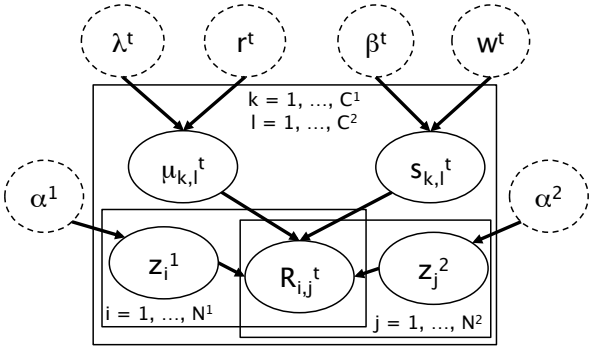

Figure 2: Graphical representation of the extended IRM for matrix data, R^t $(t \in \{1,2\})$

Figure 2 shows a graphical model representation for matrix data, R^t $(t \in \{1,2\})$. For $t = 3$, we can give the representation in a similar way. Parameters enclosed by dashed circles represent the fixed hyperparameters. Note that, in practice, we need three hyperparameters for the CRP, $\alpha^d (d \in \{1,2,3\})$, to give priors on cluster assignments of genes, samples, and modules, respectively.

2.2.4 Gibbs Sampling for Cluster Assignments

In the IRM, the cluster assignments are obtained by performing the Gibbs sampling on their conditional posterior distribution. In our case, we sampled the i-th gene's cluster assignment from Equation (1), the j-th sample's cluster assignment from Equation (2), and the m-th module's cluster assignment from Equation (3). Algorithm (1) shows the summary of Gibbs sampling in the extended IRM. As to a new gene cluster where no gene belongs yet, we can obtain the likelihood pertaining to the new cluster through the integration over the priors for the means and precisions. Unfortunately, this integral is not analytically tractable, however, we can effectively sample these parameters for unrepresented blocks just by sampling from priors, Equation (8), (9) [5].

Algorithm 1 The Gibbs sampling in the IRM

1: initialize the cluster assignments, z^1, z^2, z^3, randomly
2: update the parameters, ${\mu_{k,l}}^t$ and ${s_{k,l}}^t$ for all the (k,l) blocks
3: **repeat**
4: **for** $i = 1$ to N^1 (genes), $j = 1$ to N^2 (samples), and $m = 1$ to N^3 (modules) **do**
5: remove the cluster assignment
6: update the mean and precision parameters for biclusters
7: sample a new cluster assignment from Equation (1), Equation (2), or Equation (3)
8: update the mean and precision parameters for biclusters
9: **end for**
10: **until** convergence

2.3 Search for Drivers and Their Associated Modules

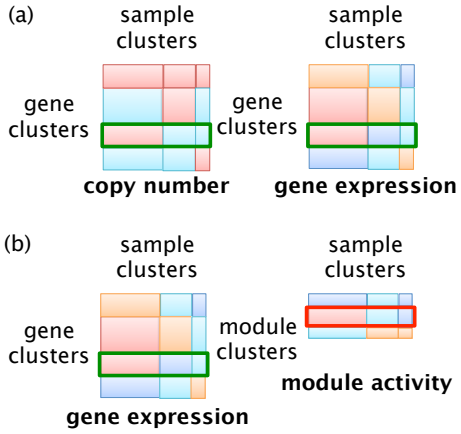

Figure 3: Search of prognostic factor candidates from the simplified clustering result

Utilizing the clustering results obtained by the extended IRM, we assigned a "*representative*" value to each bicluster in data matrices. Suppose that we obtain

C^1 gene clusters for N^1 genes, C^2 sample clusters for N^2 patient samples, and C^3 module clusters for N^3 modules as a clustering result, then we have $C^1 \times C^2$ biclusters for copy number, $C^1 \times C^2$ for gene expression, and $C^3 \times C^2$ for module activities. Let R^1_{sorted}, R^2_{sorted}, and R^3_{sorted} be data matrices obtained by sorting rows and columns of R^1, R^2, and R^3 based on the infered cluster assignments of genes, samples, and modules. Since each of the blocks in the sorted matrices is composed of the similar values, we can represent a *representative* value of the block as an average value of elements belonging to the bicluster. In this way, we applied a coarse granularization to the clustering results using $D^1 \in \mathbb{R}^{C^1 \times C^2}$, $D^2 \in \mathbb{R}^{C^1 \times C^2}$, $D^3 \in \mathbb{R}^{C^3 \times C^2}$, matrices where the (k, l) element is the average of the (k, l) bicluster of R^1_{sorted}, R^2_{sorted}, and R^3_{sorted}, respectively.

Based on the simple interpretation of clustering results, we searched driver genes and their associated modules. We distinguished drivers and others based on the following assumptions: for a driver gene (1) changes in DNA copy numbers well correlate with its gene expression levels among individuals; (2) its gene expression levels well associate with activity levels of downstream modules. In Figure 3, the more strongly reddish color blocks have, the larger the average of the blocks is, and the more strongly bluish color blocks have, the smaller it is. We selected a gene cluster (green squared) correlated positively the most between copy number and expression *representative* values (Figure 3(a)) as a group of driver candidates, and a module cluster (red squared) correlated positively or negatively the most with the *representative* values of gene expression data belonging to the selected gene cluster (Figure 3(b)) as a set of module candidates associated with the drivers.

3 Results

3.1 Analysis of CRC Patient Data by the Extended IRM

We applied our proposed model, the extended IRM, to the TCGA datasets of CRC patients in order to obtain an integrative clustering result in which three types of information, i.e., DNA copy number, gene expression, module activity, are simultaneously taken into account as the following manner. For DNA copy number and mRNA expression, the size of each data matrix was 2,000 genes × 423 patient samples. For module activity, the matrix size was 268 modules × 423 patient samples, which was obtained by using EEM method. Then we applied the extended IRM to the data matrices to estimate a set of clusters by using the Gibbs sampling algorithm. Figure 4 (a) shows the auto-correlation for log likelihood. The blue dashed line represent 95% confidence intervals based on uncorrelated series, which tells us that a correlation-length of the Markov chains is about 20. We generated total 200 MCMC samples, and to eliminate the auto-correlation among the samples, we discarded the first half, 100 MCMC samples, for "burn-in" period. Figure 4 (b) illustrates time series of log likelihood versus Monte Carlo iterations. We started from 100 clusters for all the types of clusters and selected one MCMC sample from the final 100. Figure 5 illustrates the visualization result of the selected MCMC sample. Black boundaries on each matrix separate clusters. Note that only 268 genes (the same number as modules) are shown in Figure 5 to make the boundaries for gene clusters clear. Each column in three matrices corresponds to the same patient sample, and each row in copy number and gene expression matrices represents the same gene. From the results, our extension of the IRM for continuous data work well enough to make biclusters exhibiting similar patterns. From comparison of clustering results on DNA copy number and mRNA expression, for example, when DNA copy number and mRNA expression are both high in several *gene* × *sample* blocks, we can see that copy number amplification may upregulate mRNA expression of the gene group in the patient sample cluster. In the same way, when mRNA expression of a gene cluster and module activity of a module cluster has a strong correlation, we can find that a regulatory relationship may exist between genes in the gene cluster and modules in the module cluster.

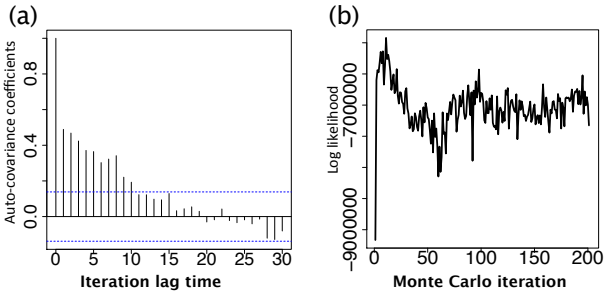

Figure 4: (a) Auto-correlation coefficient of log likelihood, (b) Log likelihood values of each MCMC sample

Regardless of the initial value, the number of estimated clusters converged to almost the same one. Figure 6 shows changes in the number of clusters versus the number of MCMC iterations starting from different values. Figure 6 shows that approximately 62 to 72 gene clusters, 38 to 45 sample clusters, and 42 to 50

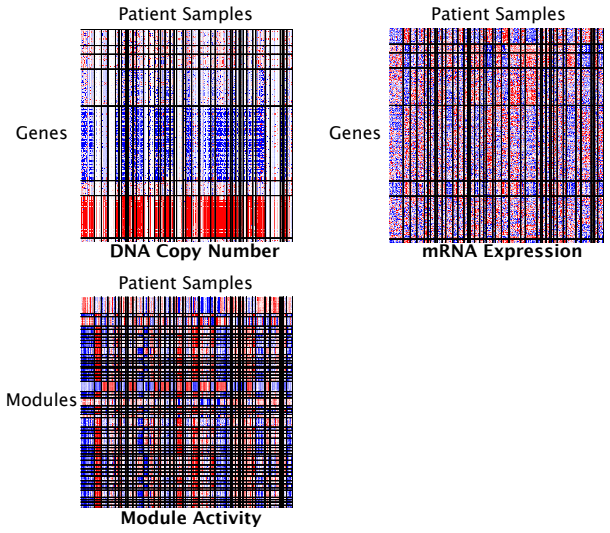

Figure 5: Result of Trans-omics clustering for TCGA CRC patients data using the extended IRM

module clusters were inferred from the data. After computing the frequency of the number of clusters and log likelihood in the final 100 MCMC samples, we selected one MCMC sample exhibiting the highest frequency in the number of clusters and log likelihood.

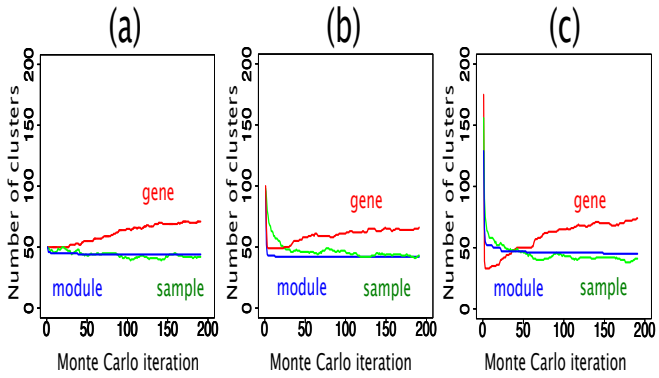

Figure 6: Changes in the number of clusters after starting from (a) 50 clusters, (b) 100 clusters, and (c) about 175 clusters

3.2 Identification of Prognostic Factors from Coarse-grained Signatures

Since clustering of patient samples was performed based on their genetic and transcriptional data, we attempted to search for prognostic factor candidates, indicators of severity of CRC in patients, from obtained gene clusters and modules ones. In order to achieve the goal, we applied to the clustering results a coarse granularization that assigns a single "*representative*" value to elements in the same cluster in order to extract essential tendencies by reducing effects of measurement noises, and represented differences of genetic and transcriptional characteristics across patient samples by coarse-grained signatures. Specifically, first, we obtained an average value for each block in the three matrices as the *representative* value. After performing the extended IRM, we got 68 gene clusters, 43 samples clusters, and 50 module clusters, respectively. Therefore, we got 68×43 blocks for DNA copy number and mRNA expression and 50×43 blocks for module activity, and prepared average values representing these blocks.

Next, in order to search for driver gene candidates showing strong association between copy number and gene expression and to infer modules that are likely to be regulated by the genes, we obtained correlation coefficients of averaged copy number and gene expression, and of averaged gene expression and module activity. Then we found a gene cluster exhibiting high correlation coefficient, 0.81, between copy number and gene expression, and a module cluster positively correlated with the gene cluster with correlation coefficient 0.75. Here, we had no module cluster showing a strongly negative correlation with the gene cluster. Figure 7 illustrates copy number and gene expression for the gene cluster, and module activity for the module cluster. 95 genes, including MAPK4 (mitogen-activated protein kinase 4), were assigned to the selected gene cluster and 9 modules were classified as the selected module cluster, and we found no overlap between genes in the selected gene cluster and genes in each module assigned to the module cluster. As the bottom Figure 7 shows, many modules in the selected module cluster are associated with the immune system, e.g., the T-Cell receptor signaling pathway and the interleukin pathway.

Finally, using the clinical information from TCGA, we investigated the relevance between these selected clusters and clinical outcome. We separated the 43 sample clusters into two groups, sample group G_{High} and G_{Low}, a highly expressed group and a lowly expressed one in the selected gene cluster, and into two groups, sample group M_{Up} and M_{Down}, an upregulated group and a downregulated one in the selected module cluster. Here, we separated sample groups whether the *representative* value is higher than the arithmetic mean or not. The Kaplan-Meier curves for group G_{High} and G_{Low} (left) and for group M_{Up} and M_{Down} (right) are shown in Figure 8. Interestingly, from the comparison of two plots in Figure 8, although we could not see the significant association of the selected gene cluster with clinical outcome (P=0.442 calculated by log-rank test), we found that the selected module cluster was associated with the predicted survival rate significantly

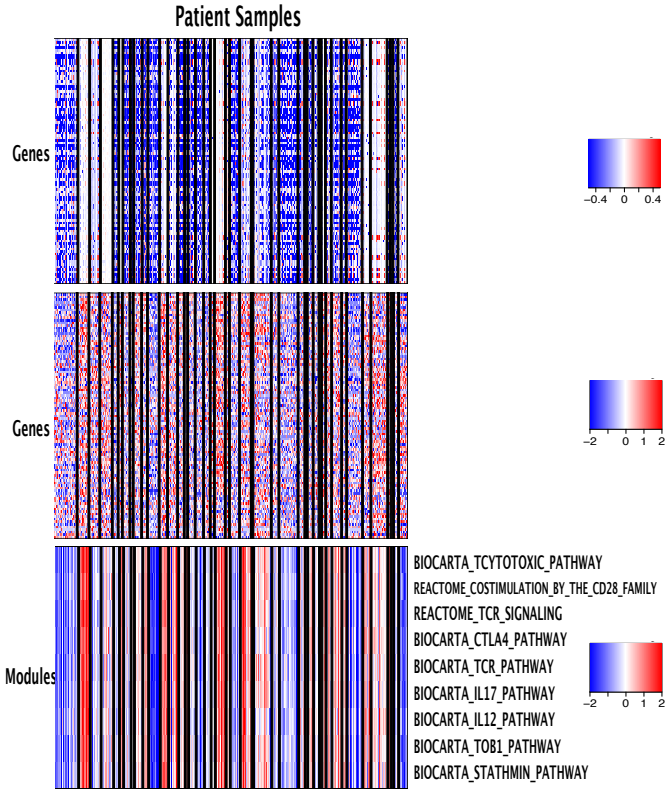

Figure 7: Extracted data of the selected gene cluster and module cluster(top: copy number, middle: gene expression, bottom: module activity)

($P=5.88\times10^{-4}$). These results indicates the clinical importance of the modules including the genes related to the T-Cell receptor and the interleukin, and we can possibly predict the prognosis of patients from the activity values in the selected modules.

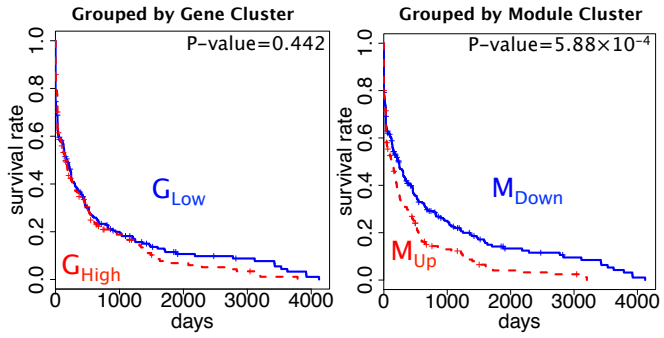

Figure 8: Kaplan Meier plots on patient sample group separated by gene expression of genes in the selected gene cluster (left) and by module activity of modules in the selected module cluster (right)

4 Discussion

In this study, we presented an integrative biclustering approach to give us an integrated view of genetic and transcriptional changes. Many methods are proposed for the identification of drivers or clinical important patterns in patients' gene signatures, however, most of them focus on only one type of features, either of genes or patient samples. We need to discover patterns on genes, samples, and modules to identify a prognostic factor from drivers and their associated modules. Although we can group each type of features sequentially (for example, we can classify patient samples after grouping genes), one of the pros in the simultaneous clustering approach is to group all the types of features without placing too high priority on either one type of them. Furthermore, of existing biclustering approaches, the IRM allows us to cover multi-dimensional matrix data without fixing the number of clusters. These advantages enabled us to extract information on genes, samples, and modules. Also, this time, we focused on the association with clinical outcome; however, we expect that this powerful tool can also provide more other types of biological insights when combined with other source of information, e.g., DNA methylation data and mutation data. Although the presence of mutation is normally represented as binary data, the combination of the original IRM and the extended IRM will easily enable us to analyze both the binary and continuous data.

After obtaining a clustering result by the extended IRM, we attempted to derive biologically insightful information from it. We applied to the clusters a coarse granularization, which assigns a single "*representative*" value to elements in the same cluster in order to extract essential tendencies by reducing effect of measurement and biological noises. Although further biological validation is needed to assert the regulatory relationship between the genes in the selected gene cluster and the modules associated with the immune system, as a result of the macroscopic viewpoint, we could identify these modules as significant prognostic factor candidates. Given the discovery that patient sample groups separated by gene expression patterns of the selected genes are not associated with the patients' survival, but those separated by module activity levels of the selected modules are correlated with it, the selected modules may be positioned downstream of regulatory pathways and the activity levels of them may reflect clinical phenotypes more accurately than the gene expression levels of the upstream genes.

Through the integrative analysis of multiple types of biological data, we could result in the identification of measurable indicators suggesting the severity of CRC.

Our modeling framework allows for other kinds of data resources, e.g., DNA methylation data, and has the potential ability to uncover hidden gene regulatory relationships. We hope that our approach will make further key contributions in revealing associations of clinical importance.

Acknowledgements. Computation time was provided by the Super Computer System, Human Genome Center, the Institute of Medical Science, University of Tokyo. This research used also computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project (Project ID: hp140230).

References

[1] Uri D. Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C. Causton, Panisa Pochanard, Eyal Mozes, Levi A. Garraway, and Dana Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, 2010.

[2] Norman Huang, Parantu K. Shah, and Cheng Li. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics*, 13(3):305–316, 2012.

[3] Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning Systems of Concepts with an Infinite Relational Model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 381–388. AAAI Press, 2006.

[4] Qianxing Mo, Sijian Wang, Venkatraman E. Seshan, Adam B. Olshen, Nikolaus Schultz, Chris Sander, R. Scott Powers, Marc Ladanyi, and Ronglai Shen. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, 2013.

[5] Radford M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[6] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[7] Atsushi Niida, Andrew D. Smith, Seiya Imoto, Hiroyuki Aburatani, Michael Q. Zhang, and Tetsu Akiyama. Gene set-based module discovery in the breast cancer transcriptome. *BMC Bioinformatics*, 10(1):71, 2009.

[8] Jim Pitman. *Combinatorial Stochastic Processes*, volume 32. Springer, 2006.

[9] Carl E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, volume 12, pages 554–560, 1999.