

Visualization of transitions of developing of hepatitis C virus-associated hepatocellular carcinoma

Takanobu Miyamoto, Yusuke Fujita, Shunji Uchimura,
Yoshihiko Hamamoto, Norio Iizuka, and Masaaki Oka
Yamaguchi University
t-miy@yamaguchi-u.ac.jp

Abstract

In our previous study, we visualized microarray data of hepatocellular carcinoma (HCC) by using self-organizing-map, and investigated molecular signature representing the development of HCC. In this study, we propose two visualization methods of microarray data with Euclidean distance classifiers and Sammon's non-linear mapping. Our proposed methods will serve as tool to discover molecular signature representing the development of HCC for molecular biologists or doctors.

1. Introduction

In our previous study [3], using high-density oligonucleotide array, we comprehensively analyzed expression levels of 12600 genes in 50 hepatocellular carcinoma (HCC) samples with positive HCV serology (well (G1), moderately (G2), and poorly (G3) differentiated tumors) and 11 non-tumorous livers (L1 and L0) with and without HCV infection. We hypothesize that HCC develops sequentially from L0 to L1 to G1 to G2 to G3. We searched for discriminatory genes of transition with a supervised learning method, and then arranged the samples by self-organizing map (SOM) with the discriminatory gene sets. The SOM correctly arranged the 5 clusters on a unique sigmoidal curve in the order L0, L1, G1, G2, and G3. However, it was difficult for molecular biologists or doctors to identify transitions because 5 clusters existed on a complex curve. Therefore, simple images are generally required in order that non-experts of pattern recognition interpret biological meanings. In this paper, we aim to easily identify transitions by visualizing 5 stages with linear and nonlinear mapping.

2. Microarray data

50 patients who underwent surgical treatment for HCC at Yamaguchi University Hospital between May 1997 and August 2000 were enrolled in this study. None of them underwent chemotherapy prior to surgery. In the 50 patients with HCV-related HCC, histopathologic examination based on TNM classification of the International Union Against Cancer revealed that 7 had well differentiated HCC (G1), 35 had moderately differentiated HCC (G2), and the remaining 8 had poorly differentiated HCC (G3). As controls, 6 non-tumorous liver samples (L0) from 6 patients who underwent hepatic resection and who had histologically normal livers were used. We also had 5 HCV-infected non-tumorous liver samples (L1) from 5 HCC samples [3]. We hypothesized that HCC develops sequentially from L0 to L1 to G1 to G2 to G3. Informed consent in writing was obtained from all of these patients before surgery. The study protocol was approved by the Institutional Review Board for the Use of Human Subjects at the Yamaguchi University School of Medicine.

We comprehensively analyzed the levels of expression of approximately 12600 genes using high-density oligonucleotide array (HuGeneFL Array, Affymetrix, Santa Clara, CA) [3].

3. Visualization with simple classifiers

First, we visualized pre-transition samples and post-transition samples for 4 subsets. Next, we visualized L0, L1, G1, G2, and G3 samples. In this section, linear mapping methods were used.

Step 1: Division of samples

In this section, we used 4 subsets consisted of pre-transition and post-transition samples as shown in Table 1. We divided 61 samples into training samples and test samples in each subset.

Table 1. Subset of training samples and test samples

	Training sample		Test sample	
	Pre-transition	Post-transition	Pre-transition	Post-transition
Subset A	L0	L1		G1, G2, G3
Subset B	L1	G1	L0	G2, G3
Subset C	G1	G2	L0, L1	G3
Subset D	G2	G3	L0, L1, G1	

Step 2: Filtering

We selected genes that had levels of expression > 40 in 50 HCV-related HCC samples. Among 12600 genes, this filtering resulted in the selection of 3559 genes.

Step 3: Calculation of the Fisher criterion

We used the Fisher criterion $J(k)$ to evaluate the potentials of the selected genes to discriminate pre-transition samples and post-transition samples.

$$J(k) = \frac{(\mu_1(k) - \mu_2(k))^2}{P(\omega_1)\sigma_1^2(k) + P(\omega_2)\sigma_2^2(k)} \quad (1)$$

where $\mu_i(k)$ is the k th component (=gene) of the sample mean vector $\boldsymbol{\mu}_i$ of Group ω_i , $P(\omega_i)$ is the a prior probability of Group ω_i , and $\sigma_i(k)$ is the k th diagonal element of the sample covariance matrix Σ_i of Group ω_i .

Step 4: Selection of the top 40 genes

We mathematically selected the 40 significant discriminatory genes out of 3559 genes in each transition. Each gene passed the random permutation test ($P < 0.005$) [3, 4] using the Fisher criterion $J(k)$.

Step 5: Comparison of classifiers

We evaluated 4 classifiers designed with the 40 significant discriminatory genes for classification between pre-transition and post-transition stages. The minimum distance classifier (MDC), the nearest neighbor classifier (1-NN), the support vector machine (SVM), and the artificial neural network (ANN) (MATLAB Neural Network Toolbox) were used. We designed classifiers on the training samples and the classification rates were estimated on the test samples in each subset. The classification rates are shown in Table 2.

Although it is generally believed that complicated classifiers such as SVM and ANN outperform a simple classifier such as MDC, MDC was the most powerful among 4 classifiers in terms of the classification rate.

Step 6: Visualization of pre-transition and post-transition samples

In each subset, we visualized the pre-transition samples and the post-transition samples by estimating a score $T(\mathbf{x})$ based on the MDC (see Figure 1).

$$T(\mathbf{x}) = dist(\mathbf{x}, \boldsymbol{\mu}_{pre}) - dist(\mathbf{x}, \boldsymbol{\mu}_{post}) \quad (2)$$

Table 2. Comparison of four classifiers

	Subset A	Subset B	Subset C	Subset D
MDC	92%	98%	84%	100%
1-NN	92%	98%	63%	89%
SVM	94%	98%	63%	100%
ANN	92%	98%	68%	94%

where $\boldsymbol{\mu}_{pre}$, $\boldsymbol{\mu}_{post}$, and $dist(\mathbf{x}, \mathbf{y})$ are respectively the mean vector of pre-transition training samples, the mean vector of post-transition training samples, and the Euclidean distance between \mathbf{x} and \mathbf{y} . A pre-transition sample and a post-transition sample are represented by an open circle (\circ) and a closed circle (\bullet), respectively.

Step 7: Visualization of 5 stage

We visualized L0, L1, G1, G2, and G3 stage samples by estimating a score $\hat{T}(\mathbf{x})$ based on $T(\mathbf{x})$ (see Figure 2).

$$\hat{T}(\mathbf{x}) = T_A(\mathbf{x}) + T_B(\mathbf{x}) + T_C(\mathbf{x}) + T_D(\mathbf{x}) \quad (3)$$

where $T_i(\mathbf{x})$ is a score $T(\mathbf{x})$ in subset i ($i = A, B, C, D$).

4. Visualization with multi-class Fisher criterion and Sammon's mapping

In this section, we visualized samples of L0, L1, G1, G2, and G3 with multi-class Fisher criterion [1] and Sammon's nonlinear mapping (Sammon's mapping) [5].

Step 1: Division of samples

First, in order to eliminate the bias of the sample size, we randomly selected 5 samples in each stage. In this section, only 25 samples were used.

Step 2: Filtering

We selected genes that had levels of expression > 40 in 25 samples selected in step 1. Among 12600 genes, this filtering resulted in the identification of 4237 genes.

Step 3: Calculation of multi-class Fisher criterion

We used the multi-class Fisher criterion $\hat{J}(k)$ to evaluate the potentials of the selected genes to discriminate

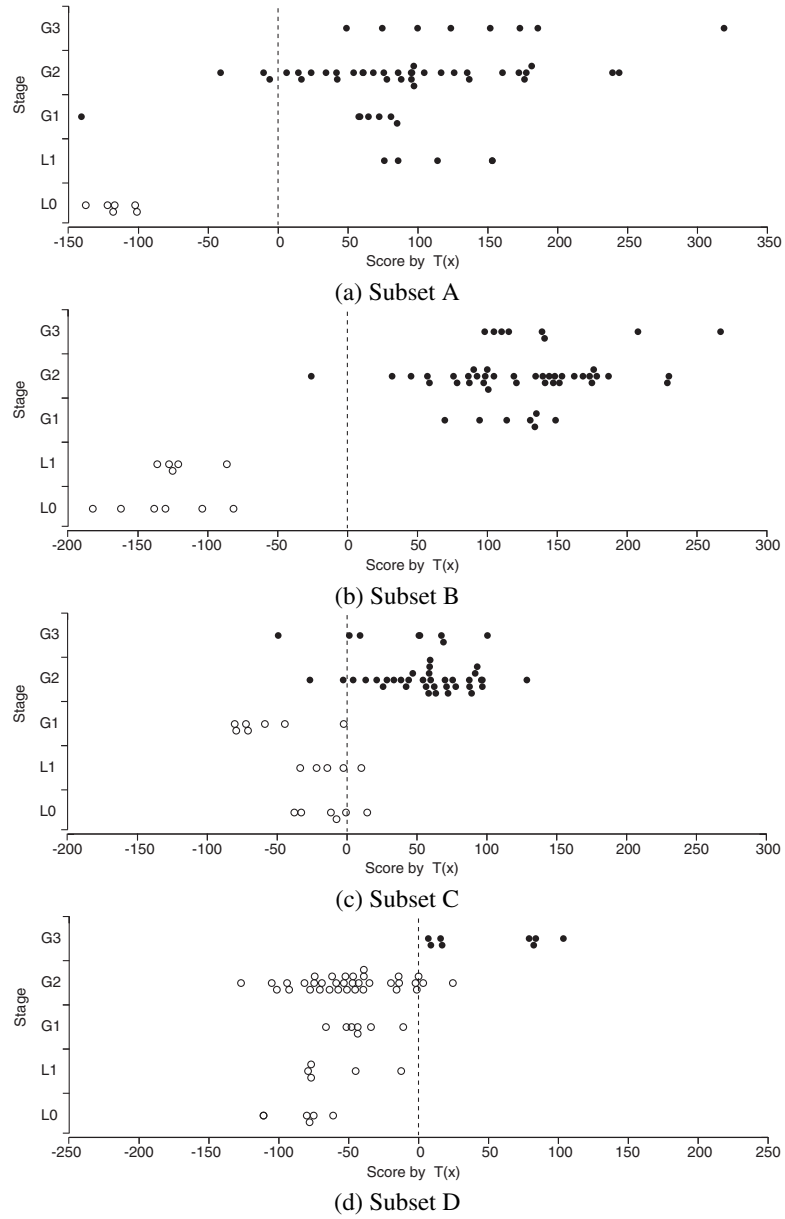


Figure 1. Visualization of pre-transition samples and post-transition samples.

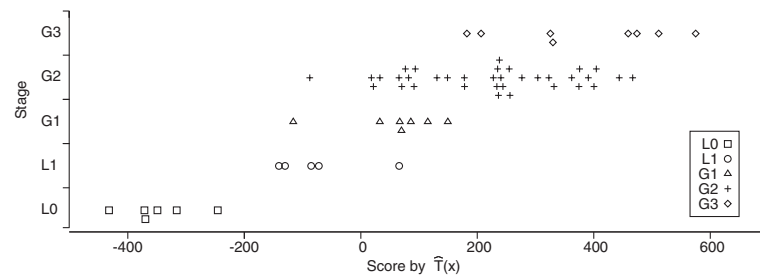


Figure 2. Visualization of samples consisted of L0, L1, G1, G2, and G3 stages with simple classifiers

L0, L1, G1, G2, and G3.

$$\hat{J}(k) = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^m P(\omega_i)P(\omega_j)(\mu_i(k) - \mu_j(k))^2}{\sum_{i=1}^m P(\omega_i)\sigma_i(k)^2} \quad (4)$$

The multi-class Fisher criterion and normal Fisher criterion measure the difference between means normalized by the averaged variance. Unlike Euclidean distance or a criterion based on a fold change, the Fisher criterion takes account of the variance.

Step 3: Selection of the top 40 genes

We selected the top 40 out of 4237 genes on the basis of their individual effectiveness. Here, we ranked the genes to be considered in the order of decreasing magnitude of the multi-class Fisher criterion and selected the top 40 genes.

Step 4: Visualization with Sammon's mapping

We visualized a subset of 25 samples by using Sammon's mapping. The $D(=40)$ -dimensional sample \mathbf{x} were reduced to the $d(=2)$ -dimensional sample $\hat{\mathbf{x}}$. In the Sammon's mapping, d -dimensional samples are arranged so as to minimize E .

$$E = \frac{1}{\sum_{i<j} [d_{ij}^*]} \sum_{i<j} \frac{[d_{ij}^* - d_{ij}]^2}{d_{ij}^*} \quad (5)$$

where d_{ij}^* is the distance between \mathbf{x}_i and \mathbf{x}_j in the D -dimensional space, and d_{ij} is the distance between $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ in the d -dimensional space.

Visualization result are shown in Figure 3

5 Discussion and Conclusions

We visualized 50 HCC samples with positive HCV serology (well (G1), moderately(G2), and poorly (G3) differentiated tumors) and 11 non-tumorous livers (L1 and L0) with and without HCV infection. First, we visualized pre-transition and post-transition samples (Figure 1) and pre-transition and post-transition samples were clearly discriminated in each transition. Using this experimental results, we visualized 61 samples consisted of L0, L1, G1, G2, and G3 stages (Figure 3). We found that stages of differentiation tend to develop as the value of $T(\mathbf{x})$ grows.

Moreover, we selected 40 significant discriminatory genes by using multi-class Fisher criterion, and visualized samples of L0, L1, G1, G2, and G3 with Sammon's mapping. From Figure 3, we found that stages of differentiation tend to develop toward the arrow direction.

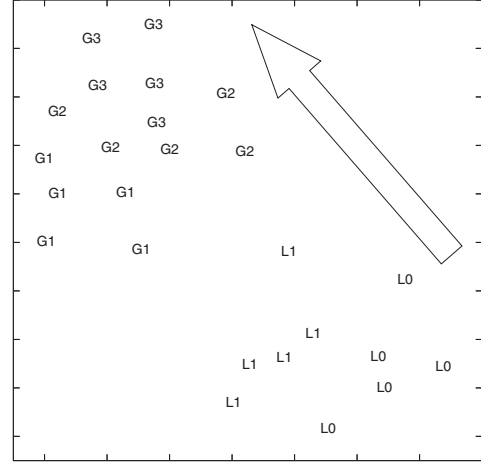


Figure 3. Visualization with Sammon's mapping

In this paper, we could successfully visualize transitions of developing of HCV-associated HCC. Our simple results will provide a framework to explain the development of HCC with positive HCV serology.

References

- [1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley-interscience publication, 1973.
- [2] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.
- [3] N. Iizuka, M. Oka, H. Yamada-Okabe, N. Mori, T. Tamesa, T. Okada, N. Takemoto, K. Sakamoto, K. Hamada, H. Ishitsuka, T. Miyamoto, S. Uchimura, and Y. Hamamoto. Self-organizing-map-based molecular signature representing the development of hepatocellular carcinoma. *FEBS Letters*, 579:1089–1100, 2005.
- [4] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, K. Hamada, H. Nakayama, H. Ishitsuka, T. Miyamoto, A. Hirabayashi, S. Uchimura, and Y. Hamamoto. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, 361:923–929, 2003.
- [5] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans.*, C-18(5):401–409, January 1969.