The Identification of a Cancer Cell Gene by Using SOM

Yasaburo Matsuura\textsuperscript{1} \hspace{1cm} Tuoya\textsuperscript{2} \hspace{1cm} Masaharu Seno\textsuperscript{2}

Heizo Tokutaka\textsuperscript{3} \hspace{1cm} Masaaki Ohkita\textsuperscript{1}

mohkita@ele.tottori-u.ac.jp

\textsuperscript{1} Department of Electrical and Electronic Engineering, Faculty of Engineering, Tottori University.
\hspace{1cm} 4-101 Koyama-Minami Tottori-City Tottori-Prefecture 680-8552-8552, Japan.
\textsuperscript{2} Graduate school of Natural Science and Technology, Okayama University, Okayama 700-8530, Japan
\textsuperscript{3} SOM Japan,Ltd.

Keywords: self-organizing maps, cancer cell, DNA microarray

1 Introduction
The structure of all genes that each living thing maintains is clarified one after another by progress of the genome sequence project for several years. DNA microarray which monitors the amount of appearance of all genes in the cell at a time is known. In this article, the method of being search about the gene which is considered to be caused by the cancer from the value of a DNA microarray is proposed.

2 Method and Results
The values of the micro-array data of 1411 genes which are obtained from 6 cells of A, B, C, D, E, and F are compared with each other. Therefore, it is important to discover the gene which is considered to be caused by the cancer. Cell A is a normal cell and the cell B, C, D, E, and F are the cells invaded by the cancer.

2.1 Method 1
Data were revised for the difference (of the values among the cells of every gene) to appear remarkably. An average of every genes for 5 brain cells invaded by the cancer is made as $G$. A standard deviation of every genes for 5 brain cells invaded by the cancer is made as $V$. The value of every genes of the normal brain cell is made as $A$. There is a possibility that the gene where $|A - G|$ is big and $V$ is small is roughly caused by the cancer manifestation.

2.2 Method 2
The Self-Organizing Maps (SOM) which is effective in the classification of the multi-dimensional data is used to classify the relation of the 6 cells more clearly.

In order to make the created map as more concise form, the condition of $|A - G| - V > 0$ is necessary. Thus, the input sample number becomes 861. Also, as an ideal combination where there is a clear possibility to be caused by the cancer manifestation, the following pseudo-data was introduced as

\begin{align*}
\text{nb1} \{1 0 0 0 0 0\} \quad \text{and} \quad \text{nb0} \{0 1 1 1 1 1\}
\end{align*}

In the nb1, the value of the micro-array of normal cell A is as 1, and the values of one of the cells of B, C, D, E, and F which are invaded by the cancer are all as 0.

On the other hand, in the nb0, the value of the micro-array of normal cell A is as 0, and the values of one of the cells of B, C, D, E, and F which are invaded by the cancer are all as 1. Incidentally, the value of the micro-array is normalized to the range from 0 to 1.

There will be the possibility that the gene near this 2 pseudo-data on the map is roughly caused by the cancer manifestation.

2.3 Result
The result of Method 1 is shown in Fig. 1. The vertical axis is the standard deviation $V$ of every
genes of the 5 brain cells invaded by the cancer. The transverse axis is the difference $|A - G|$ between the value $A$ of every genes of the normal brain cell and the average $G$ of every genes of the 5 brain cells invaded by the cancer. There will be the possibility that the gene will be caused for the cancer manifestation as to go to the lower right of Fig.1. The small number of genes are dotted distributed in the lower right corner and then, the objective genes will be easily found. There will be the possibility that the gene near this 2 pseudo-data on the map is roughly caused by the cancer manifestation.

The result of the Method 2 is shown in Fig. 2. nb1 appeared in the upper left and nb0 appeared at the center of the right. nb0 and nb1 are fully separated. Then, the classification which is not possible in Method 1 will be possible in this method.

![Fig.1 Method 1](image1)

![Fig.2 SOM by Method 2](image2)

3 Discussion

A result is shown in Table 1 by comparing both 2 methods with each other. In Method 1, the case of 4 cells of cell A, B, C, and D and the case of 6 cells of A, B, C, D, E, and F are compared. In the case of 6 cells, the precision becomes higher. By increasing the number of the cells, it will be possible to attempt for the precision to be improved in the future work. Also, a similar result with Method 1 was obtained using Method 2 (SOM). Using Method 2 (SOM), nb0 and nb1 are clearly classified. However, nb1 was clearly classified using Method 1, but nb0 wasn't so much classified. It will be left for the future work of this confirmation.

<table>
<thead>
<tr>
<th></th>
<th>Method 1</th>
<th></th>
<th>Method 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>6 cell nb1</td>
<td>454</td>
<td>3rd</td>
<td>454</td>
<td>1st</td>
</tr>
<tr>
<td>4 cell nb0</td>
<td>1365</td>
<td>422th</td>
<td>787</td>
<td>11th</td>
</tr>
<tr>
<td>6 cell nb0</td>
<td>281</td>
<td>13th</td>
<td>281</td>
<td>2nd</td>
</tr>
<tr>
<td></td>
<td>949</td>
<td>806th</td>
<td>611</td>
<td>291th</td>
</tr>
<tr>
<td></td>
<td>426</td>
<td>17th</td>
<td>545</td>
<td>840th</td>
</tr>
<tr>
<td></td>
<td>190</td>
<td>35th</td>
<td>893</td>
<td>5th</td>
</tr>
<tr>
<td></td>
<td>893</td>
<td>20th</td>
<td>678</td>
<td>8th</td>
</tr>
<tr>
<td></td>
<td>767</td>
<td>34th</td>
<td>426</td>
<td>3rd</td>
</tr>
<tr>
<td></td>
<td>206</td>
<td>783th</td>
<td>190</td>
<td>4th</td>
</tr>
</tbody>
</table>

Table 1 the gene which has the possibility to be as roughly caused for the cancer manifestation

References