Probe classification of on-off type DNA microarray images with a nonlinear matching measure

Munho Ryu
Chonbuk National University
Division of Bionics and Bioinformatics
664-14 1 Ga, Duckjin-Dong
Duckjin-Gu, Jeonju
Jeonbuk 561-756, South Korea
E-mail: mhryu@chonbuk.ac.kr

Jong Dae Kim
Hallym University
Division of Information and Communication Engineering
1 Okchon-dong
Chunchon 200-702, Korea

Byoung Goo Min
Seoul National University
College of Medicine
Department of Biomedical Engineering
and
Medical Research Center
Institute of Medical and Biological Engineering
Seoul, Korea

Jongwon Kim
BiomedLab Co., Ltd.
Gyeonggi Technopark Ste. 811
1271-11 Sa-dong, Sangnok-gu
Ansan, Gyeonggi-do
426-901 Seoul, Korea

Yong-Yook Kim
Chonbuk National University
Center for Healthcare Technology Development
664-14 1 Ga, Duckjin-dong
Jeonju, Jeonbuk 561-756
South Korea

1 Introduction

Automatic analysis is one of the main issues of DNA microarray technology.\textsuperscript{1-15} The analysis is commonly composed of two steps. First, a reference position of a spot set is located and then the signal amplitude of each spot is subsequently measured. There have been various efforts to measure signal amplitude effectively. However, most of these studies focused on ratio images, where the measured continuous signal corresponds to the gene expression profile. However, one must focus on the absolute value of the detected signal strength for on-off type DNA microarrays, such as HPVDNAChip (Biomedlab Co., Korea), because the signal is interpreted in the on-off state to detect the presence of target DNA sequence.

The HPVDNAChip, designed to detect human papillomavirus (HPV) infection, one of the main causes of cervical cancer, is configured\textsuperscript{16-26} as shown in Fig. 1. There are four chambers in one slide, one for each patient. Each chamber has two identical spot sets to increase diagnostic credibility. A spot set has four positive control markers and 22 pairs of HPV type-specific oligonucleotide probes. Each HPV type probe is also duplicated forming a pair of spots, yielding four spots for one type of probe in one chamber. The four positive control markers in each set are oligonucleotide probes for human β-globin and are used to locate the reference of a spot set and to verify the hybridization.

The target DNA is extracted from clinical sample, amplified by a single-round polymerase chain reaction (PCR), and

Abstract. We propose a nonlinear matching measure, called counting measure, as a signal detection measure that is defined as the number of on pixels in the spot area. It is applied to classify probes for an on-off type DNA microarray, where each probe spot is classified as hybridized or not. The counting measure also incorporates the maximum response search method, where the expected signal is obtained by taking the maximum among the measured responses of the various positions and sizes of the spot template. The counting measure was compared to existing signal detection measures such as the normalized covariance and the median for 2390 patient samples tested on the human papillomavirus (HPV) DNA chip. The counting measure performed the best regardless of whether or not the maximum response search method was used. The experimental results showed that the counting measure combined with the positional search was the most preferable. © 2006 Society of Photo-Optical Instrumentation Engineers.

Keywords: human papillomavirus; DNA chip; high throughput; template matching; nonlinear matching measure; probe.

Paper 05077RR received Mar. 22, 2005; revised manuscript received Sep. 13, 2005; accepted for publication Sep. 20, 2005; published online Jan. 31, 2006.
hybridized onto the chip. It randomly incorporates Cy5 during PCR amplification and visualizes the position of hybridization when the DNA chip is scanned. After locating the marker, each spot is read as present or absent to detect the existence of the corresponding HPV type. To make the chip an on-off type, all of the probe sequences are unique to the HPV type, using the thermodynamic stability between the oligonucleotide probe sequence in length of thirty base pairs and the target DNA sequence.27–30

The template matching method31–55 in our previous studies showed reasonable performance in locating markers for on-off type microarray,24–26 as has integration of prior knowledge and template-matching methods with normalized covariance as a measure help to locate markers.24 A nonlinear matching measure has been proposed and compared with normalized covariance, and successfully applied to locating markers.25,26

Fig. 1 Architecture of HPVDNAChip microarray.

The nonlinear matching measure was obtained by binary thresholding the template region and counting the white pixels inside the object region. It simulated the behavior of the expert, who decided that a spot was hybridized if the spot area was filled with a certain number of relatively white pixels. It was proven that the measure was robust especially with respect to the spot signal amplitude variation, which is common in microarray images.

This paper presents that the nonlinear matching measure, which we denote as a counting measure, also delivers better performance than the classical signal detection measures such as covariance and median for the probe classification. We also present the maximum response search method in which the expected signal is obtained by taking the maximum of the responses according to the various positions and sizes of a spot template. The counting measure was tested on 2390 patient samples in terms of its discrimination ability on classifying the probes to an on-off state. The comparison showed that the counting measure outperformed the existing measures regardless of the adapting maximum response search method used.

2 Probe Classification by Template Matching

A single binary circular template is used to simulate spot shape.30–36 It is composed of an object region and a background region. Object and background have the same area and their values are 1 and −1, respectively, to eliminate bias. Even though the spot size is determined by the dotter shape and the marker-locating step, which locates the reference position of a spot set, our method defines probe location within a reasonable range, as in Ref. 26. However, it is advantageous to search the position and size of each probe spot, i.e., fine-tuning, to achieve a more discriminating result, as is clear in Fig. 2. Three kinds of fine-tuning methods are applied. First, neither position nor size is searched, but their default values are used (no search). Second, the spot position that gives maximum response is searched (position search). Third, both the spot position and size that gives the maximum response are searched (position and size search).

For the matching measure, normalized covariance (NC), median, and the counting matching measures are compared. They are expressed as

\[
NC(i, j) = \frac{\sum_{k,l \in \text{Template}} T(k, l)I(i+k, j+l)}{\sigma_T \sigma_I(i, j)},
\]

\[
\text{Med}(i, j) = \text{median}_{k,l \in \text{Template}} [I(i+k, j+l)],
\]

\[
\text{Counting}(i, j) = \#_{k,l \in \text{Template}} \{ (k, l) | I(i+k, j+l) \in O \} - \#_{k,l \in \text{Template}} \{ (k, l) | I(i+k, j+l) \in B \},
\]

where \( T(i, j) \) and \( I(i, j) \) are the intensities of the template and the image at the pixel position \((i, j)\), respectively; \( \sigma_T \) and \( \sigma_I \) are the intensity standard deviation of the template and that of the image in the area corresponding to the template; and \#\{\}, median, Template, \( O \), and \( B \) are the number of element of the set, the median operation, the template area, the object area, and the background, respectively.

Fig. 2 Fine-tuning for probe classification. Red crosses are the initial spot positions predetermined by the located marker. White circles denote the final spot areas. (a) No search, the white circle is centered at the cross point; (b) position search, the white circle is not centered at the cross point; and (c) position and size search, the white circle is not centered at the cross point and its size is greater than the initial size.
the probe classification procedure can be composed as follows.

**Probe Classification Procedure**

For 22 probes

For four duplicate spots

Search spot position which gives maximum response.

Search spot size which gives maximum response.

End (for four duplicate spots)

Calculate average response for four duplicate spots.

Determine if this probe is hybridized or not. The probe is hybridized if the average response is higher than threshold, and not hybridized otherwise.

End (for twenty-two probes)

**End procedure**

In the preceding procedure, each probe delivers a real-valued measure response and it is compared with a given threshold to determine whether or not the probe is hybridized. Therefore, if we have the true class of each probe, that is, we know whether or not it is hybridized, then we can select an effective signal detection measure using well-defined pattern classification methods.

The counting measure of Eq. (3) and the existing measures of Eqs. (1) and (2) can be combined with three kinds of fine-tuning methods: no search, position search, and the position and size search, effectively yielding nine measures. These nine measures are compared in terms of discriminating abilities with the experts’ classification. We investigated the scatter plots of pairs of measures for subjective comparison. For quantitative comparison, the interclass distances and the partial receiver operating characteristic (ROC) analysis were employed.

3 Experimental Results

A total of 2390 patient samples were used to evaluate the three kinds of measure and three kinds of fine-tuning methods. Among them, 768 (32.1%) samples were HPV positive. All the samples were tested and each spot was determined as present or absent by an expert’s naked eye. We assumed that...
the expert’s classification is true. For easy data manipulation, a reduced set was chosen. Note that as there are 22 probes in one sample, there are 52,580 probe spots. Among them, all of 1115 hybridized spots were selected, while 2000 of nonhybridized spots were picked up randomly. The scatter plots shown in upcoming Figs. 4–6 were employed for the subjective comparison of the measures. The interclass distance (Fig. 7, shown later) and the partial ROC analysis\(^{56–58}\) (Table 1) were used to quantify the comparisons.

The counting measure provided a more stable response than the NC for the spots with different signal intensities, as shown in Fig. 3. There are three hybridized spot quadruples, and each quadruple gives a different intensity, as shown in Fig. 3(a). While the NC responses vary with the spot intensity, as in Fig. 3(b), the counting measure gives more stable response regardless of the spot intensity as in Fig. 3(c). The separation between the hybridized and nonhybridized spots does not seem good enough here, because neither spot position nor size was searched.

The NC and the counting measure are highly correlated and fine-tuning improves discriminating performance. For each fine-tuning method, the NC, median, and the counting
measure are shown in scatter plots in Figs. 4–6. From the figures, we can see that the NC and the counting measure show similar and reasonable performance, but the median shows poor discriminating ability. It can be easily noticed when we draw a vertical line at the class boundary for the NC and horizontal lines for the median and the counting measure. For example, in Fig. 4, there are not many data points (nine false negative points) crossing over the vertical line at NC = 0.43, that is, the rectangular data points on the left side, while there are too many data points (146 false negative points) crossing over the horizontal line around median = 0.51. This investigation proves that the discrimination with median is poorer than with NC.

The position search showed significant improvement, as we can see by comparing Figs. 4(b) and 5(b). However, the improvement by position and size search is not significant, as shown in Figs. 5(b) and 6(b). Note that there was no false negative spot for NC and counting measure in Figs. 5(b) and 6(b).

The feature selection criterions are compared in Fig. 7. The response values of nine measures are applied to a feature selection criterion as a feature set. Hybridized spots and nonhybridized spots are applied as different classes. The interclass distance was calculated as expressed in Eq. (4). It is an averaged Euclidean distance between samples for exclusive classes. We can confirm the same result as expected from the previous scatter plots. The discriminating capability improved as fine-tuning was applied, and the median showed the poorest performance. The counting measure with both position and size search, and that with position search, were the best and second best performances.

\[
\text{interclass distance}(i,j) = \sum_{k} \sum_{l} \sum_{j} d(k,l),
\]

where \(i\) and \(j\) are different classes; \(k\) and \(l\) are samples in classes \(i\) and \(j\), respectively; and \(d(k,l)\) is Euclidean distance between samples \(k\) and \(l\).

An ROC curve analysis was performed for each feature. Each partial area index was calculated for the range where false positive rate (FPR) is below 0.05, as shown in Table 1. This also confirms the previous result shown in the scatter plots and the interclass distance comparison.

<table>
<thead>
<tr>
<th>Fine-Tuning</th>
<th>NC</th>
<th>Median</th>
<th>Counting</th>
</tr>
</thead>
<tbody>
<tr>
<td>No search</td>
<td>0.9987</td>
<td>0.9619</td>
<td>0.9993</td>
</tr>
<tr>
<td>Position search</td>
<td>0.9995</td>
<td>0.9728</td>
<td>1.0000</td>
</tr>
<tr>
<td>Position and size search</td>
<td>0.9994</td>
<td>0.9807</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Table 1 Partial ROC area index (for FPR < 0.05).
4 Conclusion and Discussion

We proposed a nonlinear matching measure, called counting measure, and applied it to a probe classification, especially for an on-off type DNA microarray. This kind of microarray is designed in such a way that the hybridization signal is detected as present or absent. Therefore, consideration of measure selection is important. The basic underlying concept of this paper is to simulate an expert’s behavior that estimates the amount of white pixels filling the spot area.

Probe classification was applied to 2390 patient samples. The counting measure was compared with the NC and median. Three measures were combined with the fine-tuning method, where the position and both the position and size were searched to give the maximum response. The nine kinds of measures, combinations of each of three measures and fine-tuning methods, were compared subjectively and quantitatively. Pairwise scatter plots were investigated for subjective comparison and the interclass distance and ROC analysis were employed for quantitative comparison of nine measures. The counting measure performed the best regardless of the employed fine-tuning method. Even though the counting measure delivers better performance with both position and size search, only the positional search is preferable, because the performance enhancement with the additional size search is negligible in contrast to the great increase of the computation cost. Note that the position search is operated in the same template-matching measure response with a single fixed size. However, the size search requires a separate measure response for each size.

The integration of a probe classification scheme proposed here and the marker-locating method along with the proposed template-matching measure, proposed by the authors previously, provides a complete solution set to the automatic analysis of the on-off type DNA microarray image. This automatic analysis of HPV DNA microarray chip has significance in the sense that it accelerates high-throughput for cervical cancer screening. This solution can be also applied to the other on-off type DNA microarrays images.

Acknowledgment

This work was supported by the Regional Research Centers Program of the Korean Ministry of Education & Human Resources Development through the Center for Healthcare Technology Development.

References
Ryu et al.: Probe classification of on-off type DNA microarray images…

31(14), 4211–4217 (2003).