Nonlinear matching measure for the analysis of on-off type DNA microarray images

Munho Ryu
Seoul National University
Graduate School
Interdisciplinary Program in Biomedical
    Engineering Major
Seoul, Korea
E-mail: mhr@bio.bmelab.co.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

Myung-Geol Pang
Chung-Ang University
College of Industrial Sciences
Department of Animal Science and Technology
72-1 Nae-Ri, Daeduk-Myun, Ansung-Si
Gyungki-Do 456-756, Korea
    and
Seoul National University
Medical Research Center
Institute of Reproductive Medicine and Population
28 Yungun-Dong, Jongro-Gu
Seoul 110-799, Korea

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

Abstract. We propose a new nonlinear matching measure for automatic analysis of the on-off type DNA microarray images in which the hybridized spots are detected by a template-matching method. The proposed measure is obtained by binary thresholding over the entire template region and taking the number of white pixels inside the spotted area. This measure is compared with the normalized covariance method in terms of classifying the ability to successfully locate markers. The proposed measure was evaluated for scanned images of human papillomavirus (HPV) DNA microarrays where locating markers is a critical issue because of the small number of spots. The targeting spots of HPV DNA chips are designed for genotyping twenty-two types of the human papillomavirus. The proposed measure is proven to give a more discriminative response, reducing the missed cases of successful marker location. The locating accuracy of the proposed method is also shown to have the same performance as that of the normalized covariance. © *2004 Society of Photo-Optical Instrumentation Engineers.*
[DOI: 10.1117/1.1691026]

1 Introduction

The automatic analysis of microarray images is one of the main issues for high-throughput screening using DNA microarrays. The analysis is commonly composed of two steps: one finds a spot position and the other measures the signal amplitude of each spot.[1,2] It is natural to utilize an ideal template to find the position of spots because prior knowledge of the microarray is available.[1–5] The predetermined template is tested for the best match over the hybridized image to locate a relative reference for the spot positions. This process can be trivial for microarrays that have a relatively large number of hybridized spots as in Refs. 1–4, 6, 7. However, it is not always easy to find the spot position when the number of spots is too small to obtain enough response for HPVDNA-Chip (Biomedlab Co., Korea). Because this microarray con-

tains only four marker spots that can be utilized for the reference position, we need to choose the template matching measure carefully.

The HPVDNAChip is designed for the detection of human papillomavirus (HPV) infection, which is one of the main causes of cervical cancer. Several groups reported on the clinical application and the evaluation of the HPVDNAChip. Kim et al.[8] examined the use of the chip, comparing it with the well-established detection system HC-II(Hybrid Capture II) of Digene Co. In particular, they evaluated its clinical efficacy for detecting HPV in cervical neoplastic lesions in 140 specimens. The chip was highly comparable to HC-II and provided useful information on viral genotype and multiple HPV infections in HPV-related cervical lesions. Cho et al.[9] performed a comparative study with Papanicolaou diagnosis for 685 cervicovaginal swabs. HPV types 16, 18, and 58 were confirmed to be major causative factors for cervical carcino-

Address all correspondence to Munho Ryu, Seoul National University, Graduate School, Seoul 110-744, South Korea. Tel: +82-2-747-9308; FAX: +82-2-747-4395; E-mail: mhr@bio.bmelab.co.kr
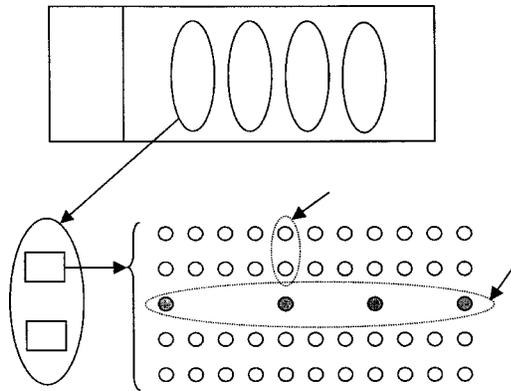
Fig. 1 The architecture of an HPV DNA microarray.

genesis, in descending order. An et al.[10] performed HPV genotyping in cervical specimens from 1983 patients and compared their cytological and histological diagnoses. They evaluated the quality of the HPVDNAChip method and identified HPV types related to cervical carcinoma and precancerous lesions. The chip provided a very sensitive method for detecting twenty-two HPV genotypes with reasonable sensitivity (96.0%) and a reasonable negative prediction value (96.9%), and it overcame the low sensitivity of cytological screening for detection of high-grade squamous intraepithelial lesions (HSIL) or carcinomas.

The HPVDNAChip has four chambers, one for each patient, as shown in Fig. 1. Each chamber is designed to have two identical sets of spots to increase the credibility of the diagnosis. A set of spots has four markers and twenty-two pairs of HPV type-specific oligonucleotide probes. The oligonucleotide probe of one HPV type is dotted twice, creating a pair of spots for the HPV type, and such probe pairs for twenty-two different HPV types are employed. The markers in each set, the oligonucleotide probe for human β-globin, are selected for the identification of probe positions as well as the verification of the hybridization success.

The target DNA of the sample is amplified by a polymerase chain reaction (PCR) and hybridized onto the chip. It randomly incorporates Cy5 during PCR amplification and visualizes the position of hybridization when the DNA chip is scanned. The DNA chip is an on-off type that can be read by simply finding the fluorescent spots with a scanner. The automatic analysis starts from scanning enough area to cover the specified set of spots. The initial scanning area can be determined from the accuracy of the arrayer and the scanner. However, the exact position of a set should be searched by locating the markers, which are always visualized if the hybridization is successful.

The template-matching method has been reported to give reasonable performance in locating markers for the on-off type microarray.[5] The normalized covariance measure works well if the template-matching method is combined with prior knowledge of the relative distances between marker spots. This is because the intrinsic local problem of the template-matching method can be overcome by utilizing the geometric relationship of the patterns.[11] However, it is not enough to distinguish the success of locating markers with the normalized covariance. Since it is normalized, not by the signal

power, but by the mixture of the signal and noise, it delivers a smaller response for smaller signal amplitudes. In the meantime, experts tend to distinguish the hybridized spots by the distribution of white pixels rather than the absolute intensity of each pixel. They consider a spot hybridized if its area is filled with a certain amount of pixels whose intensities are relatively higher than those of the background. The mechanism seems to reduce the signal variation greatly because the pixel intensities are mapped onto the binary state.

In order to simulate an expert's behavior, the template-matching method should be analyzed. A template and a target image can be in binary or gray-valued forms according to applications. However, a binary template is usually employed for detecting objects in gray-valued images because of the lack of information on image degradation in the imaging system and model.[12] Therefore, it is natural to construct a composite binary template so that the pixel value is "1" in the object area and "−1" in the background. It is also reasonable to make a template with the numbers of pixels equal in the object and the background. In that case, the covariance measure can be regarded as the mean difference of the intensity functions of the object and the background. This can be reinterpreted as the following two steps. First, the pixel intensity function of each area is mapped to its mean value. Next, the template-matching measure is the difference of the mean values.

We can deploy the matching measure to more closely simulate the expert's mechanism. Instead of choosing the mean value for the representative mapping of each area, we can take any order statistics, assuming that only the order relation of the intensity values is valid. This can be a rational choice when noise destroys the distance metric of the intensity values. For example, the gray-scale hit-or-miss transform chooses the first-order statistics for the mapping as in Ref. 13. Even if the order does not seem to be preserved, we might map the intensity values of the target image to the binary ones by a threshold. In this case, the covariance measure is the same as the hit-or-miss transform, where both the template and the target images are in the binary forms. Since it is also not easy to optimize the threshold to accommodate the bias and signal amplitude over the target image, this paper introduces a heuristic reasoning. If the image area coincides with the template pattern, then all the bright pixels in the template region exist in the corresponding object area. From this observation, we can select the threshold so that the number of the selected brighter pixels is same as that of the object's pixels.

This paper introduces a matching measure based on the preceding discussion. The target image region corresponding to the template is grouped into bright and dark pixels through the thresholding strategy mentioned earlier. After thresholding, the difference between the numbers of brighter pixels in the object area and the background can serve as a similarity measure that will deliver the maximum value only if the template exactly overlaps the object.

The proposed measure is integrated with the same strategy as in Ref. 5 to compare with the normalized covariance. The performance of both measures is evaluated for 1230 scanned microarray images of 615 patients, two images per patient. The analysis is focused on the criterion for failure to locate markers. The details of the proposed matching measure are

(a) an example scanned image
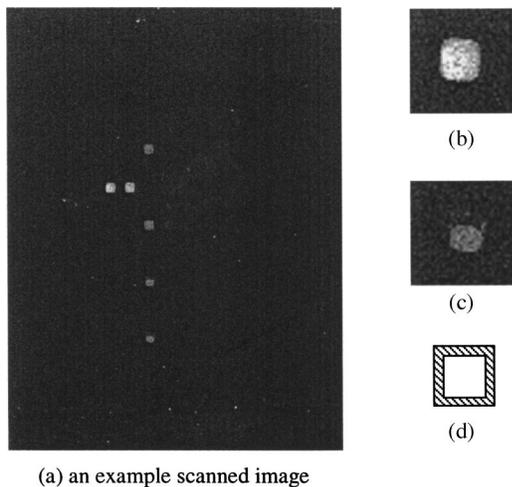
Fig. 2 (a) An example of a scanned image of a set in a chamber. (b) and (c) Enlarged images of the spots. (d) The proposed template.

described in Sec. 2, which also covers the marker locating strategy and the normalized covariance measure. In Sec. 3, both measures are compared; we give the conclusion and the discussion in the last section.

2 Locating Markers by Template Matching

Figure 2 shows a scanned image of a set in a chamber and the enlarged images of the spots. Figure 2(a) shows an image that is taken by scanning one of the predefined areas in a slide that are specified to cover each set of spots. The spots should be searched for because their positions are not guaranteed to be fixed in terms of image coordinates by the production of the microarrays. Figures 2(b) and 2(c) are enlarged views of some spots in image 2(a). The hybridized spot pattern might not simulate the square shape of the dotting pin shown in Fig. 2(d). Neither the average nor the distribution of intensity in the spot area is regular. Even though there might be several causes of these kinds of artifacts, the simple template shown in Fig. 2(d) might be reasonable because a strict model of the spot pattern is not currently available. The template intensity profile can be an average of a group of expert-selected spot-image patches as in Ref. 1 or a Gaussian function to accommodate the unknown spot size as in Ref. 2. However a unit function is proven to be enough for our application. Therefore we set the template intensity profile as follows. The intensity is 1 in the spot area (the white box) and -1 in the background area (the dashed region). The size of the background is set to be the same as that of the spot area to unbias the matching response where any hybridized spot does not exist.

We can use two additional ways to locate markers other than the size or shape of spots, as mentioned in Sec. 1. One is based on the fact that the markers are aligned vertically and the other on the fact that the relative distances between markers are known. Applying this information to the response image has been proven to be better than integrating it with the template itself.[5] In other words, the following procedure shows better performance than finding the maximum response position using a template with a global background as shown in Fig. 3.

Fig. 3 Template with global background.

2.1 Marker-Locating Procedure

1. Calculate the template matching response $m(k,j)$ for the entire search area using the template for a single spot.
2. Find the position (k,j) where the averaged measure over the relative positions of the markers, $\bar{m}(k,j) = 1/4\{m(k,j)+m(k+dx1,j)+m(k+dx2,j)+m(k+dx3,j)\}$, is maximum.
3. Test the success of locating markers.

2.2 End Procedure

In this procedure, $dx1$, $dx2$, and $dx3$ are the relative distances from the topmost marker, respectively.

In order to verify the success of locating a marker, it is preferable to investigate the distribution of the averaged measure. In other words, if there are some pixel positions where the responses are close to the maximum, then the position at maximum response is not feasible as an indication of successful marker location. To adapt for the variation of the response over the images, we calculated the standard deviation of the responses all over the image pixels and divided the maximum response by the standard deviation to test the success of the location process. We denoted the ratio as the maximum-to-sigma ratio (MSR) as follows:

$$r_{\bar{m}} = \frac{\bar{m}_{\max}}{\sigma_{\bar{m}}}, \tag{1}$$

where $\bar{m}_{\max}$ and $\sigma_{\bar{m}}$ are the maximum and the standard deviation of the responses. Now, step 3 of the procedure given in Sec. 2.1 is replaced by the following procedure: if $r_{\bar{m}}$ is greater than the failure threshold, the marker is successfully located; if it is less than the failure threshold, the marker has not been located.

The possible linear matching measures can be the covariance (C) and its normalized form (NC) as follows:

$$C(i,j) = \sum_{k,l \in \text{Template}} T(k,l)I(i+k,j+l),$$

$$NC(i,j) = \frac{\sum_{k,l \in \text{Template}} T(k,l)I(i+k,j+l)}{\sigma_T \sigma_I(i,j)}, \tag{2}$$

where σ_T and $\sigma_I(i,j)$ are the standard deviation of the template function values and the intensities of the image in the template region, respectively. Note that the covariance C is same as the mean difference between the image intensities inside the spot area and background because $T(k,l)$ has 1 in the spot area and -1 in the background. It is not adequate for our application because it is possible for a probe spot to give the maximum response, owing to the high intensity variation of the spots. In particular, any multiple hybridized probes often deliver a greater matching response than the markers.
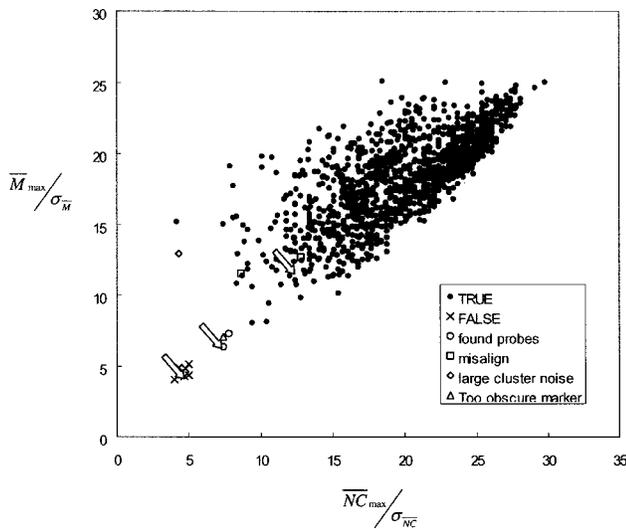
Fig. 4 Scatter plot of the MSR of the averaged normalized covariance and the average of the proposed measure. Each data point was taken from an image. The *x*-axis denotes the MSR of the normalized covariance and the *y*-axis that of the proposed measure.

The normalized covariance also can be regarded as the mean intensity difference between the spot and background area, but normalized by the signal power of the entire template region. Although the normalization contributes to the reduction of the signal intensity variation to some degree, we can make the measure more independent of the signal variation using the following equation:

$$M'(i,j) = No(\{p:p \in \text{spot area}, i(p) > \text{threshold}\})$$
$$- No(\{p:p \in \text{background}, i(p) > \text{threshold}\}),$$
$$\tag{3}$$
$$No(\{p:p \in \text{template}, i(p) > \text{threshold}\})$$
$$= No(\{p:p \in \text{object}\}),$$

where $No(\cdot)$, p and $i(p)$ are the number of elements of a set, a pixel, and the intensity at the pixel, respectively. The threshold is taken in this manner that the number of pixels greater than the threshold is the same as that of the object. The above measure indicates the amount of filling of the spot area with the relatively bright pixels and simulates the way that experts investigate the hybridized spots. Actually, the number of bright pixels only in the spot area is a sufficient and necessary quantity for the above measure because the number of the pixels is fixed as the template size, resulting in the following equation:

$$M(i,j) = No(\{p:p \in \text{spot area}, i(p) > \text{threshold}\}). \tag{4}$$

In Sec. 3 the proposed measure of Eq. (4) is evaluated for locating markers compared with using the normalized covariance of Eq. (2). To compare both measures, two aspects must be considered. One is that the failure to locate markers should be defined. The other is to predict the possible failure thresholds that are selected by the actual examiners according to their quality control strategy. In this paper, the failure was decided subjectively and the MSR of the failed images served as the measure of the failure threshold. Note that even though
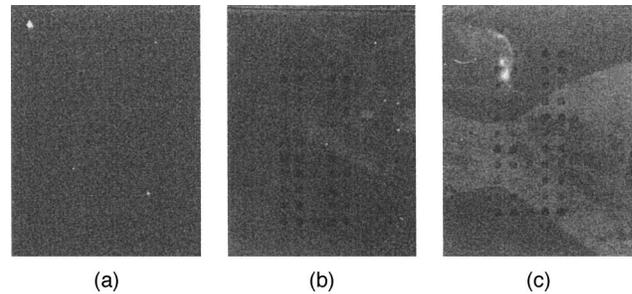
Fig. 5 Example of images that have no hybridized markers on probes. Location of markers failed with both measures. All the images are amplified 16 times and gamma-corrected for display purposes.

the markers are successfully found, it is better to abandon the procedure if the MSR is less than the threshold. If a measure is able to classify the success of locating markers more strictly, then the number of the success cases below the failure threshold will be smaller. A Gaussian distribution assumption was employed to handle the possible failure cases that were not explored in this paper. Then a similar approach to the receiver operating characteristic (ROC) was utilized to present the results of the comparison of the measures.

3 Experimental Results

The MSRs of both measures are scatterplotted in Fig. 4 for 1230 scanned images. The *x*-axis of Fig. 4 denotes the MSR of the normalized covariance and the *y*-axis that of the proposed measure. The figure shows that both measures give highly correlated results. The black circles, labeled "TRUE," show that the markers are successfully located with both measures. The other symbols were chosen for cases where at least one of the measures provided the wrong marker position. The data points of the symbol "×" are from the images with neither distinguishable markers nor probes like the images in Fig. 5. We were not able to find any distinct markers and probes in spite of 16-times amplification and gamma correction of the images.

When there were no distinct markers but some hybridized probes as shown in Fig. 6, both MSRs were somewhat higher than those without any hybridized spots (the two white circles in Fig. 4). In that case, the normalized covariance and the
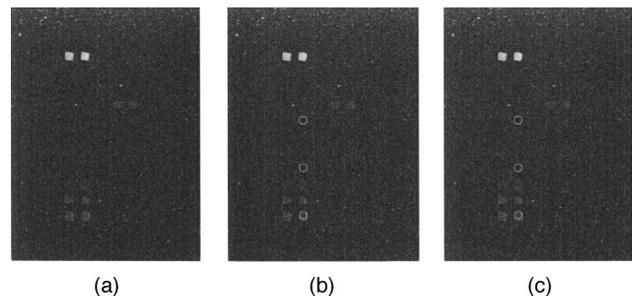
Fig. 6 Marker-locating results for an image with some hybridized probes and no distinct markers. The markers are falsely located at the probes just left of the markers using both measures. (a) The original image. (b) The result with the normalized covariance. (c) The result with the proposed measure. All of the images are amplified 8 times and gamma-corrected.
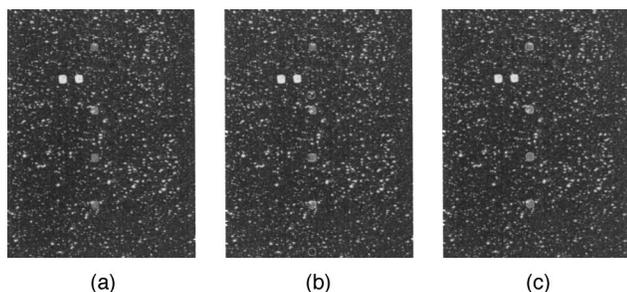
(a)   (b)   (c)
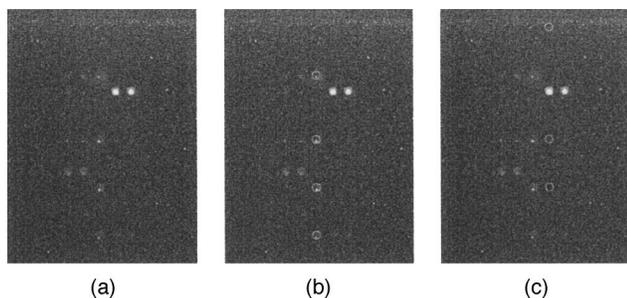
Fig. 7 Marker-locating results for an image in which the hybridized markers are too obscure. The marker was located correctly with the normalized covariance. (a) The original image. (b) The result with the normalized covariance. (c) The result with the proposed measure. All of the images are amplified 16 times and gamma-corrected.

(a)   (b)   (c)

Fig. 9 Marker-locating results for an image in which the markers are misaligned. The markers were located at a fairly reasonable position with the proposed measure (a) The original image. (b) The result with the normalized covariance. (c) The result with the proposed measure. All the images are amplified 4 times and gamma-corrected.

proposed measure found the probe spots instead of the markers, as shown in Figs. 6(b) and 6(c). In these figures, four aligned circles depict the estimated location of the markers.
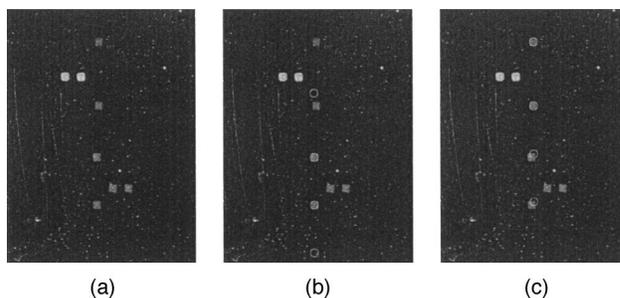
While the markers were barely distinguishable, as shown in Fig. 7, the MSRs had a value similar to that of the previous examples (the triangle symbol between the white circles in Fig. 4). Even though the normalized covariance seems to give the correct marker position, as in the center image in Fig. 7, it should be reported as a failure because the greater value of the MSR can be delivered at the false position, as in the example in Fig. 6.

Since the noise was highly clustered as in Fig. 8, the proposed measure gave the correct results while the normalized covariance did not. It was not necessary to take into account the failure threshold because the MSRs of both measures provided reasonable values (the diamond symbol in Fig. 4). They were either large or small enough to be classified exactly. Figure 8 also shows that the proposed measure is more suitable for images with highly clustered noise.

There were two unexpected cases like the example shown in Fig. 9. The markers are slightly misaligned in a vertical direction, which might be caused by the malfunction of the dotter. In that case, both MSRs were relatively large (the two rectangles in Fig. 4). Even if this misalignment were corrected by improving the quality control of the dotter, it is worth regarding the measured values as indicating failure thresholds in the performance comparison. It is interesting to note that
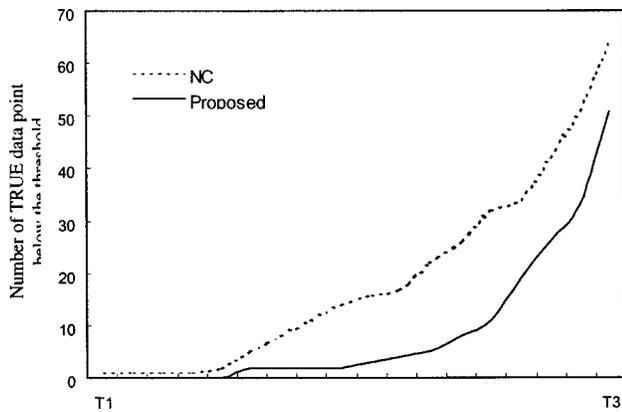
the proposed measure locate a position fairly close to the markers, in contrast to the normalized covariance.

We selected three failure thresholds to compare the proposed measure with the normalized covariance. They are shown in Fig. 4 as the arrows labeled 1 to 3 by grouping all the failures in three classes. The first group was for the failure of both measures ("×" symbols in Fig. 4), the second was for cases with only hybridized probes and markers that were too obscure because in both cases the measures delivered similar values (the white cicles and the triangle in Fig. 4). The third was for the misalignment cases (the rectangles in Fig. 4). Table 1 shows the number of "TRUE" data points below each failure threshold that was taken as the maximum MSR of each group. It also shows the percent probability that "TRUE" data points are below the failure thresholds. The probability was taken under the assumption that the MSRs of "TRUE" data points were distributed in the Gaussian form.

Table 1 shows that the proposed measure is able to classify success more strictly than the normalized covariance for all the failure cases investigated in this study. To predict the other possible failure cases, it is worth testing the increasing trend in the number of "TRUE" data points, varying the failure threshold from the first to the third threshold. If the MSR is below the first threshold, we can assume that marker location has failed. However, it currently is not clear whether the third threshold can be employed as the decision boundary for success. That is because the marker alignment can be controlled by the chip manufacturer's quality assurance. However, we can expect that the failure threshold might exist between the

(a)   (b)   (c)

Fig. 8 Marker-locating results for an image with highly clustered noise. The markers were located correctly with the proposed measure. (a) The original image. (b) The result with the normalized covariance. (c) The result with the proposed measure. All the images are amplified 4 times and gamma-corrected.

Table 1 Number of TRUE data points below each threshold shown in Fig. 4. (1218 TRUE data points).

Failure threshold position	Number of TRUE data points below the threshold		Percentage below the threshold under the Gaussian assumption	
	NC	Proposed	NC	Proposed
1	1	0	0.03	0.00
2	2	0	0.23	0.01
3	64	51	4.59	3.08

Fig. 10 The increase in the number of success cases below the failure threshold, which varies from the first failure threshold (T1) to the third failure threshold (T3).

first and third thresholds. Figure 10 depicts the variation in the "TRUE" data points below the threshold according to the failure threshold. The solid curve labeled "Proposed" is for the proposed measure and the dashed one is for the normalized covariance (labeled "NC"). The solid curve is more downwardly concave than the dashed one. It shows that we can have a better advantage with the proposed measure than with the normalized covariance whenever the failure threshold is established between the first and the third thresholds.

In order to compare the accuracy of locating a marker with both measures, we compared their estimated positions and selected cases where the distances between them were greater than two pixels. There were fifteen such images. We located the markers for them manually and calculated the distances between the resultant positions and the markers using both measures. Table 2 shows the mean and standard deviation of the distances for the measures. In the table, "NC" and "Proposed" indicate the normalized covariance and the proposed measure, respectively. Even though both mean and standard deviation are a litte bit smaller with the proposed measure, we concluded that the accuracy of locating markers was similar with both measures.

4 Discussion

This paper has presented a nonlinear matching measure that uses estimates of the number of white pixels in the spot area. This method was combined with a marker location strategy that combined template matching and integrating the knowl-

Table 2 Mean and standard deviation of the distances between the manually determined position and that found by each method.

	NC	Proposed
Mean (pixels)	2.40	2.39
Std (pixels) Dev.	1.21	1.06

edge of the relative distances between markers. The same strategy with a normalized covariance was employed for verifying the proposed measure.

A total of 1230 images of hybridized HPV microarrays were used to evaluate the marker-locating performance of the proposed measure and the normalized covariance. The failure cases were analyzed to define failure thresholds that indicated the decision boundary for success in locating markers. The performance criterion was how small the number of the success cases below the thresholds was. That was because if the measure was able to classify success more strictly, the number would be smaller. The proposed measure performed better than the normalized covariance for all the failure cases presented. It also promises to work better for the possible failures that were not found in these experiments. The location accuracy was also analyzed and showed almost the same performance.

When the markers are misaligned vertically, both measures deliver relatively high values because the marker location strategy assumes that the markers are aligned vertically. If these cases are viewed as failures, the threshold will become so high that we should abandon many successfully located markers (4.59% with the normalized covariance and 3.08% with the proposed measure). The investigation described here can be used in designing quality assurance guidelines in chip manufacture. The framework given in this paper can also be utilized to guide the other design issues for these kinds of chips.

References

1. N. Brandle, H. Bischof, and H. Lapp, "A generic and robust approach for the analysis of spot array images," *Proc. SPIE* **4266**, 1–12 (2001).
2. L. M. Kegelmeyer, L. Tomsascik-Cheeseman, M. S. Burnett, P. van Hummelen, and A. J. Wyrobek, "A groundtruth approach to accurate quantitation of fluorescence microarrays," *Proc. SPIE* **4266**, 35–45 (2001).
3. Z. Z. Zhou, J. A. Stein, and Q. Z. Ji, "GLEAMS: a novel approach to high throughput genetic micro-array image capture and analysis," *Proc. SPIE* **4266**, 13–23 (2001).
4. T. Bergermann, F. Quiaoit, J. Delrow, and L. P. Zhao, "Statistical issues in signal extraction from microarrays," *Proc. SPIE* **4266**, 24–33 (2001).
5. J. D. Kim, S. K. Kim, J. S. Cho, and J. Kim, "Knowledge-based image processing for on-off type DNA microarray," *Proc. SPIE* **4623**, 38–46 (2002).
6. Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.* **2**(4), 364–374 (Oct. 1997).
7. G. Delenstarr, H. Cattell, C. Chen, A. Dorsel, R. Kincaid, K. Nguyen, N. Sampas, S. Schidel, K. Shannon, A. Tu, and P. Wolber, "Estimation of the confidence limits of oligonucleotide arraybased measurements of differential expression," *Proc. SPIE* **4266-17**, 20–26 (Jan. 2001).
8. Chan Joo Kim, Jeongmi Kim Jeong, Misun Park, Tae Shin Park, Tae Chul Park, Sung Eun Namkoong, and Jong Sup Park, "HPV oligonucleotide microarray-based detection of HPV genotypes in cervical neoplastic lesions," *Gynecol. Oncol.* **89**, 210–217 (2003).
9. Nam Hoon Cho, Hee Jung An, Jeongmi Kim Jeong, Suki Kang, Jae Wook Kim, Young Tae Kim, and Tchan Kyu Park, "Genotyping of 22 human papillomavirus types by DNA chip in Korean women with

cytologic diagnosis," *Am. J. Obstet. Gynecol.* **188**, 56–62 (2003).

10. Hee Jung An, Nam Hoon Cho, Sun Young Lee, In Ho Kim, Chan Lee, Seung Jo Kim, Mi Sook Mun, Se Hyun Kim, and Jeongmi Kim Jeong, "Correlation of cervical carcinoma and precancerous lesions with human papillomavirus (HPV) genotypes detected with the HPV DNA chip microarray method," *Cancer* **97**, 1672–1680 (2003).

11. Hiroshi Sako, Masakazu Fujio, and Naohiro Furukawa, "The constel-lation matching and its application," *Intern. Conf. on Image Processing*, pp. 790–793 (2001).

12. J. C. Russ, *The Image Processing Handbook*, CRC Press, Boca Raton, FL (1999).

13. M. Khosravi and R. W. Schafer, "Template matching based on a gray-scale hit-or-miss transform," *IEEE Trans. Image Process.* **5**(6), 1060–1066 (June 1996).