32nd Annual International Conference of the IEEE EMBS
Buenos Aires, Argentina, August 31 - September 4, 2010

Automatic analysis of macroarrays images

C.M.R. Caridade, A.R.S. Marcal, T. Mendonça, P. Albuquerque, M.V. Mendes, F. Tavares

Abstract— **The analysis of dot blot (macroarray) images is currently based on the human identification of positive/negative dots, which is a subjective and time consuming process. This paper presents a system for the automatic analysis of dot blot images, using a pre-defined grid of markers, including a number of ON and OFF controls. The geometric deformations of the input image are corrected, and the individual markers detected, both tasks fully automatically. Based on a previous training stage, the probability for each marker to be ON is established. This information is provided together with quality parameters for training, noise and classification, allowing for a fully automatic evaluation of a dot blot image.**

I. INTRODUCTION

The fast and reliable detection of bacteria from environmental samples is of upmost importance in diagnostic microbiology. In the last years, DNA-based methods of bacteria detection have been increasingly acknowledged as trustworthy alternatives to circumvent the limitations of traditional culture-based detection approaches focused on biochemical, serological and pathogenicity tests. In fact, since DNA-loci rather than organisms are detected, molecular detection methods are unbiased by the limitations of culturability. DNA-based methods of bacterial detection rely mainly on two key factors: the selection of taxa-specific DNA signatures [1], [2] and a sensitive molecular detection technique [2]. Array-based hybridization assays, such as microarrays and macroarrays, allow the analysis of numerous molecular markers simultaneously, increasing the detection reliability. Currently, macroarrays provide a better cost-benefit for routine analysis than the costly microarray platforms, as emphasized by various examples for detection/identification of several microorganisms, including bacterial potato pathogens [3], phytopathogenic *Pseudomonas* [1], *Lactobacillus* species [4], *Pythium* species [5] and *Aeromonas* spp. [6], among others. In dot blot macroarrays, an ideal positive dot is defined as a dark area in a light gray background, whereas an ideal negative dot is undistinguishable from the background. However, the different hybridization molecular affinities between the labeled probe and the spotted marker, and the heterogeneity of the background noise, results in a grayscale image from which is not always easy to distinguish a positive from a negative dot. At present, the analysis of dot blot images is an operator dependent decision, which is subjective and therefore a key drawback to further implement macroarrays for microbial detection. There is thus a great interest in developing an automatic process for detection, analysis and classification of dot blot images, in order to allow the fully automatic processing of dot blot images and consequently enhance the potential of macroarrays for routine microbial detection.

This work describes an application for the fully automatic processing of macroarray images, using dot blot prototype containing seventeen DNA signatures for the detection of the plant pathogenic *Ralstonia solanacearum* [7].

II. METHODOLOGY

The algorithm developed receives as input a digital image and assumes prior knowledge of the grid size (number of dots per line and column). The process can be divided in five stages, illustrated schematically in figure 1: grid detection, adaptive training, noise evaluation, classification and confidence estimation. The grid detection process is described in detail in [8].

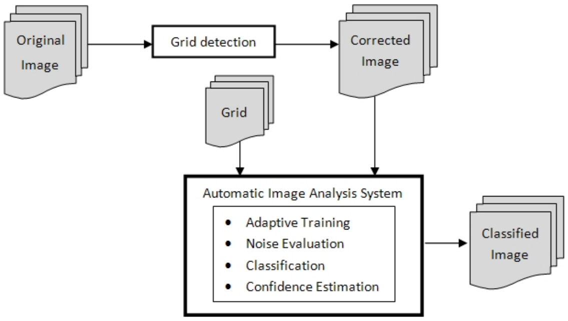


Fig. 1. Schematic representation of the automatic image analysis system.

A. Dot blot hybridization assays

For Dot blots, 100 *ng* of each heat-denatured DNA signature was spotted in a Nylon membrane, using a Bio-Dot apparatus (Bio-rad, Hercules, CA). Each dot blot had a pre-defined grid of evenly spaced dots, in this case, 48 dots arranged in 8 rows and 6 columns. Figure 2 shows the 8 test images used ($T1$, $T2$, $C1$, $C2$, $C3$, $C4$, $C5$, $C6$).

Total DNA, from different bacteria, was labeled with Digoxigenin using the DIG-High Prime labeling kit (Roche, Basel, Switzerland). Hybridization was carried out overnight at

C.M.R. Caridade is with Instituto Superior de Engenharia de Coimbra, R. Pedro Nunes, Qt. Nora, Coimbra, Portugal and Faculdade de Ciências, Univ. Porto, DM, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal caridade@isec.pt

A.R.S. Marcal is with CICGE and CMUP, Faculdade de Ciências, Univ. Porto, DM, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal andre.marcal@fc.up.pt

T. Mendonça is with Faculdade de Ciências, Univ. Porto, DM, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal tmendo@fc.up.pt

P. Albuquerque and F. Tavares are with IBMC – Instituto de Biologia Molecular e Celular, Faculdade de Ciências, Departamento de Biologia, Univ. Porto psilva@ibmc.up.pt FTavares@ibmc.up.pt

M.V. Mendes is with IBMC – Instituto de Biologia Molecular e Celular, Univ. Porto mvm@ibmc.up.pt

978-1-4244-4124-2/10/$25.00 ©2010 IEEE

$68°C$, using a final concentration of 100 ng/mL Digoxigenin-labeled DNA. DIG-labeled nucleic acids were detected by chemiluminescency using X-ray films (GE healthcare). The dot blot images were acquired with a GS-800 densitometer (Bio-rad, Hercules, CA), producing grayscale images with 1100 by 820 pixels.

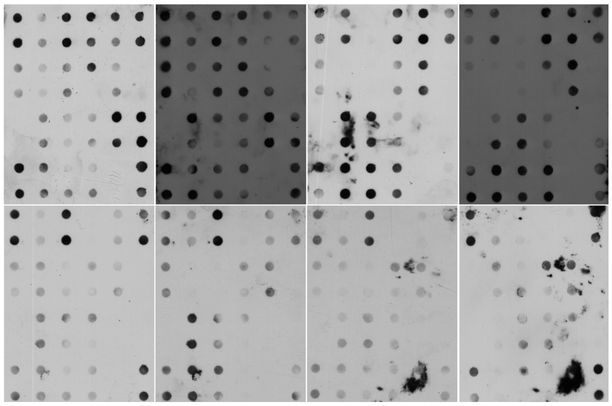


Fig. 2. Dot blot images used (from left to right): T1, T2, C1, C2 (top); C3, C4, C5 C6 (bottom).

B. Grid Detection

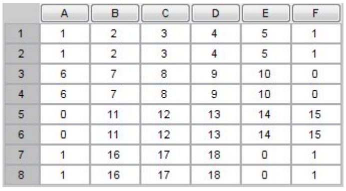
Initially, the very dark dots are identify in the original grayscale image. A binary image is obtained by thresholding the grayscale image with the Otsu method [9]. The noise in the binary image is reduced using the morphological operation opening (erosion followed by dilation) with a circular structuring element of 5 pixel radius [10]. The holes present in the original binary image are filled by the morphological reconstruction [11]. After this initial processing, the center of mass of each object is computed and the objects with less than 15 pixel radius are eliminated. The size of a circular marker diameter is $3mm$, which corresponds to approximately 56 pixels at the scanning resolution used.

The set of markers identified are used to calculate the directions between all marker pairs. The two main directions are orthogonal, so the orientation of the grid is obtained as the most frequent angle in the range $[0°, 90°[$, using a bin search of 2°. The original image is then rotated by this angle. More details about this procedure are available in [8].

The correspondence between the markers detected and the pre-defined (reference) grid is established by assigning a position in the grid for each marker. The reference grid is thus mapped into the image, correcting the effects of rotation, scale and translation [8]. Once the image to grid mapping is established, a local thresholding is performed on the original image, to detect all markers present in the image, even the lighter ones that were not detected in the initial thresholding.

C. Adaptive Training

In the dot blot images, each marker (i) is associated with a type (T_i), corresponding to a specific DNA signature. The control markers are represented by 0 and 1 ($T_{OFF} = 0$, $T_{ON} = 1$) and the other markers by integers larger than 1 ($2, ...n$). Seventeen different signatures (types 2-18) were considered for the dot blot. Figure 3 shows the marker's type matrix for the test images used. The images have six control markers OFF, eight control markers ON and two duplicate markers for all other seventeen types ($T_i : i = 2, ..., 18$). Two circles are established for each marker (figure 4). One (internal) circle with the same radius of the marker, and another (external) with a radius of one third of the minimum distance between markers. Pixel intensity measurements are computed for all markers and for the image background. For each marker, the average intensity of the pixels inside the marker (I_i) and the average intensity of pixels in the local background I_{back} (pixels in the outer ring) are computed. The normalized intensity for a marker ($\bar{I}_i$) is obtained by (1).



	A	B	C	D	E	F
1	1	2	3	4	5	1
2	1	2	3	4	5	1
3	6	7	8	9	10	0
4	6	7	8	9	10	0
5	0	11	12	13	14	15
6	0	11	12	13	14	15
7	1	16	17	18	0	1
8	1	16	17	18	0	1

Fig. 3. Marker's type matrix for the dot blot images used.

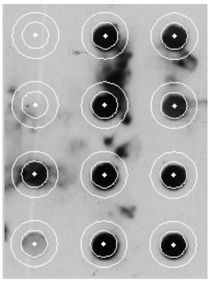


Fig. 4. Example of the circular areas (internal and external) established for each marker.

$$\bar{I}_i = 1 - \frac{I_i}{I_{back}} \tag{1}$$

For regular markers (markers that are not ON/OFF controls), the average normalized intensity ($\overline{m}_{I_i}$) is computed for each type, and compared with the average normalized intensity for all OFF markers ($\overline{m}_{OFF}$). The difference between these two values ($\Delta = \overline{m}_{I_i} - \overline{m}_{OFF}$) should be positive, with a value between 0.05 and 0.10. A probability function $P(\bar{I}_i)$ is established using $\overline{m}_{I_i}$ as the top level and $\overline{m}_{OFF}$ as the bottom level, using equation (2). The shape of this probability function is illustrated in figure 5. For values of normalized intensity below $\overline{m}_{OFF}$ the probability of the marker being ON is 0 and for values above $\overline{m}_{I_i}$, the probability is 1. For values between $\overline{m}_{OFF}$ and $\overline{m}_{I_i}$ the probability values, between 0 and 1, are given by (2).

$$P(\bar{I}_i) = \begin{cases} 0 & \text{if } \bar{I}_i < \overline{m}_{OFF} \\ \frac{\arg(\tan(u)) - \arg(\tan(-2))}{\arg(\tan(8)) - \arg(\tan(-2))} & \text{if } \overline{m}_{OFF} \leq \bar{I}_i \leq \overline{m}_{I_i} \\ 1 & \bar{I}_i > \overline{m}_{I_i} \end{cases} \tag{2}$$

where $u = -2 + \frac{10\times(\bar{I}_i - \overline{m}_{OFF})}{\Delta}$.

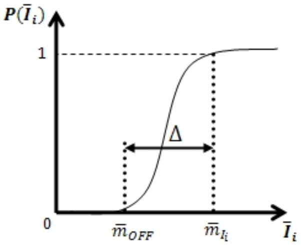


Fig. 5. General probability function established for each marker type.

The adaptive training is performed using a number of images, and provides probability functions $P(\bar{I}_i)$ adapted for each marker type. In this work, the first two images presented in figure 2 (T1 and T2) were used as training images, with the matrix type as presented in figure 3.

D. Noise evaluation

The dot blot images may have considerable noisy areas, which limits the ability to evaluate some markers. This can be observed for example in images $C1$, $C5$ and $C6$ (figure 2). The estimation of the noise level of each marker is therefore an important complementary information required to properly interpret and classify a dot blot image.

The level of noise contamination is computed for each marker (i) using the information from its neighboring pixels. The area around the marker, defined as the outer ring, is divided in eight equal sectors. Figure 6 (left) illustrates this process. The difference between the average intensity of pixels within the sector and the average intensity of the background ($\triangle_{noise}$) is used as a noise estimator. The range of potencial values for $\triangle_{noise}$ is $[0,255]$. Three different noise levels are established for a sector. For $\triangle_{noise}$ below 30, the sector is considered to have low level of noise, with a level 0 assigned. For $\triangle_{noise}$ between 30 and 100, a sector is considered to have moderate noise, and the level 1 is assigned to that sector. For $\triangle_{noise}$ above 100 the sector is considered to have hight level of noise and the level assigned is 2. The overall noise level for the marker is computed as the sum of the noise levels of the eight sectors. In figure 6 (right) two examples of markers from the same image ($C1$) are presented. The top marker in figure 6 ($3rd$ line, $1st$ column in figure 7) has two sectors with level 1 noise. The noise level of this marker is thus 2 ($2 = 0\times 6 + 1\times 2$). The bottom marker in figure 6 ($1st$ line, $2nd$ column in figure 7) has one sector with level 1 noise and two sectors with level 2 noise. The overall noise level for this marker is 5 ($5 = 0\times 5 + 1\times 1 + 2\times 2$), which is a rather noisy marker. Figure 7 (left) presents the noise estimation for the twelve markers of the bottom left section of image $C1$.

E. Classification

The classification stage is done assuming that at least one image is available to train the classifier. Each marker is classified as being ON or OFF with a probability value.

The classification probability for a marker is calculated using the information obtained in the adaptive training stage. The probability of a marker being ON is calculated by equation 2, using for $\overline{m}_{OFF}$ and $\overline{m}_{I_i}$ the values obtained in the adaptive training stage for that marker type. If the value of $P(\bar{I}_i)$ is below 0.5, then the marker is labeled as OFF, with a probability of being OFF of $1 - P(\bar{I}_i)$.

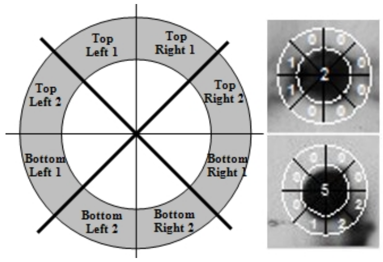


Fig. 6. Illustration of the noise estimation of process, based on 8 sectors, and two examples of markers with noise level of 2 (top) and 5 (bottom).

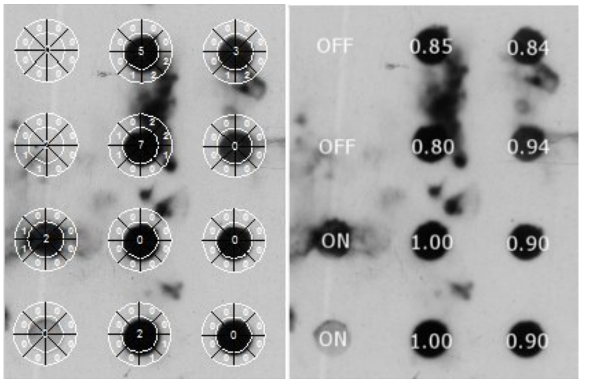


Fig. 7. Noise level estimation (left) and quality parameter Q (right) for the 12 markers of the bottom left section of image $C1$.

F. Confidence estimation

Three parameters are used to produce a confidence estimation: a training confidence parameter, a noise confidence parameter and a classification confidence parameter. All these parameters use the range 0 to 1, with a value of 1 corresponding to an optimum result. The training confidence parameter (q_t) is established for each marker type, as the average of the probabilities of all training markers used for that type. The noise confidence parameter (q_n) is obtained from the noise level, with $q_n = 1$ for noise levels of 0 or 1; and q_n=0 for a noise level of 11 or greater. A linear interpolation is used for noise levels between 2 and 10. The classification confidence parameter (q_c) is defined by equation (3). The probability of the marker being ON (3a) or OFF (3b) is a value between 0.5 and 1.

$$q_c = \begin{cases} 2P(\bar{I}_j) - 1 & \text{if } P(\bar{I}_j) \geq 0.5 \\ 1 - 2P(\bar{I}_j) & \text{if } P(\bar{I}_j) < 0.5 \end{cases} \tag{3}$$

The overall confidence estimation parameter for a regular marker is computed using (4). For control markers (ON or

TABLE I

CLASSIFICATION RESULTS FOR A SECTION OF IMAGE $C1$ (FIGURE 7).

(line,column)	*Status*	Q	q_t	q_n	q_c
(1,1)	C-OFF	1.00	X	1.00	1.00
(1,2)	ON	0.85	1.00	0.70	1.00
(1,3)	ON	0.84	0.87	0.80	1.00
(2,1)	C-OFF	0.90	X	0.90	1.00
(2,2)	ON	0.80	1.00	0.60	1.00
(2,3)	ON	0.94	0.87	1.00	1.00
(3,1)	C-ON	0.90	X	0.90	1.00
(3,2)	ON	1.00	1.00	1.00	1.00
(3,3)	ON	0.90	0.79	1.00	1.00
(4,1)	C-ON	1.00	X	1.00	1.00
(4,2)	ON	1.00	1.00	1.00	1.00
(4,3)	ON	0.90	0.79	1.00	1.00

OFF) the confidence estimation is simply obtained as $Q = q_c \times q_n$, as there is no training for these markers.

$$Q = q_c \times (\frac{q_t + q_n}{2}) \tag{4}$$

III. RESULTS

The method proposed for the automatic dot blot image analysis was evaluated using 8 test images (figure 2). The first two images ($T1$, $T2$) were used for the training stage and the other six ($C1$-$C6$) for classification. These images were obtained with the marker's type matrix presented in figure 3. The automatic processing corrects the image rotation, identifies the visible markers (location of their centers and radius) and provides an estimate of the noise level present in each marker. It uses the training images to calculate the probability function (of a marker being ON/OFF), adjusted for each marker type, and classifies each observed marker accordingly. Together with the final result for each marker (ON or OFF), a confidence estimation is also provided, based on the quality of training, the level of noise and the classification process.
An example of a classification result for the 12 markers of the bottom left section of image $C1$ is presented in figure 7 (right). In this section, there are 4 control markers (2 ON and 2 OFF) and all regular markers were classified as ON, with a probability of 1.00. The values presented in figure 7 (right) are the overall quality parameter Q, which provides an indicator of the degree of certainly of the label assignment.

As a further illustration of the final results produced, table I shows the status and confidence parameters for the 12 markers presented in figure 7. A total of 2 markers are labeled OFF (the 2 control markers C-OFF) and 10 are labeled ON (including 2 control markers C-ON). The markers C-OFF and C-ON were used as controls in the training stage, thus they do not have training confidence parameter values (q_t), while the DNA-signature markers had values of q_t close to 1. Concerning noise evaluation, 7 markers (1 OFF and 6 ON) had a high noise confidence parameter (low level of noise $q_n = 1$), 4 had a moderate value for the noise confidence parameter ($q_n = 0.7, 0.8, 0.9$) and 1 had a low level of the confidence parameter (high level of noise, $q_n = 0.6$). The classification confidence parameter (q_c) was 1 for all 12 markers. Finally, the overall quality (Q) of these markers had values of 0.80 or above. The overall quality of markers (2,2) and (1,2) was influenced by the noise confidence parameter, the overall quality of markers (2,3), (3,3) and (4,3) was influenced by the training confidence parameter, while the overall quality of marker (1,3) was influenced both by the noise and the training confidence parameters.

IV. CONCLUSIONS

The proposed method performs a fully automatic analysis of dot blot images with a pre-defined structure (grid size and location of control markers), based on a number of training images (at least one). The system provides not only the status of each marker (ON or OFF), but also three quality parameters, related to the training and classification stages, and to the noise present in the observed image. One limitation of dot blot images is the fact that the status of same marks is sometimes unclear, even for an experience user. However, the application of the proposed image analysis system will increase the reliability of macroarrays used for bacteria detection, and is therefore an important contribution in diagnostic microbiology.

REFERENCES

[1] J. Vieira, M.V. Mendes, P. Albuquerque, P. Moradas-Ferreira, and F. Tavares, "A novel approach for the identification of bacterial taxa-specific molecular markers", *Letters in Applied Microbiology*, vol. 44, 2007, pp. 506-512.

[2] P. Albuquerque, M.V. Mendes, C.L. Santos, Moradas-Ferreira, and F.P. Tavares, "DNA signature–based approaches for bacterial detection and identification", *Science of the Total Environment*, vol. 407(12), 2009, pp. 3641-3651.

[3] A. Fessehaie, S.H. De Boer, and C.A. Levesque, "An oligonucleotide array for the identification and differentiation of bacteria pathogenic on potato", *Phytopathology*, vol. 93(3), 2003, pp. 262-269.

[4] P. Poltronieri, O.F. D'Urso, G.Blaiotta and M. Morea, "DNA Arrays and Membrane Hybridization Methods for Screening of Six *Lactobacillus* Species Common in Food Products", *Food Analytical Methods*, vol. 1(3), 2008, pp. 171-180.

[5] J.T. Tambong, A.W.A.M. de Cock, N.A. Tinker and C.A. Lévesque, "Oligonucleotide array for identification and detection of *Pythium* species", *Applied and Environmental Microbiology*, vol. 72(4), 2006, pp. 2691-2706.

[6] R. khushiramani, S.K. Girisha, and I. Karunasagar, "Evaluation of a digoxigenin-labelled probe for detection of *Aeromonas* spp", *Letters in Applied Microbiology*, vol. 48(3), 2009, pp. 383-385.

[7] P. Albuquerque, F. Tavares, P. Moradas-Ferreira, and M.V. Mendes, "Twenty molecular markers for an efficient detection of R*alstonia solanacearum*", *in Proceedings of the 2nd FEMS Congress, 4-6 July 2006, Madrid, Spain*. FEMS, 2006.

[8] A.R.S. Marçal, C.M.R. Caridade, P. Albuquerque, M.V. Mendes, F. Tavares, "Automatic Detection of Molecular markers in Digital Images", *Conf Proc IEEE Eng Med Biol Soc*, vol. 1, 2009, pp. 6710-6713.

[9] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. on Systems Man Cybernetics*, vol. 9(1), 2002, pp. 62-69.

[10] R.C. Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice Hall, Upper Saddle River, New Jersey, 3rd edition, 2008.

[11] The MathWorks, *Using Matlab, Version 6.5.*, The MathWorks, Inc. Natick. MA, 2002.